Holistic prediction of enantioselectivity in asymmetric catalysis

Jolene P. Reid1 & Matthew S. Sigman1*

When faced with unfamiliar reaction space, synthetic chemists typically apply the reported conditions (reagents, catalyst, solvent and additives) of a successful reaction to a desired, closely related reaction using a new substrate type. Unfortunately, this approach often fails owing to subtle differences in reaction requirements. Consequently, an important goal in synthetic chemistry is the ability to transfer chemical observations quantitatively from one reaction to another. Here we present a holistic, data-driven workflow for deriving statistical models of one set of reactions that can be used to predict out-of-sample reactions. As a validating case study, we combined published enantioselectivity datasets that employ 1,1′-bi−2−naphth (BINOL)−derived chiral phosphoric acids for a range of nucleophilic addition reactions to imines and developed statistical models. These models reveal the general interactions that impart asymmetric induction and allow the quantitative transfer of this information to new reaction components. This technique creates opportunities for translating comprehensive reaction analysis to diverse chemical space, streamlining both catalyst and reaction development.

The efficacy of a catalytic process is dictated by the possible transition states, which feature core non-covalent interactions that determine their geometries and energies1,2. Such interactions are often difficult to identify and define because they are energetically weak and sensitive to the molecular properties of every reaction component (catalyst, substrates, reagents, solvent and so on)3,4. This overarching issue in reaction optimization is often exacerbated 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 outcome5–7. These factors, combined with the number of dimensions under study in most reactions, are the underlying reasons that optimization is traditionally empirical8,9. This situation is particularly common in the area of asymmetric catalysis, wherein seemingly minor structural variations in any reaction component can have acute and non-intuitive influences on the observed enantioselectivity10. However, it is possible that such mechanistic outliers may be concealed within larger datasets because 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. Here we develop and deploy a workflow that parameterizes all the reaction variables of more than 350 distinct reaction combinations, which allows the development of comprehensive statistical models, resulting in the 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 activation11–14. However, when ‘similar’ reactions are attempted, many changes to the precise reaction conditions are often required to obtain the desired reaction performance15,16. These changes can be subtle (that is, one aromatic solvent for another) or more profound (one catalyst class for another). This leads us to ask (1) whether mechanistic insight is transferable to a new reaction in the same subclass, given that a standard mechanistic paradigm may exist with a general mode of activation? If so, (2) how could a data-driven workflow that combines data acquisition and a description of the molecules involved mathematically be used to build a statistical model for diverse and multiple reaction profiles? And if such a workflow is achievable, (3) can the observed conditions of one or more reactions be deployed to predict the performance of another? Such analysis could provide a mechanistic understanding of why certain conditions are effective for a general reaction type and the ability to transfer this information quantitatively to out-of-sample predictions streamlining reaction optimization17,18.

To assess a specific workflow that is designed 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 qualitative rather than quantitative—have been sparse. Such an assessment would be challenging because most datasets, often generated under non-uniform conditions, are incomplete and readily comprehensible descriptors for each varying reaction component need to be developed. 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) derived19—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 (see the Supplementary Information). By combining appropriate data-organization and trend-analysis techniques, general relationships between reactions can be established. The ability of the statistical models to predict a new reaction type performance is used as a validation of mechanistic transferability (Fig. 1).

1Department of Chemistry, University of Utah, Salt Lake City, UT, USA. *e-mail: sigman@chem.utah.edu
Reaction platform selection

As a proof-of-concept reaction class, we chose the addition of various nucleophiles to imines owing to the ubiquity of this type of transformation in asymmetric catalysis. This reaction class uses imine starting materials that are easy to obtain and the resulting amine products have broad applicability in both synthetic and biosynthetic settings. As a next step, we evaluated the different catalyst chemotypes used in this reaction class, focusing on those that provide a wide range of both substrate structural types and enantioselectivity data from published sources. With these constraints in mind, we selected the field of chiral phosphoric acid (CPA) catalysis, in particular the addition of protic nucleophiles to imines catalysed by chiral 1,1′-bi-2-naphthol (BINOL)-derived phosphoric acids bearing aromatic groups at the 3 and 3′ positions.

To initiate this workflow, an expanded inventory of 367 reactions with varied components was curated from multiple reports (for a list of references, see Supplementary Information). From this survey, we categorized the dataset by imine transition-state geometry (E or Z) wherein E-imine transition states have a +e.e. value and Z-imines have a −e.e. 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, we also needed to consider substituent size if the smaller C-substituent has higher Cahn–Ingold–Prelog (CIP) priority. For the reactions we needed to consider substituent size if the smaller C-substituent has higher Cahn–Ingold–Prelog (CIP) priority. For the reactions we studied here, this affects only ketimines that have either a trifluoromethyl or ester C-substituent, which are considered to have lower priority.

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, when comparing substrates and catalyst structures, it is apparent that they have overlapping and distinctive features that are probably required for determining selectivity patterns. Extended Data Fig. 1). By contrast, the solvents do not have common substructures, yet are critical for enantioselectivity.

To address this limitation, we explored two approaches: (1) we collected parameters derived from DFT calculations, which satisfactorily describe molecules containing common structural features including Sterimol parameters, bond lengths, angle measurements, molecular vibrations and intensities, natural bond orbital (NBO) charges, polarizabilities, highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies. We collected these parameters for both the reaction partners and the catalysts. (2) We used two-dimensional descriptors (such as topology and connectivity as exemplified by molecular shape, size and number of heteroatoms) because this is a traditional method of assessing structurally disperse molecules such as solvents. Other reaction variables, such as concentration of reagents or catalysts and inclusion of molecular sieves, were also included as categorical descriptors (see Supplementary Information).
Comprehensive model development

Linear regression algorithms (see Supplementary Information) 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. $\Delta G^2 = -RT\ln(\text{e.r.})$ (where e.r. is the enantiomeric ratio, $T$ is the temperature at which the reaction was performed and $R$ is the gas constant) was regressed to an equation to reveal a surprisingly good correlation despite the large 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, suggest a relatively robust model (see Supplementary Information). 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 a goodness-of-fit $R^2$ 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 crucial 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 (see below). A comparison of two Strecker reactions performed under uniform conditions results in values ranging from $+99\%$ enantiomeric excess for the enantiomer that proceeds through the E-imine transition state and $-80\%$ enantiomeric excess for the Z-imine transition state. Remarkably, this represents a 3.5 kcal mol$^{-1}$ energy 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 across a general reaction type. Using this analysis, each distinct reaction (as determined by individual publications) in the data field was evaluated (coefficients are shown).

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 transition-state geometry (E or Z) determined by the relative sign of the enantiomeric excess defined previously, as these are hypothesized to lead to structurally distinct interactions with the other reaction components. This organizational scheme was viewed as a means of facilitating 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 resulting 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 the E-imine transition state. 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. However, this model emphasizes the importance of 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.
configurations. The most compelling difference between the two models is that the Z-imine model includes an important nucleophile steric descriptor, 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 transition state, leading to the competing product, which ultimately favours the observed enantiomer. This claim is further supported by the observation of high enantioselectivities when using catalysts with smaller substituents (for example, Ar = 3,5-(CF3)2C6H3). The proposed physical meanings of 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, and these predictions could then be further refined with the specific models (E or Z). This two-tiered workflow is imperative because the process avoids mechanistic assumptions about whether the reaction proceeds via an E or Z transition state, thus 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 Fig. 4. These are based on the amine product yielded from a reaction proceeding via an E or Z transition state and is catalysed by an (R)-CPA. Product stereochemistry is reversed if (S)-CPA is used.

**Fig. 4 | Out-of-sample predictions using two-tiered prediction workflow**. Comprehensive model first determines the E or Z transition state, configuration specific models are then used to refine predictions. A generic amine product denotes the stereochemical outcome predicted if the reaction proceeds via the E or Z transition state and is catalysed by an (R)-CPA. Product stereochemistry is reversed if (S)-CPA is used.
These examples showcase that the predictive capabilities of the model are not limited to classifying the vast literature, but can be applied to analyse 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 algorithms. The study reported by Denmark and co-workers involved the addition of thiols to benzoyl imines, a distinct reaction included in our training set. To utilize machine learning approaches, they performed 2,150 separate experiments using 43 catalysts to yield 25 different products (5 x 5 nucleophile/electrophile matrix). We postulated that our approach could reliably predict their results, including the best catalyst, TCYP (2,4,6-tricyclohexyl phenyl phosphoric acid), a CPA that 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—which lacks experimental data on this reaction—can also predict the enantioselectivities (average absolute ΔΔG° error = 0.65 kcal mol⁻¹ comprehensive model (26 examples within 5% enantiomeric excess), 0.67 kcal mol⁻¹ E-imine-only model (25 examples within 5% enantiomeric excess)), confidently determining the stereochemical outcome to be R and TCYP to be 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 that the comprehensive model is valuable because it determines which equation should be deployed.

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

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1384-z.

Received: 24 February 2019; Accepted: 29 May 2019;
Published online 17 July 2019.

1. Houk, K. N. & Cheong, P. H.-Y. Computational prediction of small-molecule catalysts. Nature 455, 309–313 (2008).
2. Davis, H. J. & Phipps, R. J. Harnessing non-covalent interactions to exert control over regioselectivity and site-selectivity in catalytic reactions. Chem. Sci. 8, 864–877 (2017).
3. Knowles, R. R. & Jacobsen, E. N. Attractive noncovalent interactions in asymmetric catalysis: links between enzymes and small molecule catalysts. Proc. Natl Acad. Sci. USA 107, 20678–20685 (2010).
4. Sigman, M. S., Harper, K. C., Bess, E. N. & Milo, A. The development of multimdimensional analysis tools for asymmetric catalysis and beyond. Acc. Chem. Res. 49, 1292–1301 (2016).
5. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C-N cross-coupling using machine learning. Science 360, 186–190 (2018).
6. Chuang, K. V. & Keiser, M. J. Comment on “Predicting reaction performance in C-N cross-coupling using machine learning”. Science 362, eaat6803 (2018).
7. Estrada, J. G., Ahneman, D. T., Sheridan, R. F., Dreher, S. D. & Doyle, A. G. Response to ‘Comment on “Predicting reaction performance in C-N cross-coupling using machine learning”’. Science 362, eaat7663 (2018).
8. Robbins, D. W. & Hartwig, J. F. A simple, multidimensional approach to high-throughput discovery of catalytic reactions. Science 333, 1423–1427 (2011).
9. McNally, A., Prier, C. K. & MacMillan, D. W. C. Discovery of an alpha-C-H arylation reaction using the strategy of accelerated serendipity. Science 334, 1114–1117 (2011).
10. Nekl, A. J., Milo, A., Sigman, M. S. & Toste, F. D. Enantiodivergent fluorination of allylic alcohols: dataset design reveals structural interplay between chiral directing group and chiral anion. J. Am. Chem. Soc. 138, 3863–3875 (2016).
11. Walsh, P. J. & Kozlowski, M. C. Fundamentals of Asymmetric Catalysis (University Science Books, 2008).
12. Yoon, T. P. & Jacobsen, E. N. Privileged chiral catalysts. Science 299, 1691–1693 (2003).
13. Yamamoto, H. Lewis Acids in Organic Synthesis (Wiley, 2000).
14. Akiyama, T. Stronger Bransted acids. Chem. Rev. 107, 5744–5758 (2007).
15. Collins, K. D. & Glorius, F. Intermolecular reaction screening as a tool for reaction evaluation. Acc. Chem. Res. 48, 619–627 (2015).
16. Gersmundo, N. J. et al. Nanoscale synthesis and affinity ranking. Nature 557, 228–232 (2018).
17. Reetz, M. T. 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, A. R., Tukutowski, B., Nobry, P.-O. & Wiest, O. Prediction of stereochemistry using QM2M. Acc. Chem. Res. 49, 996–1005 (2016).
19. Metsänen, T. 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 Prevymin(TM) (ertemorim). Chem. Soc. J. 6922–6927 (2018).
20. Robak, M. T., Herbage, M. A. & Eilman, J. A. Synthesis and applications of tert-butanesulfonamide. Chem. Rev. 110, 3600–3740 (2010).
21. Kobayashi, S., Mori, Y., Fossey, J. S. & Saltar, M. M. Catalytic enantioselective formation of C-C bonds by addition to imines and hydrazones: a ten-year update. Chem. Rev. 111, 2626–2704 (2011).
22. Nugent, T. C. Chiral Amine Synthesis: Methods, Developments and Applications (Wiley, 2010).
23. Silverio, D. L. et al. Simple organic molecules as catalysts for enantioselective synthesis of amines and alcohols. Nature 494, 216–221 (2013).
24. Parmar, D., Sugiono, E., Raja, S. & Rueping, M. Complete field guide to asymmetric BINOL-phosphate derived Bransted acid and metal catalysis: history and classification by mode of activation; Bransted acidity, hydrogen bonding, ion pairing, and metal phosphates. Chem. Rev. 114, 9047–9153 (2014).
25. Simón, L. & Goodman, J. M. 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).
26. Reid, J. P., Simón, L. & Goodman, J. M. A practical guide for predicting the stereochemistry of bifunctional phosphoric acid catalyzed reactions of imines. Acc. Chem. Res. 49, 1029 (2016).
27. Santiago, C. B., Guo, J.-Y. & Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. Chem. Sci. 9, 2398–2412 (2018).
28. Reid, J. P. & Sigman, M. S. Comparing quantitative prediction methods for the discovery of small-molecule chiral catalysts. Nat. Rev. Chem. 2, 290–305 (2018).
29. Denmark, S. E., Gould, N. D. & Wolf, L. M. 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).
30. Hansch, C. & Leo, A. Exploring QSAR: Fundamentals and Applications in Chemistry and Biology (ACS, 1995).
31. Reid, J. P. & Goodman, J. M. Goldilocks catalysts: computational insights into the role of the 3′ substitutions on the selectivity of BINOL-derived phosphoric acid catalysts. J. Am. Chem. Soc. 138, 7910–7917 (2016).
32. Terada, M., Machioka, K. & Sorimachi, K. High substrate/catalyst organocatalysis by a chiral Bransted 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 alkyl ketimines. J. Org. Chem. 83, 8688–8694 (2018).
34. Zeng, J. A. F. et al. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. Science 363, eaau5631 (2019).

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
METHODS
After the database of the reactions was 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 and the descriptor tables can be found in the Supplementary Information. The models produced were evaluated for their goodness of fit, $R^2$, and their robustness is demonstrated by external validation of the goodness of fit; the predicted $R^2$. The nearer the $R^2$ and slope values are to 1 (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. LORO analysis was performed to probe general mechanistic principles, which provides the basis for mechanistic transfer of experimental observations and tested further by predicting out-of-sample.

Data availability
All data relating to this study is available in the Supplementary Information.

Code availability
All code used for model development is available in the Supplementary Information.

Acknowledgements
J.P.R. thanks the EU Horizon 2020 Marie Skłodowska-Curie Fellowship (grant 792144) and M.S.S. thanks the NIH (grant GM-121383) 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 (grant ACI-1548562) and provided through allocation TG-CHE180003.

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

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1384-z.

Peer review information
Nature thanks Timothy Cernak, Per-Ola Norrby and Robert Paton for their contribution to the peer review of this work.

Correspondence and requests for materials
should be addressed to M.S.S.

Reprints and permissions information
is available at http://www.nature.com/reprints.
Extended Data Fig. 1 | Reaction component comparison. Parameterization challenges for the identification of numerical descriptors in reaction dimension, demonstrated using two reactions that represent the extremes of multidimensional feature space. MS, molecular sieves.