Published in final edited form as:

Nature. 2019 July ; 571(7765): 343–348. doi:10.1038/s41586-019-1384-z.

Holistic Prediction of Enantioselectivity in Asymmetric Catalysis

Jolene P. Reid and **Matthew S. Sigman***

Department of Chemistry, University of Utah, 315 South 1400 East, Salt Lake City, Utah 84112, United States

Summary:

When faced with unfamiliar reaction space, synthetic chemists typically apply reported conditions (reagents, catalyst, solvent, additives) from closely-related reactions to new substrate types. Unfortunately, this approach often fails due to subtle, albeit important, differences in reaction requirements. Consequently, a significant goal in synthetic chemistry is the ability to transfer chemical observations from one reaction to another, quantitatively. Here, we present such a platform by developing a holistic, data-driven workflow for deriving statistical models for one set of reactions that can be applied to predict out-of-sample examples. As a validating case study, published enantioselectivity data sets that employ BINOL-derived chiral phosphoric acids for a range of nucleophilic addition reactions to imines were combined and statistical models developed. These models reveal the general interactions imparting asymmetric induction and allow the quantitative transfer of this information to new reaction components. The disclosed techniques create opportunities for translating comprehensive reaction analysis to diverse chemical space, streamlining both catalyst and reaction development.

Methods Summary:

After the database of the reactions is constructed, the experimental output, enantiomeric ratios, were mathematically modelled through linear regression techniques to reveal which of the proposed parameters allow for the prediction of new outcomes. The detailed acquisition of parameters can be found in the Supplementary Information and the descriptor tables attached as an accompanying spreadsheet. The models produced were evaluated for their goodness of fit, R^2, and their robustness is demonstrated by external validations' goodness of fit, ${}_{\text{pred}}R^2$. The nearer the R^2 and slope values are to one (indicating a tight, one-to-one correlation between predicted and measured outcomes) and the nearer the intercept is to zero (indicating minimal systematic error), the more robust the model. Potential models were refined through number of parameters, because this allows for a mechanistically informative interrogation and cross-validation scores. Leave one reaction out (LORO) analysis was performed to probe general mechanistic principles, which

*Correspondence to: M.S.S. sigman@chem.utah.edu.

Author contributions: J.P.R. designed and performed all computations and statistical analyses. Both authors contributed to the analysis and writing of the manuscript.

Author information: Authors declare no competing interests.

Data and code availability: All data and code used for model development is available in the Supporting Information.

provides the basis for mechanistic transfer of experimental observations and tested further by predicting out-of-sample.

The efficacy of a catalytic process is dictated by the possible transition states (TS), which feature core non-covalent interactions that determine their geometries and energies.[1–2] Such interactions are often difficult to identify and define since they are energetically weak and sensitive to the molecular properties of every reaction component (catalyst, substrate(s), reagents, solvent, etc.).[3–4] This overarching issue in reaction optimization is often exasperated by subtle connections across several reaction variables, wherein modest structural changes to any or a few of these can have a profound effect on the experimental outcome.[5–6,7] These factors combined with the number of dimensions under study in most reactions, are the underlying reasons for why optimization is decidedly empirical.[8–9] This situation is particularly common in the area of asymmetric catalysis, wherein seemingly minor structural variations to any reaction component can have acute and non-intuitive influences on the observed enantioselectivity.[10] However, it is possible that such mechanistic outliers may be concealed within larger data sets as our pattern recognition skills do not perceive pivotal generalities when reaction situations change. On this basis, we hypothesized that connecting common mechanistic features through the simultaneous interrogation of all reaction components would provide a holistic view of the key non-covalent interactions responsible for reaction performance. This would enable the transfer of experimental observations to genuinely different substrate combinations with unique catalysts. Herein, we develop and deploy a workflow to parameterize all reaction variables of >350 distinct reaction combinations, which allows development of comprehensive correlations with the ultimate ability to predict reaction performance for entirely different structural motifs. The workflow includes techniques to probe general mechanistic principles, which provides the basis for transfer learning or generalized identification of the key interactions imparting asymmetric induction.

Asymmetric catalysis is replete with examples of catalysts that can promote disparate reactions through a common mode of activation.[11,12–13,14] However, when one surveys “similar reactions”, many changes to the precise reaction conditions are often required to obtain the desired reaction performance.[15–16] In many cases, these changes can be subtle (i.e., one aromatic solvent for another) or more profound (one catalyst class for another). This leads to the questions: 1) is mechanistic insight truly transferrable to a new reaction in the same subclass considering that a standard mechanistic paradigm may exist with a general mode of activation? 2) if so, how can a workflow be used in a quantitative manner for diverse and multiple reaction profiles? And 3) if achievable, can the observations of one or more reactions be deployed to predict the performance of another? Such analysis strategies could provide mechanistic understanding to why certain conditions are effective for a general reaction type and the ability to quantitatively transfer this information to out of sample predictions streamlining reaction optimization.[17–18]

To vet a specific workflow required to probe the questions posed above, it would be pragmatic to compare transformations within a reaction class facilitated by a single catalyst chemotype. Although multifarious reports of the same catalyst class for different

transformations exist in enantioselective catalysis, comparative studies – even in a qualitative manner – have been sparse. Such an interrogation would be challenging as a consequence of incomplete datasets generated under non-uniform conditions and the development of readily comprehensible descriptors for each varying reaction component. To address this correlation challenge, we envisioned a strategy for the interrogation of enantioselective catalysis involving the application of modern data-analysis methods and advanced parameter sets. In this approach, integrated descriptor sets (Quantitative Structure Activity Relationships (QSAR), Molecular Mechanics (MM), and Density Functional Theory (DFT) derived),[19] are related to a relatively large library of outputs collected from a general reaction and catalyst type, which are data-mined from multiple literature sources. By combining the appropriate data-organization and trend analysis techniques, general relationships between reactions can be established. The statistical models ability to predict a new reaction type performance is used as a validation of mechanistic transferability (Fig. 1).

Reaction Platform Selection:

As a proof of concept reaction class, the addition of various nucleophiles to imines was identified owing to the ubiquity of this type of transformation in asymmetric catalysis.[20–21] The commonality of this reaction is due to both the simplicity of accessing imine starting materials and the broad applicability of the resulting amines in both synthetic and biosynthetic settings.[22–23] The second criteria in reaction selection is determining a catalyst chemotype that has been widely used for these processes such that significant data exists and provides diversity in data range and structure from published sources. Considering these constraints, we selected the field of chiral phosphoric acid catalysis, particularly, the addition of protic nucleophiles to imines catalyzed by chiral 1,1’-Bi-2-naphthol (BINOL)-derived phosphoric acids bearing aromatic groups at the 3 and 3’ positions (Fig.1).[24]

To initiate this workflow, an expanded inventory of 367 reactions with varied components was curated from multiple reports (for list of references see SI). From this survey, we categorized the dataset by imine TS geometry (*E* or *Z*) wherein *E*-imine TS are grouped by a +ee value and *Z*-imines as a -ee value. Imine stereochemistry was determined by the enantiomer of the product formed if the imine was derived from an aldehyde. However, if ketimines (imines derived from ketones) were employed substituent size must also be considered if the smaller C-substituent has higher Cahn-Ingold Prelog (CIP) priority.[25–26] For the reactions under study, this only effects ketimines that have either a trifluoromethyl or ester C-substituents in which they are considered to have lower priority for the purpose of assigning *E* or *Z*-TS. This is important in understanding product enantioselectivities, because nucleophile addition to the same face will yield opposite enantiomers for *E* and *Z* configurations. Therefore, developed models will not be capable of predicting product stereochemistry but can be deployed to predict whether a reaction will proceed *via* an *E* or *Z* type mechanism and this information can be used to determine absolute configurations.

Simultaneously, a diverse array of molecular descriptor values was collected from DFT optimized geometries to describe the structural features of each imine, nucleophile, catalyst, and solvent. Unfortunately, the lack of structural commonality for particular molecular subsets creates a challenge in identifying readily comprehensible and extensive parameter

sets for each component. For example, on comparing substrates and catalyst structures, it is apparent that they have overlapping and distinctive features likely required for determining selectivity patterns (Extended Data Fig. 1). In contrast, the solvents do not have common substructures, yet are critical for enantioselectivity. To address this limitation two approaches were explored: (1) parameters derived from DFT calculations were collected, which have proven well-equipped to describe molecules containing common structural features including Sterimol parameters, bond lengths, angle measurements, molecular vibrations and intensities, Natural Bond Orbitals (NBO) charges, polarizabilities, Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbitals (LUMO) energies.[27–28] This array was collected for the reaction partners and the catalysts. (2) 2D descriptors (e.g., topological and connectivity as exemplified by molecular shape, size and number of heteroatoms) are used as this is a traditional method to assess structurally disparate molecules such as solvents.[29–30] Other reaction variables, such as concentration of reagents/catalysts and inclusion of molecular sieves, were also included as categorical descriptors (see SI).

Comprehensive Model Development:

Linear regression algorithms (see SI) were then applied to the entire dataset (367 reactions) to identify correlations between the molecular structure of every reaction variable defined by the parameters collected in the previous step of the workflow and the experimentally determined enantioselectivity. $\Delta\Delta G^{\neq}$ (where $\Delta\Delta G^{\neq} = —RT\ln(e.r.)$ and T is the temperature at which the reaction was performed) was regressed to an equation to reveal a surprisingly good correlation despite the significant structural variance included in the training set. Both cross-validation analysis (Leave-one-out (LOO) and *k*-fold) and external validation, in which the dataset is partitioned pseudorandomly into 50:50 training:validation sets suggests a relatively robust model (see SI). The model emphasizes solvent (black), imine (blue), nucleophile (green) and catalyst (red) terms distributed over six parameters, as contributors to the enantioselectivity across these seventeen reaction types (Fig. 2A). A slope approaching unity and intercept approaching zero over the training set indicates an accurate and predictive model with an R^2 value of 0.88 demonstrating a high degree of precision. The largest coefficients in this normalized model belong to the imine NBO descriptors indicating the significant role of the imine substrate in the quantification of enantioselectivity as highlighted by the formation of both enantiomeric products, a consequence of active *E* and *Z* configurations (*vide infra*). Comparing two Strecker reactions performed under uniform conditions, results in values ranging from +99% ee for the enantiomer that proceeds through the *E*-imine TS and −80% ee for the *Z*. Remarkably, this represents a 3.5 kcal/mol range based solely on imine structure.

We postulated that the ability to correlate and predict using a singular model for an array of reactions suggests that the transition state features are fundamentally similar within this reaction range. Perhaps, the best test of this hypothesis could be achieved by a “leave one reaction out” (LORO) analysis. In this statistical evaluation, the catalyst, imine, and nucleophile structures are varied as a validation set and assessed through the ability of the model to predict with sufficient accuracy. This would report on the model’s capacity to match patterns across a general reaction type. Using this analysis, each distinct reaction (as

determined by individual publications) in the data field was evaluated with most predicted well (see SI). As an illustration of model robustness, we could exclude up to seven reactions with little change in the correlation statistics (Fig. 2B). However, not surprisingly, some reactions were poorly predicted in the LORO protocol, which can be attributed to the model's inability to capture specific structure changes if they are not adequately expressed in the training set. In sum, the descriptor definitions coupled to the model and validation strategies showcase that patterns can be matched. This is consistent with the hypothesis that a defined set of key non-covalent interactions impart asymmetric induction across a general reaction type. Essentially, this workflow provides evidence that one reaction can be used to predict the results of another, quantitatively.

Trend analysis:

Although the comprehensive model in Fig. 2 establishes the capacity of the selected parameters to describe general aspects of this system, the ultimate goal of our workflow is to discern subtle underlying mechanistic phenomena. This objective could not be achieved by using the above correlation because it was produced by using the entire dataset, which provides only an overview of the mechanistic patterns. We hypothesized that a series of focused correlations, coupled with an evaluation of the overall trends, might serve to reveal fundamental features of the systems. To this end, we truncated the dataset into subsets, categorized by imine TS geometry (*E* or *Z*) determined by the relative sign of the ee determined previously, as these are hypothesized to lead to structurally distinct interactions with the other reaction components. This organizational scheme was viewed as a means to facilitate the identification of catalyst features that affect particular mechanistic pathways and therefore, reactant combinations (and *vice versa*). Linear regression algorithms were then applied to this data classification to identify correlations between molecular structure and the experimentally determined enantioselectivity. Subsequently, analysis and refinement of the resultant models were used to produce explicit mechanistic hypotheses (Fig. 3).

The correlation depicted in Fig. 3 was identified from a set of 204 reactions (evenly split into training and validation sets) that proceed via *E*-imine TS. The relationship includes two solvent, two imine, one nucleophile and three catalyst terms. Overall, the statistical model suggests a mechanistic scenario in which the imine adopts an arrangement that minimizes energetically penalizing repulsion interactions with reasonably large catalyst substituents.[31] Perhaps most telling is that the steric profile of the nucleophile does not impart a significant impact on the stereoselectivity prediction, despite the large structural variance. The included parameters (LUMO and the P−O asymmetric stretching intensity, iPO_{as}) suggest that H-bonding contacts between catalyst and nucleophile play a minor role and the use of essentially any nucleophile should be compatible with the reaction if the imine and catalyst are matched.

In evaluating the model for *Z*-imines determined by 147 reactions, a number of overlapping terms reinforce the notion that similar interactions between catalyst and substrates remain within the two geometric imine stereoisomers. Two of these terms, the size of the catalyst aryl substituent as measured by Sterimol B1 term, $B1_{cat}$, and the imine NBO_{PG} parameter essentially describes the repulsive interactions between proximal sterics and the imine *N*-

substituent, a common critical catalyst-substrate interaction with both TS imine configurations. The most significant contrast in the two models is that the Z-imine model includes a significant nucleophile steric descriptor, $B5_{Nu}$, which is the most highly weighted term in the equation. This suggests that larger nucleophiles introduce enhanced repulsive interactions with the catalyst substituents in the TS, leading to the competing product, which ultimately favors the observed enantiomer. This claim is further supported by the observation of higher enantioselectivities when using catalysts with smaller substituents (e.g., Ar = 3,5-CF_3). The proposed physical meaning behind each term in the mathematical equations have been summarized in Fig. 3.

Evaluation of prediction capabilities:

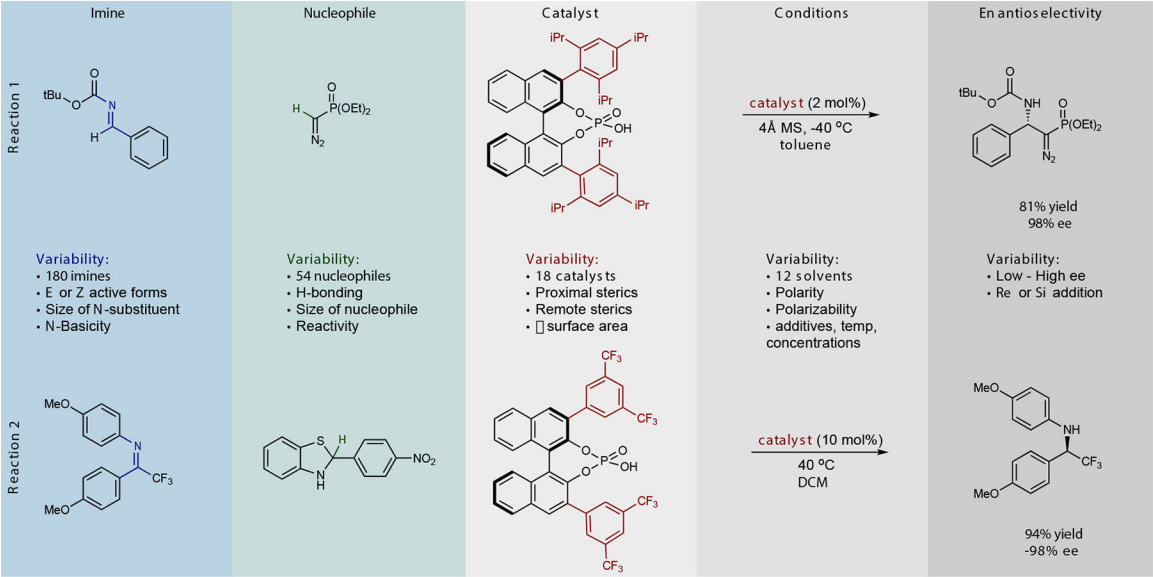
As a final step in the workflow, we evaluated the ability to transfer the mechanistic principles leading to enantioselective catalysis captured by the statistical models to genuinely different structural motifs not contained in the training dataset. If effective out-of-sample prediction were possible, the model could predict the impact of a new imine, nucleophile, and/or catalyst. Initially, reaction performance was evaluated using the comprehensive model to determine the mechanistic pathway under operation, these predictions could then be further refined with the specific models (*E* or *Z*). This two-tiered workflow is imperative as the process avoids mechanistic assumptions regarding whether the reaction proceeds via an *E* or *Z* TS and therefore ensuring that the results of the test reactions are unknown. The comprehensive model does not immediately allow prediction of stereochemistry however, product configuration can be assigned from the simple models shown in Figure 4. These are based on the amine product yielded from a reaction proceeding via an *E* or *Z* TS and catalyzed by the (*R*)-CPA. The opposite enantiomer will be formed if the (*S*)-catalyst is employed. As a first case study, we evaluated fifteen additional reactions involving enecarbamates, a nucleophile not contained in the training set, and benzoyl imines, an imine subclass that is part of our initial training set (Fig. 4).[32] Each result was predicted using the comprehensive model, with an average absolute $\Delta\Delta G^{\neq}$ error of 0.37 kcal/mol (13 examples within 5% ee) and correctly assigned the absolute stereochemistry as *R*, demonstrating the ability of the model to extrapolate effectively to a new nucleophile. A slightly improved outcome is observed using the *E*-imine mechanistic model with a 0.24 kcal/mol average error (all examples within 5% ee).

As the second case study, the hydrogenation of alkynyl ketimines catalyzed by H8-BINOL where the 3,3' groups = 3,5-CF_3(C_6H_4) was predicted.[33] This is a more challenging scenario as both imine and catalyst components are not included in the training set. Again, accurate prediction of the outcomes was construed using the *Z*-imine mechanistic model, with an average absolute error of 0.30 kcal/mol and 13 examples predicted within 2% ee (Fig. 4). The stereochemical outcome was correctly determined to be *R* with the (*S*)-catalyst. Although the comprehensive model assesses the mechanistic scenario and therefore assigns the stereochemical outcome, it was not as accurate since the nucleophile information was categorical (symmetrical or displaced). Thus, the beneficial effect of a large nucleophile for a *Z* reaction was not adequately captured. These examples showcase that the model's predictive capabilities are not limited to classifying the vast literature, but can be applied to analyze and predict new reactions even in situations where multiple components are varied.

As a final case study, we evaluated a recently reported reaction that was rendered highly predictable by application of machine learning (ML) algorithms. The study reported by Denmark and coworkers involved the addition of thiols to benzoyl imines, a distinct reaction included in our training set.[34] To utilize ML approaches, they performed 2,150 separate experiments using 43 catalysts to yield 25 different products (5×5 nucleophile/electrophile matrix). We postulated that our approach could reliably predict their results including the best catalyst, TCYP, which has cyclohexyl groups at the 2,4,6 positions of the aromatic ring and is not in our training set. To test this hypothesis, all experimental results of this reaction type were removed from our original training data, the model was retrained, and deployed to predict their new dataset (34 reactions) collected with the best catalyst, TCYP. We conclude, that our model, with no experimental data on this reaction can also predict the enantioselectivities (average absolute $\Delta\Delta G^{\neq}$ error = 0.65 kcal/mol comprehensive model (26 examples within 5% ee), 0.67 kcal/mol *E*-imine only model (25 examples within 5% ee)) confidently determining the stereochemical outcome to be *R* and TCYP as a highly selective catalyst. Overall, through the combination of results generated from the out-of-sample prediction platforms, we can conclude that the *E* and *Z* focused correlations generate more accurate predictions but the comprehensive model is valuable as it determines which equation should be deployed.

In this report, we have introduced a workflow to model enantioselectivity in assorted catalytic systems. The value of this approach is that complicated reaction conditions can be accounted and successfully evaluated for multiple and diverse reactions. The ability to correlate and predict using a single model for many reactions suggests that general transition state features are fundamentally similar across the reaction range, allowing the transfer of observations from one reaction to another. This finding highlights a likely general phenomenon in asymmetric catalysis, whereby various transformations may be found to perform in the same manner when exposed to similar reaction conditions. However, such reaction similarities may be unmasked, and reaction-specific mechanistic principles emerge from the development of focused correlations.

Extended Data



Extended Data Figure 1 |. Reaction component comparison.

Parameterization challenges for the identification of numerical descriptors in reaction dimension, demonstrated using two reactions representing the extremes of multidimensional feature space. DCM, dichloromethane; MS, molecular sieves; ee, enantioselectivity.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

J.P.R. thanks the EU Horizon 2020 Marie Skłodowska-Curie Fellowship (grant no. 792144) and M.S.S thanks the NIH (1 R01 GM121383) for support of this work. Computational resources were provided from the Center for High Performance Computing (CHPC) at the University of Utah and the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the NSF (ACI-1548562) and provided through allocation TG-CHE180003.

References and Notes:

(1). Houk KN & Cheong PH-Y Computational prediction of small-molecule catalysts. Nature 455, 309–313 (2008). [PubMed: 18800129]

(2). Davis HJ & Phipps RJ Harnessing Non-Covalent Interactions to Exert Control over Regioselectivity and Site-Selectivity in Catalytic Reactions. Chem. Sci 8, 864–877 (2017). [PubMed: 28572898]

(3). Knowles RR & Jacobsen EN Attractive Noncovalent Interactions in Asymmetric Catalysis: Links between Enzymes and Small Molecule Catalysts. Proc. Natl. Acad. Sci. U. S. A 107, 20678–20685 (2010). [PubMed: 20956302]

(4). Sigman MS, Harper KC, Bess EN & Milo A The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. Acc. Chem. Res 49, 1292–1301 (2016). [PubMed: 27220055]

(5). Ahneman DT; Estrada JG; Lin S; Dreher SD; Doyle AG “Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning” Science 360, 186–190 (2018). [PubMed: 29449509]

(6). Chuang KV; Keiser MJ Comment on “Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning” Science 362, eaat8603 (2018). 10.1126/science.aat8603 [PubMed: 30442776]

(7). Estrada JG; Ahneman DT; Sheridan RP; Dreher SD; Doyle AG Response to Comment on “Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning” Science 362, eaat8763 (2018). 10.1126/science.aat8763 [PubMed: 30442777]

(8). Robbins DW & Hartwig JF A Simple, Multidimensional Approach to High-Throughput Discovery of Catalytic Reactions. Science 333, 1423–1427 (2011). [PubMed: 21903809]

(9). McNally A, Prier CK & MacMillan DWC Discovery of an alpha-C-H Arylation Reaction Using the Strategy of Accelerated Serendipity. Science 334, 1114–1117 (2011). [PubMed: 22116882]

(10). Neel AJ, Milo A, Sigman MS & Toste FD Enantiodivergent Fluorination of Allylic Alcohols: Dataset Design Reveals Structural Interplay Between Achiral Directing Group and Chiral Anion. J. Am. Chem. Soc 138, 3863–3875 (2016). [PubMed: 26967114]

(11). Walsh PJ & Kozlowski MC Fundamentals of Asymmetric Catalysis (University Science Books, 2008).

(12). Yoon TP & Jacobsen EN Privileged Chiral Catalysts. Science 299, 1691–1693 (2003). [PubMed: 12637734]

(13). Yamamoto H Lewis Acids in Organic Synthesis (Wiley-VCH, 2000).

(14). Akiyama T Stronger Brønsted acids. Chem. Rev 107, 5744–5758 (2007). [PubMed: 17983247]

(15). Collins KD & Glorius F Intermolecular reaction screening as a tool for reaction evaluation. Acc. Chem. Res, 48, 619–627 (2015). [PubMed: 25699585]

(16). Gesmundo NJ et al. Nanoscale synthesis and affinity ranking. Nature 557, 228–232 (2018) [PubMed: 29686415]

(17). Reetz MT Laboratory Evolution of Stereoselective Enzymes: A Prolific Source of Catalysts for Asymmetric Reactions. Angew. Chem., Int. Ed 50, 138–174 (2011).

(18). Hansen E, Rosales AR, Tutkowski B, Norrby P-O & Wiest O Prediction of stereochemistry using Q2MM. Acc. Chem. Res 49, 996–1005 (2016) [PubMed: 27064579]

(19). Metsänen TT et al. Combining traditional 2D and modern physical organic-derived descriptors to predict enhanced enantioselectivity for the key aza-Michael conjugate addition in the synthesis of Prevymis- (letermovir). Chem. Sci, 9, 6922–6927 (2018). [PubMed: 30210766]

(20). Robak MT, Herbage MA & Ellman JA Synthesis and Applications of tert-Butanesulfinamide. Chem Rev 110, 3600–3740 (2010). [PubMed: 20420386]

(21). Kobayashi S, Mori Y, Fossey JS & Salter MM Catalytic Enantioselective Formation of C–C Bonds by Addition to Imines and Hydrazones: A Ten-Year Update. Chem. Rev 111, 2626–2704 (2011). [PubMed: 21405021]

(22). Nugent TC Chiral Amine Synthesis: Methods, Developments and Applications Wiley-VCH Verlag GmbH & Co. KGaA; Weinheim, Germany: 2010.

(23). Silverio DL et al. Simple organic molecules as catalysts for enantioselective synthesis of amines and alcohols. Nature 494, 216–221 (2013). [PubMed: 23407537]

(24). Parmar D; Sugiono E; Raja S & Rueping M Complete Field Guide to Asymmetric BINOL-Phosphate Derived Brønsted Acid and Metal Catalysis: History and Classification by Mode of Activation; Brønsted Acidity, Hydrogen Bonding, Ion Pairing, and Metal Phosphates. Chem. Rev 114, 9047–9153 (2014). [PubMed: 25203602]

(25). Simón L & Goodman JM Theoretical study of the mechanism of hantzsch ester hydrogenation of imines catalyzed by chiral BINOL-phosphoric acids. J. Am. Chem. Soc 130, 8741–8747 (2008). [PubMed: 18543923]

(26). Reid JP, Simón L & Goodman JM A Practical Guide for Predicting the Stereochemistry of Bifunctional Phosphoric Acid Catalyzed Reactions of Imines. Acc. Chem. Res 49, 1029 (2016). [PubMed: 27128106]

(27). Santiago CB, Guo J-Y & Sigman MS Predictive and mechanistic multivariate linear regression models for reaction development. Chem. Sci 9, 2398–2412 (2018). [PubMed: 29719711]

(28). Reid JP & Sigman MS Comparing quantitative prediction methods for the discovery of small-molecule chiral catalysts. Nat. Rev. Chem 2, 290–305 (2018).

(29). Denmark SE, Gould ND & Wolf LM A systematic investigation of quaternary ammonium ions as asymmetric phase-transfer catalysts. Application of quantitative structure activity/selectivity relationships. J. Org. Chem 76, 4337–4357 (2011). [PubMed: 21446723]

(30). Hansch C & Leo A Exploring QSAR: Fundamentals and Applications in Chemistry and Biology (ACS, Washington, 1995).

(31). Reid JP & Goodman JM Goldilocks catalysts: computational insights into the role of the 3,3' substituents on the selectivity of BINOL-derived phosphoric acid catalysts. J. Am. Chem. Soc 138, 7910–7917 (2016). [PubMed: 27227372]

(32). Terada M, Machioka K & Sorimachi K High Substrate/Catalyst Organocatalysis by a Chiral Brønsted Acid for an Enantioselective Aza-Ene-Type Reaction. Angew. Chem. Int. Ed 45, 2254–2257 (2006).

(33). Chen M-W et al. Organocatalytic Asymmetric Reduction of Fluorinated Alkynyl Ketimines. J. Org. Chem 83, 8688–8694 (2018). [PubMed: 29884023]

(34). Zahrt AF et al. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. Science 363,(2019) DOI: 10.1126/science.aau5631.

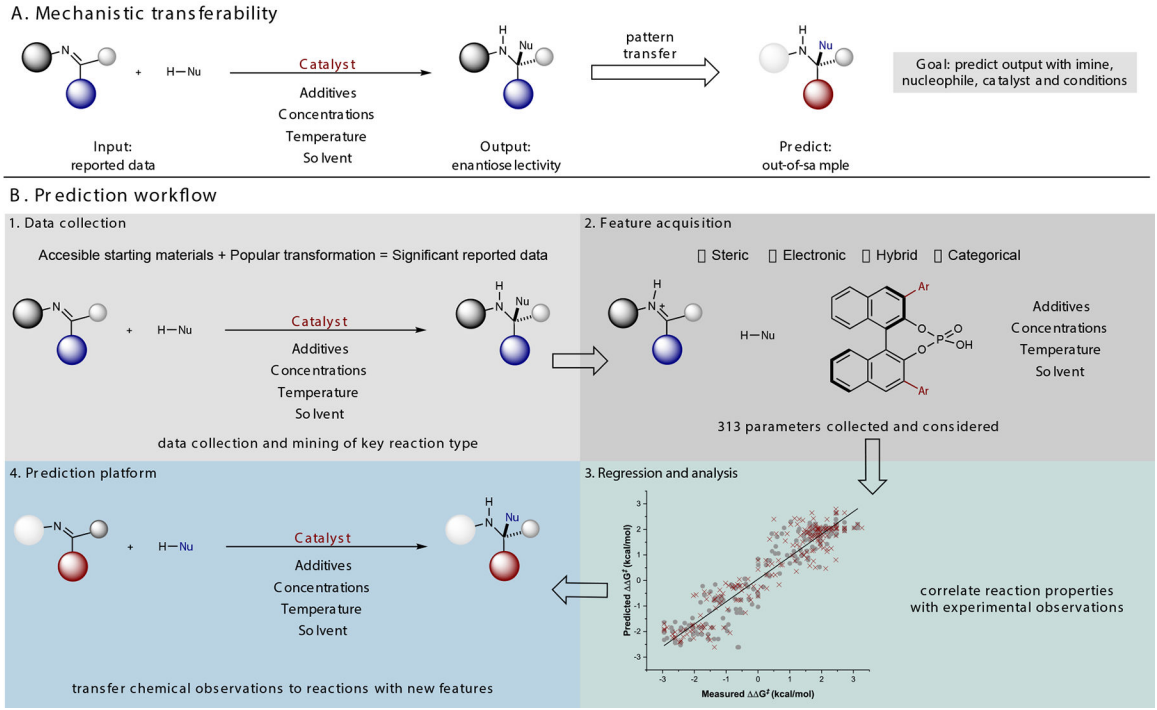


Figure 1 |. Workflow for interrogating and applying mechanistic transferability.

(A) BINOL-based phosphoric acid catalyzed nucleophilic additions to imines as a general reaction for workflow development. (B) Streamline reaction performance predictions by employing a mechanistic transferability strategy implemented through correlation of all reaction variables to enantioselectivity. General correlations can be built to reveal the interactions between any reaction component in the relevant TS and enantioselectivity. The mechanistic principles leading to enantioselective catalysis captured by the statistical models can be transferred to genuinely different structural motifs not contained in the training dataset.

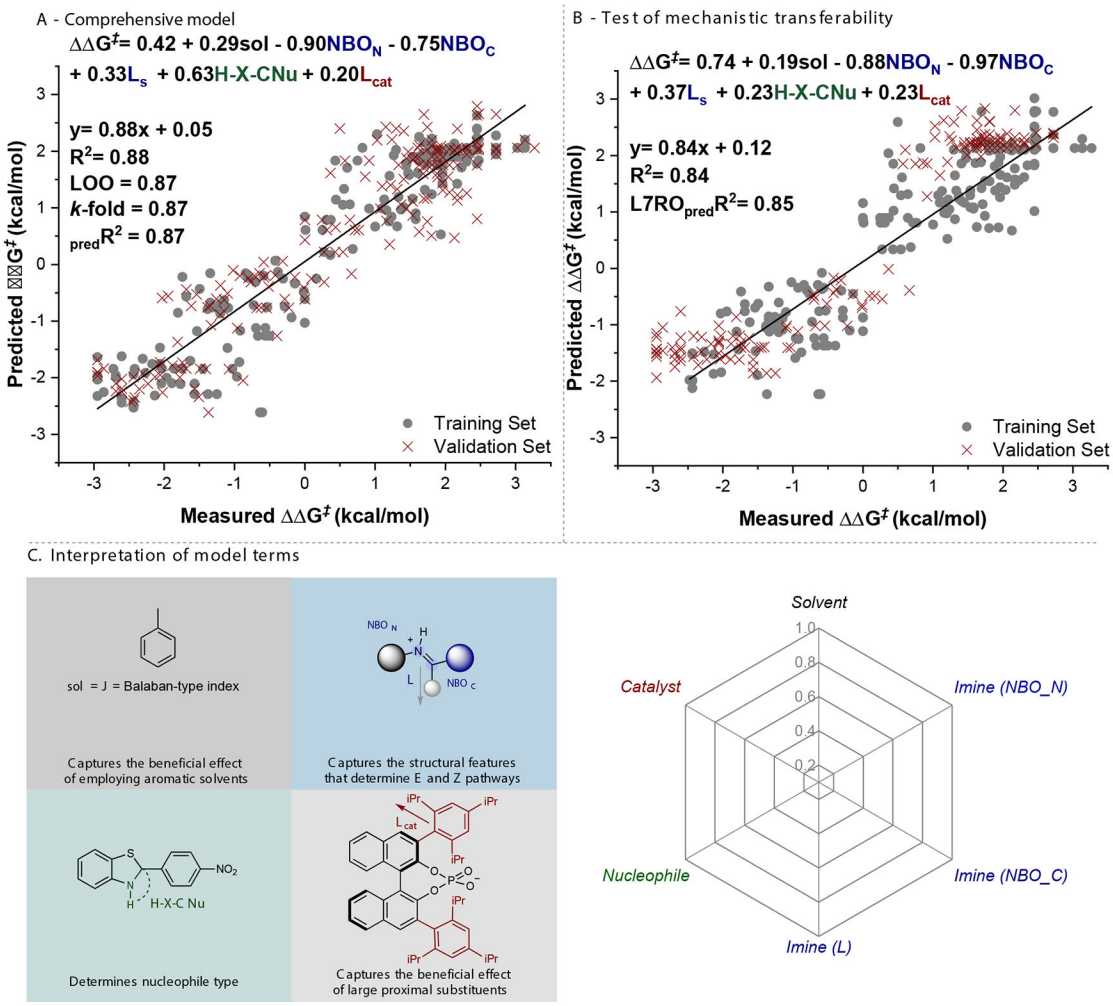


Figure 2 |. Comprehensive model development.

(A) Regression model containing 367 data entries facilitated by parameterization of every reaction variable. A positive %ee value indicates *E*-imine TS, a negative %ee *Z*-imine TS. LOO, leave-one-out cross-validation score; *k*-fold, average fourfold cross-validation score. (B) Illustration of mechanistic transferability in the data set via “leave one reaction out” (LORO) analysis. In which distinct reactions (as determined by individual publications) are defined as the validation set. (C) Visual analysis and interpretation of the model terms.

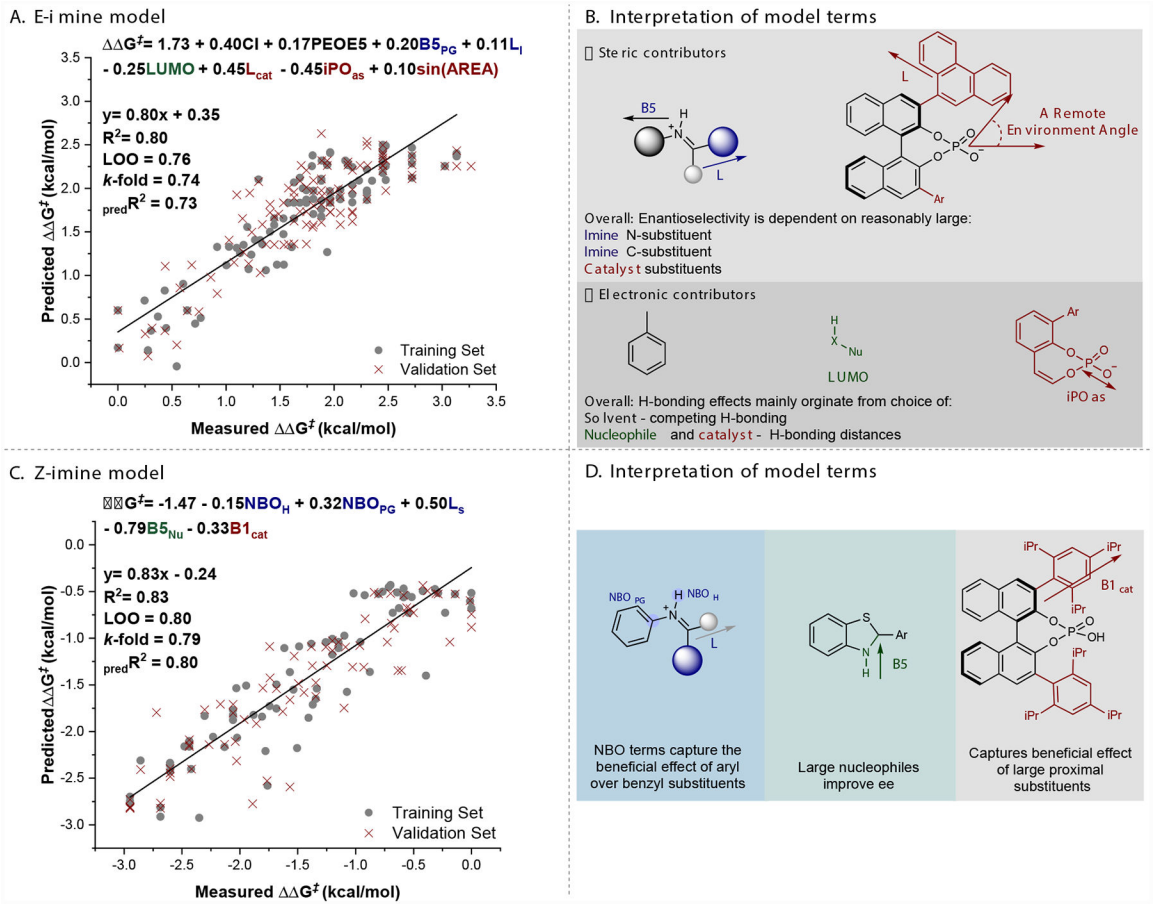


Figure 3 |. Development of focused correlations.

(A) Regression model containing 204 entries data-mined from nine literature sources. (B) Model emphasizes the importance of both steric and electronic factors. Reasonably large catalyst and imine substituents lead to high-levels of enantioselectivity, if these two components are matched any nucleophile should be compatible. (C) Regression model containing 147 entries data-mined from eight literature sources. (D) Overlapping steric terms describing catalyst and imine reinforce the notion that similar interactions remain within the two geometric imine stereoisomers. However, this model emphasizes the importance of steric contributions predominantly from the nucleophile for high enantioselectivities.

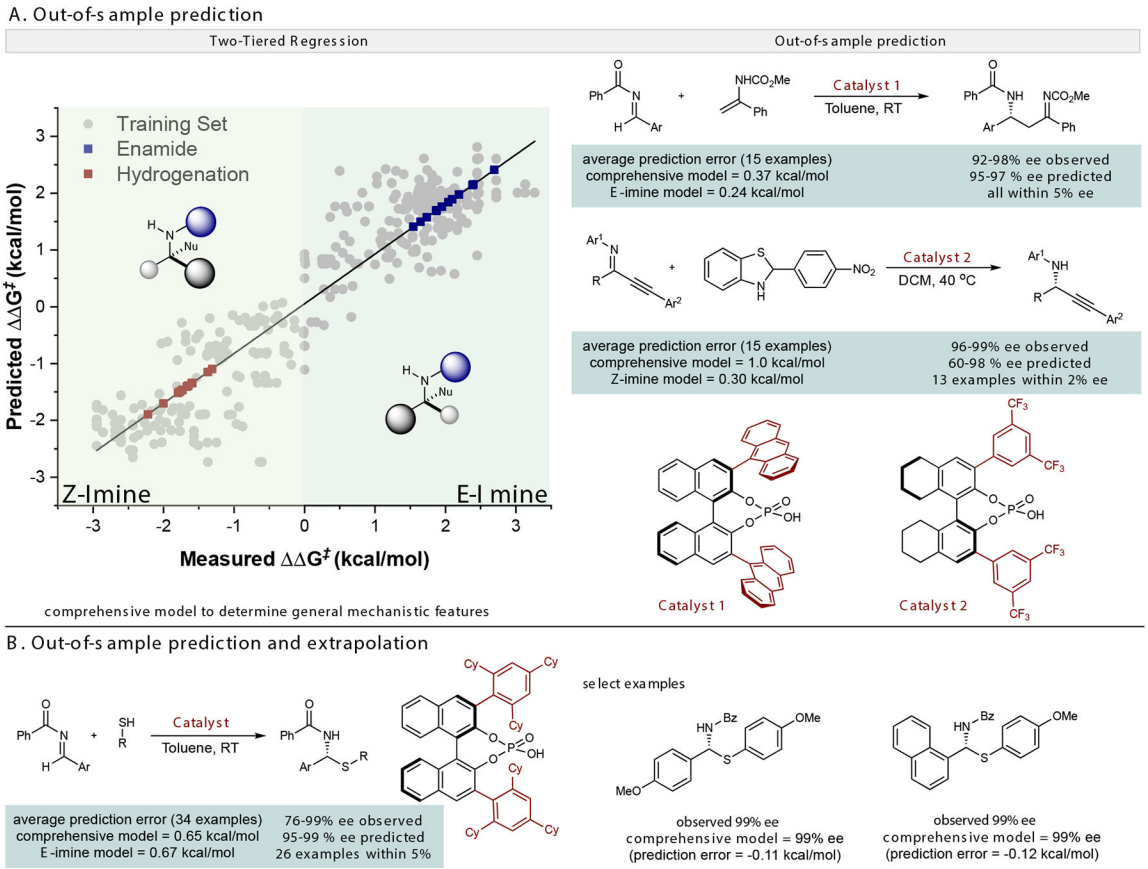


Figure 4 |. Out-of-sample predictions using two-tiered prediction workflow.

Comprehensive model first determines *E* or *Z* TS, configuration specific models are then utilized to refine predictions. A generic amine product denotes the stereochemical outcome predicted if the reaction proceeds via the *E* or *Z* TS and catalyzed by an (*R*)-CPA. Product stereochemistry is reversed if the (*S*)-catalyst is used. (A) Application to addition of enecarbamates to benzoyl imines and transfer hydrogenation of alkynyl ketimines. (B) Prediction of TCYP which has cyclohexyl groups at the 2,4,6 positions of the aromatic ring, as a highly selective catalyst for the addition of thiol to benzoyl imines.