CHEMICAL REVIEWS

Review

Cite This: *Chem. Rev.* 2020, 120, 1620−1689

pubs.acs.org/CR

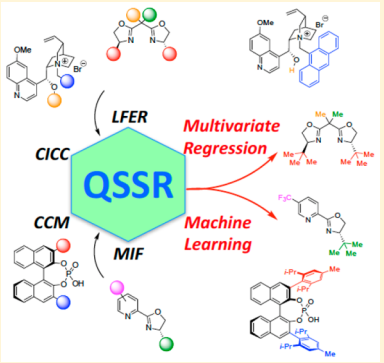
Quantitative Structure−Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future

Andrew F. Zahrt, Soumitra V. Athavale, and Scott E. Denmark*

Roger Adams Laboratory, Department of Chemistry, University of Illinois, Urbana, Illinois 61801, United States

Supporting Information

ABSTRACT: The dawn of the 21st century has brought with it a surge of research related to computer-guided approaches to catalyst design. In the past two decades, chemoinformatics, the application of informatics to solve problems in chemistry, has increasingly influenced prediction of activity and mechanistic investigations of organic reactions. The advent of advanced statistical and machine learning methods, as well as dramatic increases in computational speed and memory, has contributed to this emerging field of study. This review summarizes strategies to employ quantitative structure−selectivity relationships (QSSR) in asymmetric catalytic reactions. The coverage is structured by initially introducing the basic features of these methods. Subsequent topics are discussed according to increasing complexity of molecular representations. As the most applied subfield of QSSR in enantioselective catalysis, the application of local parametrization approaches and linear free energy relationships (LFERs) along with multivariate modeling techniques is described first. This section is followed by a description of global parametrization methods, the first of which is continuous chirality measures (CCM) because it is a single parameter derived from the global structure of a molecule. Chirality codes, global, multivariate descriptors, are then introduced followed by molecular interaction fields (MIFs), a global descriptor class that typically has the highest dimensionality. To highlight the current reach of QSSR in enantioselective transformations, a comprehensive collection of examples is presented. When combined with traditional experimental approaches, chemoinformatics holds great promise to predict new catalyst structures, rationalize mechanistic behavior, and profoundly change the way chemists discover and optimize reactions.



CONTENTS

Received: July 2, 2019
Published: December 30, 2019

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

1. INTRODUCTION

1.1. Chemoinformatics in Asymmetric Catalysis

Understanding the correlation of structure to reactivity, a central tenet of organic chemistry, provides a means to rationalize and predict chemical transformations. New reactivity, either serendipitous or hypothesized, tests, informs, complements, and improves this understanding; over time, a wide variety of transformations emerge. Within this enterprise, the synthesis of enantiomerically pure compounds with substoichiometric quantities of a catalyst is among the most significant advances in organic synthesis. As recognized by the 2001 Nobel Prize in chemistry and continuing with no surcease, advances in asymmetric catalysis are at the forefront of research in synthetic organic chemistry. The field has since continued to expand, with the number of publications containing the concept "asymmetric catalysis" increasing from 1646 from 2001 to over 2500 by 2004, remaining constant at that point ever since (Figure 1).

Despite this continuing effort, the general strategy for the development of chiral catalysts has arguably evolved at a much slower rate. Catalyst design remains primarily reliant on chemical intuition, wherein practitioners qualitatively identify relationships between catalyst enantiomeric products that may differ in only a few kcal/mol energy barrier (for example a 97.5:2.5 er corresponds to a difference of ∼2 kcal/mol at 298 K). This small energy difference presents a monumental challenge in the rational design of a catalyst; any of the myriad molecular effects including conformation, solvation, substrate interactions, steric and electronic considerations, and even temperature can alter the balance to affect selectivity. Although the intuition of a skilled experimentalist is still valued, even the most experienced practitioner is incapable of analyzing vast quantities of data and identifying the multidimensional correlations pertaining to catalyst efficacy. Inherently intuition-guided methods are qualitative; attempts to quantify relevant catalyst properties responsible for enantioinduction are typically made after the system is already optimized (if quantified at all) at which point the important aspects of catalyst structure have already been intuitively identified.

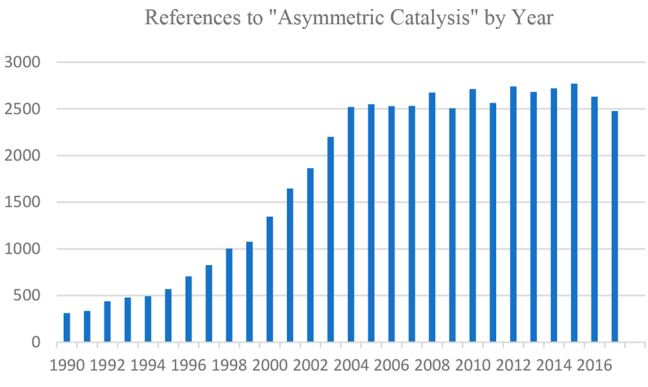


Figure 1. References obtained from a Scifinder search including the words or the concept "asymmetric catalysis" since 1990.

Quantitatively driven catalyst design is thus exceedingly rare, given the unquantifiable nature of intuition-guided methods. As discussed later in this review, quantitative methods relating selectivity to structural properties have been developed but have seen limited adoption. With the rise of "big-data" techniques, these pioneering efforts have laid the groundwork for the future of this field in the development of tools enabling expedited catalyst design. Further, more modern tools capable of analyzing large collections of data are paramount in discerning the relative importance of catalyst features.

In view of the spectacular improvements in processor speed and memory capacity, computationally guided methods for catalyst optimization have become an attractive alternative to empiricism. The last two decades have witnessed significant advances in computational methods for catalyst design, as reflected in the numerous reviews in the area.[1−22] The most common method is the application of accurate quantum mechanical calculations to provide mechanistic insight into reactions of interest that then guide experimentalists' efforts to modify the catalysts. This strategy, however, is limited in that the origin of selectivity must first be established for this method to be viable. A complementary strategy is the implementation of chemoinformatics to catalyst design. Although chemoinformatics does rely on mechanistic information, it can also be mechanism agnostic or be used to probe mechanisms of interest. Chemoinformatics-based protocols are also generally less computationally intensive. It is therefore possible to evaluate many catalyst candidates computationally before deciding which to synthesize, whereas analogous protocols using quantum chemistry and transition state analysis would be infeasible in many cases because of the greater computational resources required.

Chemoinformatics methods have long been used in the development of pharmaceutically relevant molecules.[23,24] In this application, certain structural features of molecules are correlated with biological activity using statistical methods to guide optimization. This concept, known as quantitative structure−activity relationships (QSAR), is a subfield of chemoinformatics that has been used extensively in biological systems (wherein "activity" refers to biological activity of a compound). This concept has also been employed with other applications, in which the reactivity (QSRR), a chemical property (QSPR), or, as in enantioselective catalysts, the selectivity (QSSR) of a molecular entity is probed. Although these terms are sometimes used interchangeably in the

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

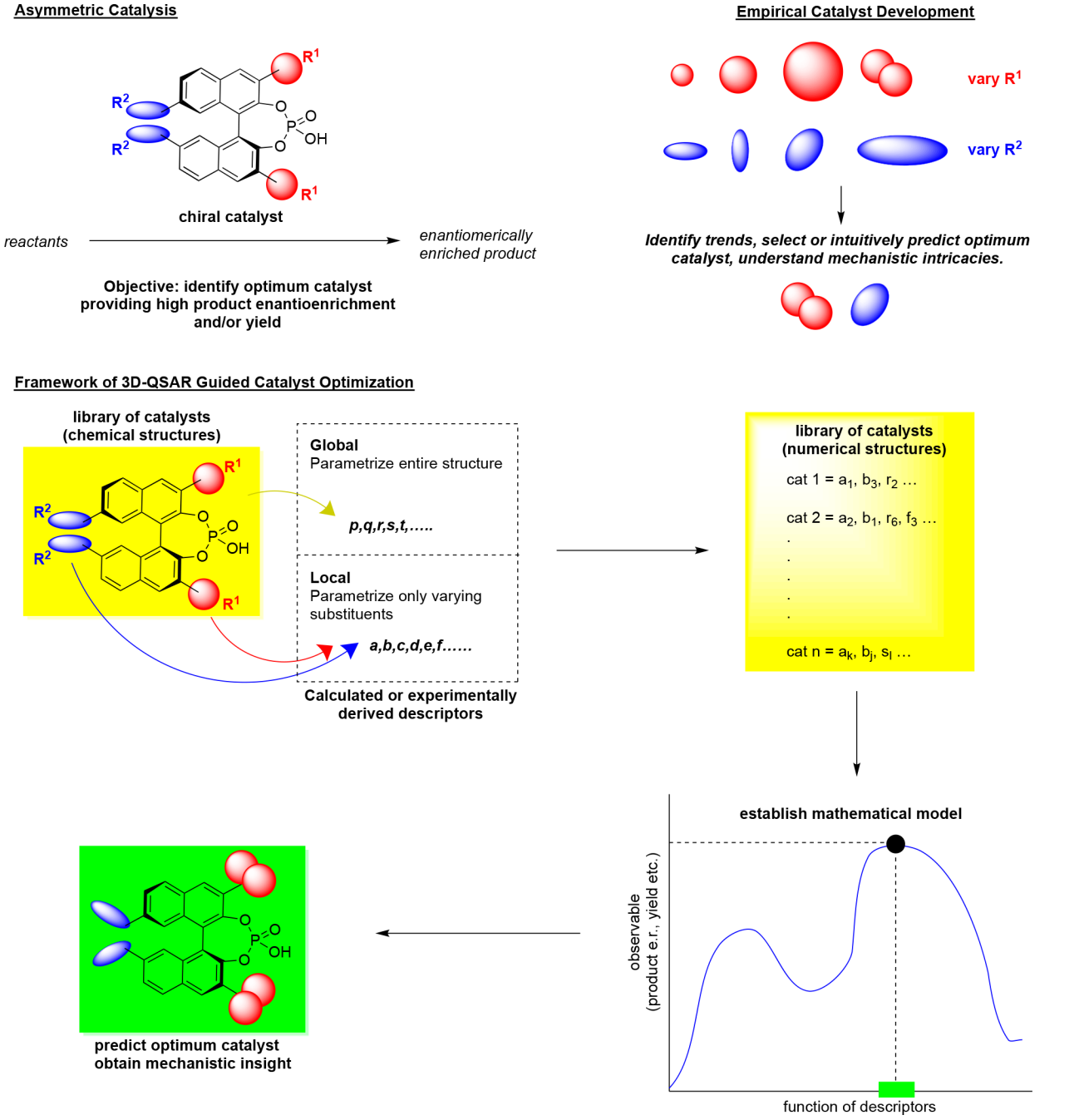


Figure 2. Overview of chemoinformatics guided catalyst optimization.

chemoinformatics literature, the subfield this review most pertains to is QSSR in enantioselective catalysis. Applications that involve the 3D structure of molecular entities benefit from a subfield of QSSR termed 3D-QSSR. In 3D-QSSR, three-dimensional descriptors are used to correlate structural features of the activity and selectivity of catalysts. It is then possible to predict the outcome of new, untested catalysts and to identify their important structural properties. In our opinion, 3D-QSSR methods have the greatest potential for capturing the subtle features of catalytic entities responsible for highly selective catalysis. Over the past two decades, a new paradigm has emerged to quantitatively supplement the "chemists' intuition". The fundamentals of this paradigm are outlined in the lower half of Figure 2. The search for an optimum catalyst begins by considering a library of potential candidates (exemplified here by a binaphthylphosphoric acid scaffold) with varying substituents (red, blue). It is assumed that there is an inherent, quantifiable correlation between catalyst structure and catalyst selectivity. The first step to uncover this correlation is the conversion of the catalyst "chemical structure" to a "numerical structure" by describing the molecule in terms of physical descriptors (for example, as Hammett parameters, Taft steric values etc.). Following this, mathematical models can be constructed to relate these descriptors to an observable (in this case, enantioselectivity). If a consistent correlation is found, a mapping of the descriptor space to selectivity is obtained, enabling the prediction of new descriptor values that correspond to catalyst structures providing improved perform-

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

ance. In this review, we will cover the various strategies by which this workflow can be implemented.

1.2. Scope and Organization

This review will concentrate on soluble, small-molecule catalysts. Experimentally measured parameters used as descriptors in QSSR will not be exhaustively discussed because the focus of this review is on theoretical molecular descriptors.[25] We have allowed some "spillover" into methods that are not rigorously considered QSSR, including most methods that correlate a calculable property derived from the 3-dimensional structure of a molecule to experimental enantioselectivity. We also offer our own opinions and perspectives on the various subfields and the direction of the QSSR-related methods and their pertinence to catalyst design. It is also noteworthy that many methods mentioned in this review were first developed in biological settings. Only small molecule catalysts are considered, with most cases of enzymatic transformations ignored. To learn more about his field, we direct the reader to other resources already available on the topic.[23,24,26]

The organization of this overview aims to introduce topics that incorporate increasingly complex methods for describing molecules. First, we begin with a brief introduction to the mathematical terms used throughout the review. Then, the review progresses from linear free energy relationships (LFER) with "local parametrization" of molecules to molecular interaction field (MIF) based methods, which we view as a "global parametrization" approach. The order of presentation has been designed not necessarily with preference to chronology, but rather with a view to assist a clear understanding of concepts, especially for nonspecialists. Indeed, we hope the review stimulates a wider adoption of chemoinformatics methods by the synthetic organic chemistry community to complement traditional discovery approaches in asymmetric catalysis.

1.3. Mathematical Background

Numerous modeling methods have been used in the studies detailed in this review. Although different experts might not agree on what constitutes a machine learning method, one could consider most methods used in this field as supervised learning methods. The goal of supervised learning is to relate independent variables to dependent variables (regressors and regressands in regression models, respectively). In the context of enantioselective catalysis, the descriptors are the regressors, and enantioselectivity is the regressand. A supervised learning model is trained by "mapping" the relationship between descriptors and selectivity in a subset of data called a training set. At the highest level of abstraction, the process by which the "mapping" occurs differentiates statistical learning techniques.

The simplest example of such a method is univariate linear regression ($y = mx + b$). In this equation, the independent variable is "mapped" to the dependent variable through a coefficient (m, the slope) and an intercept (b). An example of such a univariate relationship would be the relationship between Hammett parameters and enantioselectivity. If a Hammett plot is established, one might use this equation to predict the enantioselectivity of a new molecular entity that has not yet been explored with a corresponding Hammett parameter. Thus, there is one descriptor and it is being used to make predictions for enantioselectivity.

As the stereochemical outcome of a chemical reaction is dependent on numerous factors, many dependent variables of interest also are not dependent on only one factor. A number of statistical tools exist for dealing with such situations, many of which have been applied to chemical systems. Because detailed descriptions of these statistical learning methods are already given in these references, the discussion herein will be brief, and the interested reader is directed to these resources if they are interested in a more in-depth treatment of this subject matter.[27−32]

Multiple linear regression is the simplest approach to modeling problems with multiple independent variables. It has the same form as the univariate linear regression model, but each new descriptor is added into the model with its own coefficient ($y = ax + bz + c$). In multiple regression, the coefficients for each variable are optimized in a model training process by adjusting these coefficients to minimize a loss function. For example, in ordinary least-squares regression the loss function is the sum of squares of residuals. This model type assumes a linear relationship between the independent and dependent variables. Because this model type suffers from complications such as multicollinearity (independent variables that are correlated), other methods such as Ridge, Lasso, and ElasticNet have been developed. These models are similar in that they add a penalty function to the loss function of the ordinary least-squares regression. Lasso allows coefficients to shrink to zero, resulting in the elimination of correlated descriptors, whereas Ridge minimizes the variation in the predictions given by a model for a particular data point. ElasticNet is a hybrid of these two methods. These models still assume a linear relationship between regressand and regressors. A method by which nonlinearity is dealt with in practice is by the inclusion of interaction terms (the product of different descriptor values) or with polynomial terms (a descriptor value raised to the power of n).

Another modeling method capable of dealing with nonlinearity is the use of decision trees. Decision trees are conceptually easy to understand; they can be thought of as a flowchart of different "if/then" statements. These models can be used for both classification and regression applications but are prone to overfitting and have higher error rates owing to both bias and variance. Random forest models alleviate this problem by considering the aggregate of many decision trees, reducing overfitting and bias error.

Support vector machines (SVM) can also be used to construct regression models. Support vector regressors essentially seek to optimize a hyperplane that minimizes the error between the hyperplane and the training data (the hyperplane is conceptually similar to a best fit line on a 2D-plot). The model is optimized by generating a hyperplane that includes a maximum number of points within a certain error limit of the hyperplane. If nonlinear modeling is necessary, a kernel can be used that projects the data set in a higher dimensional space in which a hyperplane may be optimized.

Finally, neural networks are a modeling method designed to simulate the way the human brain learns and recognizes patterns. Neural networks are assembled of individual units called neurons. These neurons are assembled in layers, in which the first layer is termed the input layer and the last layer of the network is called an output layer. Any layers in between are called hidden layers, and more than one hidden layer qualifies the neural network as a deep neural network. Many network architectures have been developed for different applications and will not be covered in this review. In a simple, feed-forward neural network consisting of an input

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

layer, hidden layer(s), and an output layer, the input data are received by the input layer. Each neuron in the input layers is connected to each neuron in the next layer (i.e., the first hidden layer). An alternative way of phrasing this is that there is a unique connection between every pairwise combination of neurons from the input layer and the hidden layer. Each connection is associated with an activation function. The input value is passed through this activation function, and if it reaches a certain threshold value, the neuron fires, passing the new value to the next neuron. If the value is below the threshold value, the neuron does not fire. This process propagates through each layer until it reaches the output layer, which in turn gives the output of the regression analysis.

Often, in cases wherein many descriptors are available, it is advisible to reduce the dimensionality of the descriptor space to avoid overfitting. To achieve this goal, two general approaches are used—descriptor selection or dimensionality reduction. In the former, an algorithm is applied to select a subset of descriptors that give an acceptable model. In the latter, the descriptor space is transformed into a space with fewer dimensions while preserving the variance in the data. Some examples of descriptor selection mentioned in this review are forward selection, backward selection, stepwise selection, selection using Lasso or a linear support vector regressor (termed l1 selection), ranking feature importance in random forest models, and genetic algorithms. For interested readers, a more complete summary of these concepts is available elsewhere.[32−34]

Forward selection begins with a model that contains only an intercept which is the average of the regressands. Independent variables are added sequentially, each iteration adding the descriptor which improves the model the most. This process continues until a termination condition is met. Backward selection is the opposite, in which the model first contains all descriptors and removes the least informative descriptor iteratively. Stepwise selection is similar to forward selection, except that it also allows for variables to be removed. At each step, the significance of each variable is assessed. If a variable is identified as insignificant, it is removed from the model. A limitation of these methods is that they are dependent on sample size; too many descriptors with too few observations will likely identify a "good" model that is fit on the randomness in the data.

A more modern method for descriptor selection uses l1-regularization to eliminate descriptors. An example of l1-regularization is Lasso, in which the loss function has an added term that allows the coefficients of a given descriptor to shrink to zero, thus removing that descriptor from the model. Random forest models are capable of ranking descriptors in terms of importance and can also be used to select important descriptors.

Finally, genetic algorithms solve optimization problems in a method meant to mirror natural selection. Genetic algorithms can be used to solve a variety of optimization problems; in the case of selecting the best model given many descriptors, it identifies the optimal descriptor set to make the best model. The algorithm has a set of individuals (models in this case) that make up a population. Each individual is evaluated by some ranking metric (R^2, Q^2, etc.), wherein the best individuals are assigned a high fitness value. Individuals with high fitness values (better models) have a higher likelihood of being selected as parents for the next generation. In this step, two parents (models) are crossed to make two individuals, which potentially results in the creation of a superior individual. In this way, individuals with good traits (the best descriptors) are kept and individuals with bad traits (inferior descriptors) are eliminated, eventually converging on an acceptable solution.

Dimensionality reduction is an alternative to descriptor selection methods. A prominent example of this type of transformation is principal component analysis (PCA). In PCA, the descriptor space is transformed to a new set of uncorrelated variables in a manner which maximizes the variance per principal component. This operation mandates that the first principal component contains the most variance in the data, followed by the second, until the number of principal components is equal to the input dimensionality. The first n principal components can then be selected and used for various application such as modeling or data visualization. It is noteworthy that PCA considers only the descriptors (it is unsupervised). Projection to latent structure (PLS) can be thought of as the supervised analog of PCA. PLS constructs latent variables that are linear combinations of the original descriptors but are also related to the regressand. A more in-depth discussion of these methods is available.[35]

To generalize or interpret models, it is imperative that they are validated. Two general types of validation are used: internal and external. An example of internal validation is k-fold cross-validation. In k-fold cross-validation, the input data are divided into k equally sized folds (i.e., if there are 100 input data points, 10-fold cross-validation will divide the data randomly into 10 groups of 10). Then $k - 1$ of the folds are used as the training set, and the remaining fold is used as the test set. This process is repeated until each fold is used as the validation set one time. The average of the model during cross-validation is summarized by Q^2, which is the cross-validated R^2 (i.e., the R^2 of the predicted values). Leave-one-out cross-validation (LOO) is a specific kind of cross-validation in which k is equal to the number of samples and each sample is held out once. This method is typically considered inferior to k-fold cross-validation because the training sets in each iteration are very similar (only one point at a time is held out). Thus, the models made are strongly correlated and noise is not averaged away as it would be in, for example, 10-fold cross-validation. Thus, LOO can give overoptimistic Q^2 values. Typically, internal validation is not sufficient to fully validate a model. This assertion is especially true if cross-validation is used to tune model parameters. In these situations, it is necessary to use external validation in addition to internal validation. External validation is the use of an external test set, in which a partition of data points is withheld from the model training and cross-validation process. Once the final model is obtained after cross-validation, the test set is used only one time to evaluate the model. Note that changing model parameters after evaluating a test set to improve accuracy should be avoided. Making iterative changes to a model to increase the accuracy of an external set can cause overfitting and is not strictly an example of external validation.[36]

2. LINEAR FREE ENERGY RELATIONSHIPS (LFERS) WITH 3D-DESCRIPTORS

2.1. Introduction to LFERs

As a textbook concept in physical organic chemistry, linear free energy relationships (LFERs) represent a classical method to uncover a correlation between a substituent characteristic and reactivity. Briefly, for a reaction class, the variable substituent is

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

described (parametrized) most commonly in electronic or steric terms and is correlated directly or indirectly with reaction rates. The descriptor may be experimentally or computationally determined. For example, in a classical Hammett correlation, the σ descriptor derives from the relative acidity of substituted and unsubstituted benzoic and phenylacetic acids. The key here is that the substituent (a structural element) is effectively transformed to a "number" (the σ descriptor). The extension of this concept to asymmetric catalysis involves correlating descriptors of catalyst substituents to the enantioselectivity conferred by those catalysts in a synthetic transformation with the aim to uncover mechanistic information as well as predict the performance of untested catalysts. In general, only specific subunits of the catalysts, the varying substituent, are provided a descriptor. In that sense, such an approach involves only a "local parametrization" of the catalyst structure.

Here, we concentrate on models employing calculable, 3D descriptors. For example, Hammett parameters[37,38] and Taft steric parameters[39−41] are derived from experimental values and will not be a focus of discussion, but Charton values[42−54] will be considered (however, whether they are truly 3D descriptors is open to debate). The application of linear free energy relationships to asymmetric catalysis with calculable, tailored parameters represents the largest class of applications in which statistical methods are employed. Because multiple publications already exist detailing these endeavors, discussion of this section will be less comprehensive.[55−59] Finally, many studies included are not typically thought of as QSAR. For example, most chemists would not consider a univariate relationship between selectivity and a molecular property (e.g., selectivity vs Hammett parameters) as QSAR. However, in the broadest sense, a molecular descriptor is quantitatively correlated to catalyst activity; thus, one could argue that this type of correlation is a QSAR. To give context to the field, we elected to include some studies wherein 3D descriptors are used. First, examples of univariate LFER in asymmetric catalysis will be briefly discussed followed by a summary of multivariate methods.

2.2. Selected Univariate Free Energy Relationships in Asymmetric Catalysis

Multiple descriptors, for example, atomic charges, van der Waals radii, polarizability, cone angle, etc. can be assigned to a substituent. Subsequent LFERs may be constructed by using either a single descriptor or an algebraic combination of multiple descriptors. These two cases result in a univariate or multivariate relationship, respectively. Obtaining every example describing a linear relationship between a calculated property and enantioselectivity is a daunting task. Thus, the aim of this section is not to provide a comprehensive collection of every example of univariate LFERs but rather to select representative illustrations.

Steric effects imposed by catalyst substituents are often critical in enforcing high enantioselectivities. Such effects are easy to intuitively recognize and predict, especially once a reasonable transition state model is hypothesized. Likewise, a campaign for catalyst library synthesis commonly has the objective of covering substituents that differ widely in their steric contribution. LFERs that attempt to explicitly relate catalyst steric properties with enantioselectivity naturally become good starting points to test the application of QSSR in asymmetric catalysis. Although a variety of steric descriptors can be used, Charton values and Sterimol parameters have emerged as the most popular (Figure 3).[60,61] Charton values are derived from Taft steric parameters, which in turn are derived from the relative influence certain groups have on the rate of hydrolysis of methyl esters. Charton was able to fit these measurements to calculated values derived from the van der Waals radii of the substituent. Sterimol parameters, developed by Verloop,[62] are calculated from the size of substituents with respect to a primary axis. The B1 parameter is the smallest radius accessible to a group of interest as it rotates around a central axis, the B5 parameter is the widest radius possible to that group, and the L parameter is the length from the attachment point of the group to the distance away, linearly.

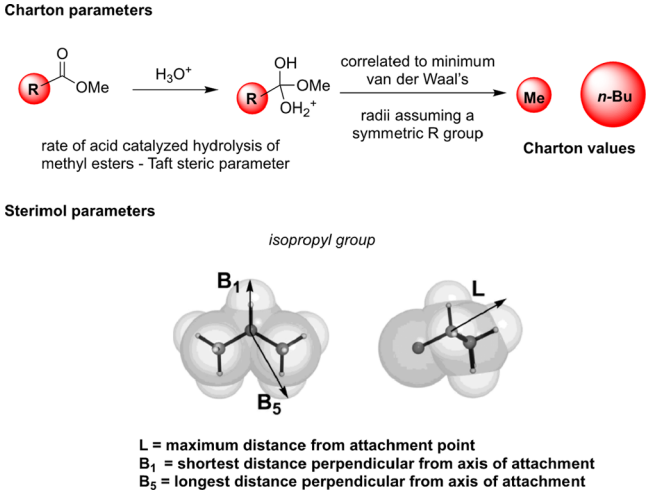


Figure 3. Charton and Sterimol steric parameters. Reproduced from Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Sterimol Steric Parameters. *ACS Catal.* **2019**, 9, 2313−2323. Copyright 2019 American Chemical Society.

2.2.1. Univariate LFERs Based on Charton Values.

Miller and Sigman reported the use of Charton values to construct LFERs in enantioselective Nozaki-Hiyama-Kishi (NHK) allylation reactions.[63] In view of the sensitivity of the proline carbamate moiety in the catalyst scaffold to conferred enantioselectivity, five catalysts with differing carbamate groups (G) were evaluated (Figure 4). In the reaction of three different substrates, a linear relationship is obtained when substituent Charton values are plotted against the logarithm of product enantiomeric ratios. This trend is consistent with the empirical observation that increasing size of the G group results in enhanced selectivity.

In the same report, the authors explored the generality of this correlation approach. Three reactions from the literature were selected to study the relationship between the Charton value of a key substituent and enantioselectivity: (1) palladium-catalyzed, enantioselective, allylic alkylation reactions,[64] (2) enantioselective cyclopropanation of allylic alcohols,[65] and (3) the Mn-salen catalyzed, enantioselective aziridination of styrenes (Figure 5).[66] In the first two examples, substituents on the catalysts are varied. In the third, substituents on the β-position of the styrene substrate are varied. All three examples produce linear relationships between the free energy differential of diastereomeric transition structures and the Charton value. Note that in the third example, the descriptor is for a variable substituent on the

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

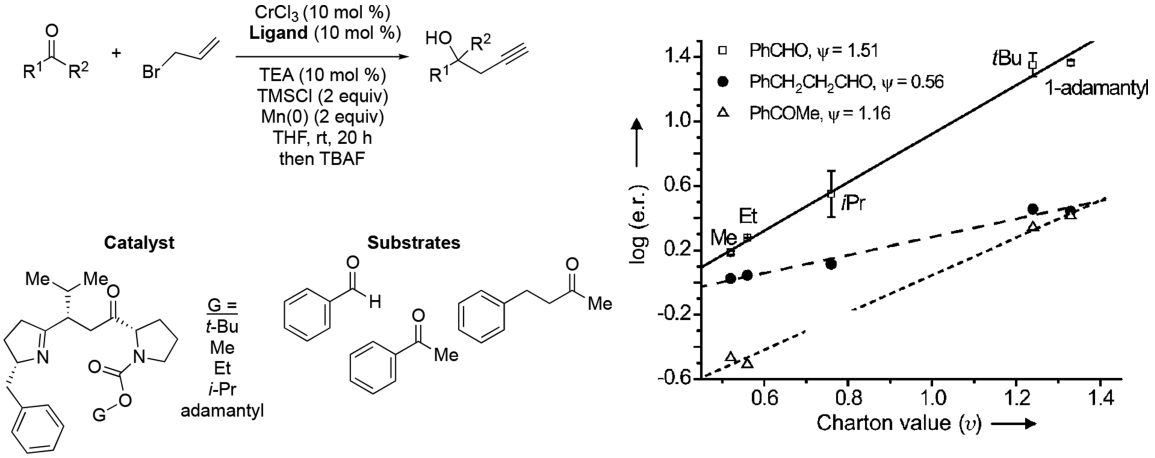


Figure 4. Catalysts and substrates used in enantioselective NHK reaction and a univariate correlation of enantioselectivity with substituent Charton values. Adapted with permission from ref 63. Copyright 2008 John Wiley and Sons.

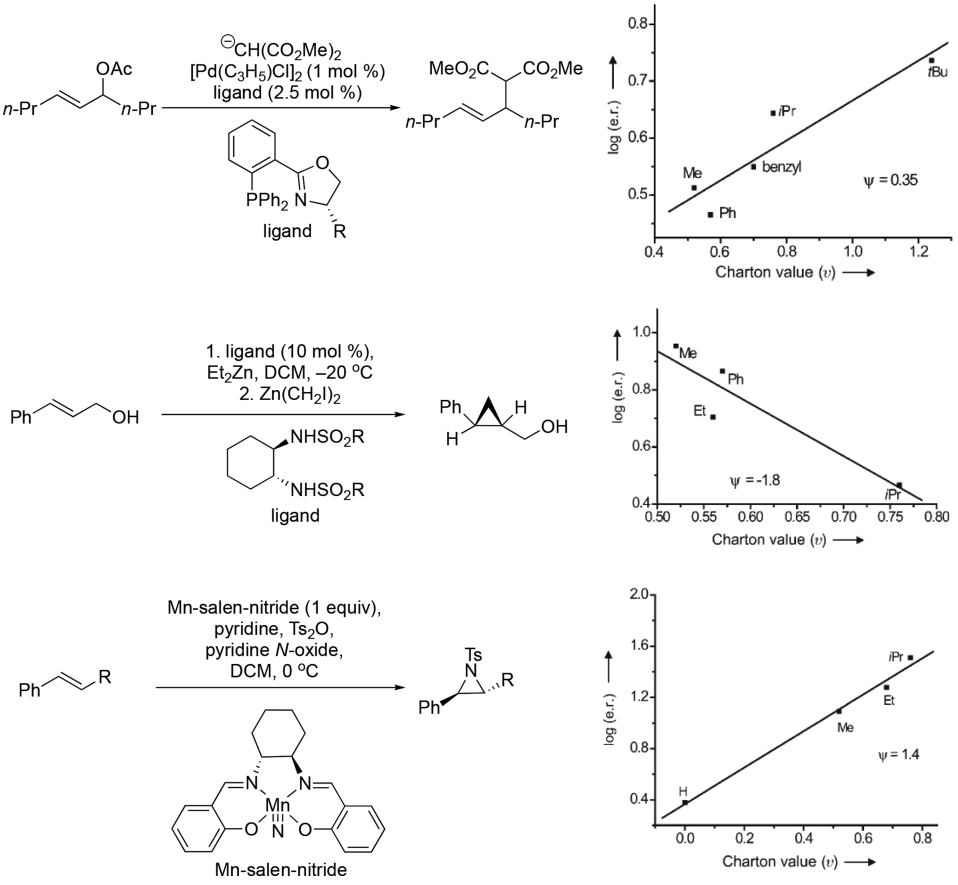


Figure 5. Top: LFER for enantioselective, palladium catalyzed alkylation of allyl acetates. Center: LFER for enantioselective cyclopropanation of allylic alcohols. Bottom: LFER for enantioselective aziridination of styrene. Adapted with permission from ref 63. Copyright 2008 John Wiley and Sons.

styrenyl substrate rather than the catalyst; in principle, a LFER can be investigated by considering any consistently varying substituent on any component in the reaction system.

In the plots in Figure 5, the slope, designated as ψ, provides information about the relative influence of the varied substituent and whether selectivity increases or decreases with substituent size. A positive slope indicates that large groups are associated with increased selectivity, whereas a negative slope indicates that smaller groups are associated with higher selectivity. Thus, this study demonstrated that in these cases, Charton values are satisfactory descriptors to recapitulate the empirically observed selectivity trends, with the ψ value as a quantitative indicator of sensitivity.

This method of analysis has been applied to other reactions. The studies from Pfaltz and co-workers[64] employing phosphine oxazoline ligands are contrasted with studies from Park and co-workers,[67] employing oxazolinylferrocene ligands in an enantioselective, palladium catalyzed, allylic alkylation

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

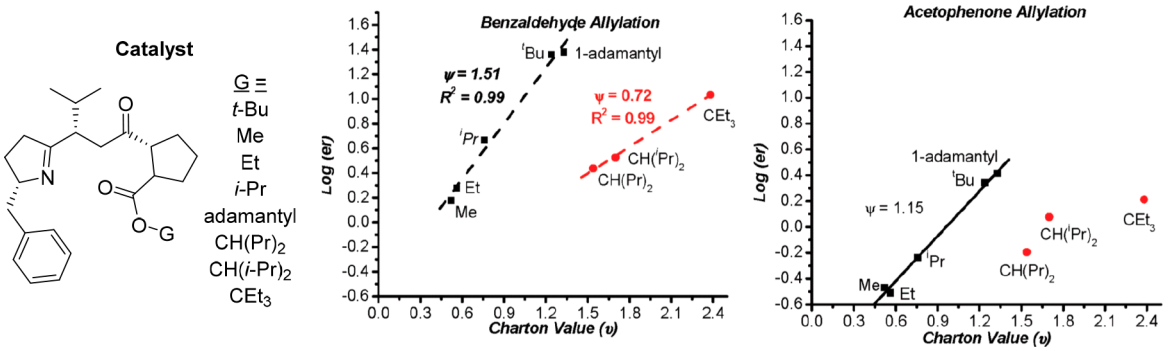


Figure 6. Catalysts with accompanying LFER plots for benzaldehyde and acetophenone allylation. Reproduced from Sigman, M. S.; Miller, J. J. Examination of the Role of Taft-Type Steric Parameters in Asymmetric Catalysis. *J. Org. Chem.* **2009**, 74, 7633−7643. Copyright 2009 American Chemical Society.

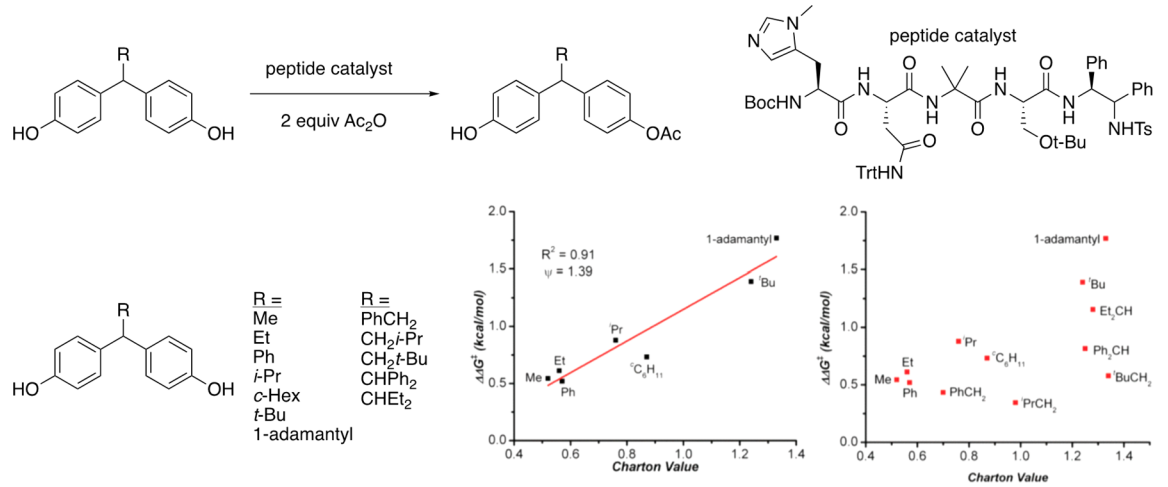


Figure 7. LFERs for desymmetrization of bisphenol substrates with a peptide catalyst (top). Reproduced from Sigman, M. S.; Miller, J. J. Examination of the Role of Taft-Type Steric Parameters in Asymmetric Catalysis. *J. Org. Chem.* **2009**, 74, 7633−7643. Copyright 2009 American Chemical Society.

reaction (in both cases the substituent varied was on the 4-position of the oxazoline). The oxazolinylferrocene ligands (ψ = 0.75) are more sensitive to steric effects than the phosphine oxazoline ligands (ψ = 0.35). The enantioselective alkylation of aryl aldehydes with chiral Ti-TADDOL complexes by Seebach has also been analyzed. A substantial LFER is found for the substituents geminal to the Ti-ligating oxygens (ψ = 1.85).[68] The enantioselective, vanadium-catalyzed epoxidation of allylic alcohols reported by Wu and Wang is demonstrated to have a negative LFER (ψ = −0.30), with smaller groups associated with more selective reactions.[69] Finally, Quintard and Alexakis investigated substrate steric effects in the enantioselective, organocatalytic addition of aldehydes to *cis*-1,2-bis-(phenylsulfonyl)ethene, which subsequently undergoes a 1,2-sulfone rearrangement to give geminal sulfones on the γ-carbon with respect to the aldehyde.[70] The Charton value of the substituent at the α-carbon of the aldehyde substrate is found to correlate with both selectivity and yield, with ψ values of 0.45 and 3.54, respectively.

2.2.2. Limitations of Charton Values. It is important to note that Charton values approximate substituents as spherical. This assumption may be reasonable for symmetric substituents such as H or Me but is clearly incorrect for anisotropic substituents like *n*-Bu or Ph. The following cases highlight situations in which Charton values prove inadequate to obtain a reliable and consistent correlation.

The enantioselective hydrogenation of α-(acylamino)acrylic derivatives with *P*-stereogenic, bidentate, C_2-symmetric phosphine ligands is observed to have a negative correlation with the Charton parameter of a substituent on phosphorus (ψ = 0.73) (the other substituents are a methyl and an ethylene linker connecting the phosphorus atoms). However, the authors also highlight important considerations while executing this protocol. For example, phenyl substituents have two Charton parameters (in-plane of 0.57 and out-of-plane of 1.66) that must be judiciously selected in addition to having other electronic influences that have not been accounted for. Furthermore, groups such as *n*-Pr, *n*-Bu, and *i*-Bu all have the same Charton parameters, but the behavior of these residues in all settings is not identical. Finally, results from LFERs with limited data sets must be analyzed with caution; it is recommended that either substituents with a broad range of Charton values are used or many subunits in a narrow range of Charton values are used to ensure the validity of the LFER.

Sigman and Miller revisited the enantioselective NHK-reaction as well as other literature examples to challenge the LFER protocol.[71] In the NHK-reaction, additional catalysts were synthesized to test if the linear relationship between Charton values and enantioselectivity persisted with larger Charton values leading to a more selective catalyst (Figure 6). Combining these new catalysts with those from the first study reveals a break in the plot interpreted to arise from a change in

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

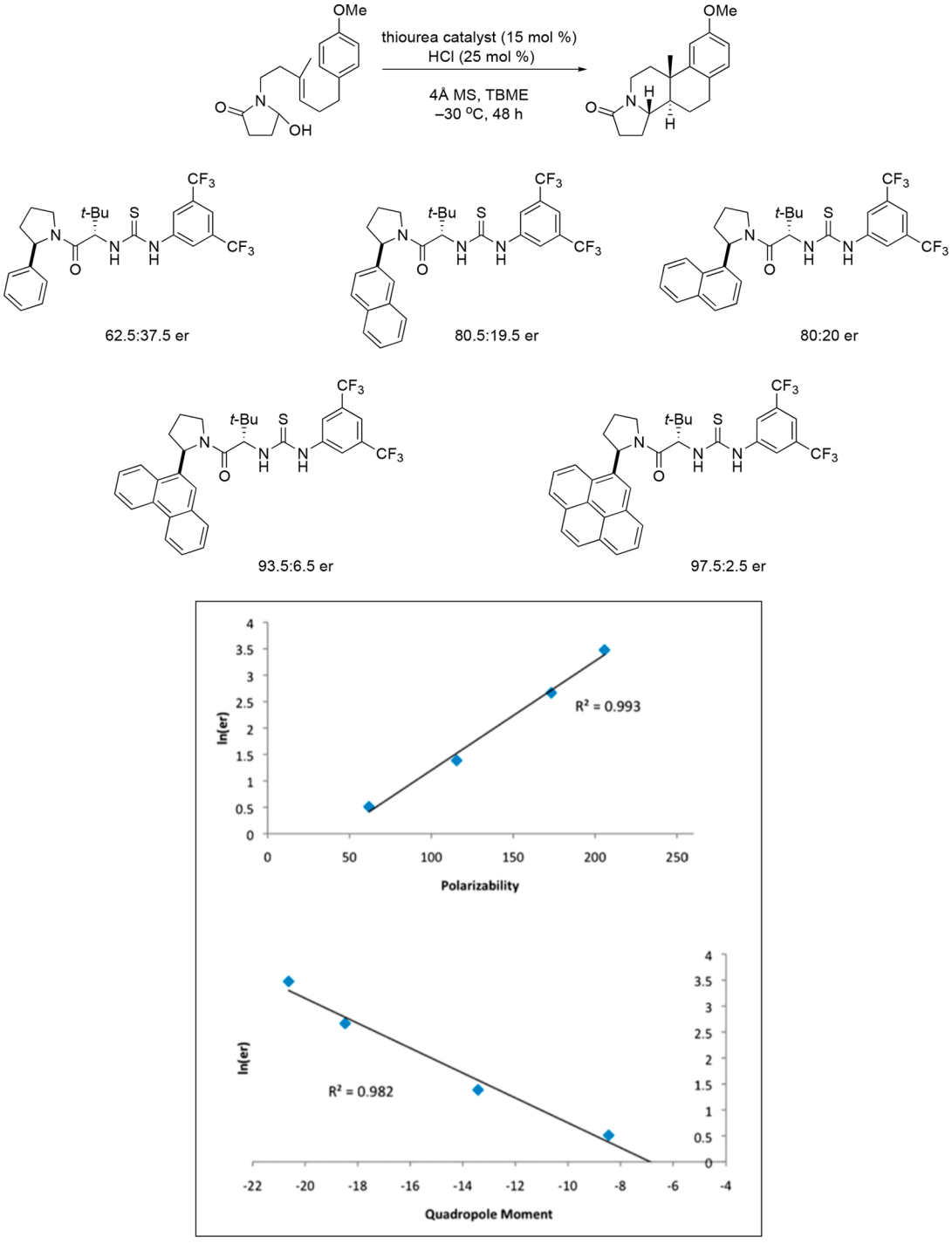


Figure 8. LFERs between enantioselectivity and polarizability (top) and quadrupole moment (bottom) of the aromatic substituent on thiourea catalysts. Adapted with permission from ref 77. Copyright 2010 National Academy of Sciences.

the conformation of the catalyst (analogous to how breaks in Hammett plots can indicate a change in mechanism). However, in later publications, the authors attribute this break to a limitation of Charton values resulting from their derivation from rates associated with a specific transformation wherein the substituents can be approximated as spherical.[60]

An example of Charton values underperforming with respect to other steric parameters is demonstrated by Gustafson, Sigman, and Miller.[72] In this work, the remarkable influence of a remote substituent in the desymmetrization of 4,4′-methylenediphenol derivatives with a peptide catalyst is analyzed with various LFERs using different steric parameters.[73] When substituents bearing steric bulk close to the benzylic carbon are present, a good correlation between the free energy differential and Charton parameter is observed (ψ = 1.39). It is postulated that steric bulk at the benzylic position orients the phenol rings in a rigid, propeller-like conformation, enabling selective enantiodifferentiation. When substituents presenting steric bulk farther away from the benzylic position are employed, the resulting correlation is poor, likely because the Charton values do not capture the perturbation of the substituent on the rings (Figure 7). Other steric parameters

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620–1689

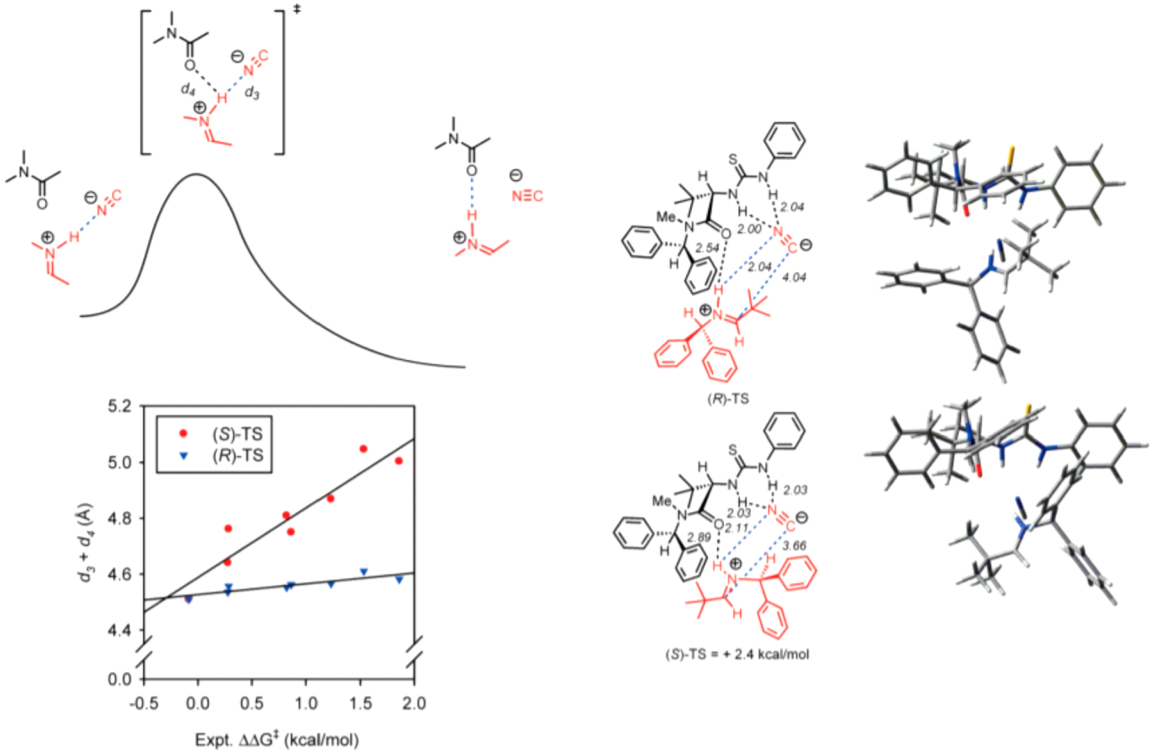


Figure 9. Interatomic distances (top left) in transition structures leading to *R* and *S* stereoisomers (right) and their correlation with enantioselectivity (bottom left). Reproduced from Zuend, S. J.; Jacobsen, E. N. Mechanism of Amido-Thiourea Catalyzed Enantioselective Imine Hydrocyanation: Transition State Stabilization via Multiple Non-Covalent Interactions. *J. Am. Chem. Soc.* **2009**, 131, 15358−15374. Copyright 2009 American Chemical Society.

such as A-values[74] and interference values[75] give stronger correlations than Charton values. However, a full analysis of all different substituents could not be carried out, as both values are experimentally measured and not available for all substituents used in this study. This exercise demonstrates that descriptor selection in LFERs can be done critically with consideration of how the values are derived.

2.2.3. Beyond Charton Values: Electronic Descriptors and Sterimol Parameters. In addition to steric factors, a catalyst substituent may influence transition state geometry and energy through charge stabilization, inductive effects, dipole minimizations or other effects. In such cases, the use of appropriate electronic descriptors is desirable to capture an electronic contribution important for catalyst selectivity.

Jacobsen and co-workers employed LFERs to identify important noncovalent interactions in thiourea-catalyzed, polyene cyclization reactions.[76,77] The authors find that incremental increases in the size of the arene unit proximal to the amide residue of the catalyst correlates with increasing enantioselectivity (Figure 8). To better understand the origin of enantioselectivity in this system, the selectivity of each catalyst is plotted against the polarizability and the quadrupole moment of the corresponding arene subunit. A strong linear relationship is found, supporting the authors hypothesis that these larger arene surfaces afford more stabilization to the transition structure leading to the major stereoisomer through a cation−π interaction, resulting in high stereoselectivity (Figure 8).

Zuend and Jacobsen provide an elegant demonstration of how a carefully considered LFER can reveal mechanistic information in the thiourea-catalyzed, enantioselective Strecker reaction.[78] Computational studies suggested that the rate-determining step of the transformation is a rearrangement of the iminium and cyanide ions stabilized by the thiourea catalyst. For eight different catalysts, cumulative H-acceptor interatomic distances for the hydrogen bond network in the disfavored transition structure (leading to the *S*-stereoisomer) correlate well with the experimental enantioselectivity (expressed as free energy) (Figure 9). This correlation suggests that in selective catalysts, the intermediate iminium ion leading to the *S*-stereoisomer is destabilized relative to the *R*-stereoisomer, and this destabilization is reflected in the respective transition structures leading to each enantiomer.

A relatively early example of univariate LFERs in enantioselective catalysis using 3D descriptors is the use of calculated atomic charges and the Sterimol B1 parameter[62] to identify important structural features in the ruthenium-catalyzed, enantioselective, transfer hydrogenation of aromatic ketones.[79] To deconvolute electronic and steric contributions, three LFERs were reported, two of which used the sum of atomic charges on the substrate arene as descriptors. The third LFER considered Sterimol parameters as steric descriptors. The authors concluded that solvation and dispersion effects are the most influential in determining selectivity, followed by electrostatic effects, with only a minor contribution from steric effects.

As mentioned above, the research summarized in this section is not an exhaustive summary of univariate LFERs using calculated parameters in enantioselective catalysis. However, this section has highlighted the value of these methods in ascertaining valuable mechanistic information about catalyst structure without the need for rigorous quantum chemical calculations and has also provided the context

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

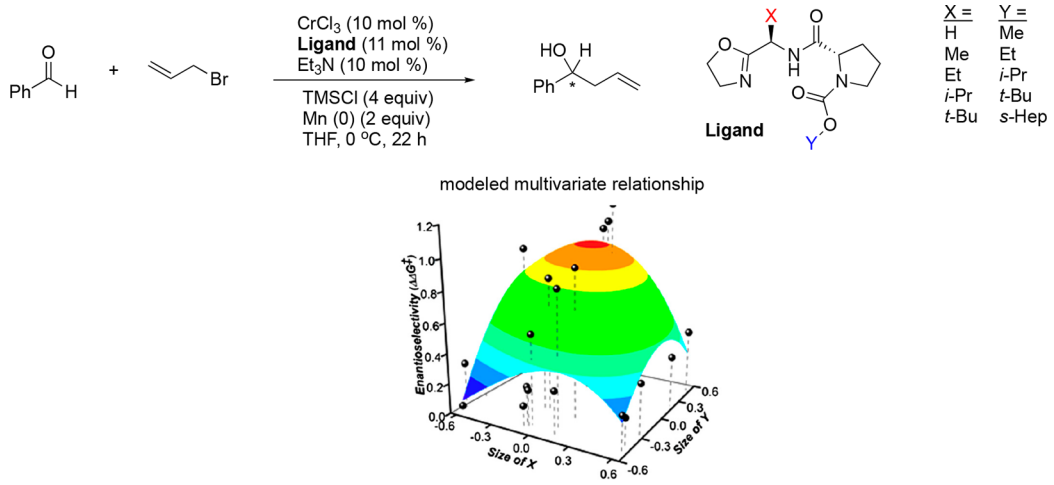


Figure 10. Pairwise combinations of X and Y substituents on the common core give 25 unique catalysts (top right). Experimentally determined enantioselectivities are used to construct a multivariate relationship between catalyst descriptors and selectivity (bottom). Predictive models are constructed by using a 9-member training set evenly covering the descriptor space. Adapted with permission from ref 81. Copyright 2011 National Academy of Sciences.

necessary for a discussion of multivariate LFERs in the following section.

2.3. Multivariate Linear Free Energy Relationships

Univariate LFERs can capture the correlation of a single structural parameter with enantioselectivity and are relatively easy to execute. However, such a reductionist approach has its limitations because multiple electronic and steric effects are often important in asymmetric catalysis, and ideally, a QSSR should take into account all such contributions. Furthermore, because selective catalysis is generally affected by an interplay of these factors, individual LFER studies with single descriptors may be ineffective or misleading. Thus, multivariate LFERs can be considered as the logical next step in QSSR studies. In a multivariate approach, a function constructed from an algebraic combination of more than one descriptor is correlated with enantioselectivity. In this way, multiple, interdependent contributory effects can be identified and the relative contributions of effects represented by these descriptors may be estimated. The multivariate LFER approach offers the potential to delineate nonobvious contributions, perhaps beyond the chemists' empirical intuition. However, caution must be exercised on the choice and number of descriptors employed; the risk of overfitting the correlation increases with increasing number of descriptors. On the contrary, inadvertently omitting the causative variable can result in erroneous interpretations of models. For example, omitting a causative variable but including a variable correlated with the omitted variable will incorrectly assign the significance to the correlated variable. In this case, accurate predictions could be made within the domain of the model but the correlated variable might be misinterpreted as causative. Further, the omission of important variables will not be readily apparent during the model development phase. Clearly this phenomenon could be detrimental to the interpretability of a model.

The first example of the application of multivariate regression analysis applied to the prediction of the free energy differential between competing, diastereomeric transition structures in enantioselective catalysis was reported by Norrby and co-workers.[80] In this work, the isomeric ratio of various nucleophilic substitution reactions on palladium η^3-allyl complexes was predicted. Descriptors used in the model included various structural features of ligated, η^3-allyl palladium complexes such as bond angles and dihedral angles. A model was generated that was used to predict the isomeric mixture of reactions resulting in different constitutional (branched vs linear) or enantiomeric ratios. The model was assessed with different internal validation scores, with $Q^2 > 0.85$ in all cases, constituting the first example of 3D-QSAR in enantioselective catalysis.

Arguably the most influential work to date in the widespread adoption of statistical methods for enantioselective catalyst development has been the application of multivariate regression techniques by Sigman and co-workers.[58] Throughout this body of work, multivariate linear regression is used to construct relationships between experimental outcome and the descriptors. In their seminal report, Harper and Sigman used multivariate regression techniques to further analyze selectivity in the NHK allylation of benzaldehyde (introduced in the previous section).[81] Charton values at two variable positions (a bivariate regression) of proline-oxazoline catalysts are correlated to enantioselectivity. In this study, 25 different catalysts are tested and a model constructed as a proof of concept, wherein the best catalysts are predicted as such. However, rigorous validation is not reported. Graphically, a multivariate regression can be imagined to represent a surface whose dimensionality depends on the number of parameters utilized. In this case, the correlation is visualized as a 3D surface (Figure 10).

After collecting preliminary results, the following design considerations are used during predictive model generation: (1) a training set of data covering the range of possible descriptor values can be used to generate models in which future predictions are interpolative, increasing the confidence in those predictions, and (2) a uniform response variable (e.g., enantioselectivity data) distribution in the training set tends to give stronger models than highly skewed data sets. With this in mind, a nine-member subset (the training set) from the full 25 member set is selected for identifying a correlation. A critical test for the validity of the derived model is to check if accurate selectivity predictions are obtained for ligands not included in the nine-member training set. From this exercise, the model does provide an accurate prediction for the best ligands (X =

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

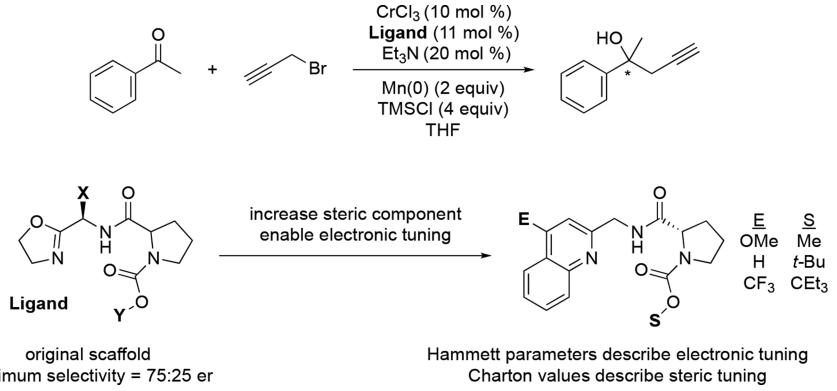


Figure 11. Scaffold comparison for the enantioselective propargylation of ketones.

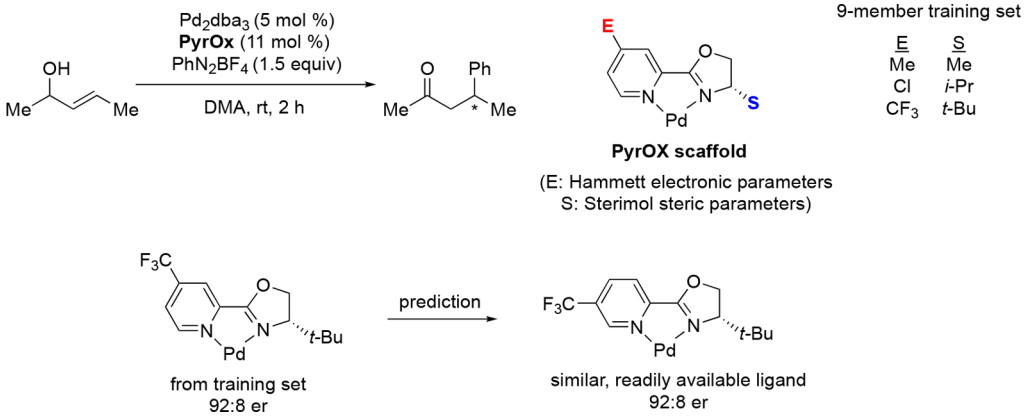


Figure 12. Palladium catalyzed enantioselective Heck arylation.

Et and *i*-Pr, Y = *t*-Bu). Predicted versus observed selectivities are 84:16 versus 90.5:0.5 and 83:17 versus 92:8, respectively, although predicted values for the other ligands are not provided.

This protocol is repeated for the allylation of acetophenone and ethyl methyl ketone; the model for the former predicts the best catalyst even though it is not included in the training set (X = *i*-Pr, Y = *t*-Bu, predicted 92:8, observed 95.5:4.5). The selectivity surface for the NHK-reaction of ethyl methyl ketone indicates that no selective catalyst derived from the proline-oxazoline scaffold exists (i.e., the selectivity surface has no high selectivity maxima), thus prompting the authors to abandon optimization of that scaffold for that substrate. Although this work does not lead to the design of a better catalyst, it serves as a proof-of-concept for the simultaneous analysis of multiple variables in linear free energy relationships.

The design considerations laid out in the first publication were later used for the optimization of enantioselective propargylation of ketones.[82] The original set of nine ligands from the previous work is used to collect experimental data and construct a mathematical model relating Charton values to enantioselectivity. However, a flat selectivity surface led the authors to abandon this scaffold (Figure 11). This negative result is a critical design element; the authors reported that obtaining experimental data for the nine training ligands and construction of a model took one week. Thus, the decision to explore new ligand scaffolds rather than continuing to try different permutations of the same scaffold is made rapidly, accelerating the rate of discovery. Guided by empirical modifications, the authors selected a quinoline-proline based architecture as the new scaffold (QuinPro), wherein electronic properties can be modified by the substituent on the quinoline ring and steric modifications can be made on the basis of the identity of the carbamate group at the proline residue. Nine training catalysts are selected by picking substituents such that the 2D chemical space constructed from Hammett parameters and Charton values is spanned evenly. Of these nine catalysts, one (E = OMe, S = *t*-Bu) that displays highly selective propargylation of aromatic, vinyl, and aliphatic ketones (17 examples, all over 85:15 e.), is chosen. A model is constructed which suggests that the training set catalyst is the most selective catalyst in the space; however, validation data are not provided. Although the model itself does not lead to identification of a selective catalyst, the experimental design of catalyst selection spanning the breadth of a meaningful chemical space facilitates optimization of a previously underperforming reaction.

This method has also been employed in the optimization of an enantioselective Heck arylation catalyzed by palladium pyridineoxazoline (PyrOX) complexes.[83] The reaction involves the coupling of aryldiazonium salts with unsaturated alcohols wherein the alkene moiety migrates to the distal hydroxyl group through a chain-walking mechanism. For modeling selectivities, a nine-member catalyst set is selected on the basis of steric parameters of the substituent at the 4-position of the oxazoline ring and the electronic nature (using Hammett parameters) of the substituent on the pyridine ring. In analogy to the previous study, a selective catalyst is identified in the training set. However, a comparable catalyst predicted by the

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

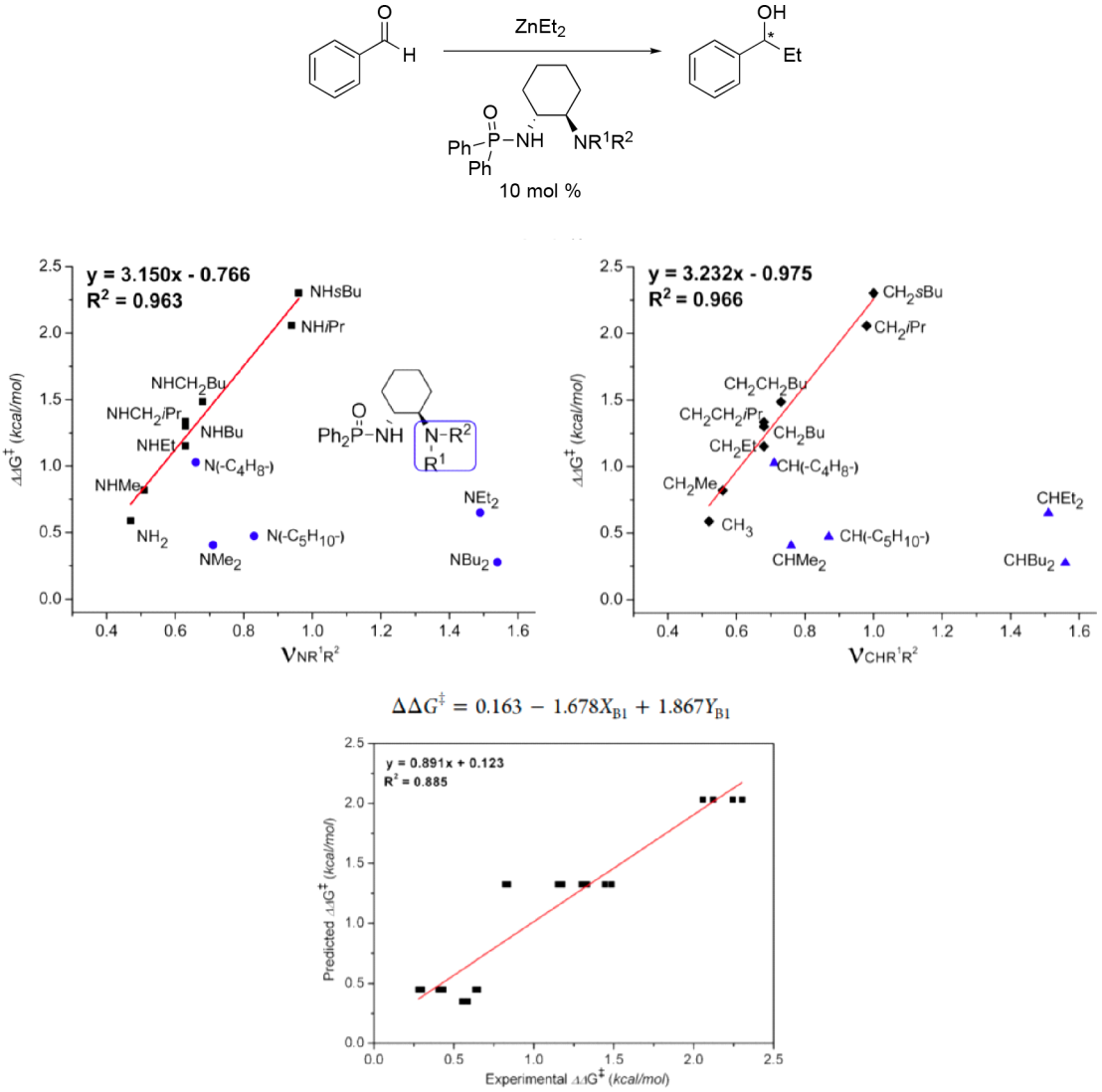


Figure 13. Enantioselective alkylation of aryl aldehydes (top). LFERs with Charton parameters (center), sterimol parameters (bottom). Reproduced from Huang, H.; Zong, H.; Bian, G.; Song, L. Constructing a Quantitative Correlation between N-Substituent Sizes of Chiral Ligands and Enantioselectivities in Asymmetric Addition Reactions of Diethylzinc with Benzaldehyde. *J. Org. Chem.* **2012,** 77, 10427−10434. Copyright 2012 American Chemical Society.

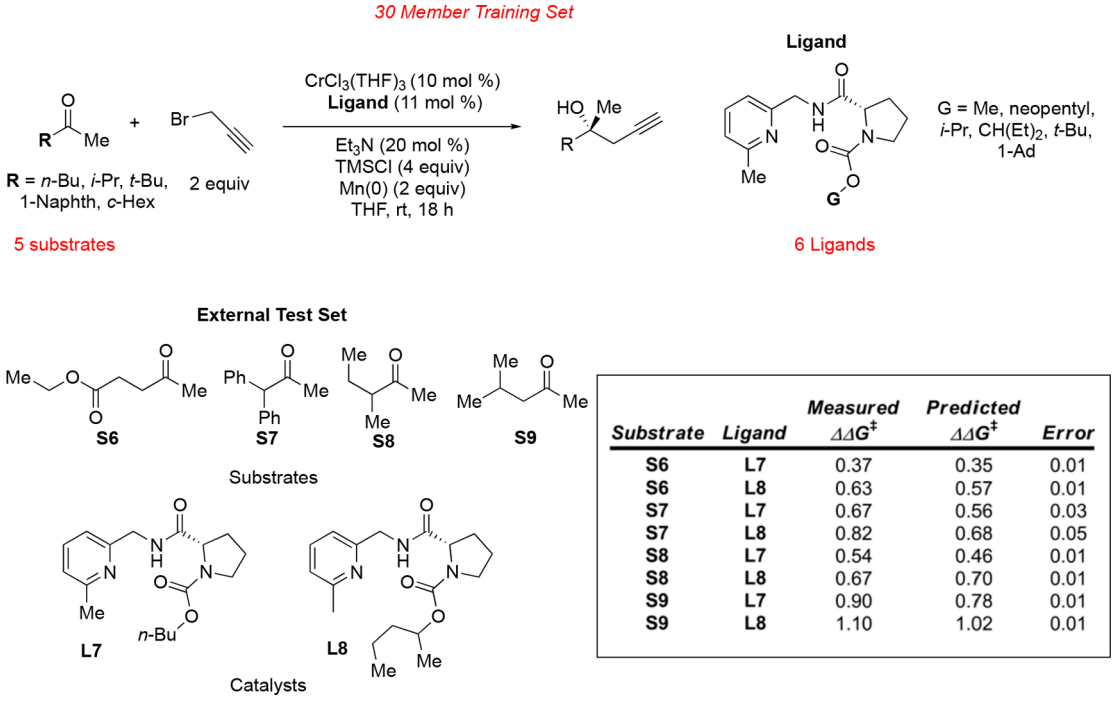
model is used because it can be prepared from a more readily available starting material (Figure 12).

As alluded to in the univariate LFER section, observation of breaks in Charton plots led to a more rigorous exploration of different steric parameters that are more broadly applicable.[60] The authors thus turned to Sterimol parameters to reexamine previous systems wherein Charton values give anomalous results. By re-examining the enantioselective NHK reaction of benzaldehyde and acetophenone, improved models are constructed wherein the Sterimol B1 parameter correlates well with the free energy difference between the competing, diastereomeric, transition structures. Although the previous study using Charton values reveals a break in the Charton plot, no break is observed when Sterimol parameters are used for both the benzaldehyde and acetophenone examples, changing the interpretation of the analysis.

Revisiting the desymmetrization of 4,4′-methylenediphenol derivatives gives similar results, wherein Sterimol B1 and L parameters are used to generate an improved model with respect to the original method. These parameters also allow for a more straightforward interpretation of the model; the importance of the B1 parameter indicates that the substituent is not freely rotatable, consistent with the hypothesis of the group biasing the phenol rings in an orientation that can be differentiated by the catalyst (such a rigid system would not be freely rotatable as depicted by Charton values). The importance of the L value suggests that groups that are too long disrupt the interaction between the catalyst and the substrate, leading to reduced selectivity. In general, Sterimol parameters can be considered to be superior to Charton values in accounting for steric contributions and should be the preferred descriptor for modeling studies.

The work described in the preceding study inspired a related investigation in which steric parameters of N-substituents are correlated to enantioselectivities in the alkylation of benzaldehyde with diethylzinc.[84] In that study, chiral 1-amino-2-phosphinamidocyclohexane ligands are employed in which the sizes of the amine substituent are represented with different steric parameters to uncover LFERs. With NR^1R^2 and CHR^1R^2 Charton values (the former derived from the

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689



Substrate	Ligand	Measured ΔΔG‡	Predicted ΔΔG‡	Error
S6	L7	0.37	0.35	0.01
S6	L8	0.63	0.57	0.01
S7	L7	0.67	0.56	0.03
S7	L8	0.82	0.68	0.05
S8	L7	0.54	0.46	0.01
S8	L8	0.67	0.70	0.01
S9	L7	0.90	0.78	0.01
S9	L8	1.10	1.02	0.01

Figure 14. Enantioselective propargylation reaction (top) and test catalyst/substrate combinations with their predicted and observed values in kcal/mol (bottom). Reproduced from Harper, K. C.; Vilardi, S. C.; Sigman, M. S. Prediction of Catalyst and Substrate Performance in the Enantioselective Propargylation of Aliphatic Ketones by a Multidimensional Model of Steric Effects. *J. Am. Chem. Soc.* **2013,** 135, 2482−2485. Copyright 2013 American Chemical Society.

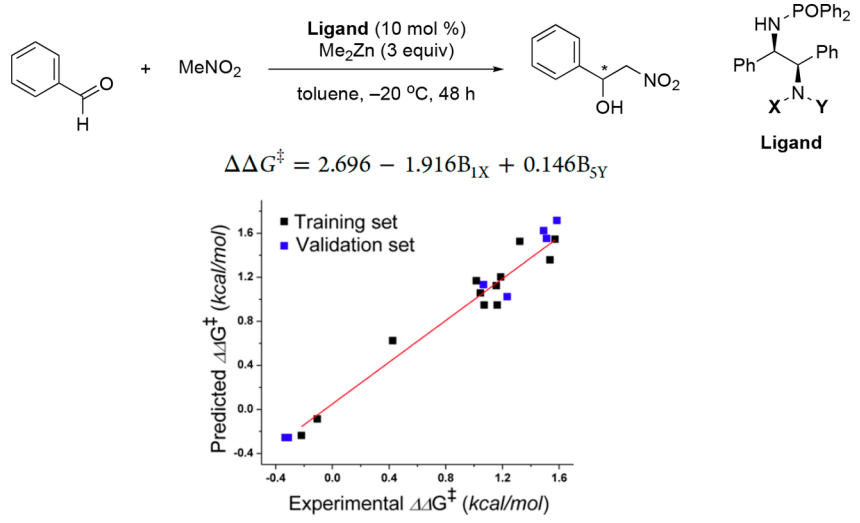


Figure 15. Enantioselective Henry reaction catalyzed by 1-amino-2-phosphinamido ligands. Reproduced from Huang, H.; Zong, H.; Bian, G.; Yue, J.; Song, L. Correlating the Effects of the N-Substituent Sizes of Chiral 1,2-Amino Phosphinamide Ligands on Enantioselectivities in Catalytic Asymmetric Henry Reaction Using Physical Steric Parameters. *J. Org. Chem.* **2014**, 79, 9455−9464. Copyright 2014 American Chemical Society.

hydrolysis of the corresponding amides and the latter linearly related to the amine hydrolysis), LFERs are constructed for secondary amines. However, when tertiary amines are included, the relationships break down. Turning to Sterimol values and using stepwise linear regression analysis, a relationship can be identified wherein the Sterimol B1 parameters for each substituent on the amine residue correlates with enantioselectivity (Figure 13).

The negative coefficient on the parameter for substituent X (−1.678) is interpreted as the interference of larger groups with ligand−metal binding, lowering the degree of enantioinduction, whereas the positive coefficient for the Sterimol parameter of substituent Y (+1.867) indicates that a large group is necessary on the amine. Another plausible explanation, which was not discussed in the manuscript, relates to the distribution of the two, in situ-generated diastereomeric (in which the coordinating nitrogen is stereogenic) ethylzinc−diamine complexes, which are the active catalysts. Perhaps the disparate sizes of X and Y substituents affects such a distribution favorably toward the isomer providing higher selectivity. The stereocenter proximal to the catalytic center could also be responsible for enhanced enantioselectivity. It is worth noting that neither internal nor external validation of the model was performed; doing so

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

would increase the confidence of these conclusions. This work was later expanded to include ketones as substrates with similar results and molar refraction was also demonstrated to be a suitable descriptor for the LFER.[85]

With new, robust steric parameters identified, Harper, Vilardi, and Sigman sought to obtain models capable of predicting reaction outcome for a range of substrates and obtain mechanistic insight on the origin of enantioinduction in new reactions.[86] As a proof of concept, the enantioselective propargylation of aliphatic ketones is used as a model system (Figure 14). Sterimol parameters (B_1, B_5, and L) and cross-terms are used for pyridine-proline based catalysts (in which the protecting group on the proline residue is varied) and the substituent of different aliphatic, methyl ketones is also varied. A combination of six catalysts and five substrates provides a training set of 30 compounds, which is used to generate a model. The model is externally validated by pairwise combination of four external substrates with two external catalysts (8 external validation reactions in total). The small error (all external validation cases are predicted within 0.2 kcal/mol) supports the hypothesis that this method can predict the outcome of new substrates excluded from the training data. To expand the utility of this protocol beyond methyl ketones, a new model is derived for predicting the results from cyclic ketones as substrates. However, because only steric parameters are employed, the model failed to predict enantioselectivities of electronically disparate substrates. The authors suggest that implementing steric and electronic descriptors would lead to more robust models.

Song and co-workers employ Sterimol parameters to aid in catalyst optimization of an enantioselective Henry reaction using chiral 1-amino-2-phosphinamido ligands in the presence of dimethylzinc.[87] An original set of 12 1-amino-2-phosphinamido ligands was used as a training set, with selectivity values ranging from 39.3:60.7 to 95.8:4.2 er. With B1, B5, and L Sterimol parameters for both substituents on the amine residue of the catalyst, a model is constructed with a strong, negative dependence on the B1 parameter of the X substituent, and a smaller, positive dependence on the B1 parameter of the Y substituent (Figure 15). On the basis of validation results with an external test set, the authors postulate that a significant size difference of the nitrogen substituents is necessary for stereoinduction. A noteworthy observation is that all selective catalysts contain a secondary amine (X = H in Figure 15), which is likely deprotonated under the reactions conditions. This process would change the coordination environment around the zinc which may also be necessary for stereoinduction. The dependence of the B1 parameter on X arises because X must be H for the catalyst to be selective, in which case the only catalyst tuning would be the identity of Y.

Another example of using LFERs to aid in mechanistic understanding of enantioselective reactions is the silylation-based kinetic resolution developed by Wiskur and co-workers.[88] In this work, LFERs are used to probe the mechanism of the tetramisole-catalyzed, kinetic resolution of secondary alcohols by selective silylation with silyl chlorides. Steric and electronic factors are probed by plotting the log of the *s*-factor against Charton and Hammett values for the substituents on the silyl chloride. The electronic effects are probed by variation of electronically disparate substituents at the 4-position of various triarylsilyl chlorides, and the Charton values for these substituents are used to probe steric effects. Electronic effects are found to be dominant, whereas steric effects are observed only when electronically similar but sterically disparate groups are employed (e.g., methyl vs *t*-butyl). By combining these terms, the authors construct a multivariate free energy relationship relating the stereospecificity to σ_{para} and Charton values ($\log(s) = -0.6\sigma_{para} + 0.09\nu$), in which the larger magnitude of the electronic term is indicative of the relative importance of the electronic effects with respect to steric effects. The authors conclude that positive charge is decreasing in the selectivity-determining transition structure, likely because the tetramisole catalyst is being displaced by the secondary alcohol. Electron donating groups on the silicon electrophile thus cause a later transition structure, in which the silicon−oxygen interatomic distance should be relatively shorter. Because the involvement of the alcohol in this transition structure is greater than in the case wherein electron-withdrawing groups are attached to silicon, the energy differential between diastereomeric transition structures is greater, corresponding to higher selectivity.

Although most parameters in physical organic chemistry seek to isolate the effects of specific interactions, Sigman and co-workers sought to devise a parameter set capable quantifying simultaneous, nonadditive interactions.[89] In this work, molecular vibrations are identified as a descriptor set that could quantify the interaction important for selectivity while capturing the interplay of multiple steric or electronic interactions (Figure 16). To evaluate the effectiveness of these features, the desymmetrization of bisphenols discussed above was studied.[72] In the original work, only steric parameters were included and the substituents studied were different in steric but not electronic character. However, when the electronically dissimilar CCl_3 substituent is tested, the enantioselectivity is much lower than would be expected from steric considerations alone. Thus, the authors include molecular vibrations of the aromatic ring (1700−1500 wavenumber region) in addition to Sterimol parameters to construct a model capable of identifying both steric and electronic properties responsible for enantioinduction. As illustrated in Figure 16, the model that includes molecular vibrations more accurately describes electronically dissimilar substrates than the original model.

In a second study from the same publication, the enantioselectivity of iridium-catalyzed hydrogenation of *α*-substituted styrenes is correlated with vibrations of the substrate molecule. In this case, intensities rather than frequencies are identified as the most important descriptors in construction of a predictive model that is externally

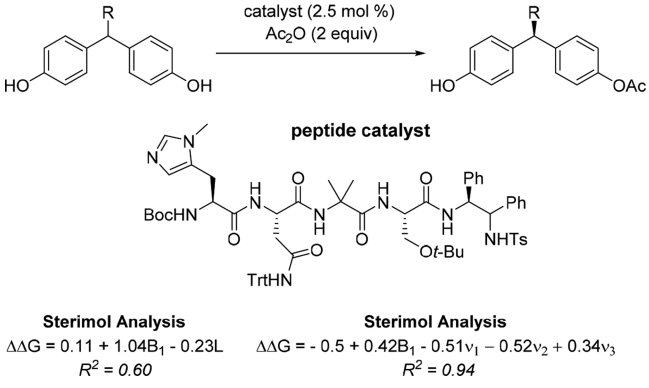


Figure 16. Models with and without molecular vibrations as descriptors.

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

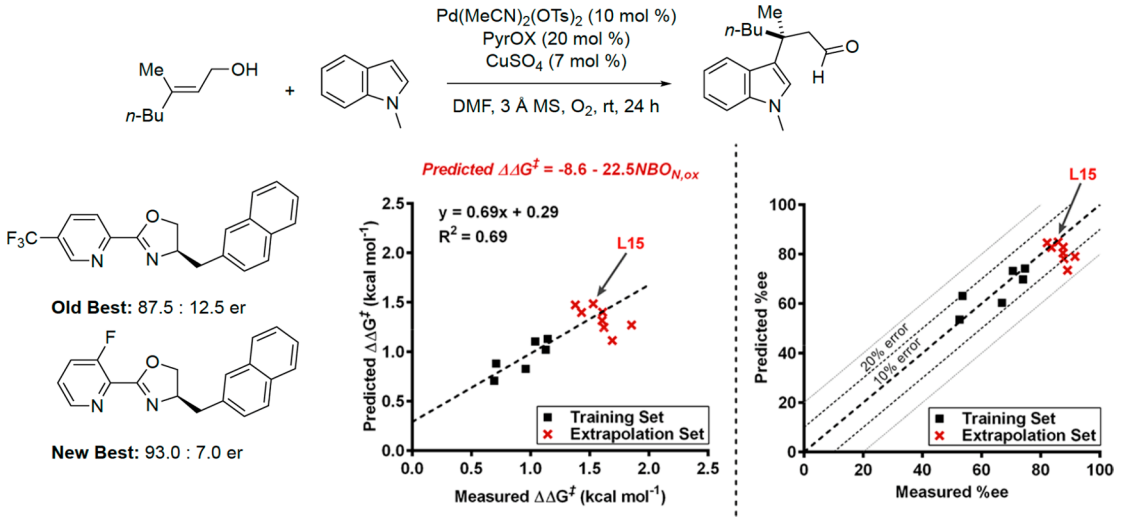


Figure 17. Dehydrogenative Heck reaction optimized using a predictive model. Reproduced from Zhang, C.; Santiago, C. B.; Crawford, J. M.; Sigman, M. S. Enantioselective Dehydrogenative Heck Arylations of Trisubstituted Alkenes with Indoles to Construct Quaternary Stereocenters. *J. Am. Chem. Soc.* **2015**, 137, 15668−15671. Copyright 2015 American Chemical Society.

validated. Finally, a third study in the same publication assesses if vibrational analysis could serve as an alternative to Hammett analysis for situations in which aromatic rings are substituted with multiple groups or bore *ortho* substituents. The reaction employed in this study is the enantioselective, redox-relay Heck reaction. Because the models are used to predict site selectivity instead of enantioselectivity, this section will not be thoroughly discussed as it is outside the scope of the review. However, the authors are able to use molecular vibrations to construct models accurately predicting the site-selectivity.

Sterimol values and molecular vibrations have also been used in the design of a substrate library to create a workflow to quantitatively assess substrate scope.[90] For development of a prototypical workflow, the enantioselective NHK reaction for the propargylation of ketones was used as a model system. This workflow consists of four steps: (1) identification of adequate, numerical representations of compounds that capture the relevant physical properties of those compounds, (2) selection of a set of substrates distributed evenly in the space constructed from those physical properties and measuring their experimental outcome, (3) construction of a mathematical model relating the descriptors to the experimental outcome, and (4) external validation of the model with new substrates.

The substrate space is first defined by tabulating Sterimol values and calculating the carbonyl IR stretching frequency for 52 different methyl ketones. Of these, eight are selected in a fashion analogous to design of experiment (DoE) sampling. The eight substituents on the ketones are then used to construct 28 differentially substituted ketones (now no longer methyl ketones, but ketones derived from pairwise combinations of the substituents), for which differential Sterimol values and IR frequencies are calculated, defining the relevant ketone space. This space is populated with substrates spanning the defined dimensions. To evaluate the library, the ketones representative of the substrate space are tested in the rhodium-catalyzed, asymmetric transfer hydrogenation (ATH) reaction, in which dialkyl and aryl/alkyl ketones are modeled separately. In model development, a different set of descriptors is used, including various IR frequencies, atomic charges of the carbonyl oxygen, carbonyl carbon, and both α-carbons, and Sterimol parameters. Predictive models for both classes of ketones are constructed and externally validated with high accuracy for both external aryl/alkyl and alkyl/alkyl ketones (R^2 = 0.97 and 0.95, respectively). Thus, by using a strategically selected initial set of substrates, it is possible to quantitatively predict reaction outcomes of new substrates.

An interesting extension of this work would be to compare different methods of defining substrate space and selecting representative substrate sets. For example, in this work, the relevant substrate space is constructed using a different set of parameters than were used to construct a model. Thus, it becomes difficult to say if new predictions are interpolative in the chemical space used by the model. It would be interesting to use descriptor inputs for the stepwise regression algorithm and select a subset from the high-dimensional space. Alternatively, principal component analysis (or other variable selection/dimensionality reduction methods) could be used on this space to reduce the dimensionality and a subset could be selected from this set. Although in this application the current method is clearly sufficient, a more thorough comparison could be worthwhile in other settings.

Using stepwise linear regression to facilitate catalyst optimization has been employed in multiple settings. One such example is optimization of a PyrOx for the dehydrogenative Heck reaction between trisubstituted alkenes and indoles reported by Sigman and co-workers.[91] In this report, the authors initially use the PyrOx ligand identified in an earlier study (Figure 17).[83] However, when this ligand is tested in the dehydrogenative Heck reaction, the selectivity is found to be sensitive to the substitution on the trisubstituted alkene. Thus, a series of ligands are tested and a relationship established in which the NBO charge on the oxazoline nitrogen is correlated with selectivity. This observation prompted the authors to design a catalyst predicted to have greater selectivity on the basis of this relationship, leading to a more selective catalyst.

An example of the use of statistical methods to identify structurally relevant features of catalysts has been described as part of a collaboration between the Toste and Sigman laboratories.[92] In this work, a chiral-phosphoric-acid catalyzed, dehydrogenative C−N coupling is investigated, wherein 12 substrates and 11 catalysts are systematically selected, and

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

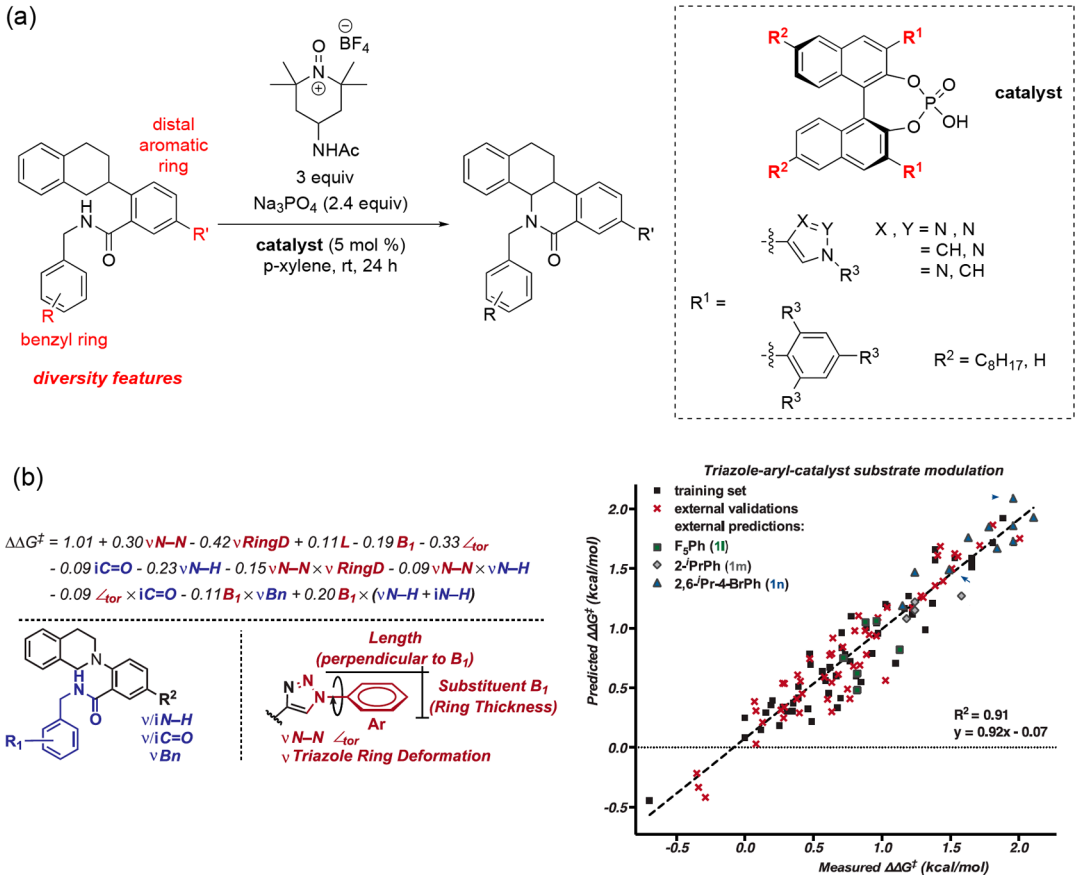


Figure 18. (a) Phosphoric acid catalyzed dehydrogenative C−N coupling indicating substituent variability in substrates and catalyst members. (b) Multivariate LFER with indicated descriptors, for the triazole substituted catalyst and the resulting correlation. Adapted with permission from ref 92. Copyright 2015 American Association for the Advancement of Science.

experimental data is collected for every pairwise combination (Figure 18). For each substrate and catalyst, descriptors are tabulated including Sterimol parameters, interatomic distances in optimized geometries, vibrational frequencies, and vibrational intensities. Empirical trends are examined for each substrate-catalyst pair. The combination of every catalyst with one substrate and *vice versa* was used to construct models. Results from this analysis stimulated the formulation of multiple mechanistic hypotheses on the origin of enantioselectivity. These hypotheses include: (1) the triazole ring at the 3,3′-positions of the BINOL-phosphoric acid (present in the most selective reactions) is engaged in a π-interaction with the substrate, (2) the strength of this interaction is modulated by the steric and electronic properties of the triazole substituents, (3) π-interactions are strengthened by heteroatoms, which is why triazole-containing catalysts are more effective than those with two or zero nitrogen atoms in the ring, and (4) benzyl and remote aryl groups on the substrate participate in π-interactions with the catalyst. Observations supporting these hypotheses include: (1) the triazole vibrational frequency and the torsion angle between the triazole ring and its substituent are selected as important parameters in catalyst models, and (2) the selection of various descriptors capturing perturbations on the benzyl or distal aromatic ring are selected in substrate models.

With 108 catalyst/substrate combinations, a model is constructed with 54 of the 108 total reactions, and the remaining 54 were used as an external validation set. This model is used to guide the selection of catalysts used to further validate the mechanistic hypotheses. First, a catalyst is selected in which the aromatic substituent on the triazole ring is a pentafluorophenyl group (R^3 on triazole substituent in Figure 18). The selectivity with this catalyst is predicted to be similar to catalysts with 2,6-difluorophenyl, 2,6-dimethoxyphenyl, and 1-adamantyl substituents, suggesting that this substituent provides only a steric contribution; the electronic contribution is negligible. This hypothesis is experimentally validated. A second catalyst is selected such that only the 2-position of the aromatic substituent of the triazole bears an isopropyl group, probing the hypothesis that bulky groups at both the 2 and 6 positions on the aromatic residue bound to the triazole interact unfavorably with benzyl groups containing substituents at their *para*-positions, thus disrupting the π-interaction responsible for enantioinduction. This catalyst is both predicted and observed to be more selective than the catalyst analogs with 2,6-disubstituted aromatic residues, supporting this hypothesis. Finally, to test if structural modifications suggested by the model could lead to a higher selectivity catalyst, namely through modulation of the torsional angle between triazole and its substituent, a catalyst is selected with a torsional angle close to 90°, which is more selective than a previous catalyst for the class of substrates with 4-benzyl substituents, validating the hypothesis that the torsional angle between the triazole and its substituent is necessary for high selectivity.

To more accurately probe noncovalent contributions, new parameters have been developed to quantify π-interactions.[93] These parameters, termed Eπ and Dπ, can be calculated for π-stacking interactions (^{S}Eπ and ^{S}Dπ) and T-shaped C−H−π

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

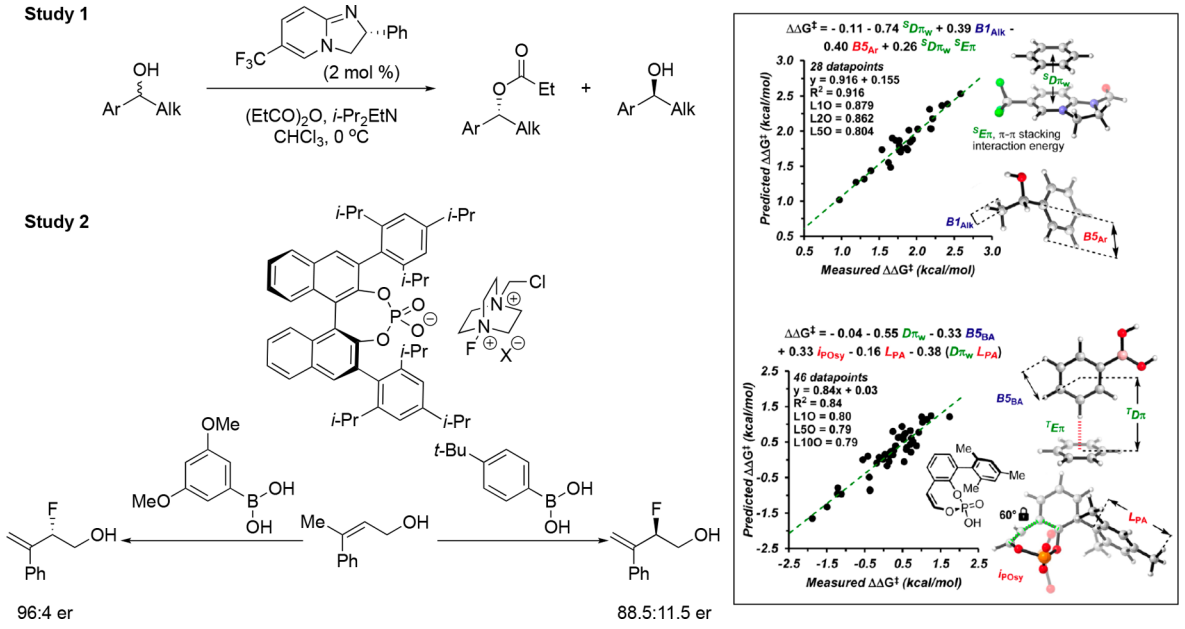


Figure 19. Predictive models for the kinetic resolution of secondary alcohols (top) and the enantioselective fluorination of allylic alcohols (bottom). Reproduced from Orlandi, M.; Coelho, J. A. S.; Hilton, M. J.; Toste, F. D.; Sigman, M. S. Parameterization of Noncovalent Interactions for Transition State Interrogation Applied to Asymmetric Catalysis. *J. Am. Chem. Soc.* **2017**, 139, 6803−6806. Copyright 2017 American Chemical Society.

interactions ($^{T}E\pi$ and $^{T}D\pi$), wherein $E\pi$ is the interaction energy between the arene substituent in question and a probe π-surface (such as benzene) and $D\pi$ is the distance between the midpoints of the arenes in the optimized geometry. These descriptors are used to analyze important noncovalent interactions in a well-understood kinetic resolution of chiral benzylic alcohols[94,95] as well as the enantioselective fluorination of allylic alcohols in which the mechanism of enantioinduction is less well understood (Figure 19). In the first reaction, a multivariate relationship is identified in which the $^{S}D\pi$ parameter and the $^{S}D\pi$ $^{S}E\pi$ cross-term are found to be significant, consistent with what is previously reported, thus validating the new descriptors. The authors then apply these descriptors to understand the enantioselective fluorination of allylic alcohols. Multidimensional modeling reveals the significance of the $^{T}D\pi$ term, suggesting the importance of a T-shaped C−H-π interaction in the stereodetermining step. This interaction is validated by DFT studies, which indicate that aromatic C−H bonds interact with the π-system in the case of 2- or 4-substituted boronic acids and that C−H bonds of the 3-methoxy substituent on boronic acids are involved in this interaction, thus rationalizing the inversion of selectivity upon inclusion of boronic acids as additives.

Yu, Sigman, and co-workers report the use of these multivariate regression methods to identify a descriptor set for amino-acid-palladium complexes as well as to select a set of compounds with which to start screening campaigns for diverse reactions (Figure 20).[96] Descriptors are derived from either amino acid derivatives bound to the palladium complex or from the free amino acids. Parameters including Sterimol parameters, torsional angles, percent buried volume, NBO charges, and vibrational frequencies are used to construct predictive models in each case. These studies include the enantioselective C_{sp2}-H activation and functionalization of substituted pyridines and carboxylates as well as the enantioselective C_{sp3}-H activation of triflimide or amide substituted cyclopropanes. In each case, predictive models are calculated, which are internally or externally validated, and descriptors from the amino-acid-bound complex give the best performance. Using the information from these models, the authors select five amino acid side chains representing a broad range in the dihedral angle of the amino acid backbone as well as five N-protecting groups according to NBO charges of the corresponding carbonyl oxygens. Assuming the pairwise combinations of these two groups span the breadth of descriptors space relevant to stereoinduction, the authors suggest a subset of compounds with which to begin screening campaigns.

Sigman, Toste, and co-workers have applied multivariate methods to the analysis of chiral, phase transfer catalysts in the enantioselective Pummerer reaction (Figure 21).[97] Using the standard set of descriptors applied in previous publications, the authors probe substrate and catalyst features. Steric parameters of the substituent at the 3,3′-positions of the chiral catalyst are identified to be important for catalyst selectivity. Similarly substrates are evaluated, and the average charges on aliphatic atoms and the size of the aromatic substituent on the N-protecting group are found to be associated with catalyst efficacy.

Linear free energy relationships have also been applied to the enantioselective, palladium-catalyzed substitution of allylic alcohols with unsymmetrical 1,3-dicarbonyl nucleophiles and chiral diamine catalysts (Figure 22).[98] The authors devise what is described as a double layer Sterimol model in which parameters of the diamine catalyst subunit are described with two parameters intended to represent the “inner sphere” and “outer sphere” steric effects of the substituent of interest. The inner sphere is a single substituent on the carbon attached to the nitrogen atom whereas the outer sphere refers to the Sterimol parameter of the entire substituent on the tertiary amine (each described with the B1 parameter). Multiple models are constructed with the best model trained on seven

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

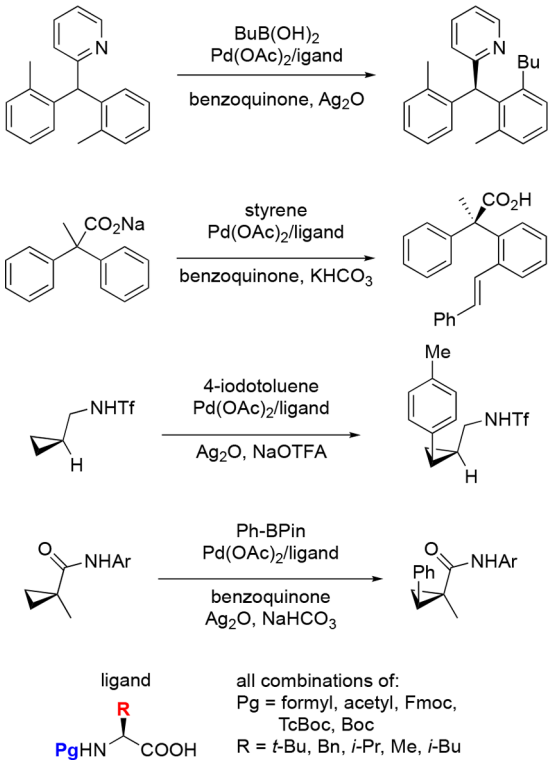


Figure 20. Four case studies to identify structural effects of amino acid ligands in palladium catalyzed C−H activation reactions and the resulting suggested ligand set. Reproduced from Park, Y.; Niemeyer, Z. L.; Yu, J.-Q.; Sigman, M. S. Quantifying Structural Effects of Amino Acid Ligands in Pd(II)-Catalyzed Enantioselective C−H Functionalization. *Organometallics*. **2018**, 37, 203-210. Copyright 2018 American Chemical Society.

data points and externally validated with five data points (two as an external test set and three synthesized as an expansion of the external set after construction of the model), although the structural diversity of both sets is limited. The authors suggest the larger "inner sphere" group reflects the inability of the amine substituent to rotate away from the nucleophile, thus blocking one face more effectively and leading to increased stereoinduction.

A recent example in which subunit-derived descriptors are used to predict the enantioselectivity of reactions is in the enantioselective benzoin condensation catalyzed by *N*-heterocyclic carbenes (NHCs) (Figure 23).[99] In this work, dynamic modulation of the chiral environment around the catalyst is achieved by complexation of a free hydroxyl residue with the boronic acid additive under the reaction conditions. Multivariate regression with stepwise variable selection was employed to understand the influence of substrate structure on enantioselectivity. Features including Hammett values, Sterimol parameters, vibrational frequencies, vibrational intensities, NBO charges, and components of the dipole moment in an aligned coordinate system were considered as independent variables. Three variables are selected to form a cross-validated model (Q^2 = 0.83, R^2 = 0.92 for predicted vs observed): (1) the Sterimol B_5 parameter, (2) the Sterimol L parameter and, (3) the dipole moment on the axis of maximal length. It was postulated that the dipole moment component would reflect the electron withdrawing nature of the aromatic residue, thus correlating to the configurational lability of the product. Hence, as the value of this variable increases selectivity should go down. This the B_5 parameter is actually related to an electrostatic interaction with the aromatic residue of the catalyst. This hypothesis is experimentally probed by comparing the selectivity values of two substrates, 3-carbomethoxybenzaldehyde and 4-cyanobenzaldehyde, with three catalysts with different aromatic residues: phenyl, pentafluorophenyl, and mesityl. The results with 3-carbomethoxybenzaldehyde are influenced by the identity of the aromatic residue and the phenyl-substituted catalyst is more selective than the pentafluorophenyl substituted catalyst. In contrast, the reaction outcome with 4-cyanobenzaldehyde is not influenced. Perhaps an electrostatic descriptor that captures this interaction can be added to the analysis; such a feature may serve as a mediator of the interaction and identifying it would strengthen this hypothesis.

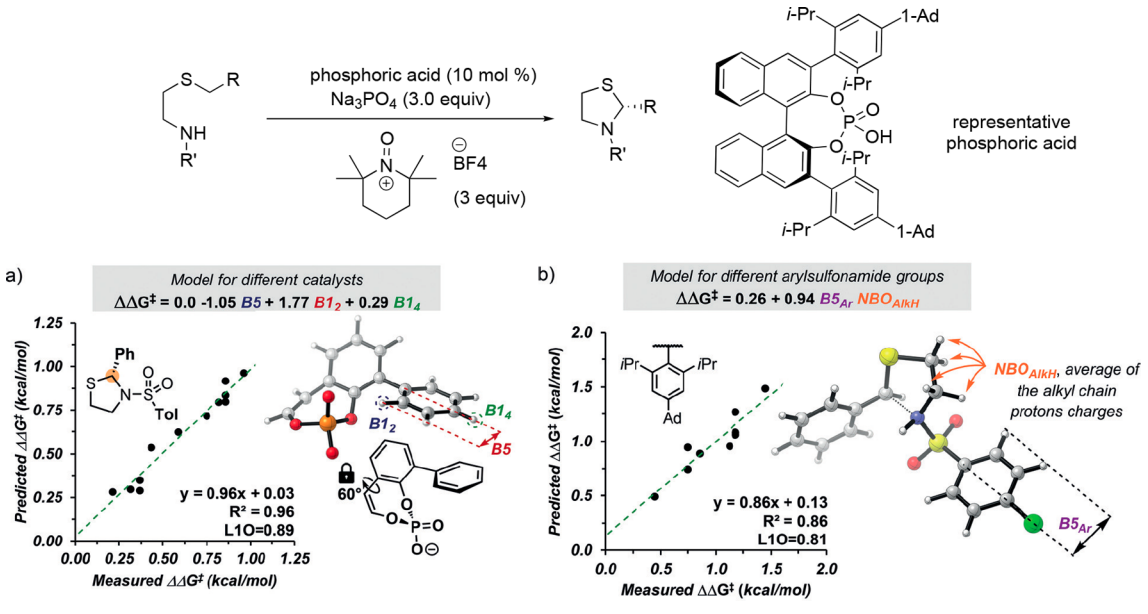


Figure 21. Multivariate LFER for an enantioselective Pummerer reaction. Adapted with permission from ref 97. Copyright John Wiley and Sons.

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

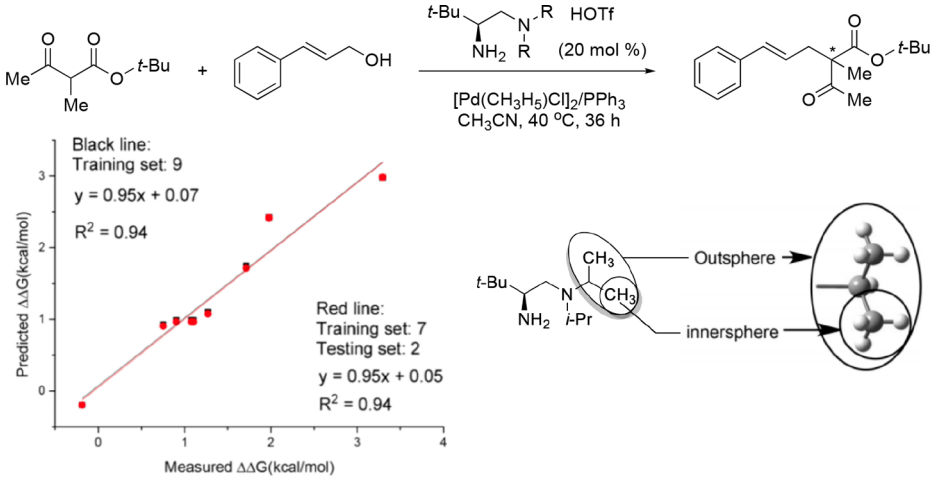


Figure 22. Enantioselective, palladium catalyzed substitution of allylic alcohols. Predicted versus observed selectivities for various substrates for two different catalyst classes. Reproduced from Wang, Y.; Zhou, H.; Yang, K.; You, C.; Zhang, L.; Luo, S. Steric Effect of Protonated Tertiary Amine in Primary-Tertiary Diamine Catalysis: A Double-Layered Sterimol Model. *Org. Lett.* **2019**, 21, 407−411. Copyright American Chemical Society.

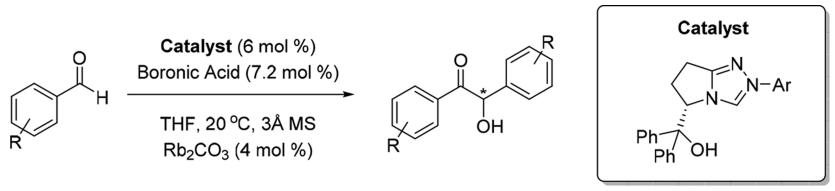


Figure 23. Enantioselective benzoin reaction.

The inclusion of the L parameter is not discussed, and it would be interesting to see if this parameter were selective if a different variable selection algorithm were used. In the analysis of boronic acid additives, a linear model could not be identified; instead, a regression tree was constructed. Catalyst features identified as the most important include: (1) the torsion angle between the boronic acid moiety and the aromatic residue, (2) the dipole component along the width axis, and (3) the dipole component along the axis of longest length. The authors comment that although the model is predictive, it is difficult to obtain mechanistic information from this model given that two factors influence selectivity—benzoin formation and racemization of the product. This study serves as an excellent example in which statistical methods and experimentation are used in concert to probe the origin of stereoinduction.

Carbo and co-workers describe an interesting strategy for devising steric descriptors for chiral oxazoline ligands in the copper-catalyzed, enantioselective cyclopropanation of styrenes with diazo esters.[100] The authors suggest that quadrant diagrams,[101] which are often used in asymmetric catalysis to understand catalyst selectivities, may also be employed to quantify steric properties around a reactive center. Data obtained from previous experimental studies[102−104] were used to develop the parametrization of the chiral complexes (Figure 24).

Construction of the descriptors for the quantitative quadrant model proceeds first by placing the ligand−metal complex (using a cationic copper complex bound the = CH_2 *in silico*) in the *xz* plane of a Cartesian coordinate system, with the copper atom placed at the origin and the coordinating nitrogen atoms in the *xz* plane with negative values of *z*. The *xy* plane can then be divided into four quadrants (each combination of positive and negative values of *x* and *y*). The steric parametrization of each quadrant is then dictated by the distance-weighted volume parameter (V_w), described by the following equation:

$$V_{w,k,l} = \sum_{i=1}^{n} \frac{r_i^k}{d_i^l} \quad (1)$$

In this expression, $V_{w,k,l}$ is the distance-weighted volume, *r* is the van der Waals radius of a given atom in a quadrant, *d* is the distance between that atom and a reference atom (in this case, the copper atom), and *k* and *l* are exponents that can have values 0, 1, 2, or 3. Thus, the distance-weighted volume for a quadrant is the summation of the van der Waals radii divided by the respective distance between that atom and the copper center for every atom in the quadrant. In this analysis, only atoms with positive values along the *z* coordinate are considered (in which the copper atom is the origin and the ligand nitrogen atoms are in the negative *z*-direction). For comparison, Taft-Charton values are also used as steric parameters wherein the value of the substituent at the 4-positions of the oxazoline is used for the corresponding quadrant (Figure 24).

Three PLS regression models are constructed from 30 experimental results using either Charton parameters, $V_{W,0,3}$, or $V_{W,1,3}$ for each quadrant as descriptors. Using Q^2 as a metric to assess model efficacy, the model with Charton values (Q^2 = 0.78) outperformed those constructed with $V_{W,0,3}$, or $V_{W,1,3}$ (Q^2 = 0.76 and 0.70, respectively). Similarly, models including cross-terms between quadrants followed similar trends, with Charton values outperforming $V_{W,0,3}$, or $V_{W,1,3}$ (Q^2 = 0.88, 0.76, and 0.70, respectively). Thus, using steric parameters that characterized the environment of specified regions around a reactive center, QSSR models are constructed relating the

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

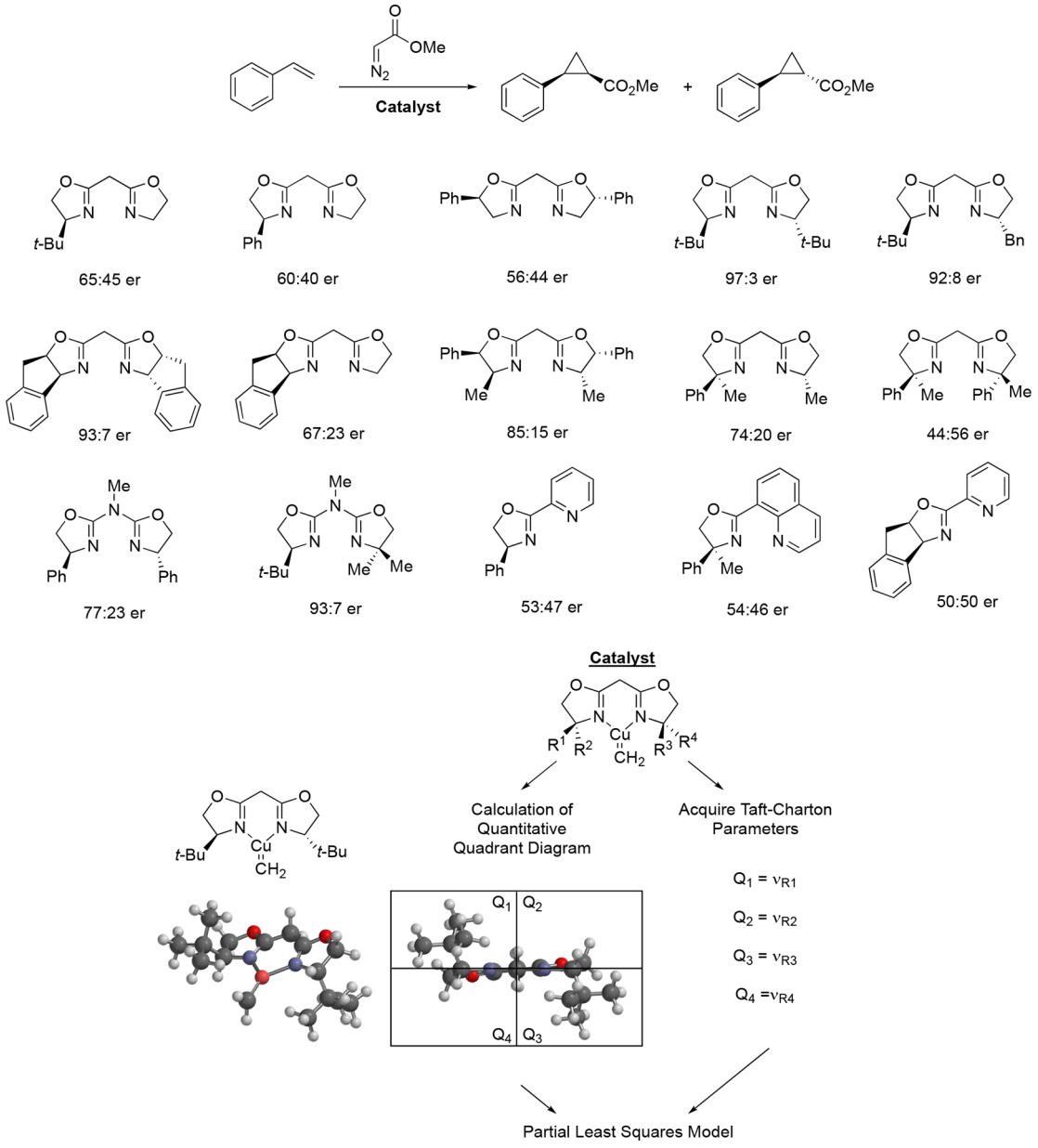


Figure 24. Copper catalyzed, enantioselective cyclopropanation with a representative oxazoline ligand set (top) and generation of distance-weighted volume and Charton-Taft values (bottom).

energy differential between diastereomeric transition structures to these structural features. This proof-of-concept study has interesting implications—for example, it is possible that the reason models with Charton values give the best performance is that the value inherently considers flexibility of the substituents residing in a particular quadrant. A suggested future experiment would be to use Sterimol parameters for each quadrant or to use a conformer-dependent V_W parameter to see if improved models could be constructed thus increasing the accuracy of the quantitative quadrant model. This procedure would facilitate calculation of the parameter (an advantage of V_W over Taft-Charton values) without loss of accuracy.

A final, recent advance is the development of conformationally weighted Sterimol parameters, developed by Fletcher and Patton.[61] In this work, open source software is used to calculate parameters that that are derived from a conformer distribution of structures. For rigid systems, the descriptors perform similarly to previously published work, but for flexible scaffolds a significant improvement is observed. Further, the software is easy to use and freely available thus allowing access to the calculable parameters for nonexperts. This enabling technology will likely increase the use of LFERs in catalyst design, owing to the new found accessibility of the parameters.

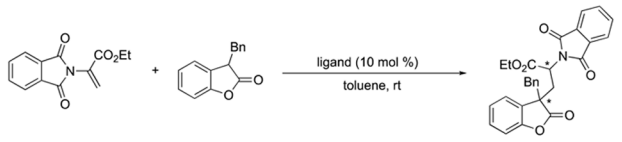
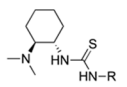
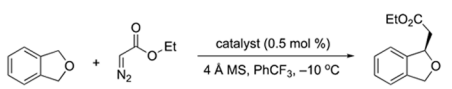
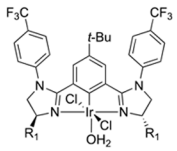
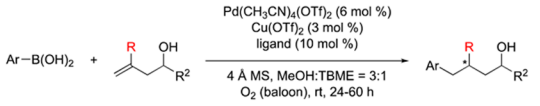
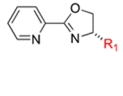
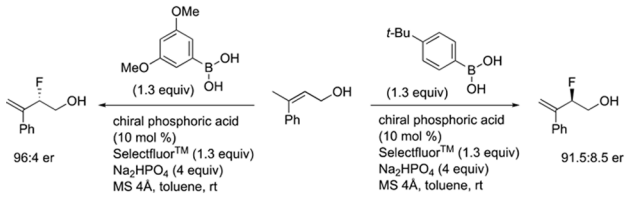
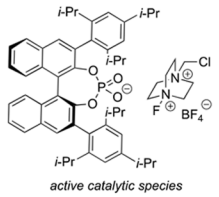
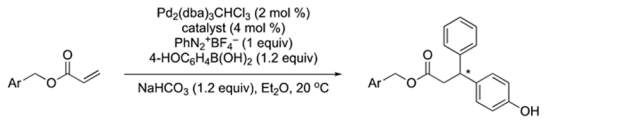
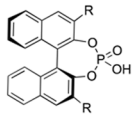
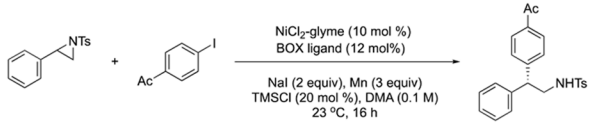
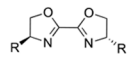
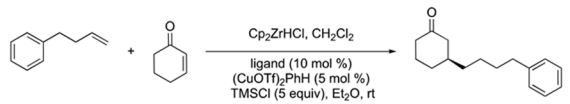
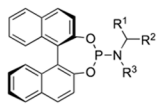
Selected cases described in this section illustrate the method and results obtained from a multivariate LFER analysis of enantioselective transformations. Additional examples are listed in Table 1; these reports are instructive for readers but are not described in detail because they are constructed using the same workflow.[105−118]

2.4. Perspectives on Linear Free Energy Relationships

The previous examples illustrate how descriptors for catalyst substituents can be used to identify a correlation of catalyst structure to selectivity. In these cases, only the variable

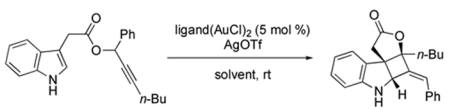
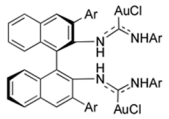
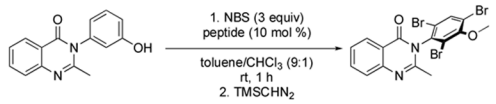
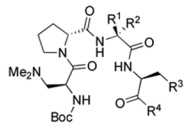
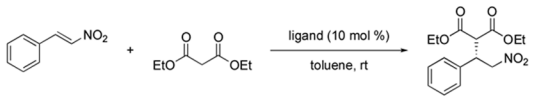
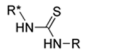
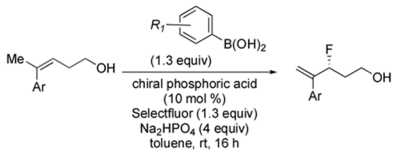
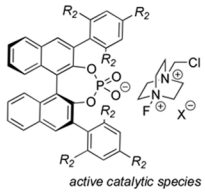
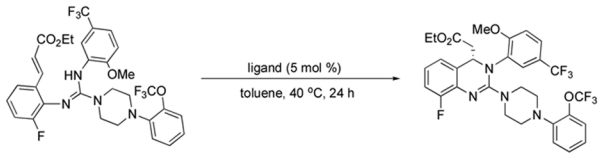
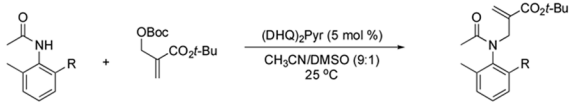
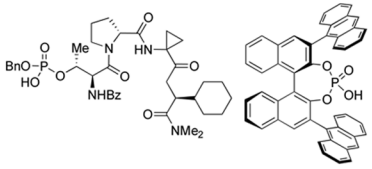
DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

Table 1. Additional Examples of Multivariate LFER Analysis for Correlation of Product Enantioselectivity with Catalyst/Substrate Descriptors

REACTION	CATALYST/LIGAND	DESCRIPTORS	REMARKS	REFERENCE
		Thiourea catalyst structure: Sterimol B1, N-H stretching frequencies NBO charges	No validation was provided	105
	 Test Set: R_1 = *i*-Pr, *i*-Bu, *n*-Bu, *s*-Bu, *c*-Hex, CH_2(*c*-Hex) Validation: R_1 = CH_2(*t*-Bu)	Bis(imidazolinyl)phenyl catalyst structure: Charton values, Austel's branching index	Externally validated with a single data point, optimal ligand already present in the training set	106
		Pyridine-oxazoline catalyst structure: stretching frequencies, polarizability Substrate alkene structure: alkene C-C vibration, O-H stretching frequency, NBO charges	This analysis led to the proposal of a transition-state structure in which the lone pair on the carbinol oxygen interacts with the phenyl substitutent on the oxazoline	107
	 *active catalytic species*	Boronic acid structure: B1, B5 Sterimol, phosphoric acid catalyst structure: torsional angle	In this reaction, the choice of boronic acid determines absolute configuration of the product.	108
		Acrylate structure: NBO charges of the aryl substituent vibrational frequencies Phosphoric acid catalyst structure: ^{S}E parameter, NBO charge, torsional angle, Sterimol B1, vibratonal frequencies	A comprehensive model with 119 data points encompassing varying catalysts and substrates was utilized for the analysis. A π-stacking interaction between the catalyst and substrate is suggested to contribute to enhanced selectivity.	109
		bisoxazoline catalyst structure: Sterimol, NBO charges, molecular vibrations, polarizability	A dispersive interaction between the ligand and substrate was postulated on the basis of the multivariate analysis performed using seventeen catalysts	110
		phosphoramidite catalyst structure: Sterimol parameters for R^1 and R^2, HOMO energy of R^1	On the basis of the analysis, a favorable Cu-arene interaction was proposed as a determinant of selectivty. New catalysts with higher predicted selectivities were succesfully identified.	111

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

Table 1. continued

REACTION	CATALYST/LIGAND	DESCRIPTORS	REMARKS	REFERENCE
		substrate structure: NBO charge of the carbonyl carbon. catalyst structure: interatomic bond distances.	The analysis is supported by external and internal validation.	112
		peptide catalyst structure: Sterimol, NBO charges, torsional angles, molecular vibrations,	The analysis is supported by external and internal validation.	113
	 R* = Chiral Substituent containing a free amine	Thiourea catalyst structure: Sterimol B1, Sterimol B5		114
	 *active catalytic species*	Boronic acid structure: B5 Sterimol, Phosphoric acid catalyst structure: ^{T}E, D parameter, NBO charge, torsional angle, Sterimol B1, vibratonal frequencies Solvent was also modeled	The analysis is supported by external validation.	115
		Diamine catalyst structure: Sterimol parameters, quantum chemistry derived descriptors, additional 2D and hybrid descriptors	-	116
	-	Substrate structure: B1 Sterimol parameters of R, chemical shift of amide proton	-	117
		substrate and catalyst structure: Sterimol parameters, NBO charges, molecular vibrations	Two different catalyst scaffolds - peptide based phosphoric acids and BINOL derived phosphoric acids were compared and evaluated as catalysts.	118

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

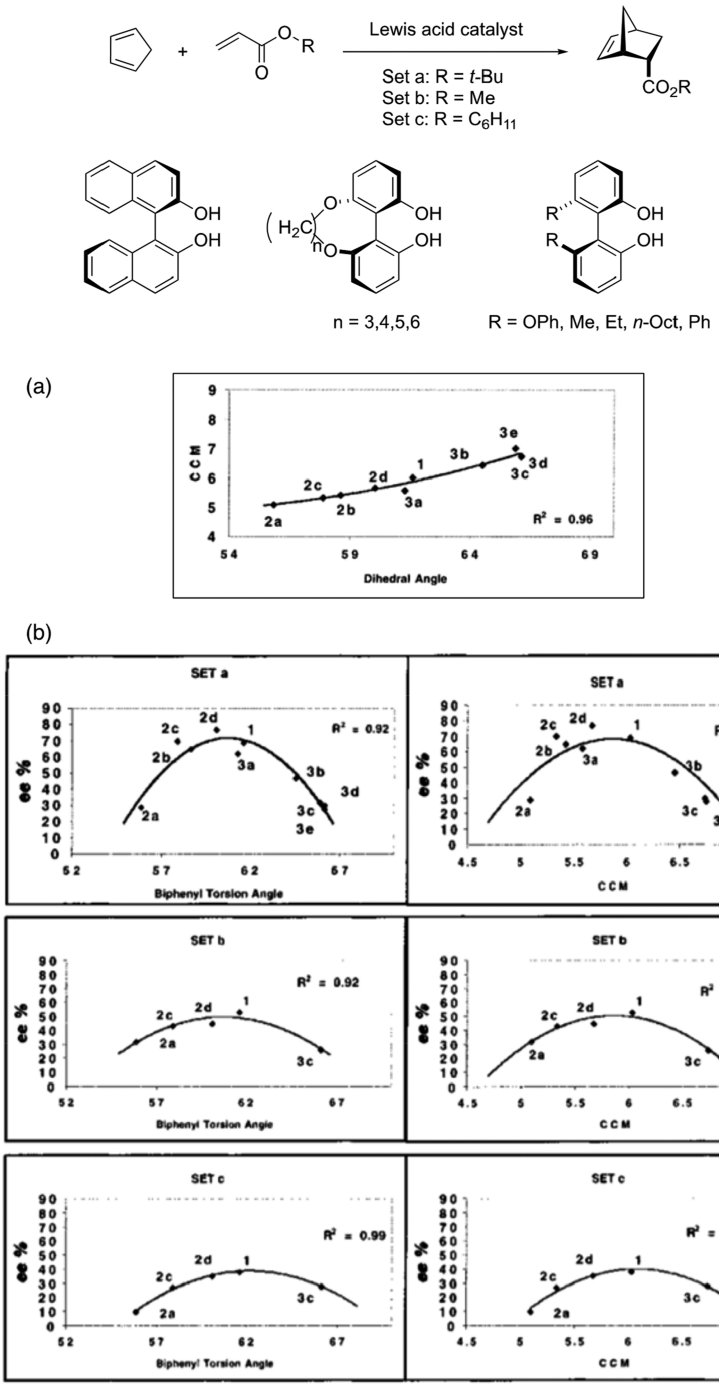


Figure 25. Chiral Lewis acid catalyzed, enantioselective Diels–Alder reactions. (a) CCM versus dihedral angle for 2,2′-biaryldiols. Sets a, b, and c refer to the different ester residues in the starting material. (b) Biphenyl dihedral angle and CCM as they relate to enantioselectivity. Reproduced from Gao, D.; Schefzick, S.; Lipkowitz, K. Relationship between Chirality Content and Stereoinduction: Identification of a Chiraphore. *J. Am. Chem. Soc.* **1999**, 121, 9481–9482. Copyright 1999 American Chemical Society.

substituents on a fixed structural scaffold are parametrized (local parametrization, see Figure 2). Such an approach assumes that the overall catalyst structure remains constant despite the changing substituents. Although this assumption may hold in certain cases, the parametrization of the entire catalyst structure may be ideal in other cases. Further, this approach assumes that all descriptors relevant to enantioinduction have been incorporated in the model. However, as previously described, omitting variables contributing to enantioinduction can still result in predictive models that do not have straightforward interpretations. Additionally, it is possible that different selection algorithms could identify different variables, thus resulting in predictive models generated with different interpretations. Finally, it is unlikely that the investigator is aware that an important variable has been omitted; if this were easily assessed, the variable likely would have not been omitted in the first place. These concerns do not imply that the use of subunit-derived descriptors to construct LFERs is not a valuable method by which mechanistic information can be garnered, simply that

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620–1689

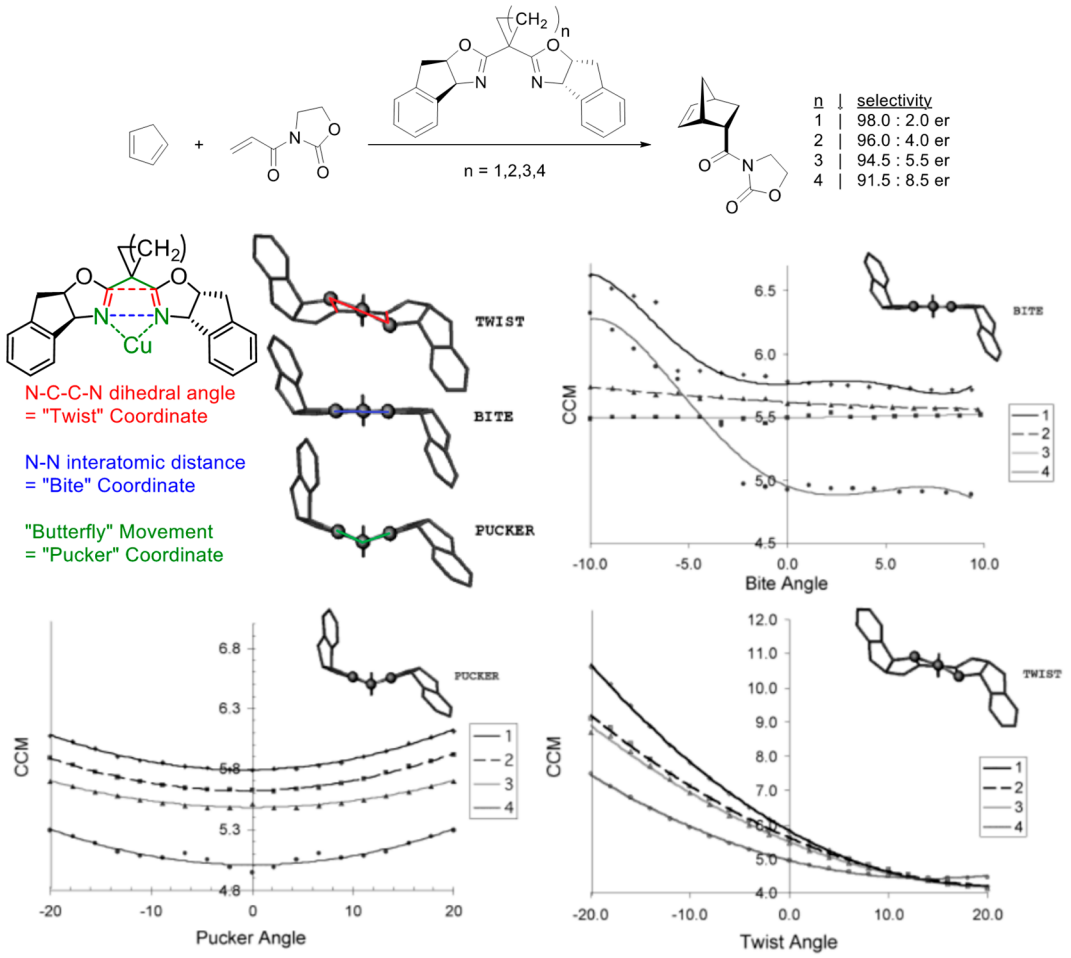


Figure 26. Illustration of twist, bite, and pucker distortions and their relation to CCM. Reproduced from Lipkowitz, K.; Schefziek, S.; Avnir, D. Enhancement of Enantiomeric Excess by Ligand Distortion. *J. Am. Chem. Soc.* **2001**, 123, 6710−6711. Copyright 2001 American Chemical Society.

researchers interested in using these techniques must be aware of the limitations and analyze their models critically. A global parametrization (see Figure 2) can potentially account for effects caused by subtle changes in the overall catalyst structure when various substituents are incorporated and can avoid the problems associated with missing variables. Strategies reliant on this approach are introduced in the next section.

3. CONTINUOUS CHIRALITY MEASURE

3.1. Background of the Continuous Chirality Measure

Continuous chirality measure (CCM) was first described by Avnir and co-workers as an extension of the related continuous symmetry measure (CSM).[119−122] The crux of this concept eschews the classical definition of chirality as a binary property of a molecule (i.e., either present or not present), but instead posits that the "degree of chirality" is a quantifiable property of a chiral molecule. Lipkowitz eloquently describes this concept using unsymmetrically substituted aryls (e.g., BINOL) as an example.[123] When the biphenyl is perfectly planar (dihedral angle of 0° between the two aromatic rings), the molecule is achiral. If this dihedral is rotated an infinitesimally small amount away from 0°, the molecule becomes chiral. Intuitively, if a molecule possessing such a small structural perturbation away from planarity (~1°) were isolable, the structure would likely be ineffective in stereodifferentiating reactions. However, as the dihedral angle increases, efficient stereodifferentiation becomes more likely because the molecule is becoming "more chiral" up to a certain point, until continuing to increase the dihedral approaches an achiral molecule once again.

Avnir constructed a mathematical formula capable of calculating this degree of chirality, derived from CSM as

$$S'(G) = \frac{1}{n}\sum_{i=1}^{n} ||P_i - \hat{P}_i||^2 \qquad (2)$$

wherein G is a given symmetry group, P_i is the original set of points, $\hat{P}_i$ is the corresponding points in the nearest G-symmetric configuration, and n is the total number of configuration points. The interpretation is best described by the original authors:[122] "The meaning of eq 2 is the following: find a set of points $\hat{P}_i$ which possess the desired symmetry (G symmetry), such that the total (normalized) distance from the original shape P_i is minimal." Because chirality is the absence of improper symmetry, searching over all achiral symmetry groups will give a minimal distance to achirality. Thus, molecules with a greater minimal distance to achirality (larger values of S'), that is a higher CCM, are more chiral.[122] CCM is thus a conceptually simple approach to provide a global parametrization of a chiral molecule.

3.2. Continuous Chirality Measure in Asymmetric Catalysis

Intuitively, it may be hypothesized that selective catalysis requires the catalyst to have at least a certain value of CCM, with larger values suitable for higher selectivity. Lipkowitz and

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

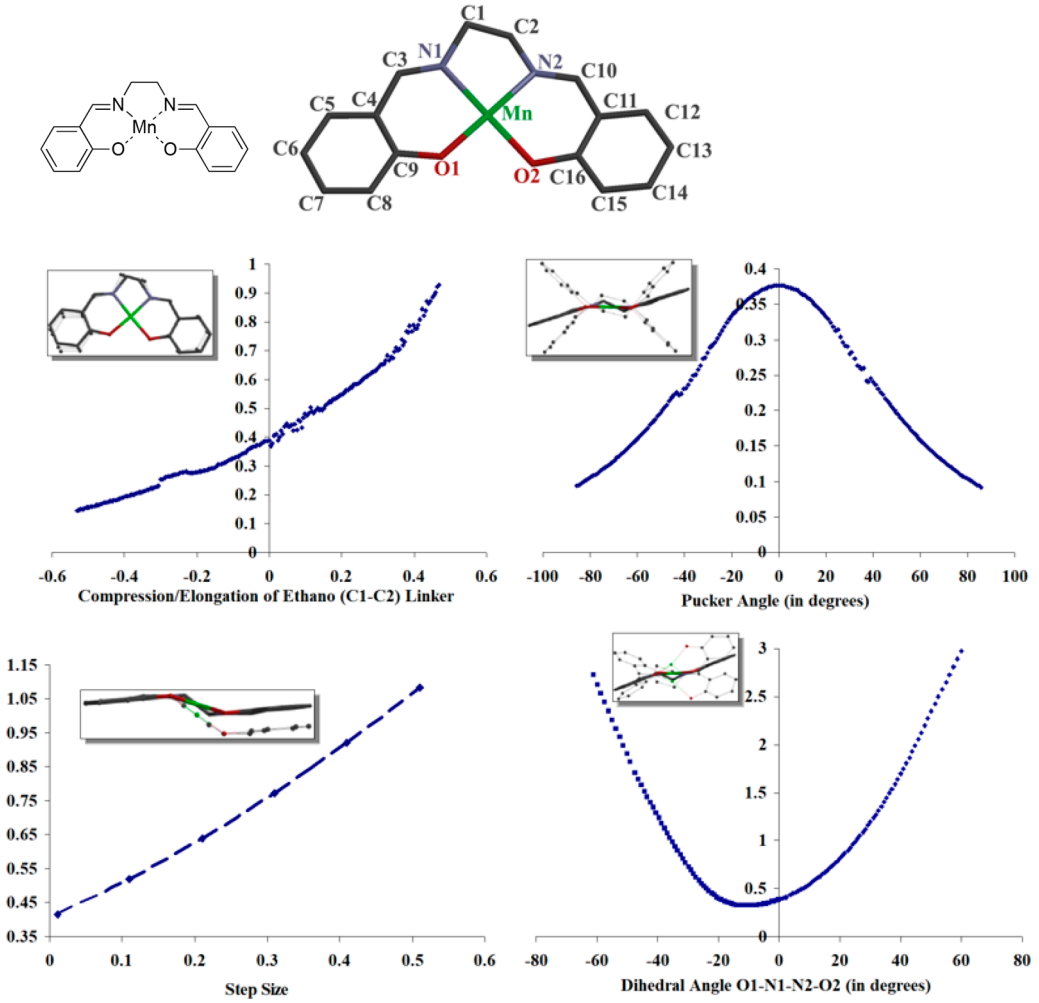


Figure 27. Four distortions studied in ref 128 and their relation to CCM. The numbering system described in the original work has been included for reference.

co-workers were the first to explore the relationship between CCM and enantioselectivity.[123] Under the premise that the chirality content of a molecule should correlate with enantioselectivity in asymmetric reactions, the enantioselective Diels–Alder reaction developed by Harada and co-workers, employing a chiral 2,2′-biaryldiol-ligated Lewis acid catalyzed was studied (Figure 25).[124] In this work, ten different catalysts were evaluated with three sets of substrates.

In both the original work and the work by Lipkowitz and co-workers, the calculated dihedral angle between the two arenes is strongly correlated to the observed enantioselectivity, with an optimal angle identified as ∼60°. As the dihedral angle is correlated linearly with CCM, the maximum selectivity is observed along the CCM coordinate, followed by a subsequent decrease in catalyst selectivity (Figure 25). The authors attribute this observation to the fact that not all atoms in a molecule contributing to overall chirality also contribute to enantiodifferentiation. Identifying subunits of the molecule with CCM values that best correlate to enantioselectivity thus identifies a "chiraphore", the subunit of a molecule responsible for its selectivity analogous to a pharmacophore. In this seminal work, the authors identified the biaryl moiety and its immediately attached atoms as the chiraphore responsible for enantioinduction.

Although this was intentionally obvious, it demonstrates the development of a tool capable of identifying which structural elements of catalysts are responsible for stereoinduction.

A similar approach was pursued by Lipkowitz, Schefzick, and Avnir in which CCM was employed to identify the structural features of bisoxazoline ligands responsible for stereoinduction in an enantioselective Diels–Alder reaction (Figure 26).[125] Calculation of the CCM of the four ligands reveals a linear relationship between CCM and enantioselectivity (R^2 = 0.98). With this relationship identified, coordinates of ligand distortions are scanned to identify which structural perturbations are most associated with CCM. Three distortions are (1) twist, (2) bite, and (3) pucker. The bite distortion is attributed to deformations associated with backbone identity (hence the nonuniform CCM differentials between catalysts), whereas the twist angle distortion causes the largest change in variance, leading the authors to suggest that design of new ligands should focus on exaggerating this twist motion to maximize enantioselectivity.

A later study examining subunits of the whole structure identified the chiraphore to be the copper, the nitrogen atoms, the substituents on the bisoxazoline core, and parts of the two triflate units ligated to the copper center.[126] This method identifies the important subunits and molecular distortions correlated to stereoinduction for this specific reaction.

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620–1689

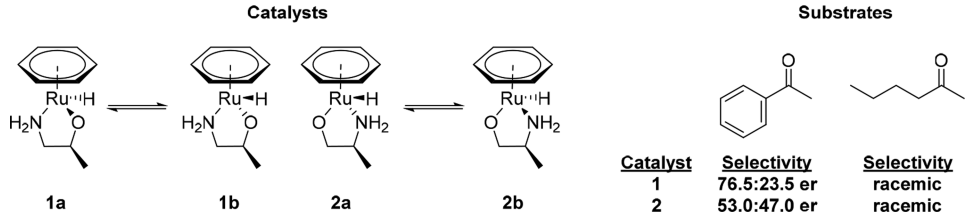


Figure 28. Ruthenium-catalyzed, enantioselective, transfer hydrogenation reaction.

However, application of such an analysis to any other asymmetric reaction has not yet been examined to assess the generality of these findings for other bisoxazoline-catalyzed transformations.[127]

This method has also been employed in an analysis of the Katsuki-Jacobsen epoxidation reaction (Figure 27).[128] The catalyst is a manganese−salen complex,[129,130] the selectivity of which is perplexing given the planar structure of the catalyst.[131] Wiest and Plattner found that the triplet and quintet spin states of the complexes are geometrically distorted with respect to the singlet state,[132] as reflected in the chirality content of the compounds. Under the assumption that these distortions are responsible for the high enantioselectivity, the effects of these distortions on CCM were examined. Four distortions are identified to have a substantial impact on the chirality content of the salen complex: (1) the C(1)−C(2) bond length, (2) puckering (cup−up or cup−down) of the complex, (3) a "step-induced" distortion (minimization or exaggeration of the step-like geometry), and (4) the dihedral angle between the two aromatic planes (Figure 27). The authors conclude that elongating the linker (C(1)−C(2)) increases the chirality content, puckering decreases CCM, step-like distortions increase CCM, and the twist motion increases CCM, with twist and step distortions having the greatest influence, followed by linker distortions and then puckering.

Bellarosa and Zerbetto introduced a modification to the CCM method termed electronic chirality measures (ECM) to asymmetric catalysis wherein the chirality is measured from the electronic wave function.[133] Although the concept of ECM was suggested earlier,[134,135] this first application sought to evaluate the "amount" of electronic chirality in structures and relate it to the enantioselectivity of asymmetric aminohydroxylation reactions.[133] The authors assumed that the chirality content of the products reflected the chirality content of the stereo-determining transition structures and calculated chirality content for six products of varying experimentally observed enantiomeric purity. The ECM had a much stronger correlation with enantiomeric purity than the analogous CCM values, thus validating ECM as a calculable feature capable of quantitatively reflecting the chiral character of molecules.

Continuous chirality measures have also been used to analyze stereodifferentiation at critical points along the reaction coordinate of a ruthenium-catalyzed, enantioselective, transfer hydrogenation reaction.[136] The authors examined two different substrates and four different permutations of catalyst structure. With acetophenone as the substrate, all eight possible reaction coordinates were examined (each catalyst approaching the *Re* and *Si* faces of the ketone), whereas for 2-hexanone, only two coordinates were examined (Figure 28).

In the reactions with acetophenone, catalyst **1b** is calculated to have much higher energy barriers than the other catalysts, and was removed from analysis. Therefore, the remainder of the discussion includes only catalysts **1a**, **2a**, and **2b**. The CCM is calculated at each stationary point along the energy profile for each of these catalysts, starting from isolated starting materials and products, then the precoordination complex of the substrate to the catalyst, followed by the hydrogenation transition structure, and finally the posthydrogenation complex. The quantum chemical calculations alone are sufficient to reveal the greater observed selectivity of **1** compared with **2**. The two competing diastereomeric complexes **2a** and **2b** are in equilibrium and both are catalytically competent. However, these two diastereomers of the active catalyst lead to different stereoisomers of the product. The observed selectivity is the average of the two catalysts and is thus lower compared to the results from catalyst **1**. As previously stated, complex **1** has only one catalytically competent diastereomer (**1a**). Because there is no competition with the other diastereomer of the catalyst, the observed selectivity is higher with respect to complex **2**.

To understand the differences in substrate selectivity, a more thorough analysis is necessary. The CCM values of both acetophenone and 2-hexanone are calculated using the geometry of the starting materials in the stereodetermining transition structure. Comparison of CCM values for acetophenone and 2-hexanone suggests that the acetophenone is forced into a state of higher chirality than 2-hexanone and is thus more amenable to enantiodifferentiation. Therefore, the ideal catalyst for the enantioselective, transfer hydrogenation reaction with 2-hexanone would distort the substrate such that chirality is maximized in the transition structure. By scanning the dihedral angle between the hydrogen being transferred, the carbonyl carbon, and the two subsequent methylene units of the butyl subunit of 2-hexanone, the authors identified a H−$C_{carbonyl}$−C-C dihedral angle between −20° and 30° in which the substrate is most chiral. The authors go on to suggest that the design of a catalyst forcing the substrate into this conformation in the transition structure should enhance the selectivity of the reaction. Unfortunately, this prediction was not validated experimentally.

Denmark and Zahrt have investigated the use of CCM as a single parameter in predicting reaction outcomes of enantioselective reactions.[137] In this work, a data set from a previous study on the enantioselective addition of thiols into aldimines with chiral phosphoric acid catalysts was used in an attempt to construct a linear relationship between CCM and enantioselectivity of the chiral N,S-acetal formed (Figure 29).[138] In contrast to previous studies, this univariate representation of molecules did not correlate to enantioselectivity (Figure 30). Additionally, conformer-dependent CCM parameters were developed, in which the average CCM of a conformer ensemble (with and without Boltzmann weighting) was used as the univariate measure. The standard deviation of the ensemble was also tabulated, considering large variation in CCM could be a measure of molecular flexibility. No

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

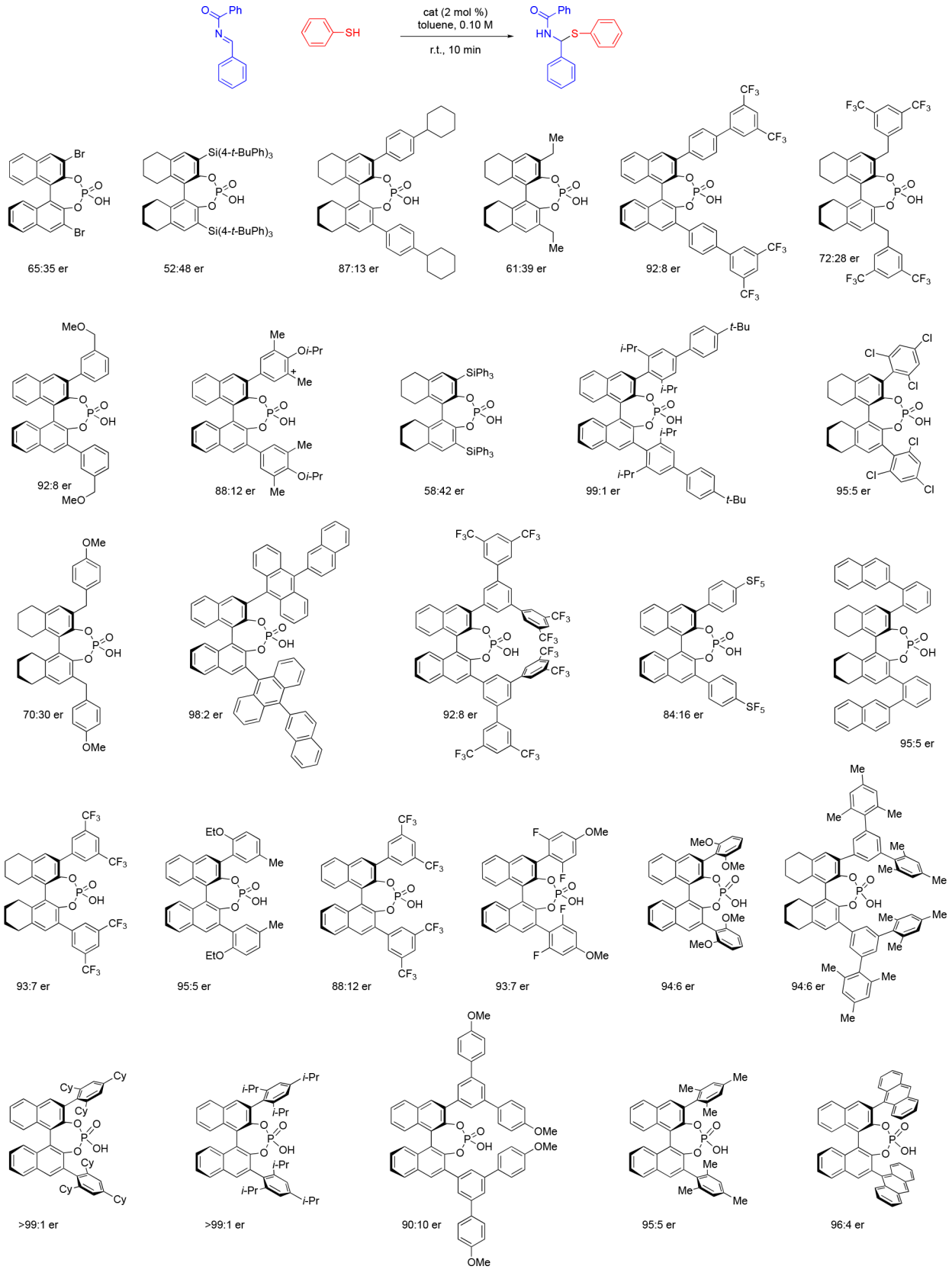


Figure 29. continued

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

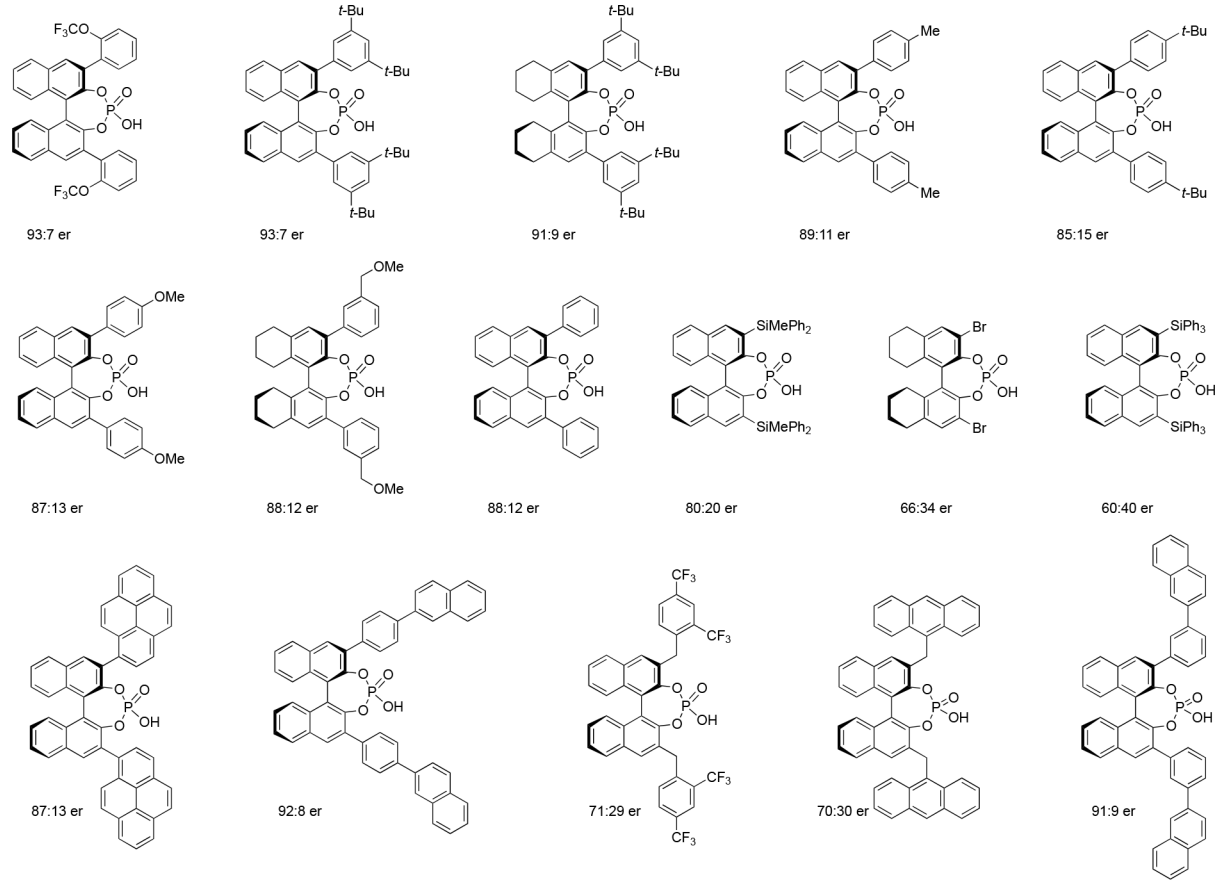


Figure 29. Selectivity of phosphoric acid catalysts in the synthesis of chiral, N,S-acetals.

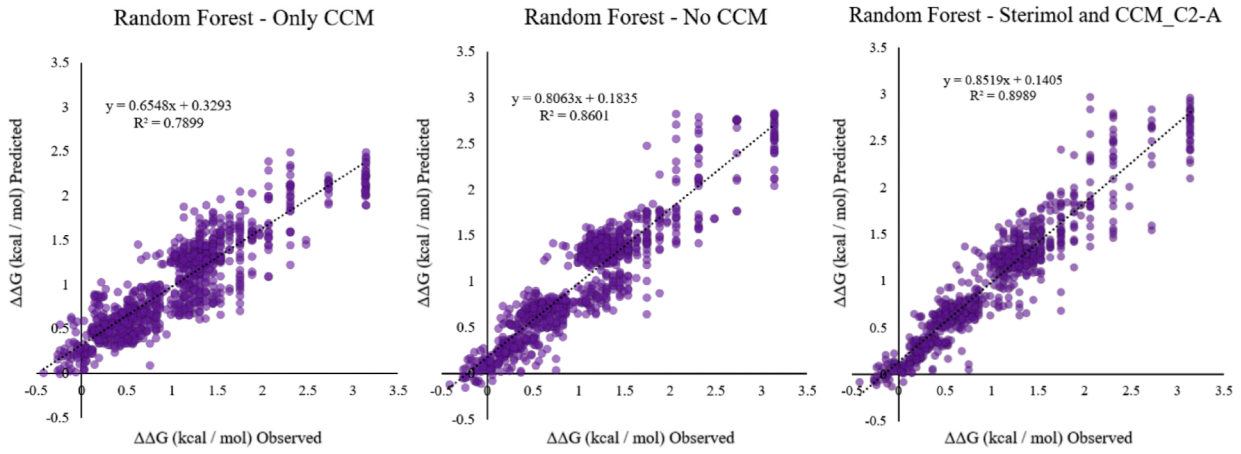


Figure 30. External test sets for models generated with CCM parameters, Sterimol parameters, and both set representing catalysts. Adapted with permission from ref 137. Copyright 2019 Elsevier.

conformer dependent method had a univariate, linear relationship with enantioselectivity. It was postulated that when vastly different molecular subunits critical for enantioselectivity are present in the data set, this simple representation does not contain the requisite information to make accurate predictions of selectivity. However, it was postulated that this measure could be a representation of molecular shape and thus could be used to augment other descriptor sets capable of representing subunit steric and electronic parameters.

Thus, Sterimol parameters are used to represent the key subunits of the 3,3′-positions of the phosphoric acid catalyst and random forest models are constructed to evaluate if the inclusion of CCM parameters result in significant improvement in the predictive performance of the model. Models are constructed using only CCM derived parameters, only Sterimol parameters, and both CCM and Sterimol parameters as catalyst features. The Sterimol parameter model is significantly better than the CCM parameter model (determined by ANOVA with Tukey posthoc test) with MADs of 0.21 and 0.29 kcal/mol, respectively. Including both CCM and Sterimol parameters results in the construction of significantly more accurate models with MADs as low as 0.176

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

kcal/mol. However, it was noted that models with conformer dependent or one-conformer CCM representations are generally not significantly different from one another. This study thus demonstrated three primary points: (1) CCM cannot be used as a univariate method to predict catalyst efficacy in some enantioselective reactions, particularly when large differences between important subunits dictate reaction outcome, (2) it may be possible to use CCM to augment other subunit-based descriptors to improve the predictive performance of models, although immediate mechanistic interpretation of the model will likely not be possible, and (3) CCM can perhaps be treated as a shape index for chiral molecules, but mechanistic interpretations of the significance of this index are not straightforward.

3.3. Perspective on CCM

Although CCM has received limited application in asymmetric catalysis, using CCM seems to have potential in identifying important structural features of catalysts responsible for stereoinduction. However, perhaps the most serious limitation of CCM is that it is not necessarily linearly related to enantioselectivity across the entire range of enantioselectivity values. Thus, studies in which authors make extrapolative predictions with respect to CCM in univariate models must be validated experimentally in each unique case. To date, no experimental validation has been reported. Further, no examples are on record in which the predictions made on the basis of CCM measures resulted in the design of a more highly selective catalyst. However, this is not to say that CCM is not useful in catalyst design, rather, that the applications of CCM in the design of more selective catalysts remain to be demonstrated. Experimental validation of this method is facilitated by the availability of a Web site for the calculation of CSM and CCM.[139] This program therefore could easily be used by others interested in implementing CCM-guided workflows for the development of new, more selective catalysts.

4. CHIRALITY CODES

4.1. Introduction to Chirality Codes

Aires-de-Sousa and Gasteiger have developed chirality codes to represent chiral compounds.[140,141] Depending on the need to consider specific molecular conformations, these representations are termed conformer independent chirality codes (CICC) and conformer dependent chirality codes (CDCC). Both chirality codes are constructed by transforming the 3D molecular structure into a fixed-length vector.[142] A design element in these molecular representations is the ability to account for the absolute configuration of a chiral molecule. In CICC, this feature is achieved by including a chirality signal (represented as S_{ijkl}), which is derived from user-specified properties (e.g., atomic charge) and atomic coordinates of sets of four atoms. The molecular environment is then represented by a term E_{ijkl}, which is derived from user-specified properties of atoms and the distances between those atoms. For the representation to be conformer independent, the distance between two atoms is defined as the summation of the bonds connecting those two atoms rather than interatomic distance. Using these two terms, a function is constructed. This function is scanned at uniform increments to calculate the descriptors; thus, the number of increments dictates the dimensionality of the molecular representation (Figure 31).

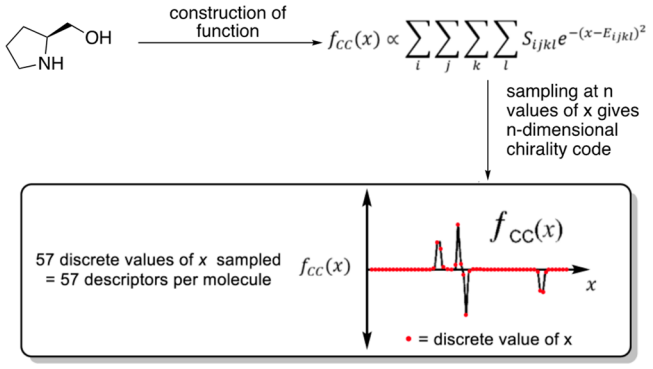


Figure 31. CICC calculations. Reproduced from Aires-de-Sousa, J.; Gasteiger, J. New Description of Molecular Chirality and Its Application to the Prediction of the Preferred Enantiomer in Stereoselective Reactions. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 369−375. Copyright American Chemical Society.

A more detailed description of the calculation of these descriptors is available in the Supporting Information. Further, because only CICC has been applied to asymmetric catalysis CDCC will be discussed as a future direction and is also available in the Supporting Information.

4.2. Application of CICC

Aires-de-Sousa and Gasteiger first reported the use of CICC to predict the absolute configuration of secondary alcohols resulting from the enantioselective addition of diethylzinc to benzaldehyde and for the enantioselective reduction of ketones by (−)-DIP-chloride ((−)-*B*-chlorodiisopinocampheylborane).[140] For each example, literature data are used to train a counterpropagation neural network that is then used to predict the major enantiomer of the transformation. A counterpropagation network is constructed of two parts, a Kohonen layer (the input layer) and an output layer. The Kohonen layer and output layer are linked and thus can be used as a look-up table; a neuron in one layer (e.g., the Kohonen layer) is linked to a corresponding vector in other layer (e.g., the output layer). A more detailed description of a counterpropagation network is available in the Supporting Information.

For the enantioselective addition of diethylzinc to benzaldehyde, CICC is calculated at 75 evenly distributed values for a series of 50 amino alcohol catalysts for which literature data are available for the transformation of interest (Figure 32). Thus, each molecule is represented by a 75-dimensional vector. These vectors are used as input into a counterpropagation neural network, which is trained with 45 catalysts and tested with five catalysts. For the training set, catalysts that give the (+)-enantiomer of the product are given an output value of +1 and catalysts that give the (−)-enantiomer of the product are given an output of −1. For the test set, the absolute configuration of the product is predicted on the basis of the sign of the output value. In each case, the network is able to successfully predict the major stereoisomer from the reaction (Figure 33).

For the enantioselective reduction of ketones by (−)-DIP-chloride, literature data for 50 different ketones are obtained for which the absolute configuration of the corresponding secondary alcohol is available (Figure 34). The input vectors are derived by calculating CICC at 31 evenly spaced points for each enantiomer of the product alcohol. If the configuration of the input alcohol structure corresponds to the major isomer

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

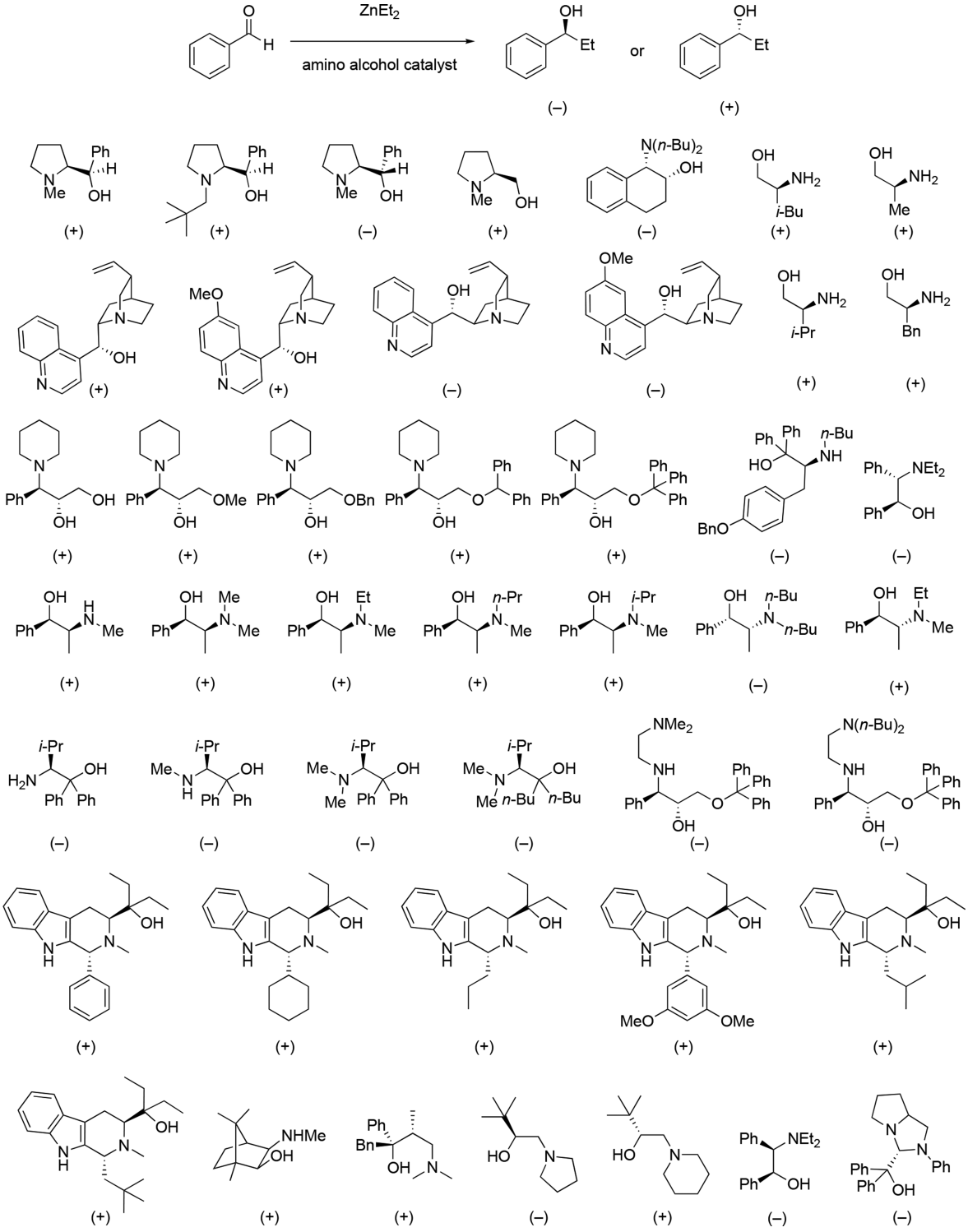


Figure 32. Enantioselective addition of diethylzinc to benzaldehyde and 45 training catalysts. The major isomer formed when the catalyst is used in the reaction is designated by the (+) or (−) under the catalyst structure.

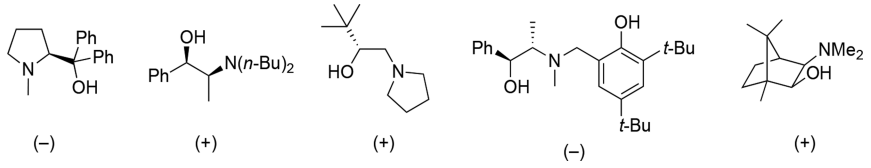


Figure 33. Test catalysts for the enantioselective addition of diethylzinc to benzaldehyde. The major isomer formed when the catalyst is used in the reaction is designated by the (+) or (−) under the catalyst structure.

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

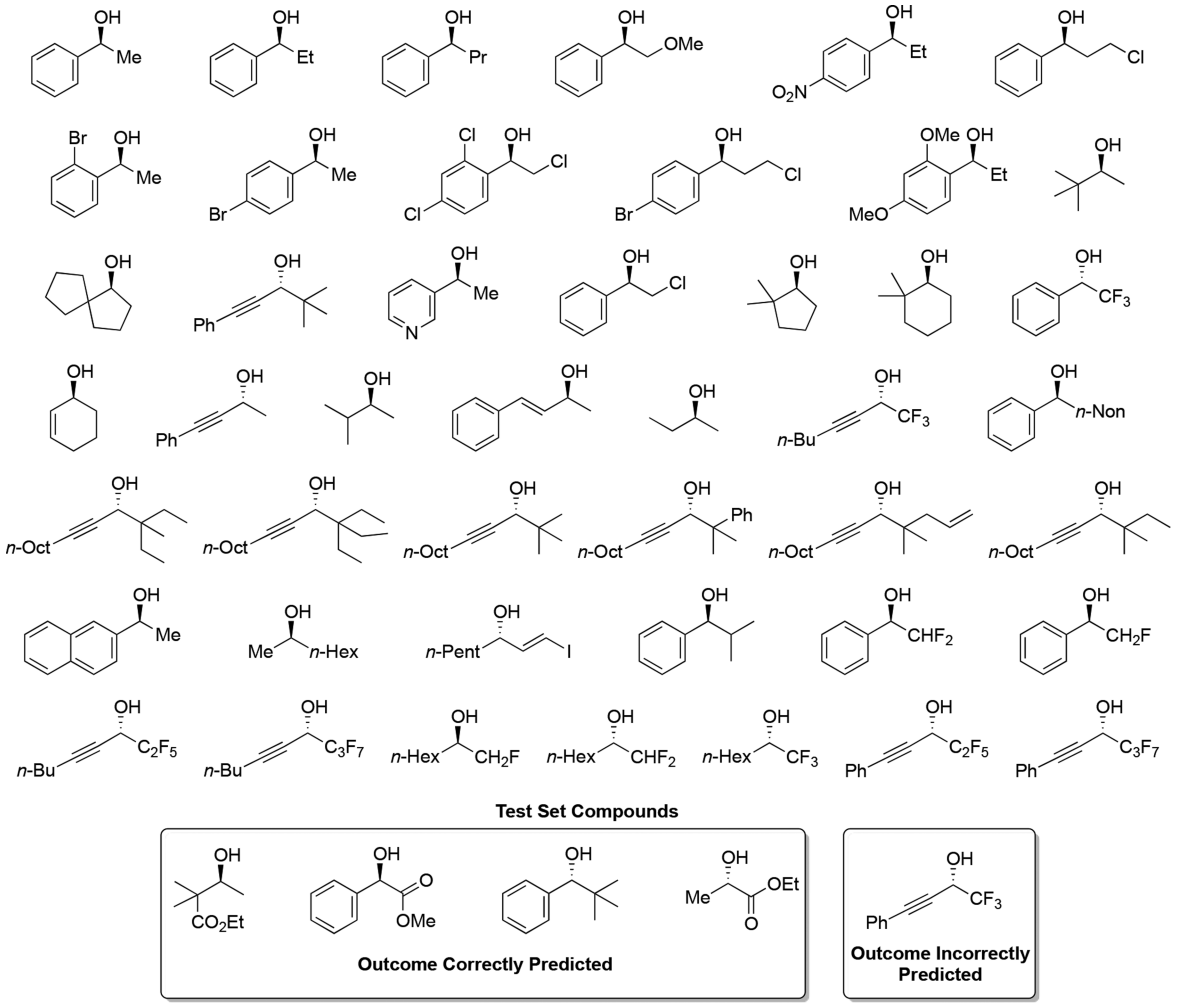


Figure 34. Secondary alcohols synthesized by the reduction of the corresponding ketone with (−)-DIP-chloride. The major isomer of the product is depicted in each case.

formed, the output value i set to +1. If the opposite configuration of the alcohol (with respect to the input structure) is formed preferentially, the output value is set to −1. This training protocol is used for 45 alcohol pairs (the two possible enantiomers for each parent ketone). The trained model is evaluated with the five remaining alcohols pairs and four of the five test cases are predicted correctly. The incorrect prediction is attributed to the configurational variability obtained from the reduction of fluorinated substrates by DIP-chloride.

This seminal publication on the utility of CICC to predict the absolute configuration of products represents a unique application of chemoinformatics in asymmetric catalysis. Typically, selectivity is predicted as a continuous variable wherein experimentalists search for a maximum. This example highlights the ability of chirality codes to store information related to the absolute configuration of a catalyst in a way that is conformation independent and alignment independent, a unique capability absent from other descriptor classes.[143,144]

Aires-de-Sousa and Gasteiger later applied CICC in a regression analysis to predict enantioselectivity in the addition of diethylzinc to benzaldehyde, wherein enantiomeric excess (% ee) is the regressand.[145] In this work, the authors employ a number of modeling methods including feed-forward neural networks, perceptrons (neural networks with no hidden layers), multilinear regression, and support vector machines to predict the continuous selectivity output. In the reaction system developed in a previous study,[146] five different racemic amino alcohol ligands are used simultaneously with 13 different enantioenriched amino alcohol additives (Figure 35). This set, containing 65 experimental data points, is used to evaluate if CICC could predict a continuous output. The enantioenriched chiral additives are represented by 101-length chirality codes, whereas the racemic catalysts are represented by the absolute value of the 101-length chirality code of a single enantiomer. Prior to modeling, all low variance features are eliminated, leaving only 28 parameters per reaction. Using these descriptors, a neural network is used to identify the relative weights of each individual feature, allowing the number of variables to be reduced further to 11 features per reaction. Following this, the different modeling methods are evaluated, with feed-forward neural networks providing the best performance with a combined three-fold cross-validated R^2 of 0.923 and an RMSE of 6.9% ee. Notably, this example uses % ee rather than e.r. or free energy differential as the regressand. Thus, in this case and all other cases in which this is true, it is possible to calculate predicted values of % ee that are over 100%. Because this value exceeds the theoretical maximum, these are simply interpreted as very selective predictions. Alternatively, generating models with other regressands such as

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

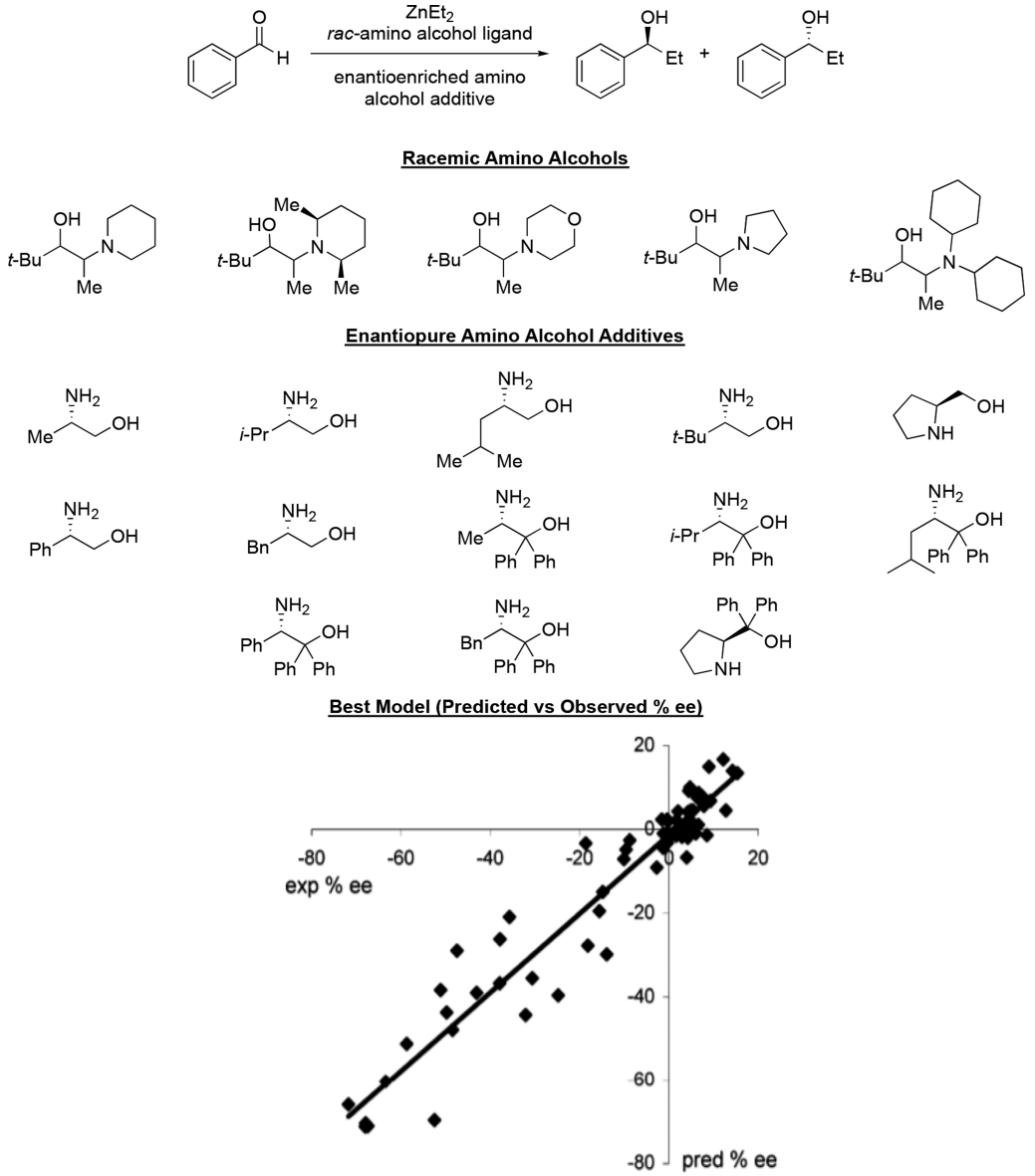


Figure 35. Enantioselective diethylzinc alkylation of benzaldehyde with the predicted versus observed plot. Reproduced from Aires-de-Sousa, J.; Gasteiger, J. Prediction of Enantiomeric Excess in a Combinatorial Library of Catalytic Enantioselective Reactions. *J. Comb. Chem.* **2005**, 7, 298−301. Copyright American Chemical Society.

e.r. or free energy differential removes the possibility of producing physically meaningless values. This work thus demonstrates the capacity of CICC to predict a continuous output using data from combinatorial experimentation. The approach is particularly appealing given the conformer and alignment independence of CICC.

Another example of the application of CICC to asymmetric catalysis is enantioselective transfer hydrogenation.[147] In this work, a published data set[148] is used to determine if CICC with counterpropagation neural networks could be used to identify an optimal catalyst by experimentally testing only a small number of catalysts. In the original work by Bellefon and co-workers,[148] selectivity and conversion data for a combinatorial library of 1914 catalysts are experimentally measured in the enantioselective transfer hydrogenation of acetophenone catalyzed by chiral, metal-amino alcohol complexes (Figure 36). A genetic algorithm is then used to identify selective catalysts on the basis of a normalized performance factor (NPF), calculated by multiplying the conversion by two then adding that product to the enantiomeric excess. The value is then normalized to the catalyst with the highest NPF; thus, each catalyst has an NPF value between 0 and 1. By using a genetic algorithm to guide catalyst selection, Bellefon and co-workers are able to identify at least five of the top ten catalysts, on average, by only testing 10% of the total library.

Subsequently, Xu and co-workers calculated CICC for the complexes in the following manner: (1) the CICC for the amino alcohol portion are encoded with a 51-dimensional vector (corresponding to CICC with 51 different increments), (2) the CICC corresponding to the N-protecting group, (B) are calculated using 63 increments, yielding a 63-dimensional vector, (3) the metal complexes are encoded by a binary, 6-dimensional indicator vector in which each dimension corresponds to the presence of one metal precatalyst (thus, for each complex, five dimensions zero and one dimension is 1), and (4) these descriptors are then concatenated

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

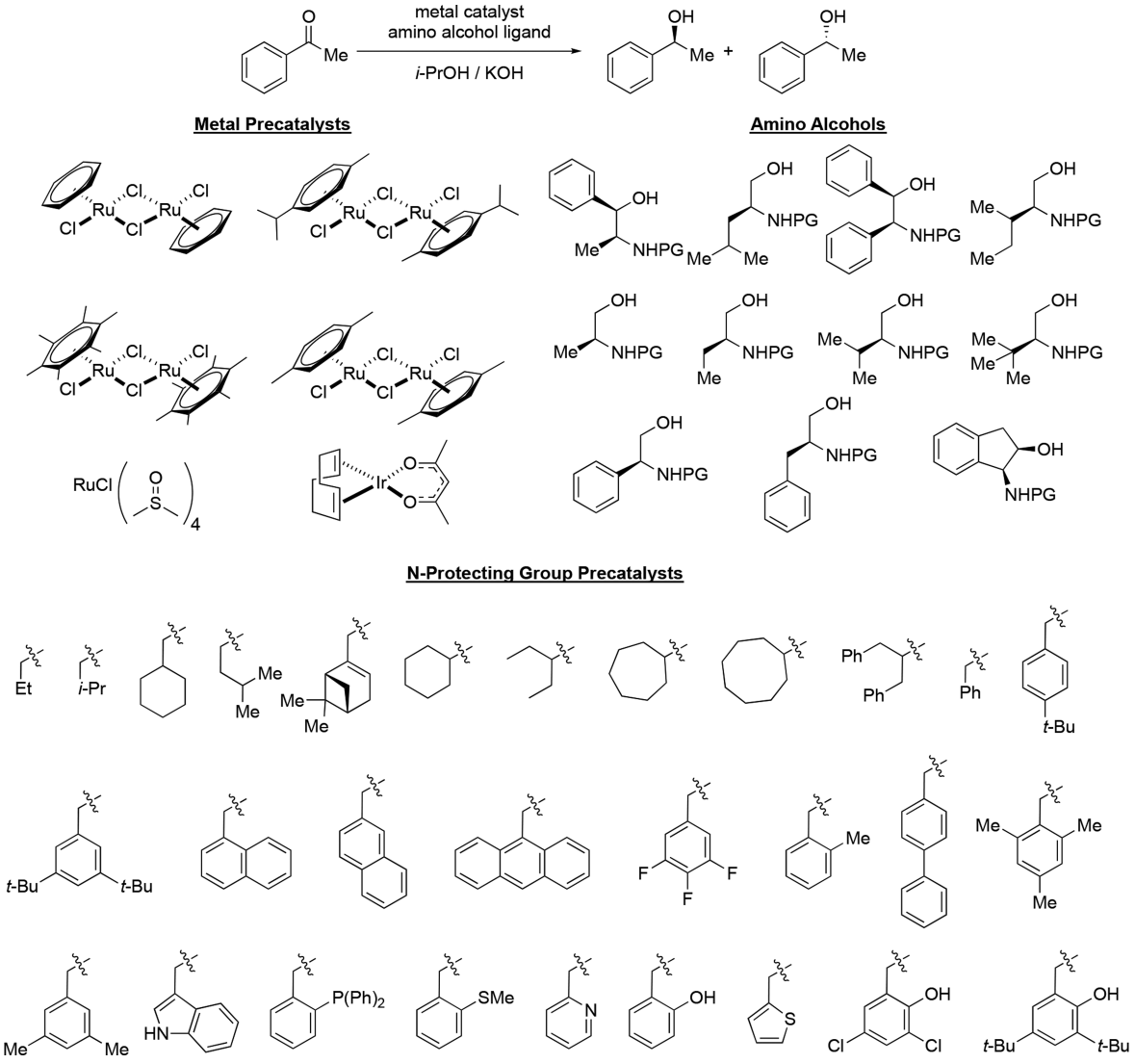


Figure 36. Enantioselective transfer hydrogenation for the reduction of acetophenone.

combinatorially, yielding 120-length vectors for each member of the combinatorial library. Dimensions in which every catalyst had a value of zero are removed, reducing the dimensionality of the final vector to 108. These vectors can then be used as input to a counterpropagation network, in which the weight of the output layer associated with the winning node moves toward the NPF associated with the input catalyst vector.

The authors select a training set of 198 compounds semirandomly, with the conditions that each metal complex appears 33 times, each amino alcohol portion appears 18 times, and each N-protecting group appears either 6 or 7 times to ensure an even representation of the different possible catalyst permutation. The remaining 1716 catalysts are then used as a test set. Performance is evaluated by calculating a hit number (N), defined as the percentage of the top 10 catalysts (termed target catalysts) in the combinatorial library that would have been uncovered by the simulated optimization. For example, if 198 catalysts are used to train a network, the network could then rank the remaining 1716 catalysts. The top 50 are then "selected", simulating the next set of catalysts that would be synthesized and tested in a real optimization campaign. Inclusion of one of the top ten catalysts in the data set, as according to NPF, is defined as a successful end to the simulated optimization and the number of top ten catalysts identified in the 50 selected catalysts is thus a metric of success. As an example, inclusion of one of the top 10 catalysts would give N = 10% for 246 reactions (198 training and 50 top predicted). This process is repeated multiple times with different selections of training data to remove error associated with random selection of training data.

On average, random selection of 198 (10% of the combinatorial library) gives a success rate of N = 9%, far below the success rate of N = 50% observed in the original report employing a genetic algorithm. Using a counterpropagation neural network, success rate is increased to N = 78.5% on average by "screening" the top 200 predicted catalysts from the test set (totaling 20.8% of the entire combinatorial library). The dimensionality of the input vectors are reduced by a genetic algorithm, and these new descriptors are used with 20.8% of the combinatorial library to achieve a success rate of N = 85.5%. The number of catalysts surveyed could be reduced to 13% of the total library (198 training catalysts and top 50 test catalysts) to achieve a success rate of 55%, a similar level of success when compared with the original study. Because the training data is 10% of the combinatorial

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

library, a direct comparison with this method and the original method is not possible. However, this report does represent an alternative method by which new enantioselective methods could be optimized with significantly greater success over random sampling alone.

In a continuation of this work, Xu and co-workers use a subset of this data set in regression modeling to predict the performance of an external test set of catalysts.[149] Only reactions in which the yield is over 5% and some enantioinduction is observed are considered in the study, limiting the data set size to 296 catalysts. Both a regression tree and a random forest regressor are trained on 237 members of the data set and used to predict the remaining 59 catalysts, which are used as a test set. For the regression tree, the predictivity is low with R^2 = 0.71 and 0.56 for training and test sets, respectively. Random forest models performed significantly better, with R^2 = 0.71 and 0.77 for training and test sets, respectively. Using a genetic algorithm to reduce the dimensionality of the input vectors to 28 further increased the predictivity of the resulting model, with R^2 = 0.77 and 0.82 for training and test sets, respectively, with RMSE = 9.96% ee. Predictivity observed with this method is limited, but it demonstrates the capability to use machine learning methods with a vector representation of a molecule that (1) is not fragment based, (2) is alignment independent, (3) is conformer independent, and (4) can account for the absolute configuration of the catalyst.

4.3. Other Chirality Codes

Zhang and co-workers have also developed variants of chirality codes for the prediction of the major isomer of enantioselective reduction of ketones to form secondary alcohols.[150] The authors propose a physicochemical atomic stereodescriptor derived from numerous topological properties that are taken from the groups attached at the stereogenic carbon of the secondary alcohol. In particular, the two substituents other than the −H and −OH groups are described arbitrarily as "right" and "left" groups, and are used to generate the individual codes. The vector representation of the molecule is constructed from the concatenation of right and left groups, which in turn comprises: (1) the number of atoms in the group, (2) the number of atoms three bonds away from the stereocenter, (3) the distance (in number of bonds) the farthest atom is away from the stereocenter, (4) the maximum distance (in number of bonds) between two atoms in the group, (5) atomic charge, (6) sum of atomic charges, (7) atomic polarizability, (8) electronegativity, (9) charge density, (10) total charge density, and (11) a steric hindrance parameter. To account for chirality, chiral connectivity indices[151] and chiral topological charge indices[152,153] are employed.

Data sets used to benchmark this new chirality code are the enantioselective reduction of ketones by (−)-DIP chloride[140] and the enzymatic resolution of racemic alcohols with a lipase.[154] In case of the former, the 100 possible stereoisomers derived from the 50 parent ketones are divided into two groups: Group A contains the experimentally observed, major stereoisomer of the alcohol whereas Group B consists of the minor stereoisomer. A random forest classifier is used to predict which would be the major isomer of the reduction, which could be improved by feature reduction with a genetic algorithm. After being trained on 40 alcohol pairs, the classifier achieves a 90% success rate classifying the remaining 10 pairs of alcohols. The same workflow is implemented for the enzymatic resolution of alcohols with a lipase, which achieved an 87% success rate classifying a test set of 15 pairs after being trained on 52 pairs.

Zhang and co-workers later developed a simpler permutation of chirality codes designed to facilitate the generation of an empirical rule governing the preferred enantiomer of the product formed in the same reactions discussed above.[155] Using the same data sets, codes for chiral secondary alcohols are generated as 12-dimensional bits derived from the carbons attached to the stereogenic carbon of the secondary alcohol (6-bits for each substituent). The bits are binary indicators categorizing the carbon atom into one of six groups: (1) sp carbon, (2) sp^2 carbon, (3) sp^3 carbon with four degrees of branching (sp^3D4, i.e., *tert*-butyl group), (4) sp^3 carbon with three degrees of branching (sp^3D3, i.e., *iso*-propyl group), (5) sp^3 carbon with two degrees of branching (sp^3D2, i.e., an ethyl group), and (6) sp^3 carbon with one degree of branching (sp^3D1, i.e., a methyl group). These vectors are used as input for Fischer linear discriminant analysis, a pattern recognition supervised method, to build a classification model. Using 40 alcohol pairs as training data, the model is able to categorize the 10 test alcohol pairs with 100% accuracy, marking the best performance on the data set to date. The authors use the model to inform a ranked list of significance, as follows: sp^3D4 > sp^2 > sp^3D3 > sp^3D1 > sp^3D2 > sp. Using this list, the authors suggest what they term the "PT rule", consisting of two situations: (1) if the category for the "left" substituent is greater for than the "right" substituent ("left" and "right" defined with respect to the positions of the alcohol residue and hydrogen atom on the stereogenic carbon atom, depicted in Figure 37), the isomer of the alcohol being analyzed is predicted to be the major product, and (2) if the values of "left" and "right" substituents are identical, the relative sizes of those substituents dictate the enantiomer of the product formed wherein a larger "left" substituent indicates the isomer of the alcohol being analyzed is predicted to be the major enantiomer.

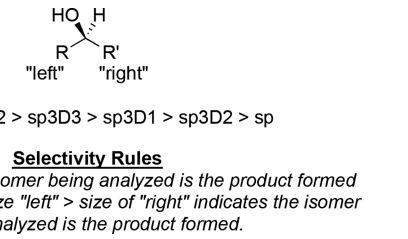


Figure 37. Designation of "left" and "right" substituents used in ref 155.

4.4. Conclusion and Perspective

The primary limitation of CICC is that it necessitates the definition of neighborhoods, which in turn necessitate the presence of a tetrahedral, stereogenic center in the molecules of interest. Thus, certain classes of chiral compounds (i.e., atropisomers) cannot be described by the above representation. To address this limitation, CDCC was developed. CDCC differs from CICC in that a chirotopic atom is not explicitly considered; rather all atoms are considered. Further, the interatomic distances in the form of through-space Cartesian distances are used, rather than the summation of bond length separating the atoms. As a result, the chirality code becomes

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

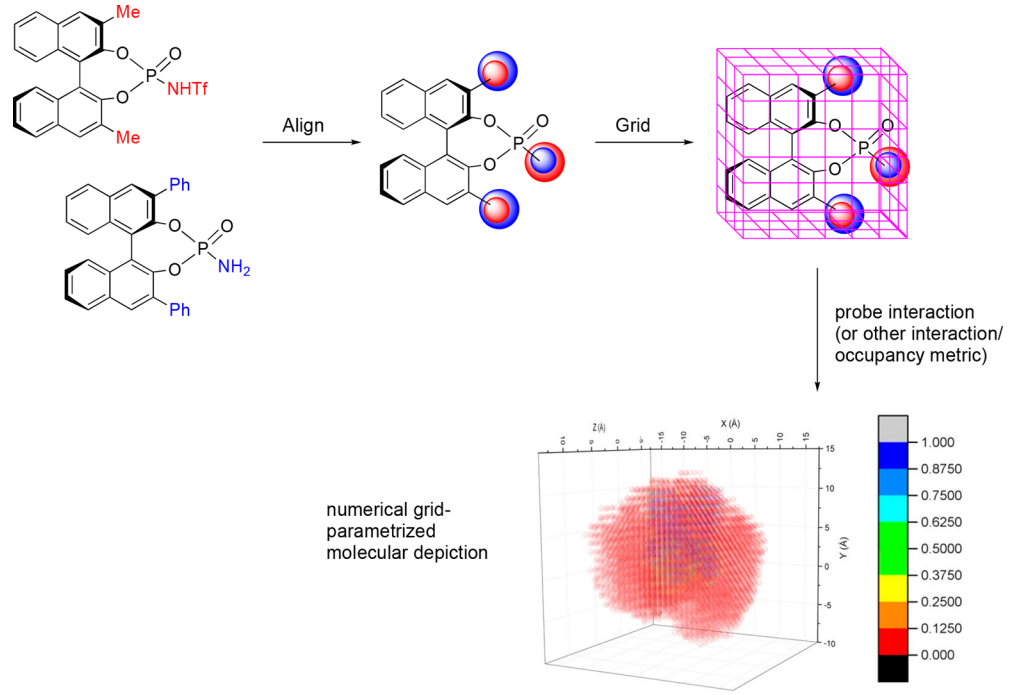


Figure 38. Alignment-dependent MIF workflow to represent a molecule with a grid based descriptor.

dependent on molecular conformation. A more detailed description of CDCC is available in the Supporting Information.

The application of CDCC to asymmetric catalysis would expand the scope of compounds that can be described using chirality codes. However, no application of this representation in asymmetric catalysis has been demonstrated. Given the initial success of CICC, it is surprising that CICC and CDCC have not attracted more widespread implementation in enantioselective catalysis. A possible explanation for this is the inaccessibility of these descriptors; to the best of our knowledge, no open source implementations of this representation exist. We anticipate that an open source version (i.e., a downloadable module on GitHub) would facilitate more widespread adoption of this molecular representation. Another limitation of CDCC is its interpretability. This representation is not intuitive—thus, it is difficult to conceive of a way to garner insight regarding the origin of stereoinduction with CDCC. Other chirality codes may be more interpretable but are not readily applicable to a wide array of chemical systems.

5. MOLECULAR INTERACTION FIELD (MIF) BASED METHODS

Among strategies to provide a "global" description of catalyst structures, comparative molecular field analysis (CoMFA)[156] has emerged as a popular 3D-QSAR method. CoMFA aims to enable a common description of structures in a catalyst library regardless of specific substitution patterns or ideally, even gross structural changes. Such a description should implicitly account for varying steric and electronic effects in the library members. Although other molecular interaction field (MIF)-based descriptors are known, most conceptually resemble CoMFA; thus, an overview of CoMFA will be given here to provide a reference for other methods. This section will be divided into two subsections: alignment dependent methods and alignment independent methods.

5.1. Alignment Dependent MIF Methods

5.1.1. Background to Alignment Dependent MIF Methods. The comparison of two library members is naturally simplified if a common reference frame is employed. For example, in a library of BINOL derived phosphoric acids, the common BINOL core can be used as a fixed reference. Intuitively, it is expected that two members differing only in substitution pattern around the core will share many of the descriptors for common structural regions. For this to occur in practice, it is necessary to align the structures and attain a common reference frame (in this case, the BINOL core). A general description of such alignment dependent protocols is depicted in Figure 38. First, the molecules of interest are aligned to allow comparison of analogous regions of space around the core structures. Next, molecules are placed into common grids with defined grid spacing. Probes are placed at each grid point to calculate the steric or electronic interaction between the probe and the molecule at a specific point in space, thus achieving a "global", grid-based molecular description. These interaction energies are then used as descriptors to make a mathematical model relating the calculated properties to an outcome of interest, and this model is validated either by internal or external validation (or both). On the basis of a validated model, it is then possible to either identify important catalyst properties for enantioinduction or to predict the activity of catalysts that have not yet been synthesized. For the steric interaction energies, traditionally Lennard-Jones potentials[157] are employed with some reference atom at each grid point. The electronic MIF is typically constructed from Coulombic interaction of each structure with a charged particle at each grid point. Model validation can be performed with either cross-validation methods (internal), wherein a set number of entries from the training set (the set of compounds originally used to make the model) is excluded

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

from model generation iteratively until all entries have been excluded once, or by attempting to predict the observed properties of a test set not used in model generation (external). However, best practices recommend using both internal and external validation for evaluating models.[36]

5.1.2. Applications of Alignment Dependent MIF-Based Methods in Asymmetric Catalysis. The first example of the application of CoMFA to asymmetric catalysis was reported by Lipkowitz and co-workers in 2003.[158] The aims of this study were to demonstrate that "out-of-the-box" (meaning readily available from commercial software packages without extensive optimization) CoMFA could be used to generate models capable of predicting catalyst selectivity and to identify which catalyst features were important for enantioinduction in an enantioselective Diels–Alder reaction (Figure 39). The authors calculate descriptors using the workflow described above for 23 different catalysts whose selectivity values are readily available from the literature, and span a wide range (55:45 e.r. to 99.05:0.5 e.r.).[159−163] The common oxazoline core is used for reference alignment. The authors perform two different validation protocols: partial least-squares (PLS) modeling with internal validation (leave-one-out (LOO) cross-validation) and external validation (modeling with 18 catalysts, then using this model to predict the selectivity of the remaining five). Internal validation methods yield high Q^2 values ranging from 0.533 to 0.840 (Q^2 values greater than 0.5 are typically considered to be acceptable)[158] depending on factors such as grid spacing, field type, probe type, dielectric function, and number of latent variables used in PLS modeling. Using external validation (18-member training set), the authors are able to predict the external catalysts with exceptional accuracy (R^2 = 0.94 in the predicted vs observed plot with slope near unity and y-intercept = 7.5), thus demonstrating the ability to use readily available software to make validated QSAR models for chiral catalysts. The authors then demonstrate the ability to obtain information pertaining to which structural features of catalysts are responsible for enantioinduction. Specifically, the aim is to quantify the relative importance of steric and electronic effects in determining the reaction outcome. Using the best PLS models obtained, it is found that 60−70% of the variance in the data is described by steric effects, whereas 30−40% is described by electronic effects, suggesting that steric properties of the catalyst are relatively more important for enantiodifferentiation. From examination of the steric MIF, two important regions of space are identified where steric occupancy either enhances selectivity or is detrimental to selectivity (Figure 39). The region where increased steric bulk enhances enantioselectivity is green, whereas the region which must be devoid of occupancy for high-selectivity is yellow. The complex depicted in Figure 39 is the most selective catalyst in the study that fits these guidelines, supporting the hypothesis that CoMFA can be used to obtain useful structural information about how catalyst structure relates to selectivity.

Contemporaneous with Lipkowitz's report, Kozlowski and co-workers employed an MIF-based method to predict the selectivity of β-amino-alcohol-catalyzed alkylation of aldehydes with organozinc reagents.[164] Rather than using classically calculated MIFs, semiempirical methods (PM3) are used to calculate approximate transition structures for the reaction. These structures are aligned and used to calculate an electronic MIF again at the PM3 level of theory. A quantitative structure-selectivity relationship (QSSR) is then calculated using

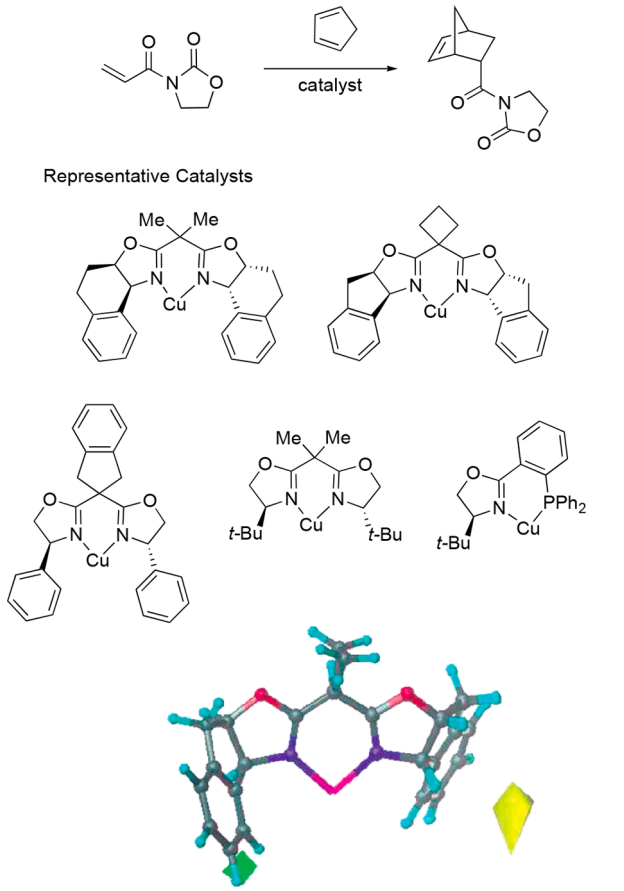


Figure 39. CoMFA for enantioselective, Diels–Alder reaction. Most selective catalyst in ref 158 with areas where high occupancy corresponds to selectivity (green) and where low occupancy corresponds to selectivity (yellow). Reproduced from Lipkowitz, K.; Pradhan, M. Computational Studies of Chiral Catalysts: A Comparative Molecular Field Analysis of an Asymmetric Diels–Alder Reaction with Catalysts Containing Bisoxazoline or Phosphinooxazoline Ligands. *J. Org. Chem.* **2003**, 68, 4648−4656. Copyright 2003 American Chemical Society.

different combinations of only two grid points at a time, resulting in the generation of a "best" two-variable model and an "average" model constructed by weighting all accepted two-variable models. The model is validated externally (R^2 = 0.90 for averaged model), demonstrating the ability to use semiempirical calculations to construct MIFs capable of generating models that predict enantioselectivity to comparable accuracy[165] as other high-level (e.g., DFT) methods (root–mean–square error (RMSE) = 0.29 kcal/mol).

Kozlowski and co-workers later used a similar method to evaluate the importance of the A-ring of sparteine for enantioselectivity in enantioselective lithiation reactions.[166] Prior work had demonstrated that the D-ring of sparteine had little impact on the enantioselectivity of lithiation.[167] However, the A-ring was important in enantioinduction because omitting the A-ring resulted in a reduction in selectivity from 98.5:1.5 er to 60.5:39.5 er (Figure 40).[168]

The researchers sought to determine if the entire A-ring is necessary for enantioinduction and thus prepared an analog devoid of the A-ring leaving only the *N*-methyl carbon and the carbon affixed to the corresponding stereogenic center on the B-ring. Interestingly, this analog gives low, opposite selectivity. To explain this unexpected reaction outcome, a QSSR model is

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

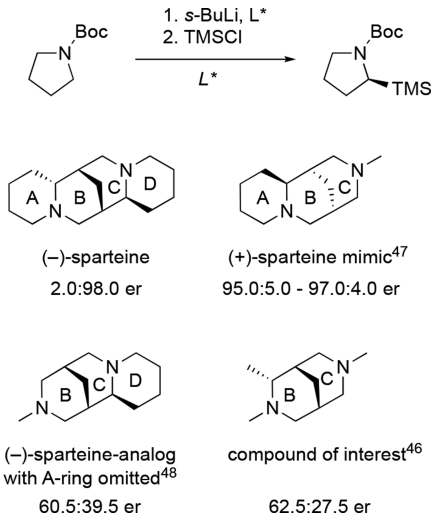


Figure 40. Sparteine and sparteine analogs employed as ligands in enantioselective lithiation of *N*-Boc-pyrrolidine.

generated using 16 chiral diamine ligands, employing a similar method to their original work (*vide supra*). However, PM3 electrostatic potential MIFs did not give satisfactory results correlating the calculated descriptors to activity. Thus, the authors developed G-QSAR in which the electrostatic potential energy (ESP) MIFs are derived from higher-level methods (DFT, HF, MP2, etc.) using Gaussian 98. By employing more accurate ESP calculations (BLYP/6-31G**), predictive models are generated again using two-point linear regression models (Q^2 = 0.68, R^2 = 0.82). Probe energies at defined regions of space with respect to the chelated lithium ion correlate well with observed selectivity.

Figure 41 (sparteine) depicts one example of many possible chelating diamine ligands, in which the red and blue spheres represent analogous regions of space in the common core scaffold. As the interaction energy at the center of the red sphere decreases (owing to either fewer steric interactions or the presence of a relatively electronegative group), the enantioselectivity increases. Conversely, high interaction energies at the center of the blue sphere are found to be associated with high selectivity. From these observations, the authors suggest three guidelines: (1) larger groups below the ring (e.g., near the blue region) result in high selectivity, (2) aromatic groups above the ring are associated with higher selectivity, and (3) large alkyl groups above the ring give rise to low selectivity. The authors are also able to accurately predict the outcome of new catalysts using their calculated descriptors.

In a study of the enantioselective addition of organozinc reagents to aldehydes catalyzed by amino alcohols, Kozlowski and co-workers used a similar, modified workflow for the accurate prediction of reaction outcome for novel catalysts (Figure 42).[169] Rather than use calculated transition structures as inputs as in the original study,[164] a zinc dimer representing the ground state is used for descriptor calculations to make the method more agnostic to mechanism. GQSSR and QMQSAR (MIF calculated with semiempirical methods, i.e., PM3) methods are used for calculating the electronic MIF for 18 training compounds, and *k*-fold cross-validation (k = 2) are used in model construction (Q^2 = 0.85), in which the models are relatively unaffected by training set compounds (independence on training set selection is indicative of robust models). The model is then used to predict the selectivity of 17 compounds for which experimental data was unavailable, 13 of which were then synthesized and evaluated. The model accurately predicts the enantioselectivity of new catalysts prior to their synthesis, an important "first" in this field. However, it is worth noting that employing external validation sets (a test set, demonstrated previously)[158,164,166] is identical to predicting the selectivity of novel compounds, then synthesizing them and collecting data; data collection before or after model generation are irrelevant because that data are not used in model generation in either situation (both are external sets, only the order of the workflow is different).

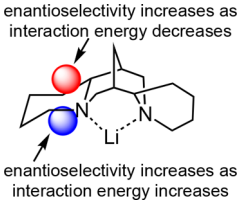


Figure 41. Regions with grid points (red and blue) correlated with catalyst enantioselectivity.

This method was again used by Kozlowski, Hsung, and co-workers to evaluate the selectivity of a new class of chiral amino alcohols ligands for the addition of diethylzinc to aryl aldehydes (Figure 43).[170] 3D-QSAR models with semi-empirical (QMQSAR) and DFT methods (GQSAR) were employed. With a training set of 17 compounds and the same computational workflow described above, accurate models could be generated. These models are then used to predict the performance of catalysts not yet synthesized at a range of temperatures by constructing different models using training data collected at each of the three temperatures.

Kozlowski and co-workers have also used quantum electronic MIFs to predict the enantioselectivity of a single, common catalyst with different substrates in the enantioselective addition of diethylzinc to aldehydes.[171] With the same QMQSAR method as previously described,[169,170] two different sets of aldehydes are used to generate models predicting the reaction outcome. Models are generated for two data sets, the first with 11 aryl aldehydes and the second with eight aryl aldehydes, furfural, seven α,β-unsaturated aldehydes, and two aliphatic aldehydes (Figure 44).

Both data sets generate cross-validated models (Q^2 = 0.67 for the first, Q^2 = 0.61 for both models in the second); however, in the second data set, the aryl aldehydes and nonaryl aldehydes must be treated separately to generate robust models. Although no external validation is done in this study, it does represent an important proof of concept that QSSR models could be used to predict reaction outcome for novel substrates in established systems.

Lower-level methods (using classical rather than semi-empirical or *ab initio* methods), such as "traditional" CoMFA, have also been applied to asymmetric catalysis. Hirst and co-workers used CoMFA to predict the outcome of enantioselective, phase-transfer-catalyzed reactions.[172] This method combines high-throughput screening with computational catalyst evaluation wherein 88 cinchonidinium catalysts are combinatorially synthesized and evaluated in asymmetric alkylation reactions (Figure 45).

To rapidly generate data, catalysts are synthesized *in situ* through sequential N- and O-alkylation of dihydrocinchoni-

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620–1689

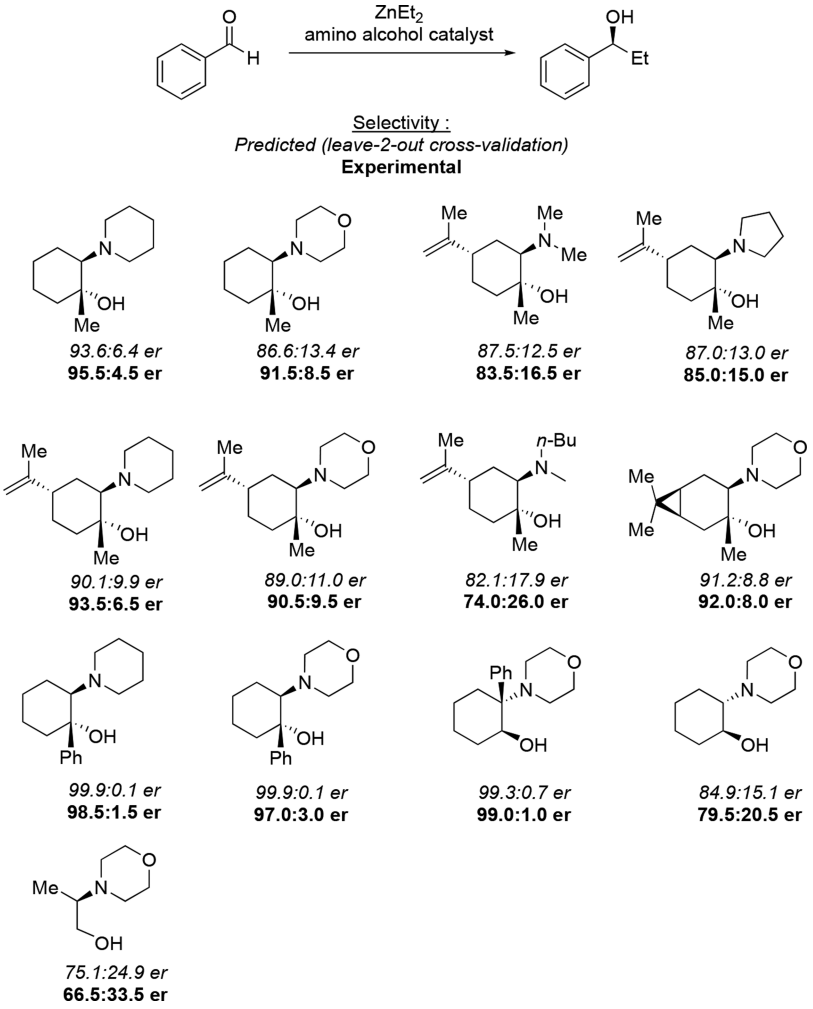


Figure 42. Sparteine and sparteine analogs employed as ligands in enantioselective lithiation of *N*-Boc-pyrrolidine.

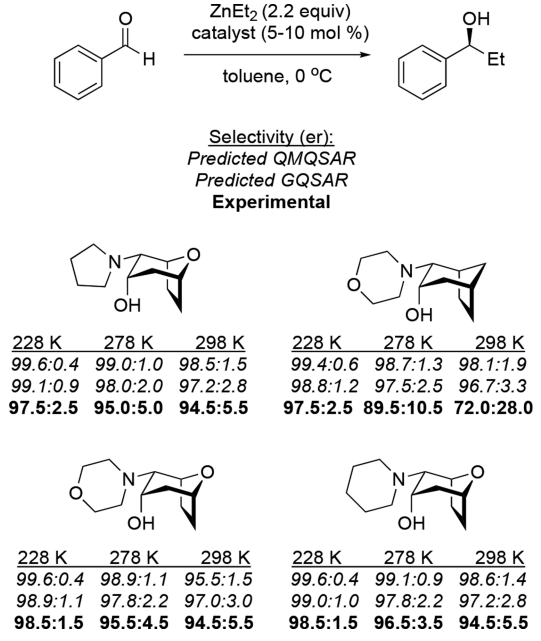


Figure 43. Temperature dependent selectivity predictions in the enantioselective diethylzinc addition to benzaldehyde.

dine. Because catalyst permutations are on a common core scaffold, the authors model only the substituents rather than the entire catalyst scaffolds in an attempt to decrease the amount of "noise" in the data resulting from low-variance points near common parts of the core and to speed up the total calculation time by reducing the number of features. The authors train the reaction on 70 of the catalysts (LOO Q^2 = 0.72), withholding all catalysts with 4-iodobenzyl substituents as an external test set. By using this strategy, accurate 3D-QSAR models could be generated (RMSE = 0.13 on external training set) with only substituent descriptors (predicted vs observed plot in Figure 45). Most importantly, the report demonstrates a catalyst design approach using rapid catalyst generation combined with chemoinformatic analysis.

Similar substituent-based analyses have been performed with a different class of biaryl-derived phase transfer catalysts.[173] In this study, Hirst and co-workers use selectivity data measured from a combinatorial library of 40 catalysts in an enantioselective alkylation reaction (Figure 46).

Two catalyst scaffolds are identified, denoted as **i** and **ii**. The (+) and (−) forms of the biaryl are assumed to be in equilibrium in solution and have been designated only for future discussion—the biaryl precursor used in catalyst synthesis is achiral. For scaffold **i**, substituent combinations from **S2Sa** through **S8Se** are synthesized (wherein the **Sn** refers to the R′ substituent and **Sx** refers to the corresponding

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

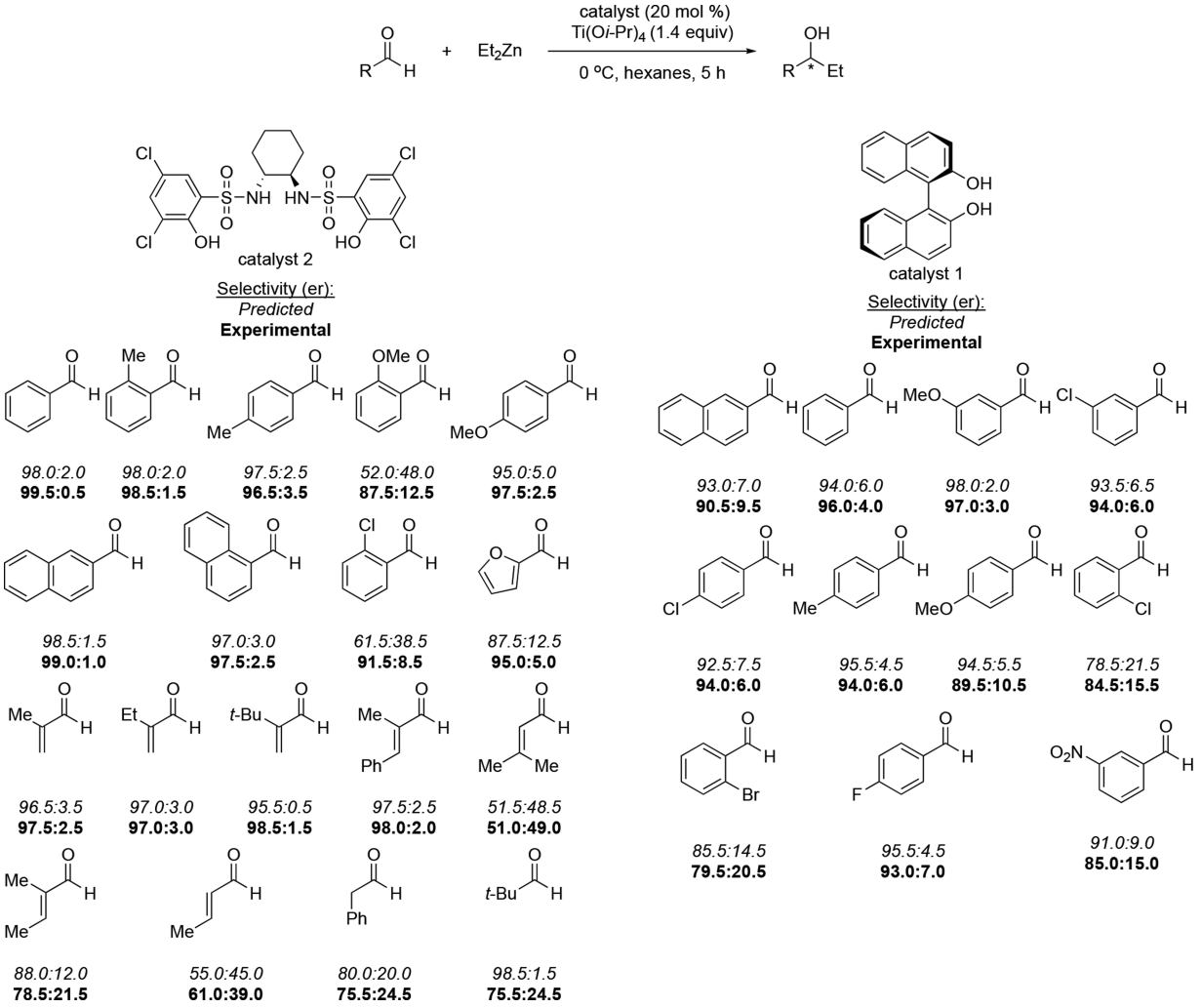


Figure 44. Catalysts for the enantioselective addition of diethylzinc to aryl aldehydes.

chiral, secondary amine shown in Figure 46). For scaffold **ii**, catalysts **S1Sa** through **S1Se** are synthesized. This 40 member set gives an observed selectivity range from 65:35 to 4.5:95.5 (*S*:*R*) er.

With this set of catalysts, the authors sought to evaluate the predictive capabilities of 3D-QSAR (CoMFA-like), 3.5D-QSAR,[174] and 4D-QSAR[175] methods. Both 3.5D- and 4D-QSAR methods take the conformer distribution of molecules into account. 4D-QSAR uses a molecular dynamics trajectory to calculate time-averaged occupancy values at each grid point. 3.5D-QSAR is so named because it can be thought of as a hybrid between 3D-QSAR and 4D-QSAR. In 3.5D-QSAR, different conformations are taken from a molecular dynamics trajectory. These structures are minimized and are all used to calculate descriptors in a MIF.

To obtain the best 3D-QSAR models, many attempts were made to select a single, representative conformer. First, five different conformer selection methods are tested: (1) the lowest energy conformation of each catalyst, (2) the lowest energy (+)-backbone configuration for each catalyst, (3) the lowest energy (−)-backbone configuration for each catalyst, (4) the opposite backbone configuration of each catalyst with respect to the lowest energy conformer for each catalyst, and (5) random conformer selection for each catalyst. It is worth noting that only the substituent coordinates are used in descriptor calculation for 3D, 3.5D, and 4D methods. Considering only (+) or (−) backbone configurations produce inaccurate models (R^2 = 0.68 and 0.54, respectively), suggesting that both (+) and (−) forms of the backbone are in equilibrium in solution. The best model is obtained from considering the global minimum energy conformer (R^2 = 0.94, Q^2 = 0.78). The opposite backbone configuration gives a similarly correlated, but perhaps overfit model (R^2 = 0.88, Q^2 = 0.65) whereas random selection (on average over 100 different selections) give generally weak models (R^2 = 0.59, Q^2 = 0.22). The best model is compared with 3.5D and 4D methods.

Descriptors for both 3.5D and 4D approaches are calculated using two methods — MIFs and indicator fields. Indicator fields differ from MIFs in that rather than calculating a molecular interaction (MIF), simple metrics are used to identify occupied regions of space. For example, the steric indicator field is calculated with a binary metric of occupancy (1 if a grid point of a conformer overlaps with the van der Waals radius of an atom, 0 if it does not). Similarly, the electronic indicator fields assign the value of the atomic charge to a grid point within the van der Waals radius of an atom. For the 3.5D method, a hydrogen bond acceptor indicator field is also employed, calculated the same way as the steric indicator field but populated only when the grid point overlaps with a hydrogen bond donor. When compared with MIFs, these

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

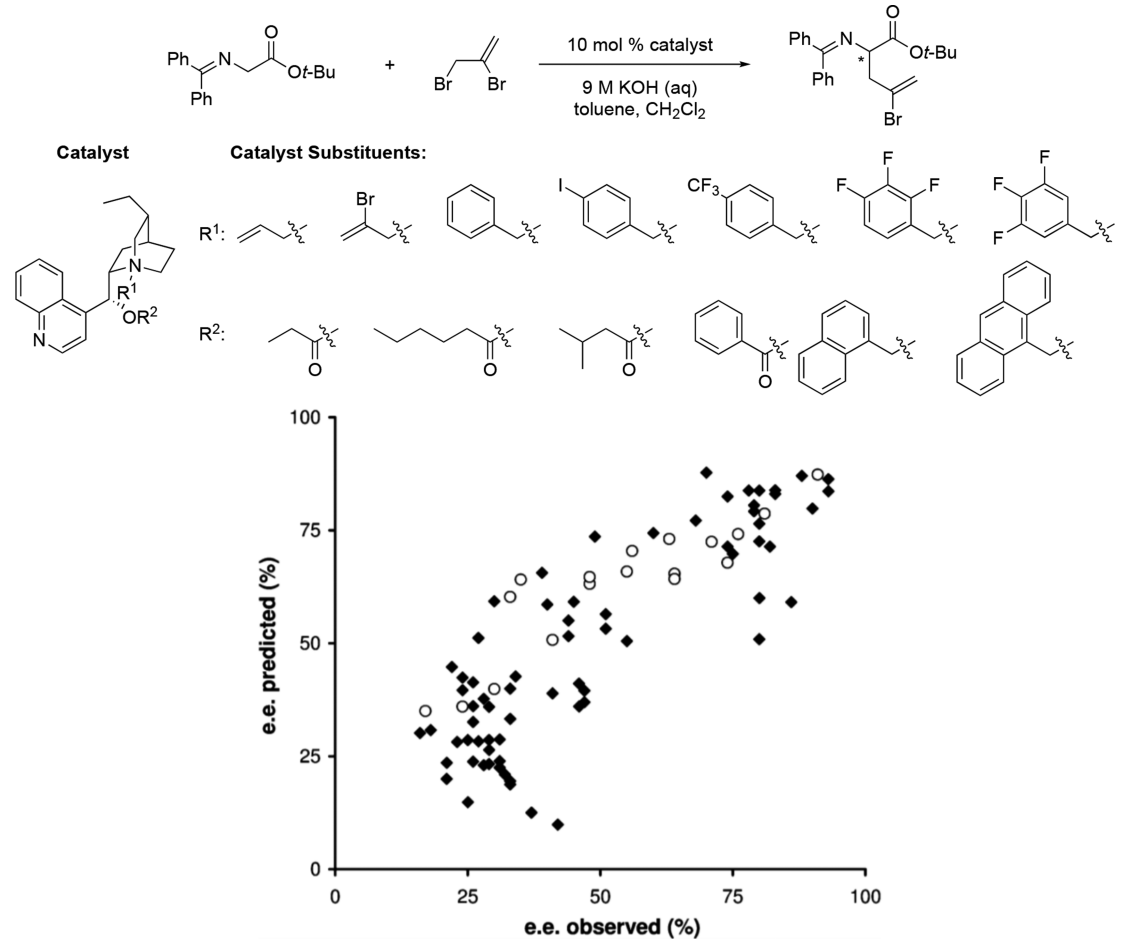


Figure 45. Cinchonidinium-alkaloid-catalyzed, enantioselective, phase transfer alkylations along with predicted versus observed plot for training (diamonds) and test (circles) catalyst sets. Adapted with permission from ref 172. Copyright Royal Society of Chemistry.

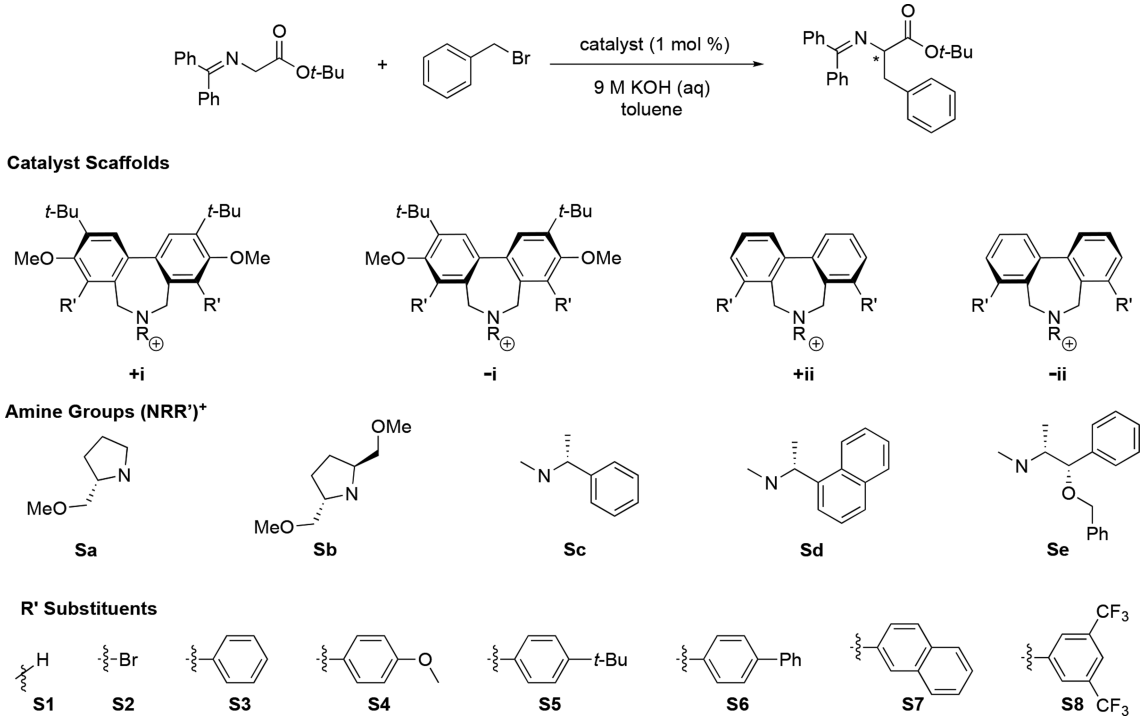


Figure 46. Enantioselective alkylation with biaryl-derived phase transfer catalysis.

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

simpler descriptors performed worse in 3D-QSAR, but better in 3.5D- and 4D-QSAR, likely because they minimized noise in the data for the latter methods.

To compare 3D-, 3.5D-, and 4D-QSAR methods, two-deep cross-validation was used.[176,177] In this study, the authors first used four-fold (four sets of 10 catalysts), wherein 30 are used to create a model and 10 were used to validate the model. The 30 training catalysts are then used to make a model, using LOO cross-validation, and the residuals of the remaining 10 catalysts are used to validate the model. This process is repeated until all 40 catalysts have been used in the first validation set. With this more robust method of cross-validation, 3.5D descriptors including the hydrogen bond acceptor indicator field give the best models (Q^2 = 0.73) followed by 4D descriptors (Q^2 = 0.71) followed by 3.5D descriptors with only steric and electronic indicator fields (Q^2 = 0.70), with 3D descriptors giving the weakest models (Q^2 = 0.64). Thus, 3.5D and 4D descriptors may be computationally more intensive, but also lead to more robust models.

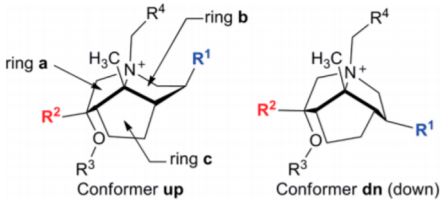
The application of 3D-QSAR in the study of chiral phase transfer catalysts has been implemented by Denmark and co-workers.[178,179] A novel class of cyclopentapyrrolizidinium catalysts was synthesized by the tandem inter [4 + 2]/intra[3 + 2] cycloaddition of nitroalkenes with chiral enol ethers followed by hydrogenolysis. This route allows the synthesis of 160 catalysts that are then tested in an asymmetric alkylation reaction.

To elucidate the features important for enantiodifferentiation, 3D-QSSR models (CoMFA) were generated. Chemoinformatics methods using 0D, 1D, 2D, and 3D descriptors were also calculated and correlated to catalyst activity. However, because the focus of this review is 3D-QSSR and its application to asymmetric catalysis, only the study of enantioselectivity will be discussed here; interested readers are directed to the original work for more information regarding catalyst activity.[178,179]

Following a similar workflow, a global minimum conformer was located for each catalyst with molecular mechanics that was later verified with DFT (B3LYP/6-31G*). Because the global minimum conformer may not be the relevant conformer in the stereodetermining transition structure, multiple conformer classes were identified and categorized on the basis of the scaffold conformation (Figure 47). The core structures are assigned to different libraries considering their substituents and the relative conformation of the b ring. To evaluate the best set of conformers, each library is used to generate QSSRs, in which the best conformer set is selected according to which library gives the strongest models.

Molecular charges are calculated with MNDO, and MIFs are calculated using Coulombic potentials and Lennard-Jones potentials. Cutoff energies are applied to avoid extraordinarily large values for grid points within the van der Waals radius of an atom. Indicator fields have also been explored; instead of a binary indication of occupancy, grid points overlapping with the van der Waals radius of an atom are assigned the cutoff energy value and other are assigned a value of zero. The conformers in library D (defined in Figure 47) with indicator fields at a 30 kcal/mol steric cutoff and 15 kcal/mol electronic cutoff give the best models (R^2 = 0.924, Q^2 = 0.778).

QSSR models can be used to identify regions of space around the catalyst where increasing/decreasing steric or electronic effects impact enantioselectivity. These contour maps (Figure 48) are discussed with reference to the



	R²/R¹				
library	H/H	H/Me	Me/Me	(*i*-Pr or *t*-Bu)/Me	aryl[a]/Me
A[b]	up[c]	up	dn	dn	dn
B	dn	dn	dn	dn	dn
C	up	up	up	up	up
D	up	up	up	dn	up
E	dn	dn	dn	up	up

Figure 47. Different possible conformations of the catalyst scaffold. (a) aryl = Ph, 1-naphthyl, mesityl. (b) Library containing different conformer combinations, (c) Conformation of scaffold. Reproduced from 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.* **2011**, 76, 4337−4357. Copyright 2011 American Chemical Society.

perspective of the actual structure shown on the far left. The green surfaces on the bottom right next to the methyl group indicate that steric bulk in that region is necessary to shield that portion of the b-ring. The green contours overlapping with the 3,5-positions of the arene ring are consistent with substitutions here increasing selectivity. The green contour next to the benzyl substituent (R^3) on the oxygen atom is consistent with the observation that having hydrogen at the R^2 position decreases selectivity; if this position is a hydrogen atom, the arene ring rotates into this region of space. This arrangement is associated with low selectivity. The yellow contour near the nitrogen atom indicates that bulky groups at R^4 leads to low selectivity, presumably because of shielding of the electrostatic interactions. The electrostatic potential map corroborates this analysis, with the blue contour over the a-ring indicating this region as a productive binding site for the substrate.

These observations are consistent with the following stereochemical rationale: each face favors association with either the *Re* or *Si* face of the enolate. Although the extent of selectivity for *Re* or *Si* face may not be perfect, enolate association is likely at either the a-ring or the b-ring and the competition in binding is responsible for diminished selectivity. These findings led to the following design criteria to optimize this catalyst scaffold: (1) the monopole is not variable, (2) the dipole can be reinforced by installing polarizing groups near the nitrogen, (3) R^2 = aryl may result in favorable π-interactions that could increase selectivity, (4) the addition of steric bulk to three of the four faces of the core should disfavor association to all faces except one, and (5) the removal of steric bulk from the remaining face should facilitate selective enolate association.

CoMFA models have also been used to study enantioselective ketone hydrogenation reactions.[180] In this study, 25 ruthenium complexes with both chiral diamine ligands and chiral bisphosphine ligands with known experimental selectivity values for the enantioselective hydrogenation of acetophenone are selected and split into a training set of 20 members and an external test set of five members.[181−185] The data in this

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

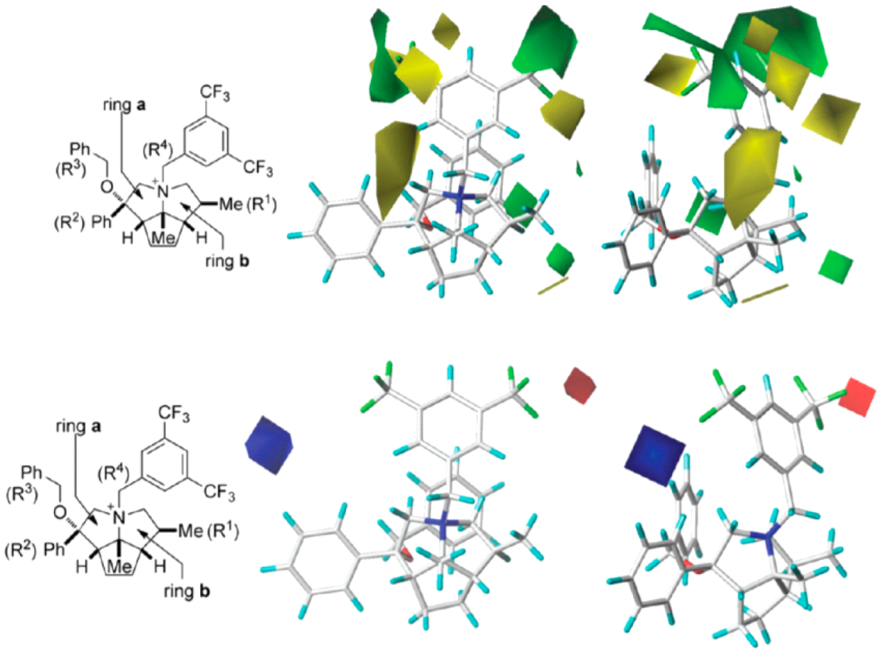


Figure 48. Top: steric contour maps from two different perspectives. Green contours indicate regions where steric bulk leads to increased enantioselectivity, whereas yellow regions indicate regions where less steric bulk leads to increased enantioselectivity. Bottom: Electrostatic contour maps from two different perspectives. Blue contours indicate regions where increased positive charge leads to greater enantioselectivity, whereas red contours indicate regions where decreased positive charge (or increased negative charge) leads to increased enantioselectivity. Reproduced from 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.* **2011**, 76, 4337−4357. Copyright 2011 American Chemical Society.

set ranges from 99.5:0.5 er (*R* is the major enantiomer) to 0.5:99.5 er (*S* is the major enantiomer). Steric, electronic, and H-bond donor MIFs as well as indicator fields are calculated for each catalyst. The performance of the five member external test is accurately predicted (R^2 = 0.974 for the predicted vs observed plot for the external test set), indicating a strong model. From this model, contour maps could be developed to elucidate the structural features of the catalyst responsible for enantioinduction (Figure 49).

The relative contributions from steric and electronic effects are found to be 80% and 20%, respectively. The green contour, where more steric bulk leads to greater selectivity, is localized around the diamine ligand. The N−H−O interaction postulated in the stereodetermining transition structure suggests that increasing steric bulk around the amide residue will bias one diastereomeric transition structure over another, leading to enhanced selectivity. Examination of the electronic MIF suggests that introducing negative charge in the vicinity of the aromatic rings of the diamine should also increase selectivity. To test this hypothesis, the authors increased the electron density in the aromatic ring of catalyst **A1** (experimental selectivity = 87% ee, predicted selectivity = 78.1% ee) by installing NH_2 units on the *para* positions (Figure 49). This new catalyst is predicted to have a selectivity of 84.3% ee.

Unfortunately, the authors did not synthesize this catalyst, but instead calculated the free energy differential between the competing diastereomeric transition structures leading to enantiomers of the product. Whereas the calculated free energy differential for the two transition structures employing **A1** is 1 kcal/mol, the free energy differential for the corresponding transition structures employing **C1** is 2 kcal/mol. Although experimental validation would provide more compelling evidence to unambiguously prove the conclusions drawn in this report, this example represents a case in which the selectivity of a chiral catalyst is improved theoretically by making modifications suggested by 3D-QSSR.

Except for a few reports, most studies described in this section use projection to latent structure (PLS) modeling for the QSSR. However, many of the regression coefficients for the descriptors used in CoMFA and related methods could be assigned to zero, such as in grid points that reside far away from the catalytically active entity or points with variance approaching zero. Yamaguchi and co-workers have used least absolute shrinkage and selection operator (LASSO)/Elastic net regressions to construct QSSRs assigning values of zero to unimportant coefficients.[186] These authors used steric indicator fields as descriptors as well as calculated electronic descriptors,[187] including only a global minimum energy conformer for esterification reactions and the enantioselective addition of phenylboronic acid to 1-napththaldehyde. However, as the focus of this review is on asymmetric catalysis, only the latter will be discussed.

A summary of the catalysts employed and selectivity values obtained are given in Figure 50. Eighteen catalysts are used in the training set and five in the external test set. LASSO/Elastic net regression analysis gives good models (R^2 = 0.92 for the predicted vs observed plot of the external test set). In analogy to the previous discussion, regions in space where steric occupancy is either beneficial or detrimental to enantioselectivity can be visualized using these methods (Figure 50). Parts b and d of Figure 50 correspond to a catalyst in which R = R′ = 3,5-dimethylphenyl. The red region in Figure 50b designates that occupancy in this region of space is associated with diminishing enantioselectivity. Thus, the authors synthesized a catalyst in which the methyl groups are removed (R =

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

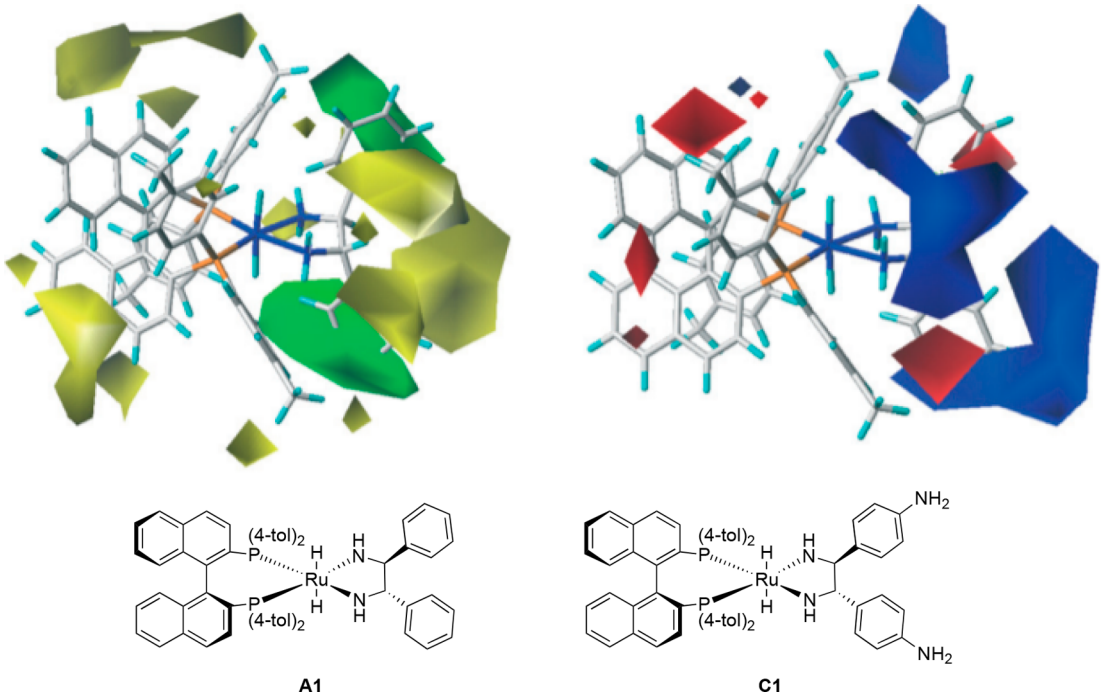


Figure 49. Top left: steric contour map for catalysts in enantioselective ketone hydrogenation reactions. Green contours indicate regions where steric bulk leads to increased enantioselectivity, whereas yellow regions indicate regions devoid of steric bulk which lead to increased enantioselectivity. Top right: Electrostatic contour map. Blue contours indicate regions where increased positive charge leads to greater enantioselectivity whereas red contours indicate regions where decreased positive charge (or increased negative charge) lead to increased enantioselectivity. Experimental catalyst (*A1*) and theoretically improved catalyst (*C1*). Adapted with permission from ref 180. Copyright Royal Society of Chemistry.

R′ = 3-methylphenyl, shown in parts a and c of Figure 50. The selectivity of this catalyst is improved with respect to the original, validating that LASSO/Elastic Net can be used to generate QSSRs and the physical information in these relationships can be used to enhance selectivity. No attempt was made in this work to further optimize the reaction.

5.1.3. Complete, Chemoinformatics Guided, Catalyst Discovery Workflow. Recently, our laboratory has developed a computer-guided workflow that uses chemoinformatics at all stages of development.[138] This workflow consists of the following components: (1) construction of an *in silico* library of a large collection of conceivable, synthetically accessible catalysts of a particular scaffold, (2) calculation of robust chemical descriptors for each scaffold, and (3) selection of a representative subset of the catalysts in this space.

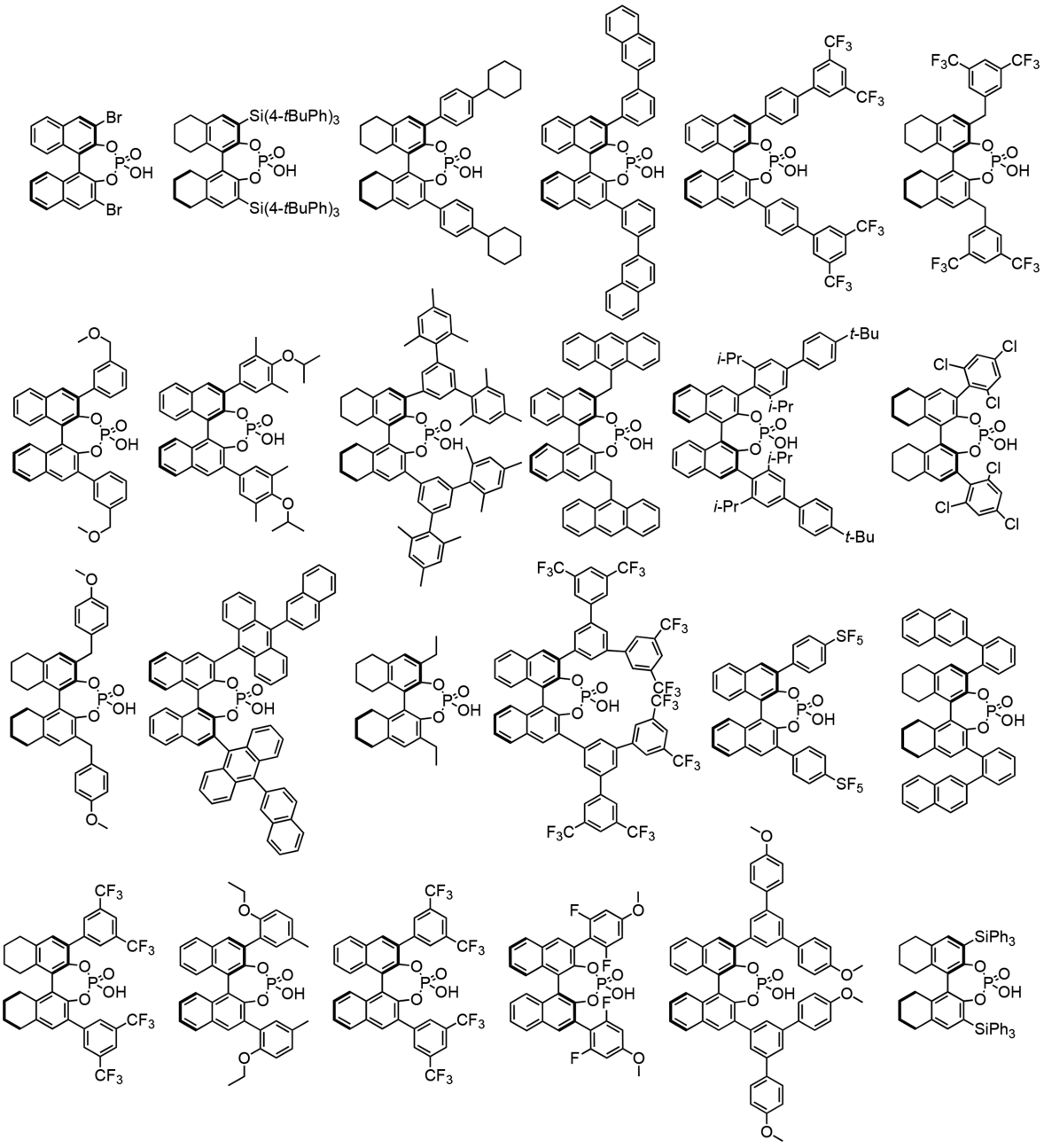
This subset is agnostic to reaction or mechanism as the only input in the selection algorithm is the intrinsic properties of the catalysts. Accordingly, it is named a Universal Training Set (UTS). The next steps are: (4) collection of the training data and (5) application of modern machine learning methods to generate models that predict the enantioselectivity of each member of the *in silico* library. These models are evaluated with an external test set of catalysts (predicting selectivities of catalysts outside of the training data). The validated models can then be used to select the optimal catalyst for a given reaction.

As a prototype of this workflow, an *in silico* library of 806 chiral phosphoric acids is generated. For each member, steric and electronic descriptors are calculated. The newly developed steric descriptors, called average steric occupancy (ASO) descriptors are used which are conceptually similar to Hirst's 3.5D descriptors.[172] First, a conformer distribution for each catalyst in the *in silico* library is obtained. Second, for each catalyst, all of the conformers are aligned and individually placed in identical grids. If a grid point is within the van der Waals radius of an atom, it is assigned a value of 1; otherwise it is assigned a value of 0. This process is repeated for *n* conformers and upon completion each grid point has a cumulative value ranging from 0 to *n*. The values are then normalized by dividing by *n* such that all grid points have a value between 0 and 1. These values comprised the steric descriptors for the structures. For electronic descriptors, a calculable parameter arising from the perturbation of the electrostatic potential energy of trimethylammonium ions by substituents has been developed which correlates well with Hammett parameters. Calculation of these descriptors for each catalyst affords the chemical space on which the Kennard-Stone subset selection algorithm is applied, yielding the UTS (Chart 1). These descriptors are also used to digitize the reactants and products; concatenation of catalyst, reactant, and product descriptors combinatorially yields an *in silico* library of unique reactions.

To validate this workflow, the training set was evaluated on a previously optimized model reaction. The enantioselective formation of N,S-acetals (Figure 51A) developed by Antilla and co-workers was selected.[188] For the selected model reaction, a four by four grid of imines and thiols is chosen, resulting in 16 reactions per catalyst (Figure 51A). Evaluating the 24-member training set with each substrate combination then results in 384 training reactions that are used for model development. The range of selectivities covered by the UTS in the 16 training reactions spans from 28.5:71.5 to >99.5:0.5 er with the same enantiomer of catalyst, further supporting the

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

Chart 1. UTS for Chiral, Phosphoric Acid Catalysts



hypothesis that this training set selection method covers a broad range of selectivity-space.

A suite of models is generated and used to predict the selectivity of three families of test sets (Figure 51B), namely: (1) a "substrate test set" of reactions generating new products (i.e., those formed from substrates not included in the training set), (2) a "catalyst test set" of reactions generating the same products in the training set but with catalysts not included in the training data (Figure 51C), and (3) a "substrate/catalyst (sub/cat) test set" of reactions creating new products also using catalysts not included in the training set.

Support vector regressors with a second order polynomial kernel gave the highest performance on the basis of the mean absolute deviation (MAD) from the combined external test sets (Figure 52A). The first test set evaluated only the ability of the models to predict the selectivity of reactions forming new products. In this role, the model excelled with an MAD of 0.161 kcal/mol. Next, the same model is used to predict the selectivity of the external test set of catalysts. The performance of the model is still highly accurate, with an MAD of 0.211 kcal/mol. Finally, reactions forming new products with the external test catalysts are predicted with an MAD of 0.236 kcal/mol. The chemical space constructed from the first three principal components of descriptor space also reveals regions of high-, medium-, and low- selectivity space, indicating the ASO descriptors are accurately capturing the catalyst features responsible for enantioinduction (Figure 52B).

To demonstrate the potential to identify new, selective catalysts, a situation was simulated in which highly selective reactions have not been achieved. To do this, only reactions below 80% ee were used as training data. Deep feedforward neural networks accurately reproduce the experimental

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

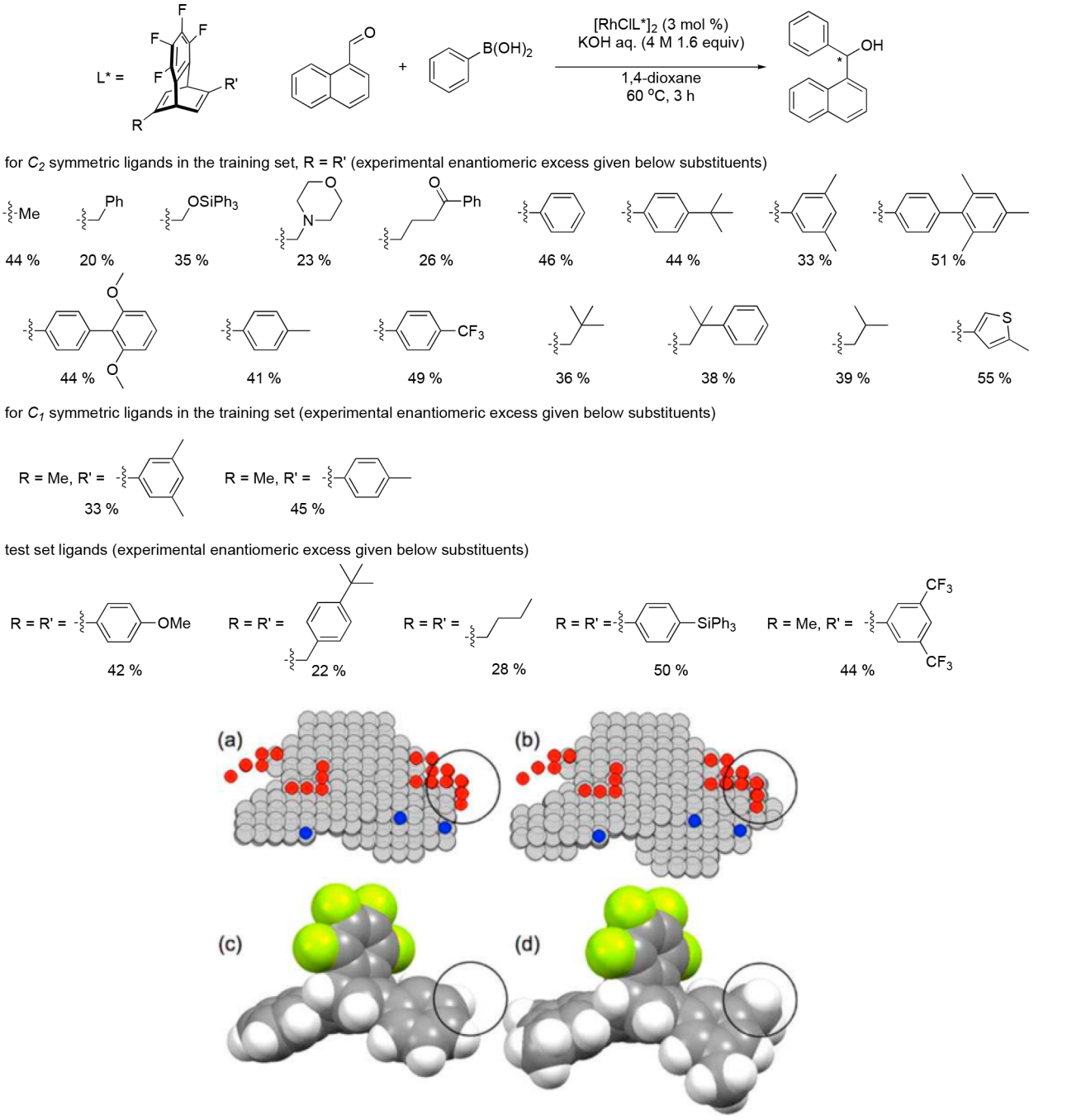


Figure 50. Rhodium catalyzed enantioselective addition of phenylboronic acid to 1-naphthaldehyde with depiction of ligand library. (a–d) Space filling models and digitized structures of two catalysts. The red areas designate regions where steric bulk is associated with diminished selectivity, and the blue regions designate areas where steric bulk is associated with high selectivity. The circled region in b and d contains the methyl unit that was removed (it is absent in a and c). Reprinted from with permission from ref 186. Copyright 2017 John Wiley and Sons.

selectivities (MAD = 0.33 kcal/mol, Figure 53A), and the general trends in selectivity on the basis of average catalyst selectivity. As shown in Figure 53B, the most selective catalyst evaluated is predicted as the most selective catalyst. The next two catalysts are also the second and third most selective catalysts (the order is inverted, but they are within experimental error of each other).

5.1.4. Perspective on Alignment Dependent MIF-Based Methods. Alignment dependent MIF-based methods and related protocols have the capacity to identify how perturbations of the steric and electronic environments around the active catalytic entity can influence the enantioselectivity of that scaffold. Moreover, the studies detailed in this subsection demonstrate the general applicability of this method across multiple catalyst architectures. Thus, MIF-based methods are indeed a promising avenue of research if one wishes to identify a general, chemoinformatics-driven protocol to catalyst optimization. The most serious limitation of these methods is the alignment dependency. For comparing variations of catalysts of the same basic core scaffold, alignment is trivial. However, if one wishes to compare multiple catalyst scaffolds, alignment can quickly become challenging. To address this

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620–1689

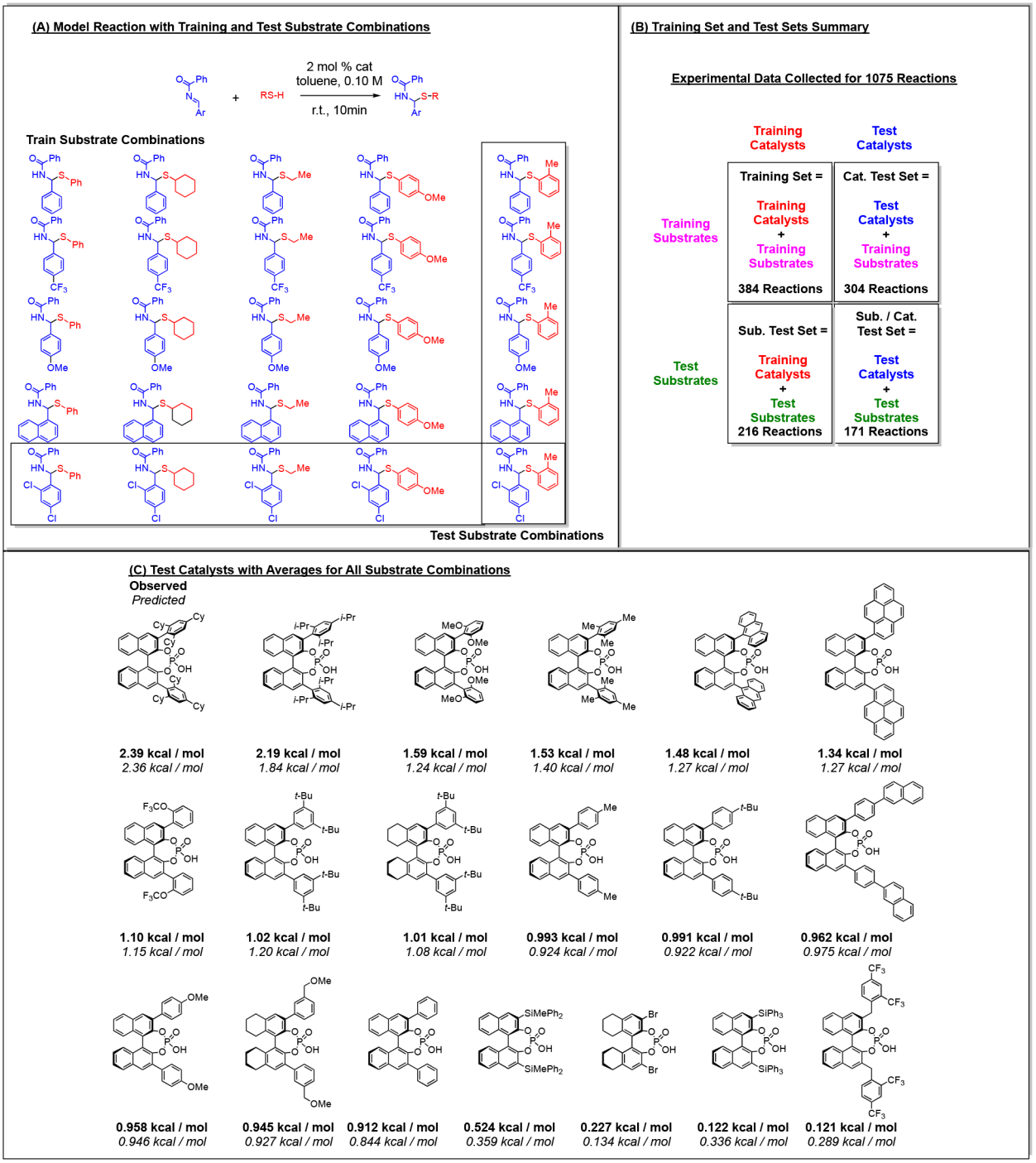


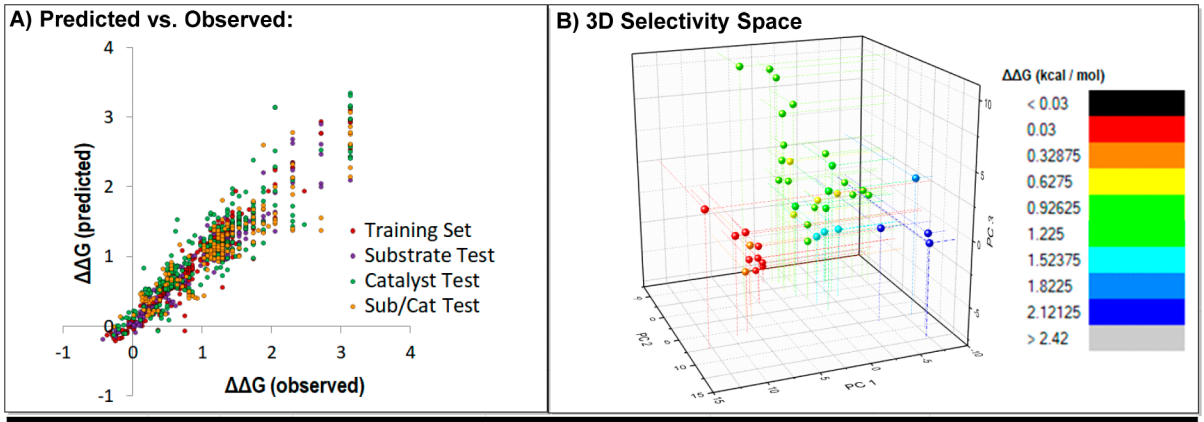
Figure 51. (A) Formation of chiral N,S-acetals with train and test substrate combinations. (B) Catalyst and substrate combinations forming different train and test sets. (C) External test catalysts. Adapted with permission from ref 138. Copyright 2019 American Association for the Advancement of Science.

limitation, grid independent descriptors (GRIND) have been developed, which constitute the topic of the next section.

5.2. Grid Independent Descriptors (GRIND)

5.2.1. Introduction to GRIND. To address the necessity of alignment for the previously described MIF-based methods, the GRIND class of descriptors was developed.[189] These descriptors are derived from MIFs that then undergo processing to generate GRIND. The first step is a simplification of the MIF of interest by identifying important regions in space where interaction energies are of large magnitude (either large negative for favorable interactions or large positive for unfavorable interactions). These regions are identified by points (termed nodes in the GRIND nomenclature) of high interaction energy. Subsequent selection of new regions is derived from additional points of high interaction energy outside a predefined distance away from previously selected regions. Then, the grid points surrounding the high interaction energy grid points are included to identify the important regions in space around the molecule. This process is called filtering and only the selected grid points (high interaction energy points and

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689



Model	Q^2	R^2	MAD (Train, ΔΔG)	MAD (SubTest, ΔΔG)	MAD (CatTest, ΔΔG)	MAD (Sub/CatTest, ΔΔG)
SVR_POLY2	0.748	0.953	0.096	0.161	0.211	0.236

Figure 52. (A) Predicted versus observed free energies (kcal/mol) of the train and test sets overlaid for a support vector machine using a second order polynomial kernel. Accuracy metrics are listed in the table below. (B) Selectivity space as represented by the first three principal components of the full feature-space. Adapted with permission from ref 138. Copyright 2019 American Association for the Advancement of Science.

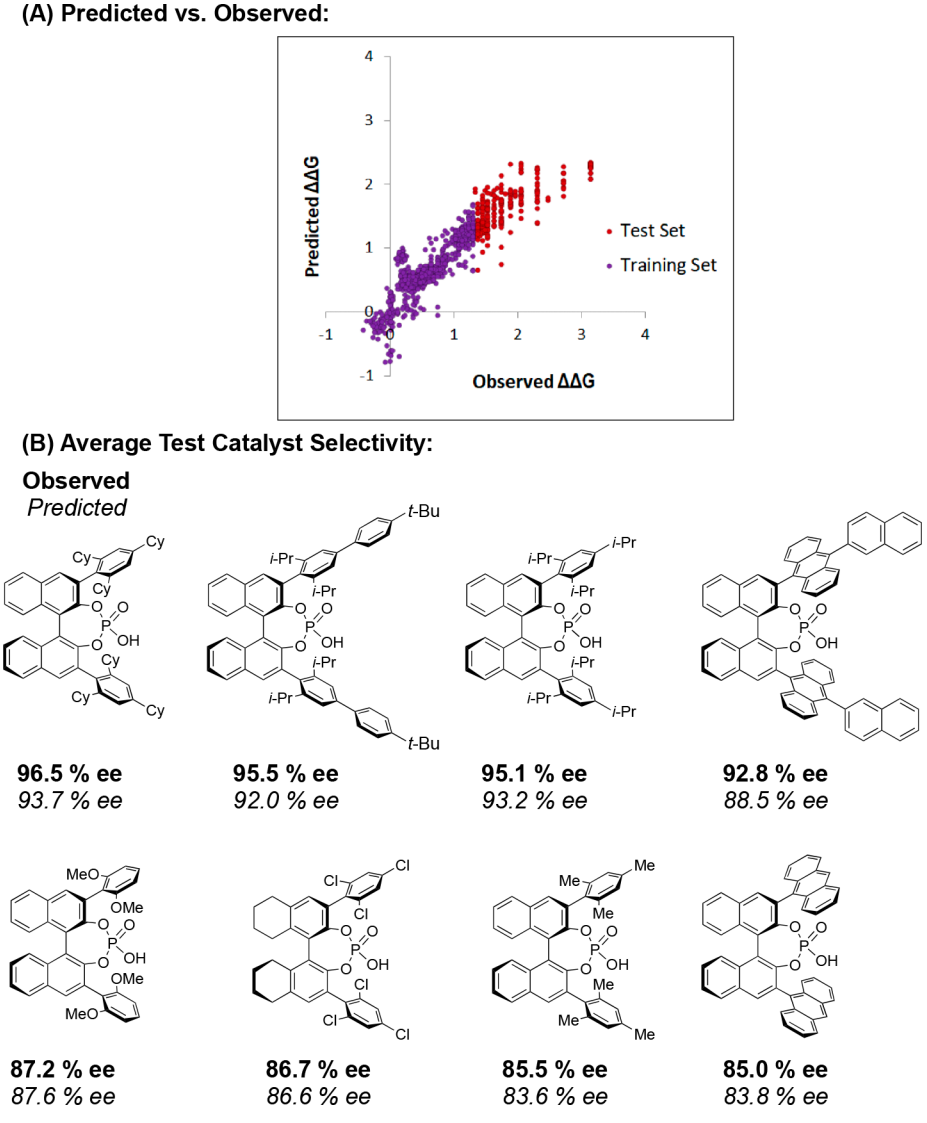


Figure 53. (A) Predicted versus observed plot for simulated reaction optimization. (B) Average predicted and observed selectivity data for all catalysts with average selectivity over 80% ee. Adapted with permission from ref 138. Copyright 2019 American Association for the Advancement of Science.

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

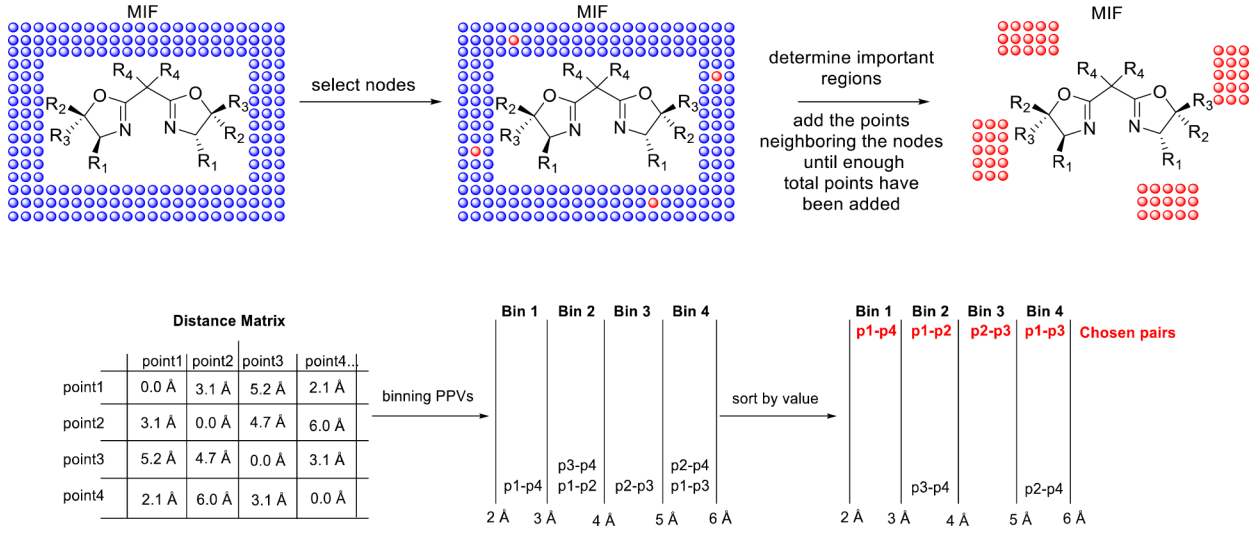


Figure 54. Graphical representation of the process for the calculation of GRIND.

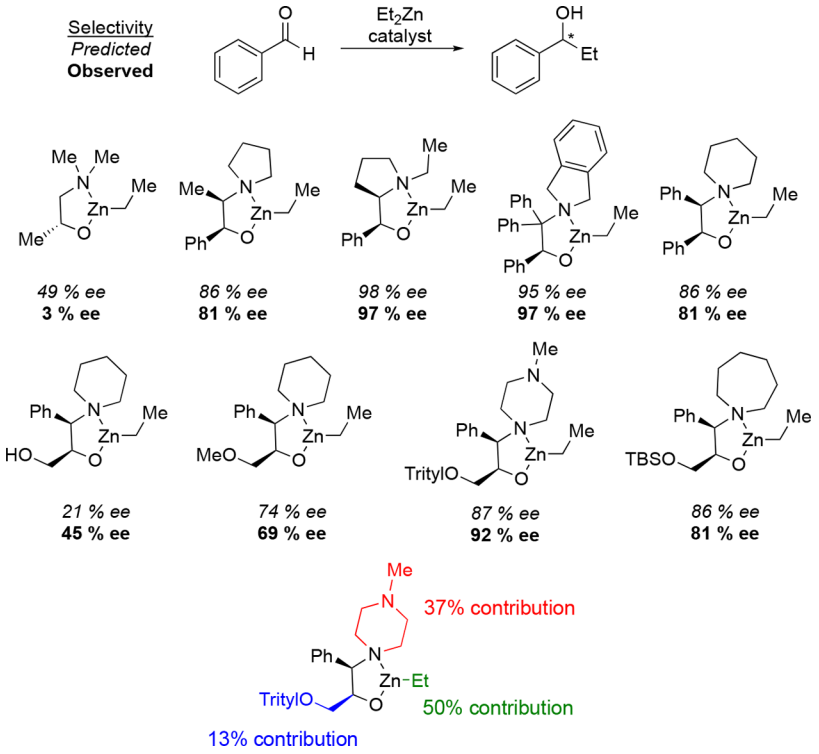


Figure 55. Predicted and observed selectivities for enantioselective addition of diethylzinc to benzaldehyde. GSI ratios for different structural motifs are depicted in color.

surrounding points) are taken on to future steps. For all the selected grid points, the pairwise products of all the interaction energies are calculated and sorted on the basis of the distance between the two grid points multiplied. For example, all products from grid point pairs in which the distance between points is 1−2 Å would be one class, whereas products from grid point pairs in which the distance between points is less than 1 Å could be another class. The highest value(s) in each class are taken forward and used as a descriptor. Thus, the number of descriptors per molecule is determined by the number of distance ranges used and the number of values kept per distance range (Figure 54). This method was later augmented with a molecular shape field.[190]

5.2.2. Applications of GRIND in Asymmetric Catalysis.

The first application of this method to asymmetric catalysis was reported by Higginson and Morao[191] and sought to benchmark descriptor performance with Lipkowitz's CoMFA models,[158] Kozlowski's QMQSAR models,[164] and an enantioselective reduction of acetophenone with borane catalyzed by chiral β-amino alcohols reported by Damen and co-workers.[192] GRIND was calculated following the protocol above with steric, electronic, and hydrogen-bonding MIFs, in which only the highest value per distance range in the field is kept. When compared with Lipkowitz's models, the cross-validated correlation coefficient from LOO cross-validation is lower than reported in the original study (Q^2 = 0.52 vs 0.84,

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

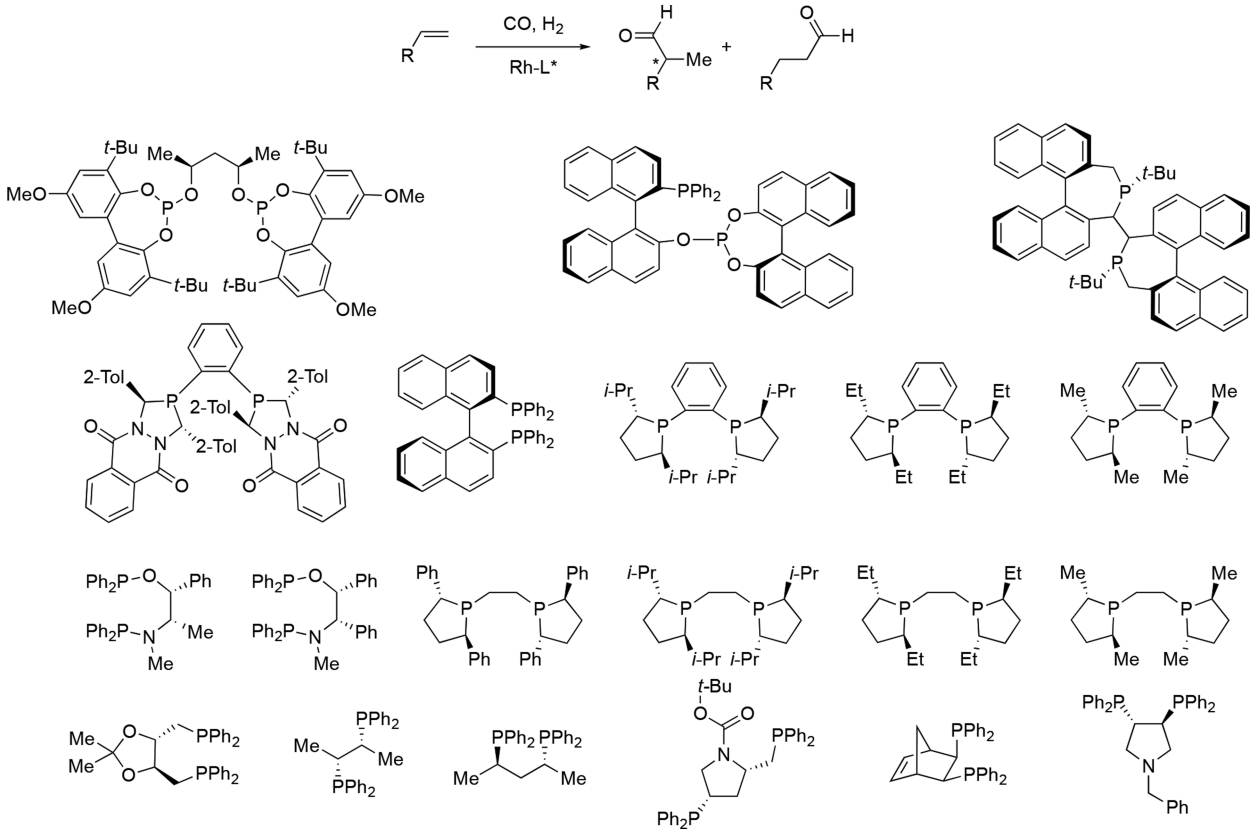


Figure 56. Rhodium-catalyzed hydroformylation of terminal alkenes.

respectively). Although Kozlowski does not provide a Q^2, the two methods predict similar external validation sets, with Kozlowski's MAD of 6.25% and Higginson and Morao's MAD of 6.75%. It is noteworthy that different modeling methods are used in each of these studies, confounding whether the descriptors themselves are comparable or a more advanced modeling method paired with inferior descriptors gives results similar to the original study. For the final study, the Q^2 value for both the GRIND descriptors and the Damen study are approximately 0.8, but given that the original report did not disclose their descriptor calculation method, it is difficult to draw conclusions from this result. Examining the relative performance of the GRIND with respect to the three original studies, the GRIND descriptors at best give similar results, as suggested in the second two comparisons. When compared with Lipkowitz's work, the GRIND gives somewhat diminished, although still potentially meaningful results.

This protocol has been modified by the use of values obtained from quantum mechanics from which the descriptors were calculated.[193] Two fields are calculated from which to derive GRIND, a molecular shape field (MSF) and a molecular electrostatic potential field (MEP). The MEP field is calculated using DFT methods in a way similar to the aforementioned GQSAR method.[166] The MSF is calculated by using the local curvature of the molecular surface.[190] Convex areas range from −1 to 0, whereas concave areas range from 0 to 1. These MIFs are then subjected to filtering and pairwise multiplication, and the products sorted by interpoint distance as described in the general GRIND method. These descriptors are then evaluated in the enantioselective addition of dialkylzinc reagents to aldehydes. This reaction is chosen because the set of 18 catalysts evaluated by Kozlowski[166,169] is available for comparison and a 40 member set is available using other literature sources.[194,195] When training on the same 14 catalysts as Kozlowski and Higginson and Morao with the same four remaining catalysts held out as an external test set, a MAD of 15.7% ee between the predicted and observed values for the four test catalysts is found, significantly higher than the two previous cases (6.25 and 6.75% ee respectively). Moving on to the 40-member set, this group is divided into a training set of 30 members and a test set of 10 members. The predicted versus observed values found in the test set are compared and the model predicts moderately to highly selective catalysts accurately but performs poorly when predicting low selectivity catalysts (Figure 55). The authors attribute this disparity to the skew of the data set, wherein most training catalysts have an energy differential between the diastereomeric transition structures leading to different enantiomers of the product greater than 1 kcal/mol.

The relative importance of each selected grid point can also be calculated to identify important regions of space around that catalytic entity. The individual contribution of each node can be calculated by first tabulating the product of the following three values for each node−node interaction: (1) QSSR coefficient, (2) the correlogram value (the product of the interaction energies at two nodes), and (3) the relative contribution of the node to the correlogram value across all pairwise combinations of which that node is a part. The summation of these product values for each node quantifies the relative importance of that node. These relative importance values are then attributed to structural motifs on the basis of the relative proximity of the node to different substituents. This value is termed the Group Structural Influence (GSI) ratio, which quantifies the overall contribution of the individual

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

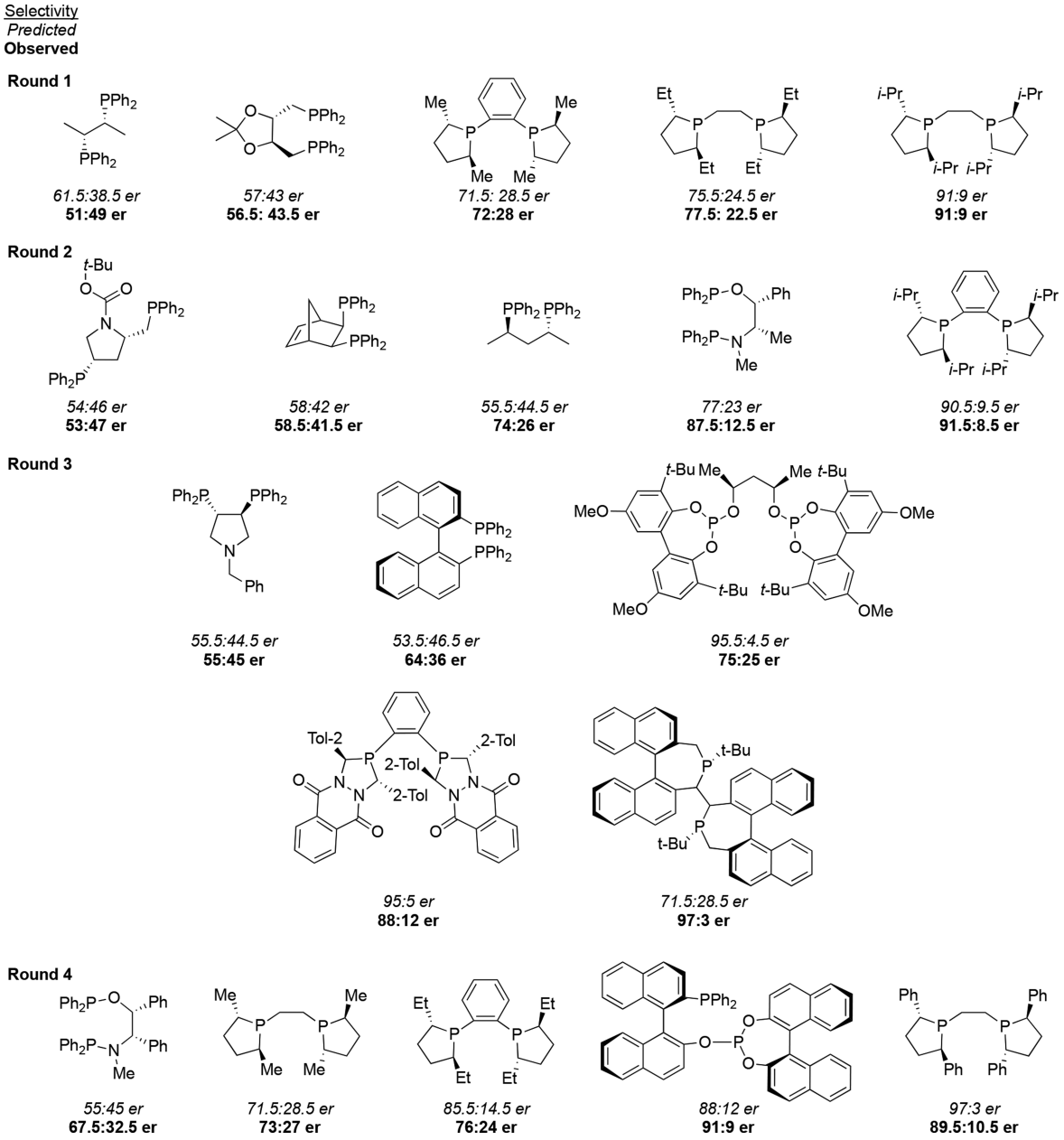


Figure 57. Predicted and observed selectivity data for selected catalysts.

group to the enantioselectivity. An example of relative contributions of each group for selectivity is shown for the diethylzinc alkylation of benzaldehyde (Figure 55).

For this structure, the finding that the phenyl substituent makes no contribution to the observed enantioselectivity is surprising. Unfortunately, the authors did not report the synthesis of a catalyst with a different group at this position to test this conclusion. It is also possible that the phenyl substituent has a secondary effect of changing the equilibrium conformation of the O-trityl group and thus this contribution cannot be detected by the GSI ratio. Another interesting observation is the high contribution attributed to the ethyl ligand, given that this residue is constant across all catalysts examined. Nevertheless, this method does represent an attempt to quantify the relative importance of groups. Further development and validation of GSI ratios could facilitate the extraction of useful structural information pertaining to catalyst selectivity.

This method of using GRIND calculated from quantum mechanically derived MIFs has also been applied to the study of enantioselective, rhodium-catalyzed hydroformylation reactions.[196] This work also studied catalyst activity; however, because the focus of this review is on asymmetric catalysis only this section of the work will be discussed. The original workflow was modified by a second filtering method, in which the most negative values in the MEP field (points representing the most basic areas of the catalyst denoted as BAS) are removed.[197] Thus, four descriptor types could be calculated, two for each filtering method (filtered on the basis of the most convex areas of the isosurface as in the original publication or on the basis of the most negative values in the MEP field). These include: (1) MSF-MSF$_{convex}$, (2) MEP-MEP$_{convex}$, (3) MSF-MSF$_{BAS}$, and (4) MEP-MEP$_{BAS}$. Twenty catalysts were

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

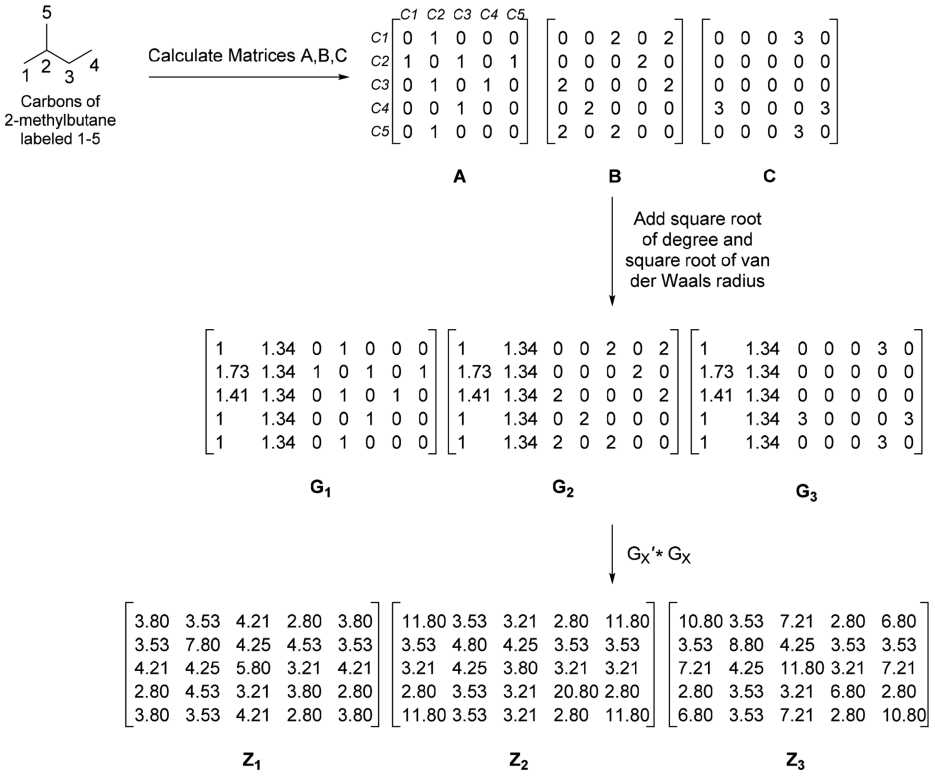


Figure 58. Construction of matrices Z_1, Z_2, and Z_3.

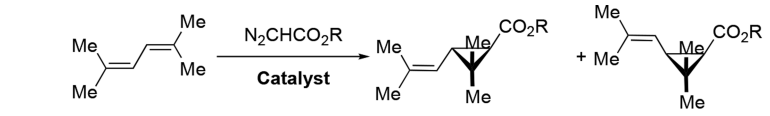
selected from the literature to use in modeling and model validation (Figure 56).[198,199]

The four descriptor types listed previously are each tested by modeling the above data set and validating with LOO cross-validation. Neither the MSF-MSF nor the MEP-MEP descriptors generated with the BAS filtering method give acceptable models (Q^2 = 0.12 and 0.41, respectively). Both MSF-MSF and MEP-MEP descriptors derived from the filtering method derived from the most convex regions give good results (Q^2 = 0.68 and 0.60, respectively). With the MSF-MSF descriptors from this filtering method five catalysts are selected at random from the twenty-member set to be withheld as an external validation set with the remaining 15 used to construct a model. This process is repeated four-times such that the effect of which training catalysts are used on model efficacy could be examined (Figure 57). Rounds 1, 2, and 4 afford moderate success when predicting ligand selectivity in that the models generated can typically predict whether a catalyst will give high, medium, or low selectivity. Round 3 shows the worst performance, likely because only one catalyst left in the training set had over 55:45 er. Thus, this region of high selectivity chemical space may not have been well described by the model. The authors are also able to use the coefficients from the QSSR to extract some physical information pertaining to how catalyst structure influences catalyst performance, identifying regions where steric encumbrance is related to high enantioselectivity.

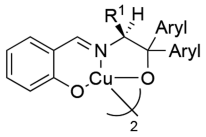
5.2.3. Perspective on GRIND. In general, grid-independent MIF-based methods underperform with respect to their alignment dependent counterparts. We hypothesize that during the construction of the correlogram, information pertaining to high interaction energies in different regions of space is lost. For example, one can imagine a hypothetical situation wherein a catalytic entity has two regions of high-interaction energy—one at the "front" of the molecule, near the catalytically active center, and one at the "back". First consider the situation in which a substituent at the front of the catalyst has the highest interaction energy whereas the substituent in the back has the second highest interaction energy. Second, consider the same core catalyst scaffold in which the highest interaction energy is now at the back of the catalyst and the second highest interaction energy is at the front. Because the sorting of the node−node products is distance based, these catalysts will appear nearly identical in GRIND-space. Moreover, if groups responsible for producing the highest and second highest interaction energies give different selectivity values, the calculated descriptors will not capture the important relative spatial information necessary for differentiating the two compounds. One could argue that selection of the correct number of high interaction energy regions or use of the right MIF or indicator field could result in stronger models. However, this correction will likely require individualized optimization for each new system studied.

Another possible explanation could be that constructing the correlogram introduces more noise into the descriptors. Thus, when GRIND descriptors are used on small data sets, it is much more difficult to construct meaningful QSSRs than with their alignment-dependent counterparts. Perhaps using the standard GRIND-based method on larger data sets could increase the performance of these methods. Despite being somewhat less explored than alignment dependent methods, GRIND has the potential to be very useful in that it becomes possible to bypass the alignment process. Although alignment dependent MIF methods may have an advantage over GRIND when it is easy to align the molecules of interest, GRIND may be the best tool available when evaluating more structurally diverse scaffolds in which alignment is not straightforward. In this sense, GRIND, a much younger technology than CoMFA,

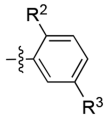
DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689



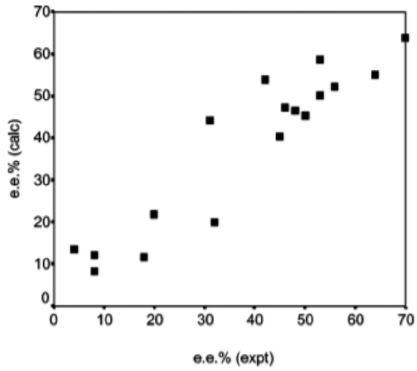
Catalyst



Aryl =



R1	R2	R3	Predicted er	Observed er
Me	H	H	52:48	56:44
Me	OMe	H	60:40	61:39
Me	O-*n*-Bu	H	75:25	72:28
Me	O-*n*-Bu	Me	74:26	73:27
Me	O-*n*-Bu	*n*-Pr	78:22	76:24
Me	O-*n*-Bu	*t*-Bu	82:18	77:23
Me	O-*n*-Octyl	*t*-Bu	85:15	82:18
CH_2Ph	H	H	54:46	54:46
CH_2Ph	OBu	H	72:28	70:30
CH_2Ph	OBu	*n*-Pr	73:27	73:27
CH_2Ph	OBu	*t*-Bu	76:24	75:25
CH_2Ph	O-*n*-Octyl	*t*-Bu	76:24	79:21
$CHMe_2$	H	H	54:46	56:44
$CHMe_2$	O-*n*-Bu	H	65:35	72:28
$CHMe_2$	O-*n*-Bu	*t*-Bu	71:29	76:24
CHMeEt	H	H	59:41	56:44
CHMeEt	OMe	H	66:44	60:40



$R^2 = 0.823$
standard error = 8.255 % *ee*

$$\%\ ee = 13.7(4.3) - 0.3(0.3)A_{X1}^{R1} + 11.1(4.3)A_{X2}^{R2} - 7.9(3.8)A_{X3}^{R3} + 1.2(0.8)A_{X1}^{R3}$$

Figure 59. Enantioselective cyclopropanation and predictive model of enantioselectivity. Adapted from Jiang, C.; Li, Y.; Tian, Q.; You, T. QSAR Study of Catalytic Asymmetric Reactions with Topological Indices. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 1876−1881. Copyright 2003 American Chemical Society.

represents a promising avenue of research that should be investigated further.

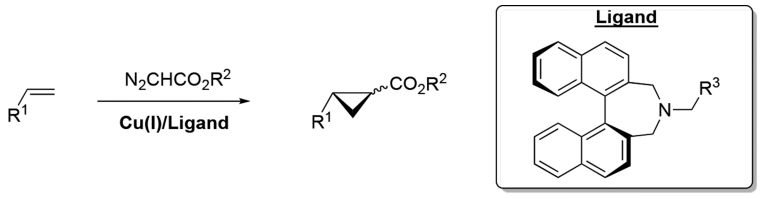
6. OTHER APPLICATIONS OF CHEMOINFORMATICS IN ENANTIOSELECTIVE CATALYSIS

The previous sections in this review illustrate descriptors derived from the 3D representation of molecular structures. Alternatively, lower-dimensional representations can be used to formulate descriptors for applications in chemoinformatics. Such representations have several advantages over 3D descriptors including (but not necessarily limited to) the following: (1) they can be calculated much more rapidly than their 3D counterparts, (2) they do not require optimization of molecular structures, and (3) they do not require conformational analysis. However, one could argue that these representations inherently lack critical information about the 3D structure of the molecules of interest and are thus inferior representations. As the following section will demonstrate, it is possible to construct predictive models using such descriptors to represent variable subunits of a common core in a similar way Sterimol parameters are used in the LFER section. A commentary on the interpretability of these models will be discussed throughout, if applicable, and in the Perspectives section.

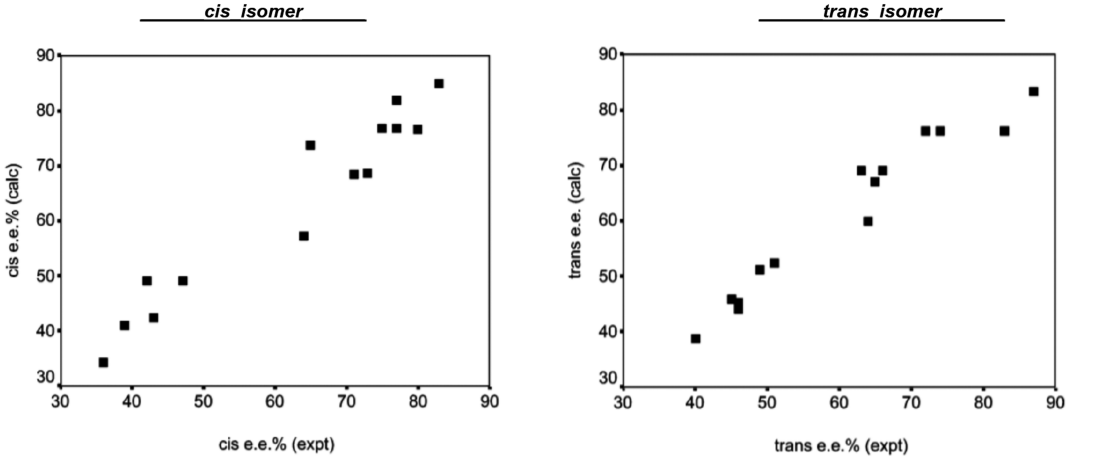
6.1. Topological Indices as Descriptors in Enantioselective Catalysis

Topological indices, such as those developed by Xu and coworkers, are an example of 2D descriptors.[200] These descriptors are calculated for a molecule by first defining three path matrices, termed A, B, and C. A path matrix is a two-dimensional representation of the connectivity of a molecule determined by which atoms are bonded to each other. Thus, each dimension of a matrix is an identical list of atoms, wherein the matrix element corresponding to the intersection between two different atoms receives some value representing the relationship between those two atoms. In this particular case, matrix A is constructed such that each matrix element receives a value of 1 if the path between vertices has a value of 1 (i.e., the atoms are separated by one bond). If the path between vertices is not 1, it receives a value of zero. For matrix B, if the path between vertices is equal to 2, the matrix element receives a value of 2 and otherwise receives a value of 0. The same is true for matrix C, wherein a path length of 3 corresponds to a matrix element of 3 and otherwise a value of 0

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689



			cis isomer		trans isomer	
R^1	R^2	R^3	Predicted er	Observed er	Predicted er	Observed er
C_6H_5	Et	pyridinyl	68:32	67:33	70:30	69:31
C_6H_5	Et	quinolinyl	71:29	71:29	72:28	73:27
C_6H_5	Dicyclohexylmethyl	pyridinyl	85:15	84:16	81:19	84:16
C_6H_5	Dicyclohexylmethyl	quinolinyl	90:10	88:12	91:9	88:12
C_6H_5	L-menthyl	pyridinyl	88:12	88:12	87:13	88:12
C_6H_5	L-menthyl	quinolinyl	91:9	92:8	93:7	91:9
C_6H_5	D-menthhyl	pyridinyl	69:31	70:30	73:27	72:28
C_6H_5	D-menthyl	quinolinyl	73:27	74:26	75:25	76:24
4-MeC_6H_4	L-menthyl	pyridinyl	82:18	87:13	83:17	83:17
4-MeC_6H_4	L-menthyl	quinolinyl	88:12	91:9	86:14	88:12
4-$MeOC_6H_4$	L-menthyl	pyridinyl	86:14	84:16	82:18	79:21
4-$MeOC_6H_4$	L-menthyl	quinolinyl	87:13	88:12	82:18	83:17
4-ClC_6H_4	L-menthyl	pyridinyl	71:29	74:26	73:27	72:28
4-ClC_6H_4	L-menthyl	quinolinyl	82:18	78:22	74:26	75:25



% *ee* = -526.4(93.6)+50.9(8.6)A_{X1}^{R1} - 14.6(2.4)A_{X3}^{R1}+1.4(0.2)A_{X1}^{R2}+0.7(0.3)A_{X1}^{R3}-17.9(2.5)C_{R2}

R^2 = 0.919

standard error = 5.045 % *ee*

% *ee* = -533.5(86.8)+53.3(8.0)A_{X1}^{R1} - 15.9(2.2)A_{X3}^{R1}+1.2(0.2)A_{X1}^{R2}+0.6(0.2)A_{X1}^{R3}-15.5(2.3)C_{R2}

R^2 = 0.908

standard error = 4.676 % *ee*

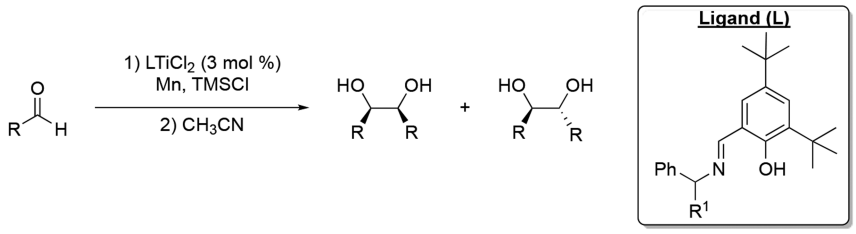
Figure 60. Enantioselective cyclopropanation and enantioselectivity predictive model. Reproduced from Jiang, C.; Li, Y.; Tian, Q.; You, T. QSAR Study of Catalytic Asymmetric Reactions with Topological Indices. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 1876−1881. Copyright 2003 American Chemical Society.

is assigned. To these matrices are added two columns: (1) the square root of the vertex degree (i.e., square root of the number of non-hydrogen atoms bound to the atom), and (2) the square root of the van der Waals radius of the atom. This operation results in the formation of three new matrices, termed G_1-G_3. These matrices are then multiplied by the corresponding transpose matrices (i.e., G_1G_1'), which gives three new matrices Z_1−Z_3. The topological indices (A_{X1}, A_{X2}, and A_{X3}) are then defined as the largest eigenvalue of each respective matrix divided by 2 (Figure 58).

You and co-workers used these molecular representations to predict reaction outcomes in three enantioselective reactions;[201] enantioselective cyclopropanation,[202] enantioselective pinacol coupling[203] and, enantioselective cross-coupling reactions with Grignard reagents.[204] In each case, multiple regression analysis is employed.

The enantioselective cyclopropanation employs a copper-salen-derived complex (Figure 59). The ligand contains three groups that are varied, the substituent on the stereogenic carbon atom of the ligand (R^1 in Figure 59), the substituent on the 2-position of the aromatic residue attached to the oxygen-bearing carbon (R^2 in Figure 59), and the substituent on the 5-position of the aromatic residue affixed to the oxygen-bearing carbon (R^3 in Figure 59). Topological descriptors are calculated for each of the three substituents, generating a total of nine descriptors for each catalyst. The 17 different catalyst structures employed are divided such that 16 are used to develop the linear regression model and the remaining catalyst is used to validate the model. The regression analysis yielded a reasonable correlation (R^2 = 0.823, Figure 59) with a standard error of 8.3% ee. A simple neural network provides slightly diminished performance.[205]

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689



		dl / (dl+meso) %	
R	R^1	Observed	Predicted
C_6H_5	H	99	100
C_6H_5	Ph	97	97
C_6H_5	Me	92	90
1-naphthyl	Me	70	74
4-$MeOC_6H_4$	H	92	93
2-BrC_6H_4	Ph	58	93
3-BrC_6H_4	H	96	97
4-BrC_6H_4	H	95	97
2-furyl	H	88	86
4-PhC_6H_4	H	91	88
4-$AcOC_6H_4$	Ph	91	91
$CH{=}CHC_6H_4$	Me	84	82
$CH_2CH_2C_6H_4$	Me	66	82

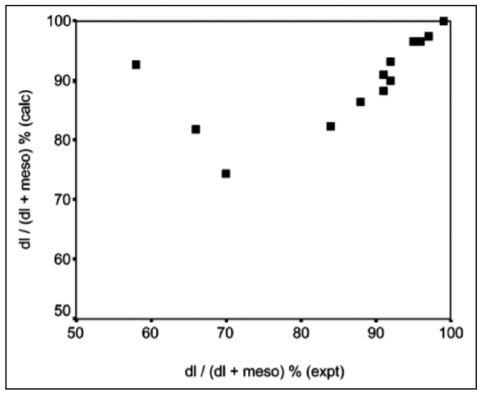


Figure 61. Diastereoselective pinacol coupling and predictive model of diastereoselectivity. Reproduced from Jiang, C.; Li, Y.; Tian, Q.; You, T. QSAR Study of Catalytic Asymmetric Reactions with Topological Indices. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 1876−1881. Copyright 2003 American Chemical Society.

A related analysis is then performed using a different cyclopropanation reaction (Figure 60). In this case, data for both diastereomers of the product are available; thus, two different models are constructed to predict the enantiomeric composition of each diastereomer. In this model, three topological descriptors are derived from each the styrene aromatic residue, the identity of the substituent on the ester residue of the diazo compound, and the identity of the heteroaromatic residue on the catalyst. In addition to these nine descriptors, an indicator of either −1 or +1 is given to specify the configuration of the L- or D-menthyl substituent on the ester residue, whereas achiral residues receive a value of 0. Using these 10 descriptors, regression analysis is performed to predict the reaction outcomes for both possible diastereomers, the results of which are given in Figure 60. Of the 14 available data points, 13 are used to construct the model and one additional data point is used for external validation. This model is quite accurate giving an R^2 = 0.908 and a standard error of 4.8% ee for the *trans*-isomer and an R^2 = 0.919 and standard error of 5.0% ee for the *cis*-isomer. The large coefficient of the A_{x3}^{R1} term in the regression analysis is interesting. Because this matrix is associated with three-bond connections, the authors suggest that this result indicates the importance of the steric bulk of the styrene substrate to achieve high selectivity (larger substituents lead to higher selectivity). This interpretation is surprising given the similar sizes of the R^1 substituents $-C_6H_6$, 4-MeC_6H_5, 4-$MeOC_6H_5$, and 4-ClC_6H_5. By most metrics, the latter three substituents would not be considered significantly sterically different; the most obvious difference among the groups is the electronic character. Because no electronic information is available in the descriptors, it is surprising that the model can accurately identify this trend and make accurate predictions. One possible explanation is that because the van der Waals radius is present in the original A, B, and C matrices, the model developed some method of "penalizing" the compounds with chlorine atoms (or atoms with larger van der Waals radii) present and can thus reproduce the trend. An interesting challenge to this model would be to prepare the 4-$CF_3C_6H_5$ or 4-FC_6H_5 analogs as external tests for further validation of this interpretation and potentially exposes shortcomings in the descriptors used. If the model fails to predict the 4-FC_6H_5 case, this could suggest the correlation (and its interpretation) is not founded in the underlying physics and that the variable with a large coefficient is also correlated with the actual causative property within the domain of the model. Such a situation would both confound extraction of physical meaning from the model and limit the domain of applicability. Additionally, it would be interesting to include Hammett parameters in the model and see if the significance of the other variables decreases. If this is the case, it is likely that the physical interpretation of the current model is unfounded.

The authors revisited this system to compare the performance of models generated with linear regression with those generated by neural networks.[205] The neural network contained two output neurons and could thus predict the enantiomeric composition for both diastereomers simultaneously. Unfortunately, different combinations of the aforementioned topological indices are used in each case making direct comparisons between the models difficult. Similar predictive capabilities are observed in each case (*trans*- and *cis*-products) for both modeling methods.

In the second case study, the diastereoselectivity of a pinacol coupling is predicted using the percentage of *dl*/(*dl*+*meso*) as the dependent variable (Figure 61). Of the 13 data points available, 12 are used as training data with one validation case. Topological indices are calculated for both a variable catalyst substituent of interest and the variable portion of the aldehyde substrate (Figure 61). During the regression analysis, all except two cases are predicted with exceptional accuracy. These two cases highlight limitations of the topological descriptors used in the analysis. In the first case, a large error is present in predicting the selectivity of the 2-BrC_6H_4-substituted aldehyde substrate. Because only the aromatic residue is treated in the analysis, the 2-Br, 3-Br, and 4-Br substituted phenyl rings are

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

identical in the molecular representation. Thus, predicting the influence of an *ortho* substituent is impossible. In the second case, two methylene substituents linking the phenyl ring and the aldehyde are erroneously predicted to be a competent substrate and very similar in selectivity to cinnamaldehyde. This outcome is unsurprising given the nearly identical topological representations of these groups. Thus, the descriptors used here are incapable of identifying such changes and are unable to predict their influence on enantioselectivity.

Finally, the authors examine an enantioselective Kumada cross-coupling (Figure 62). In this case study, the aromatic substituent on the phosphine ligand, the chiral amine-containing substituent of the ligand, and the identity of the Grignard reagent are used to obtain nine descriptors. The planar chirality of the ferrocene is included as an indicator such that +1 designates *R* and −1 designates *S* configurations, respectively. Using these descriptors, 13 data points are used to construct a multivariate model with good correlation (R^2 = 0.915) and a standard error of 12.3% ee (one additional point is used for validation). However, this model excluded two points with high deviation which the authors postulate to have mechanistic differences and are therefore difficult to predict with the current method. An alternative hypothesis could be that even though the process is mechanistically the same, the descriptors do not adequately describe an important feature of the catalyst structure for those two points and thus cannot accurately predict them. The model is slightly improved with by the use of neural networks but could still not accurately predict the problematic points.[205]

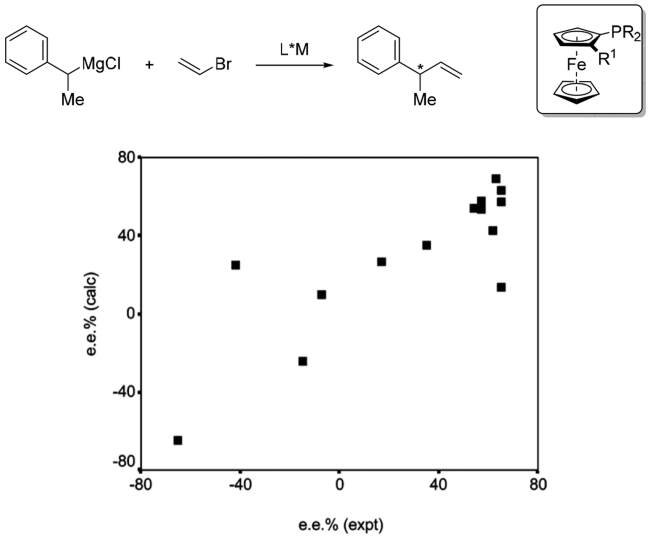


Figure 62. Enantioselective Kumada coupling reaction. Reproduced from Jiang, C.; Li, Y.; Tian, Q.; You, T. QSAR Study of Catalytic Asymmetric Reactions with Topological Indices. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 1876−1881. Copyright 2003 American Chemical Society.

Topographical descriptors are beneficial in that they require minimal computational cost with respect to other 3D descriptors. However, the descriptors in this work give a limited representation of the molecules of interest. It is probable that using more physically meaningful representation (i.e., that contain information about shape and electronic characteristics of molecules) could lead more robust models. Further, although the models discussed give reasonable predictions, they lack rigorous validation likely owing to the lack of available experimental data.

Later work by You and co-workers examined the enantioselective addition of diethylzinc to aldehydes.[206] In this work, the topological descriptors employed are the Randić index,[207] Kier and Hall index,[208] and the Kier shape index.[209] A more detailed discussion of how these descriptors are calculated is available in the Supporting Information. In addition to these topological descriptors, AM1[210] charges for key atoms are also used. The authors used three-layer feedforward neural networks to construct models in four different case studies.

The first data set is taken from Pericás and co-workers, in which a single amino alcohol ligand is used to catalyze the addition of diethylzinc into a series of aldehydes (Figure 63).[211] Nineteen different aldehydes are represented with Randić order 2, Kier and Hall order 2, and Kier shape index order 3. A neural network with three input neurons, three nodes in a hidden layer, and one output neuron is used to produce a model to predict enantioselectivity. The model had acceptable accuracy over the narrow range of observed selectivity in the training data, with Q^2 = 0.5245 (5-fold cross-validation) and R^2 = 0.8575. It is interesting that the model can accurately predict the outcome of the reaction with 2-methoxybenzaldehyde as the substrate. Similar substrates, including 2-chloro and 2-methylbenzaldehyde, give high selectivity. Given that the descriptors contain no direct information about electronic contributions, the origin of this ability to accurately predict the reaction outcome for this substrate is mysterious.

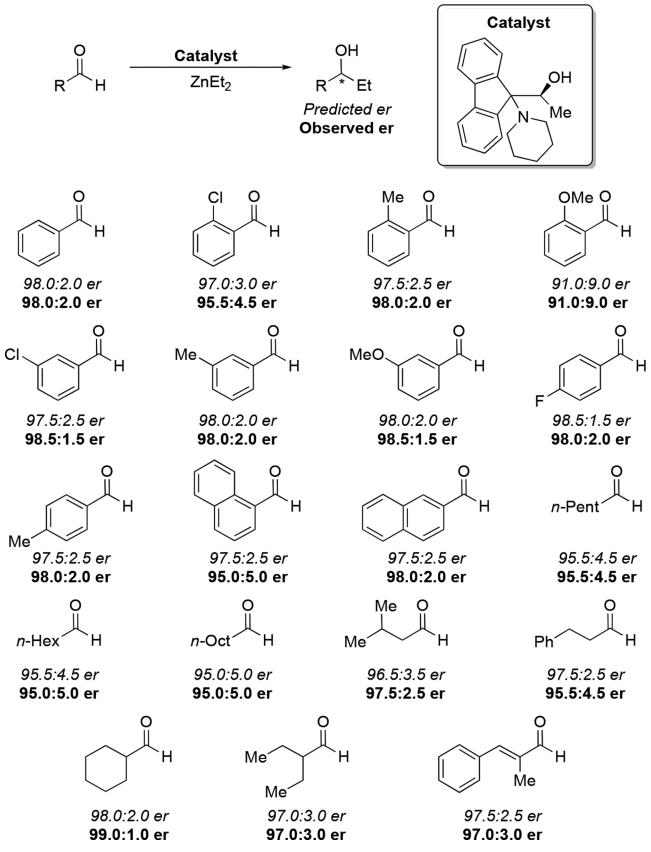


Figure 63. Enantioselective addition of diethylzinc to aldehydes using an amino alcohol ligand.

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

The second case study[206] employs a data set from Kang and co-workers in which various amino thiol ligands with different *N*-substituents are evaluated as catalysts (Figure 64).[212,213] Aldehyde substrates are represented with the Kier and Hall order 2 index whereas the catalysts are represented with the Kier and Hall order 2 index for the *N*-substituent and the partial charge (AM1) on the nitrogen atom. A set of 28 reactions is used to train a network with three input neurons, three hidden neurons, and one output neuron. The model is internally validated with seven-fold cross-validation (R^2 = 0.8580, Q^2 = 0.6376). The researchers attempted to use the model to probe if the size of the R-group on the aldehyde substrate is necessary for selectivity as suggested by Noyori for amino alcohol catalysts.[214,215] Reaction selectivities are predicted for substrates with increasing linker distance between the aldehyde residue and a phenyl substituent to probe the expected influence of steric bulk of the aldehyde in the reaction. As expected, less steric bulk correlates to lower predicted selectivity. However, no experiments are performed to validate these predictions. Although there is no reason to believe amino thiols require different substrate features to achieve high selectivity when compared with amino alcohol catalysts, experimental validation would be valuable to further test model robustness by treating the predicted values as an external test set and assess if the model is still applicable in the domain of less sterically biased substrates.

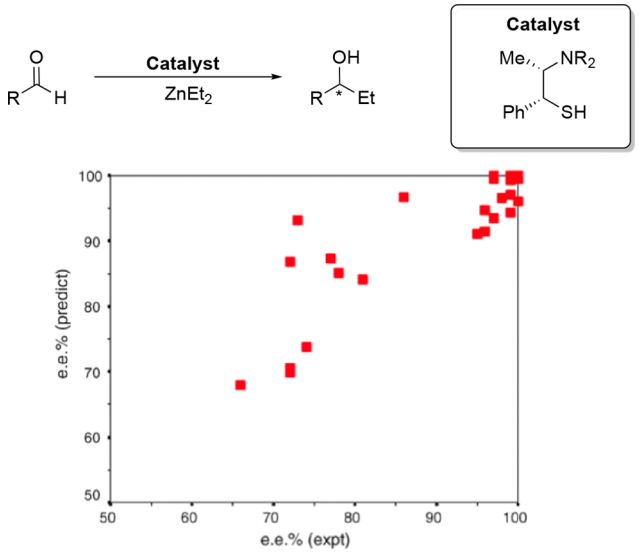


Figure 64. Enantioselective addition of diethylzinc to aldehydes using amino thiol ligands. Adapted with permission from ref 206. Copyright 2006 Elsevier.

In a third case study,[206] You and co-workers employed a data set from Falorni and co-workers[216,217] that uses a 2-(2-pyridyl)pyrrolidine ligand. A set of 15 aldehydes are represented by Randić order 3 index, Kier and Hall order 3 index, and the partial charge (AM1) on the oxygen of the aldehyde (Figure 65). Additional features included for each reaction are the temperature and reaction time. A neural network with five input nodes, two nodes in a hidden layer, and one output neuron is trained and internally validated with five-fold cross-validation (Q^2 = 0.8570, R^2 = 0.9334). Using this model, the authors plotted the relationship between predicted selectivity over time. A maximum is identified at 40 h; however, it is likely that this occurs simply because the only reactions available in the training data that required long reaction times also had low selectivity. Thus, the longer reaction time regime for selective reactions is likely outside the domain of applicability of the model.

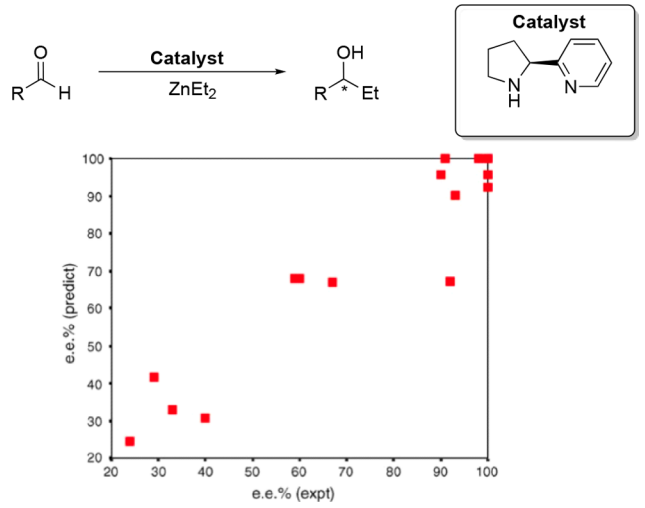


Figure 65. Enantioselective addition of diethylzinc to aldehydes using a 2-(2-pyridyl)pyrrolidine ligand. Adapted with permission from ref 206. Copyright 2006 Elsevier.

Finally, the You and co-workers experimentally evaluated histidine-derived catalysts with various *N*-substituents (Figure 66).[206] This particular scaffold is of interest because of its multiple potential binding sites. Aldehydes and the catalyst N-substituent are represented with the Randic order 2 index. Further, the partial charge of the oxygen atom of the aldehyde is selected as an input variable. Experimental data is collected for 11 unique reactions, for which a model is constructed with internally validation (Q^2 = 0.5451, R^2 = 0.8833). The authors intended to predict a more selective catalyst using this information, however, the predicted values for two additional N-substituents (naphthyl and *t*-butyl) suggested no improvement beyond what is already observed. This conclusion would benefit from experimental justification—it is likely that the two new N-substituents are outside the domain of applicability of the model, especially considering the very small data set used to train the model. Thus, it is not possible to definitively say whether the new catalysts could be more selective. The authors also suggest that further experimental work is necessary; however, to our knowledge no further experimental study has been published.

6.2. Other Applications of QSSR in Enantioselective Catalysis

Damen and co-workers have reported the application of QSSR to predict the reaction outcome of the enantioselective reduction of acetophenone with chiral oxazaborolidine reagents (Figure 67).[192] A 24-member training set of different amino alcohol catalysts is used to construct a model using partial least-squares regression which is then validated with a four member external test set (R^2 = 0.978, Q^2 = 0.797). However, the specific descriptors used in the study are not specified. The study served as a proof of concept exemplifying the application of QSSR to enantioselective catalysis.

A similar study performed by these researchers has examined the catalytic, enantioselective hydrogenation of ketones with

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

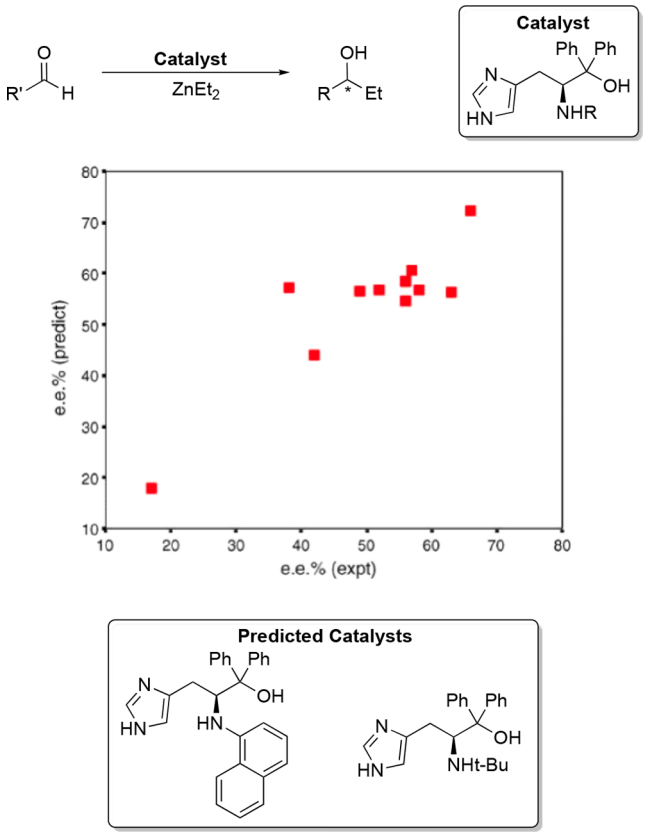


Figure 66. Enantioselective diethylzinc alkylation, case study 4 from ref 206. Adapted with permission from ref 206. Copyright 2006 Elsevier.

chiral ruthenium complexes (Figure 68).[218] Thirteen unsymmetrical benzophenone derivatives are subjected to Noyori enantioselective hydrogenation and the results are used to construct a PLS model correlating structure to reaction outcome. The descriptors employed are obtained using DRAGON software[219] with 3D-structures as input, then removing descriptors that are highly skewed. However, the descriptors actually included are not specified. Two models are constructed each with different relative conformers of *ortho*-substituted aromatic rings in the benzophenone starting material. A dependence on conformation is observed in which the first model underperforms with respect to the second model (average Q^2 = 0.58 and 0.66, respectively). The improved model is able to more accurately predict substrates with *ortho*-substituents by using a conformation with the substituent canted toward the ketone residue rather than away.

With improved models identified, numerous predictions are made for other *in silico* substrates. The predictions follow reasonable trends; substrates containing one *ortho*-substituted aromatic residue are the most selective followed by *meta*-substituted and *para*-substituted as the least selective. However, the predicted values are not validated experimentally.

Carnell and co-workers published the development of quantitative structure−property relationship (QSPR) models in the study of rhodium catalyzed, enantioselective conjugate addition of arylboronic acids to acyclic enones (Figure 69).[220] Eighteen different diene catalyst structures are synthesized and evaluated in the model system. Descriptors are calculated using DRAGON software from which an optimal three-component model is identified using a genetic algorithm for descriptor selection. At most, a maximum of three dimensions are used in the model to maximize model robustness by minimizing overfitting resulting from the inclusion of many parameters. The model constructed is validated using leave-one-out cross-validation (LOO), as well as 10-fold cross-validation and the bootstrap method (Q^2 = 0.70, 0.70, and 0.76, respectively). The specific descriptors used in the model are the total number of tertiary sp^3 carbon atoms in the catalyst structure, which is thought to reflect steric parameters, and MATS6i and MATS3m, which are 2D descriptors thought to reflect the electronic characteristics of the ligands.[221] The authors note that interpretation of the significance of these parameters is not straightforward and thus discerning mechanistic rational from the model is not possible. However, the model could be used predictively with substrates for which an ideal catalyst has not been identified.

6.3. Perturbation Theory QSSR

An interesting subfield of QSSR applied to enantioselective catalysis is perturbation theory QSSR (PT-QSSR) developed by González-Díaz and co-workers.[222] Interested readers are referred to the original work for the mathematical formulation of this theory. The significance of this method is the capability to predict multiple output efficiencies (e.g., yield and enantioselectivity) simultaneously, if desired. In this method, the goal is to predict the efficiency of a new chemical transformation with respect to a known chemical transformation. Thus, the new transformation can be treated as a perturbation of the original one (for example, a change in chemical structure). Using this workflow, general equations for chemical systems can be constructed relating input structures to performance. As an example of this workflow over 9000 predictions are made for the outcome of the enantioselective carbolithiation of olefins. However, to the best of our knowledge no predictions are validated experimentally. The reader is referred to the original work for a comprehensive list of predicted values.

A later example of this method is the analysis of enantioselective Heck−Heck cascade reactions.[223] In this work, literature data[224−231] is used to construct a model capable of predicting both yield and enantioselectivity outcomes of the Heck−Heck cascade reactions. Descriptors are calculated with DRAGON software: (1) substrates are described by their hydrophobic index and topological polar surface area (N and O contributions only), (2) products are described by logP, (3) bases are described by logP, (4) ligands are described by topological polar surface area (N, O, P, and S contributions), and (5) solvents are described by their dipole moment. Other nonstructural parameters are included by multiplying their values by the corresponding descriptors (for example, amount of base is multiplied by logP of base to obtain the value of one independent variable). These features are used to construct a PT-QSRR equation such that subsequent multivariate linear regression analysis is used to create a quantitative relationship to yield and selectivity. Multiple models are generated in this manner and are trained with 520 reactions the best of which had an R^2 of 0.79, Q^2 of 0.79, and standard error of 1.19% ee.

Using this model, the authors performed a simulated optimization for a nonoptimal reaction (Scheme 1). First, 2000 reaction conditions are screened computationally each with identical reaction components but different catalyst, base,

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

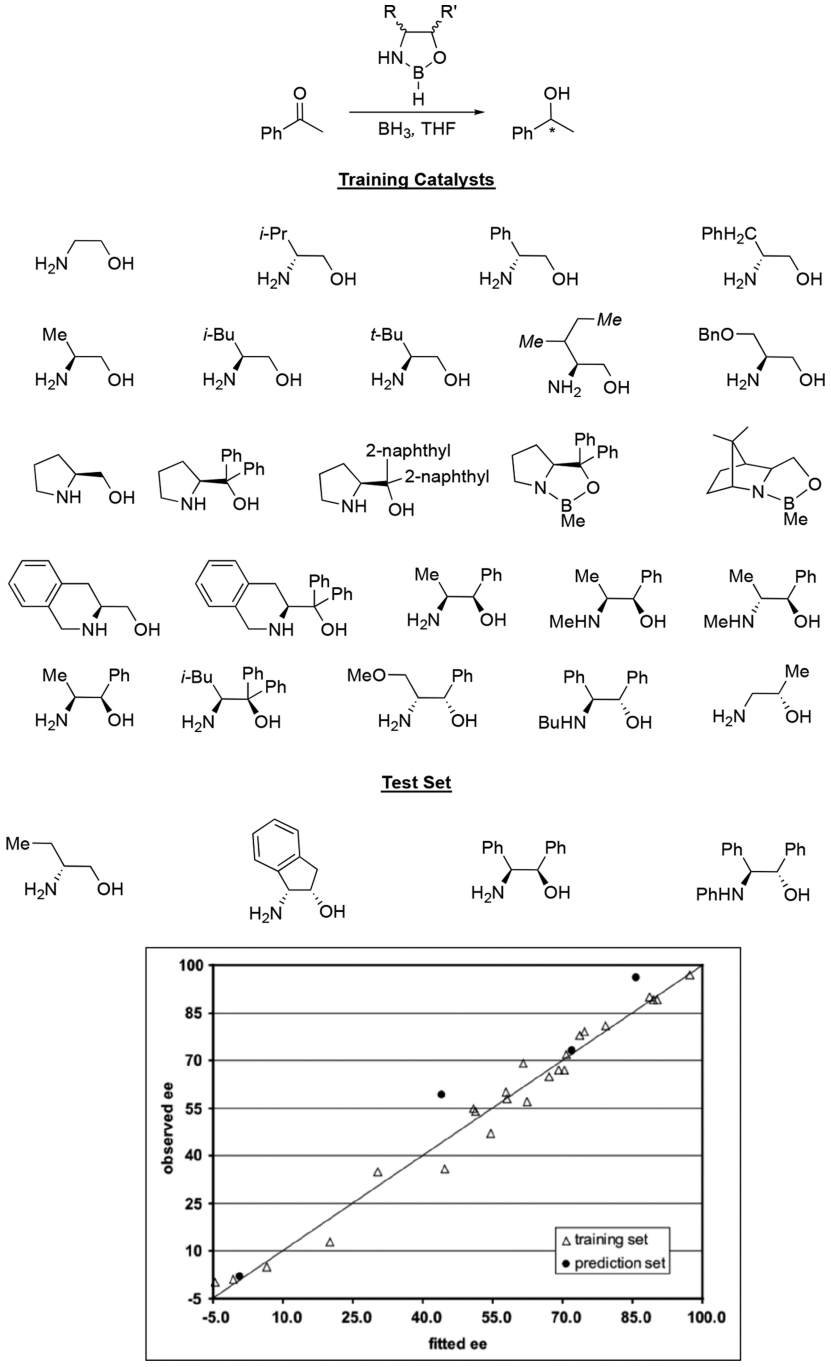
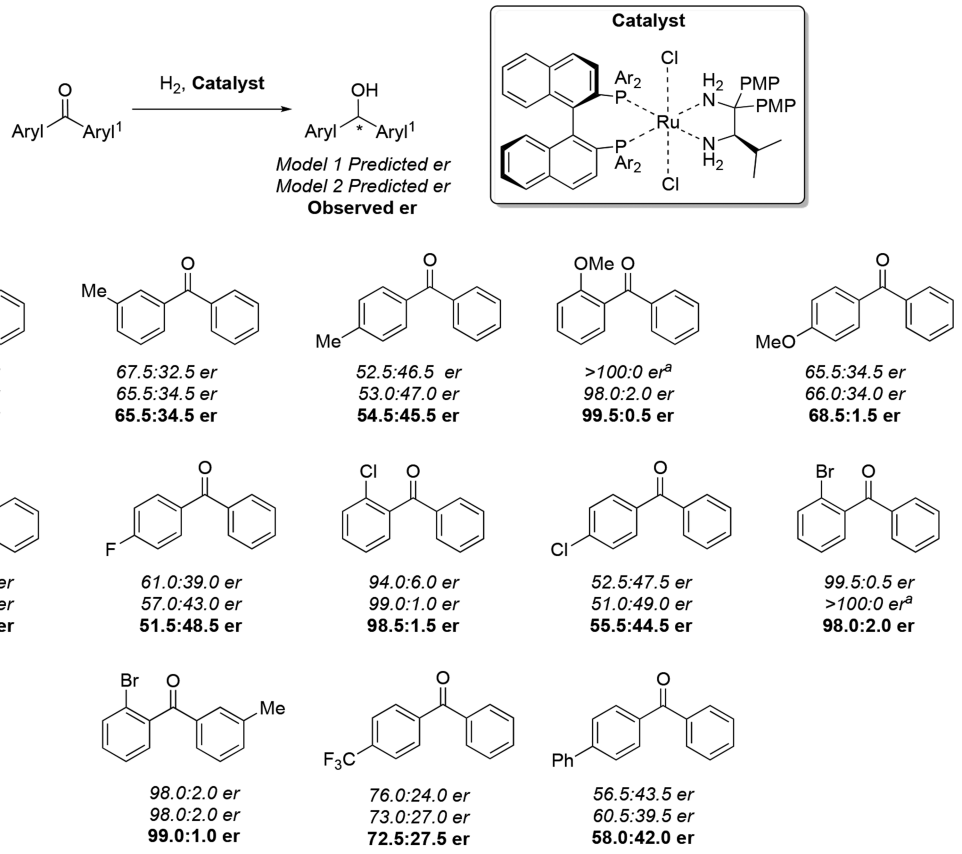


Figure 67. Reaction and catalysts studies in enantioselective acetophenone reduction. Adapted with permission from reference.[192] Copyright 2004 Elsevier.

and ligand loadings. A ternary diagram is constructed with this data revealing high, predicted selectivity in the region with 2.5−7.5 mol % of palladium catalyst, 7.5−20 mol % of ligand, and 2.5 to 7.5 equiv of base. It is a curious observation that the highest catalyst and ligand loadings (10 mol % and 30 mol %, respectively) are left out of this highest predicted range as one would not expect diminished performance with increased catalyst loading while maintaining the catalyst/ligand ratio. Ligand and substrate structures are then varied resulting in predicted selectivity values up to 100:0 er. Unfortunately, these predictions are not validated experimentally. It is unclear if these predictions are within the domain of the model; thus, further experimentation is needed to assess the validity of this approach.

PT-QSSR has also been applied to the chiral Brønsted acid catalyzed addition of enecarbamates to acyliminium ions (Scheme 2).[232] This reaction is performed with numerous chiral phosphoric acids and triflimide catalysts; BINOL-derived phosphoric acids catalysts yield the best enantioselectivity but BINOL-derived triflimides produce the highest yield. Further, unusual temperature effects are observed; improved enantioselectivity values are obtained around 40 °C with respect to the room temperature reaction. Descriptors are calculated for reactants, catalysts, and solvents using DRAGON software and multivariate models are constructed

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689



[a] Value from the regression output was over 100 % *ee*, which is interpreted only as a very selective reaction.

Figure 68. Enantioselective hydrogenation reaction studied in ref 218.

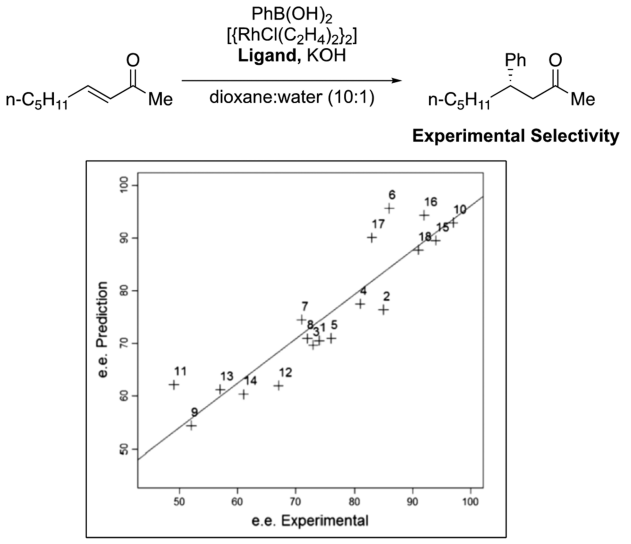
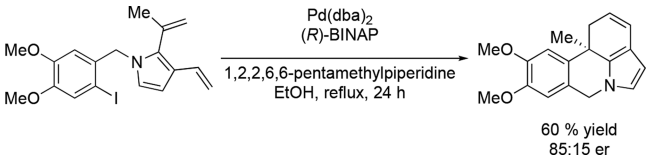


Figure 69. Enantioselective *β*-arylation reaction. Adapted with permission from ref 220. Copyright 2012 Royal Society of Chemistry.

Scheme 1. Enantioselective Heck–Heck Cascade Reaction



using STATISTICA[233] software. A significant multivariate regression model is constructed with descriptors for substrates, products, catalysts, nucleophiles, and solvent. Virtual screens are conducted varying the catalyst structure and the nucleophile structure to identify a suggestion of optimal reaction conditions for a target reactant. Though an intriguing approach, the suggested reaction is not experimentally validated.

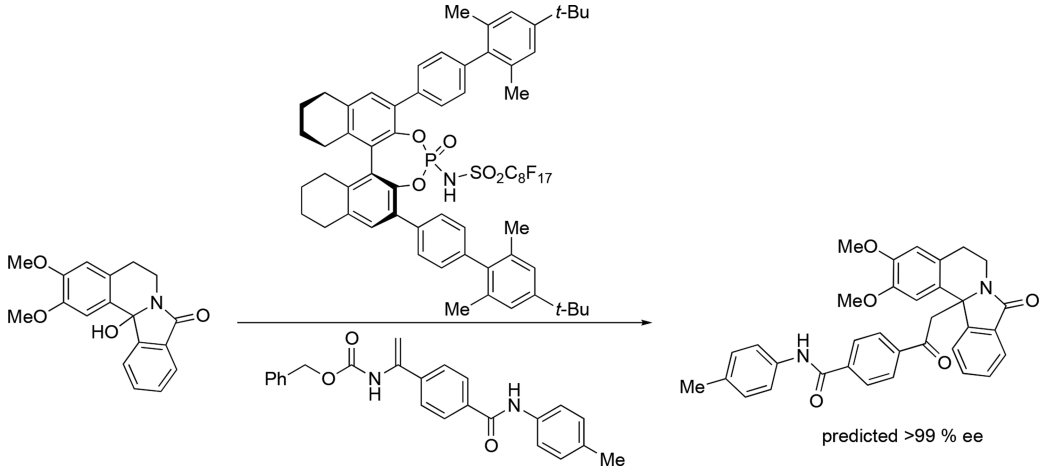
Although the capability of predicting both enantioselectivity and yield is exciting, rigorous validation of this approach has not been realized. For example, experimentally challenging the predictions with an external test set or experimentally testing the *in silico* optimized conditions is necessary to demonstrate the viability of this approach. This validation is especially important when predictions are made for structures distinct from those found in the training data, which likely constitute extrapolative predictions that may not be well described by the model.

6.4. Perspectives on 0–2D Descriptors in QSSR Applied to Enantioselective Catalysis

Descriptors not requiring a 3D molecular structure have multiple advantages over 3D descriptors. These advantages include rapid calculation and conformer independence. However, these rapidly calculable descriptors frequently do not contain the chemical information necessary for many applications, including for some systems in enantioselective catalysis.[234] Further, when correlations are constructed that are predictive it is likely that the model is simply recognizing patterns in features and relating them to experimental performance. One could argue that this approach is

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620–1689

Scheme 2. Chiral Bronsted Acid Catalyzed Addition of Enecarbamates to Acyliminium Ions



fundamentally different than having the model learn the underlying physics responsible for enantioinduction; the former does not allow for understandable models that can be used to form hypotheses about mechanism whereas the latter could help guide mechanistic investigation. Further, one would expect a model that has learned the underlying physics of a system to be more successful in domain adaptation (i.e., making predictions for novel substrates, extrapolating into novel catalyst space, etc.). Despite these limitations, multiple applications have demonstrated that these readily calculable descriptors can be used to make predictive models in certain situations and may be sufficient for specific applications.

6.5. Related Fields

Several areas of research related to using computational methods to assist catalyst design in enantioselective catalysis are beyond the scope of this review but warrant mention. Force field methods represent a class of prominent examples for which excellent summaries are already available.[235−238] These methods are particularly appealing because they allow a virtual screening of new catalysts and do not require generation of experimental data. However, the stereodetermining step of the transformation typically must be known. Similarly, modern computation and recent advances have enabled screening campaigns using quantum chemical methods that are reliable, proceed in a reasonable time frame, and are accessible to the general community.[239] Other areas of research include various mapping strategies, wherein molecular properties or structures are mapped to allow for a visual representation of important molecular influences, comparison between catalyst structures, and searches on the basis of similar molecular properties.[240−243] These methods are particularly useful for understanding stereoinduction and could likely be adapted in some way to merge with QSAR (e.g., derivation of new descriptors) but are typically not used to make quantitative predictions and are thus outside the scope of this review.

7. PERSPECTIVES

The dawn of the 21st century brought with it an increased interest in the application of statistical methods to catalyst design. In the last two decades, the field has matured beyond the proof-of-concept stage. Looking to the future, two primary goals can be identified to motivate the phase of development: (1) accurate, predictive identification of new, optimal catalysts, and (2) retrospective rationalization of structural features responsible for enantioinduction. Although examples to achieve these goals exist, most focus on the latter. The groundwork to achieve the former goal is accumulating but a true realization of predictive design remains an outstanding challenge. In our view, success in this endeavor must satisfy three conditions: (1) an unoptimized reaction needs to be selected, (2) available catalysts must affect these reactions in poor to moderate enantioselectivities, (3) a chemoinformatics model must enable identification of nontrivial determinants of selectivity from the catalyst performance data and predict new structures that affect high enantioselectivities, and (4) this method must be made accessible to nonexperts such that it can impact routine decision making processes encountered by bench chemists. A general demonstration of this strategy is a holy grail in the field and will lead to a revolutionary change in the way researchers approach catalyst optimization for asymmetric reactions.

One interesting observation regarding the application of chemoinformatics to predict more selective catalysts is the number of publications in which suggestions are made to achieve a more selective transformation compared to the number that are experimentally evaluated. For new methods to be accepted as tools for catalyst optimization they must demonstrate the capability to optimize a real system which in turn must be supported by experimentation. One method to facilitate experimental validation of new computational predictions would be to construct large data sets that can be used to benchmark new descriptor sets and modeling methods. This resource would alleviate the computational scientists' need for an experimental collaborator in addition to enabling experimentalists develop new computational tools that can be directly compared with the same data set. Of course, the interface between statistical analysis and enantioselective catalysis is an exciting opportunity for collaboration given the interdisciplinary nature of the field. With increasing interest, the emergence of tools tailored to optimization rather than interpretation can be expected in the immediate future.

Given the preliminary success of a variety of approaches toward making predictions of more selective processes, it seems likely that the adoption of such tools by those engaged in asymmetric catalysis is imminent. The need for expert knowledge for the implementation of chemoinformatics is perhaps the greatest limitation to its widespread adoption. However, given the recent work focused on making computa-

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

tionally guided methods more accessible to the bench chemist, optimization campaigns employing these methods will assist more widespread adoption.[21,61,238,243] Optimization strategies, however, are typically designed for implementation over a relatively small domain of chemical space. For example, they are often limited to a specific class of catalyst structures and rarely include reaction conditions. A greater challenge, and an arguably more significant advance, would be the ability to use statistical methods for comparison of vastly different catalyst structures. Engineering descriptors to capture the differences between such disparate structures would be a monumental task. The alternative approach of implementing modern deep learning methods may be a more a promising direction for the solution to this problem. However, the amount of data necessary to train the machine learning networks to optimize new systems could be prohibitive, particularly given the absence of low and moderately selective reactions from the literature. It is possible that using machine learning with high throughput screening methods could alleviate deficiencies in available data. Although challenging, solving such a problem now seems within the grasp of the community and would represent a leap forward for this field. It also follows then that the ability to propose new *in silico* structures on the basis of ideal chemical descriptors would be the pinnacle achievement in this endeavor. Such technology already finds precedent,[244] and adapting similar technology toward this goal has exciting implications for enantioselective catalysis.

The realm of garnering new mechanistic insights using statistical methods has already been established as a "launching pad" for generating new mechanistic hypotheses. As chemoinformatics tools become increasingly accessible, it is likely that such methods will find their way into the routine toolkit of synthetic chemists. However, as accessibility increases so must the necessity to scrutinize mechanistic claims made on the basis of these methods. Constructing a model with good predictive power in a particular domain does not guarantee that the model is interpretable. For example, confounding variables can conspire to confound the interpretation of models. Further, descriptor selection methods are not infallible; many types of descriptor selection methods exist many of which have not been widely implemented. For example, embedded methods (e.g., LASSO) have some precedent in catalysis but have not been widely adapted.[186,245] Random forest models have also been used recently to identify important features[137,246,247] as have neural networks.[145] Genetic algorithms are able to reduce the number of descriptors necessary to produce good models.[149] It is possible that using a collection of methods and comparing which descriptors are selected with each method could provide more trustworthy results. Unfortunately, the methods cited above are data intensive. In any case, mechanistic conclusions should not be made using only this information. Rather, such methods serve as a powerful construct to formulate mechanistic insights, which are then corroborated with complementary investigations.

The interpretability of a model is also dependent on the descriptors used to construct it. Subunit-derived, local descriptors are particularly appealing given the feasibility of direct interpretation. In contrast, global descriptors like CCM, chirality codes, and GRIND have more convoluted interpretations. Alignment-dependent grid descriptors have the benefit of a potentially straightforward interpretation, but the number of descriptors present impacts which modeling methods can be used and can consequently necessitate more data. However, subunit-derived descriptors are also theoretically more susceptible to the omitted variable bias[248,249] because it is unlikely that every molecular property that contributes to enantioinduction is adequately described. Whole-molecule global representations are less likely to be missing information and therefore less likely to erroneously assign the significance to another descriptor. However, this situation is only true if the global representation adequately represents the important structural features of interest. For example, CCM is a whole-molecule representation but also a single number—it is unlikely that this single number represents every feature responsible for enantioinduction and in some cases is indeed inferior to subunit derived descriptors.[137] It is possible that some methods such as alignment-dependent grid methods might give a more accurate representation of a molecule, but the high dimensionality of the molecular representations necessitate some method of dimensionality reduction, which thus omits information from the raw descriptors. The number of descriptors can also necessitate larger data sets (this problem can potentially be circumvented by preprocessing descriptors, but that discussion is outside the scope of this review).

Comparing the accuracy of models derived from different molecular parametrizations may suggest which descriptors are superior. However, a representation, just because it leads to more accurate models in a given situation, is not necessarily superior in terms of the interpretability of the model. The best representation, at least with current methods, is likely to be case dependent. Further, it is possible that the model could be "right for the wrong reasons", that is, the model could make accurate predictions within its domain but not actually be generally interpretable. Realizing domain adaptation could also be another way to assess if a model is founded in the underlying chemistry of a transformation which could give more credence to formulating mechanistic hypotheses. From the perspective of a chemist, the best approach is to validate the hypotheses with other methods. A related concept is that most of the aforementioned methods do not explicitly deal with the absolute configuration of the catalyst. Thus, the model is limited to predict values for only one enantiomer of the catalyst. This issue is not necessarily a problem, but it does pose a limitation of the method. Alternatively, some approaches that could explicitly treat the absolute configuration of the catalyst might not do so in practice. In these cases, models with an intercept in predicted versus observed plots will fail to give an equal but opposite magnitude of selectivity for enantiomeric catalyst pairs. Again, this problem is easily solved by considering only one enantiomer of the catalyst, but the model is then limited to making predictions only for that enantiomer. An alternative to engineering descriptors such as those mentioned above could be to use more advanced machine learning algorithms. It is possible that such advanced methods could achieve improved performance with simpler representations. However, as discussed above, this would likely also be much more data intensive.

The use of chemoinformatics to optimize and understand new enantioselective reactions has advanced from its infancy to adolescence in the past 20 years. This field holds tremendous promise for formalizing and extending the chemists' intuition to allow predictions of new, unknown catalyst structures. With increasing interest in the field, more advanced computational models, and more accessible protocols combined with high-

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

throughout data generation, we can confidently anticipate the transformation of this emergent field into standard operating procedures used by the practicing chemist in the not-too-distant future.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemrev.9b00425.

Conformer dependent chirality codes, description of counterpropagation network, additional information regarding topological descriptors (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: sdenmark@illinois.edu. Web: http://faculty.scs.illinois.edu/denmark/.

ORCID

Scott E. Denmark: 0000-0002-1099-9765

Author Contributions

A.F.Z. compiled the reference list and drafted the introduction, linear free energy relationships, continuous chirality measures, chirality codes, molecular interaction fields, miscellaneous, and perspectives sections. S.V.A. provided critical review and commentary, contributed to creating final drafts of every section, including major revisions of the linear free energy relationships section, and drafted the final version of most figures and the abstract. S.E.D. provided financial support, critical review and commentary, and edited text and graphics of the final version.

Notes

The authors declare no competing financial interest.

Biographies

Andrew F. Zahrt completed his B.S. degrees in chemistry and biology at Aquinas College in Grand Rapids, MI in 2014. In the same year, he commenced his Ph.D. studies at the University of Illinois Urbana–Champaign in the laboratory of Prof. Scott E. Denmark. Andrew's research involves the application of computational methods to problems in organic chemistry. Specifically, his research includes using applied quantum chemistry to garner mechanistic insights in catalytic reactions and developing strategies incorporating chemo-informatics and machine learning to optimize enantioselective catalysts.

Soumitra V. Athavale completed his B.S.-M.S. degrees in chemistry and biology from the Indian Institute of Science Education and Research (IISER) Pune in 2013, with research in chemical biology and protein engineering. In August 2014, he began Ph.D. studies at the University of Illinois Urbana–Champaign in the laboratory of Prof. Scott E. Denmark. Soumitra's research involved the mechanistic investigation of catalytic, asymmetric transformations, with a specific concentration on the autocatalytic Soai reaction. In Fall 2019, he joined Prof. Frances Arnold's group at the California Institute of Technology as a postdoctoral research associate.

Scott E. Denmark obtained an S.B. degree from MIT in 1975 (working with Richard H. Holm and Daniel S. Kemp) and his D. Sc. Tech. (under the direction of Albert Eschenmoser) from the ETH Zürich in 1980. That same year, he began his career at the University of Illinois. He was promoted to associate professor in 1986 and to full professor in 1987, and since 1991, he has been the Reynold C. Fuson Professor of Chemistry. His research interests include preparative and mechanistic aspects of synthetic organic chemistry. Professor Denmark was the Editor in Chief of *Organic Reactions* Volumes 71–100 and edited Volume 85 of *Organic Syntheses*. He served for six years as an Associate Editor of *Organic Letters* and for nine years as Editor of *Topics in Stereochemistry*. He is a Fellow of the Royal Society of Chemistry and was selected as an ACS Fellow in the inaugural year, 2009. He was elected to the American Academy of Arts and Sciences in 2017 and the National Academy of Sciences in 2018.

ACKNOWLEDGMENTS

We are grateful to the National Science Foundation for generous financial support (NSF CHE1900617). We thank Kenny Lipkowitz for providing Figure 27. We thank Jeremy Henle for useful discussion and providing Figure 54. A.F.Z. and S.V.A. thank the University of Illinois for graduate fellowships.

REFERENCES

(1) Lipkowitz, K.; Kozlowski, M. Understanding Stereoinduction in Catalysis via Computer: New Tools for Asymmetric Synthesis. *Synlett* **2003**, *10*, 1547–1565.

(2) Ahn, S.; Hong, M.; Sundararajan, M.; Ess, D. H.; Baik, M.-H. Design and Optimization of Catalysts Based on mechanistic Insights Derived from Quantum Chemical Reaction Modeling. *Chem. Rev.* **2019**, *119*, 6509–6560.

(3) Burello, E.; Rothenberg, G. *In Silico* Design in Homogeneous Catalysis Using Descriptor Modelling. *Int. J. Mol. Sci.* **2006**, *7*, 375–404.

(4) Reid, J. P.; Sigman, M. S. Comparing Quantitative Prediction Methods for the Discovery of Small-Molecule Chiral Catalysts. *Nat. Rev. Chem.* **2018**, *2*, 290–305.

(5) Cheong, P. H.-Y.; Legault, C. Y.; Um, J. M.; Çelebi-Ölçüm, N.; Houk, K. N. Quantum Mechanical Investigations of Organocatalysis: Mechanisms, Reactivities, and Selectivities. *Chem. Rev.* **2011**, *111*, 5042–5137.

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

(7) Peng, Q.; Paton, R. S. Catalytic Control in Cyclizations: From Computational Mechanistic Understanding to Selectivity Prediction. *Acc. Chem. Res.* **2016**, *49*, 1042–1052.

(8) Poree, C.; Schoenebeck, F. A. Holy Grail in Chemistry: Computational Catalyst Design: Feasible or Fiction. *Acc. Chem. Res.* **2017**, *50*, 605–608.

(9) Peng, Q.; Duarte, F.; Paton, R. S. Computing Organic Stereoselectivity – from Concepts to Quantitative Calculations and Predictions. *Chem. Soc. Rev.* **2016**, *45*, 6093–6107.

(10) Wheeler, S. E.; Seguin, T. J.; Guan, T.; Doney, A. C. Noncovalent Interactions in Organocatalysis and the Prospect of Computational Catalyst Design. *Acc. Chem. Res.* **2016**, *49*, 1061–1069.

(11) Tantillo, D. J. Faster, Catalyst! React! React! Exploiting Computational Chemistry for Catalyst Development and Design. *Acc. Chem. Res.* **2016**, *49*, 1079.

(12) Balcells, D.; Maseras, F. Computational Approaches to Asymmetric Synthesis. *New J. Chem.* **2007**, *31*, 333–343.

(13) Houk, K. N.; Cheong, P. H-Y. Computational Prediction of Small-Molecule Catalysts. *Nature* **2008**, *455*, 309–313.

(14) Fey, N.; Orpen, A. G.; Harvey, J. N. Building Ligand Knowledge Bases for Organometallic Chemistry: Computational Description of Phosphorus(III)-Donor Ligands and the Metal-Phosphorus Bond. *Coord. Chem. Rev.* **2009**, *253*, 704–722.

(15) Corbeil, C. R.; Moitessier, N. Theory and Application of Medium to High Throughput Prediction Method Techniques for

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620–1689

Asymmetric Catalyst Design. *J. Mol. Catal. A: Chem.* **2010**, *324*, 146−155.
(16) Fey, N. The Contribution of Computational Studies to Organometallic catalysis: Descriptors, Mechanisms and Models. *Dalton Trans* **2010**, *39*, 296−310.
(17) Maldonado, A. G.; Rothenberg, G. Predictive Modeling in Homogeneous Catalysis: a Tutorial. *Chem. Soc. Rev.* **2010**, *39*, 1891−1902.
(18) Neel, A. J.; Hilton, M. J.; Sigman, M. S.; Toste, F. D. Exploiting Non-Covalent π-Interactions for Catalyst Design. *Nature* **2017**, *543*, 637−646.
(19) Baskin, I. I.; Madzhidov, T. I.; Antipin, I. S.; Varnek, A. Artificial Intelligence in Synthetic Chemistry: Achievements and Prospects. *Russ. Chem. Rev.* **2017**, *86*, 1127−1156.
(20) Engkvist, O.; Norrby, P.-O.; Selmi, N.; Lam, Y.-H.; Peng, Z.; Sherer, E. C.; Amberg, W.; Erhard, T.; Smyth, L. A. Computational Prediction of Chemical Reactions: Current Status and Outlook. *Drug Discovery Today* **2018**, *23*, 1203−1218.
(21) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and Mechanistic Multivariate Linear Regression Models for Reaction Development. *Chem. Sci.* **2018**, *9*, 2398−2412.
(22) Durand, D. J.; Fey, N. Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019**, *119*, 6561−6594.
(23) Verma, J.; Khedkar, V. M.; Coutinho, E. C. 3D-QSAR in Drug Design − a Review. *Curr. Top. Med. Chem.* **2010**, *10*, 95−115.
(24) Roy, K.; Kar, S.; Das, R. N. *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*; Academic Press: London, 2016.
(25) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics: Vol. 1: Alphabetical Listing/Vol. II: Appendices, References*; Mannhold, R., Kubinyi, H., Folkers, G., Eds.; WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim, 2009; Vol. *41*.
(26) Roy, K. *Advances in QSAR Modeling. Applications in Pharmaceutical, Chemical, Food, Agricultural, and Environmental Sciences*; Springer International Publishing: Cham, 2019.
(27) Goldman, B. B.; Walters, W. P. Chapter 8 Machine Learning in Computational Chemistry. In *Annual Reports in Computational Chemistry*; Spellmeyer, D. C., Ed.; Elsevier, 2006; Vol. *2*, pp 127−140.
(28) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep Learning for Molecular Design − a Review of the State of the Art. *Mol. Sys. Des. Eng.* **2019**, *4*, 828−849.
(29) Geurts, P.; Irrthum, A.; Wehenkel, L. Supervised Learning with Decision Tree-Based Methods in Computational and Systems Biology. *Mol. BioSyst.* **2009**, *5*, 1593−1605.
(30) Ivanciuc, O. Applications of Support Vector Machines in Chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Cundari, T. R. Eds.; Wiley-VCH: Weinheim, 2007; Vol. *23*, pp 291−400.
(31) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545−2559.
(32) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, 2016.
(33) Harrell, F. E. *Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression, and Survival Analysis*; Springer: New York, 2001.
(34) Konishi, S.; Kitagawa, G. *Information Criteria and Statistical Modeling*; Springer: New York, 2008.
(35) Maitra, S.; Yan, J. Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression. *Appl. Multivar. Stat. Model* **2008**, *79*, 79−90.
(36) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.
(37) Hansch, C.; Leo, A.; Taft, R. W. A Survey of Hammett Substituent Constants and Resonance and Field Parameters. *Chem. Rev.* **1991**, *91*, 165−195.
(38) Hammett, H. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96−103.
(39) Taft, R. W. Linear Free Energy Relationships from Rates of Esterification and Hydrolysis of Aliphatic and Ortho-substituted Benzoate Esters. *J. Am. Chem. Soc.* **1952**, *74*, 2729−2732.
(40) Taft, R. W. Polar and Steric Substituent Constants for Aliphatics and o-Benzoate Groups from Rates of Esterification and Hydrolysis of Esters. *J. Am. Chem. Soc.* **1952**, *74*, 3120−3128.
(41) Taft, R. W. Linear Steric Energy Relationships. *J. Am. Chem. Soc.* **1953**, *75*, 4538−4539.
(42) Charton, M. Steric Effes. I. Esterification and Acid-Catalyzed Lysis of Esters. *J. Am. Chem. Soc.* **1975**, *97*, 1552−1556.
(43) Charton, M. Steric Effects. II. Base-catalyzed Ester Hydrolysis. *J. Am. Chem. Soc.* **1975**, *97*, 3691−3693.
(44) Charton, M. Steric Effects. III. Bimolecular Nucleophilic Substitution. *J. Am. Chem. Soc.* **1975**, *97*, 3694−3697.
(45) Charton, M. Steric Effects. IV. E1 and E2 Eliminations. *J. Am. Chem. Soc.* **1975**, *97*, 6159−6161.
(46) Charton, M.; Charton, B. Steric Effects. V. Barriers to Internal Rotation. *J. Am. Chem. Soc.* **1975**, *97*, 6472−9473.
(47) Charton, M. Steric Effects. 6. Hydrolysis of Amides and Related Compounds. *J. Org. Chem.* **1976**, *41*, 2906−2910.
(48) Charton, M. Steric Effects. 7. Additional V Constants. *J. Org. Chem.* **1976**, *41*, 2217−2220.
(49) Charton, M. Steric Effects. 8. Racemization of Chiral Biphenyls. *J. Org. Chem.* **1977**, *42*, 2528−2529.
(50) Charton, M. Steric Effects. 9. Substituents at Oxygen in Carbonyl Compounds. *J. Org. Chem.* **1977**, *42*, 3531−3535.
(51) Charton, M. Steric Effects. 10. Substituents at Nitrogen in Carbonyl Compounds. *J. Org. Chem.* **1977**, *42*, 3535−3538.
(52) Charton, M.; Charton, B. Steric Effects. 11. Substituents at Sulfur. *J. Org. Chem.* **1978**, *43*, 1161−1165.
(53) Charton, M.; Charton, B. Steric Effects. 12. Substituents at Phosphorus. *J. Org. Chem.* **1978**, *43*, 2383−2386.
(54) Charton, M. Steric effects. 13. Substituents at Sulfur. *J. Org. Chem.* **1978**, *43*, 3995−4001.
(55) Bess, E. N.; Sigman, M. S. *Asymmetric Synthesis*; Christmann, M., Bräse, S., Eds.; Wiley-VCH Verlag & Co. KGaA: Weinheim, Germany, 2012; pp 363−370.
(56) Harper, K. C.; Sigman, M. S. Using Physical Organic Parameters To Correlate Asymmetric Catalyst Performance. *J. Org. Chem.* **2013**, *78*, 2813−2818.
(57) Toste, F. D.; Sigman, M. S.; Miller, S. J. Pursuit of Noncovalent Interactions for Strategic Site-Selective Catalysis. *Acc. Chem. Res.* **2017**, *50*, 609−615.
(58) Sigman, M. S.; Harper, K. C.; Bess, E. N.; Milo, A. The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Acc. Chem. Res.* **2016**, *49*, 1292−1301.
(59) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and Mechanistic Multivariate Linear Regression Models for Reaction Development. *Chem. Sci.* **2018**, *9*, 2398−2412.
(60) Harper, K. C.; Bess, E. N.; Sigman, M. S. Multidimensional Steric Parameters in the Analysis of Asymmetric Catalytic Reactions. *Nat. Chem.* **2012**, *4*, 366−374.
(61) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Sterimol Steric Parameters. *ACS Catal.* **2019**, *9*, 2313−2323.
(62) Verloop, A.; Hoogenstraten, W.; Tipker, A. Development and application of new steric parameters in drug design. In *Drug Design*; Ariens, E.J., Ed.; Academic Press: New York, 1976; Vol. *7*, pp 165−207.
(63) Miller, J. J.; Sigman, M. S. Quantitatively Correlating the Effect of Ligand-Substituent Size in Asymmetric Catalysis Using Linear Free Energy Relationships. *Angew. Chem., Int. Ed.* **2008**, *47*, 771−774.
(64) Von Matt, P.; Pfaltz, A. Chiral Phosphinoaryldihydrooxazoles as Ligands in Asymmetric Catalysis: Pd-Catalyzed Allylic Substitution. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 566−568.
(65) Denmark, S. E.; Christenson, B. L.; O'Connor, S. P. Catalytic Enantioselective Cyclopropanation with Bis(halomethyl)zinc Reagents. II. The Effect of Promoter Structure on Selectivity. *Tetrahedron Lett.* **1995**, *36*, 2219−2222.

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

(66) Minakata, S.; Ando, T.; Nishimura, M.; Ryu, I.; Komatsu, M. Novel Asymmetric and Stereospecific Aziridination of Alkenes with a Chiral Nitridomanganese Complex. *Angew. Chem., Int. Ed.* **1998**, *37*, 3392−3394.
(67) Park, J.; Quan, Z.; Lee, S.; Han Ahn, K.; Cho, C.-W. Synthesis of Chiral 1'-Substituted Oxazolinylferrocenes as Chiral Ligands for Pd-Catalyzed Allylic Substitution Reactions. *J. Organomet. Chem.* **1999**, *584*, 140−146.
(68) Ito, Y. N.; Zriza, X.; Beck, A. K.; Bohac, A.; Ganter, C.; Gawley, R. E.; Kuehnle, F. N. M.; Tuleja, J.; Wang, Y. M.; Seebach, D. Preparation and Structural Analysis of Several New α,α,α',α'-Tetraaryl-1,3-dioxolane-4,5-dimethanols (TADDOL's) and TADDOL Analogs, their Evaluation as Titanium Ligands in the Enantioselective Addition of Methyltitanium and Diethylzinc Reagents to Benzaldehyde, and Refinement of the Mechanistic Hypothesis. *Helv. Chim. Acta* **1994**, *77*, 2071−2110.
(69) Wu, H.-L.; Uang, B.-J. Asymmetric Epoxidation of Allylic Alcohols Catalyzed by New Chiral Vanadium(V) Complexes. *Tetrahedron: Asymmetry* **2002**, *13*, 2625−2628.
(70) Quintard, A.; Alexakis, A. 1,2-Sulfone Rearrangement in Organocatalytic Reactions. *Org. Biomol. Chem.* **2011**, *9*, 1407−1418.
(71) Sigman, M. S.; Miller, J. J. Examination of the Role of Taft-Type Steric Parameters in Asymmetric Catalysis. *J. Org. Chem.* **2009**, *74*, 7633−7643.
(72) Gustafson, J. L.; Sigman, M. S.; Miller, S. J. Linear Free-Energy Relationship Analysis of a Catalytic Desymmetrization Reaction of a Diarylmethane-bis(phenol). *Org. Lett.* **2010**, *12*, 2794−2797.
(73) Lewis, C. A.; Chiu, A.; Kubryk, M.; Balsells, J.; Pollard, D.; Esser, C. K.; Murry, J.; Reamer, R. A.; Hansen, K. B.; Miller, S. J. Remote Desymmetrization at near-Nanometer Group Separation Catalyzed by a Miniaturized Enzyme Mimic. *J. Am. Chem. Soc.* **2006**, *128*, 16454−16455.
(74) Derived from the energy differential between equatorial and axial conformers of substituted cyclohexanes: Winstein, S.; Holness, N. J. Neighboring Carbon and Hydrogen. XIX. *t*-Butylcyclohexyl Derivatives. Quantitative Conformational Analysis. *J. Am. Chem. Soc.* **1955**, *77*, 5562−5578.
(75) Derived from the effect *ortho*-substituents have on the barrier to rotation in an engineered biaryl system: Bott, G.; Field, L. D.; Sternhell, S. Steric effects. A Study of a Rationally Designed System. *J. Am. Chem. Soc.* **1980**, *102*, 5618−5626.
(76) Knowles, R. R.; Lin, S.; Jacobsen, E. N. Enantioselective Thiourea-Catalyzed Cationic Polycyclizations. *J. Am. Chem. Soc.* **2010**, *132*, 5030−5032.
(77) Knowles, R. R.; Jacobsen, E. N. Attractive Noncovalent Interactions in Asymmetric Catalysis: Links Between Enzymes and Small Molecule Catalysts. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 20678−20685.
(78) Zuend, S. J.; Jacobsen, E. N. Mechanism of Amido-Thiourea Catalyzed Enantioselective Imine Hydrocyanation: Transition State Stabilization via Multiple Non-Covalent Interactions. *J. Am. Chem. Soc.* **2009**, *131*, 15358−15374.
(79) Brandt, P.; Roth, P.; Andersson, P. G. Origin of Enantioselectivity in the Ru(arene)(amino alcohol)-Catalyzed Transfer Hydrogenation of Ketones. *J. Org. Chem.* **2004**, *69*, 4885−4890.
(80) Oslob, J. D.; Åkermark, B.; Helquist, P.; Norrby, P.-A. Steric Influences on the Selectivity in Palladium-Catalyzed Allylation. *Organometallics* **1997**, *16*, 3015−3021.
(81) Harper, K. C.; Sigman, M. S. Predicting and Optimizing Asymmetric Catalyst Performance using the Principles of Experimental Design and Steric Parameters. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 1279−1283.
(82) Harper, K. C.; Sigman, M. S. Three-Dimensional Correlation of Steric and Electronic Free Energy Relationships Guides Asymmetric Propargylation. *Science* **2011**, *333*, 1875−1878.
(83) Werner, E. W.; Mei, T.-S.; Burckle, A. J.; Sigman, M. S. Enantioselective Heck Arylations of Acyclic Alkenyl Alcohols Using a Redox-Relay Strategy. *Science* **2012**, *338*, 1455−1458.
(84) Huang, H.; Zong, H.; Bian, G.; Song, L. Constructing a Quantitative Correlation between N-Sbstituent Sizes of Chiral Ligands and Enantioselectivities in Asymmetric Addition Reactions of Diethylzinc with Benzaldehyde. *J. Org. Chem.* **2012**, *77*, 10427−10434.
(85) Huang, H.; Zong, H.; Shen, B.; Yue, H.; Bian, G.; Song, L. QSAR Analysis of the Catalytic Asymmetric Ethylation of Ketone Using Physical Steric Parameters of Chiral Ligand Substituents. *Tetrahedron* **2014**, *70*, 1289−1297.
(86) Harper, K. C.; Vilardi, S. C.; Sigman, M. S. Prediction of Catalyst and Substrate Performance in the Enantioselective Propargylation of Aliphatic Ketones by a Multidimensional Model of Steric Effects. *J. Am. Chem. Soc.* **2013**, *135*, 2482−2485.
(87) Huang, H.; Zong, H.; Bian, G.; Yue, J.; Song, L. Correlating the Effects of the N-Substituent Sizes of Chiral 1,2-Amino Phosphinamide Ligands on Enantioselectivities in Catalytic Asymmetric Henry Reaction Using Physical Steric Parameters. *J. Org. Chem.* **2014**, *79*, 9455−9464.
(88) Akhani, R. K.; Moore, M. I.; Pribyl, J. G.; Wiskur, S. L. Linear Free-Energy Relationship and Rate Study on a Silylation-Based Kinetic Resolution: Mechanistic Insights. *J. Org. Chem.* **2014**, *79*, 2384−2396.
(89) Milo, A.; Bess, E. N.; Sigman, M. S. Interrogating Selectivity in Catalysis using Molecular Vibrations. *Nature* **2014**, *507*, 210−2014.
(90) Bess, E. N.; Bischoff, A. J.; Sigman, M. S. Designer Substrate Library for Quantitative, Predictive Modelling of Reaction Performance. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 14698−14703.
(91) Zhang, C.; Santiago, C. B.; Crawford, J. M.; Sigman, M. S. Enantioselective Dehydrogenative Heck Arylations of Trisubstituted Alkenes with Indoles to Construct Quaternary Stereocenters. *J. Am. Chem. Soc.* **2015**, *137*, 15668−15671.
(92) Milo, A.; Neel, A.; Toste, D. F.; Sigman, M. S. A Data-Driven Approach to Mechanistic Elucidation in Chiral Anion Catalysis. *Science* **2015**, *347*, 737−743.
(93) Orlandi, M.; Coelho, J. A. S.; Hilton, M. J.; Toste, F. D.; Sigman, M. S. Parameterization of Non-covalent Interactions for Transition State Interrogation Applied to Asymmetric Catalysis. *J. Am. Chem. Soc.* **2017**, *139*, 6803−6806.
(94) Birman, V. B.; Uffman, E. W.; Jiang, H.; Li, X.; Kilbane, C. J. 2,3-Dihydroimidazo[1,2-a]pyridines: A New Class of Enantioselective Acyl Transfer Catalysts and Their Use in Kinetic Resolution of Alcohols. *J. Am. Chem. Soc.* **2004**, *126*, 12226−12227.
(95) Birman, V. B.; Jiang, H. Kinetic Resolution of Alcohols Using a 1,2-Dihydroimidazo[1,2-a]quinoline Enantioselective Acylation Catalyst. *Org. Lett.* **2005**, *7*, 3445−3447.
(96) Park, Y.; Niemeyer, Z. L.; Yu, J.-Q.; Sigman, M. S. Quantifying Structural Effects of Amino Acid Ligands in Pd(II)-Catalyzed Enantioselective C-H Functionalization. *Organometallics* **2018**, *37*, 203−210.
(97) Biswas, S.; Kubota, K.; Orlandi, M.; Turberg, M.; Miles, D. H.; Sigman, M. S.; Toste, D. F. Enantioselective Synthesis of N,S-Acetals by an Oxidative Pummerer-Type Transformation using Phase-Transfer Catalysis. *Angew. Chem., Int. Ed.* **2018**, *57*, 589−593.
(98) Wang, Y.; Zhou, H.; Yang, K.; You, C.; Zhang, L.; Luo, S. Steric Effect of Protonated Tertiary Amine in Primary-Tertiary Diamine Catalysis: A Double-Layered Sterimol Model. *Org. Lett.* **2019**, *21*, 407−411.
(99) Dhayalan, V.; Gadekar, S. C.; Alassad, Z.; Milo, A. Unravelling Mechanistic Features of Organocatalysis with in situ Modifications at the Secondary Sphere. *Nat. Chem.* **2019**, *11*, 543−551.
(100) Aguado-Ullate, S.; Urbano-Cuadrado, M.; Villalba, I.; Pires, E.; García, J. I.; Bo, C.; Carbó, J. J. Predicting the Enantioselectivity of the Copper-Catalysed Cyclopropanation of Alkenes by Using Quantitative Quadrant-Diagram Representations of the Catalysts. *Chem. - Eur. J.* **2012**, *18*, 14026−14036.
(101) Aguado-Ullate, S.; Saureu, S.; Guasch, L.; Carbó, J. J. Theoretical Studies of Asymmetric Hydroformylation Using the Rh-(R,S)-BINAPHOS Catalysts − Origin of Coordination Preferences and Stereoinduction. *Chem. - Eur. J.* **2012**, *18*, 995−1005.

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

(102) Fraile, J. M.; García, J. I.; Gissibl, A.; Mayoral, J. A.; Pires, E.; Reiser, O.; Roldán, M.; Villalba, I. C1-Symmetric Versus C2-Symmetric Ligands in Enantioselective Copper-Bis(oxazoline)-Catalyzed Cyclopropanation Reactions. *Chem. - Eur. J.* **2007**, *13*, 8830−8839.
(103) García, J. I.; López-Sánchez, B.; Mayoral, J. A.; Pires, E.; Villalba, I. Surface Confinement Effects in Enantioselective Catalysis: Design of New Heterogeneous Chiral Catalysts based on C1-Symmetric Bisoxazolines and their Application in Cyclopropanation Reactions. *J. Catal.* **2008**, *258*, 378−385.
(104) García, J. I.; Jiménez-Osés, G.; López-Sánchez, B.; Mayoral, J. A.; Vélez, A. Stereoselectivity Induced by Support Confinement Effects. Aza-pyridinoxazolines: A new Family of C1-symmetric Ligands for Copper-Catalyzed Enantioselective Cyclopropanation Reactions. *Dalton Trans* **2010**, *39*, 2098−2107.
(105) Yang, C.; Zhang, E.-G.; Li, X.; Chen, J.-P. Asymmetric Conjugate Addition of Benzofuran-2-ones to Alkyl 2-Phthalimidoacrylates: Modeling Structure-Stereoselectivity Relationships with Steric and Electronic Parameters. *Angew. Chem., Int. Ed.* **2016**, *55*, 6506−6510.
(106) Weldy, N. M.; Schafer, A. G.; Owens, C. P.; Herting, C. J.; Varela-Alvarez, A.; Chen, S.; Niemeyer, Z.; Musaev, D. G.; Sigman, M. S.; Davies, H. M. L.; Blakey, S. B. Iridium(III)-bis(imidazolinyl)-phenyl Catalysts for Enantioselective C-H Functionalization with Ethyl Diazoacetate. *Chem. Sci.* **2016**, *7*, 3142−3146.
(107) Chen, Z.-M.; Hilton, M. J.; Sigman, M. S. Palladium-Catalyzed Enantioselective Redox-Relay Heck Arylation of 1,1-Disubstituted Homoallylic Alcohols. *J. Am. Chem. Soc.* **2016**, *138*, 11461−11464.
(108) Neel, A. J.; Milo, A.; Sigman, M. S.; Toste, F. D. Enantiodivergent Fluorination of Allylic Alcohols: Data Set Design Reveals Structural Interplay between Achiral Directing Group and Chiral Anion. *J. Am. Chem. Soc.* **2016**, *138*, 3863−3875.
(109) Yamamoto, E.; Hilton, M. J.; Orlandi, M.; Saini, V.; Toste, F. D.; Sigman, M. S. Development and Analysis of a Pd(0)-Catalyzed Enantioselective 1,1-Diarylation of Acrylates Enabled by Chiral Anion Phase Transfer. *J. Am. Chem. Soc.* **2016**, *138*, 15877−15880.
(110) Woods, B. P.; Orlandi, M.; Huang, C.-Y.; Sigman, M. S.; Doyle, A. G. Nickel-Catalyzed Enantioselective Reductive Cross-Coupling of Styrenyl Aziridines. *J. Am. Chem. Soc.* **2017**, *139*, 5688−5691.
(111) Ardkhean, R.; Mortimore, M.; Paton, R. S.; Fletcher, S. P. Formation of Quaternary Centres by Copper Catalyzed Asymmetric Conjugate Addition to β-Substituted Cyclopentenones with the aid of a Quantitative Structure-Selectivity Relationship. *Chem. Sci.* **2018**, *9*, 2628−2632.
(112) Niemeyer, Z. L.; Pindi, S.; Khrakovsky, D. A.; Kuzniewski, C. N.; Hong, C. M.; Joyce, L. A.; Sigman, M. S.; Toste, F. D. Parameterization of Acyclic Diaminocarbene Ligands Applied to a Gold(I)-Catalyzed Enantioselective Tandem Rearrangement/Cyclization. *J. Am. Chem. Soc.* **2017**, *139*, 12943−12946.
(113) Crawford, J.; Stone, E. A.; Metrano, A. J.; Miller, S. J.; Sigman, M. S. Parameterization and Analysis of Peptide-Based Catalysts for the Atroposelective Bromination of 3-Arylquinazolin-4(3H)-ones. *J. Am. Chem. Soc.* **2018**, *140*, 868−871.
(114) Yang, C.; Wang, J.; Liu, Y.; Ni, X.; Li, X.; Cheng, J.-P. Study on the Catalytic Behavior of Bifunctional Hydrogen-Bonding Catalysts Guided by Free Energy Relationship Analysis of Steric Parameters. *Chem. - Eur. J.* **2017**, *23*, 5488−5497.
(115) Coelho, J. A.; Matsumoto, A.; Orlandi, M.; Hilton, H. J.; Sigman, M. S.; Toste, F. D. Enantioselective Fluorination of Homoallylic Alcohols Enabled by the Tuning of Non-Covalent Interaction. *Chem. Sci.* **2018**, *9*, 7153−7158.
(116) Metsänen, T. T.; Lexa, K. W.; Santiago, C. B.; Chung, C. K.; Xu, Y.; Liu, Z.; Humphrey, G. R.; Ruck, R. T.; Sherer, E. C.; Sigman, M. S. Combining Traditional 2D and Modern Physical Organic-Derived Descriptors to Predict Enhanced Enantioselectivity for the Key Aza-Michael Conjugate Addition in the Synthesis of PrevymisTM (letermovir). *Chem. Sci.* **2018**, *9*, 6922−6927.
(117) Li, S.-L.; Yang, C.; Wu, Q.; Zheng, H.-L.; Li, X.; Cheng, J.-P. Atroposelective Catalytic Asymmetric Allylic Alkylation Reaction for Axially Chiral Anilides with Achiral Morita-Baylis-Hillman Carbonates. *J. Am. Chem. Soc.* **2018**, *140*, 12836−12843.
(118) Kwon, Y.; Li, J.; Reid, J. P.; Crawford, J. M.; Jacob, R.; Sigman, M. S.; Toste, F. D.; Miller, S. J. Disparate Catalytic Scaffolds for Atroposelective Cyclodehydration. *J. Am. Chem. Soc.* **2019**, *141*, 6698−6705.
(119) Zabrodsky, H.; Peleg, S.; Avnir, D. Continuous Symmetry Measures. *J. Am. Chem. Soc.* **1992**, *114*, 7843−7851.
(120) Zabrodsky, H.; Peleg, S.; Avnir, D. Continuous Symmetry Measures. 2. Symmetry groups and the tetrahedron. *J. Am. Chem. Soc.* **1993**, *11*, 8278−8289.
(121) Zabrodsky, H.; Avnir, D. Measuring Symmetry in Structural Chemistry. *Adv. Mol. Struct. Res.* **1995**, *1*, 1−34.
(122) Zabrodsky, H.; Avnir, D. Continuous Symmetry Measures. 4. Chirality. *J. Am. Chem. Soc.* **1995**, *117*, 462−473.
(123) Gao, D.; Schefzick, S.; Lipkowitz, K. Relationship between Chirality Content and Stereoinduction: Identification of a Chiraphore. *J. Am. Chem. Soc.* **1999**, *121*, 9481−9482.
(124) Harada, T.; Takeuchi, M.; Hatsuda, M.; Ueda, H.; Oku, A. Effects of Torsional Angles of 2,2'-Biaryldiol Ligands in Asymmetric Diels-Alder Reaction of Acrylates Catalyzed by their Titanium Complexes. *Tetrahedron: Asymmetry* **1996**, *7*, 2479−2482.
(125) Lipkowitz, K.; Schefziek, S.; Avnir, D. Enhancement of Enantiomeric Excess by Ligand Distortion. *J. Am. Chem. Soc.* **2001**, *123*, 6710−6711.
(126) Alvarez, S.; Schefzick, S.; Lipkowitz, K.; Avnir, D. Quantitative Chirality Analysis of Molecular Subunits of Bis(oxazoline)copper(II) Complexes in Relation to Their Enantioselective Catalytic Activity. *Chem. - Eur. J.* **2003**, *9*, 5832−5837.
(127) It is worth clarifying that the authors do not claim their analysis has implications beyond the reaction in the study.
(128) Lipkowitz, K.; Schefzick, S. Ligand Distortion Modes Leading to Increased Chirality Content of Katsuki-Jacobsen Catalyst. *Chirality* **2002**, *14*, 677−682.
(129) Zhang, W.; Loeback, J. L.; Wilson, S. R.; Jacobsen, E. N. Enantioselective Epoxidation of Unfunctionalized Olefin Catalyzed by Salen Manganese Complexes. *J. Am. Chem. Soc.* **1990**, *112*, 2801−2803.
(130) Irie, R.; Noda, K.; Ito, Y.; Matsumoto, N.; Katsuki, T. Catalytic Asymmetric Epoxidation of Unfunctionalized Olefins. *Tetrahedron Lett.* **1990**, *31*, 7345−7348.
(131) Strassner, T.; Houk, K. N. Predictions of Geometries and Multiplicities of the Manganese-Oxo Intermediates in the Jacobsen Epoxidation. *Org. Lett.* **1999**, *1*, 419−422.
(132) El-Bahraoui, J.; Wiest, O.; Feichtinger, D.; Plattner, D. A. Rate Enhancement and Enantioselectivity of the Jacobsen-Katsuki Epoxidation: The Significance of the Sixth Corrdination Site. *Angew. Chem., Int. Ed.* **2001**, *40*, 2073−2076.
(133) Bellarosa, L.; Zerbetto, F. Enantiomeric Excesses and Electronic Chirality Measure. *J. Am. Chem. Soc.* **2003**, *125*, 1975−1979.
(134) Lipkowitz, K.; Gao, D.; Katzenelson, O. Computation of Physical Chirality: An Assessment of Orbital Desymmetrization Induced by Common Chiral Auxiliaries. *J. Am. Chem. Soc.* **1999**, *121*, 5559−5564.
(135) Grimme, S. Continuous Symmetry Measures for Electronic Wavefunctions. *Chem. Phys. Lett.* **1998**, *297*, 15−22.
(136) Handgraaf, J.-W.; Reek, J. N. H.; Bellarosa, L.; Zerbetto, F. Continuous Chirality Measure in Reaction Pathways of Ruthenium-Catalyzed Transfer Hydrogenation of Ketones. *Adv. Synth. Catal.* **2005**, *347*, 792−802.
(137) Zahrt, A. F.; Denmark, S. E. Evaluating Continuous Chirality Measure as a 3D Descriptor in Chemoinformatics Applied to Asymmetric Catalysis. *Tetrahedron* **2019**, *75*, 1841−1851.
(138) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363*, eaau5631.
(139) Zayit, A.; Pinsky, M.; Elgavi, H.; Dryzun, C.; Avnir, D. A Web Site for Calculating the Degree of Chirality. *Chirality* **2011**, *23*, 17−23.
(140) Aires-de-Sousa, J.; Gasteiger, J. New Description of Molecular Chirality and Its Application to the Prediction of the Preferred Enantiomer in Stereoselective Reactions. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 369−375.
(141) Aires-de-Sousa, J.; Gasteiger, J. Prediction of Enantiomeric Selectivity in Chromatography: Application of Conformation-Dependent and Conformation-Independed Descriptors of Molecular Chirality. *J. Mol. Graphics Modell.* **2002**, *20*, 373−388.
(142) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. The Prediction of the 3D Structure of Organic Molecules from Their infrared Spectra. *Vib. Spectrosc.* **1999**, *19*, 151−164.
(143) Another example of classification not covered in this review: Zhang, Q.-Y.; Aires-de-Sousa, J. Physicochemical Stereodescriptors of Atomic Chiral Centers. *J. Chem. Inf. Model.* **2006**, *46*, 2278−2287.
(144) Another example of classification not covered in this review: Carbonell, P.; Carlsson, L.; Faulon, J.-L. Stereo Signature Molecular Descriptor. *J. Chem. Inf. Model.* **2013**, *53*, 887−897.
(145) Aires-de-Sousa, J.; Gasteiger, J. Prediction of Enantiomeric Excess in a Combinatorial Library of Catalytic Enantioselective Reactions. *J. Comb. Chem.* **2005**, *7*, 298−301.
(146) Long, J.; Ding, K. Engineering Catalysts for Enantioselective Addition of Diethylzinc to Aldehydes with Racemic Amino Alcohols: Nonlinear Effects in Asymmetric Deacivation of Racemic Catalysts. *Angew. Chem., Int. Ed.* **2001**, *40*, 544−547.
(147) Zhang, Q.-Y.; Zhang, D.-D.; Li, J.-Y.; Zhou, Y.-M.; Xu, J. Virtual Screening of a Combinatorial Library of Enantioselective Catalysts with Chirality Codes and Counterpropagation Neural Networks. *Chemom. Intell. Lab. Syst.* **2011**, *109*, 113−119.
(148) Vriamont, N.; Govaerts, B.; Grenouillet, P.; de Bellefon, C.; Riant, O. Design of a Genetic Algorithm for the Simulated Evolution of a Library of Asymmetric Transfer Hydrogenation Catalysts. *Chem. - Eur. J.* **2009**, *15*, 6267−6278.
(149) Zhang, Q.-Y.; Zhang, D.-D.; Li, J.-Y.; Zhou, Y.-M.; Xu, J. Prediction of Enantiomeric Excess in a Catalytic Process: A Chemoinformatics Approach Using Chirality Codes. *MATCH Commun. Math. Comput. Chem.* **2012**, *67*, 773−786.
(150) Suo, J.-J.; Zhang, Q.-Y.; Li, J.-Y.; Zhou, Y.-M.; Xu, L. The Derivation of a Chiral Substituent Code for Secondary Alcohols and its Application to the Prediction of Enantioselectivity. *J. Mol. Graphics Modell.* **2013**, *43*, 11−20.
(151) Golbraikh, A.; Bonchev, D.; Tropsha, A. Novel Chirality Descriptors Derived from Molecular Topology. *J. Chem. Inf. Comp. Sci.* **2001**, *41*, 147−158.
(152) Galvez, J.; Garcia-Domenech, R.; De Julián-Ortiz, J.; Soler, R. Topological Approach to Drug Design. *J. Chem. Inf. Model.* **1995**, *35*, 272−284.
(153) De Julián-Ortiz, J. D.; De Gregorio Alapont, C.; Rios-Santamarina, I.; Gracia-Domenech, R.; Gálvez, J. Prediction of Properties of Chiral Compounds by Molecular Topology. *J. Mol. Graphics Modell.* **1998**, *16*, 14−18.
(154) Kazlauskas, R. J. W.; Alexandra, N. E.; Rappaport Aviva, T.; Cuccia Louis, A. A Rule to Predict which Enantiomer of a Secondary Alcohol Reacts Faster in Reactions Catalyzed by Cholesterol Esterase, Lipase from Pseudomonas Cepacian, and Lipase from *Candida rugosa*. *J. Org. Chem.* **1991**, *56*, 2656−2665.
(155) Zheng, F.; Zhang, Q.; Li, J.; Juo, J.; Wu, C.; Zhou, Y.; Liu, X.; Xu, L. Machine Learning Induction of Chemically Intuitive Rules for the Prediction of Enantioselectivity in the Asymmetric Synthesis of Alcohols. *Chemom. Intell. Lab. Syst.* **2015**, *145*, 39−47.
(156) Kim, K. H. Comparative Molecular Field Analysis (CoMFA). In *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Springer: Dordrecht, 1995.
(157) Jones, J. E. On the Determination of Molecular Fields. *Proc. R. Soc. London, Ser. A* **1924**, *106*, 463−477.
(158) Lipkowitz, K.; Pradhan, M. Computational Studies of Chiral Catalysts: A Comparative Molecular Field Analysis of an Asymmetric Diels-Alder Reaction with Catalysts Containing Bisoxazoline or Phosphinooxazoline Ligands. *J. Org. Chem.* **2003**, *68*, 4648−4656.
(159) Davies, I. W.; Gerena, L.; Cai, D.; Larsen, R. D.; Verhoeven, T. R.; Reider, P. J. A Conformational Toolbox of Oxazoline Ligands. *Tetrahedron Lett.* **1997**, *38*, 1145−1148.
(160) Ghosh, A. H.; Mathivanan, P.; Cappielo, J. Conformationally Constrained Bis(oxazoline)derived Chiral Catalyst: A Highly Effective Enantioselective Diels-Alder Reaction. *Tetrahedron Lett.* **1996**, *37*, 3815−3818.
(161) Evans, D. A.; Miller, S. J.; Lectka, T.; von Matt, P. Chiral Bis(oxazoline)copper(II) Complexes as Lewis Acid Catalysts for the Enantioselective Diels-Alder Reaction. *J. Am. Chem. Soc.* **1999**, *121*, 7559−7573.
(162) Sagasser, I.; Helmchen, G. (Phosphino-oxazoline)copper(II) Complexes as Chiral Catalysts for Enantioselective Dields-Alder Reactions. *Tetrahedron Lett.* **1998**, *39*, 261−264.
(163) Evans, D. A.; Lectka, T.; Miller, S. J. Bis(imine)-copper(II) Complexes as Chiral Lewis Acid Catalysts for the Diels-Alder Reaction. *Tetrahedron Lett.* **1993**, *34*, 7027−7030.
(164) Kozlowski, M. C.; Dixon, S. L.; Panda, M.; Lauri, G. Quantum Mechanical Models Correlating Structure with Selectivity: Predicting the Enantioselectivity of β-Amino Alcohol Catalysts in Aldehyde Alkylation. *J. Am. Chem. Soc.* **2003**, *125*, 6614−6615.
(165) Bahmanyar, S.; Houk, K. N.; Martin, H. J.; List, B. Quantum Mechanical Predictions of the Stereoselectivities of Proline-Catalyzed Asymmetric Intermolecular Aldol Reactions. *J. Am. Chem. Soc.* **2003**, *125*, 2475−2479.
(166) Phuan, P.-W.; Ianni, J. C.; Kozlowski, M. C. Is the A-Ring of Sparteine Essential for High Enantioselectivity in the Asymmetric Lithiation-Substitution of N-Boc-pyrrolidine? *J. Am. Chem. Soc.* **2004**, *126*, 15473−15479.
(167) Dearden, M. J.; Firkin, C. R.; Hermet, J.-P. R.; O'Brian, P. A Readily-Accessible (+)-Sparteine Surrogate. *J. Am. Chem. Soc.* **2002**, *124*, 11870−11871.
(168) Danieli, B.; Lesma, G.; Passarella, D.; Piacenti, P.; Sacchetti, A.; Silvani, A.; Virdis, A. Synthesis of Enantiopure Diamine Ligands Related to Sparteine, via Scandium Triflate-Catalyzed Imino Diels-Alder Reactions. *Tetrahedron Lett.* **2002**, *43*, 7155−7158.
(169) Ianni, J. C.; Annamalai, V.; Phuan, P.-W.; Panda, M.; Kozlowski, M. A Priori Theoretical Prediction of Selectivity in Asymmetric Catalysis: Design of Chiral Catalysts by Using Quantum Molecular Interaction Fields. *Angew. Chem.* **2006**, *118*, 5628−5631.
(170) Huang, J.; Ianni, J. C.; Antoline, J. E.; Hsung, R. P.; Kozlowski, M. C. De Novo Chiral Amino Alcohols in Catalyzing Asymmetric Additions to Aryl Aldehydes. *Org. Lett.* **2006**, *8*, 1565−1568.
(171) Kozlowski, M.; Ianni, J. Quantum Molecular Interaction Field Models of Substrate Enantioselection in Asymmetric Processes. *J. Mol. Catal. A: Chem.* **2010**, *324*, 141−145.
(172) Melville, J. L.; Andrews, B. I.; Lygo, B.; Hirst, J. D. Computational Screening of Combinatorial Catalyst Libraries. *Chem. Commun.* **2004**, *0*, 1410−1411.
(173) Melville, J. L.; Lovelock, K. R. J.; Wilson, C.; Allbutt, B.; Burke, E. K.; Lygo, B.; Hirst, J. D. Exploring Phase-Transfer Catalysis with Molecular Dynamics and 3D/4D Quantitative Structure-Selectivity Relationships. *J. Chem. Inf. Model.* **2005**, *45*, 971−981.
(174) Broughton, J. B.; Gordaliza, M.; Castro, M.-A.; Miguel del Corral, J. M. San Feliciano, A. Modified CoMFA Methods for the Analysis of Antineoplastic Effects of Lignan Analogues. *J. Mol. Struct.: THEOCHEM* **2000**, *504*, 287−294.
(175) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B. Q.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.
(176) Stone, M. J. Cross-Validatory Choice and Assessment of Statistical Predictions. *Royal Stat. Soc. B* **1974**, *36*, 111−133.

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

(177) Jonathan, P.; Krzanowski, W. J.; McCarthy, W. V. On the use of Cross-Validation to Assess Performance in Multivariate Prediction. *Stat. Comput.* **2000**, *10*, 209−229.
(178) Denmark, S. E.; Gould, N. D.; Wolf, L. M. A Systematic Investigation of Quaternary Ammonium Ions as Asymmetric Phase-Transfer Catalysts. Synthesis of Catalyst Libraries and Evaluation of Catalyst Activity. *J. Org. Chem.* **2011**, *76*, 4260−4336.
(179) 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.* **2011**, *76*, 4337−4357.
(180) Li, L.; Pan, Y.; Lei, M. The Enantioselectivity in Asymmetric Ketone Hydrogenation Catalyzed by RuH_2(diphosphine)(diamine) Complexes: Insights from a 3D-QSSR and DFT Study. *Catal. Sci. Technol.* **2016**, *6*, 4450−4457.
(181) Xie, J.-H.; Zhou, Q.-L. Chiral Diphosphine and Monodentate Phosphorus Ligands on a Spiro Scaffold for Transition-Metal-Catalyzed Asymmetric Reactions. *Acc. Chem. Res.* **2008**, *41*, 581−593.
(182) Wu, J.; Chen, H.; Kwok, W.; Guo, R.; Zhou, Z.; Yeung, C.; Chan, A. S. Air-Stable Catalysts for Highly Efficient and Enantioselective Hydrogenation of Aromatic Ketones. *J. Org. Chem.* **2002**, *67*, 7908−7910.
(183) Ohkuma, T. Asymmetric Hydrogenation of Hetones: Tactics to Achieve High Reactivity, Enantioselectivity, and Wide Scope. *Proc. Jpn. Acad., Ser. B* **2010**, *86*, 202−219.
(184) Grasa, G. A.; Zanotti-Gerosa, A.; Hems, W. P. A Chiral [(dipyridylphosphine)RuCl2(1−3-diphenylpropanediamine)] Catalyst for the Hydrogenation of Aromatic Ketones. *J. Organomet. Chem.* **2006**, *691*, 2332−2334.
(185) Ohkuma, T.; Koizumi, M.; Muñiz, K.; Hilt, G.; Kabuto, C.; Noyori, R. *trans*-RuH(η1-BH4)(binap)(1,2-diamine): A Catalyst for Asymmetric Hydrogenation of Simple Ketones under Base-Free Conditions. *J. Am. Chem. Soc.* **2002**, *124*, 6508−6509.
(186) Yamaguchi, S.; Nishimura, T.; Hibe, Y.; Nagai, M.; Sato, H.; Johnston, I. Regularized Regression Analysis of Digitized Molecular Structures in Organic Reactions for Quantification of Steric Effects. *J. Comput. Chem.* **2017**, *38*, 1825−1833.
(187) Perrin, L.; Clot, E.; Eisenstein, O.; Loch, J.; Crabtree, R. H. Computed Ligand Electronic Parameters from Quantum Chemistry and Their Relation to Tolman Parameters, Lever Parameters, and Hammett Constants. *Inorg. Chem.* **2001**, *40*, 5806−5811.
(188) Ingle, G. K.; Mormino, M. G.; Wojtas, L.; Antilla, J. C. Chiral Phosphoric Acid-Catalyzed Addition of Thiols to *N*-Acyl Imines: Access to Chiral *N,S*-Acetals. *Org. Lett.* **2011**, *13*, 4822−4825.
(189) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. Grid-INdependent Descriptors (GRIND): A novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.
(190) Fontaine, F.; Pastor, M.; Sanz, F. Incorporating Molecular Shape into the Alignment-free Grid-INdependent Descriptors. *J. Med. Chem.* **2004**, *47*, 2805−2815.
(191) Sciabola, S.; Alex, A.; Higginson, P. D.; Mitchell, J. C.; Snowden, M. J.; Morao, I. Theoretical Prediction of the Enantiomeric Excess in Asymmetric Catalysis. An Alignment-Independent Molecular Interaction Field Based Approach. *J. Org. Chem.* **2005**, *70*, 9025−9027.
(192) Hoogenraad, M.; Klaus, G. M.; Elders, N.; Hooijschuur, S. M.; McKay, B.; Smith, A. A.; Damen, E. W. P. Oxazaborolidine Mediated Asymmetric Ketone Reduction: Prediction of Enantiomeric Excess Based on Catalyst Structure. *Tetrahedron: Asymmetry* **2004**, *15*, 519−523.
(193) Urbano-Cuadrado, M.; Carbó, J. J.; Maladonado, A. G.; Bo, C. New Quantum Mechanics-Based Three-Dimensional Molecular Descriptors for Use in QSSR Approaches: Application to Asymmetric Catalysis. *J. Chem. Inf. Model.* **2007**, *47*, 2228−2234.
(194) Pu, L.; Yu, H. B. Catalytic Asymmetric Organozinc Additions to Carbonyl Compounds. *Chem. Rev.* **2001**, *101*, 757−824.
(195) Soai, K.; Niwa, S. Enantioselective Addition of Organozinc Reagents to Aldehydes. *Chem. Rev.* **1992**, *92*, 833−856.
(196) Aguado-Ullate, S.; Guasch, L.; Urbano-Cuadrado, M.; Bo, C.; Carbó, J. J. 3D-QSPR Models for Predicting the Enantioselectivity and the Activity for the Asymmetric Hydroformylation of Styrene Catalyzed by Rh-diphosphane. *Catal. Sci. Technol.* **2012**, *2*, 1694−1704.
(197) See the original work for more detail on computational methods.
(198) Axtell, A. T.; Klosin, J.; Abboud, K. A. Evaluation of Asymmetric Hydrogenation Ligands in Asymmetric Hydroformylation Reactions. Highly Enantioselective Ligands Based on Bisphosphacycles. *Organometallics* **2006**, *25*, 5003−5009.
(199) Ewalds, R.; Eggeling, E. B.; Hewat, A. C.; Kamer, P. C. J.; van Leeuwen, P. W. N. M.; Vogt, D. Application of *P*-Stereogenic Aminophosphine Phosphinite Ligands in Asymmetric Hydroformylation. *Chem. - Eur. J.* **2000**, *6*, 1496−1504.
(200) Yao, Y.-Y.; Xu, L.; Yang, Y.-Q.; Yuan, X.-S. Study on Structure-Activity Relationships of Organic Compounds: Three New Topological Indices and Their Applications. *J. Chem. Inf. Model.* **1993**, *33*, 590−594.
(201) Jiang, C.; Li, Y.; Tian, Q.; You, T. QSAR Study of Catalytic Asymmetric Reactions with Topological Indices. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1876−1881.
(202) Aratani, T. Catalytic Asymmetric Synthesis of Cyclopropanecarboxylic Acids: an Application of Chiral Copper Carbenoid Reaction. *Pure Appl. Chem.* **1985**, *57*, 1839−1844.
(203) Bandini, M.; Cozzi, P. G.; et al. Highly Diastereoselective Pinacol Coupling of Aldehydes Catalyzed by Titanium-Schiff Base Complexes. *Tetrahedron Lett.* **1999**, *40*, 1997−2000.
(204) Hayashi, T.; Konishi, M.; Fukushima, M.; Mise, T.; Kagotani, M.; Tajika, M.; Kumada, M. Asymmetric Synthesis catalyzed by Chiral Ferrocenylphosphine-Transition Metal Complexes: Nickel and Palladium Catalyzed Asymmetric Grignard Cross-Coupling. *J. Am. Chem. Soc.* **1982**, *104*, 180−186.
(205) Jiang, C.; Li, D.; Wen, J.; You, T. QSAR Study on the Enantiomeric Excess in Asymmetric Catalytic Reactions with Topological Indices and an Artificial Neural Network. *J. Mol. Model.* **2006**, *13*, 91−97.
(206) Chen, J.; Jiewu, W.; Mingzong, L.; You, T. Calculation on Enantiomeric Excess of a Catalytic Asymmetric Reactions of Diethylzinc Addition to Aldehydes with Topological Indices and Artificial Network. *J. Mol. Catal. A: Chem.* **2006**, *258*, 191−197.
(207) Randic, M. Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(208) Kier, L. B.; Hall, L. M. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(209) Kier, L. B. A. Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109−116.
(210) Dewar, M. J.; Zoebisch, E. G.; Healy, F. E. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: a New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(211) Reddy, K. S.; Solá, L.; Moyano, A.; Pericás, M. A.; Riera, A. Synthesis of a 9-Fluorenone Derived β-Amino Alcohol Ligand Depicting High Catalytic Activity and Pronounced Non-linear Stereochemical Effects. *Synthesis* **2000**, *1*, 165−176.
(212) Kang, J.; Kim, J. W.; Lee, J. W.; Kim, D. S.; Kim, J. I. Chiral β-Amino Thiol Catalysts for the Enantioselective Addition of Diethylzinc to Aldehydes. *Bull. Korean Chem. Soc.* **1996**, *17*, 1135−1142.
(213) Kang, J.; Kim, D. S.; Kim, J. I. Enantioselective Addition of Diethylzinc to Aldehydes Catalyzed by A Drug-Unrelated Chiral Amino Thiol and the Corresponding Disulfide. *Synlett* **1994**, *10*, 842−844.
(214) Yamakawa, M.; Noyori, R. Asymmetric Addition of Dimethylzinc to Benzaldehyde Catalyzed by (2S)-3-exo-(Dimethylamino)isobornenol. A Theoretical Study on the Origin of Enantioselection. *Organometallics* **1999**, *18*, 128−133.

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

(215) Yamakawa, M.; Noyori, R. An Ab Initio Molecular Orbital Study on the Amino Alcohol-Promoted Reaction of Dialkylzincs and Aldehydes. *J. Am. Chem. Soc.* **1995**, *117*, 6327−6335.
(216) Chelucci, G.; Conti, S.; Falorni, M.; Giacomelli, G. Chiral Ligands Containing Heteroatoms. 8. 2-[(2s)-2-pyrrolidinyl]pyridine as a Novel Catalyst in the Enantioselective Addition of Diethylzinc to Aldehydes. *Tetrahedron* **1991**, *47*, 8251−8258.
(217) Conti, S.; Falorni, M.; Giacomelli, G.; Soccolini, F. Chiral Ligands Containing Heteroatoms. 10. 1-(2-pyridyl)alkylamines as Chiral Catalysts in the Addition of Diethylzinc to Aldehydes: Temperature Dependence on Enantioselectivity. *Tetrahedron* **1992**, *48*, 8993−9000.
(218) van der Linden, J. B.; Ras, E.-J.; Hooijschuur, S. M.; Klaus, G. M.; Luchters, N. T.; Dani, P.; Verspui, G.; Smith, A. A.; Damen, E. W. P.; McKay, B.; Hoogenraad, M. Asymmetric Catalytic Ketone Hydrogenation: Relating Substrate Structure and Product Enantiometic Excess Using QSPR. *QSAR Comb. Sci.* **2005**, *24*, 94−98.
(219) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; et al. Virtual Computational Chemistry Laboratory - Design and Description. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 453−463.
(220) Luo, Y.; Berry, N. G.; Carnell, A. J. Chiral Bicyclic [2.2.2] Octadiene Ligands for Rh-Catalysed Catalytic Asymmetric Conjugate Additions to Acyclic Enones: a Quantitative Structure-Property Relationship. *Chem. Commun.* **2012**, *48*, 3279−3281.
(221) Todeschini, R.; Consonni, V. *Molecular Descriptors*; Wiley-VCH, 2009.
(222) González-Díaz, H.; Arrasate, S.; Gómez-San Juan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. M. General Theory for Multiple Input-Output Perturbations in Complex Molecular Systems. 1. Linear QSPR Electronegativity Models in Physical, Organic, and Medicinal Chemistry. *Curr. Top. Med. Chem.* **2013**, *13*, 1713−1741.
(223) Blázquez-Barbadillo, C.; Aranzamendi, E.; Coya, E.; Lete, E.; Sotomayor, N.; González-Díaz, H. Perturbation Theory Model of Reactivity and Enantioselectivity of Palladium-Catalyzed Heck-Heck Cascade Reactions. *RSC Adv.* **2016**, *6*, 38602−38610.
(224) Lage, S.; Martinez-Estibalez, U.; Sotomayor, N.; Lete, E. Intramolecular Palladium-Catalyzed Direct Arylation vs. Heck Reactions: Synthesis of Pyrroloisoquinolines and Isoindoles. *Adv. Synth. Catal.* **2009**, *351*, 2460.
(225) Maddaford, S. P.; Andersen, N. G.; Cristofoli, W. A.; Keay, B. A. Total Synthesis of (+)-Xestoquinone Using an Asymmetric Palladium Catalyzed Polyene Cyclization. *J. Am. Chem. Soc.* **1996**, *118*, 10766−10773.
(226) Lau, S. Y. W.; Keay, B. A. Remote Sybstituent Effects on the Enantiomeric Excess of Intramolecular Asymmetric Palladium-Catalyzed Polyene Cyclizations. *Synlett* **1999**, *5*, 605−607.
(227) Lau, S. Y. W.; Andersen, N. G.; Keay, B. A. Optimization of Palladium-Catalyzed Polyene Cyclizations: Suppression of Competing Hydride Transfer from Tertiary Amines with Dabco and an Unexpected Hydride Transfer from 1,4-Dioxane. *Org. Lett.* **2001**, *3*, 181−184.
(228) Gorobets, E.; Sun, G.-R.; Wheatly, B. M. M.; Parvez, M.; Keay, B. A. Synthesis, Resolution and Applications of 3,3′-bis(RO)-MeO-BIPHEP Derivatives. *Tetrahedron Lett.* **2004**, *45*, 3597−3601.
(229) Rankic, D.; Lucciola, D.; Keay, B. A. Application of 3,3′-disubstituted xylBINAP Derivatives in Inter- and Intramolecular Asymmetric Heck/Mizorki Reactions. *Tetrahedron Lett.* **2010**, *51*, 5724−5727.
(230) Lucciola, D.; Keay, B. A. Further Developments of an Enantioselective Palladium-Catalyzed Polyene Cyclization: Surprising Solvent and Ligand Effects. *Synlett* **2011**, *2011*, 1618−1622.
(231) Miyazaki, F.; Uotsu, K.; Shibasaki, M. Silver Salt Effects on an Asymmetric Heck Reaction. Catalytic Asymmetric Total Synthesis of (+)-xestoquinone. *Tetrahedron* **1998**, *54*, 13073−13078.
(232) Aranzamendi, E.; Arrasate, S.; Sotomayor, N.; González-Díaz, H.; Lete, E. Chiral Brønsted Acid-Catalyzed Enantioselective α-Amidoalkylation Reactions: A Joint Experimental and Predictive Study. *ChemistryOpen* **2016**, *5*, 540−549.
(233) Hill, T.; Lewicki, P. *Statistics, Methods, and Applications*; Stat Soft Inc.: Tulsa, OK, 2006.
(234) Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E. P.; Grzybowski, B. A.; Bambin, A. Predicting the Outcomes of Organic Reactions via Machine Learning: are Current Descriptors Sufficient? *Sci. Rep.* **2017**, *7*, 3582.
(235) Eksterowicz, J. E.; Houk, K. N. Transition-State Modeling with Empirical Force Fields. *Chem. Rev.* **1993**, *93*, 2439−2461.
(236) Hansen, E.; Rosales, A. R.; Tutkowiski, B.; Norrby, P.-O.; Wiest, O. Prediction of Stereochemistry using Q2MM. *Acc. Chem. Res.* **2016**, *49*, 996−1005.
(237) Rosales, A. R.; Quinn, T. R.; Wahlers, J.; Tomberg, A.; Zhang, X.; Helquist, P.; Wiest, O.; Norrby, P.-O. Application of Q2MM to Predictions in Stereoselective Synthesis. *Chem. Commun.* **2018**, *54*, 8294−8311.
(238) Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; Wiest, O.; Norrby, P.-O. Rapid Virtual Screening of Enantioselective Catalysts Using CatVS. *Nature Catalysis* **2019**, *2*, 41−45.
(239) Guan, Y.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E. AARON: An Automated Reaction Optimizer for New Catalysts. *J. Chem. Theory Comput.* **2018**, *14*, 5249−5261.
(240) Lipkowitz, K.B.; D'Hue, C. A.; Sakamoto, T.; Stack, J. Stereocartography: A computational Mapping Technique That Can Locate Regions of Maximum Stereoinduction around Chiral Catalysts. *J. Am. Chem. Soc.* **2002**, *124*, 14255−14267.
(241) Kozlowski, M. C.; Panda, M. Computer-Aided Design of Chiral Ligands. Part 2. Functionality Mapping as a Method To Identify Stereocontrol Elements for Asymmetric Reactions. *J. Org. Chem.* **2003**, *68*, 2061−2076.
(242) Poater, A.; Ragone, F.; Mariz, R.; Dorta, R.; Cavallo, L. Comparing the Enantioselective Power of Steric and Electrostatic Effects in Transition-Metal-Catalyzed Asymmetric Synthesis. *Chem. - Eur. J.* **2010**, *16*, 14348−14353.
(243) Falivene, L.; Credendino, R.; Poater, A.; Petta, A.; Serra, L.; Oliva, R.; Scarano, V.; Cavallo, L. SambaVca 2. A Web Tool for Analyzing Catalytic Pockets with Topographic Steric Maps. *Organometallics* **2016**, *35*, 2286−2293.
(244) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatics Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(245) Zhou, Z.; Li, X.; Zare, R. N. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **2017**, *3*, 1337−1344.
(246) 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* **2018**, *360*, 186−190.
(247) Nielsen, M.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with Sylfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **2018**, *140*, 5004−5008.
(248) For articles related to the omitted variable bias: (a) Clarke, K. A. The Phantom Menace: Omitted Variable Bias in Econometric Research. *Conflict Management and Peace Science* **2005**, *22*, 341−352.
(249) Clarke, K. A. Return of the Phantom Menace: Omitted Variable Bias in Political Research. *Conflict Management and Peace Science* **2009**, *26*, 46−66.

NOTE ADDED IN PROOF

Since the submission of this manuscript, several relevant publications have appeared which could not be included. (a) Falivene, L.; Cao, Z.; Petta, A.; Serra, L.; Poater, A.; Oliva, R.; Scarano, V.; Cavallo, L. Towards the Online Computer-Aided Design of Catalytic Pockets. *Nat. Chem.* **2019**, *11*, 872−879. (b) Brethomé, A. V.; Paton, R. S.; Fletcher, S. P. Retooling

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689

Asymmetric Conjugate Additions for Sterically Demanding Substrates with an Iterative Data-Driven Approach. *ACS. Catal.* **2019**, *9*, 7179−7187. (c) Yamaguchi, S.; Sodeoka, M. Molecular Field Analysis Using Intermediates in Enantio-Determining Steps Can Extract Information for Data-Driven Molecular Design in Asymmetric Catalysis. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701−1706. (d) Reid, J. P.; Proctor, R. S. J.; Sigman, M. S.; Phipps, R. J. Predictive Multivariate Linear Regression Analysis Guides Successful Catalytic Enantioselective Minisci Reactions of Diazines. *J. Am. Chem. Soc.* **2019**, *141*, 19178−19285. (e) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571*, 343−348.

DOI: 10.1021/acs.chemrev.9b00425
Chem. Rev. 2020, 120, 1620−1689