

Selectivity in organocatalysis—From qualitative to quantitative predictive models

Alistair J. Sterling  | Stamatia Zavitsanou  | Joseph Ford  |
 Fernanda Duarte 

Chemistry Research Laboratory,
 University of Oxford, Oxford, UK

Correspondence

Fernanda Duarte, Chemistry Research
 Laboratory, University of Oxford,
 Mansfield Road, Oxford OX1 3TA, UK.
 Email: fernanda.duarte@chem.ox.ac.uk

Funding information

Engineering and Physical Sciences
 Research Council, Grant/Award Number:
 EP/L015838/1; Vertex; UCB; Takeda
 Pharmaceutical Company; Syngenta;
 Pfizer; Novartis; Janssen Pharmaceuticals;
 GlaxoSmithKline; Evotec; Defense Science
 and Technology Laboratory; Diamond
 Light Source; AstraZeneca

Edited by: Peter Schreiner, Editor-in-
 Chief

Abstract

Recent advances in both experimental and computational techniques pose an exciting time for chemistry. Computational tools traditionally used to interpret experimental trends have now evolved into predictive models able to guide the design of novel catalysts. This review discusses the evolution of these models, as well as challenges and future avenues in the field of organocatalysis. Through representative examples we demonstrate how traditional physical organic chemistry tools in combination with machine learning models provide a powerful approach to achieve deeper understanding alongside greater predictive power.

This article is categorized under:

Structure and Mechanism > Reaction Mechanisms and Catalysis
 Electronic Structure Theory > Density Functional Theory
 Data Science > Artificial Intelligence/Machine Learning

KEYWORDS

machine learning, organocatalysis, physical organic chemistry, predictive models, reaction mechanism

1 | INTRODUCTION

Control over all manner of selectivity—from chemo- and regioselectivity to diastereo- and enantioselectivity—is at the core of current efforts in modern organic chemistry. Truly revolutionary advances have been made over the last few decades in the areas of organocatalysis, biocatalysis, photoredox, and transition metal catalysis which have led to increasing levels of sophistication and scope.¹ Organocatalysis, in particular, has become a powerful tool for modern synthetic chemistry due to its inherent mild reaction conditions, broad applicability, low toxicity, and cost.²

In parallel to these efforts, promising developments in computational organic chemistry have brought these two fields closer together. While computational modeling was originally considered as a simple interpretative tool to elucidate reaction mechanisms and infer reactivity trends, its predictive power has become increasingly appreciated. Advances in this direction have been made possible by ever-increasing computer power, the development of more

Alistair J. Sterling and Stamatia Zavitsanou contributed equally to this work.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *WIREs Computational Molecular Science* published by Wiley Periodicals LLC.

accurate and affordable electronic structure methods and software,^{3–5} and the introduction of statistical analyses and machine learning (ML) techniques.^{6,7}

Synergy between experimental and computational investigations has led to the development of simple qualitative and, more recently, quantitative models to understand the features that account for selectivity and predict properties such as catalytic activity and new reactivity (Figure 1).⁸ A remarkable example of such efforts has been the experimental and computational investigation of the intra- and intermolecular aldol reaction catalyzed by proline and its derivatives by List and coworkers.^{9–12} Houk and List provided the first stereoselectivity model commonly referred to as the Houk–List model, while Blackmond defined the kinetic rate law of the reaction. Since then, numerous studies have followed, providing a better understanding about the nature of this reaction and its broader applicability.¹³ This reaction has also served as a benchmark to improve and develop quantitative kinetic and computational approaches.^{14,15}

While in practice the discovery of new catalysts still relies on resource-intensive efforts, often guided by trial-and-error approaches, their rational and efficient design through computation has become a “Holy Grail” in organic chemistry.¹⁶ In this review, we summarize these efforts by briefly introducing the basic principles underlying selectivity and outline the computational challenges associated with their prediction. We discuss qualitative and quantitative approaches developed over the years to rationalize and predict selectivity. Prominent examples have been chosen in the area of organocatalysis, demonstrating the power of both classical empirical “rules” and more recent data-driven approaches in predicting reaction outcomes and novel designs.

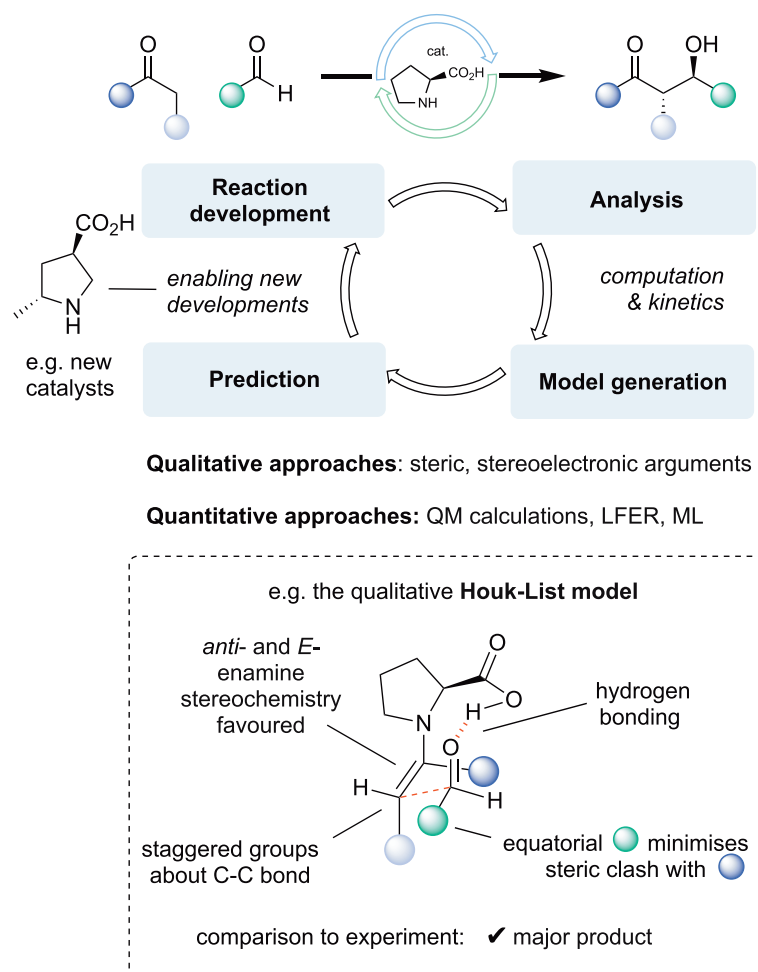


FIGURE 1 Synergistic relationship between the fields of organocatalysis and computational/kinetic analysis leading to the development of predictive models by qualitative and quantitative approaches. This relationship is exemplified by the Houk–List model for the proline-catalyzed Aldol reaction of carbonyl compounds. QM, quantum mechanics; LFER, linear free energy relationship; ML, machine learning

1.1 | Origin of selectivity

On a fundamental level, selectivity arises from the favorable formation of one product over another in the presence of several competing processes. This selectivity can be characterized into four groups (Figure 2(a)): (i) chemoselectivity, in which the manipulation of one functional group is favored over another; (ii) regioselectivity, in which the relative position of a functional group in a molecule determines its propensity to be manipulated; (iii) diastereoselectivity, in which the formation of one diastereomeric form of a molecule is favored over another; and (iv) enantioselectivity, in which the formation of one enantiomer of a molecule is favored over another. Moreover, selectivity can generally be controlled by either kinetics (relative rate constants for the formation of each product, Figure 2(b)), thermodynamics (relative stability of the products at equilibrium); or a combination of both, for example in a reaction with rapidly-interconverting intermediates (Curtin-Hammett control).^{17–19}

1.2 | Hierarchy of control of selectivity in organocatalysis

Traditional design approaches have placed sterics as the major controlling factor in organocatalysis. This is understandable as the electronic energy of two nonbonded atoms rises approximately exponentially with decreasing interatomic distance, making it possible to block all but one preferred pathway. However, a paradigm shift from sterics to attractive

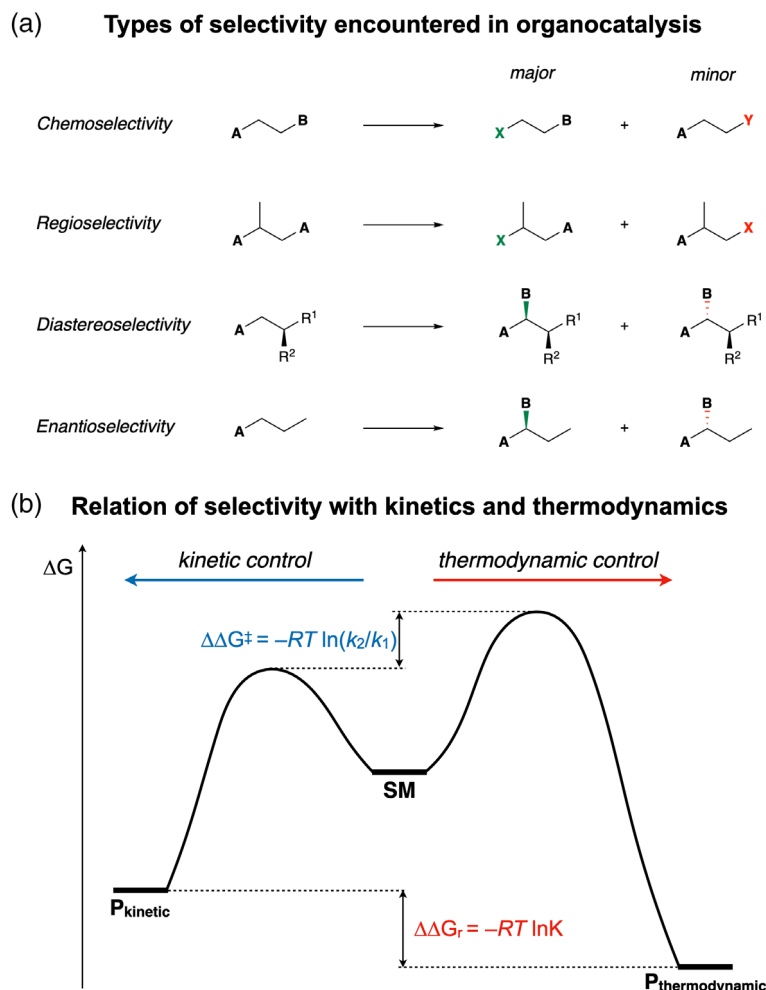


FIGURE 2 (a) The four types of selectivity encountered in organocatalysis. (b) Schematic representation of kinetic and thermodynamic control. k_i is the rate of formation of the i th species, K is the equilibrium constant, $\Delta\Delta G^\ddagger$ and $\Delta\Delta G_r$ are the difference in the free energy of the lowest energy transition states and products, respectively, R is the gas constant, and T is the temperature. P, product; SM, starting material

noncovalent interactions (NCIs) as major factors that control selectivity has occurred in the last few decades.^{20,21} For example, this has been achieved in the form of hydrogen and halogen bonding, electrostatic stabilization, π -stacking, and dispersion interactions.^{22,23} These interactions are in general weak and may not be large enough to overturn innate selectivity, however, may be significant in increasingly apolar environments and larger structures, where they can be thought of as “tuning knobs” that are twiddled to either favor or disfavor a given pathway. This paradigm shift has also led to a greater appreciation of the challenges associated with accurate computational predictions of these phenomena.

1.3 | Computational accuracy

State-of-the-art computational methods aim to achieve errors of less than ~ 1 kcal mol⁻¹ compared to experimental free energies, on systems of hundreds of atoms. Such accuracy can generally only be achieved through the accurate description of NCIs, which is necessary to achieve quantitative predictions given the exponential relationship between ΔG^\ddagger and the rate. Advances in linear scaling and local correlated techniques have allowed computational chemists to use the best possible form of electronic structure theory available. For example, the introduction of local natural orbital coupled cluster methods allows one to perform highly accurate energy calculations for large systems, including entire proteins.³ However, these advances have not guaranteed quantitative predictions in solution. In fact, many other aspects such as the treatment of entropy, thermal effects, solvation, and conformational flexibility have been found to introduce substantial errors when computing free energies (on the order of 10 kcal mol⁻¹).^{24,25} Efficient sampling techniques, compatible with the electronic structure methods, have been developed to account for conformational flexibility.²⁶ However, tackling other factors such as entropic effects in solution remains controversial; some authors argue that ideal gas expressions are sufficient and shortcomings arise from continuum solvent models,²⁵ while others suggest that different approaches are required to estimate (or scale) the translational component of the entropy.²⁷ Computational organocatalytic studies in explicit solvent are expected to become more accessible to achieve greater accuracy in the representations of experimental systems,^{28,29} however, their consideration in a quantitative manner remains to be seen.

Organocatalytic processes, and generally enantioselective reactions under kinetic control, are often sensitive to errors; small errors in the computed kinetics can even lead to the predicted enantiomer being the opposite to that observed experimentally. Figure 3 shows the relationship between the difference in experimental free energies of two competing processes and the observed selectivity, where an energy span of just 6 kcal mol⁻¹ is sufficient to overturn the selectivity from a $\sim 99:1$ to $\sim 1:99$ ratio of products at 25°C. Furthermore, in a mildly enantioselective reaction with an experimental *ee* of 70%, an error of ± 1 kcal mol⁻¹ could lead to predictions ranging from 0% to 93% *ee* at 25°C (Figure 3, gray shaded region). Fortunately, calculations of competing processes within an enantioselective reaction often benefit from error cancellation due to the structurally similar competing transition states (TSs), especially in the

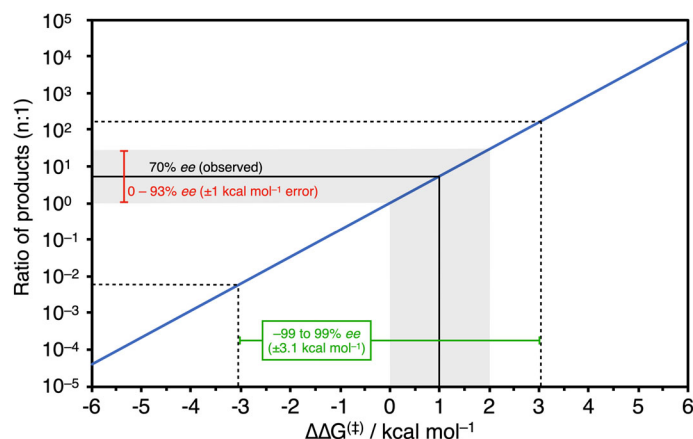


FIGURE 3 Relationship between selectivity ($\Delta\Delta G^\ddagger$ /kcal mol⁻¹) and the observed ratio of products (n:1) for two competing reactions at 25°C (blue line). Values in green show the span in $\Delta\Delta G^\ddagger$ to overturn selectivity from 99% to -99% *ee* (dashed lines). Solid black line and gray shaded region show the range in *ee* with an error in $\Delta\Delta G^\ddagger$ of ± 1 kcal mol⁻¹ for an observed *ee* of 70%

thermal and solvation correction terms, and simply using differences in electronic energy alone can sometimes be sufficient to find quantitative agreement with experiment.

Despite the challenges mentioned above, and as shown in the next paragraphs, there is a great deal of additional information gained from calculations that can aid in catalyst design, such as distortion/interaction analysis (also known as activation/strain analysis),³⁰ NCI analyses,³¹ localized orbital analysis,³² and changes in electron density from atomic charges and electrostatic potentials. Each of these analyses may be less sensitive to the level of theory employed in the calculation such that qualitative/semiquantitative predictions can be made with confidence.

2 | MODELS FOR PREDICTING SELECTIVITY

In principle, to predict selectivity, all TSs for each potential pathway need to be characterized, and the energy difference between the relevant states calculated with high accuracy. While this has been routine for computational chemists for decades,³³ locating critical points along reaction pathways remains a time-consuming and nonsystematic endeavor. Alternative approaches that avoid the direct calculation of TSs have relied on empirical rules derived from experiments, such as electronic Hammett³⁴ and steric Taft parameters,³⁵ and more recently statistical analyses.⁶ Recently, automated TS-finding algorithms have also provided systematic access to potential energy surfaces for catalyst design,³⁶ however, these are still limited by the accuracy of the electronic and thermodynamic corrections employed. Based on the type of data used and the range of applicability, predictive models may be divided into three categories:

1. Qualitative descriptor-based models.
2. Quantitative energy-based models.
3. ML-based models.

This separation is far from clear-cut, and it is often possible (and advantageous) to construct a model at the interface of these categories.

2.1 | Qualitative descriptor-based models

The development of empirical models for selectivity predictions is at the core of physical organic chemistry. These models can be generally discussed within the principles of the Klopman–Salem equation (Equation (1)). Formally described independently by Klopman and Salem in 1968,^{37,38} it expresses the energy increment incurred in a chemical reaction as the sum of three terms: closed-shell repulsion (sterics), Coulombic attraction or repulsion, and all possible interactions between pairs of occupied and vacant molecular orbitals of the reactants (stereoelectronic effects). The first term is often omitted from discussion, despite often being a controlling factor behind selectivity in organocatalysis. The remaining terms in the equation are used to describe reactions as “charge-controlled” or “orbital-controlled” according to whether $\Delta\Delta E_{\text{Coulomb}}$ or $\Delta\Delta E_{\text{stereoelectronic}}$ is dominant.

$$\Delta\Delta E^\ddagger = \Delta\Delta E_{\text{steric}} + \Delta\Delta E_{\text{Coulomb}} + \Delta\Delta E_{\text{stereoelectronic}} \quad (1)$$

Early predictive models of this form in organocatalysis were based on relatively simple back-of-the-envelope models supported by calculations. While these models do not necessarily reflect the active species in solution, they provided working principles that allowed the field to rapidly evolve. These include the steric-based model developed by MacMillan and coworkers³⁹ (Figure 4(a)) to rationalize the enantioselectivity and diastereoselectivity in a Diels–Alder reaction catalyzed by a chiral imidazolidinone. This model was based on force field-optimized catalyst–substrate complexes, where visual inspection revealed two salient stereocontrolling elements: (i) minimization of a steric clash between the substrate and the geminal methyl substituents, which is achieved with the (*E*)-iminium isomer and (ii) shielding of the *Re* face of the dienophile due to presence of the benzyl group on the catalyst, leaving the *Si* face exposed to cycloaddition. Following from this model, Ishihara and coworkers considered both sterics ($\Delta\Delta E_{\text{steric}}$) and electronic effects ($\Delta\Delta E_{\text{Coulomb}}$, Figure 4(b)) to rationalize enantio- and diastereoselectivity in a Diels–Alder reaction catalyzed by a chiral ammonium salt.⁴⁰ The design was based on the formation of tight ion pairs, where the steric clash between the anion

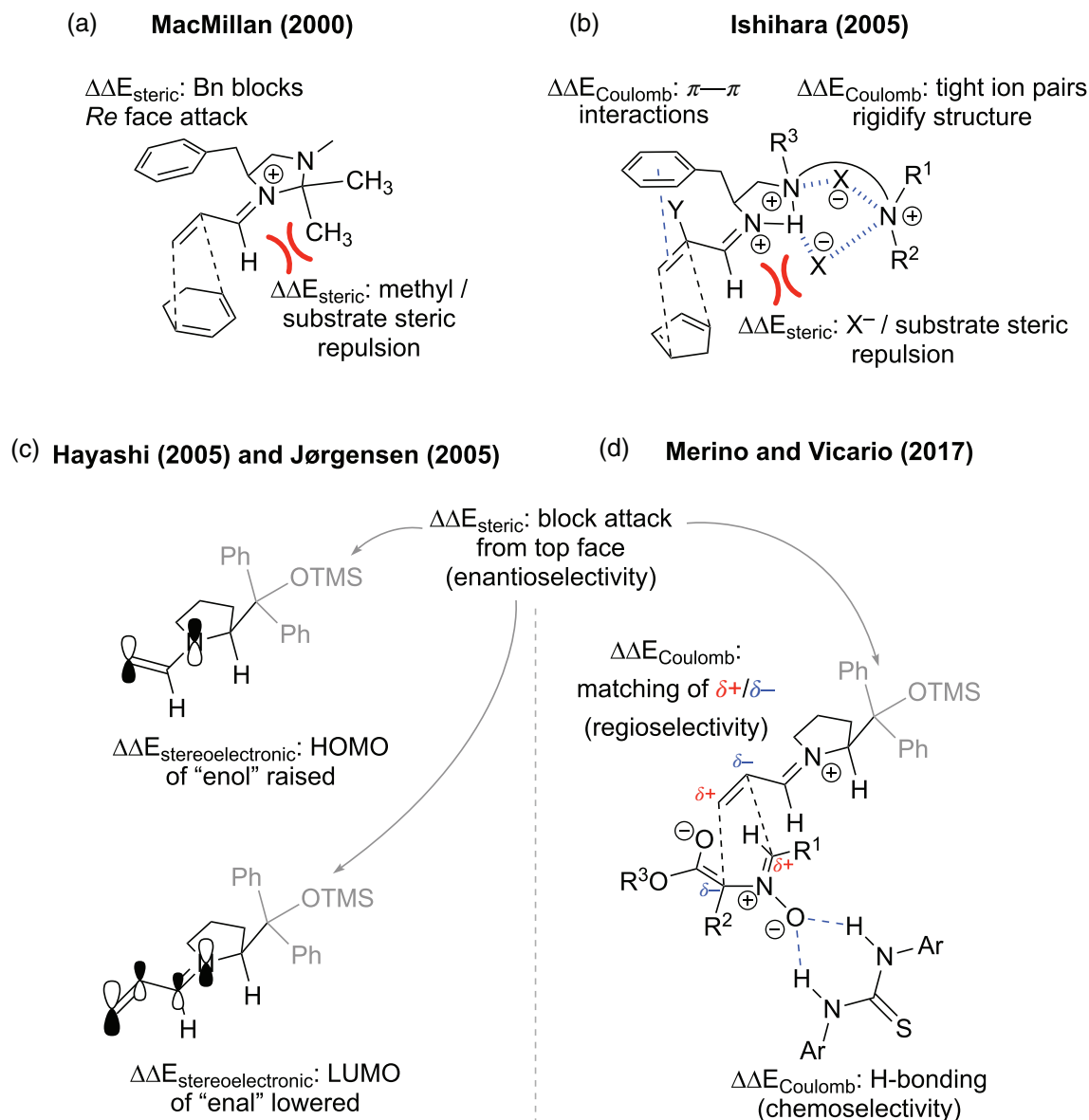


FIGURE 4 Qualitative models for organocatalytic selectivity based on the Klopman–Salem equation: Macmillan’s (a) and Ishihara’s (b) models for Diels–Alder reactions. (c) Mode of action of proline-derived organocatalysts developed by Hayashi and Jørgensen, and application in a 1,3-dipolar cycloaddition by Merino and Vicario (d)

and substrate once again favors the (*E*)-iminium isomer. Moreover, stabilizing $\pi-\pi$ interactions between the benzyl group and the iminium motif were proposed to favor the *Si-s-trans* TS conformation.

Stereoelectronic factors have also featured in qualitative models using the Hayashi–Jørgensen proline-derived catalytic system (Figure 4(c)), for example, where raising the highest occupied molecular orbital of an enol through enamine formation was suggested to induce chemoselectivity in electrophilic additions by favoring α -functionalization over competing nonproductive reactions.^{41,42} Conversely, enantioselective conjugate additions and Diels–Alder reactions could be catalyzed by lowering the lowest unoccupied molecular orbital of enals and enones through iminium ion formation. In each of these cases, the diarylprolinol silyl ether offered enantiodifferentiation through the steric bulk of the proline α -substituent blocking addition to the top face of the structure. Extending beyond the $\Delta\Delta E_{\text{steric}}$ term, Merino and Vicario designed a dual-organocatalytic system in which chemo-, regio-, diastereo-, and enantioselectivity is controlled (Figure 4(d)).⁴³ Using a chiral proline derivative and a hydrogen bond donor, a selective [3 + 2] cycloaddition between an enal and a 1,3-nitrone was developed in which nitrone 1,3-C,C attack was selected over 1,3-C,O attack

by forming hydrogen bonding interactions with the nitron oxygen. Regioselectivity can be rationalized using a hard/soft-acid/base-type model where the polarity of each atom in the cycloaddition is matched.

These simple selectivity models have also found use in enamine/iminium photoredox catalysis alongside mechanistic experiments (e.g., quantum yield measurements, trapping of intermediates, measurement of reduction potentials), notably by MacMillan and coworkers who have developed enantio- and diastereoselective functionalizations of aldehydes with both imidazolidinone and diarylprolinol catalysts.⁴⁴ Stereochemical outcomes of these reactions are generally rationalized based on the same steric models employed above, however, a combination of experimental and computational investigations have further characterized the controlling factors in this type of catalysis, including frontier orbital energies and spin density distributions.⁴⁵

Even though the classification of selectivity through sterics, charge- or orbital-controlled processes is widely used by organic chemists, this approach has been criticized due to the absence of the description of London dispersion, which can play an important role in determining selectivity.⁴⁶ In addition, this model is only valid to describe reactions where selectivity is determined by the kinetics of a single elementary step, which is not necessarily true for complex multistep reactions. This point has been raised by Mayr and coworkers, who instead propose the use of Marcus theory for both the qualitative and quantitative prediction of selectivity.⁴⁷ Using a combination of calculated free energy differences and empirical electrophilicity and nucleophilicity parameters, reactions can be classified as being thermodynamically controlled, diffusion controlled, or activation controlled, giving a more general basis from which to explain and predict reactivity trends.

2.2 | Quantitative energy-based models

Over the last two decades, advances in computer power and electronic structure methods have facilitated the extension of qualitative models to quantitative predictions. While the accuracy of these predictions remains to be seen, they have provided powerful working models capable of rationalizing experimental results and providing insights for new designs. Quantitative studies often focus on the calculation of reaction mechanisms using density functional theory (DFT) approaches, with a general workflow shown in Figure 5(a) that consists of (i) DFT optimizations of minima and transition states, from which thermodynamic corrections are obtained, (ii) conformational sampling at each identified stationary point; (iii) corrections to electronic energies using more accurate methods; and (iv) calculation of molecular properties, such as the molecular electrostatic potential (MEP), NCI plots,³¹ and natural bond orbital (NBO) analysis.³² Using the workflow mentioned above (or variants of it), several computational studies have elucidated the origins of selectivity in relevant organocatalytic reactions and achieved near-quantitative agreement with experiments. A recent example along these lines is the work of Bistoni and coworkers on a Diels-Alder reaction catalyzed by BALT and IDPi chiral anions developed by the List group (Figure 5(b)).⁴⁶ Following exhaustive conformational searches, the selectivity from their calculated TSs were found to be in reasonable agreement with experiment; while for the IDPi catalyst an error of just 0.3 kcal mol⁻¹ to experiment was observed, a 1.5 kcal mol⁻¹ difference was seen for the BALT catalyst, highlighting the fact that even state-of-the-art computational methods may not provide quantitative agreement with experiment. However, by employing a combination of MEP calculations and quantification of London dispersion interactions, they could reveal the factors controlling the binding modes of the activated species in the active site of the catalyst, as well as the delicate balance between sterics and dispersion forces for the control of enantioselectivity. While this approach might appear the most informative (although potentially the most time-consuming), there is often heavy reliance on experiment to understand which steps are the most important to model, through kinetic studies and by-product isolation. The generated models also may only be quantitative for the system under study and extension to other systems is not always guaranteed.

Another approach to obtaining quantitative predictions of enantioselectivity employs a combination of calculated TS barriers and geometric parameters which, in combination with linear regression, can be used to directly generate a quantitative predictive model. For example, Himo,⁵⁰ Terada,⁵¹ and Goodman^{52,53} have independently generated a model to predict selectivity in the nucleophilic addition to imines catalyzed by a chiral phosphoric acid (CPA). Goodman's model (Figure 5(c)) divides sterics into proximal and remote effects, where the former is determined by a parameter (A) based on the barrier to rotation of the 3,3' substituents, and the latter is based on the remote environment angle (AREA(θ)). The higher the rotation barrier, or the smaller the AREA parameter, the more sterically crowded the active pocket of the catalyst is. The preferred pathway can then be found by considering the configuration (*E/Z*) and the orientation of the imine when interacting with the catalyst cavity, resulting in four possible binding modes, with the

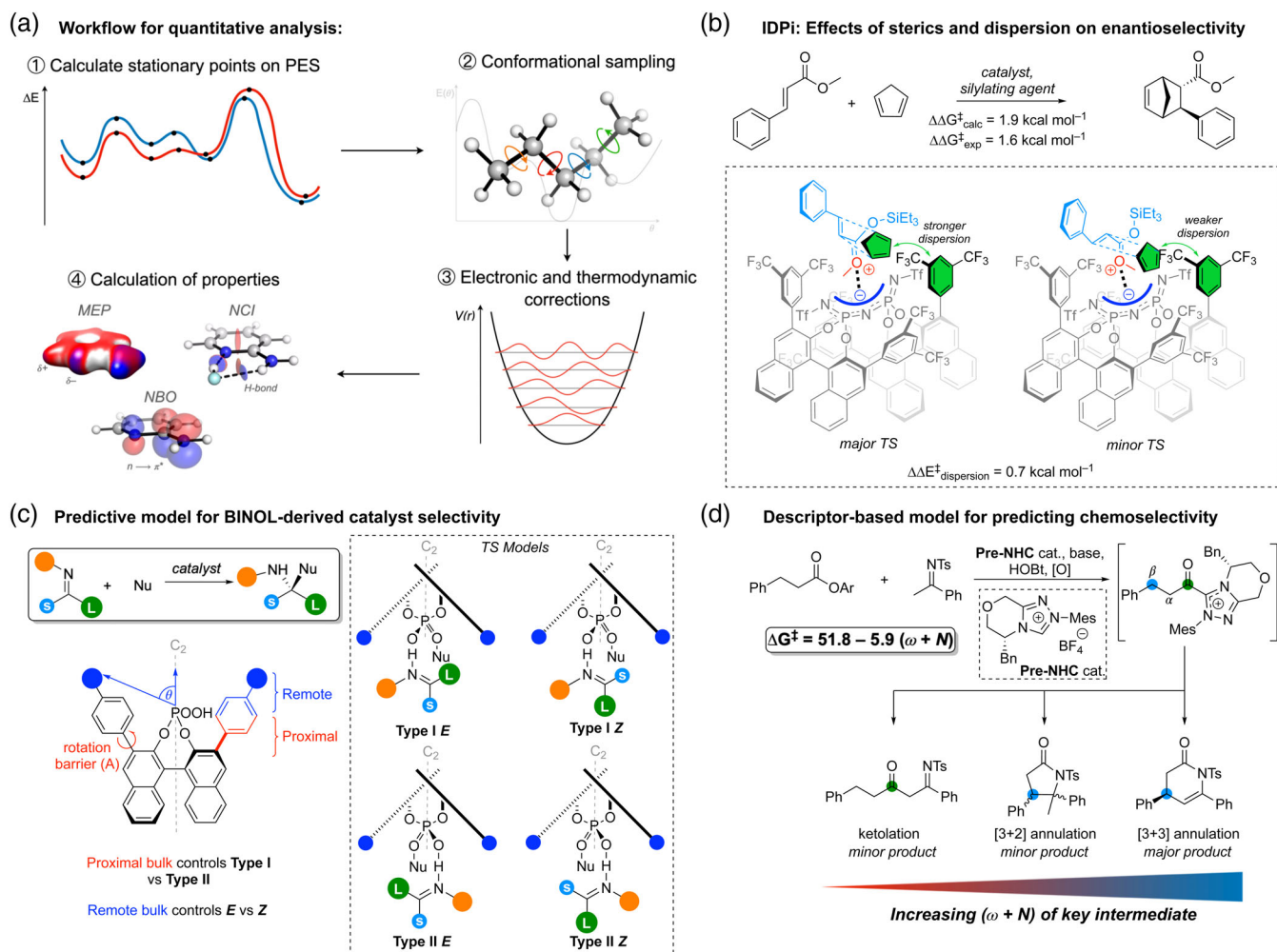


FIGURE 5 (a) Workflow for obtaining quantitative models for organocatalytic processes. (b) List's IDPi-catalyzed Diels-Alder reaction studied by Bistoni and coworkers⁴⁶ Tf = SO₂CF₃. (c) Goodman's model for the prediction of the favored TS in 1,1'-Bi-2-naphthol (BINOL)-derived catalytic reactions.⁴⁸ (d) Wei and Lan's model for an N-heterocyclic carbene (NHC)-catalyzed imine condensation using electrophilicity (ω) and nucleophilicity (N) parameters.⁴⁹

preferred one corresponding to that in which sterics are minimized between the substrate and catalyst substituents. It was found that the orientation of the imine was controlled by proximal sterics (i.e., correlates with parameter A) and the *E/Z* selectivity was controlled by remote sterics (AREA(θ) parameter). While originally developed for imine addition reactions, the simplicity and general applicability of this model has made it applicable to a much wider range of CPA-catalyzed reactions.⁴⁸

Beyond simple steric parameters to predict stereoselectivity, calculated electronic properties can be used to predict chemoselectivity, as shown by Wei and Lan.⁴⁹ They developed a quantitative model to predict the chemoselectivity of an N-heterocyclic carbene (NHC)-catalyzed ester functionalization reaction based on the global nucleophilicity (N) and electrophilicity (ω) indices for the nucleophile and electrophile involved in the product-determining step (PDS, Figure 5(d)). An inverse linear relationship was found between the sum of these two parameters and the calculated activation energy of the PDS for each competing process. Moreover, this model was successfully used to predict the major product in three similar reactions catalyzed by NHCs. It is worth noting that the calculation of the full PES was necessary to identify the PDSs for one catalytic system, however, once this information was obtained, the linear model could be applied to similar substrates and catalysts to predict the correct ordering of product distribution in each case.

Each of the models described above involved the correlation of a small number of features of the catalytic system with changes in selectivity. A natural extension of this method is to increase the number of features that are described, allowing greater variance in larger data sets with the hope of producing generalizable models for catalyst design.

2.3 | Machine learning-based models

The calculation of TSs and identification of the key steps of a chemical reaction can often be time-consuming. ML approaches have provided an alternative solution to overcome the current challenges relating to accuracy and computational cost from traditional quantum mechanics (QM)-based approaches. The prediction of reactivity and selectivity based on features of a set of molecules is the cornerstone of physical organic chemistry. Early studies from Hammett laid the groundwork for this approach, where linear free energy relationships (LFERs) were used to develop a quantitative relationship between structure and activity.³⁴ While the Hammett equation uses a single parameter (σ_x) for a given substituent, which accounts for how electronic effects influence reaction rates, variations of it also include a separation of these effects into field inductive (F) and resonance (R) effects through the Swain–Lupton equation⁵⁴ or steric effects via the Taft equation.^{35,55} Charton, influenced by Taft's work introduced a steric parameter, v , derived from van der Waals radii rather than kinetic data.⁵⁶ These early correlation analyses saw a renaissance with the development of quantitative structure-reactivity relationship (QSRR) methods and more recently with the introduction of ML approaches in chemistry.

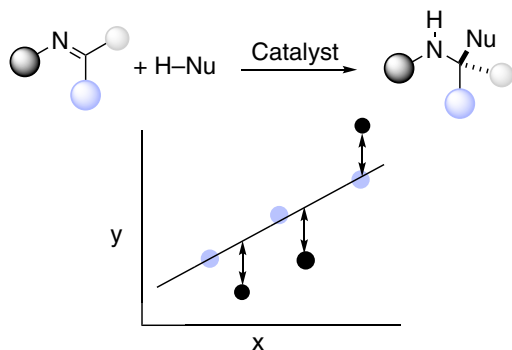
Today, a variety of chemical features (also commonly referred to as parameters or descriptors) are used to describe molecules and their reactions. These can be in the form of one-dimensional (InChI, SMILES, or SELFIES) notation, topological indices derived from chemical graph theory, or chemical descriptors obtained from electronic structure calculations. The latter includes partial atomic charges, electrostatic potentials, orbital energies, ionization energies and electron affinities, bond orders, and geometrical descriptors at minima and transition states, which can now be calculated at a low cost. With this number and breadth of descriptors and the development of widely accessible tools, like SciKit-Learn, TensorFlow, or Keras predictive models generated using ML have emerged as a promising alternative for quantitative prediction.⁵⁷

In the following paragraphs, we discuss how ML approaches, including multivariate linear regression (MLR), support vector machines (SVMs), decision trees (DTs), random forests (RFs), and neural networks (NNs),^{58,59} have been used to develop predictive selectivity models based on experimental data and molecular descriptors. Representative examples are discussed along with guidelines for the selection of a particular approach. There is no one-size-fits-all methodology, and the method of choice will depend on the research question and the data and computational resources available. While some of the ML approaches mentioned above have not yet been used in organocatalysis, we present representative examples of their use in predicting chemical selectivity. For each of the methods detailed above it is important to split the data set of interest into three subsets: *training set* (data used to fit the model), *validation set* (data used to fine-tune the model hyperparameters), and *test set* (data used to evaluate the final model). Alternatively, the data can be split into two subsets, training and test, with the former repeatedly split into a training and a validation data set in a process known as *cross-validation*. These techniques are used for assessing the generalizability of a predictive model and identifying problems such as overfitting, where the analysis of a particular set of data is not transferable to other data sets.^{60,61}

2.3.1 | MLR models

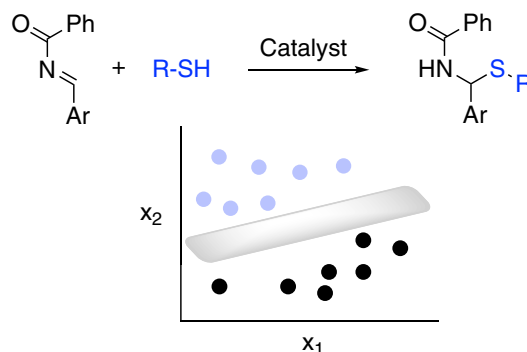
MLR models assume a linear relationship between the dependent variable y and the independent variables x_1, x_2, \dots, x_n . Accurate (R^2 as close to 1 as possible) QSRR and MLR models have been constructed for predicting reaction yields, rates, and (enantio)selectivities.⁶ While these models have the advantage of providing interpretable, easy to implement and robust models, they are sensitive to outliers and multicollinearity, where two or more variables are highly related. Sigman and coworkers have popularized the use of MLR for prediction in organocatalysis.⁶² For example, they developed predictive models for the CPA-catalyzed nucleophilic additions to imines using steric (e.g., Sterimol, bond lengths, bond angles) and electronic (e.g., vibrational frequencies and intensities, NPA charges, polarizabilities, orbital energies) parameters to describe the substrates and catalysts (Figure 6(a)). Despite the large structural variance in the data set, they constructed an accurate model using 313 descriptors to describe 350 reactions ($R^2 = 0.88$). To obtain chemical insights from the statistical models, the data set was divided into subsets categorized by imine transition-state geometry (E or Z). The model recovers the importance of sterics, with large catalyst and imine substituents leading to higher levels of enantioselectivity for the E -imine model. For the Z -imine model, catalysts with smaller substituents were found to be preferred. This approach is complementary to that employed by Goodman (vide supra), and places a similar importance of steric bulk on enantioselectivity.

(a) Multivariate linear regression



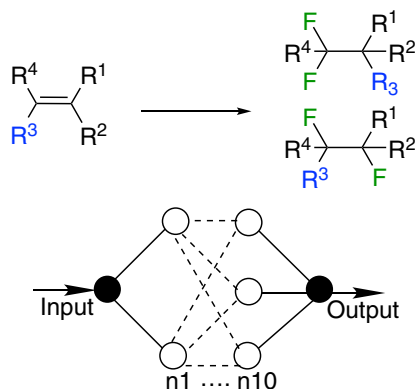
- ✓ Easy to interpret
- ✓ Low computational cost
- ✓ Any amount of data
- ✗ Multicollinearity
- 313 parameters
- 350 reactions
- $R^2 = 0.88$

(b) Support vector machines



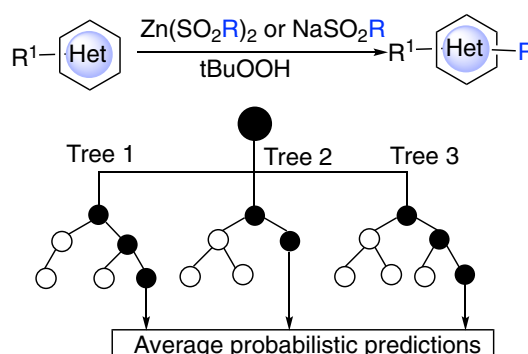
- ✓ Low computational cost
- ✓ High accuracy
- ✗ Difficult to represent
- ✗ Slow for large datasets
- 2000 parameters
- 1075 reactions
- $R^2 = 0.91$

(c) Neural networks



- ✓ Regression and classification
- ✓ High accuracy
- ✗ Hardware dependence
- ✗ Difficult to interpret
- 11 parameters
- 66 reactions
- $R^2 = 0.90$

(d) Random forests



- ✓ Easy to interpret
- ✓ High accuracy
- ✗ High computational cost
- ✗ Slow to train
- 32 parameters
- 8580 reactions
- $R^2 = 0.90$

FIGURE 6 Examples of the ML models and their respective applications in chemistry. (a) Prediction of enantioselectivity for CPA-catalyzed additions to imines.⁶² (b) Prediction of enantioselectivity for the addition of thiols to *N*-acyl imines.⁶³ (c) Prediction of regioselectivity for difluorination reactions catalyzed by hypervalent iodine.⁶⁵ (d) Prediction of regioselectivity in radical C–H functionalization.⁶⁶ CPA, chiral phosphoric acid; ML, machine learning

2.3.2 | Support vector machines

SVM models are binary classifiers that have been used for the prediction of enantioselectivity in organocatalytic reactions. They have low computational cost, but as the number of descriptors increases, the visualization of the boundary between classes becomes difficult. Denmark and coworkers developed an SVM model to predict enantioselectivity (*ee*) for the nucleophilic addition of thiols to *N*-acyl imines (Figure 6(b)).⁶³ The model was built using data from 1075 reactions, employing a new shape descriptor, the average steric occupancy (ASO), alongside electronic parameters (16,384 features). To reduce the huge number of parameters, they performed feature selection and principal component analysis (PCA). After testing different ML approaches, they concluded that a second-order polynomial kernel SVM provided the best model to describe their data ($R^2 = 0.91$) and achieved accurate prediction of higher-performing catalysts, even when only reactions below 80% *ee* were used in the training set. The model predicted the most selective catalyst (96.5% *ee*) within 3% *ee* of the experimental value, suggesting the applicability of the model beyond the training set in finding more selective catalysts.

2.3.3 | Neural networks

NNs aim to mimic the sequence of actions taking place in the human brain. First, an *input* layer receives the descriptors for the model and passes the information on to the *hidden* layers, which apply various transformations to these descriptors, through an activation function. Finally, an *output* layer returns the value that corresponds to the prediction. NNs are widely used in both regression and classification analysis due to their ability to process data quickly and yield good performance and high accuracy. However, they can be difficult to interpret. In the context of organocatalysis, few examples have been reported so far, however, promising ones have been shown in predicting selectivity in reactions more generally. Jensen and coworkers employed NNs to predict the major product of 15,000 reactions (chlorination, amide synthesis, isoxazole synthesis, sulfamide synthesis, etherification, Suzuki coupling, azidation, and alkylation, among others).⁶⁴ The model was constructed with 1055 features, including Morgan fingerprints, number of hydrogen atoms and atomic number. An accuracy of 71.8% was achieved in predicting the major product. Another example demonstrating the use of NNs in this field was reported by Sunoj and coworkers, where the regiochemical outcomes of 66 regioselective difluorination reactions of alkenes catalyzed by hypervalent iodine were predicted (Figure 6(c)).⁶⁵ Employing 63 features, such as charges, nuclear magnetic resonance shifts, electrophilic and nucleophilic Fukui indices, steric parameters, (reduced to 11 after analysis), they achieved an average accuracy of 90%. In this work, the authors noted that NNs have limited interpretability and they therefore generated a DT model to identify the descriptors responsible for the regioselectivity outcomes. From this model, it was found that 1,2-difluorination is favored with electron deficient benzylic carbons, whereas electron-rich terminal carbons exclusively lead to 1,1-difluorinated products.

2.3.4 | Decision trees/Random forests

DTs can be thought of as a flowchart of different “if/then” statements or a set of questions that calculate the probability of a data point belonging to a class. The deeper the tree, the more complex the decision rules and, in principle, the more accurate the model. DT models are conceptually easy to understand; however, they are prone to overfitting due to the specificity and strict rules followed by the statements (a data point has to follow every rule to be sorted into a specific class), and they generally have higher error rates than other methods.⁵⁸ RF models alleviate these problems by considering the aggregate of many DTs, each making their own predictions, which are then merged for a final collective decision. While RFs models show good performance, they are computationally costly and slow to train. Hong and coworkers trained a RF model to predict regioselectivity in radical C–H functionalization reactions (Figure 6(d)).⁶⁶ In total, 50 types of descriptors were used to describe both steric and electronic effects (smooth overlap of atomic positions (SOAPs), buried volume, molecular fingerprints, frontier molecular orbital (FMO) energies, and atomic charge, among others). After analysis, the authors ultimately implemented a predictive model using 32 physical organic descriptors to achieve an accuracy of 90%, advising against the use of SOAP (15,876 descriptors) and molecular fingerprints (1358 descriptors) which would require a significantly larger training set to match the feature space. The use of a smaller set of physical organic descriptors also facilitated interpretation of the model, which revealed that the innate reactivity of heteroarenes played a dominating role in the regiocontrol, and the radical acts as an additional important descriptor on the innate regioselectivity of heteroarenes.

The examples above demonstrate the potential but also the limitations of ML methods in the prediction of selectivity in chemical reactions; further investigations to improve interoperability and identify the applicability regime of these models are necessary. In general, we suggest that MLR and SVM are more appropriate for small data sets, as they are more robust and are less prone to overfitting. In contrast, for larger data sets, RFs and NNs can be powerful approaches as they can identify patterns in complex data. In this context, Jorner et al. have identified separate data regimes for the accurate prediction of S_NAr reaction barriers using a hybrid mechanistic/ML model.⁶⁷ They showed that for databases smaller than 50 samples traditional DFT modeling is the more accurate choice, while for data sets between 50 and 150 samples hybrid approaches are suitable. Only for diverse data sets with >300 samples models based on purely structural information become competitive. It remains to be seen if such a regime can be generalized to other reaction classes.

Despite limitations on the size of currently used data sets, we anticipate that the use of deep learning in organocatalysis will continue to expand in the coming years; especially, considering its recent use to predict molecular energies at the DFT level for the computational cost of a traditional force field.⁶⁸ However, these promising examples remain to prove their applicability in more complex catalytic systems. While current efforts have focused on developing

ML models that can reproduce QM accuracies, more recent developments focus on using experimental data as it becomes more readily available to the chemistry community, bypassing the challenges associated with QM approaches, such as thermal and solvation corrections. While deep learning models may be trained on reaction databases containing millions of patent and literature data, with active learning and reinforcement learning there is the potential to construct accurate predictive models with limited data. We are hopeful that these techniques will become more prevalent in the near future, leading to models capable of accurately predicting experimental outcomes within seconds.

3 | CONCLUSIONS AND OUTLOOK

Recent developments in electronic structure methods, software, and the increasing level of automation in synthesis have the potential to revolutionize the way catalyst design is performed, freeing chemists from repetitive tasks and allowing them to explore a much wider chemical space. These opportunities have brought the synthetic and computational chemistry communities together, and demonstrate the power of integrating empirical knowledge and automation in catalyst design. In this review, we have summarized the traditional computational models to elucidate the origins of selectivity in organocatalysis, which have evolved into quantitative predictive models able to tackle challenging systems and guide synthetic efforts. No longer just relying on simple steric models, the current state-of-the-art in the generation of these predictive models can involve quantum-mechanical calculations on hundreds of atoms, or the use of thousands of chemical descriptors to elucidate patterns in multidimensional data. While challenges still remain in terms of the predictivity, interpretability and generalizability of these models, it is clear that their use in the field of organocatalysis will increase in the near future. The revolution in computer power and ML is here to stay; it will be the role of computational chemists to work at the interface of those fields and deliver solutions that advance developments in chemistry and address current societal challenges.

ACKNOWLEDGMENTS

A. J. S. and J. F. thank the EPSRC Centre for Doctoral Training in Synthesis for Biology and Medicine for studentships (EP/L015838/1), generously supported by AstraZeneca, Diamond Light Source, Defense Science and Technology Laboratory, Evotec, GlaxoSmithKline, Janssen, Novartis, Pfizer, Syngenta, Takeda, UCB, and Vertex. A. J. S. also thanks the Oxford-Radcliffe Scholarship for a studentship. We also thank Tanya Rogova and Tom Young for comments on the manuscript.

AUTHOR CONTRIBUTIONS

Alistair Sterling: Writing-original draft. **Stamatia Zavitsanou:** Writing-original draft. **Joseph Ford:** Writing-review and editing. **Fernanda Duarte:** Writing-original draft.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ORCID

Alistair J. Sterling  <https://orcid.org/0000-0002-3571-1094>

Stamatia Zavitsanou  <https://orcid.org/0000-0002-0571-2264>

Joseph Ford  <https://orcid.org/0000-0001-7959-2085>

Fernanda Duarte  <https://orcid.org/0000-0002-6062-8209>

RELATED WIREs ARTICLES

[In silico toxicology: computational methods for the prediction of chemical toxicity](#)

[Hybrid discrete-continuum solvation methods](#)

[NCIPLOT and the analysis of noncovalent interactions using the reduced density gradient](#)

[Proline-derived organocatalysis and synergism between theory and experiments](#)

REFERENCES

1. Busacca CA, Fandrick DR, Song JJ, Senanayake CH. The growing impact of catalysis in the pharmaceutical industry. *Adv Synth Catal.* 2011;353(11–12):1825–64.

2. Xiang S-H, Tan B. Advances in asymmetric organocatalysis over the last 10 years. *Nat Commun.* 2020;11(1):3786.
3. Nagy PR, Kállay M. Approaching the basis set limit of CCSD(T) energies for large molecules with local natural orbital coupled-cluster methods. *J Chem Theory Comput.* 2019;15(10):5275–98.
4. Gordon MS, Barca G, Leang SS, Poole D, Rendell AP, Galvez Vallejo JL, et al. Novel computer architectures and quantum chemistry. *J Phys Chem A.* 2020;124(23):4557–82.
5. Dral PO. Quantum chemistry in the age of machine learning. *J Phys Chem Lett.* 2020;11(6):2336–47.
6. Santiago CB, Guo J-Y, Sigman MS. Predictive and mechanistic multivariate linear regression models for reaction development. *Chem Sci.* 2018;9(9):2398–412.
7. Foscatto M, Jensen VR. Automated in silico design of homogeneous catalysts. *ACS Catal.* 2020;10(3):2354–77.
8. Peng Q, Duarte F, Paton RS. Computing organic stereoselectivity—from concepts to quantitative calculations and predictions. *Chem Soc Rev.* 2016;45(22):6093–107.
9. Bahmanyar S, Houk KN, Martin HJ, List B. Quantum mechanical predictions of the stereoselectivities of proline-catalyzed asymmetric intermolecular Aldol reactions. *J Am Chem Soc.* 2003;125(9):2475–9.
10. Hoang L, Bahmanyar S, Houk KN, List B. Kinetic and stereochemical evidence for the involvement of only one proline molecule in the transition states of proline-catalyzed intra- and intermolecular Aldol reactions. *J Am Chem Soc.* 2003;125(1):16–7.
11. Zotova N, Franzke A, Armstrong A, Blackmond DG. Clarification of the role of water in proline-mediated Aldol reactions. *J Am Chem Soc.* 2007;129(49):15100–1.
12. Zotova N, Broadbelt LJ, Armstrong A, Blackmond DG. Kinetic and mechanistic studies of proline-mediated direct intermolecular Aldol reactions. *Bioorg Med Chem Lett.* 2009;19(14):3934–7.
13. Yang JW, Chandler C, Stadler M, Kampen D, List B. Proline-catalysed Mannich reactions of acetaldehyde. *Nature.* 2008;452(7186):453–5.
14. Armstrong A, Boto RA, Dingwall P, Contreras-García J, Harvey MJ, Mason NJ, et al. The Houk–List transition states for organocatalytic mechanisms revisited. *Chem Sci.* 2014;5(5):2057–71.
15. Orlandi M, Ceotto M, Benaglia M. Kinetics versus thermodynamics in the proline catalyzed Aldol reaction. *Chem Sci.* 2016;7(8):5421–7.
16. Houk KN, Liu F. Holy grails for computational organic chemistry and biochemistry. *Acc Chem Res.* 2017;50(3):539–43.
17. Seeman JJ. Effect of conformational change on reactivity in organic chemistry. Evaluations, applications, and extensions of Curtin–Hammett Winstein–Holness kinetics. *Chem Rev.* 1983;83(2):83–134.
18. Berson JA. Kinetics, thermodynamics, and the problem of selectivity: the maturation of an idea. *Angew Chem Int Ed.* 2006;45(29):4724–9.
19. Burés J, Armstrong A, Blackmond DG. Curtin–Hammett paradigm for stereocontrol in organocatalysis by diarylprolinol ether catalysts. *J Am Chem Soc.* 2012;134(15):6741–50.
20. Wagner JP, Schreiner PR. London dispersion in molecular chemistry—reconsidering steric effects. *Angew Chem Int Ed.* 2015;54(42):12274–96.
21. Wheeler SE, Seguin TJ, Guan Y, Doney AC. Noncovalent interactions in organocatalysis and the prospect of computational catalyst design. *Acc Chem Res.* 2016;49(5):1061–9.
22. Phipps RJ, Hamilton GL, Toste FD. The progression of chiral anions from concepts to applications in asymmetric catalysis. *Nat Chem.* 2012;4(8):603–14.
23. Sutar RL, Huber SM. Catalysis of organic reactions through halogen bonding. *ACS Catal.* 2019;9(10):9622–39.
24. Plata RE, Singleton DA. A case study of the mechanism of alcohol-mediated Morita Baylis–Hillman reactions. The importance of experimental observations. *J Am Chem Soc.* 2015;137(11):3811–26.
25. Harvey JN, Himo F, Maseras F, Perrin L. Scope and challenge of computational methods for studying mechanism and reactivity in homogeneous catalysis. *ACS Catal.* 2019;9(8):6803–13.
26. Pracht P, Bohle F, Grimme S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys Chem Chem Phys.* 2020;22(14):7169–92.
27. Wertz DH. Relationship between the gas-phase entropies of molecules and their entropies of solvation in water and 1-octanol. *J Am Chem Soc.* 1980;102(16):5316–22.
28. Duarte F, Paton RS. Molecular recognition in asymmetric counteranion catalysis: understanding chiral phosphate-mediated desymmetrization. *J Am Chem Soc.* 2017;139(26):8886–96.
29. Pupo G, Ibba F, Ascough DMH, Vicini AC, Ricci P, Christensen KE, et al. Asymmetric nucleophilic fluorination under hydrogen bonding phase-transfer catalysis. *Science.* 2018;360(6389):638–42.
30. Bickelhaupt FM, Houk KN. Analyzing reaction rates with the distortion/interaction-activation strain model. *Angew Chem Int Ed.* 2017;56(34):10070–86.
31. Johnson ER, Keinan S, Mori-Sánchez P, Contreras-García J, Cohen AJ, Yang W. Revealing noncovalent interactions. *J Am Chem Soc.* 2010;132(18):6498–506.
32. Weinhold F. Natural bond orbital analysis: a critical overview of relationships to alternative bonding perspectives. *J Comput Chem.* 2012;33(30):2363–79.
33. Houk KN, Paddon-Row MN, Rondan NG, Wu YD, Brown FK, Spellmeyer DC, et al. Theory and modeling of stereoselective organic reactions. *Science.* 1986;231(4742):1108–17.
34. Hammett LP. Some relations between reaction rates and equilibrium constants. *Chem Rev.* 1935;17(1):125–36.

35. Taft RW. Linear free energy relationships from rates of esterification and hydrolysis of aliphatic and ortho-substituted benzoate esters. *J Am Chem Soc.* 1952;74(11):2729–32.
36. Guan Y, Ingman VM, Rooks BJ, Wheeler SE. AARON: an automated reaction optimizer for new catalysts. *J Chem Theory Comput.* 2018;14(10):5249–61.
37. Klopman G. Chemical reactivity and the concept of charge- and frontier-controlled reactions. *J Am Chem Soc.* 1968;90(2):223–34.
38. Salem L. Intermolecular orbital theory of the interaction between conjugated systems. I. General theory. *J Am Chem Soc.* 1968;90(3):543–52.
39. Ahrendt KA, Borths CJ, MacMillan DWC. New strategies for organic catalysis: the first highly enantioselective organocatalytic Diels–Alder reaction. *J Am Chem Soc.* 2000;122(17):4243–4.
40. Ishihara K, Nakano K. Design of an organocatalyst for the enantioselective Diels–Alder reaction with α -acyloxyacroleins. *J Am Chem Soc.* 2005;127(30):10504–5.
41. Franzén J, Marigo M, Fielenbach D, Wabnitz TC, Kjærsgaard A, Jørgensen KA. A general organocatalyst for direct α -functionalization of aldehydes: stereoselective C–C, C–N, C–F, C–Br, and C–S bond-forming reactions. Scope and mechanistic insights. *J Am Chem Soc.* 2005;127(51):18296–304.
42. Hayashi Y, Gotoh H, Hayashi T, Shoji M. Diphenylprolinol silyl ethers as efficient organocatalysts for the asymmetric Michael reaction of aldehydes and nitroalkenes. *Angew Chem Int Ed.* 2005;44(27):4212–5.
43. Prieto L, Juste-Navarro V, Uriá U, Delso I, Reyes E, Tejero T, et al. Regioselectivity change in the organocatalytic enantioselective (3+2) cycloaddition with nitrones through cooperative hydrogen-bonding catalysis/iminium activation. *Chem Eur J.* 2017;23(12):2764–8.
44. Nicewicz DA, MacMillan DWC. Merging photoredox catalysis with organocatalysis: the direct asymmetric alkylation of aldehydes. *Science.* 2008;322(5898):77–80.
45. Um JM, Gutierrez O, Schoenebeck F, Houk KN, MacMillan DWC. Nature of intermediates in organo-SOMO catalysis of α -arylation of aldehydes. *J Am Chem Soc.* 2010;132(17):6001–5.
46. Yepes D, Neese F, List B, Bistoni G. Unveiling the delicate balance of steric and dispersion interactions in organocatalysis using high-level computational methods. *J Am Chem Soc.* 2020;142(7):3613–25.
47. Mayr H, Breugst M, Ofial AR. Farewell to the HSAB treatment of ambident reactivity. *Angew Chem Int Ed.* 2011;50(29):6470–505.
48. Reid JP, Goodman JM. Selecting chiral BINOL-derived phosphoric acid catalysts: general model to identify steric features essential for enantioselectivity. *Chem-Eur J.* 2017;23(57):14248–60.
49. Li X, Xu J, Li S-J, Qu L-B, Li Z, Chi YR, et al. Prediction of NHC-catalyzed chemoselective functionalizations of carbonyl compounds: a general mechanistic map. *Chem Sci.* 2020;11(27):7214–25.
50. Marcelli T, Hammar P, Himo F. Phosphoric acid catalyzed enantioselective transfer hydrogenation of imines: a density functional theory study of reaction mechanism and the origins of enantioselectivity. *Chem Eur J.* 2008;14(28):8562–71.
51. Gridnev ID, Kouchi M, Sorimachi K, Terada M. On the mechanism of stereoselection in direct Mannich reaction catalyzed by BINOL-derived phosphoric acids. *Tetrahedron Lett.* 2007;48(3):497–500.
52. 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.* 2008;130(27):8741–7.
53. Reid JP, Simón L, Goodman JM. A practical guide for predicting the stereochemistry of bifunctional phosphoric acid catalyzed reactions of imines. *Acc Chem Res.* 2016;49(5):1029–41.
54. Swain CG, Lupton EC. Field and resonance components of substituent effects. *J Am Chem Soc.* 1968;90(16):4328–37.
55. Taft RW, Lewis IC. Evaluation of resonance effects on reactivity by application of the linear inductive energy relationship. V. Concerning a σ_R scale of resonance effects 1,2. *J Am Chem Soc.* 1959;81(20):5343–52.
56. Charton M. Steric effects. I. Esterification and acid-catalyzed hydrolysis of esters. *J Am Chem Soc.* 1975;97(6):1552–6.
57. Géron A. *Hands-on machine learning with Scikit-learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems.* Vol XXV. 2nd ed. Beijing: O'Reilly; 2019 819 pages.
58. Shalev-Shwartz S, Ben-David S. *Understanding machine learning: from theory to algorithms.* Vol XVI. Cambridge University Press: New York; 2014 397 pages.
59. Elton DC, Boukouvalas Z, Fuge MD, Chung PW. Deep learning for molecular design—a review of the state of the art. *Mol Syst Design Eng.* 2019;4(4):828–49.
60. Hawkins DM. The problem of overfitting. *J Chem Inf Comput Sci.* 2004;44(1):1–12.
61. Zahrt AF, Athavale SV, Denmark SE. Quantitative structure–selectivity relationships in enantioselective catalysis: past, present, and future. *Chem Rev.* 2020;120(3):1620–89.
62. Reid JP, Sigman MS. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature.* 2019;571(7765):343–8.
63. Zahrt AF, Henle JJ, Rose BT, Wang Y, Darrow WT, Denmark SE. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science.* 2019;363(6424):eaau5631.
64. Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF. Prediction of organic reaction outcomes using machine learning. *ACS Central Sci.* 2017;3(5):434–43.
65. Banerjee S, Sreenithya A, Sunoj RB. Machine learning for predicting product distributions in catalytic regioselective reactions. *Phys Chem Chem Phys.* 2018;20(27):18311–8.
66. Li X, Zhang S-Q, Xu L-C, Hong X. Predicting regioselectivity in radical C–H functionalization of heterocycles through machine learning. *Angew Chem Int Ed.* 2020;59(32):13253–9.

67. Jorner K, Brinck T, Norrby P-O, Buttar D. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chem Sci.* 2020;9(99):1163–75.
68. Smith JS, Isayev O, Roitberg AE. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem Sci.* 2017;8(4):3192–203.

How to cite this article: Sterling AJ, Zavitsanou S, Ford J, Duarte F. Selectivity in organocatalysis—From qualitative to quantitative predictive models. *WIREs Comput Mol Sci.* 2021;11:e1518. <https://doi.org/10.1002/wcms.1518>