Catalyzing Change: The Power of Computational Asymmetric Catalysis

Sharon Pinus,[1,ǂ] Jérôme Genzling,[1] Mihai Burai-Patrascu,[2] Nicolas Moitessier,[1,]*

[1] Department of Chemistry, McGill University, 801 Sherbrooke St. W., Montréal, Québec, H3A 0B8, Canada

[2] Molecular Forecaster Inc., 910-2075 Robert Bourassa St., Montreal, Quebec, H3A2L1, Canada.

[ǂ] First author. * Correspondence: Nicolas Moitessier, nicolas.moitessier@mcgill.ca

Abstract

Computational asymmetric catalysis has seen an impressive rise in the last twenty years, thanks to advancements in algorithm and method development for predicting catalyst enantioselectivity. These methods/algorithms describe reactions that can be categorized into two groups: reactions where **1)** knowledge of the mechanism is not required and where leveraging experimental data to establish correlations between reaction descriptors and enantioselectivity is imperative, and **2)** the mechanism (or transition state (TS) for the enantioselective step) is known and used to determine catalyst stereoselectivity by modeling the diastereomeric TSs. Although these methods have reached an important level of proficiency for enantioselectivity prediction, this field remains largely obscured for experimental chemists. In this review, we aim to shed light on models, methods, and applications used in asymmetric synthesis, with accessible language suited for experimental chemists. Our hope is that these methods will ultimately be adopted by synthetic chemists for the design of novel catalysts.

Introduction

Asymmetric Catalysis. The field of asymmetric catalysis has revolutionized organic synthesis in the last 50 years. Catalysts have been developed to introduce stereogenic centers into molecules primarily through the formation of new C-C bonds and reduction of unsaturated bonds (e.g., carbonyls, alkenes). These chiral catalysts can take many forms, including transition metal complexes, organocatalysts, and biocatalysts (**Figure 1**). With transition metal complexes, the reaction is often catalyzed by the metal itself while stereochemistry is introduced in the form of metal ligands (e.g., chiral phosphines). In organocatalysis and biocatalysis, small organic molecules and enzymes function as both catalysts and chiral directing groups. The latter two forms of asymmetric catalysts have been seen as very promising alternatives to transition metal catalysts, due to their reduced costs and toxicity. In fact, the potential impact of these alternatives was recently recognized when the Nobel Prize in Chemistry was awarded to Frances Arnold (2018, directed evolution of enzymes), then Benjamin List and David MacMillan (2020, asymmetric organocatalysis), after Barry Sharpless, Ryoji Noyori and William Knowles shared the Nobel Prize in 2001 for asymmetric transition metal catalyzed reactions.

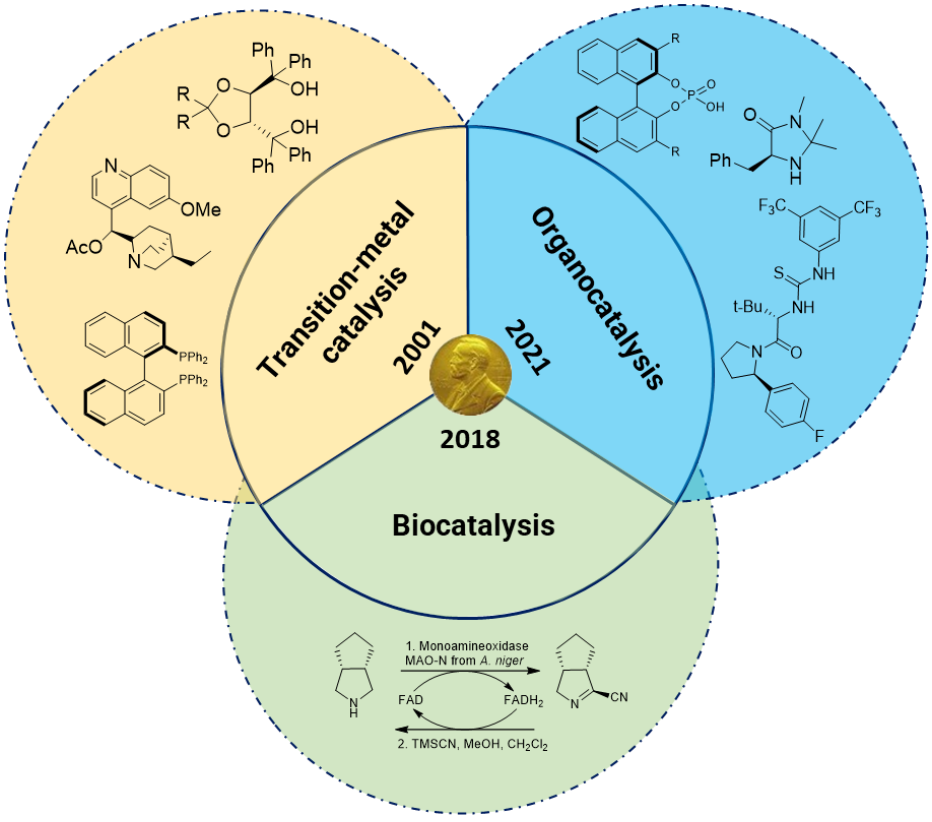


Figure 1. The three main asymmetric catalysis fields.

Developing new catalysts. The relatively simple mechanisms of organocatalyzed reactions are a significant advantage for their development and optimization. However, the use of these catalysts has been hampered by their lower stereoselectivities and the need for higher loading. Alternatively, development of transition metal catalysts is facing complex mechanisms often involving multiple possible transition states (TSs), metal coordinations, additives, and ligands. As a result, the very tedious "trial-and-error" approach is still commonly used. To address environmental concerns, catalysts based on greener and cheaper metals (bismuth, iron, copper) have been developed, yet the most used transition metal catalysts are built around

palladium, rhodium, and other toxic and expensive metals. Thus, despite tremendous progress in the field of small molecule catalysis, development of new chiral catalysts remains quite challenging and often calls for stepwise optimization. This is often a time-consuming, labour-intensive process. It requires the synthesis and evaluation of multiple novel ligands/catalysts, in an iterative process that is often pursued in an empirical fashion with little guidance other than simple models and intuition. Computational methods guiding the design of new catalysts are sought after and should address this major issue.

Computational methods for catalyst design and discovery. As an analogy, over the past decades, docking-based virtual screening has found extensive use and acceptance as a design tool in medicinal chemistry.[1] The low computational demands of these methods and user-friendly interfaces removed hurdles towards their widespread adoption. In contrast, computational tools that could improve the process of chemical reaction development remain underutilized as predictive/design methods. The power of quantum mechanics (QM) calculations, particularly density functional theory (DFT), is primarily used in a retrospective, *post-hoc* fashion for understanding reaction mechanisms and for rationalizing observed selectivities, rather than in the prediction/design of new catalysts. In fact, the computational cost associated with *ab initio* QM or DFT methods, let alone the required expert knowledge, makes them unsuitable for the screening of large libraries of potential catalysts. However, major efforts are currently ongoing to develop computational tools to assist organic chemists, and integration into organic chemistry laboratories is imminent.[2]

We believe that the successes of virtual screening in medicinal chemistry should be adaptable to reaction prediction, with the challenge being the availability of computationally inexpensive, rapid, and accurate methods for predicting stereoselectivities associated with complex TS structures and energies.

State-of-the-art approaches. In recent years, computer-assisted synthesis has gained significant momentum and several computational methods (most commonly machine learning (ML) methods and statistical models) applied to organic synthesis problems have been reported. For example, computer-aided synthesis planning has advanced rapidly and can propose reaction mapping[3] and realistic retrosynthesis[2, 4-6] (e.g., Chematica/Synthia[7, 8] and AIZynthFinder[9]), can predict yields,[10-12] catalyst inhibition,[13] regioselectivity,[14] and chemical reactivity.[15] In practice, these predictive trained methods have been successfully used to design new catalysts[16] and predict the stereochemical outcome of asymmetric reactions.[17] For the latter, ML techniques are advantageous over QM and molecular mechanics (MM) for two reasons: speed (orders of magnitude faster than QM) and their application to reactions with unclear or complex mechanisms (ML models developed from catalyst structures only).[18] However, training ML models requires a significant amount of experimental data for training[19] and can hardly be applied to new reactions, as emphasized by Norrby in a viewpoint.[20] Reaction/chemical mapping is more general; for example, physicochemical and QM descriptors of over 300,000 monodentate phosphine ligands have been added to a database named Kraken,[21] which will certainly be particularly useful for new reactions relying on phosphine ligands. However, despite all these successful developments, much is left to be done in the field, including the use of these methods in prospective studies.

In this review we will describe the development and application of computational methods for the design of asymmetric catalysts with a particular focus on organocatalysis and transition metal catalysis. For more information about biocatalysis (primarily enzymes), the readers are referred to the excellent overviews of this field by *Bell et al.*[22] and *Pyser et al.*[23] Thus, while computational methods for biocatalysis have been developed,[24] these will not be discussed herein.

In the context of computational asymmetric catalysis, an overarching goal of the computational methods is the prediction of enantioselectivity of asymmetric reactions, enabling computational reaction

optimization (i.e., optimal catalysts, substrates, ligands, and conditions). Different methodologies for achieving these goals exist, and they can be divided in two main categories: **1)** methods requiring sufficient knowledge on the reaction mechanism (primarily QM and MM-based methods) and **2)** data-driven methods (primarily databases and ML methods). Irrespective of category, the overall principles are similar and involve data collection, identification of meaningful patterns, parameters or features, model building and testing, followed by model refinement (**Figure 2**).

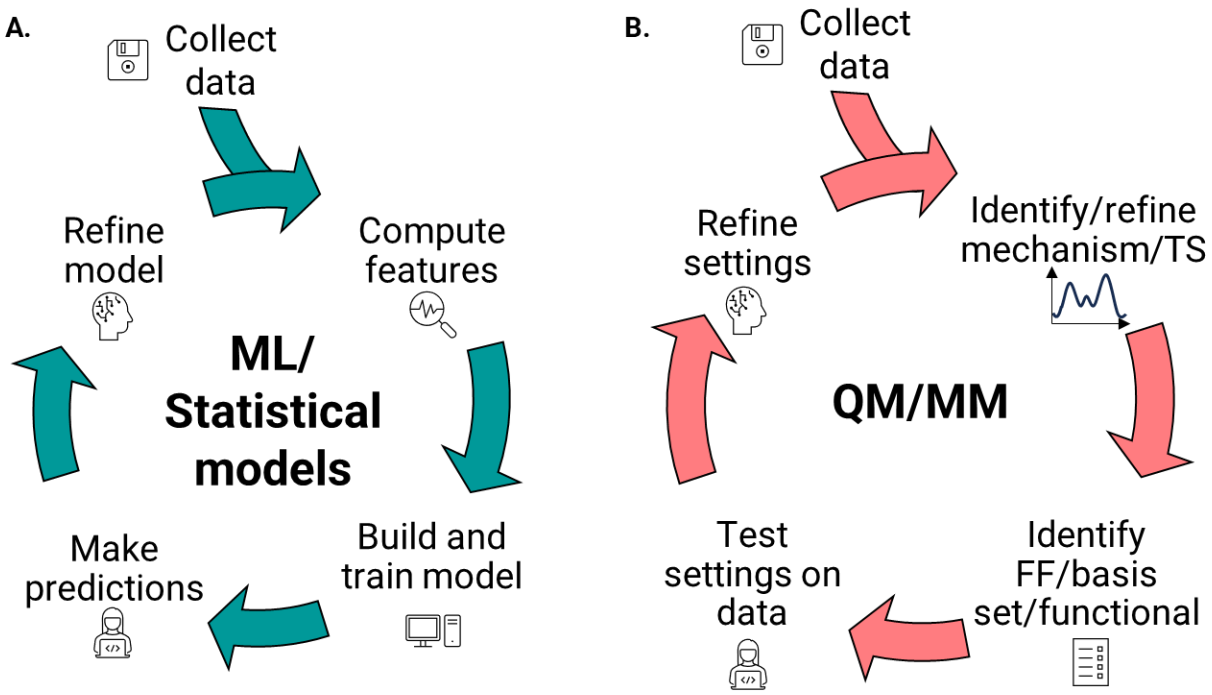


Figure 2. Overall workflow for A) ML methods and B) QM/MM methods.

Data.

Developing datasets. The first step in any modeling project is often ***data collection*** from the scientific literature (or generating data in the wet lab). Since representative systems are selected and modeled based on these data (ML) and/or used to test the method (QM, MM), ***data curation*** is essential, as literature data may be riddled with issues (e.g., misassigned stereocenters, incomplete experimental data, reproducibility issues). This process is a challenging task that may be prone to human error. Thus, model or method development and/or evaluation require the availability of consistent and reliable experimental enantioselectivities. The type and amount of data required to develop a predictive model depends on the approach used to model the reaction (evaluation on small sets or training on large sets) - Box 1.

Below, we list several guidelines that we believe to be key when developing a robust dataset.

1) For most methods, it is assumed that the set of substrates and catalysts collected follow a similar (if not identical) mechanism. If more than one competing mechanism is to be considered, substrates and/or catalysts favoring alternative mechanisms should be sufficiently represented. For each mechanism, one should also assume that the enantioselectivity is affected by the same factors.

2) In the context of catalyst design using ML methods, the model is more likely transferable to new sets of catalysts and substrates when trained on datasets with larger chemical diversity. If a component is kept constant (e.g., all the reactions in the dataset are applied to the same substrate), then the model will not be

trained to understand the impact of this component (e.g., not applicable to the search for ideal substrates for a given catalyst or prediction of the substrate scope of an asymmetric reaction).

3) Training a model or testing a method requires not only information on highly selective catalysts, but also on poorly selective catalysts. For optimal accuracy, ML models should be trained to correlate "good" features (e.g., chemical groups) to high enantioselectivity, as well as "bad" features to poor enantioselectivity, while QM and MM methods must be evaluated for their ability to distinguish good from poorly stereoselective catalysts.

4) TS-based methods derive the enantioselectivity from the energy difference of the diastereomeric TSs, requiring some experimental knowledge of the mechanism. More specifically, the stereoselectivity determining step (or steps) must be known, and at the very least, a good hypothesis for the TS structure must be available. If not, a separate investigation of the mechanism must be carried out, which can be quite time-consuming.

5) In general, the larger (and more diverse) the dataset is, the more information ML models may learn. However, building a large dataset should be done with care as adding data may also results in loss of diversity and introduction of biases. For example, overrepresenting a class of catalysts may result in the model learning (or even memorizing) mostly about this chemical series (e.g., phenyl better than methyl at a given position on a ring), rather than learning general rules (e.g., steric effects). This can result in significant biases of the model and poor accuracy in the search for novel catalysts. At this point, it is important to note that information following all these criteria is rarely available in a form that can be used immediately (such as a text files or formatted tables). For QM and MM methods, the size of the set is not as relevant, but the diversity should still be a focus of the data collection, as these methods must be able to capture various effects experienced by various catalysts (e.g., hydrogen bonds, cation-π interactions). In practice, should all this information be available (several catalysts from different classes already developed), one may question the need for a model to design novel catalysts for this reaction. Thus, the transferability of the method to other reactions may be investigated (see ***Applications*** section). In the case several catalysts are already available but none providing the level of stereoselectivity targeted, an ML model may be required. However, one may question the ability of a model trained on poor to good catalysts to identify excellent catalysts (e.g., based on a different mechanism, on an interaction not experienced by other catalysts). An important aspect is that the model would be as general as the data set is: *if the data set contains little variability, the model would likely fail to predict an out of set example*.

6) Depending on the method used, additional information may be needed. For example, in the case of MM, relevant force fields (FFs, Box 1) are necessary (e.g., for transition metals). This will be discussed further under the ***Descriptors*** section.

7) In practice, ML models are trained on a first set (referred to as training set) and tested on a second distinct set (testing set), while hyperparameters may be optimized using a third set (validation set). While simple random splitting is still often used, the similarity between these sets must be monitored and minimized. If the model memorizes input data ("*CH_3COOH: pKa 4.75*") rather than learning to predict a property (e.g., "*electronic effects make acetic acid acidic, with a pKa of 4.75*"), the model tends to be poorly predictive (e.g., "*CF_3COOH looks like CH_3COOH and is predicted to have a pKa of 4.75*", as opposed to "*the fluorine inductive effect reduces the pKa to 0.2*"). If similar compounds are kept in the testing set, memorization would still yield a good prediction ("*C_2H_5COOH looks like CH_3COOH and is predicted to have a pKa 4.75*"). This would lead to an overestimation of the real accuracy of the model.

Available datasets. When MM or QM methods are used, developers often rely on sets of a few dozens of systems to test their methods. However, as more data-intensive ML-based methods become more prevalent, and the need for datasets arises, a few curated datasets have been reported and made available. While these datasets contain information about catalysts/ligands and computed descriptors, the reaction-related information still needs to be manually collected.

We have built a non-exhaustive collection of available datasets (see Supporting Information for a detailed breakdown) which may be useful for method developers. We expand on a selection of datasets below:

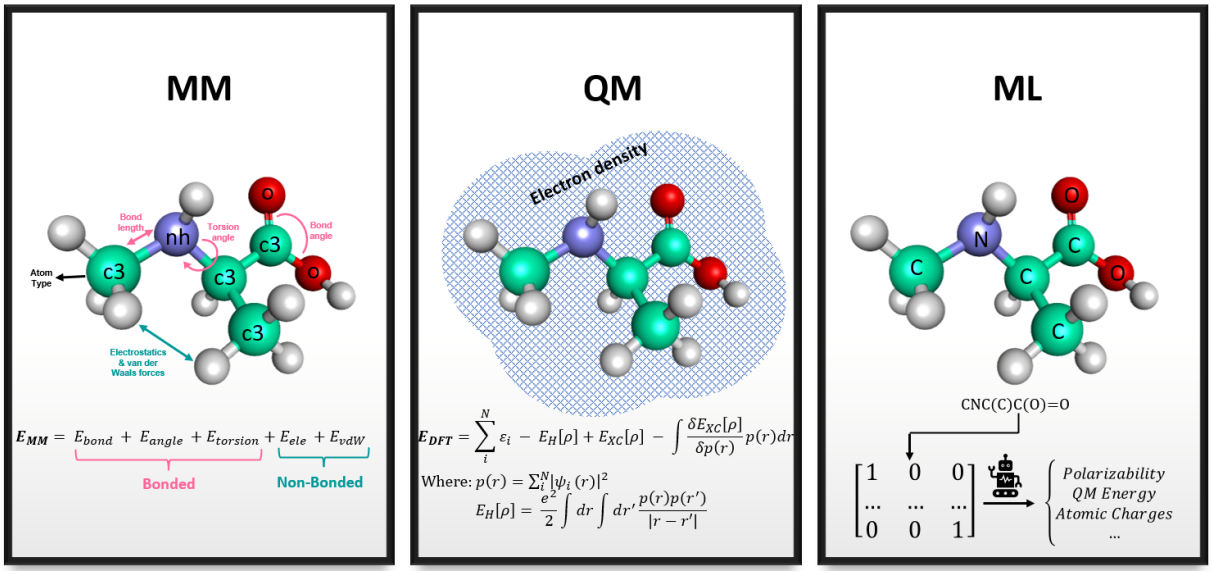
1. **Kraken.** This database, curated by Gensch *et al.*,[21] contains ~300,000 virtual monodentate organophosphorous (III) ligands for asymmetric catalysis. These ligands were combinatorially enumerated *in silico* using a set of 1,558 experimental ligands (including commercially available compounds) and 576 unique, diverse substituents. For the set of 1,558 ligands, physicochemical descriptors were calculated on conformer ensembles using QM methods. These descriptors were then used as input for ML models trained to predict the physicochemical profiles of the entire virtual library of 300,000 ligands.
2. **OSCAR**. This dataset of organocatalysts assembled by Gallarati *et al.*[25] is available online (see Supporting Information) and contains 4,000 catalysts collected either from literature or the Cambridge Structural Database (CSD), along with combinatorially enriched sets for carbene catalysts (over 8,000), and non-covalent dual-hydrogen-bond donor catalysts (ca. 1.5 million). All catalysts have QM-computed stereoelectronic descriptors and DFT-optimized structures available.
3. **VIRTUAL CHEMIST.** Upon the publication of the VIRTUAL CHEMIST platform for asymmetric catalysis, Burai-Patrascu *et al.*[26] made available the data collected for the platform validation. These data include experimental conditions and %*ee* (experimental and computed) for over 350 reactions across 7 reaction classes, involving both transition metal and organocatalysis. The data are available in table format (see Supporting Information in reference [26]) and on the research group website (http://www.moitessier-group.ca/) for structures.

Box 1. *A closer look at MM, QM, and ML.*

MM. In MM, the molecules are somewhat represented as charged points (atoms) connected by springs (bonds) and interacting by other means (e.g., angles, torsions, van der Waals, electrostatic) in a 3-dimensional Cartesian space. Because MM does not consider electrons or nuclei, atoms (and the interactions between them) must be parametrized *a priori*. These parameters may be obtained either experimentally (e.g., van der Waals radius of an atom) or from high level QM calculations (e.g., atomic charge, force constants) and stored in lists termed force fields (FFs). Importantly, in MM, atoms in different chemical environments are distinguished by introducing the concept of atom types (i.e., an oxygen atom in water vs. in a carbonyl group would have different atom types), each bearing specific bonding and non-bonding parameters. It is thus important to define the correct atom type for a system under scrutiny, as the total energy of the system is calculated as a sum of contributions from both bonded and non-bonded terms. Due to the size of the chemical space - >10^{60} small organic molecules - the number of possible atom types is virtually extremely large.[89] To make the development of FFs tractable, the number of atom types is limited yet must be accurately assigned (together with the corresponding parameters) to each atom of any molecule. On the one hand, although the problem of accurately atom typing every molecule seems problematic, significant efforts have been made to exhaustively assign atom types to a large number of small molecules.[90] On the other hand, others have proposed approaches discarding atom types altogether.[91, 92] Thus, the accuracy of MM calculations is strongly impacted by the choice of an FF and the assignment of atom types.

QM. In contrast to MM, QM methods consider both electrons and nuclei. The orbitals required to calculate electronic terms are described by basis sets. Many different types of basis sets have been developed, and their usage depends on the system under scrutiny (i.e., metal complexes, organic molecules, etc.). As a rule of thumb, the larger the basis set, the more computationally expensive the simulations. The explicit treatment of electrons significantly increases the computational cost compared to MM, due to the resource-intensive computations of electronic integrals. Depending on the desired level of accuracy, significant speed-ups are possible, either through elegant algorithms for computing electronic integrals (see the ORCA SHARK integral engine[93]), lower-cost semi-empirical (SE) QM methods where some integrals are neglected (e.g., PM6,[94] GFN2-XTB[95]), composite methods such as HF-3c/PBEh-3c, or DFT. By far the most popular method is DFT, which has seen a tremendous uptake from organic chemists. Most often employed to rationalize reaction mechanisms, DFT has more recently been used to try and explain the reactivity of different types of chemicals (through the conceptual DFT – cDFT – framework). Higher accuracy algorithms such as Møller–Plesset (MP) or coupled cluster (CC) are generally reserved for advanced theoretical work and are not commonly employed in routine organic chemistry simulations. Overall, the accuracy of the calculations is highly dependent on the choice of method (e.g., MP2, DFT, SE) and basis sets.

ML. ML methods differ significantly from both MM and QM. ML algorithms (and the more interpretable statistical models) aim at finding and quantifying patterns in vast amounts of data for predicting a given outcome. In chemistry, this outcome can come in the form of a single number (e.g., pKa, solubility, enantioselectivity), while the data can be in the form of structural data, reaction data, etc. As a major advantage, ML algorithms are orders of magnitude faster than both MM and QM. In contrast to both QM and MM, most ML models are "black box", meaning that the way in which an algorithm arrives at a prediction is unknown, although explainable artificial intelligence (AI) is emerging. There are multiple flavours of ML: **supervised, unsupervised, and reinforcement**, each with their own advantages and disadvantages. For example, **supervised learning** requires labelled input and output training data (e.g., catalyst A provides an enantioselectivity of 45 %ee), whereas **unsupervised learning** deals with raw data. In **reinforcement learning**, a feedback loop is employed so that the model can learn from its environment and maximize its correct predictions. Irrespective of flavour or model, the key ingredients for a predictive ML model are *data quality and quantity.*



Descriptors.

Encoding catalysts, substrates and/or TS structures. There are numerous methods through which molecular complexes and associated information (e.g., solvent, temperature, counter ions) can be represented numerically. Since enantioselectivity is affected by multiple factors, these representations usually include a combination of steric, electronic, and geometric information - Box 2.[27, 28]

In TS-based methods molecules can be represented in several ways, depending on the approach. In MM methods, molecules are represented by Cartesian coordinates, atom types and/or FF parameters (Box 1). The challenge is to optimize TS structures when most FFs have been developed for ground structures. To solve this, specific FFs have been developed (i.e., ReaxFF), while other methods rely on a combination of ground state reactants and products (i.e., empirical valence bond (EVB)),[29] multi configurational molecular mechanics (MCMM),[30]and SEAM[31]). Although these methods are available to the scientific community at large, they still require expertise in computational chemistry, scripting, and/or computer environments to be truly usable in chemistry laboratories. In this context, another two approaches, ACE and Q2MM, were implemented into user friendly platforms (VIRTUAL CHEMIST and CatVS, Box 3).

In QM methods, molecules are more accurately described than in MM through the usage of *atomic and molecular orbitals* to describe atoms in molecules (Box 1). In terms of computational cost and time requirements, QM calculations are orders of magnitude more costly than MM. However, the major advantages of QM over MM when it comes to predicting TS structures are **1)** the ability to optimize structures without the need for specific parameters (assuming all the necessary elements - in particular transition metals - are included in the basis set, see Box 1) and **2)** obtaining accurate energies and geometries that describe the bond breaking/forming process. While some QM methods (i.e., *ab initio*: Hartree-Fock or post-Hartree-Fock, DFT) are highly accurate, they can be used effectively for catalyst design primarily in a *post-hoc* manner and on a limited number of systems.[32] Alternatively, SEQM methods, although less accurate than the former, are significantly faster and can be envisioned as a useful tool in prospectively screening libraries of hundreds of potential catalysts due to their relatively low computational cost. However, the accuracy of SEQM in transition metal catalysis is yet to be demonstrated.[33]

In ML-based approaches, the information is usually represented by descriptors (also referred to as features). Different molecular representations exist (e.g., graph, simplified molecular-input line-entry system (SMILES)) and are linked with different molecular descriptors.[34] Among those are *system descriptors* (temperature, concentration, etc.), *steric descriptors*[35] (e.g., Sterimol parameters, average steric occupancy (ASO), %buried volume), *electronic descriptors* (e.g., Natural Bond Orbital (NBO)- charges, polarizability, Frontier Molecular Orbital (FMO)-gap), and *geometric descriptors* (e.g., bond lengths, dihedral angles). The latter are often obtained from QM calculations (see Box 2).[36]

Selection of descriptors and their computation. In TS-based models, descriptors are often chosen to describe the steric and electronic effects governing the reaction. It is important to note here that these descriptors generally have ***chemical meaning*.** For example, in both QM and MM, electronic descriptors such as *atomic charges* and *dipole moments* may be used to understand catalyst/ligand reactivity.[37, 38] Additionally, more advanced QM descriptors such as *local* and *global reactivity parameters* (obtainable in the cDFT framework - Box 1, see Supporting Information for detailed list of parameters), have often been used to rationalize the reactivity and selectivity of various chemical series in numerous reaction classes.[39]

In the case of statistical or machine learning models, thousands of descriptors may be computed and used for building the model, especially with the advent of specialized software for computing descriptors.[40, 41]

However, without carefully choosing only the most important descriptors that significantly contribute to the prediction, the model is bound to contain a large amount of noise. This affects the accuracy of the predictions. Indeed, different methods to select descriptors[42, 43] (either supervised or unsupervised) have been developed to address this exact issue. These techniques (for a complete breakdown with examples see Supporting Information) include *filtering methods* (selection based on statistical methods like the chi-squared test), *wrapper methods* (selection based on a predictive model to generate the best descriptor combinations), *embedded methods* (selection is made by learning the importance of each feature during model training), *hybrid methods* (combination of filtering and wrapper methods), and *dimensionality reduction techniques* (selection of features after dimensionality reduction of the data).[44] Perhaps the most widespread method is principal component analysis (PCA), a technique that reduces data dimensions to fewer components while retaining essential information about its diversity. This allows for simpler visualization, although interpretation might not always be straightforward.[45]

If one of the goals is not only to develop a predictive model, but also to have chemically interpretable descriptors that can shed light on the mechanism (statistical models), then the descriptor selection requires additional attention. First, chemical knowledge may be used when selecting these. For example, if properties profoundly influencing the reaction outcome have been identified experimentally, descriptors of these properties may be considered. In this case, the ML or statistical model will eventually quantify the impact of these properties. Unfortunately, irrelevant descriptors may still coincidentally correlate with the property the model aims to predict, leading to poorly predicting models. Generally, the assumption is that the different descriptors are independent of each other (as in, the change of one will not influence the other). However, a counter example can be seen in the work of *Werth et al.*[46] on bifunctional hydrogen bond donor (BHD) catalysts, where the NBO charge of the catalyst was indirectly correlated with the pKa value via the LUMO energy and a separate steric parameter.

How the descriptors are computed is another fundamental aspect. Many descriptors are conformation-dependent: descriptors computed only for a single conformer may not adequately represent more flexible ligands. Should a Boltzmann population average be used instead, a conformational search (hence time) must be added to the computation. A significant advantage of some available databases (e.g., Kraken and Oscar described above) is their computed descriptors, which may be used by other model developers.

Number of descriptors. With the help of numerous cheminformatics tools[40, 47, 48] thousands of descriptors can be computed for each model, although the final version of the model will ideally contain less than 10. While ***more*** descriptors are expected to provide a more complete representation of a reaction, they may also lead to overfitting, a common issue when developing ML or statistical models. As an indication of overfitting, the model performs well on the training set but poorly on the testing set, hence the need for significantly dissimilar training and testing sets to detect overtraining. Moreover, not all the descriptors will have a significant enough influence on the accuracy of the model and will add unnecessary noise. In general, the simpler the model (i.e., the fewer descriptors), the easier (and faster) it is to train and often the better (more generalizable) and interpretable the model will be. In practice, many descriptor combinations are evaluated for model training, and the most predictive set of descriptors is chosen for the final version of the model. However, a careful evaluation of the relevance of these descriptors should be carried out.

Box 2. *Different types of descriptors/methods are used to describe unique atomic and molecular properties.*

Steric parameters. Steric effects play an important role in catalyst reactivity and selectivity. In the context of asymmetric catalysis, a catalyst may have different substituents, with each substituent taking up a different volume of the space around the reactive part of the catalyst, influencing the shape, stereoinduction, and reactivity of the catalyst. To account for these effects, descriptors representing the steric character of the molecule and its substituents have been developed. An important aspect of some steric parameters is that they are conformation dependent, making the selection of the conformation impactful. In practice, chemists often rely on C2 symmetrical catalysts (e.g., BINAP) to reduce the conformational space (e.g., the number of possible conformations, the number of different faces of nucleophilic attack), which in turn simplifies their optimization. Depending on the system, these descriptors may be calculated for the lowest energy conformer, the catalytically relevant conformer, or as a Boltzmann-weighted conformation average. **Mentioned in text:** Sterimol parameters, ASO, %buried volume.

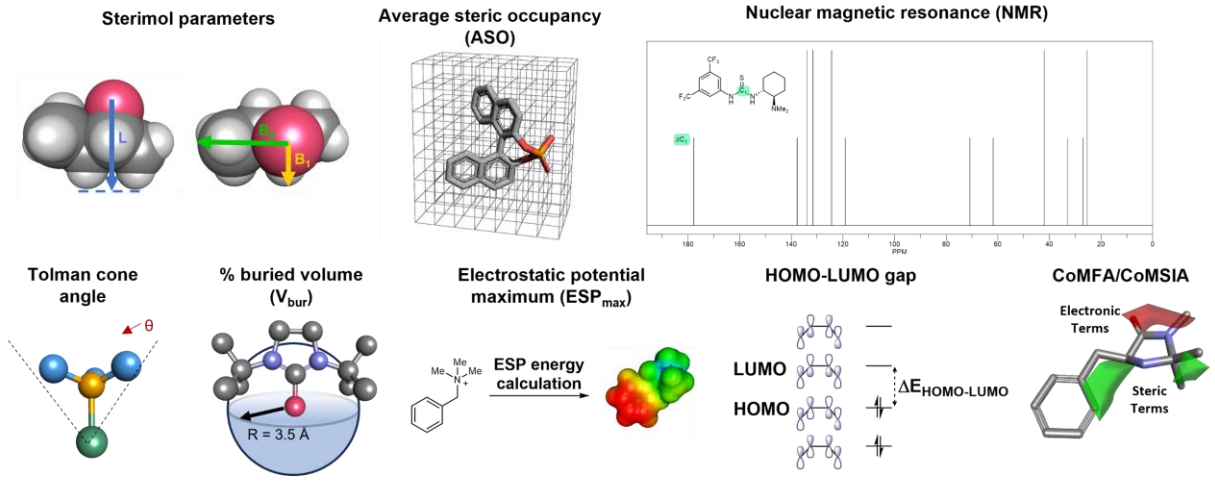
Geometric parameters. In addition to sterics, one can compute descriptors providing information on the 3D shape (geometry) of the molecule. The geometry of a catalyst affects the selectivity and rate of catalyzed reactions; therefore, the selection of the conformation is critical. In addition to overall shapes, geometric parameters may also contain information such as bond lengths or bond-angle and torsion values. **Mentioned in text:** cone angles.

Electronic parameters. These descriptors aim to represent the electronic properties of the molecule, often of a reactive centre (e.g., the atoms participating in the reaction) or of a ligand (e.g., a phosphine ligand modulating the reactivity of metal centers). In organic chemistry, electronic properties describe information such as ability to donate or accept electrons (Lewis basicity and acidity, respectively), nucleophilicity and electrophilicity, hyperconjugation, and more. Unlike steric and geometric parameters, electronic parameters are often less influenced by conformation, and they can either be derived from a single (optimized) conformation or as a Boltzmann-weighted conformational average. Most electronic parameters describe a fragment of the molecule, such as a substituent or reactive atom, rather than the entire molecule. **Mentioned in text:** NBO charges, FMO gap.

Empirical parameters. Descriptors do not always have to be computed theoretically. A variety of empirical descriptors can be used in model development, such as nuclear magnetic resonance (NMR) chemical shifts, NMR ^{31}P tensors, and infrared (IR) spectrum frequencies and vibrations.

Interaction Fields. Comparative molecular field analysis (CoMFA). A method developed for ligand 3D-QSAR studies in drug discovery, which has since been implemented in asymmetric catalysis modeling.[96] The method aims to correlate reaction outcomes to molecular fields described by the steric and electronic properties of a molecule. To achieve this, molecules are first aligned and then placed in a three-dimensional energy grid. A probe atom is then added at strategic points on this grid and the interaction energy (Van der Waals and electrostatic) between the molecules and the probe atom is calculated at each grid point. These energies are the descriptors that input into regression models, most often partial least squares (PLS), which correlate catalytic activity/enantioselectivity with the computed descriptors.[97, 98]

Comparative molecular similarity indices analysis (CoMSIA). CoMSIA is a 3D-QSAR method developed as a natural extension of CoMFA by including molecular similarity in the computation of the molecular fields. In addition to the steric and electronic parameters captured by CoMFA, CoMSIA also includes a hydrophobicity term. The calculation of descriptors is then performed in a similar manner to CoMFA.



Models.

Unknown mechanism/stereoselective step (statistical and ML models). Modeling enantioselectivity for reactions and catalysts with unknown mechanisms relies on quantitative structure-selectivity relationships (QSSR), where statistical models can be used to correlate enantioselectivity to the structure of the catalyst and reaction components.[49-51] In addition to these, the field has evolved to also incorporate ML methods for the development of predictive models.[52, 53] While sometimes used interchangeably in the scientific literature, statistical and ML models are in fact distinct. Statistical models are derived from the whole data (where statistical relevance is measured), while ML models are trained and tested on separate sets. Additionally, statistical models are often more interpretable than ML models.

Depending on the desired outcome, different methods can be used to relate the catalyst (and substrate, in some cases) structure to enantioselectivity. For example, Sigman and co-workers have been developing statistical models based on linear free energy relationships (LFER), where they aim to achieve both an increase in prediction accuracy as well as an intuitive understanding of the potential mechanism.[51, 54] This type of model aims to find a linear correlation between a variable and free energy, which in asymmetric catalysis is often the energy difference between the diastereomeric TSs ($\Delta\Delta G^{\ddagger}$). Since stereoselectivity is dependent on multiple variables, the model would usually be a multivariate linear regression (MLR). The input would be the numerical representations of the reaction components (descriptors), and the output would be the free energy difference.[52, 55] These models have been successfully applied on several metal catalyzed reactions such as the Pd-catalyzed enantioselective aryl-carbonylation of sulfonimidamides,[56] Pd-catalyzed Hayashi-Heck reaction,[57] Negishi coupling,[58] and different Pd- and Ni-catalyzed cross-coupling reactions.[59] Additionally, they have been applied to organocatalyzed reactions including the Mannich reaction,[43] chiral phosphoric acid catalyzed nucleophilic addition to iminiums,[17, 60, 61] HBD catalyzed addition of nucleophiles to nitro alkenes,[46] and others.[54, 55]

A different type of model used successfully in asymmetric catalysis is support vector regression (SVR). With such an ML algorithm, the data may be correlated to the energy difference (enantioselectivity). In SVR the data points are correlated by a linear line, or a higher dimension curve, within a predetermined margin. The goal of the model is to identify the line that fits most data points which falls within a predetermined margin (prediction error).[62-64] Such models have been successfully applied to organocatalyzed reactions such as the chiral phosphoric acid catalyzed thiol-nucleophiles addition to imines.[65]

The major difference between MLR and SVR is that SVR does not have an underlying assumption of a linear relationship between the data points (descriptors) and the outcome (enantioselectivity), and therefore, can be more suitable for modeling complex reactions. On the other hand, MLR can be more informative for the interpretation of the influence of each descriptor (e.g., catalyst structure, solvent, temperature), as the coefficients also contribute to their weight of influence. An SVR model is used for predicting an outcome, and the model itself, often used as a black box, is not easily interpreted. Each model has its strengths and weaknesses, and they are used based on the main goals of the modeling project as well as the existing data. It is worth noting that other regression models such as kernel ridge and partial least squares (PLS, also referred to as projection to latent structures) have also been used.[63, 65-67]

Once a predictive model has been developed and tested, it ***should*** then be adopted by organic chemists. As the field remains in its infancy, most published ML models and methods are only made available in the form of scripts and methodologies. However, there is no available user-friendly package that can be used or models that can easily be trained for novel reactions by organic chemists with minimal expertise in computer science, although this may soon change.[68]

Using knowledge of potential TS structures (mechanism-based approach). When a hypothesis exists for the diastereomeric TSs of the stereoselective step, these TSs may be modeled and used directly to extract the energy difference between the diastereomeric TSs ($\Delta\Delta G^{‡}$), and consequently to compute selectivity. Although many methods to model TSs exist, this review is not meant to be exhaustive and will focus on the most recent and advanced methods available, as well as their applications to stereoselectivity predictions. For more detailed information on TS modeling for the prediction of enantioselectivity, we refer the readers to previous reviews.[69-71]

Calculating the stereoselectivity from the energy difference $\Delta\Delta G^{‡}$ assumes that the reaction is under Curtin-Hammett control (**Figure 3**). Simply put, the product ratio (e.g., *R*:*S*) reflects the energy difference between the two competing and irreversible diastereomeric TSs (i.e., $\Delta\Delta G^{‡}$).

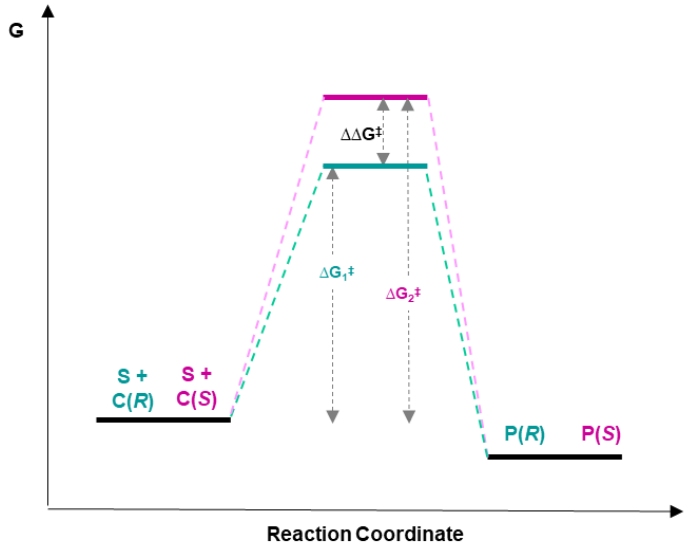


Figure 3. Curtin-Hammett principle in the context of asymmetric catalysis. C: catalyst, S: substrate, P: product, either *R* or *S* enantiomer. S+C(*S*) and S+C(*R*) are the catalyst substrate complexes leading to either the *S* or the *R* product, respectively.

Since the difference in $\Delta\Delta G^{‡}$ between a moderately selective catalyst (~80 %*ee*) and an excellent catalyst (>97 %*ee*) is about 1 kcal/mol, there is a requirement for the methods to be accurate enough to be able to distinguish between them. Hence, a prediction error within 1 kcal/mol is targeted. While high level calculations may fulfill this criterion, an objective is to obtain predictions quicker than experimental data and using less expensive equipment. Unfortunately, the most accurate methods for modelling TSs (e.g., MP2, DFT) are demanding in both computational resources and time. It follows that they can hardly compete with high throughput experimentation which provides true data (not predictions) in a time-efficient manner. As a result, these high-level calculation methods are primarily used to investigate reactions *post facto* rather than for designing novel catalysts.[72] Thus, faster alternative methods are necessary.

The energy difference can be calculated from single, lowest lying conformers for each diastereomeric TS identified using a conformational search algorithm.[73] Alternatively, multiple thermally accessible conformations, and their Boltzmann population distribution, can be used to obtain the energy difference. In both cases, the challenge of TS structure-based approaches is to identify and optimize ***all*** potential diastereomeric TSs of the reaction under investigation. In practice, the simplicity of the Curtin-Hammet principle is overshadowed by the number of possible TS conformations, or reactions for which multiple

steps could be rate-determining, or for which multiple competing mechanisms leading to opposite enantiomeric products exist.[74] For these methods to be user-friendly, an automated conformational search algorithm is needed.

Automated conformational sampling of TSs lies at the heart of tools like VIRTUAL CHEMIST/ACE, CatVS/Q2MM, QChASM/AARON (Quantum Chemistry Automation and Structure Manipulation/An Automated Reaction Optimizer for New catalysts),[75, 76] or the chemical steering wheel.[77] VIRTUAL CHEMIST/ACE, developed in our research group, is a self-contained, graphics user interface (GUI)-based asymmetric catalyst design platform.[26] Designed with a chemist's needs in mind, the underlying MM methodology has been thoroughly tested on seven widespread metal and organocatalyzed reactions, with an overall accuracy of ~ 1 kcal/mol (Box 3). To note, VIRTUAL CHEMIST was also applied to several scenarios that an experimental chemist might face in his project: one-by-one catalyst design, screening a library of catalysts, catalyst lead optimization through analogue search (detailed in the ***Evaluation of the Models, Methods, and Applications*** section), and identifying the substrate scope of a known catalyst, with demonstrated advantages over traditional asymmetric catalyst design.

A similar asymmetric catalyst design platform is CatVS/Q2MM (Box 3), primarily focused on organometallic catalysts. Like VIRTUAL CHEMIST/ACE, CatVS/Q2MM is an MM-based method and was first benchmarked on known metal-catalyzed reactions (see Box 3), followed by its application in a "real-world" scenario, which yielded stereoselective ligands for the Rh-catalysed asymmetric hydrogenation of enamides (discussed in more detail in the ***Evaluation of the Models, Methods, and Applications*** section). Unlike VIRTUAL CHEMIST, the free version of CatVS does not include an interface and the calculations must be run from the command line environment.

Apart from VIRTUAL CHEMIST and CatVS, another virtual platform for catalyst design is QChASM/AARON (Box 3). Contrary to both VIRTUAL CHEMIST and CatVS, QChASM/AARON is an interface to various open-source tools for structural manipulation, TS search and optimization, as well as free energy calculations for *%ee* determination. While VIRTUAL CHEMIST and CatVS are primarily based on MM, the geometry optimization and energy calculations available through the QChASM/AARON interface are based on QM methods (either SEQM or DFT) accessible through software such as Gaussian,[78] Psi4,[79] or ORCA.[80] QChASM employs a GUI plugin for the Chimera visualizer,[81] which benefits experimental chemists with little to no expertise in command line environments.

Box 3. *Computational platforms for asymmetric catalyst design.*

Asymmetric Catalyst Evaluation (ACE). ACE is an MM-based software that predicts the stereochemical outcome of asymmetric reactions by modeling the relevant TSs of ligand/substrate/catalyst systems. The stereoinducing step for these reactions must be known *a priori*. Part of the larger **VIRTUAL CHEMIST** platform for asymmetric catalyst design, ACE is built on two fundamental organic chemistry principles: **1)** the Hammond-Leffler postulate and **2)** the Curtin-Hammett principle. The TSs are built in accordance with principle **1)** (i.e., the TS is most similar to the species to which it is closest in energy, either reactants or products, hence is a linear combination of reactant and product structures), while the enantiomeric excess is calculated according to principle **2)** (i.e., the %ee is determined according to the difference in energies between diastereomeric TSs). The preferred stereoisomers are determined through a genetic algorithm that efficiently samples the conformational space around the ligand/substrate/catalyst system. ACE has been successfully tested on seven organo- and metal-catalyzed reactions commonly employed in asymmetric synthesis: Diels-Alder cycloaddition (with chiral auxiliaries and organocatalysts), Aldol reaction, Shi epoxidation, OsO_4-based dihydroxylation of alkenes, $ZnEt_2$-addition to aldehydes, and Rh-catalyzed hydrogenation of enamides, achieving accuracies of ~ 1 kcal/mol compared to experimental values.[26]

Quantum-guided molecular mechanics (Q2MM). Q2MM is an MM-based methodology that uses automated FF parametrization to describe TSs and predict the outcome of stereochemical reactions. Similar to ACE, Q2MM is part of a larger catalyst design platform called **CatVS**. To date, CatVS/Q2MM has been primarily employed for organometallic catalysis, with the tested reactions involving Rh-catalyzed hydrogenation of enamides, OsO_4 dihydroxylation of alkenes, $ZnEt_2$-addition to aldehydes, Pd-catalyzed allylation, asymmetric redox-relay Heck reaction, and Ru-catalyzed hydrogenation of ketones.[83, 99, 100] The FFs generated by Q2MM are reaction-specific and are known as TSFFs. Similar to ACE, the stereoinducing step must be known *a priori*. However, in contrast to ACE, Q2MM relies on reference data for a training set of model TSs that is subjected to QM calculations to determine the necessary parameters for FF parametrization. The uniqueness of Q2MM relies on the usage of the QM-derived Hessian matrix (i.e., the variations in energy with respect to geometry changes) to fit TSFF force constants for bonded parameters. Once the TSFF has been generated and validated for a reaction, Monte Carlo (MC) conformational searches are employed to find the relevant TSs and stereoisomers. The %ee's are calculated by Boltzmann-averaging the relative energies of the identified conformations. Q2MM has been tested on four metal-catalyzed reactions, achieving correlation coefficients between 0.8-0.9 between predicted and experimental data.

An Automated Reaction Optimizer for New catalysts (AARON). In contrast to both ACE and Q2MM, AARON is an open-source framework that interfaces various tools for structural manipulation, TS searches, and energy calculations. However, like ACE and Q2MM, AARON is part of a larger toolkit named **QChASM**. Designed with ease-of-use in mind, AARON uses a library of TS templates to construct TSs of novel ligand/substrate/catalyst systems, followed by TS optimization at a desired level of theory (semiempirical methods or DFT). Once the TSs have been located, conformational sampling is performed using a rule-based methodology that accounts for the torsional preferences of each substituent. These conformers are then subjected to thermochemistry calculations to obtain free energies, which are Boltzmann-averaged over the populations of conformers leading to specific enantiomers to predict the %ee. Representative applications of AARON include Pd-catalyzed Heck allenylation, Rh-catalyzed hydrogenation of enamides, and the Lewis-base promoted propargylation of aromatic aldehydes.

Evaluation of the Models, Methods, and Applications

This section is not meant to be exhaustive but, rather, to illustrate different uses of these methods.

Catalyst design. The use of computers for asymmetric catalyst design has been a very promising field for two decades. For example, as early as 2003, Kozlowski and co-workers developed a model based on interaction fields (see Box 2) for dialkylzinc addition to aldehydes catalyzed by β-amino alcohols and applied it to identify novel catalysts.[82] However, unexpectedly, twenty years later, while more validated methods are now available to the organic chemistry community, the applications to new catalysts design by groups other than the developers are still scarce.

A representative example of catalyst design is the application of CatVS to the investigation and discovery of novel catalysts.[20] TSFFs for several reactions including asymmetric dihydroxylation and rhodium-catalyzed asymmetric hydrogenation had previously been developed using Q2MM.[83] When Q2MM was integrated into CatVS, real-world case studies were carried out. Prediction of the (DHQD)2PHAL-catalyzed dihydroxylation of a dozen substrates revealed a mean unsigned error of about 0.6 kcal/mol, while the screening of rhodium ligands for Rh-catalyzed asymmetric hydrogenation of enamides was also

performed. Remarkably, CatVS/Q2MM was able to distinguish between highly stereoselective and poorly stereoselective ligands. For one substrate, the four most stereoselective ligands were among the top 5 predictions. Interestingly, the use of implicit solvent was not found to improve the accuracy, in line with what was observed with ACE.[73]

Stereoselectivity and catalytic activity optimization. A key aspect of reaction optimization is tuning the enantioselectivity without compromising reactivity. This is especially important in the synthesis of active pharmaceutical ingredients (APIs), as the final product has strict purity requirements. For this reason, high yields of the desired product with low catalyst loading are of significance. An interesting application of this concept has been developed by Dotson *et al.*,[57] who designed a computational workflow to fine-tune enantioselectivity while simultaneously accounting for catalyst/ligand reactivity in two metal-catalyzed reactions: **1)** Pd-catalyzed Hayashi-Heck and **2)** Rh-catalyzed alkene hydroformylation. These reactions use chiral bisphosphine ligands and are pharmaceutically relevant due to their use in the synthesis of a transient receptor potential ankyrin 1 (TRPA1) antagonist (**Figure 4A**, Compound 20).[84]

The Dotson workflow began with assembling a set of over 550 chiral bisphosphine ligands, for which steric, electronic, and geometric descriptors were calculated with QM. For each reaction, a subset of ligands was selected for experiments to determine the regio-, enantioselectivity, and reaction yields/conversion. The latter was used to discriminate between reactive and unreactive complexes using two classification algorithms: **a)** a single-node decision tree for reaction **1)**[55] and **b)** a logistic regression classification algorithm for reaction **2)**. Consequently, for each reaction, the descriptors and associated experimental data of the reactive complexes were used to train a reaction/metal-agnostic MLR model capable of correlating input data to regio- and enantioselectivity. To verify whether the workflow can be used to prospectively screen for high conversion/high enantioselectivity ligands, the last step involved a virtual screen on the database of ligands not involved in training of the classification or MLR models. Applying the developed classification and MLR models on this database led to the identification of several ligands with excellent experimental conversion and enantioselectivity (**Figure 4A**).

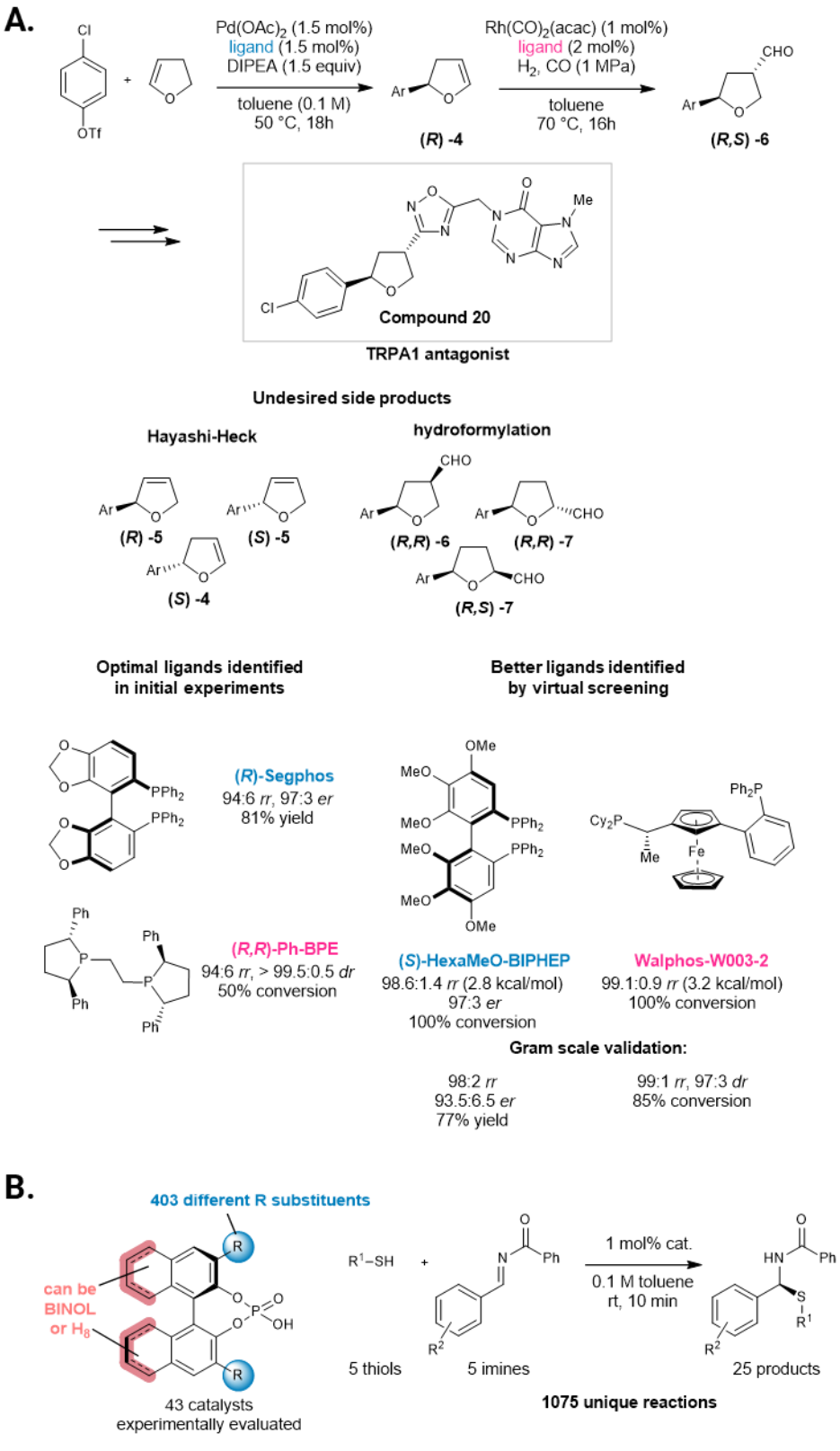


Figure 4. A) Dotson workflow. B) Rinehart workflow.

Similar to Dotson *et al.*, Rinehart *et al.* developed a model to predict enantioselectivity, with a focus on the chiral phosphoric acid (CPA) catalyzed thiol nucleophilic addition to imines.[63, 65] The goals of this work were to **1)** describe the components of the reactions using descriptors that are agnostic to the mechanism and **2)** develop a predictive SVR model without an assumption of a shared mechanism between the different data points. The final model related the catalyst structure to its function (enantioselectivity), in any reaction catalyzed by the input catalyst scaffold.

For this reason, the descriptors of choice were ASO for the portrayal of steric information, and ESP_{max} for the electronic information.[19, 63, 85] These descriptors are more abstract than the ones commonly used when building an MLR model (e.g., NBO charges, cone angles). However, the emphasis of this work is on descriptors that have been previously shown to work well,[85] as opposed to exploring a wide range of descriptors that relate to the reaction mechanism. Therefore, the choice of an SVR model was also appropriate. As mentioned in the section ***Models***, the descriptors appearing in the final SVR model cannot generally be used to gain insights into the reaction mechanism and the factors that influence enantioselectivity. Thus, there is no underlying assumption of a shared mechanism/stereoselective step. Once the descriptor library was ready, the next step in the workflow was the use of an algorithm to divide the dataset into training and testing sets, as the more diverse the data the model is trained on, the more likely it is to be transferable to new data points. The training set, termed a universal training set (UTS), represents the variability of the chemical space of the full library.[66]

With a library of 1,075 unique reactions consisting of 43 chiral phosphoric acids, 5 thiol nucleophiles, and 5 imine electrophiles (25 possible products) the bulk of the work consisted of developing and testing different descriptor combinations (steric and electronic) with different models, as well as developing the algorithm for training set selection. Eventually the best performing model and descriptors (SVR with ASO and ESP_{MAX}) were chosen. Interestingly, Rinehart *et al* demonstrated the ability of the model to predict highly enantioselective catalysts, even when the training set consists of data points of 80 %*ee* or less. This is a significant achievement, as most asymmetric catalyst developments start with only lower selectivity catalysts.

The last representative example we shall discuss in this sub-section was described by our research group during the validation of VIRTUAL CHEMIST.[26] In this example, we replicated *in silico* the excellent experimental study by Gerosa *et al.*[86] that aimed to identify selective chiral pyrrolidines as organocatalysts for the Diels-Alder cycloaddition. In this report, 22 catalysts were synthesized and tested for their ability to catalyze the Diels-Alder cycloaddition between (E)-cinnamaldehyde and cyclopentadiene. The preparation of these potential levoglucosenone-derived organocatalysts required complex synthesis, separation, and characterization of stereoisomers. We developed a workflow using the modular workflow interface in VIRTUAL CHEMIST to simulate the entire process including the **1)** parallel synthesis of a small library (ca. 500+) of these organocatalysts, including the ones tested by Gerosa *et al.* and **2)** evaluation of the induced stereoselectivities. We demonstrated that VIRTUAL CHEMIST not only reproduced the process successfully within just a few days, but accurately identified the most stereoselective catalysts determined experimentally.

Guiding asymmetric synthesis. An important aspect about model development is its transferability to out-of-set reaction components, and most importantly the application of the model for a synthesis project. The work by Betinol *et al.*,[60] exemplifies this scenario, by demonstrating how previously developed statistical models can be extrapolated to structurally diverse substrates. Four goals were set at the onset of this research project: **1)** application of an existing model to the synthesis of complex scaffolds; **2)** application to the synthesis of a natural product; **3)** preferred catalyst for a given reaction; **4)** preferred route (reaction

type) for the synthesis of a given product. For Goal **#1** *Betinol et al.* investigated the CPA catalyzed nucleophilic addition to iminiums.[17, 61] The model was tested on three reported reactions for the functionalization of indoles (as relevant scaffolds for biological compounds) that are catalyzed by CPAs. Importantly, these three reactions are not represented in the training set. Nonetheless, the model was able to predict the enantioselectivity with excellent accuracies (average errors between 0.29-0.54 kcal/mol). Next, for Goal **#2** Betinol *et al.* tested a model designed for secondary amine catalyzed reactions,[87] and demonstrated the transferability of their model to new reactions with more complicated conditions that were not represented in their training set.

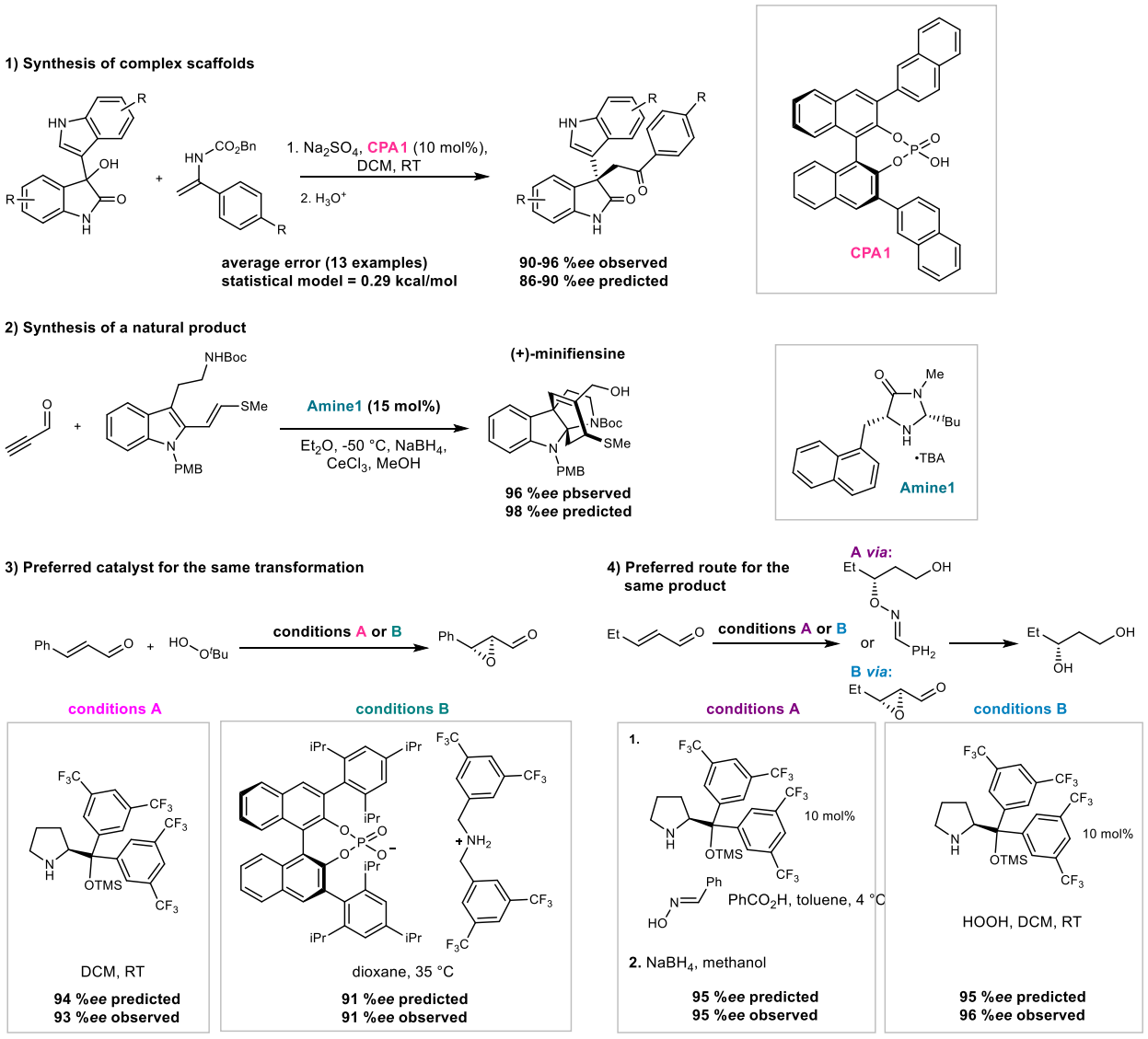


Figure 5. Betinol and co-workers' work on the application of existing models to different scenarios.

With both models successfully extrapolated to more complex reactions, Betinol *et al.* moved on to goals **#3** and **#4**. Selecting the optimal catalyst for a given reaction (Goal **#3)** was tested on the asymmetric epoxidation of cinnamaldehyde, while selecting the optimal catalyst to synthesize a product (Goal **#4**) was evaluated on the synthesis of diols via two different pathways. The results of both studies were highly encouraging, with predictions being within 1 %*ee* of experimental results.

Conclusions and Perspectives

With the advancement of methods and algorithms for predicting catalyst enantioselectivities, computational asymmetric catalysis is a ripe area for further research. In general, the different approaches used to predict these enantioselectivities are broadly distinguished as ones where the mechanism is unknown (statistical models and ML), and ones where TS information for the enantioselective step is (partially) known (QM, MM). Regrettably, the field still requires some understanding of the underlying computational methods and theory before being adopted integrally by experimentalists. Consequently, practical applications of these methods are yet to come. This situation leads most organic chemistry laboratories to continue with employing the conventional, albeit laborious and time-intensive technique of stepwise optimization. As computational methods continue to evolve and their accessibility improves, we envision a future where these tools will be completely integrated in the toolbox of experimentalists, and where the trained models or platforms will be able to improve the discovery rate and unveil new insights in asymmetric catalysis.

With the help of high-throughput experimentation (HTE) for asymmetric catalysis, reproducibility and reliability of data will increase and facilitate the development and integration of predictive computational tools. Automated HTE systems have played a crucial role in accelerating catalyst screening processes, generating vast datasets for diverse reaction conditions.[88] Integrating computational models, as we have described throughout this review, alongside high-throughput screening will not only help overcome limitations related to the number of variables that can be tested (temperature, solvent etc.), but will also contribute to the generalization and robustness of the models developed on datasets gathered under the same experimental conditions. This intersection of automated experimental and computational approaches can then enhance the global efficiency of asymmetric catalysis research, leading to more accurate catalyst design strategies and predictive models.

Supporting Information

Commonly computed descriptors in the conceptual DFT framework (Table S1). Examples of methods to selected features for training ML models (Table S2). A list of available datasets for asymmetric catalysis, curated from the literature (Excel format).

Acknowledgements

We thank NSERC (Discovery programme) for financial support.

Author contributions

SP, MBP, and NM devised the structure of the review. SP collected, curated, sorted, and categorized the references. SP, JG, MBP, NM curated the datasets. SP led the writing of the review (including creating the figures), while JG, MBP and NM contributed to the writing of this manuscript.

ORCID

Sharon Pinus – 0000-0001-9771-3098
Jerome Genzling – 0009-0007-4728-1478
Mihai Burai-Patrascu – 0000-0001-9289-7887
Nicolas Moitessier – 0000-0001-6933-2079

Conflict of Interest

VIRTUAL CHEMIST is distributed by Molecular Forecaster (free of charge for academic research) co-founded by NM. MBP is a senior scientist at Molecular Forecaster.

References

(1) Moitessier, N.; Pottel, J.; Therrien, E.; Englebienne, P.; Liu, Z.; Tomberg, A.; Corbeil, C. R. Medicinal Chemistry Projects Requiring Imaginative Structure-Based Drug Design Methods. *Accounts of Chemical Research* **2016**, *49* (9), 1646-1657. DOI: 10.1021/acs.accounts.6b00185.

(2) Shen, Y.; Borowski, J. E.; Hardy, M. A.; Sarpong, R.; Doyle, A. G.; Cernak, T. Automation and computer-assisted planning for chemical synthesis. *Nature Reviews Methods Primers* **2021**, *1* (1). DOI: 10.1038/s43586-021-00022-5.

(3) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the space of chemical reactions using attention-based neural networks. *Nature machine intelligence* **2021**, *3* (2), 144-152.

(4) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Accounts of Chemical Research* **2018**, *51* (5), 1281-1289. DOI: 10.1021/acs.accounts.8b00087.

(5) Molga, K.; Szymkuć, S.; Grzybowski, B. A. Chemist Ex Machina: Advanced Synthesis Planning by Computers. *Accounts of Chemical Research* **2021**, *54* (5), 1094-1106. DOI: 10.1021/acs.accounts.0c00714.

(6) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical Science* **2020**, *11* (12), 3316-3325, 10.1039/C9SC05704H. DOI: 10.1039/C9SC05704H.

(7) Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; et al. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **2018**, *4* (3), 522-532. DOI: https://doi.org/10.1016/j.chempr.2018.02.002.

(8) Mikulak-Klucznik, B.; Gołębiowska, P.; Bayly, A. A.; Popik, O.; Klucznik, T.; Szymkuć, S.; Gajewska, E. P.; Dittwald, P.; Staszewska-Krajewska, O.; Beker, W.; et al. Computational planning of the synthesis of complex natural products. *Nature* **2020**, *588* (7836), 83-88. DOI: 10.1038/s41586-020-2855-y.

(9) Genheden, S. E., O.; Bjerrum, E. J. A Quick Policy to Filter Reactions Based on Feasibility in AI-Guided Retrosynthetic Planning. *ChemRxiv. This content is a preprint and has not been peer-reviewed.* **2020**. DOI: https://doi.org/10.26434/chemrxiv.13280495.v1.

(10) Żurański, A. M.; Martinez Alvarado, J. I.; Shields, B. J.; Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Accounts of Chemical Research* **2021**, *54* (8), 1856-1865. DOI: 10.1021/acs.accounts.0c00770.

(11) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Machine Learning: Science and Technology* **2021**, *2* (1), 015016. DOI: 10.1088/2632-2153/abc81d.

(12) Anstine, D. M.; Isayev, O. Generative Models as an Emerging Paradigm in the Chemical Sciences. *Journal of the American Chemical Society* **2023**, *145* (16), 8736-8750. DOI: 10.1021/jacs.2c13467.

(13) 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* (6385), 186-190. DOI: 10.1126/science.aar5169.

(14) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chemical Science* **2021**, *12* (6), 2198-2208, 10.1039/D0SC04823B. DOI: 10.1039/D0SC04823B.

(15) Jorner, K.; Tomberg, A.; Bauer, C.; Sköld, C.; Norrby, P.-O. Organic reactivity from mechanism to machine learning. *Nature Reviews Chemistry* **2021**, *5* (4), 240-255. DOI: 10.1038/s41570-021-00260-x.

(16) Zhao, S.; Gensch, T.; Murray, B.; Niemeyer, Z. L.; Sigman, M. S.; Biscoe, M. R. Enantiodivergent Pd-catalyzed C–C bond formation enabled through ligand parameterization. *Science* **2018**, *362* (6415), 670-674. DOI: 10.1126/science.aat2299

(17) Reid, J. P.; Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **2019**, *571* (7765), 343-348. DOI: 10.1038/s41586-019-1384-z.
(18) Reid, J. P.; Sigman, M. S. Comparing quantitative prediction methods for the discovery of small-molecule chiral catalysts. *Nature Reviews Chemistry* **2018**, *2* (10), 290-305. DOI: 10.1038/s41570-018-0040-8.
(19) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **2019**, *363* (6424), eaau5631. DOI: 10.1126/science.aau5631.
(20) Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; et al. Rapid virtual screening of enantioselective catalysts using CatVS. *Nature Catalysis* **2018**, *2* (1), 41-45. DOI: 10.1038/s41929-018-0193-3.
(21) Gensch, T.; dos Passos Gomes, G.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S.; et al. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *Journal of the American Chemical Society* **2022**, *144* (3), 1205-1217. DOI: 10.1021/jacs.1c09718.
(22) Bell, E. L.; Finnigan, W.; France, S. P.; Green, A. P.; Hayes, M. A.; Hepworth, L. J.; Lovelock, S. L.; Niikura, H.; Osuna, S.; Romero, E.; et al. Biocatalysis. *Nature Reviews Methods Primers* **2021**, *1* (1), 46. DOI: 10.1038/s43586-021-00044-z.
(23) Pyser, J. B.; Chakrabarty, S.; Romero, E. O.; Narayan, A. R. H. State-of-the-Art Biocatalysis. *ACS Central Science* **2021**, *7* (7), 1105-1116. DOI: 10.1021/acscentsci.1c00273.
(24) Finnigan, W.; Hepworth, L. J.; Flitsch, S. L.; Turner, N. J. RetroBioCat as a computer-aided synthesis planning tool for biocatalytic reactions and cascades. *Nature Catalysis* **2021**, *4* (2), 98-104. DOI: 10.1038/s41929-020-00556-z.
(25) Gallarati, S.; van Gerwen, P.; Laplaza, R.; Vela, S.; Fabrizio, A.; Corminboeuf, C. OSCAR: an extensive repository of chemically and functionally diverse organocatalysts. *Chemical Science* **2022**, *13* (46), 13782-13794, 10.1039/D2SC04251G. DOI: 10.1039/D2SC04251G.
(26) Burai Patrascu, M.; Pottel, J.; Pinus, S.; Bezanson, M.; Norrby, P. O.; Moitessier, N. From desktop to benchtop with automated computational workflows for computer-aided design in asymmetric catalysis. *Nat. Catal.* **2020**, *3* (7), 574-584, Article. DOI: 10.1038/s41929-020-0468-3.
(27) Gallegos, L. C.; Luchini, G.; St. John, P. C.; Kim, S.; Paton, R. S. Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties. *Accounts of Chemical Research* **2021**, *54* (4), 827-836. DOI: 10.1021/acs.accounts.0c00745.
(28) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catalysis* **2019**, *9* (3), 2313-2323. DOI: 10.1021/acscatal.8b04043.
(29) Warshel, A.; Weiss, R. M. An empirical valence bond approach for comparing reactions in solutions and in enzymes. *Journal of the American Chemical Society* **1980**, *102* (20), 6218-6226.
(30) Kim, Y.; Corchado, J. C.; Villà, J.; Xing, J.; Truhlar, D. G. Multiconfiguration molecular mechanics algorithm for potential energy surfaces of chemical reactions. *The Journal of Chemical Physics* **2000**, *112* (6), 2718-2735. DOI: 10.1063/1.480846.
(31) Jensen, F. Locating minima on seams of intersecting potential energy surfaces. An application to transition structure modeling. *Journal of the American Chemical Society* **1992**, *114* (5), 1596-1603.
(32) Pairault, N.; Zhu, H.; Jansen, D.; Huber, A.; Daniliuc, C. G.; Grimme, S.; Niemeyer, J. Heterobifunctional Rotaxanes for Asymmetric Catalysis. *Angewandte Chemie International Edition* **2020**, *59* (13), 5102-5107. DOI: https://doi.org/10.1002/anie.201913781
(33) Minenkov, Y.; Sharapa, D. I.; Cavallo, L. Application of Semiempirical Methods to Transition Metal Complexes: Fast Results but Hard-to-Predict Accuracy. *Journal of Chemical Theory and Computation* **2018**, *14* (7), 3428-3439. DOI: 10.1021/acs.jctc.8b00018.
(34) Gallarati, S.; Laplaza, R.; Corminboeuf, C. Harvesting the fragment-based nature of bifunctional organocatalysts to enhance their activity. *Organic Chemistry Frontiers* **2022**, *9* (15), 4041-4051, 10.1039/D2QO00550F. DOI: 10.1039/D2QO00550F.
(35) Harper, K. C.; Bess, E. N.; Sigman, M. S. Multidimensional steric parameters in the analysis of asymmetric catalytic reactions. *Nature Chemistry* **2012**, *4* (5), 366-374. DOI: 10.1038/nchem.1297.

(36) 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 Prevymis™ (letermovir). *Chemical Science* **2018**, *9* (34), 6922-6927, 10.1039/C8SC02089B. DOI: 10.1039/C8SC02089B.
(37) Maji, R.; Mallojjala, S. C.; Wheeler, S. E. Electrostatic Interactions in Asymmetric Organocatalysis. *Accounts of Chemical Research* **2023**, *56* (14), 1990-2000. DOI: 10.1021/acs.accounts.3c00198.
(38) Wang, S.; Jiang, J. Interpretable Catalysis Models Using Machine Learning with Spectroscopic Descriptors. *ACS Catalysis* **2023**, *13* (11), 7428-7436. DOI: 10.1021/acscatal.3c00611.
(39) Liu, S. *Conceptual density functional theory: Towards a new chemical reactivity theory*; John Wiley & Sons, 2022.
(40) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **2011**, *32* (7), 1466-1474. DOI: https://doi.org/10.1002/jcc.21707.
(41) *Open-Source Cheminformatics Software*. https://www.rdkit.org/.
(42) See, X. Y.; Wen, X.; Wheeler, T. A.; Klein, C. K.; Goodpaster, J. D.; Reiner, B. R.; Tonks, I. A. Iterative Supervised Principal Component Analysis Driven Ligand Design for Regioselective Ti-Catalyzed Pyrrole Synthesis. *ACS Catalysis* **2020**, *10* (22), 13504-13517. DOI: 10.1021/acscatal.0c03939.
(43) Betinol, I. O.; Lai, J.; Thakur, S.; Reid, J. P. A Data-Driven Workflow for Assigning and Predicting Generality in Asymmetric Catalysis. *Journal of the American Chemical Society* **2023**, *145* (23), 12870-12883. DOI: 10.1021/jacs.3c03989.
(44) Pudjihartono, N.; Fadason, T.; Kempa-Liehr, A. W.; O'Sullivan, J. M. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics* **2022**, *2*, Review. DOI: 10.3389/fbinf.2022.927312.
(45) Bro, R.; Smilde, A. K. Principal component analysis. *Analytical Methods* **2014**, *6* (9), 2812-2831, 10.1039/C3AY41907J. DOI: 10.1039/C3AY41907J.
(46) Werth, J.; Sigman, M. S. Connecting and Analyzing Enantioselective Bifunctional Hydrogen Bond Donor Catalysis Using Data Science Tools. *Journal of the American Chemical Society* **2020**, *142* (38), 16382-16391. DOI: 10.1021/jacs.0c06905.
(47) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; et al. Virtual Computational Chemistry Laboratory – Design and Description. *Journal of Computer-Aided Molecular Design* **2005**, *19* (6), 453-463. DOI: 10.1007/s10822-005-8694-y.
(48) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *Journal of Cheminformatics* **2018**, *10* (1), 4. DOI: 10.1186/s13321-018-0258-y.
(49) Zahrt, A. F.; Athavale, S. V.; Denmark, S. E. Quantitative Structure-Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future. *Chem Rev* **2020**, *120* (3), 1620-1689. DOI: 10.1021/acs.chemrev.9b00425.
(50) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *Journal of the American Chemical Society* **1997**, *119* (43), 10509-10524. DOI: 10.1021/ja9718937.
(51) Williams, W. L.; Zeng, L.; Gensch, T.; Sigman, M. S.; Doyle, A. G.; Anslyn, E. V. The Evolution of Data-Driven Modeling in Organic Chemistry. *ACS Central Science* **2021**, *7* (10), 1622-1637. DOI: 10.1021/acscentsci.1c00535.
(52) Sigman, M. S.; Harper, K. C.; Bess, E. N.; Milo, A. The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Accounts of Chemical Research* **2016**, *49* (6), 1292-1301. DOI: 10.1021/acs.accounts.6b00194.
(53) Zahrt, A. F.; Henle, J. J.; Denmark, S. E. Cautionary Guidelines for Machine Learning Studies with Combinatorial Datasets. *ACS Combinatorial Science* **2020**, *22* (11), 586-591. DOI: 10.1021/acscombsci.0c00118.
(54) Crawford, J. M.; Kingston, C.; Toste, F. D.; Sigman, M. S. Data Science Meets Physical Organic Chemistry. *Accounts of Chemical Research* **2021**, *54* (16), 3136-3148. DOI: 10.1021/acs.accounts.1c00285.

(55) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chemical Science* **2018**, *9* (9), 2398-2412, 10.1039/C7SC04679K. DOI: 10.1039/C7SC04679K.
(56) van Dijk, L.; Haas, B. C.; Lim, N.-K.; Clagg, K.; Dotson, J. J.; Treacy, S. M.; Piechowicz, K. A.; Roytman, V. A.; Zhang, H.; Toste, F. D.; et al. Data Science-Enabled Palladium-Catalyzed Enantioselective Aryl-Carbonylation of Sulfonimidamides. *Journal of the American Chemical Society* **2023**, *145* (38), 20959-20967. DOI: 10.1021/jacs.3c06674.
(57) Dotson, J. J.; van Dijk, L.; Timmerman, J. C.; Grosslight, S.; Walroth, R. C.; Gosselin, F.; Püntener, K.; Mack, K. A.; Sigman, M. S. Data-Driven Multi-Objective Optimization Tactics for Catalytic Asymmetric Reactions Using Bisphosphine Ligands. *Journal of the American Chemical Society* **2023**, *145* (1), 110-121. DOI: 10.1021/jacs.2c08513.
(58) Xu, J.; Grosslight, S.; Mack, K. A.; Nguyen, S. C.; Clagg, K.; Lim, N.-K.; Timmerman, J. C.; Shen, J.; White, N. A.; Sirois, L. E.; et al. Atroposelective Negishi Coupling Optimization Guided by Multivariate Linear Regression Analysis: Asymmetric Synthesis of KRAS G12C Covalent Inhibitor GDC-6036. *Journal of the American Chemical Society* **2022**. DOI: 10.1021/jacs.2c09917.
(59) Newman-Stonebraker, S. H.; Smith, S. R.; Borowski, J. E.; Peters, E.; Gensch, T.; Johnson, H. C.; Sigman, M. S.; Doyle, A. G. Univariate classification of phosphine ligation state and reactivity in cross-coupling catalysis. *Science* **2021**, *374* (6565), 301-308. DOI: 10.1126/science.abj4213.
(60) Betinol, I. O.; Kuang, Y.; Reid, J. P. Guiding Target Synthesis with Statistical Modeling Tools: A Case Study in Organocatalysis. *Organic Letters* **2022**, *24* (7), 1429-1433. DOI: 10.1021/acs.orglett.1c04134.
(61) Shoja, A.; Zhai, J.; Reid, J. P. Comprehensive Stereochemical Models for Selectivity Prediction in Diverse Chiral Phosphate-Catalyzed Reaction Space. *ACS Catalysis* **2021**, *11* (19), 11897-11905. DOI: 10.1021/acscatal.1c03520.
(62) Liu, X. H.; Song, H. Y.; Ma, X. H.; Lear, M. J.; Chen, Y. Z. Virtual screening prediction of new potential organocatalysts for direct aldol reactions. *Journal of Molecular Catalysis A: Chemical* **2010**, *319* (1), 114-118. DOI: 10.1016/j.molcata.2009.12.008.
(63) Henle, J. J.; Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Wang, Y.; Denmark, S. E. Development of a Computer-Guided Workflow for Catalyst Optimization. Descriptor Validation, Subset Selection, and Training Set Analysis. *Journal of the American Chemical Society* **2020**, *142* (26), 11578-11592. DOI: 10.1021/jacs.0c04715.
(64) Rinehart, N. I.; Zahrt, A. F.; Denmark, S. E. Leveraging Machine Learning for Enantioselective Catalysis: From Dream to Reality. *Chimia (Aarau)* **2021**, *75* (7), 592-597. DOI: 10.2533/chimia.2021.592.
(65) Rinehart, N. I.; Zahrt, A. F.; Henle, J. J.; Denmark, S. E. Dreams, False Starts, Dead Ends, and Redemption: A Chronicle of the Evolution of a Chemoinformatic Workflow for the Optimization of Enantioselective Catalysts. *Accounts of Chemical Research* **2021**, *54* (9), 2041-2054. DOI: 10.1021/acs.accounts.0c00826.
(66) Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Henle, J. J.; Denmark, S. E. Computational methods for training set selection and error assessment applied to catalyst design: guidelines for deciding which reactions to run first and which to run next. *Reaction Chemistry & Engineering* **2021**, *6* (4), 694-708, DOI: 10.1039/D1RE00013F.
(67) Lipkowitz, K. B.; 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. *The Journal of Organic Chemistry* **2003**, *68* (12), 4648-4656. DOI: 10.1021/jo0267697.
(68) Dalmau, D. A. R., J. V. ROBERT: Bridging the Gap between Machine Learning and Chemistry. *ChemRxiv 2023. This content is a preprint and has not been peer-reviewed.* **2023**.
(69) Peng, Q.; Duarte, F.; Paton, R. S. Computing organic stereoselectivity - from concepts to quantitative calculations and predictions. *Chem Soc Rev* **2016**, *45* (22), 6093-6107. DOI: 10.1039/c6cs00573j.
(70) Pottel, J.; Moitessier, N. Efficient Transition State Modeling Using Molecular Mechanics Force Fields for the Everyday Chemist. In *Reviews in Computational Chemistry*, Reviews in Computational Chemistry, 2016; pp 152-185.
(71) Maloney, M. P.; Stenfors, B. A.; Helquist, P.; Norrby, P.-O.; Wiest, O. Interplay of Computation and Experiment in Enantioselective Catalysis: Rationalization, Prediction, and—Correction? *ACS Catalysis* **2023**, *13* (21), 14285-14299. DOI: 10.1021/acscatal.3c03921.

(72) Bolitho, Elizabeth M.; Coverdale, J. P. C.; Wolny, J. A.; Schünemann, V.; Sadler, P. J. Density functional theory investigation of Ru(ii) and Os(ii) asymmetric transfer hydrogenation catalysts. *Faraday Discussions* **2022**, *234* (0), 264-283, 10.1039/D1FD00075F. DOI: 10.1039/D1FD00075F.
(73) Weill, N.; Corbeil, C. R.; De Schutter, J. W.; Moitessier, N. Toward a computational tool predicting the stereochemical outcome of asymmetric reactions: Development of the molecular mechanics-based program ACE and application to asymmetric epoxidation reactions. *Journal of Computational Chemistry* **2011**, *32* (13), 2878-2889, https://doi.org/10.1002/jcc.21869.
(74) Verdolino, V.; Forbes, A.; Helquist, P.; Norrby, P.-O.; Wiest, O. On the mechanism of the rhodium catalyzed acrylamide hydrogenation. *Journal of Molecular Catalysis A: Chemical* **2010**, *324* (1), 9-14. DOI: https://doi.org/10.1016/j.molcata.2010.02.026.
(75) Ingman, V. M.; Schaefer, A. J.; Andreola, L. R.; Wheeler, S. E. QChASM: Quantum chemistry automation and structure manipulation. *WIREs Computational Molecular Science* **2021**, *11* (4), e1510. DOI: https://doi.org/10.1002/wcms.1510.
(76) Guan, Y.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E. AARON: An Automated Reaction Optimizer for New Catalysts. *Journal of Chemical Theory and Computation* **2018**, *14* (10), 5249-5261. DOI: 10.1021/acs.jctc.8b00578.
(77) Steiner, M. R., Markus. Navigating chemical reaction space with a steering wheel. *arXiv. This article is a preprint and has not been peer reviewed* **2023**. DOI: https://doi.org/10.48550/arXiv.2308.16499.
(78) *Gaussian 16 Rev. C.01*; Wallingford, CT, 2016.
(79) Smith, D. G. A.; Burns, L. A.; Simmonett, A. C.; Parrish, R. M.; Schieber, M. C.; Galvelis, R.; Kraus, P.; Kruse, H.; Remigio, R. D.; Alenaizan, A.; et al. Psi4 1.4: Open-source software for high-throughput quantum chemistry. *The Journal of Chemical Physics* **2020**, *152* (18). DOI: 10.1063/5.0006002.
(80) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA quantum chemistry program package. *The Journal of Chemical Physics* **2020**, *152* (22), 224108. DOI: 10.1063/5.0004608.
(81) Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. (2004) UCSF Chimera—A Visualization System for Exploratory Research and Analysis. Journal of Computational Chemistry, 25, 1605-1612.
https://doi.org/10.1002/jcc.20084.
(82) 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. *Journal of the American Chemical Society* **2003**, *125* (22), 6614-6615. DOI: 10.1021/ja0293195.
(83) Hansen, E.; Rosales, A. R.; Tutkowski, B.; Norrby, P.-O.; Wiest, O. Prediction of Stereochemistry using Q2MM. *Accounts of Chemical Research* **2016**, *49* (5), 996-1005. DOI: 10.1021/acs.accounts.6b00037.
(84) Terrett, J. A.; Chen, H.; Shore, D. G.; Villemure, E.; Larouche-Gauthier, R.; Déry, M.; Beaumier, F.; Constantineau-Forget, L.; Grand-Maître, C.; Lépissier, L.; et al. Tetrahydrofuran-Based Transient Receptor Potential Ankyrin 1 (TRPA1) Antagonists: Ligand-Based Discovery, Activity in a Rodent Asthma Model, and Mechanism-of-Action via Cryogenic Electron Microscopy. *Journal of Medicinal Chemistry* **2021**, *64* (7), 3843-3869. DOI: 10.1021/acs.jmedchem.0c02023.
(85) Zahrt, A. F.; Denmark, S. E. Evaluating continuous chirality measure as a 3D descriptor in chemoinformatics applied to asymmetric catalysis. *Tetrahedron* **2019**, *75* (13), 1841-1851. DOI: https://doi.org/10.1016/j.tet.2019.02.007.
(86) Gerosa, G. G.; Spanevello, R. A.; Suárez, A. G.; Sarotti, A. M. Joint Experimental, in Silico, and NMR Studies toward the Rational Design of Iminium-Based Organocatalyst Derived from Renewable Sources. *J Org Chem* **2015**, *80* (15), 7626-7634. DOI: 10.1021/acs.joc.5b01214.
(87) Kuang, Y.; Lai, J.; Reid, J. P. Transferrable selectivity profiles enable prediction in synergistic catalyst space. *Chemical Science* **2023**, *14* (7), 1885-1895, 10.1039/D2SC05974F. DOI: 10.1039/D2SC05974F.
(88) Isbrandt, E. S.; Sullivan, R. J.; Newman, S. G. High Throughput Strategies for the Discovery and Optimization of Catalytic Reactions. *Angewandte Chemie International Edition* **2019**, *58* (22), 7180-7191. DOI: https://doi.org/10.1002/anie.201812534.
(89) Kirkpatrick, P.; Ellis, C. Chemical space. *Nature* **2004**, *432* (7019), 823-823. DOI: 10.1038/432823a.
(90) Lu, C.; Wu, C.; Ghoreishi, D.; Chen, W.; Wang, L.; Damm, W.; Ross, G. A.; Dahlgren, M. K.; Russell, E.; Von Bargen, C. D.; et al. OPLS4: Improving Force Field Accuracy on Challenging Regimes of Chemical Space. *Journal of Chemical Theory and Computation* **2021**, *17* (7), 4291-4300. DOI: 10.1021/acs.jctc.1c00302.

(91) Qiu, Y.; Smith, D. G. A.; Boothroyd, S.; Jang, H.; Hahn, D. F.; Wagner, J.; Bannan, C. C.; Gokey, T.; Lim, V. T.; Stern, C. D.; et al. Development and Benchmarking of Open Force Field v1.0.0—the Parsley Small-Molecule Force Field. *Journal of Chemical Theory and Computation* **2021**, *17* (10), 6262-6280. DOI: 10.1021/acs.jctc.1c00571.
(92) Wei, W.; Champion, C.; Barigye, S. J.; Liu, Z.; Labute, P.; Moitessier, N. Use of Extended-Hückel Descriptors for Rapid and Accurate Predictions of Conjugated Torsional Energy Barriers. *Journal of Chemical Information and Modeling* **2020**, *60* (7), 3534-3545. DOI: 10.1021/acs.jcim.0c00440.
(93) Neese, F. The SHARK integral generation and digestion system. *Journal of Computational Chemistry* **2023**, *44* (3), 381-396. DOI: https://doi.org/10.1002/jcc.26942.
(94) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *Journal of Molecular Modeling* **2007**, *13* (12), 1173-1213. DOI: 10.1007/s00894-007-0233-4.
(95) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation* **2019**, *15* (3), 1652-1671. DOI: 10.1021/acs.jctc.8b01176.
(96) 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. *Journal of Computational Chemistry* **2017**, *38* (21), 1825-1833. DOI: https://doi.org/10.1002/jcc.24791.
(97) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society* **1988**, *110* (18), 5959-5967. DOI: 10.1021/ja00226a005.
(98) Lipkowitz, K. B.; Kozlowski, M. C. Understanding Stereoinduction in Catalysis via Computer: New Tools for Asymmetric Synthesis. *Synlett* **2003**, *10*, 1547–1565.
(99) 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. *Chemical Communications* **2018**, *54* (60), 8294-8311, 10.1039/C8CC03695K. DOI: 10.1039/C8CC03695K.
(100) Rosales, A. R.; Ross, S. P.; Helquist, P.; Norrby, P.-O.; Sigman, M. S.; Wiest, O. Transition State Force Field for the Asymmetric Redox-Relay Heck Reaction. *Journal of the American Chemical Society* **2020**, *142* (21), 9700-9707. DOI: 10.1021/jacs.0c01979.