EPFL

Thèse n° 10 610

Enriching the Computational Toolbox for Organocatalysis

Présentée le 12 octobre 2023

Faculté des sciences de base
Laboratoire de design moléculaire computationnel
Programme doctoral en chimie et génie chimique

pour l'obtention du grade de Docteur ès Sciences

par

Simone GALLARATI

Acceptée sur proposition du jury

Prof. V. Hatzimanikatis, président du jury
Prof. A.-C. Corminboeuf, directrice de thèse
Prof. K. Jorner, rapporteur
Prof. A. Doyle, rapporteuse
Prof. J. Waser, rapporteur

École polytechnique fédérale de Lausanne

2023

We can't just stop. We're not rocks.
Progress, migration, motion is… modernity.
It's animate, it's what living things do.

Tony Kushner, Angels in America (1992)

Acknowledgements

I would like to express the deepest gratitude to my advisor, Prof. Clémence Corminboeuf, for having given me the opportunity of becoming a member of LCMD and studying in EPFL under her supervision. Her passion for research, commitment to rigor and excellence, eagerness, directness, and availability have been profoundly inspirational. Thanks to her guidance and constant feedback, I feel I have deeply grown as a scientist. I am extremely grateful for the many opportunities she has given me, including attending and presenting at international conferences and meeting and discussing with other members of the scientific community.

The past four years in Switzerland have been a privilege. Among the many people that have contributed to making this time so thriving, I would like to thank past and present LCMD members, starting with Dr Raimon Fabregat, Dr Veronika Juraskova, and Ksenia Briling, whom I had the pleasure of collaborating with at the beginning of my PhD and experience many adventures together, memories I will always cherish. Dr Ruben Laplaza, Puck van Gerwen, and Dr Matthew Wodrich have also been great collaborators and friends, and have made my time in and out of the lab so special. A special acknowledgement to Alexandre Schöpfer for all the great weekends spent skiing and hiking together. The LCMD family has grown and changed over the years, and I would like to thank Terence (Terry) Blaskovits, Dr Shubhajit Das, Dr Boodsarin Sawatlon, Dr Alberto Fabrizio, Dr Benjamin Meyer, Dr Kun-Han Lin, Dr Sergi Vela, Dr Maria Fumanal, Dr Marc Garner, Dr Frédéric Célerse, Yannick Calvino Alonso, Juliette Schleicher, Yuri Cho, Matthieu Haeberle, Osvaldo Hernandez Cuellar, Liam Marsh, Sara Bassetta, and Dr Jan Weinreich for our time together. I would also like to thank Lukas Lätsch and Dr Jordan de

Jesus Silva from the Copéret group with whom we organized the EPFL/ETHZ summer school on “Big Data and Machine Learning for Chemistry”. Furthermore, my PhD would have been far less smooth had I not been in the hands of Véronique Bujard and received technical support from Dr Daniel Jana.

I would also like to express the deepest love and thankfulness to my family, who has always supported and encouraged me to purse my passions, wherever they might take me. You have given me everything and I will strive to continue making you proud.

I am also grateful to Prof. Vassily Hatzimanikatis, Prof. Jérôme Waser, Prof. Abigail Doyle, and Prof. Kjell Jorner for having accepted to review my thesis and being on my evaluating committee. Finally, I would like to acknowledge the EPFL, the European Research Council (ERC, Grant Agreement No. 817977) within the framework of the European Union’s H2020, and the National Center of Competence in Research (NCCR) “Sustainable chemical process through catalysis (Catalysis)” of the Swiss National Science Foundation (SNSF, grant number 180544) for financial support.

Abstract

Organocatalysis has evolved significantly over the last decades, becoming a pillar of synthetic chemistry, but traditional theoretical approaches based on quantum mechanical computations to investigate reaction mechanisms and provide rationalizations of catalyst performance have failed to keep pace with experiment. This thesis focuses on developing tailored yet transferable data-driven tools and concepts to accelerate organocatalyst discovery, going beyond state-of-the-art computational methods, by addressing three aspects: (1) reaction optimization using closed-loop workflows and strategies based on molecular building blocks for generating candidate species from fragments, (2) establishing cost-effective ways of evaluating how close a prospective catalyst is to achieving optimal performance (*i.e.*, fitness functions), and (3) facilitating and improving the prediction of enantioselectivity and generality through accurate machine learning algorithms and efficient inverse design pipelines.

The first aspect examines the under-exploited modularity of organocatalysts to enable bottom-up database construction, accelerated activity-based screening, and inverse catalyst design. By defining structural components that encapsulate a catalyst's functionalities, we were able to curate a database of thousands of structures mined from the literature or generated combinatorially. These building blocks may be assembled on-the-fly to suggest prospective species with improved performance.

The second aspect focuses on harnessing the structure–activity relationship offered by molecular volcanos as a way to establish a catalyst's "fitness" in closed-loop optimizations. To this end, we developed a genetic algorithm package, NaviCatGA, and showed that it is an efficient tool to

streamline computer-aided catalyst discovery. Multi-objective problems *e.g.*, activity–selectivity tradeoffs, may also be solved with evolutionary experiments by considering, and scalarizing, more than one target simultaneously.

In the final section, we address current limitations of machine learning and generative models in predicting and optimizing challenging targets, specifically enantioselectivity and catalyst generality. We design reaction-inspired representations to improve the accuracy of physics-based models and show how evolutionary experiments may be planned to find catalysts displaying high performance across a broad substrate scope.

Overall, this thesis demonstrates how tailored data-driven tools and concepts that are able to address the unique properties and structures of organocatalysts streamline reaction optimization and the discovery of prospective new species.

Keywords: organocatalysis, inverse design, closed-loop optimization, molecular volcano plots, machine learning, structure–activity relationships, enantioselectivity

Riassunto

Nel corso degli ultimi decenni, l'organocatalisi si è evoluta in modo significativo, diventando un pilastro della sintesi organica, ma approcci teoretici tradizionali fondati su calcoli quantomeccanici per studiare meccanismi di reazione e razionalizzare la prestazione dei catalizzatori non sono stati in grado di rimanere al passo con approcci sperimentali. Questa tesi si concentra sullo sviluppo di strumenti e concetti specifici, ma allo stesso tempo trasferibili, per accelerare la scoperta di organocatalizzatori basandosi sull'analisi di grandi quantità di dati, superando altri metodi computazionali all'avanguardia, considerando tre aspetti: (1) ottimizzazione di reazioni usando flussi di lavoro "a circuito chiuso" e strategie basate su blocchi molecolari per assemblare possibili candidati usando frammenti; (2) stabilire modi economici per valutare quanto un possibile catalizzatore sia vicino al raggiungimento di prestazioni ottimali (*i.e.*, funzioni di idoneità); (3) facilitare e migliorare la previsione di enantioselettività e generalità tramite algoritmi di apprendimento automatico e canali di "progettazione inversa" efficienti.

Il primo aspetto esamina la poco sfruttata modularità degli organocatalizzatori per permettere la costruzione di database "dal basso verso l'alto", accelerare il loro lo screening in termini di attività catalitica, e la loro "progettazione inversa". Definendo componenti strutturali che incapsulano le funzionalità di un catalizzatore, siamo stati in grado di curare un database di migliaia di strutture estratte da pubblicazioni o generate in modo combinatorio. Questi elementi strutturali possono essere assemblati al volo per suggerire nuove specie con miglior performance.

Il secondo aspetto si concentra sullo sfruttamento delle relazioni fra struttura e attività offerte da grafici a vulcano molecolare in qualità di tramite per stabilire l'idoneità di un catalizzatore in

ottimizzazioni a “circuito chiuso”. A tal fine, abbiamo sviluppato un algoritmo genetico, NaviCatGA, e dimostrato quanto sia efficiente per accelerare la scoperta computazionale di catalizzatori. Problemi con più di un obbiettivo *e.g.*, compromessi fra attività e selettività, possono essere risolti con esperimenti evolutivi considerando, e scalarizzando, più di un target contemporaneamente.

Nella sezione finale, abbiamo considerato i limiti attuali di modelli generativi e di apprendimento automatico nel predire e ottimizzare obiettivi difficili, specificatamente l’enantioselettività e la generalità di un catalizzatore. Abbiamo progettato rappresentazioni ispirate a reazioni chimiche per migliorare la precisione di modelli basati sulle leggi della fisica e mostrato come esperimenti evolutivi possano essere pianificati per trovare catalizzatori con prestazione elevata su un’ampia gamma di substrati.

Nel complesso, questa tesi dimostra come strumenti e concetti su misura basati su grandi quantità di dati in grado di affrontare le proprietà e le strutture uniche degli organocatalizzatori semplifichino l'ottimizzazione di reazione chimiche e la scoperta di potenziali nuove specie.

Parole chiave: organocatalisi, progettazione inversa, ottimizzazione a circuito chiuso, grafici a vulcano molecolare, apprendimento automatico, relazioni struttura–attività, enantioselettività

Contents

1 Introduction

Over the last 20 years organocatalysis *i.e.*, the use of small and medium-sized organic molecules to catalyze chemical reactions, has matured from a few mechanistically ill-defined niche transformations to one of the most thriving research domains in chemistry.[1] Along with transition-metal and enzyme catalysis, it is an established pillar of organic synthesis and, accordingly, it was elected by IUPAC as one of the ten emerging technologies with the potential to make our planet more sustainable.[2] Some of the features that have attracted the community's interest are the catalysts' availability from renewable sources as single enantiomers, their robustness, non-toxicity, and the operational simplicity of organocatalytic reactions, which are often air and/or water tolerant.[3] In a recent perspective,[4] the field was deemed ripe to become broadly applicable, even in industrial settings, despite current concerns regarding high catalyst loading and recyclability.[5] Overcoming some of the existing limitations requires continuous advances in catalyst design and the development of new activation strategies.[6,7] So far, this has primarily been achieved *via* experimental screening and rational or empirical design. High-level quantum chemical (QC) methods have provided valuable insight into reaction mechanisms and occasionally helped perform a fine tuning of catalyst structure.[8,9] However, owing to the substantial computational cost and inherent complexity of catalytic cycles,[10] the ability of traditional approaches based on Density Functional Theory computations of potential energy surfaces to unravel mechanistic details and offer quantitative reactivity predictions is failing to keep pace with experiment.[11] Alternative, well-established strategies[12–14] aimed at obtaining correlations between empirical data and DFT/experimentally-based molecular descriptors provide rapid insights into the structural features are necessary for good performance, but can be

challenging if descriptors simulating the interactions at play are not found or if the empirical observations are the result of more than one fundamental process (*e.g.*, nonlinearity due to a change in mechanism).

Since both DFT computations and multivariate linear regression analysis are often limited to case-by-case studies and dominated by trial-and-error, there is a need to develop conceptual and data-driven tools that facilitate both the exploration of a wider range of organocatalyst space and the automated optimization of reaction properties. Applications of machine learning (ML) in chemistry have demonstrated its potential to propel the next leap forward in the discovery of functional molecules and materials,[15] principally because the number of prospective species that can be examined greatly exceeds that amenable to more traditional theoretical or experimental approaches.[16,17] Going beyond direct screening, coupling supervised ML algorithms that estimate catalytic properties with generative models[18] that bias chemical space exploration towards areas of interest enables the inverse design of entirely new species.[19] It is therefore reasonable to envision that, in today's world, any strategy aiming to design new catalysts will inevitably involve some form of ML. In particular, data mining and analysis techniques, together with statistical and generative models, are ideally suited to complement quantum chemical and mechanistic-guided methods, which are often faced with numerous challenges,[20] and transform the nature, scale, and complexity of the problems tackled.[21]

One aspect that makes organocatalytic reactions particularly difficult to study by means of traditional QC methods[8] is the catalysts' flexibility and the existence of many low-energy, thermally accessible conformations that may potentially be reactive, which must be exhaustively searched for and evaluated for accurate stereoselectivity predictions.[22] Furthermore, the catalyst often engages the substrate in multiple, subtle non-covalent interactions (NCIs) that ultimately determine the reaction outcome but whose extent is difficult to quantify.[23] The free energy difference between two conformational states or the strength of NCIs like hydrogen-bonding, π-stacking, or CH/π interactions may account to only a few kcal/mol,[24,25] which poses a significant challenge to standard DFT. Additionally, to expand the scope of possible transformations, the

structure of organocatalysts is becoming increasingly complex and multifunctional yet, unlike organometallic compounds, well-defined building blocks[26] are harder to enumerate for functionally diverse species across wide regions of chemical space, making the implementation of otherwise successful fragment-based strategies[27] less routine. Given these caveats, individual computations on a single catalytic system or existing tools are not necessarily transferable to other (even related) systems.

The overarching theme of this thesis involves addressing these limitations by developing tailored yet transferable data-driven tools and concepts to accelerate organocatalyst discovery, going beyond state-of-the-art methods.[28] The material is organized following three aspects: (1) reaction optimization using closed-loop workflows and strategies based on molecular building blocks for generating candidate species from fragments, (2) establishing cost-effective ways of evaluating how close a prospective catalyst is to achieving optimal performance (*i.e.*, fitness functions), and (3) facilitating and improving the prediction and optimization of challenging targets (*e.g.*, enantioselectivity, generality).

Chapter 2 provides the reader with a brief overview of the approaches used to study and optimize organocatalytic reactions, divided between mechanistic-guided or -agnostic.

Many of these tools have focused on specific reaction classes or structurally related catalysts. Consequently, there is a dearth of general strategies and platforms for organocatalysts comparison, fragmentation into building blocks, and assembly across different regions of catalyst space, encompassing functionally and chemically diverse species. To address this, **Chapter 3** introduces OSCAR (Organic Structures for CAtalysis Repository), a database of 4000 experimentally derived organocatalysts along with their corresponding building blocks, enriched with combinatorially generated structures (up to 1.5 million). The fragment-based approach used for dataset curation is outlined and the repository's diversity, in terms of functions and molecular properties, is showcased with chemical space maps, which help establish structure–reactivity relationships for reaction optimization. This article has been published in *Chemical Science*.[29]

Following the creation of OSCAR, we exploit the same modular strategy and the type of fragments contained in the repository to improve the activity of bifunctional hydrogen-bond donor/amines in combination with statistical modelling[14] and molecular volcano plots. The latter are data-driven tools useful for rationalizing trends in catalytic behavior and predicting the performance of untested candidates.[30] Originally developed for applications in electro- and heterogeneous catalysis, this concept was successfully transferred to organometallic reactions by our group.[31] In **Chapter 4**, we demonstrate how the automated construction of volcano plots and activity maps[32] may be integrated into a bottom-up protocol that leverages the organocatalyst's modularity and guides the choice of ideal building blocks for rate enhancement. This article has been published in *Organic Chemistry Frontiers*.[33]

While the study above involves direct activity-based screening, molecular volcanos allow for an inexpensive mapping between structure and reactivity and thus constitute an ideal way of estimating how close a prospective species is to achieving maximum performance. **Chapter 5** describes how volcano plots are incorporated into a pipeline for inverse design,[34,35] which relies on our genetic algorithm[36] NaviCatGA and the previously curated molecular fragments libraries from OSCAR. As a validating case study, we perform multi-objective optimization of bipyridine *N,N*′-dioxides Lewis bases in the propargylation reaction of benzaldehyde.[37] This article has been published in *Chemistry—Methods*.[38]

Evaluating the fitness function during genetic optimization with NaviCatGA is accelerated *via* machine learning (ML) predictions. Among the many different "flavors" of ML, physics-based models[39] accurately predict molecular and atomic properties while offering a great deal of generality and transferability, but can struggle with more challenging reaction-based targets, such as enantioselectivity.[40] In **Chapter 6**, we outline a strategy for improving molecular representations within such atomistic model and accurately predict the enantiomeric excess of bipyridine *N,N*′-dioxides in the aforementioned propargylation. This article has been published in *Chemical Science*.[41]

In **Chapter 7**, we show how the NaviCatGA pipeline may be adapted to optimize catalyst generality[42] *i.e.*, exhibiting both high turnover and enantioselectivity across a broad substrate scope, as primary target. The workflow combines data mining to curate an experimental database of 820 Pictet–Spengler reactions, used to train statistical models for *e.e.* and TOF predictions (in combination with molecular volcanos), structure manipulation to define a combinatorial space of building blocks from OSCAR, and evolutionary experiments performed across a virtual catalyst–substrate landscape of millions of possibilities.

Finally, **Chapter 8** concludes the thesis by summarizing the main findings of the tools and concepts developed herein with respect to the three overarching objectives. Future work is suggested regarding further developments, serving as stimuli to accelerate the discovery of organocatalysts for new transformations.

2 Computational Tools to Study and Optimize Organocatalytic Reactions

2.1 Historical Overview of Computational Organocatalysis

Organocatalysis' contribution to modern society has recently been recognized on account of the Royal Swedish Academy of Sciences awarding the 2021 Nobel Prize in Chemistry to Benjamin List and David W.C. MacMillan "for the development of asymmetric organocatalysis". Since its inception, tremendous advancement in reactivity, activation modes, and stereoinduction has been attained.[1,43,44] In 2000, two seminal, independent publications by List, Barbas III, and Lerner on the proline-catalyzed intermolecular aldol reaction[45] and by MacMillan on iminium catalysis[46] set the stage for a new pillar of asymmetric synthesis. Despite several contributions on the use of small organic molecules as catalysts having appeared long before,[6] these papers described general activation strategies that could be extended to a broad range of reaction classes, conceptualizing the field of "organocatalysis" (a term introduced by MacMillan) and highlighting its potential environmental, economic, and scientific advantages.[3]

The first organocatalytic reaction to be studied computationally was the Hajos–Parrish reaction by Cheong *et al*. in 2004.[47] This work showed that theoretical tools could be used to investigate and even predict the performance of organocatalysts[48] however, since then, quantum mechanical (QM) methods have mostly been used to rationalize experimental observations *ex post facto* rather than to make true predictions. Few successful early examples of computationally-led organocatalyst optimization exist, including predictions on *anti*-selective Mannich-type reactions

by Houk, Tanaka, and Barbas III (2006),[49] the development of (thio)urea analogues for epoxide ring-opening (2007),[50] and the design of a simplified yet highly enantioselective primary amine by Paton and Dixon (2015).[51] These studies involve modifying molecules that show some catalytic activity using knowledge gained from experiments or computational models to attain second-generation catalysts with improved performance. A complementary approach consists in screening a virtual catalyst library curated manually with well-known compounds or generated combinatorially from molecular fragments. This has been made possible owing to the past decade's advances in computing power and the development of software packages to automate routine computational tasks, such as the optimization of the hundreds of transition states (TSs) required to accurately predict the stereochemical outcome of a reaction.[52]

In 2015, Neel and Toste investigated a phase transfer chiral anion catalysis system[53] using a new set of data science tools being developed by Milo and Sigman, including computational featurization of reaction components, linear regression modelling, statistical classifications, and data set design.[54] They showed that enantioselectivity data from organocatalytic reactions where selectivity arises from differential non-covalent interactions (NCIs) can be quantitatively connected to the attributes (molecular descriptors) of the reaction components.[55–60] Since then, multivariate linear regression (MLR) analysis of activity/selectivity measures has gained enormous popularity for the optimization of (organo)catalytic systems.[14] This approach differs from those described above in that it is mechanistically agnostic at the outset of the investigation, while heavily relying on an initial set of experimental data. Mechanistic hypotheses may be formulated *a posteriori* on the basis of which features are highly correlated with reaction performance. On the other hand, Denmark and co-workers have demonstrated a purely data-driven method[61] whereby a library of organocatalyst candidates is evaluated in order to optimize a reaction without simultaneously exploring its mechanism.[62] Their pioneering 2019 study showed how support vector machine and deep feed-forward neural network models trained using electronic descriptors combined with a newly designed steric descriptor of conformer ensembles, the Average Steric Occupancy (ASO), are able to predict the selectivity of higher-performing

chiral phosphoric acids (CPAs).[63] This work provided a large dataset (1075 reactions) that has become popular for subsequent machine learning (ML) studies[64–68] and fueled the application of artificial intelligence techniques in organocatalysis.

2.2 Mechanistic-Guided Approaches

Broadly speaking, the tools that have been used in computational organocatalysis can be divided into two categories, depending on the researchers' primary objective: (1) tools for mechanistic interpretation and catalyst fine-tuning, and (2) tools for performance prediction and reaction optimization.

The mechanistic-guided approach mostly relies on quantum chemical modelling techniques and the calculation of potential energy profiles along the reaction pathways, with Density Functional Theory having become the *de facto* standard,[8,69–71] although highly accurate wavefunction methods *e.g.*, DLPNO-CCSD(T), are occasionally used for single point energy computations.[72–74] When reactivity is believed to be highly influenced by solvation, molecular dynamics simulations with explicit solvent molecules may be performed.[75] This approach is mostly aimed at helping interpret experimental observations by verifying if a proposed mechanism is plausible *i.e.*, energetically viable. Information gained from these results is sometimes used to suggest modifications to the catalytic system to improve its performance,[76,77] however it is still more efficient to experimentally screen a range of potential organocatalysts than to "test" them *in silico*. That is because QM-based reactivity predictions often require computing the complete potential energy profile associated with a catalytic cycle, taking into account the complex conformational space associated with large and flexible molecules, which becomes extremely expensive for more than a handful of systems.[20]

When the nature of the selectivity-determining step is well-known, catalyst optimization may be carried out using automated toolkits. They help generate the structure of new catalyst-substrate combinations and optimize hundreds of stereodetermining TS geometries. Among the fully QM-based tools, AARON[78] and QChASM[79,80] have been successfully applied to organocatalysis.[37,52]

Despite significant advances in the speed of QM and hybrid QM/MM[81] methods, DFT-based predictions may still be intractable for high-throughput screenings or large and flexible systems. Norrby and co-workers have therefore developed alternative approaches using transition state force fields (TSFF)[82] and QM-derived molecular mechanics force fields (Q2MM),[83] which provide accuracy rivaling DFT but at a drastically reduced computational cost. Their tools ACE[84] and CatVS[85] are now integrated into the platform VIRTUAL CHEMIST.[86] A similar technique whereby a rigid TS model of the substrate is docked into a flexible organocatalyst, modelling the conformational space of the non-covalently bound complex with FFs, is reverse docking.[87] Applications of this approach to peptide catalysis,[87] TADDOL-catalyzed asymmetric hetero-Diels–Alder,[88,89] and Strecker hydrocyanation[90] have been reported.

A challenge inherently faced in the mechanistic-guided approach is the conformational flexibility of organocatalysts. Accurate stereoselectivity predictions require evaluating all the thermally accessible catalyst-substrate conformations.[91] A number of programs (*e.g.*, Crest,[92,93] RDKit,[94] Balloon,[95] wSterimol[96], and Molassembler[97]) have been developed to perform conformational sampling together with QM methods;[98] however, static DFT computations of isolated minima are sometimes insufficient to describe chemical events occurring in experimental settings,[99] especially when the entropic contribution to the stability of different molecular states must be taken into account. In this context, our group has proposed enhanced sampling techniques, in particular replica exchange molecular dynamics (REMD),[100] to address organic chemistry problems connected to fluxional molecules, including organocatalysts.[101,102] While these studies have demonstrated the importance of thoroughly mapping the conformational landscape of flexible catalysts, structure–activity relationships were only indirectly inferred, as the substrates were excluded from the simulations.

2.3 Regression Methods for Reaction Optimization

Tools for the second approach are aimed at obtaining quantitative predictions of catalytic performance, typically enantioselectivity or reaction rates, even in the absence of mechanistic hypotheses.[103] Sometimes, computed reaction profiles and stereoelectronic parameters, in combination with linear regression, are used to generate predictive models.[104] Examples include the work of Goodman on CPA selection in nucleophilic additions to imines based on steric assessments[105–107] and models developed by Wei and Lan for predicting chemoselectivity in N-heterocyclic carbene[108] and Lewis base-catalyzed[109] reactions from the global nucleophilicity and electrophilicity indices of the species involved in the product-determining step. While those descriptors were obtained from DFT computations,[110] Mayr and co-workers have conducted extensive research on establishing experimental nucleophilicity and electrophilicity scales, including in nucleophilic organocatalysis,[111–114] and reaction rates can be calculated by the Mayr–Patz Linear Free Energy Relationship (LFER).[115]

Traditional univariate LFERs,[116] such as the Hammett equation,[117] are often used to gain mechanistic information, but their simplicity can limit the obtainable insight in more complex scenarios.[118] Multiparameter approaches have therefore emerged to correlate chemical reactivity to structure using physically interpretable quantities and find better-performing catalysts.[119,120] "Holistic" MLR models to transfer chemical observations from one reaction to another and improve selectivity have now been proposed for CPA and hydrogen-bond donating (HBD) catalysis.[121–123] Beyond intuitive physical organic and spectroscopic parameters, enantioselectivity has been modelled *via* molecular interaction filed (MIF), comparative molecular filed analysis (CoMFA),[124–128] and Continuous Chirality Measure (CCM).[64] While MLR is typically used with small to medium-sized datasets,[129] nonlinear methods are leveraged when large databases are available.[118] In such cases, topological (*i.e.*, 2D) descriptors have been found to be cost-efficient alternatives to the more expensive 3D-based representations.[67,68,130,131] If predictive models cannot be developed using the entire chemically diverse dataset available,

"catalyst selection by committee" has been proposed to exploit multiple, data-limited models generated on different substrates and recommend novel organocatalysts.[132]

2.4 Other State-of-the-Art Data-Driven Approaches

Apart from supervised ML methods to regress activity/selectivity data, unsupervised learning techniques have recently found use in organocatalysis. These include dimensionality reduction (*e.g.*, PCA, UMAP, t-SNE) and clustering (*e.g.*, *k*-means) algorithms to visualize *in silico* catalyst libraries and select subsets for experimental screening or training sets to probe the accessible chemical space.[133] Recently, Reid *et al*. used such unsupervised learning methods to assess and quantify the generality (*i.e.*, broadness of the substrate scope) of asymmetric organocatalysts.[134]

Another emerging statistical analysis tool is univariate classification. Pioneering work by Chen and co-workers involved support-vector machine (SVM)–based virtual screening of the PubChem repository to distinguish primary and secondary amine catalysts for the direct intermolecular aldol reaction from drug-like molecules.[135] More recently, Doyle and Sigman have reported a classification algorithm leveraging a single-node decision tree[136] to explain the origin of "activity cliffs" and establish which molecular descriptor could predict whether an organocatalyst will be enantioselective in a given reaction ("selectivity cliffs").[133,137]

Finally, alternatives to QM- or ML-assisted high-throughput screening are emerging, specifically the use of generative models, including generic algorithms (GA), in the spirit of "inverse design".[34] Using a GA, Jensen *et al*. discovered new azetidine catalysts for the alcohol-mediated Morita–Baylis–Hillman reaction by searching the chemical space of the ZINC database.[138] This constitutes the first experimentally verified *de novo* design of an efficient organocatalyst with generative models.

2.5 Outlook

The computational tools described above represented the state-of-the-art at the time this thesis was started. However, both the mechanistic-guided (*i.e.*, the creation of potential energy surfaces) and -agnostic approach (regression of experimental outcomes) are often limited in their transferability to, at best, closely related systems. Clearly, tools that can combine large quantities of data, connect seemingly disparate reactivity profiles, and establish trends that facilitate the rationalization and prediction of organocatalysts' properties are highly desirable. Furthermore, developing fast and accurate ML algorithms that are compatible with generative models, as well as robust structure generation workflows from well-defined building blocks and efficient ways of scoring candidates according to their activity/selectivity, would enable *de novo* organocatalyst discovery. The following chapters describe our efforts in developing such data-driven tools to expand the computational toolbox for organocatalysis.

3 OSCAR: An Extensive Repository of Chemically and Functionally Diverse Organocatalysts

This chapter is based on following publication:
Gallarati S., van Gerwen P., Laplaza R., Vela S., Fabrizio A., and Corminboeuf C., OSCAR: An Extensive Repository of Chemically and Functionally Diverse Organocatalysts. *Chem. Sci.* **2022**, *13*, 13782.

3.1 Introduction

Constructing extensive yet tailored databases is crucial for the successful development and application of data-driven tools in catalysis and materials science.[139,140] The way datasets are generated largely reflects how chemists think about the structure of a catalyst. In turn, this not only influences the way improved molecular systems are searched, but also how their structure is manipulated, for example through trial-and-error,[141] fine-tuning according to mechanistic insight,[49–51,142] or generating compound libraries for activity/selectivity screening.[37,76,77,85]

Transition-metal catalysts are naturally viewed in a modular fashion as a combination of active metal center and ligands, which are further decomposed into metal-coordinating groups, backbone/bridging units, and substituents.[27] This simple, yet powerful fragment-based strategy has enabled tremendous advancements in computer-aided catalyst design,[143,144] from the exploration of the chemical space of inorganic species curated through bottom-up or top-down approaches,[145–150] the construction of ligand databases with associated steric and electronic descriptors,[151–157] to the development of algorithms for the assembly of metal complexes from fragments and evolutionary experiments.[158–160] Modularity is even more apparent in biocatalysts,[161] which combine a limited number of building blocks, the amino acids; inspired by

natural evolution, strategies such as combinatorial backbone assembly[162] have allowed to generate libraries of structurally diverse enzymes with altered catalytic properties.

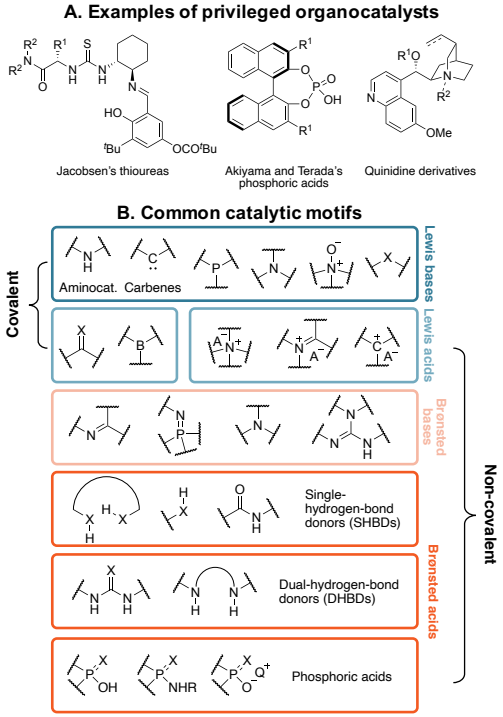


Figure 3.1 (A) Prototypical privileged chiral frameworks for asymmetric catalysis. (B) Classification of organocatalysts according to their catalytic motifs (X = O, S).

Organocatalysts are far less frequently classified according to fragment-based schemes. Instead, they are typically grouped into families of "privileged catalysts",[163,164] or according to the functional components that encapsulate their catalytic power (Figure 3.1).[165] Privileged catalysts are those species possessing certain chiral scaffolds that have proven to be effective at inducing high levels of enantioselectivity across a wide range of mechanistically unrelated reactions.[163,164] Some effort has been made to summarize these catalytic motifs,[166] however their comprehensive enumeration across all of chemical space is challenging due to the large possible variations in functionalities. This problem is exacerbated by the fact that organocatalysts are essentially a subclass of organic molecules, whose space is estimated to exceed 10^{60},[167,168] and chemical expertise is required to evaluate whether an organic molecule could function as a catalyst in a reaction. Therefore, *de novo* organocatalyst design is a formidable, seldomly approached task,

primarily due to the lack of robust ways of defining and assembling their building blocks,[169,170] and reaction optimization is dominated by testing closely related analogues of a known privileged catalyst.[171]

Similarly, the field of data-driven organocatalysis has been dominated by efforts, either on the automation side[107] (*e.g.,* AARON[78] or ACE/Virtual Chemist[84,86]) or on the development of statistical models for enantioselectivity prediction,[41,63,105,121,122,130,172] that have focused on specific reaction classes or structurally related catalysts. There is currently a dearth of general strategies and platforms for organocatalysts comparison, fragmentation into building blocks, and assembly across a wide region of catalyst space, encompassing functionally and chemically diverse molecules with a multitude of catalytic functions.

In this work, we propose a solution in the form of OSCAR (Organic Structures for CAtalysis Repository), a database of experimentally derived or combinatorially enriched organocatalysts and of the corresponding molecular fragments that are extracted from them. Not only OSCAR constitutes a map to navigate organocatalyst space and potentially enable informed catalyst design, but the modular strategy behind its construction paves the way to a multitude of data-driven and fragment-based reaction optimization methods.[33,173] Herein, we show how such a dataset is curated and augmented with crystallographically determined structures using a combination of top-down and bottom-up approaches, and how the fragments are assembled in a combinatorial fashion to generate thousands of species. In its current forms, OSCAR contains 4,000 catalysts, whose use has either been documented in the literature for organic synthesis or with chemically analogous structure reported in the Cambridge Structural Database (CSD), spanning various catalytic functions (Lewis/Brønsted acids and bases), and two exemplary enriched combinatorial supersets, OSCAR!(NHC) and OSCAR!(DHBD). The former consists of over 8,000 carbenes for covalent catalysis, the latter contains *ca*. 1.5 million non-covalent dual-hydrogen-bond donors. The approaches used to generate these combinatorial databases (*vide infra*) are however transferable to other classes, implying the possibility of further extending OSCAR. A selection of stereoelectronic molecular descriptors, including reactivity indices

derived from conceptual DFT, are provided and may help establishing structure–reactivity relationships for reaction optimization. All structures and properties are publicly available on the Materials Cloud for interactive visualization with Chemiscope (https://doi.org/10.24435/materialscloud:gy-3h).[174] They could serve as the starting point to define the combinatorial space for evolutionary experiments,[38] as well as the basis for dataset curation to train machine learning models for applications in organic synthesis.[68]

3.2 Results and Discussion

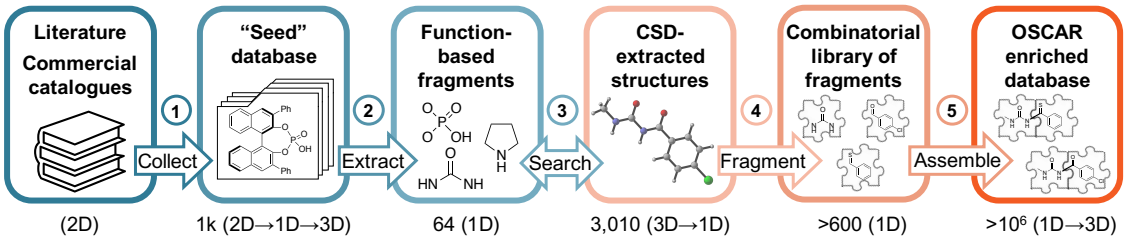


Figure 3.2 Graphical summary of the steps followed for the curation of OSCAR.

3.2.1 Database Curation

No comprehensive repository of organocatalysts' structures covering all of the functionalities summarized in Figure 3.1B currently exists. Most frequently, they are reported in the literature in 2D format (*i.e.*, ChemDraw pictures) with associated experimental characterization data in the Supporting Information (NMR and IR spectra and, less often, crystal structure information), but molecular geometries are not easily accessible. To construct OSCAR, we followed a five-step protocol (Figure 3.2), which starts with the manual collection of catalysts (as 2D objects) from reviews,[43,165,175–185] journal articles,[186–189] books,[190–194] and commercial catalogues[195,196] into a "seed" database (step 1). Each of the 1,000 2D entries in this library is labelled according to the classes in Figure 3.1 and converted into a 1D (*i.e.*, SMILES strings) and subsequently 3D (*i.e.*, optimized XYZ geometry) structure (see the Computational methods). Given that more than ~1,500 publications on organocatalysis are published each year,[1] it is virtually impossible to curate an exhaustive library of all existing catalysts. Nonetheless, the seed database aims at

covering the chemical diversity observed across all of organocatalyst space in terms of chemical functionalities, catalytic motifs and scaffolds/substituents, with the added bonus of each structure either being commercially available or synthetically accessible, having being mined from the literature.

This top-down approach ensures that only organic molecules that have been reported to display, or be tested for, catalytic activity are included in the database. However, it is a slow, human error-prone process that cannot be automated and might either introduce in the repository erroneous or mislabeled structures or lead to chemically interesting ones being excluded. Existing crystallographic databases (*e.g.*, CSD,[197,198] COD[199]) offer the most comprehensive collection of organic (and inorganic) molecules that have been synthetized. Although it not possible to filter out *a priori* those compounds that have not been tested as organocatalysts, CSD offers the chance to significantly augment the seed database with more, chemically diverse structures, provided that the right chemical motifs, which might make a molecule catalytically active, are searched. To achieve this goal we enumerated, in 1D format, 64 "function-based fragments" included in the seed database (step 2 Figure 3.2 and Figures S1–S2). Although not exhaustive, they represent the most common catalytic motifs and ensure that the species that contain them are relevant to the task at hand. In step 3, these fragments are searched in CSD and the corresponding whole molecules extracted. After retrieving the 3D geometries from the cif files with the *cell2mol* software,[200] 3,010 compounds are added to the seed database, yielding a total of 4,000 entries (after filtering out identical ones, see the ESI). All 3D entries are then converted into 1D format for subsequent fragmentation and recombination (steps 4 and 5, *vide infra*).

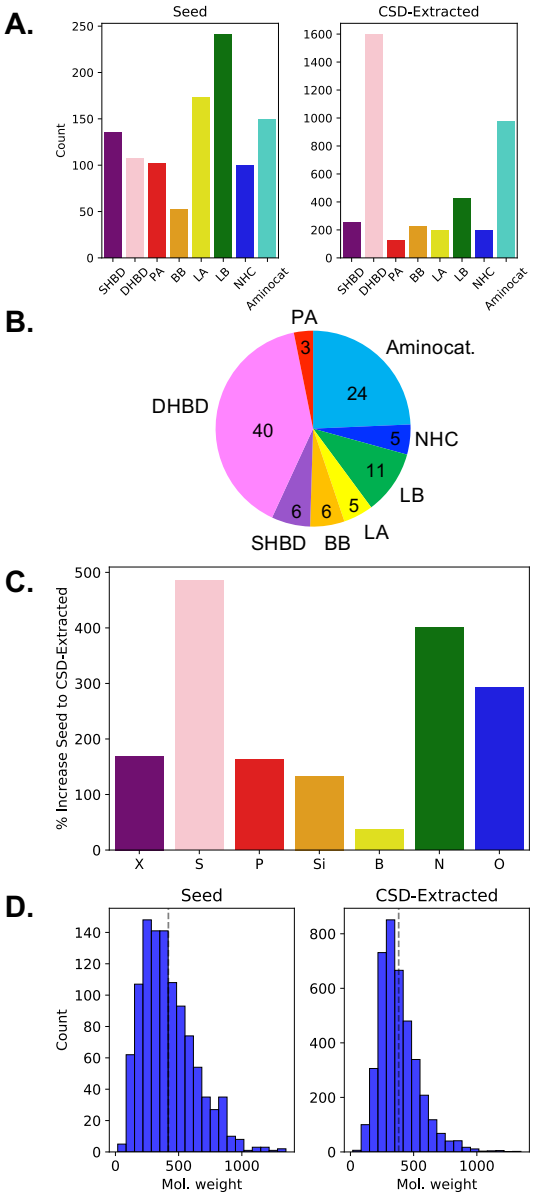


Figure 3.3 (A) Distribution histograms of catalytic motifs in the seed database and in the CSD-extracted structures. (B) Pie chart showing percentages of catalytic motifs in the seed and CSD-extracted datasets. (C) Distribution histograms of heteroatom types (X = halogens), and (D) molecular weight in the seed and in the CSD-extracted sets.

With respect to the catalytic motifs (*cf* Figure 3.1B), the distribution of the CSD-extracted structures changes significantly from the seed database (see the two histograms in Figure 3.3A). In OSCAR, the majority of species (40%) are classified as dual-hydrogen-bond donors; their large increase in number upon CSD extraction is likely due to the popularity of the (thio)urea moiety as pharmacophore[201–203] and for anion recognition.[204] The second most popular class

(24%) is aminocatalysts based on the pyrrolidine motif: in the early days of organocatalysis, the vast majority of reactions were indeed amine-based[176,205] and five-membered (polycyclic) secondary amines are widely encountered in natural products, as well as being a preferred scaffold in pharmaceutical science and drug design.[206] The other classes are more or less equally represented (~5-6%, Figure 3.3B), with a slight predominance of Lewis bases (11%), given the large variety of N(O)-, P(O)-, and S(O)-nucleophilic organocatalysts. If we consider the increase in type of heteroatoms from the seed to the CSD-extracted database (Figure 3.3C), sulfur and nitrogen are the most abundant due to the predominance of the thiourea and pyrrolidine catalytic motifs. The amount of P, Si, X, and especially B atoms increases to a significantly lesser extent. In the case of phosphorous, even though we seek to augment the quantity of P-containing motifs, only a limited number of phosphoric acids (*ca*. 25) are extractable from CSD. On the other hand, no catalytic unit that specifically contains halogens, silicon or boron is searched. An exhaustive description of the functional groups present in OSCAR is given in the Supporting Information (Table S2 and Figure S4). Finally, the catalysts in the two datasets have a similar distribution of molecular weights (Figure 3.3D), with the seed database containing on average slightly larger molecules (~ 430 u) and displaying a smoother decrease in occurrences as their size increases.

3.2.2 Structure and Property Maps

The chemical and structural diversity contained in OSCAR is visualized in Figure 3.4A with a 2D t-SNE map[207] based on FCHL19[208] of the 4,000 organocatalysts from the seed and CSD databases. Alternative representations and dimensionality reduction can be found in the ESI (Figures S5–S7). Although the two axes (dimensions) of this structure map have no formal physical meaning, it is possible to establish a qualitative relationship between them and chemical properties. In particular, species found higher in the map are bigger (higher molecular weight/surface area), whereas the degree of conjugation and the presence of aromatic scaffolds and substituents decreases left to right. For example, diol-based catalysts,[209] which act as single-hydrogen-bond donors, and phosphoric acids[210] are found along the upper edge of the map, with the fully aromatic BINOL derivatives on the left, the H_8-BINOL core in the middle, and

BAMOLs on the right. Dual-HBDs, especially diaryl (thio)ureas, occupy the lower left corner of the map, while simple proline derivatives the bottom right, with larger and more complex aminocatalysts in the upper left region. Other noticeable clusters correspond to the ketone epoxidation catalysts developed by Shi and Shu (covalent Lewis acidic carbohydrate derivatives),[211,212] and to iminophosphorane Brønsted bases.[213]

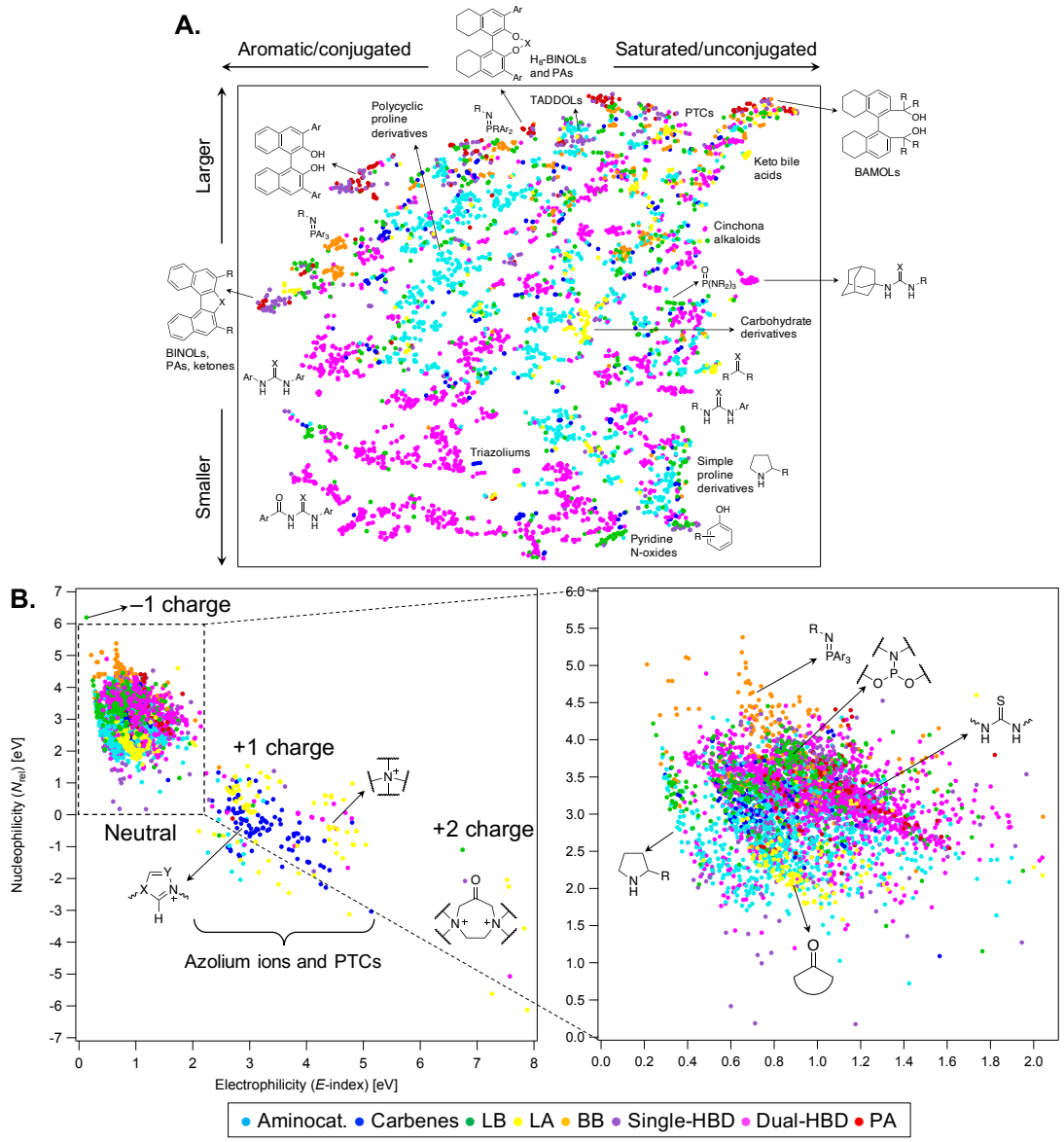


Figure 3.4 2D t-SNE map of OSCAR on the basis of the FCHL19 representation.[208] Each point represents an organocatalyst, colored by the corresponding catalytic motif. Each cluster contains catalysts with similar structure, with some examples being shown R = alkyl group; Ar = aromatic group; PTC = phase-transfer catalyst. (B) Property map: computed (ωB97X-D/Def2-TZVP//B97-D/Def2-TZVP) nucleophilicity (N_{rel}) *vs*. electrophilicity (*E*-index) parameters.[110] A zoom-in of the map is provided on the right hand side.

The structure map is complemented by a "property map" (Figure 3.4B) in which the organocatalysts are evaluated in terms of their DFT-computed global electro/nucleophilicity indices (see the Computational methods), which assume that, when these catalysts react, they do so cumulatively and simultaneously at all their atomic sites.[214] The largest influence on the descriptors is exerted by the total molecular charge, and three regions are found (four if the green point corresponding to the phosphorylated sulfonimidamide[215] with –1 charge is considered). *E*-index increases with the charge, while N_{rel} decreases. Highly electrophilic and charged species include phase transfer catalysts[216] (PTCs, non-covalent Lewis acids) and azolium ions, which are the conjugate acid precursors of carbene organocatalysts.[217] The zoom-in on the right hand side of Figure 3.4B shows the spread of *E*-index and N_{rel} values for neutral organocatalysts. Among the most nucleophilic species, Brønsted bases, in particular iminophosphoranes, and phosphoramidite Lewis bases are found towards the top of the map ($\overline{N_{rel}}$ = 3.8 eV, σ = 0.6 eV), while ketone epoxidation electrophiles are at the bottom ($\overline{E_{index}}$ = 1.0 eV, σ = 0.2 eV, $\overline{N_{rel}}$ = 2.4 eV, σ = 0.5 eV). Some families of catalysts, such as DHBDs containing the thiourea motif and aminocatalysts, cover a wide range of values ($0.4 < E\text{-index}_{DHBD} < 2.1$ eV), indicating that their electronic properties are highly dependent on the nature of the substituents bound to the catalytic motif. Although it is unlikely that these simple reactivity indices can accommodate a robust and universal scale for electrophilicity and nucleophilicity of such diverse molecules with a varied range of structural, electronic, and bonding properties, the property map in Figure 3.4B and the set of descriptors provided with OSCAR may supplement existing structure–reactivity scales in organocatalysis,[218–224] such as the ones developed by Mayr *et al.*[111–114]

3.2.3 Combinatorial Datasets

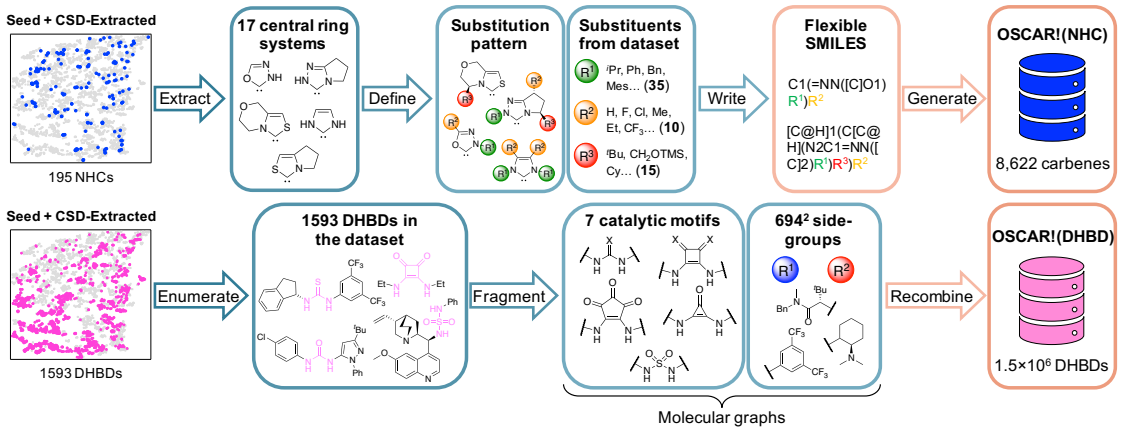


Figure 3.5 Graphical summary of the steps followed to generate the combinatorial databases OSCAR!(NHC) (top) and OSCAR!(DHBD) (bottom). X = O/S.

OSCAR currently covers a significant part of organocatalyst space and a large pool of chemically and functionally diverse catalytic motifs. However, given the nearly infinite number of possible derivatives of each catalyst, only relatively few examples are included. Harnessing the fragment-based strategy used to enrich the seed database with structures from CSD in a bottom-up fashion, we exponentially increase the size of OSCAR by building combinatorial databases from molecular fragments. The exact nature of the fragments depends on the family of organocatalysts, but they can be grouped into two categories: catalytic motifs (*i.e.*, the chemical groups that contain the reactive components) and structural substituents (which modulate their stereoelectronic properties). If the catalytic motif is easily distinguishable from the rest of the molecule (*e.g.*, for dual-hydrogen-bond donors, *vide infra*), it is extracted as a subgraph of the whole catalyst, and the rest handled as structural substituents. If the catalytic motif exhibits larger chemical diversity and substitution patterns (*e.g.*, carbenes, *vide infra*), the possible functional units and substituents are curated manually based on chemical expertise. Herein, we show how to do this for two types of covalent and non-covalent organocatalysts, specifically *N*-heterocyclic carbenes [OSCAR!(NHC)] and dual-hydrogen-bond donors [OSCAR!(DHBD)]. In the first case, a relatively "small" database (8,622 catalysts) is curated by carefully selecting catalytic motifs

and substituents found in OSCAR. In the second, we adopt a graph-based approach to generate 1,573,015 DHBDs.

In the first example (Figure 3.5, top), 17 cores/scaffolds are extracted from the seed and CSD libraries (Figure 3.4A and S9, most central ring system generated with DataWarrior[225]); based on structural features reported in the literature,[226–229] 60 substituents grouped into three categories ($R^{1–3}$, Figures S10–S12) and appropriate substitution patterns are defined. They are then translated into flexible SMILES strings (Table S4), written in such a way that different $R^{1–3}$ in each core can easily be introduced and exchanged. Finally, 3D structures are generated from the SMILES and fully optimized, yielding a database of 8,622 species. In the second example (Figure 3.5, bottom), all the organocatalysts containing one DHBD unit in the seed and CSD-extracted datasets (1,593) are interpreted as molecular graphs[94] (*i.e.*, undirected multigraph with RDKit) and fragmented into the central catalytic motif and the two substituents on either side ($R^{1,2}$), affording a combinatorial space of 7×694^2 groups. After duplicate removal and recombination with RDKit, they yield a total of 1,573,015 species (all optimized at the xTB level); 1,000 structures per each DHBD motif are selected and optimized with DFT, and 6,994 are shown in Figure 3.6B.

The two combinatorial datasets are visualized with chemical space maps (Figure 3.6),[230] which are typically constructed from steric and electronic molecular descriptors. Based on their popularity and chemical meaningfulness, the percent buried volume[231,232] $\%V_{buried}$ and nucleophilicity *N*-index (see the Computational methods) are the parameters chosen for OSCAR!(NHC), while the LUMO energy ε_{LUMO} and the HNNH dihedral angle of the HBD unit (θ) are plotted for OSCAR!(DHBD). The electronic descriptors provide an indirect estimate of the catalysts' Brønsted acidity/basicity: analysis of the experimental equilibrium acidities of 23 NHCs[233] shows that the pK_a values of their precursors (the azolium ions) are directly proportional to the *N*-index of the carbene ($R^2 = 0.80$, $2\sigma = 0.72$, Figure S8), while the LUMO energies of 74 DHBDs[186] scale linearly ($R^2 = 0.92$, $2\sigma = 2.32$, Figure S13) with their experimental pK_a's (as

previously noted by Sigman for a smaller subset).[122] $\%V_{buried}$ and θ quantify the steric influence exerted by the catalysts' core and substituents.

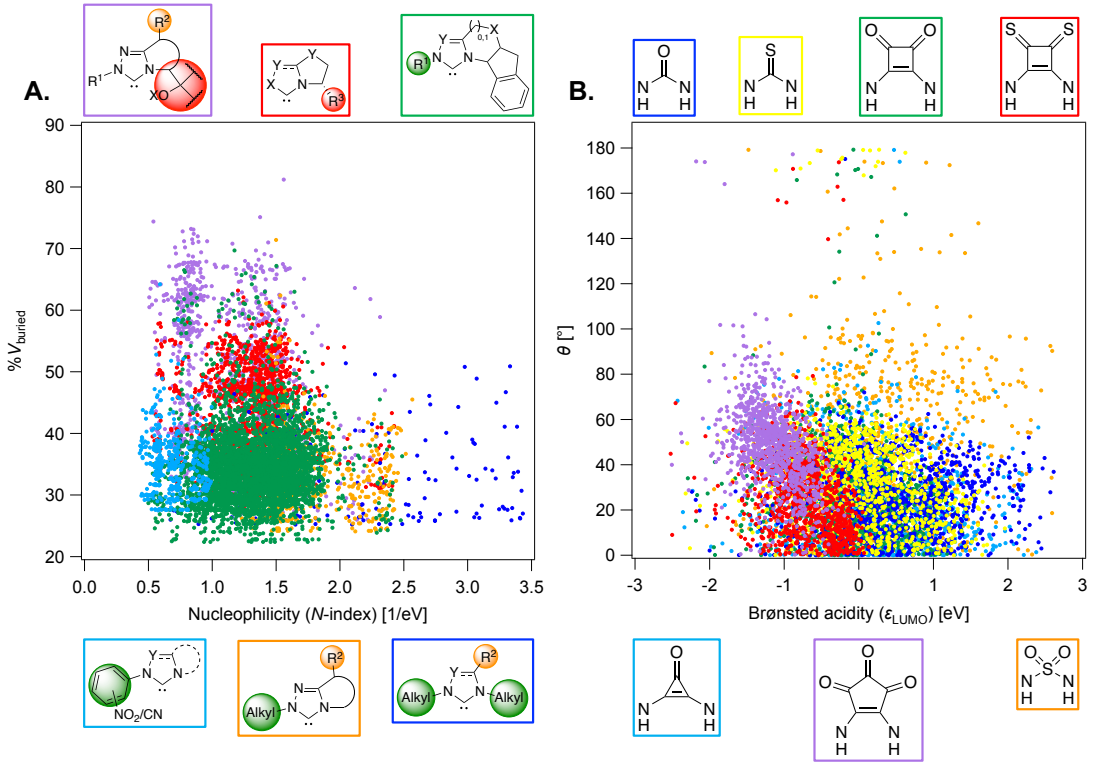


Figure 3.6 A) Percentage buried volume vs. N-index of combinatorial NHC organocatalysts. N-index is found to scale linearly with known experimental pK_a values of azolium ions (Figure S8). B) HNNH dihedral angle (θ) vs. LUMO energy (ωB97X-D/Def2-TZVP//B97-D/Def2-TZVP) of dual-hydrogen-bond donor species. Good linear correlation between ε_{LUMO} and the pK_a's of DHBDs has been found (Figure S13).

The NHCs in Figure 3.6A are colored according to common structural features. The *N*-substituent (R^1 in Figure 3.5, Figure S10) has the greatest effect on nucleophilicity, with catalysts bearing electron-donating alkyl groups [*i.e.*, Me, Et, ^{i}Pr, Cy, and C(Me)Cy] having the highest *N*-index (blue points). These species are predicted to be the most reactive towards electrophilic attack, however their precursors have pK_a's over 20,[233] meaning that relatively strong bases must be used for active catalyst generation. The steric demand of the carbene is mostly influenced by R^3 (Figure S12): L-pyroglutamic acid-derived bicyclic NHCs[234] with diaryl- and diaryl(hydroxy)methyl substituents[235,236] (red and purple points) are located towards the top of the map (large $\%V_{buried}$). Despite their ability to enforce a rigid asymmetric environment, which

could be beneficial in enantioselective reactions, these catalysts are poorly nucleophilic and predicted to be less reactive. Green and orange species, based on the tetracyclic amino-indanol-derived core developed by Rovis and Bode[237,238] and on morpholine- and pyrrolidine-based triazoliums, have more balanced steric and electronic properties and indeed are among the most popular and versatile NHCs used in organocatalysis.[217] Analysis of the descriptors provided with the 8,622 carbenes in OSCAR!(NHC) could eventually be used to tune the catalyst's composition for performance improvement in specific reactions, as outlined in structure–activity–stereoselectivity studies using similar physical organic parameters.[239–241] For example, Rovis, Lee, and co-workers found correlations between the computed gas-phase acidity of a series of triazolium cations and their selectivity in two *Umpolung* reactions,[242] while Wei and Lan developed a linear model to predict the chemoselectivity of an NHC-catalyzed ester functionalization based on the global nucleophilicity and electrophilicity indices of the species involved in the product-determining step.[108]

In Figure 3.6B, each point is colored according to the nature of the central DHBD unit. Based on ε_{LUMO}, and in agreement with pK_a measurements,[243,244] croconamides and thiosquaramides (purple and red species) are more acidic than thioureas, ureas, and deltamides (yellow, blue, and light blue). Sulfamides (orange points) cover a relatively wider range of ε_{LUMO} values, implying that the higher electron-withdrawing ability of the sulfonyl group, which should result in stronger acidity of the N–H bonds compared to ureas,[245] is significantly modulated by the substituents. The rapid estimation and comparison of the acidity of various DHBDs is useful for reaction optimization, as dual-hydrogen-bond donors with lower pK_a's have been found to give better enantioselectivities and faster reaction times.[116] Sulfamides are also the most flexible species, as indicated by the large number of catalysts with $\theta > 80°$. In OSCAR!(DHBD), the majority of structures generated and selected for DFT optimization are in the *anti-anti* or *syn-syn* conformation ($\theta < 80°$, Figure 3.7D and Figure S17),[246] the former being the most relevant to catalysis, since the hydrogens point in the same direction.[247]

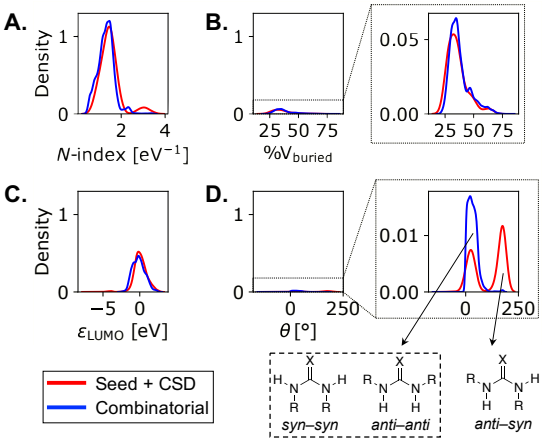


Figure 3.7 Distribution plots (y-axis: normalized probability density) of molecular descriptors for NHCs (A and B) and DHBDs (C and D) in the seed + CSD-extracted (red curves) and combinatorial databases (blue). X = O/S.

If we compare the distribution of θ values in the "original" and combinatorial datasets (Figure 3.7D), we see that many CSD-extracted DHBDs adopt the *anti-syn* conformation ($\theta > 80°$). In a comprehensive study of diaryl(thio)ureas from CSD, Paton *et al*. found that the majority (99%) of ureas exist as *anti-anti* conformers, whereas about 60% thioureas are in the *anti-syn* form.[248] These results agree with our own, with thioureas extracted from CSD having large θ's (Figure 3.7D). The "original" and combinatorial sets are more similarly distributed in terms of the other molecular descriptors (Figure 3.7A–C, *N*-index, $\%V_{buried}$, and ε_{LUMO}), suggesting that the recombination of the same fragments does not significantly alter the property space covered; instead, the combinatorial strategy provides more instances/structures for each property value.

3.3 Conclusions

We have introduced OSCAR (Organic Structures for CAtalysis Repository), a database of 4,000 organocatalysts mined from the literature and CSD and enriched with several thousand species generated from fragments in a combinatorial fashion. We have developed a transferable fragment-based strategy for dataset generation, which exploits the modularity of organocatalysts by defining function-based catalytic motifs and structural substituents. OSCAR covers a wide region of catalyst space with incomparable chemical diversity, and includes a selection of steric

and electronic molecular descriptors useful for catalytic properties estimation and performance prediction. All content (geometries, stereoelectronic parameters) is publicly available on the Materials Cloud for interactive visualization with Chemiscope[174] (https://doi.org/10.24435/materialscloud:gy-3h) and fully searchable and interoperable with chemoinformatics software (*e.g.*, RDKit, SMILES-based tools); the corresponding chemical space maps could be used for many potential applications, including data and training set curation, organocatalyst inverse design through evolutionary experiments,[38] and mechanistic understanding. We expect OSCAR, and its future extensions and refinements, to assist in the establishment of data-driven and fragment-based reaction optimization methods in organic synthesis.[33]

3.4 Computational Methods

3.4.1 Quantum Chemistry

All DFT computations were performed with the Gaussian16 software package.[249] Geometry optimizations were carried out at the B97-D/Def2-TZVP level[250–252] in the gas-phase applying density fitting techniques. ωB97X-D/Def2-TZVP single-point energies[253] were computed in the gas-phase at the B97-D geometries. The ionization potential and electron affinity of a subset 2,060 organocatalysts from the seed and CSD datasets were also computed at the IP/EA-EOM-DLPNO-CCSD[254]/cc-pVTZ level as implemented in Orca 5.0.[255] All coupled cluster computations used the RIJCOSX approximation[256] with the cc-pVTZ/C and the Def2/J auxiliary basis sets for correlation and resolution of identity. This high-level data is available and can be used for the training of ML models. The structures in the combinatorial databases were pre-optimized with the semiempirical GFN2-xTB Hamiltonian[257] in the gas-phase, followed by DFT optimizations and single-points, as described above.

The initial set of Cartesian coordinates for each organocatalyst was either obtained by converting SMILES formats[258] into three-dimensional structures with the 3D structure generator operation (*i.e.*, gen3d operation) implemented in the OpenBabel software,[259] or applying *cell2mol*[200] on

selected CSD entries exported with ConQuest (version 5.42), included in the CCSD software, from the CSD database updated to May 2021. The tSNE map[207] for the 4,000 catalysts in OSCAR was computed on the basis of the FCHL19 representation[208] of each molecule. The perplexity used to generate the structure map was set to 20 and the maximum number of optimization iterations was fixed at 5,000.

Open shell single-point computations (n–1 and n+1 electrons) were also performed at the optimized n-electron B97-D geometries and uωB97X-D/Def2-TZVP level for the 4,000 catalysts in the seed + CSD dataset and for the 8,622 carbenes in OSCAR!(NHC). These energies provide an alternative way of estimating the organocatalysts' ionization potential [IP = E(n-1) – E(n)] and electron affinity [EA = E(n) – E(n+1)] (see the ESI for further details).[260]

3.4.2 Reaction Indices

The organocatalysts' ionization potential (IP) and electron affinity (EA) were estimated from the frontier molecular orbital energies (FMOs) of the n-electron species (in the gas-phase, at the ωB97X-D/Def2-TZVP level) using Koopman's theorem[261] within a Hartree–Fock scheme and used to calculate the conceptual DFT descriptors[110,262,263] chemical potential (μ), hardness (η), E-index, N-index, and relative nucleophilicity (N_{rel}) as follows:

$$\mu = \frac{(\varepsilon_{LUMO} + \varepsilon_{HOMO})}{2} \quad (1)$$

$$\eta = \frac{(\varepsilon_{LUMO} - \varepsilon_{HOMO})}{2} \quad (2)$$

$$E\text{-index} = \frac{\mu^2}{2\eta} \quad (3)$$

$$N\text{-index} = \frac{1}{E-\text{index}} \quad (4)$$

$$N_{rel} = \varepsilon_{HOMO} - \varepsilon_{HOMO(TCNE)} \quad (5)$$

where TCNE is tetracyanoethylene. Note that, based on the different formalisms for defining nucleophilicity,[264] a distinction has been made between *N*-index (the reciprocal of the *E*-index) and relative nucleophilicity (N_{rel}).[265]

3.5 Supporting Information

The Supporting Information for this Chapter may be found at

https://www.rsc.org/suppdata/d2/sc/d2sc04251g/d2sc04251g1.pdf

4 Harvesting the Fragment-Based Nature of Bifunctional Organocatalysts to Enhance their Activity

This chapter is based on following publication:
Gallarati S., Laplaza R., and Corminboeuf C., Harvesting the Fragment-Based Nature of Bifunctional Organocatalysts to Enhance their Activity. *Org. Chem. Front.* **2022**, *9*, 4041.

4.1 Introduction

Organocatalysts often incorporate different motifs for divergent catalytic functionalities, connected by flexible (a)chiral linkers, to simultaneously activate reaction partners through covalent and non-covalent interactions (NCIs).[166,266,267] Bifunctional hydrogen-bond donor (HBD)–primary/secondary amine catalysts are among the most common arrangements and have found widespread use in enantioselective transformations due to the large variations in functionalities that can be accommodated into their structure.[268–270] They are assembled on a modular basis, which allows for easy synthetic modifications, but can complicate optimizing their structure.[271] Multiple catalyst components must be evaluated, potentially in a combinatorial fashion, and the most apt combinations identified.[272] This is typically accomplished by making small modifications on a privileged motif guided by chemical intuition or trial-and-error.[141,273] Computational approaches help in identifying structural aspects pertinent to reactivity and inform the design of improved catalysts.[9,103,118,171,274,275] A strategy popularized by Sigman and co-workers is to correlate physical organic descriptors to experimental activity or selectivity outcomes *via* multivariate regression analysis.[12,13,28,54,55,119,121,276,277] When applied to

organocatalysts, the descriptors are typically evaluated on the catalyst structure as a whole or, occasionally, from truncated versions of it.[278] Despite the locality of certain parameters employed (*e.g.*, NBO charges, IR stretching frequencies, NMR chemical shifts, *etc.*, see Figure 4.1), this approach does not fully exploit the fragment-based nature of organocatalysts and has limited transferability because, even when small modifications on part of the catalyst are made, its entire structure must be re-optimized and the parameters collected. On the other hand, modularity is a well-leveraged feature of transition-metal catalysts, whose structure is easily separated into active metal center, metal-coordinating groups, backbone/bridging fragments, and inert substituents.[27] Therefore, catalyst design strategies that rely on ligand molecular electronic and steric descriptors are widely documented in the literature.[150,154,155,157,279–282]

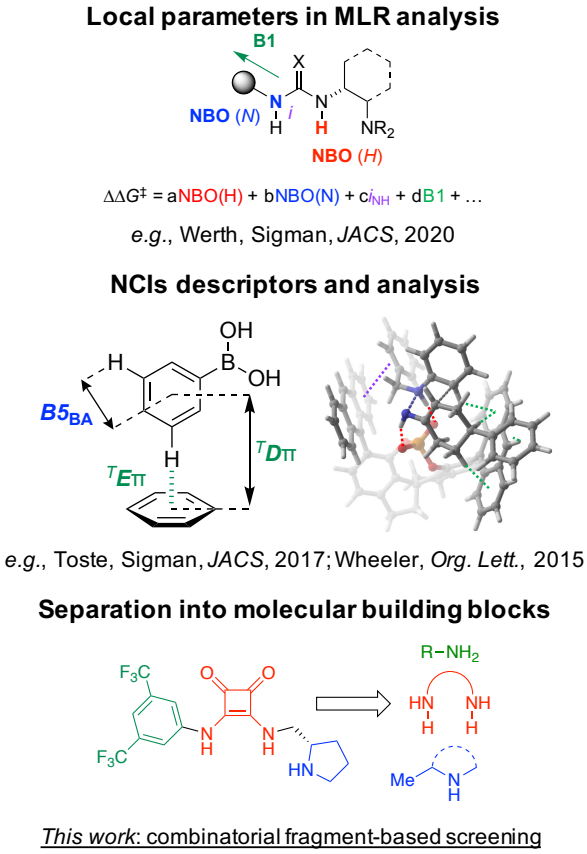
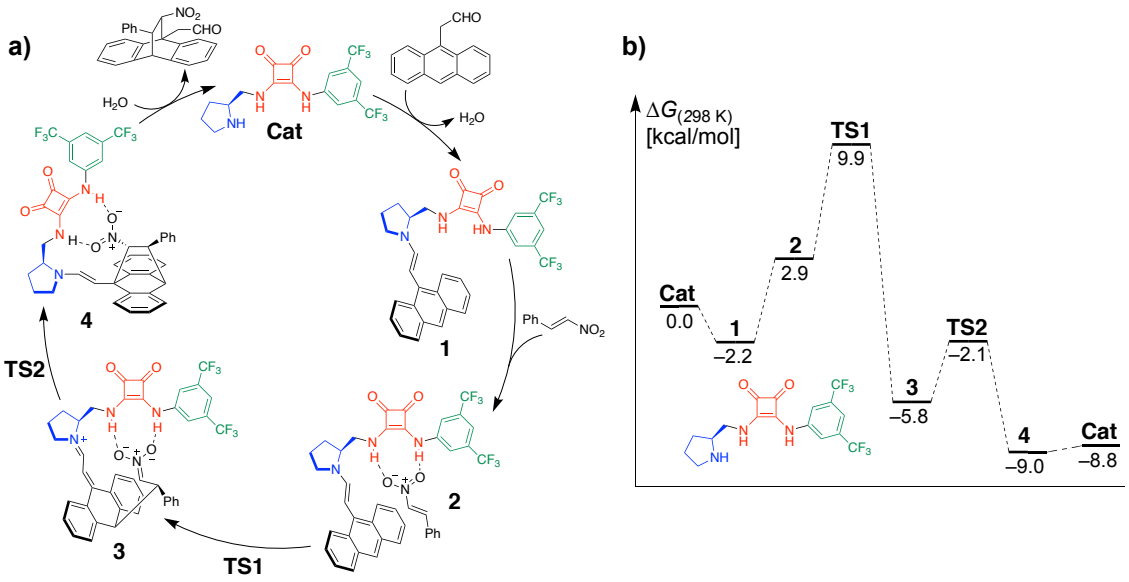


Figure 4.1 Fragment-based approaches for organocatalyst design.

Alternative approaches aim at identifying exactly which NCI motifs between the catalyst and the substrate drive the selectivity and probe their individual contribution to the reaction

outcome.[8,23,71] To this end, the energy difference between stereocontrolling transition states has been correlated with interaction energies associated with the relevant chemical motifs or distances using truncated structures[53,56,57,60,283,284] or decomposition schemes.[25,285] Despite the success of these fragment-based methods, the modular nature of multifunctional organocatalysts is not fully taken into account and the approach is not applicable to cover properties that also rely on the influence of covalent bonding.

In this work, we present a strategy to improve the performance of bifunctional organocatalysts using a fragment-based feedback approach. Since tools to rationalize/predict the enantioselectivity of organocatalysts are well-established,[22,41,286] here we focus on enhancing activity, which is often suboptimal (*vide infra*) and more ambiguous to optimize computationally. To achieve this goal, we exploit the volcano plot concept[30] and define a library of catalytic motifs *i.e.*, the chemical groups that present the catalytic functional components,[166] which is used to assemble a first class of organocatalysts. Evaluating the building blocks individually allows us to extract fragment-based design principles, as well as making the catalyst screening process faster and more transferable. In turn, the trends gathered from the generated plot and the fragment parameters evaluated through statistical modeling are used to enrich the library with additional fragments and suggest subsequent catalyst designs. Finally, an activity map provides feedback on the combinations of fragments chosen, and whether the maximum theoretically achievable turnover frequency (TOF) has been reached.[16,287] We investigate the Diels–Alder cycloaddition of anthracene and nitrostyrene (Scheme 4.1)[288,289] to verify whether the popular pyrrolidine/squaramide organocatalyst is actually optimal for this reaction, and find that further activity enhancement is still achievable. Additionally, this strategy is generalizable to any system composed of different organocatalytic functionalities.



Scheme 4.1 a) Catalytic cycle of Diels–Alder cycloaddition of 2-(anthracene-9-yl)ethanal and β-nitrostyrene catalyzed by a bifunctional organocatalyst. b) Computed Gibbs free energy profile (at the ωB97X-D/Def2-TZVP//B97-D/Def2-SVP level) of the pyrrolidine/squaramide catalyst.

4.2 Computational Details

Based on the work by Wheeler *et al.*,[289] structures were optimized at the B97-D/Def2-SVP level of theory[250–252] using the PCM method[290,291] to account for the impact of solvent (dichloromethane) and applying density fitting techniques. PCM(DCM)/ωB97X-D/Def2-TZVP single-point energies[253] were computed at the B97-D geometries. Transition state structures of class 0 organocatalysts (*vide infra*) were located using the AARON[78] toolkit starting from a TS library assembled using the structures from the lowest-energy pathway reported by Wheeler *et al.*[289] AARON automates the generation and investigation of the conformational space of flexible organocatalysts, helping to locate low-energy pathways.[78] Intermediates were obtained from IRC computations.[292] Stationary points were characterized on the basis of their vibrational frequencies (minima with zero imaginary frequencies, TSs with one imaginary frequency). Free energy corrections (298 K) were determined using the quasi rigid-rotor harmonic oscillator model using the GoodVibes program with a frequency cut-off value of 100 wavenumbers.[293] The relative Gibbs free energies were automatically post-processed using the toolkit volcanic[32] to establish linear free energy scaling relationships (LFESRs), determine the choice of the descriptor variable [the relative energy of intermediate **3**, ΔG_{RRS}(**3**)], and construct the TOF-volcano plots and

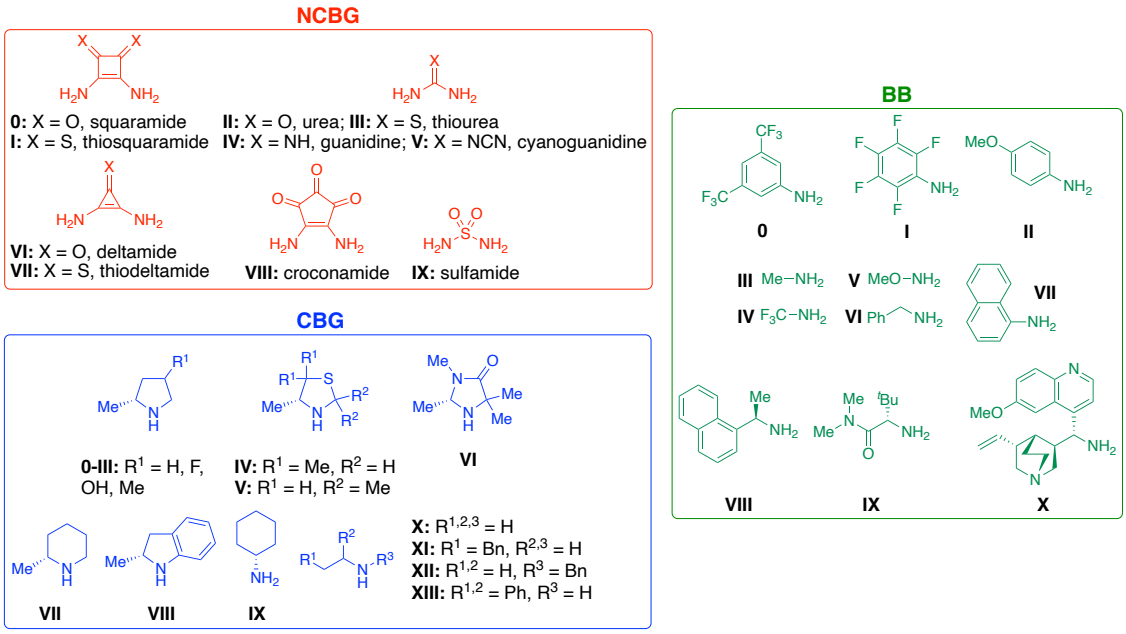
activity maps. All DFT computations were performed with the Gaussian16 (revision A.03) software package.[249]

Multidimensional regression analyses were performed on the TOF-volcano plot descriptor variable [ΔG_{RRS}(**3**)] using the collected fragment parameters (see Tables S3–S5) and an in-house Python script openly available on GitHub (https://github.com/lcmd-epfl/mlr_organocatalyst_fragments). As ease of interpretation of the model was considered valuable for mechanistic understanding, cross-terms of parameters were removed during model search. Variance inflation factors were computed to ensure that no collinearity was present. Normalized parameters were employed so that the coefficients would reflect the importance of each descriptor. Good linear correlation (R^2 close to 1.00 and intercept close to 0.00) between predicted and measured ΔG_{RRS}(**3**) was considered to evaluate whether a model adequately describes the system under study. Several linear equations were tested (some are given in Figures S3–S4); the MAE, number of parameters, and the adjusted R^2 were decisive factors in choosing the final form. Cross-validation analysis and external validation were used to indicate a robust model. Leave-One-Out (LOO) cross validation of the model in Figure 4.2d led to a MAE of 1.78 ± 1.54 kcal/mol (R^2 = 0.72), while 5-fold cross-validation led to a MAE of 1.76 ± 0.32 kcal/mol (R^2 = 0.72). External validation was performed by pseudorandom partitioning the class 0 data set into 50:50 training set/validation sets, leading to a MAE of 1.71 ± 0.06 kcal/mol (R^2 = 0.71).

4.3 Results and Discussion

Compared to transition-metal catalysts, the applicability of organocatalysts, especially in industrial settings, remains limited by their low activity.[1,5] Cycloaddition reactions catalyzed by pyrrolidine/squaramide catalysts generally require high loadings (16-20 mol%).[294–299] Interestingly, even in the most efficient one developed by Jørgensen and co-workers (the dearomatization of anthracenes, see Scheme 4.1),[288] the activity can still be improved, as we show in this work. The reaction mechanism was studied by Wheeler *et al.*[289] and consists of three main steps: HOMO-raising activation of the diene through condensation of the aminocatalyst with the

aldehyde to give enamine **1**, stepwise [4+2] cycloaddition with nitrostyrene *via* zwitterionic intermediate **3**, and hydrolysis of the post-reaction complex **4** (Scheme 4.1). The squaramide unit helps forming the pre-reaction complex **2** by lowering the LUMO of the dienophile and directing its attack through double hydrogen-bonding, ensuring high stereoselectivity. The bifunctional organocatalyst combines three types of fragments: an amino function for enamine catalysis (henceforth called the covalent binding group, or CBG), a unit capable of forming dual hydrogen-bonds (the non-covalent binding group, or NCBG), and a structural backbone (BB).



Scheme 4.2 Library of function-based fragments for "class 0" organocatalysts. NCBG = non-covalent binding group; CBG = covalent binding group; BB = backbone. Note that, in the bifunctional organocatalyst, CBG is bound to NCBG via the methyl group.

A library of molecular building blocks is curated, consisting of 10 NCBGs (red species in Scheme 4.2), 14 CBGs (primary and secondary amines, blue species) and 11 (a)chiral BBs (in green). The fragments are chosen to include some of the most chemically relevant and frequently seen motifs in enamine and HBD organocatalysis, as well as groups with different enough electronic and steric properties to afford robust LFESRs. They are combined to form a "class 0" of 101 organocatalysts (33 catalysts with one unit in Jørgensen's original compound kept constant and the other two systematically varied, 68 catalysts with random NCBG–CBG–BB combinations). On the basis of Wheeler's computed mechanism (Scheme 4.1),[289] the relative energies of

intermediates **1**–**4** and transition states **TS1**–**2** of these 101 organocatalysts are computed and used to construct the TOF-volcano plot shown in Figure 4.2 (see the Computational Details). The volcano plot has a twofold purpose: first, it helps rationalize the effect of each individual fragment on activity and establish reactivity trends. Second, it provides an estimation of the maximum achievable theoretical TOF and a direct link between an easy-to-determine descriptor variable [ΔG_{RRS}(**3**)] and a measure of activity. A volcano plot or activity map can be used as a roadmap towards optimized catalysts; computing (or predicting) the descriptor variable for new fragment combinations allows larger databases to be screened. Details on how the volcano plot is constructed, including how the cusp is determined, are given in ref. 287,300 and in the SI.

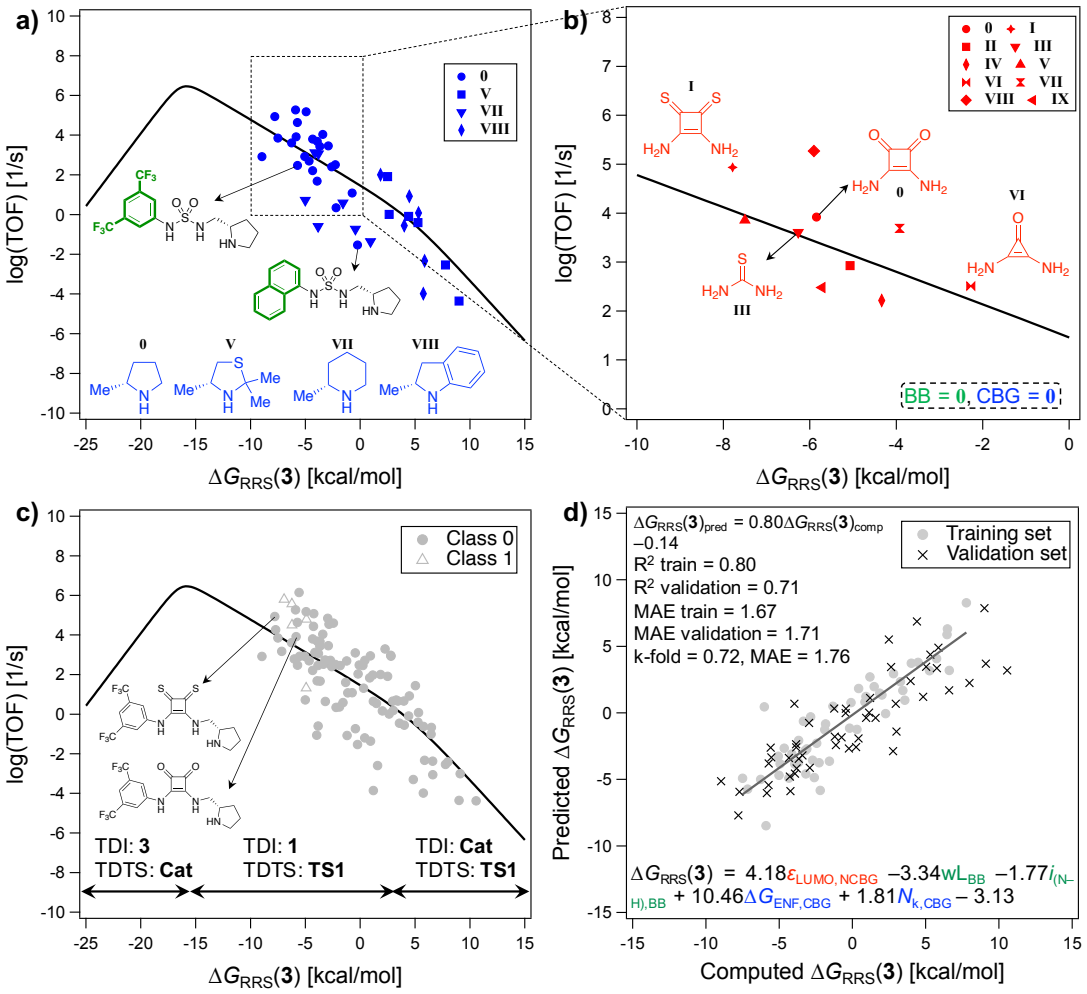


Figure 4.2 a) Influence of CBG and BB on TOF (only CBG fragments **0**, **V**, **VII**, and **VIII** displayed). b) Zoom-in of pyrrolidine organocatalysts with different NCBGs (only catalysts with CBG **0** and BB **0** displayed). c) Class 0 *vs*. 1 (all catalysts displayed). d) MLR analysis of the descriptor variable in terms of parameters of the individual molecular building blocks.

The CBG (blue fragment) has the largest influence in determining where on the volcano an organocatalyst falls (Figure 4.2a). This is not surprising since covalent interactions are stronger than non-covalent ones and the substate is primarily activated through HOMO-raising *via* enamine formation. The CBG trend follows the order pyrrolidine (**0**, closer to the volcano peak) > piperidine (**VII**) > indoline (**VIII**) ~ 2,2-dimethylthiazolidine (**V**). Note that, even though the volcano plot was constructed using all 101 organocatalysts (Figure 4.2c), Figure 4.2a only shows trends for CBGs **0**, **V**, **VII**, and **VIII** for ease of understanding. In the case of **V** and **VIII** (right slope of the volcano), enamine formation (*i.e.*, **Cat** → **1**, Scheme 4.1) is unfavorable: the steric influence exerted by substituents on the α-position and the low Lewis basicity of the N-atom

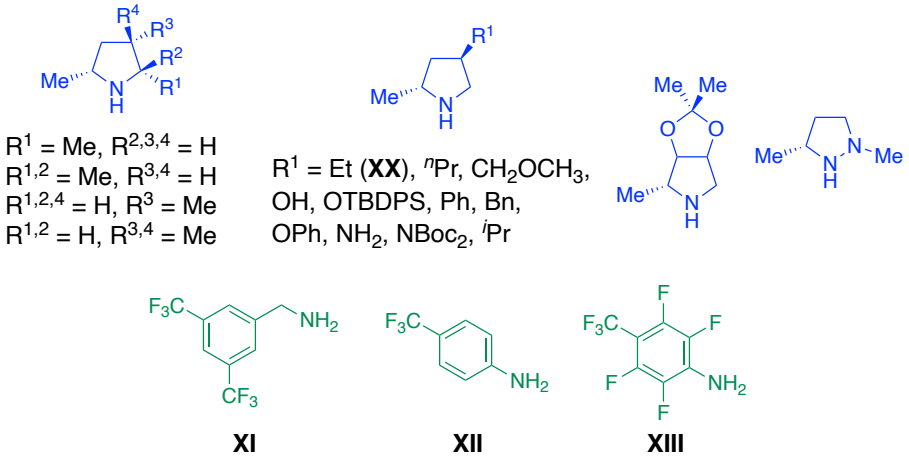
make enamine **1** unstable. Since forming **1** is energetically uphill, the catalytic resting state (the TOF-determining intermediate, or TDI)[301] is **Cat**, while the TOF-determining transition state (TDTS) is forming the first C–C bond in the cycloaddition reaction (**TS1**), which is particularly high in energy for these poorly nucleophilic CBGs. For **0** and **VII**, enamine formation is favorable (*i.e.*, **1** becomes the resting state since **Cat** → **1** is exergonic), and these fragments are limited by conjugate addition (TDTS: **TS1**). Pyrrolidine **0** lies higher on the volcano than **VII** due to the higher sp^2–character of the N-atom, leading to better donation of the electron density from the non-bonding N-orbital into the π* orbital of the enamine C=C bond and to higher nucleophilicity.[302,303] Pyrrolidine **0** could be improved by making it more nucleophilic (*i.e.*, better delocalization of the N-lone pair and higher enamine reactivity); however, since stabilizing **TS1** (TDTS) also makes **1** (TDI) more stable, to achieve higher turnover the barrier for conjugate addition (**TS1**) must be reduced more significantly than the concurrent stabilization of **1**. In this sense, the choice of BB (green) is also important because it can tune the effect of CBG (Figure 4.2a) and of NCBG.

Since NCBG (red fragment) non-covalently binds the substrate, it exerts a more subtle influence on activity than CBG. Closer inspection (Figure 4.2b) at variations of Jørgensen's catalyst (*i.e.*, CBG and BB kept constant) shows that more acidic and stronger hydrogen-bond donors display slightly higher TOF. The trend in catalytic activity thiosquaramide (**I**) > squaramide (**0**) ~ thiourea (**III**) > deltamide (**VI**) reflects that in their Brønsted acidity.[243,244,246,304,305] Better HBDs more significantly reduce the LUMO of nitrostyrene, yielding a lower **TS1**. Indeed, HBD organocatalysts with lower pK_a's have been found to give faster reaction times (and better enantioselectivities).[116]

Jørgensen's pyrrolidine/squaramide catalyst, or Wheeler's thiosquaramide derivative (Figure 4.2c), are actually not predicted to display the maximum achievable TOF (and neither are the other 99 candidates examined), inciting for further improvement. To find better-performing organocatalysts, the remaining 1,439 combinations in class 0 are screened by computing the

value of the descriptor variable [ΔG_{RRS}(**3**)]. To accelerate the process and avoid generating and optimizing the structure of intermediate **3** (and of **Cat**) for all 1,439 catalysts, we exploit the organocatalysts' modular nature by evaluating ΔG_{RRS}(**3**) in terms of molecular descriptors of the individual fragments (Figure 4.2d). ΔG_{RRS}(**3**) is thus estimated through multivariate linear regression (MLR) from only five steric and electronic fragments parameters with sufficient accuracy (R^2 = 0.80, MAE = 1.7 kcal/mol): the LUMO energy of NCBG (ε_{LUMO}), the N–H IR stretching intensity ($i_{N–H}$) and the Boltzmann-weighted L Sterimol parameter[96] of BB (wL), the local nucleophilicity[110] at the N-atom (N_k) and the free energy of enamine formation (ΔG_{ENF}) of CBG (Figure S2).

The presence of two CBG parameters and the highest coefficient value carried out by ΔG_{ENF} (Figure 4.2d) highlights the importance of the amino-motif. Smaller N_k and ΔG_{ENF}, corresponding to highly nucleophilic amines that form stable enamines, make ΔG_{RRS}(**3**) closer to its optimal value. In agreement with the trends extracted from the volcano plot (Figure 4.2a), pyrrolidine has the lowest ΔG_{ENF}; electron-donating groups make the N-atom more nucleophilic and help stabilize both **TS1** (the TDTS) and **1** (the TDI), whereas CBGs with electron-withdrawing or larger substituents that cause unfavorable steric clashes have larger ΔG_{ENF} and N_k, and a higher barrier for conjugate addition (**TS1**).[306,307] Only one parameter describes NCBG, carrying the second highest coefficient; Werth and Sigman recently observed excellent correlation between LUMO energies and experimental pK_a values of bifunctional HBD-tertiary amine organocatalysts.[286] This suggests that ε_{LUMO} provides an indirect measure of the acidity of the non-covalent motif and that more acidic groups make ΔG_{RRS}(**3**) closer to the optimum (thus reducing ΔG_{RRS}(**TS1**), Figure 4.2b). Finally, electron-withdrawing BBs (as indicated by $i_{N–H}$) that enhance the acidity of NCBG and are long enough to avoid clashes with the anthracene substrate give lower ΔG_{RRS}(**3**) values (*i.e.*, **1** and **TS1** become both stabilized, Figure 4.2a).

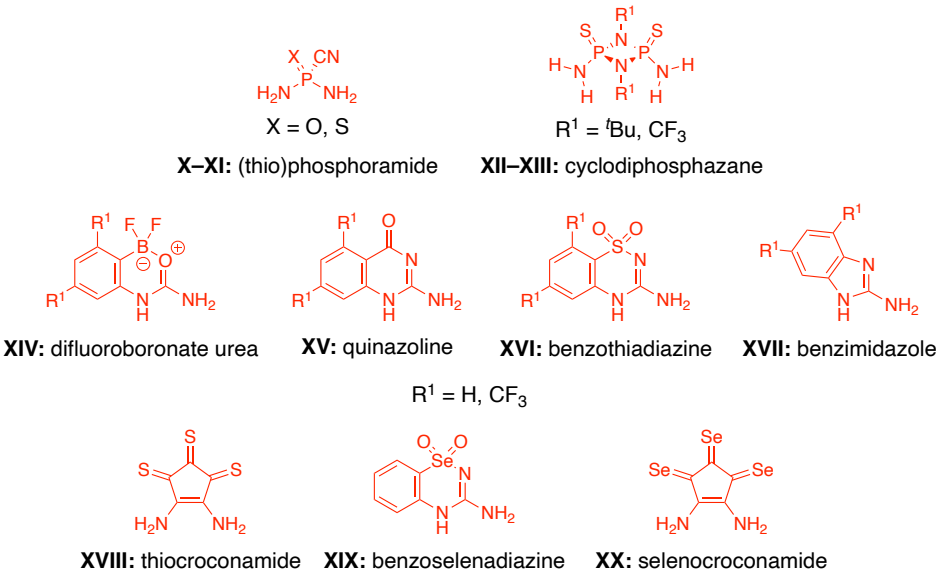


Scheme 4.3 Additional CBG (blue) and BB (green) fragments investigated in the search of improved organocatalysts.

Taking advantage of the fact that the parameters for all 35 building blocks were computed to fit the MLR expression, and that the remaining 1,439 catalysts are simply new combinations of the fragments in Scheme 4.2, the entire fragments combinatorial database is then screened without the need of further computations, simply inputting the corresponding descriptor values in the MLR equation. Surprisingly, no combination is found to correspond to the volcano peak [ΔG_{RRS}(**3**) ≈ –16 kcal/mol]. The 5 combinations of fragments predicted to give the lowest ΔG_{RRS}(**3**) values (*i.e.*, NCBG **VIII**; CBG **0**, **II**–**III**; BB **0**–**I**, Scheme 4.2) are assembled into a "class 1" of organocatalysts (Figure 4.2c). Despite minor improvements, these croconamide (**VIII**) species are still predicted to be suboptimal with respect to the theoretical TOF maximum. Clearly, additional fragments must be introduced to push the catalyst's activity even further.

Since CBG (blue) exerts the greatest influence on activity, better amino-fragments than pyrrolidine must be identified for major TOF enhancements. Therefore, 17 additional pyrrolidine derivatives (Scheme 4.3) are analyzed in terms of their nucleophilicity and ability to form stable enamines (Figure S6). Based on ΔG_{ENF}, only the aminocatalyst ethylated at the β-position (**XX**) surpasses CBG **0**. Vilarrasa *et al*. similarly found that none of the *sec*-amines they examined yielded more stable enamines than pyrrolidine, with steric effects generally counteracting the increase in nucleophilicity afforded through the introduction of electron-donating groups.[307]

Three longer BB (green) fragments with electron-withdrawing substituents are also evaluated (Scheme 4.3) because they are predicted to make ΔG_{RRS}(**3**) closer to the optimum. Unfortunately, they are found to significantly stabilize intermediate **1**, causing an overall decrease in TOF (Figure S10–11). In agreement with results by Schreiner *et al.*,[308] the 3,5-bis(trifluoromethyl)phenyl group (BB **0**, Scheme 4.2) is found to be the privileged BB required for high levels of activity.



Scheme 4.4 Additional non-covalent binding groups (NCBGs) introduced. Combined with the best-performing CBGs and BBs, these fragments constitute an enhanced "class 2" of organocatalysts.

As CBG and BB are essentially optimized, further improvements can be achieved by enhancing the acidity of NCBG (red). To this end, less established H-bonding motifs (NCBGs **X**–**XX,** Scheme 4.4) are considered. These include experimentally reported preorganizing and acidifying linkers,[309–312] and three other sulfur- and selenium-based fragments. Because thio-derivatives typically outperform their oxygen counterparts in terms of acidity and activity,[289] we sought to further increase the efficiency of the catalyst by introducing a *thiocroconamide* NCBG (**XVIII**, Scheme 4.4), which is predicted to possess even lower ε_{LUMO} (hence making ΔG_{RRS}(**3**) closer to the optimum, according to MLR analysis). Given the growing interest in the application of selenium-containing HBD analogues,[313–315] two more NCBG designs, **XIX** and **XX**, are included.

Selenium has similar properties to sulfur but with a larger atomic radius: it can accommodate more negative charge, which could induce a stronger N–H acidity.[316] NCBGs **X**–**XX** (Scheme 4.4) are combined with the best-performing CBG and BB fragments to yield 40 additional candidates grouped in "class 2" (see Figures S6–7). The relative energies of their corresponding stationary points are computed and used to construct the maps in Figure 4.3 (see the Supporting Information).

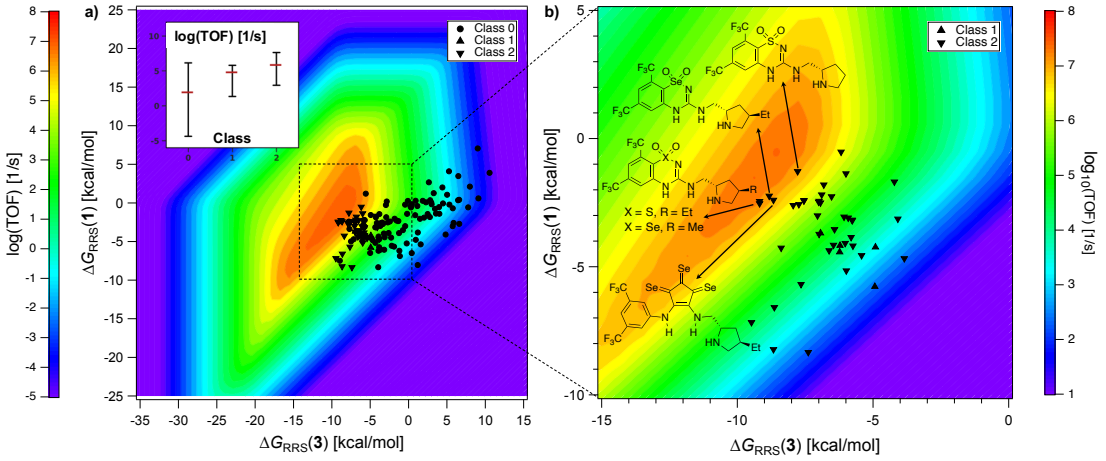


Figure 4.3 Activity map. b) Zoom-in on classes 1–2 showing the best-performing organocatalysts.

The performance of all organocatalysts is compared based on the activity map in Figure 4.3a. The inset shows the progressive improvement in the median of computed TOF values from class 0 to class 2. The purpose of the map is to evaluate whether a combination of fragments corresponding to the TOF maximum (most central region) has been found. The introduction of a second descriptor variable [ΔG_{RRS}(**1**)] aims at separating some of the points that would otherwise be clustered on a 1D volcano by highlighting the adverse effect of fragments that over-stabilize enamine **1**. Fragments that lead to ΔG_{RRS}(**1**) < –3 kcal/mol [and ΔG_{RRS}(**3**) > –7 kcal/mol] result in reduced turnover. The best combinations of fragments are located in the central region of the map (*i.e.*, highest turnover, Figure 4.3b). These include catalysts bearing benzothio- and benzoselenadiazine NCBGs (**XVI**, **XIX**, Scheme 4.4), as well as the highly acidic selenocroconamide species **XX** (Scheme 4.4). Clearly, increasing the Brønsted acidity of the

HBD moiety affords higher activity. Although selenium-containing NCBGs **XIX**–**XX** (Scheme 4.4) might not be synthetically accessible, the presence of the benzothiadiazine **XVI** species in the same area of the map is encouraging. This scaffold was reported by Takemoto and co-workers to outperform other common HBD organocatalysts in the isomerization of alkynoates to allenoates,[312] the intramolecular oxa-Michael reaction of *α*,*β*-unsaturated amides,[317] and in an asymmetric Mannich-type reaction,[318] and has the additional advantage of inducing higher structural rigidity, fixing the catalyst conformation in a catalytically active form.[309] The combination of this NCBG with pyrrolidine-based fragments bearing small electron-donating alkyl groups (Me, Et) at the *β*-position and Schreiner's 3,5-bis(trifluoromethyl)phenyl backbone leads to enhanced activity, balancing the opposing effects of a stabilized rate-limiting transition state (which increases turnover) and an over-stabilized resting state (which decreases it).

While the map in Figure 4.3a is limited to the description of the reaction under study (Scheme 4.1), the protocol for its construction, and the fragment-based strategy that has led to the identification of better-performing building blocks, is generalizable to other transformations. The chemical trends gathered from the analysis of the Figure 4.2 volcano plots and from statistical modelling are likely transferable to different reactions catalyzed by bifunctional hydrogen-bonding catalysts. Pyrrolidine (CBG) and the 3,5-bis(trifluoromethyl)phenyl group (BB) are essentially optimal, and maximum activity can be reached by modulating the acidity of the HBD bonding unit (NCBG). Simple fragment parameters that estimate, for example, the nucleophilicity of the covalent group, or the Brønsted acidity of the HBD, will be useful to predict the performance of organocatalysts in mechanistically-related transformations.

4.4 Conclusions

In summary, we have shown how the modular nature of organocatalysts can be exploited through a fragment-based computational approach to suggest structural modifications for enhanced activity. This strategy relies on curating chemically diverse families of catalytic motifs and evaluating catalyst activity in terms of the individual fragment contributions. Trends and

optimum regions extracted from volcano plots and statistical modeling help choosing which additional fragments must be added to the library for improved turnover. An activity map shows whether a combination of fragments corresponding to the TOF maximum has been found. Specific to the cycloaddition reaction studied here, enhancing the Brønsted acidity of the HBD unit and enforcing some conformational rigidity is essential to push the activity limit of the organocatalyst, whereas further optimizing the covalent chemical motif or the side group is hindered by the increasing stability of the catalytic resting state, which reduces turnover. We have shown how in even seemingly optimal catalytic systems involving the commonplace pyrrolidine/squaramide bifunctional organocatalyst there is room for improvement, and that the development of innovative HBD scaffolds with increased acidity is a focal point in hydrogen-bonding catalysis. We expect this approach to be broadly applicable and beneficial for the optimization of other organocatalytic systems.

4.5 Supporting Information

The Supporting Information for this Chapter may be found at

https://www.rsc.org/suppdata/d2/qo/d2qo00550f/d2qo00550f1.pdf

5 Genetic Optimization of Homogeneous Catalysts

This chapter is based on following publication:
Laplaza R., Gallarati S., and Corminboeuf C., Genetic Optimization of Homogeneous Catalysts. *Chem. Methods* **2022**, *2*, e202100107.

5.1 Introduction

This work introduces NaviCatGA, a software package capable of optimizing catalysts by exploiting any suitable fitness function that describes their catalytic performance. It manipulates catalyst structures generated *in situ* from a user-defined library of molecular fragments (metal centers, ligands or ligand substituents, scaffolds, *etc*.); structures can be assembled from the respective components using any representation, including SMILES strings and XYZ coordinates, and evaluated according to any fitness function (*e.g.*, molecular volcano plot descriptors,[30,31] multivariate linear regression expressions[13]). NaviCatGA is a modular part of the broader NaviCat (**Navi**gating **Cat**alysis) platform for catalyst discovery (https://github.com/lcmd-epfl/NaviCat), which includes other utilities and tools (*e.g.*, database constructors,[29] automatic volcano plot builder,[32] *etc*.).

In the spirit of inverse design,[34,35,86,157,319] NaviCatGA uses a Genetic Algorithm (GA)[15,27,320,321] to find optimal catalysts (Figure 5.1). This pipeline represents a complementary approach to high-throughput screening[85,136,322–324] that becomes comparatively more efficient as the dimensionality of the combinatorial space of catalyst components grows. Furthermore, evolutionary experiments with GAs lead to alternative chemical insight into catalyst performance, as demonstrated hereafter. They have been shown to be well-suited for molecular optimization[320,325,326] because

they are able to address discontinuities in structure-property space (*e.g.*, activity cliffs)[136,327] and, more importantly, do not require meaningful gradients for the optimization. Nonetheless, flexible and robust implementations of GA algorithms tailored for homogeneous catalysis were lacking.

The versatility and efficiency of NaviCatGA are illustrated with two representative applications to transition-metal and organocatalyzed reactions. The goal is to show that closed-loop optimization with genetic algorithms is an efficient strategy to streamline computer-aided catalyst discovery. The code, documentation, and examples are openly available at https://github.com/lcmd-epfl/NaviCatGA.

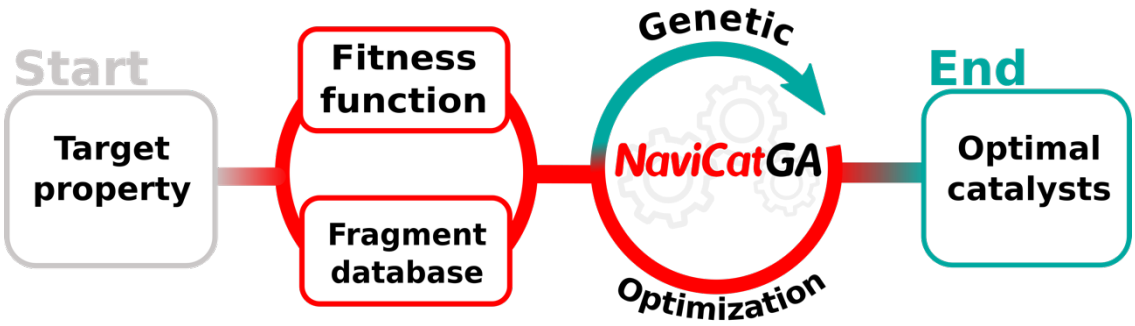


Figure 5.1 Schematic catalyst optimization pipeline powered by NaviCatGA.

5.2 Computational Methods

5.2.1 Overview of the NaviCatGA package

NaviCatGA is a lightweight genetic algorithm package that offers a simple, versatile and scalable solution to catalyst optimization problems. Simplicity is given by its Python structure and small number of dependencies, facilitating its adaptation and modification with minimal coding skills. Versatility comes from its modular design, which allows the user to define the optimization problem with utmost flexibility. For scalability, NaviCatGA relies upon the main strengths of genetic algorithms: the ability to tackle a large number of dimensions that are run in parallel. The genetic optimization loop is shown in Figure 5.2a.

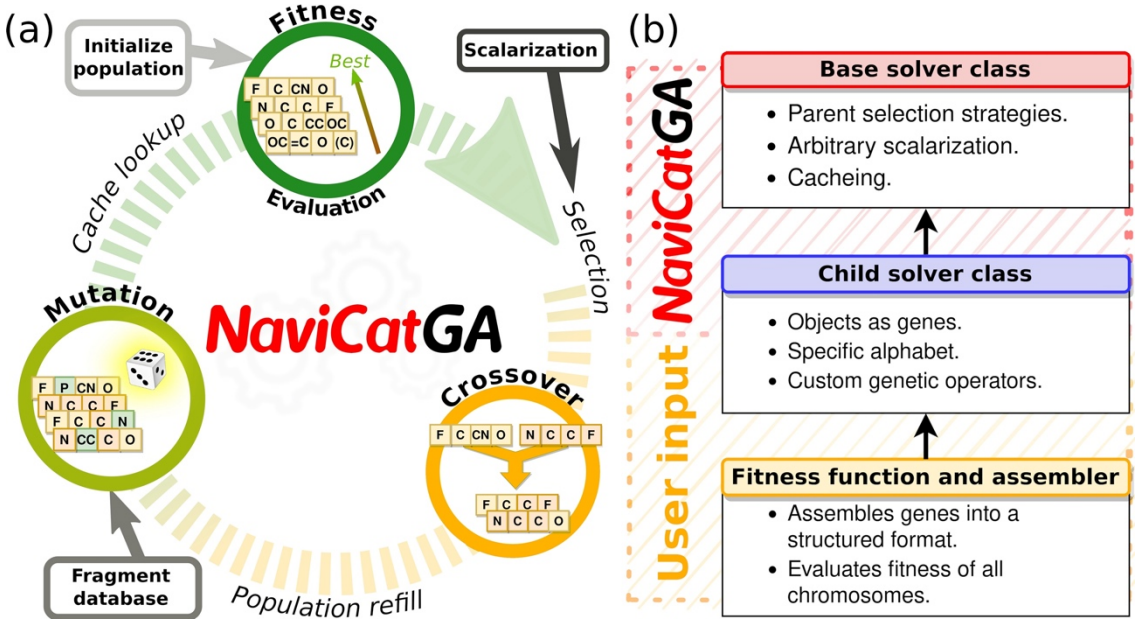


Figure 5.2 (a) Optimization loop followed by NaviCatGA. (b) Schematic representation of the user input and the functionalities implemented.

The three distinct levels in which the NaviCatGA package is structured are represented in Figure 5.2b: the base solver class with all core functionalities, the child solver class (several of which are provided) that defines the problem type (crossover and mutation), and the user input

(assembler and fitness function). This structure allows for significantly increased flexibility and adaptability, whereas adapting existing optimization tools could be difficult.[27]

5.2.2 Base Solver Class

The core genetic loop is provided by the GenAlgSolver base class (see Figure 5.2b). By design, the base class is data-type agnostic, with individuals represented by flexible lists of elements, and contains the solve method, which performs the optimization run (fitness evaluation, crossover, and mutation). Five different selection strategies to decide which individuals to recombine are provided (*i.e.*, two-by-two, roulette wheel, pairwise tournament, Boltzmann-weighted, and random). This choice regulates the greediness of the optimization by defining a number of individuals for cross-over. The number of selected individuals is limited to a percentage of the total population (*i.e.*, the selection rate). Additional features such as pruning of duplicates in each successive generation, a least-recently used cache of fitness evaluations, and *in situ* scalarization of fitness, are implemented (see Figure 5.2a–b for an overview). It is also possible to lock specific genes, so that they remain unchanged during the optimization procedure.

5.2.3 Implemented Solvers

The specificities of the optimization problem are imposed by a child class (Figure 5.2b), which defines the way mutation and cross-over are performed. Three child solver classes are provided: the SmilesGenAlgSolver, based on SMILES strings,[328] the SelfiesGenAlgSolver, based on SELFIES strings,[329,330] and the XYZGenAlgSolver, which uses AaronTools.py geometry objects,[79] representing a 3D molecular fragment. In these respective solvers, each gene has the corresponding data type. The SMILES and SELFIES solvers are suited for systems that can be readily represented as strings. On the other hand, the XYZ solver allows for detailed 3D control, as each gene contains a set of coordinates. As child classes define the data type of genes, they also contain all the possible values any given gene on an individual can take, which in NaviCatGA parlance is called an “alphabet” (Figure S1 in the Supporting Information). Genes

with the same alphabet are considered to be equivalent (*i.e.*, they can be replaced and mixed with one another).

In the implemented child solver classes, mutation is defined as substitution of a randomly chosen percentage of genes, or mutation rate, by random elements of the respective alphabets (Figure 5.2a). In turn, cross-over is achieved by combining the equivalent genes over one or more randomly determined crossover points (single-point cross-over is exemplified in Figure 5.2a but additional crossover operators could be considered in the future).

Defining new child solver classes is simple, as the core shared functionalities are kept in the base solver class. Different data structures, supported by other libraries (*e.g.*, Molassembler[97] or molSimplify[331]) could be used as alternative back-ends. Additionally, child classes can be inherited to incorporate additional definitions of mutation and crossover without substantial modifications.

5.2.4 Fragmentation Scheme

The fragmentation scheme and the corresponding alphabets define the total catalyst components combinatorial space to be explored. This step has two goals: avoiding the consideration of catalysts that are not expected to be stable and/or synthetically accessible,[332] and ensuring the domain of applicability and transferability of the fitness function (see below).

5.2.5 Assembler and Fitness Function

Once an appropriate catalyst space is defined through the fragmentation scheme, the user is required to input the fitness function and the assembler function into the solver (Figure 5.2). The assembler function takes a given individual (a list of genes of the specified data-type) and assembles them into a potential catalyst. In the case of SMILES, assembly can be as simple as concatenation of characters. In the XYZGenAlgSolver child class, the fragments must be suitably

assembled in 3D. The user is free to define any assembler function in order to generate more complex graph structures from the underlying chromosomes.

Finally, the fitness function takes as argument an individual as interpreted by the assembler function and returns a fitness value. By default, NaviCatGA attempts to maximize fitness, although internal scalarizers can be used to change the default (see Example 2 below for a complex demonstration of multi-objective optimization).

5.2.6 Choosing a Fitness Function

The choice of fitness function for catalyst optimization depends on the specific application. In a broad sense, NaviCatGA favors fitness functions that map a candidate catalyst's chemical structure to a measure of its performance in a given reaction.

Molecular volcano plots, which have been favored by us,[30] provide a way to connect a descriptor variable, typically the energy change associated with a step of the reaction mechanism (*x*-axis), to the overall catalytic performance (*y*-axis, expressed in terms of the energy span or TOF).[287] Some of us previously trained kernel-based ML models to predict the volcano descriptor variables for large pool of catalysts, from an approximate intermediate structure.[16,333] As demonstrated in Example 1, this inexpensive mapping between chemical structure and reactivity constitutes a natural fitness function that can be exploited for the GA optimization. An alternative approach to rapidly evaluate the catalytic properties and thus the fitness function consists in fitting Multivariate Linear Regression (MLR) expressions.[13] In Example 2, we fit and use MLR expressions to relate both the activity (*i.e.*, the volcano descriptor) and the selectivity, expressed in terms of $\Delta\Delta E^{\ddagger}$, to an intermediate structure. However, NaviCatGA imposes no constraint on the form of the fitness function and any alternative defined by the user is possible. In general, any ML-based model tailored for the prediction of catalytic properties constitutes a powerful alternative.[41,334]

In order to help users defining fitness functions and assemblers conveniently, a number of predefined wrapper functions are provided, built around RDKit[94] and pySCF.[335,336] Frequent

descriptors, such as frontier molecular orbital energies or molecular volumes, are provided through wrappers from multiple molecular formats, including SMILES. Coupling any of the solvers to production-level quantum chemical computations is equally possible. Thus, the set of wrappers allows users to define highly customized fitness functions with minimal coding effort.

5.3 Results and Discussion

Example 1: Exploration of Ligand Space for Ni-Catalyzed Aryl-Ether Cleavage

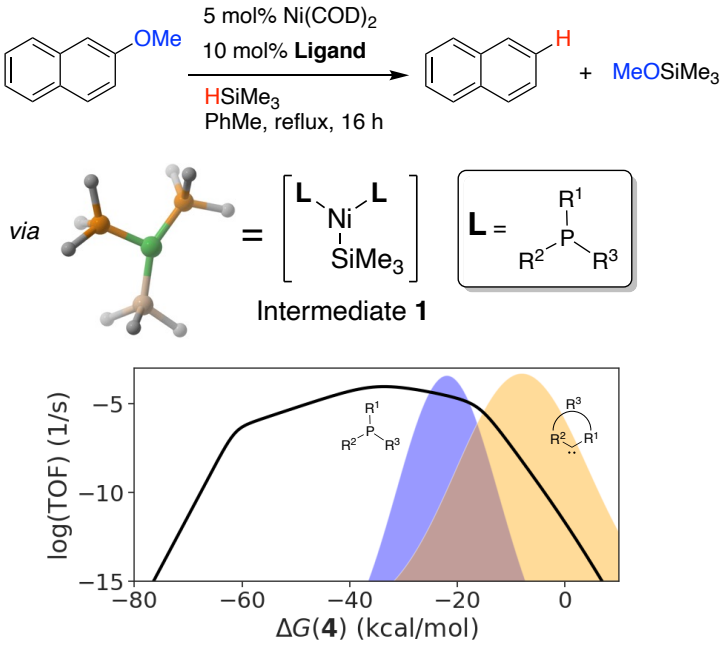


Figure 5.3 Reductive Ni–catalyzed cleavage of the 2-methoxynaphthalene C(sp^2)–O bond with trimethylsilane. The volcano plot predicts optimal catalytic activity at $\Delta G(\mathbf{4}) = -33$ kcal/mol. The blue and orange curves represent the approximate distribution of phosphine and carbene ligands, respectively (adapted from ref. 16).

One of us recently explored the ligand space for Ni-catalyzed aryl-ether reductive cleavage (Figure 5.3) relying upon a tandem volcano plot-ML approach to screen over 10^5 Ni catalysts bearing over 140 000 different phosphine or carbene ligands.[16] The volcano peak (maximum activity, see Figure 5.3) was found to correspond to $\Delta G(\mathbf{4}) = -33$ kcal/mol, where $\Delta G(\mathbf{4})$ is the free energy change associated with the formation of intermediate **4** (see Figure 5.3), used as descriptor variable. Interestingly, very few phosphine and carbene ligands lead to high turnover

frequencies, as they are spread in two gaussian distributions approximately centered on $\Delta G(\mathbf{4})$ = –20 kcal/mol (blue curve) and $\Delta G(\mathbf{4})$ = –5 kcal/mol (orange curve), respectively.

Based on the aforementioned exhaustive screening, we validate the capability of NaviCatGA by identifying the best phosphine ligands for the Ni catalyst with minimal computational cost. Additionally, we demonstrate how evolutionary experiments provide additional chemical insight and how they can be used to purge the pool of bad candidates from the database prior to further exploration. We finally demonstrate the versatility of the assembler function in exploiting the same procedure for the carbene ligands which, unlike phosphines, are composed of a backbone and two side groups.

Problem Definition

In this example, chromosomes are composed of three genes, accounting for the three different substituents in the phosphine ligands, all represented by SMILES strings using the SmilesGenAlgSolver class. The assembler is a function that generates the complete SMILES of intermediate **4** (Figure 5.3) from the chromosome information. The combinatorial space, which was taken from[16] (see it listed in the Supporting Information), comprises a set of 68 possible substituents for the phosphine ligands, as well as 77 ring and 30 backbone substituents for carbene ligands. Note that these numbers could further increase by including more exotic ligands or by decomposing the fragments into smaller components. Yet, this extension would potentially compromise both the experimental relevance of the generated intermediates **4**, a typical flaw of generative models, and the accuracy of our fitness function (see below).

Fitness Function

Following our previous work,[16] a kernel ridge regression model is trained to predict $\Delta G(\mathbf{4})$ from the approximate 3D structure of intermediate **4** using the same database of 1473 catalysts. The trained model has a cross-validated MAE of < 4 kcal/mol. Details of the ML model can be found in the Supporting Information. For prediction, the SMILES in the GA is embedded to 3D

coordinates using RDKit, then its SLATM representation[337] is obtained, which leads to its predicted $\Delta G(\mathbf{4})$ through the trained regression coefficients and kernel. For a candidate i, the final fitness score f_i is obtained by evaluating its $\Delta G(\mathbf{4})_i$ value compared to a normalized gaussian distribution centered on the target value x, $f_i = \exp\left(-\frac{1}{2}\left(\frac{\Delta G(\mathbf{4})_i - x}{\sigma}\right)^2\right)$ where $\sigma = \frac{|x|}{2}$.

Optimization

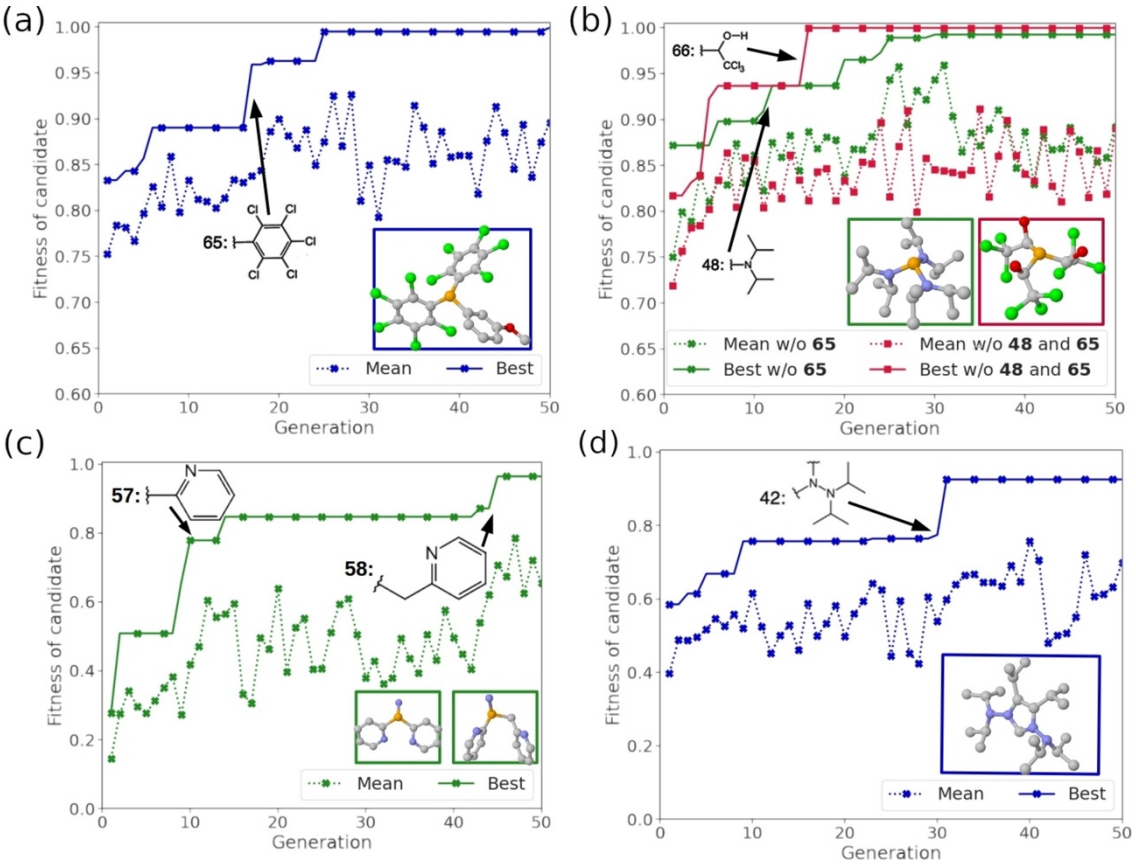


Figure 5.4 Evolution of mean population fitness and best candidate fitness over the optimization runs. Fitness is defined as $f_i = \exp\left(-\left(\frac{\Delta G(\mathbf{4})_i - x}{|x|}\right)^2\right)$. The most fit ligands from each run are highlighted in the corresponding boxes (H atoms omitted for clarity). (a) Complete run over the whole combinatorial space with x = –33 kcal/mol. (b) Ablation experiments in which fragments with x = –33 kcal/mol are removed from the combinatorial space; removal of 65 is represented with green lines, removal of both 48 and 65 is represented in red. (c) Complete run over the whole combinatorial space with x = –10 kcal/mol. (d) Complete run over the carbene combinatorial space with x = –33 kcal/mol.

The genetic optimization is initiated with a population of 10 randomized ligands (individuals) and a mutation rate of 10% for 50 generations. The maximum number of evaluations, 500, is

infinitesimal w.r.t. the combinatorial search space of 3×10^5 (68^3). The first run is set up with a target value of $x = -33$ kcal/ mol, the peak of the volcano (maximum activity). Results are shown in Figure 5.4a (blue curves and frame). The GA is able to identify top candidates, lying exactly on the volcano top, within the first 30 iterations – under 300 total evaluations. The top candidate contains a bis(pentachlorophenyl)phosphine ligand, in agreement with our previous screening,[16] in which the pentachlorophenyl (65) substituent was identified as one of the best options. The overall increase in fitness coinciding with the selection of the pentachlorophenyl substituent by the optimizer occurs in generation 20, as illustrated by the sharp increase in the best fitness curve in Figure 5.4a. It is important to stress that the GA takes three orders of magnitude less evaluations than our previous screening approach to identify it.

Given the low computational cost of the run, ablation evolutionary experiments are performed to obtain additional insight and explore different possible local fitness maxima. First, the pentachlorophenyl (65) substituent is removed from the database and the optimization is run again. This run leads to the identification of isopropylamino (48) as a good substituent, shown in Figure 5.4b as the green curve and frame, again in agreement with our previous work. Removing the aforementioned substituent and re-running leads to an increasingly difficult start for the optimization run, as less good options are available, but nevertheless ultimately identifying the 2,2,2-trichloro-1-hydroxyethyl substituent (66) as a good candidate in less than 20 iterations (Figure 5.4b, red curve and frame). Overall, the three best substituents that had previously been identified (diisopropylamino, pentachlorophenyl, and 2,2,2-trichloro-1-hydroxyethyl) are correctly and systematically located by NaviCatGA in less than 600 evaluations.

A similar optimization run is performed for a target of $\Delta G(\mathbf{4}) = -10$ kcal/mol. This value, which corresponds to the right-hand-side of the volcano plot, results in negligible catalytic activity. The GA identify ligands with a predicted $\Delta G(\mathbf{4})$ close to the targeted value, which leads to the identification of the least optimum substituents for the phosphine ligands, in this case N-

containing heterocycles (Figure 5.4c). Both good and poor candidates are identified with the same setup.

Finally, we optimize a N-heterocyclic carbene ligand using the same parameters with a target of $x = -33$ kcal/mol. The flexibility of NaviCatGA facilitates alternative definition of the fragment combinatorial space (in this case, the N-atom substituents, see the Supporting Information for details). The results, shown in Figure 5.4d, capture a key observation in line with previous work: unlike phosphine ligands, N-heterocyclic carbene ligands are generally unable to reach the top of the volcano. The optimization problem thus becomes harder as illustrated by the significantly lower fitness scores. Nevertheless, the genetic algorithm finds the best possible candidates within the combinatorial space, achieving a remarkably close value to the top using diisopropylamino substituents.[16] This optimization procedure provides a traceable evolution for every fit candidate and for the relative preference of the different substituents.

Example 2: Achieving the Activity/Selectivity Trade-Off with Enantioselective Organocatalysts

While Example 1 focuses on validation and comparison with high-throughput screening, this second example is chosen to illustrate the convenience of NaviCatGA to explore a large combinatorial space and optimize several properties simultaneously. Whenever several properties are to be optimized, there is often a trade-off between two or more targets preventing the existence of optimum solutions. In such cases, a large number of solutions to the optimization problem, the so-called Pareto front, can be identified depending on the criteria selected by a decision maker.

In catalyst design, a classic example of multi-objective optimization is the activity *versus* selectivity conundrum, where increased activity of a catalyst generally leads to decreased selectivity. A good catalyst should be both as active and as selective as possible. A pragmatic way to decide over this particular Pareto front is to search specifically for catalysts that retain

noticeable activity while prioritizing selectivity, as opposed to compromising selectivity for increased activity, or reducing activity to a negligible level in search of perfect selectivity.

Given the flexible structure of NaviCatGA, the user imposes selected criteria on the optimization problem by assigning weights to different properties (*e.g.*, the final fitness is defined 60% by selectivity and 40% by activity), or using step functions to define hard boundaries (*e.g.*, give a fitness of 0 whenever selectivity drops under some value). However, translating human criteria into mathematical functions is difficult. NaviCatGA thus supports fitness functions that return several values, which are then processed by a scalarizer to derive the final, singular fitness value. Although any internal scalarizer object can be used, we recommend the achievement scalarizing function Chimera[338] to process multi-objective fitness functions within the optimization run. Chimera requires a priority ranking and a degradation threshold to be assigned for each optimization objective and generates a score for each candidate by assessing its relative performance in the population (for further details, we refer the reader to the original publication[338]). Chimera's versatility matches NaviCatGA's and allows for the effortless formalization of complex human criteria.

To demonstrate conflicting multi-objective optimization, this example exploits a Chimera scalarizer to find optimal Lewis base organocatalysts for the enantioselective propargylation of benzaldehyde (Figure 5.5).[37,41,339,340]

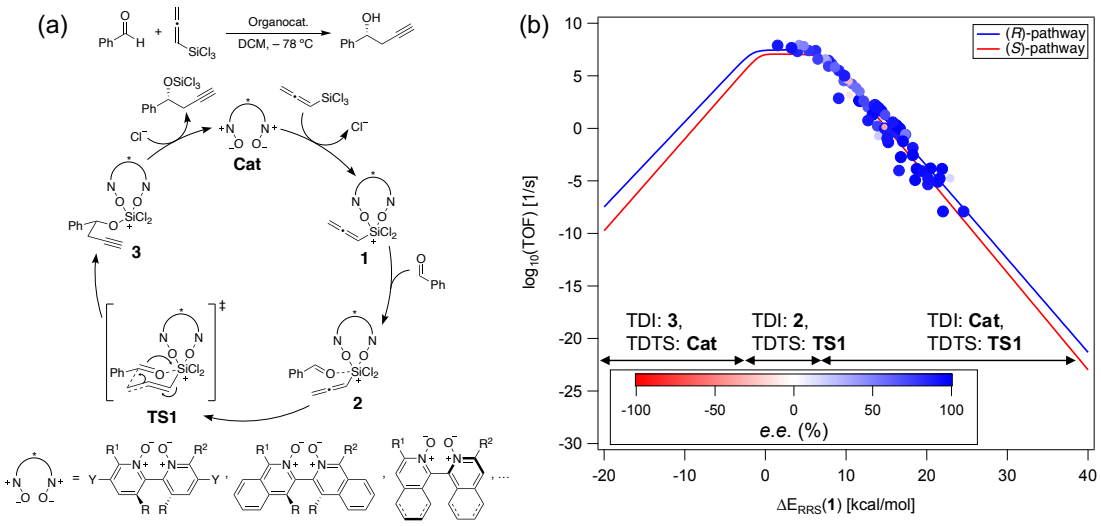


Figure 5.5 (a) Catalytic cycle for the bipyridine *N*,*N*′-dioxide-catalyzed enantioselective propargylation of benzaldehyde (R = H or Me).[37,340] (b) Enantioselectivity TOF-molecular volcano plot for the 78 test set organocatalysts depicted in Figure S12. Larger and darker blue spheres indicate catalysts with higher ee values favoring (*R*)-product formation, smaller and red spheres indicate catalysts favoring (*S*)-product formation. The different slopes of the volcano correspond to different TOF-determining intermediates (TDI) and transition states (TDTS).

Problem Definition

In this case, chromosomes are composed of three genes: a chiral scaffold (*e. g.*, the parent scaffold **S1** is a (*S*)-2,2-bipyridine *N*,*N*′-dioxide, **S2** and **S3** include additional Ph or ^{t}Bu substituents at the 6,6′-positions, **S4** is a (*S*)-8,8′-disubstituted 2,2-biquinoline *N*,*N*′-dioxide, *etc*.) and two substituents at the 6,6′-positions (see the Supporting Information for the full list of scaffolds **S1**-**S14**). The 3D coordinates of all substituents and scaffolds are obtained from DFT computations, and thus the XYZGenAlgSolver class is used. The assembler in this case is a function capable of building the 3D structure of intermediate **1** (Figure 5.5a) from a given chromosome by substituting the 3D structures of the two substituents in the 6,6′-positions of the scaffold (R^1 and R^2 in Figure 5.5a), with no re-optimization necessary. The combinatorial space is given by 14 *N*,*N*′-dioxide scaffolds and 34 different substituents (16 184 combinations, see Supporting Information for details). Note that, to increase the size of the combinatorial space, catalysts with different 6 and 6′ substituents, in addition to the more synthetically accessible symmetrically substituted ones, are considered.

Fitness Function

Based on previous work,[37,41] reference energies of intermediates **1**–**3** and of **TS1** are computed at the PCM(dichloromethane)/B97-D/TZV(2p,2d) level for 78 different organocatalysts using structures optimized at the same level of theory. Relative energies (*i.e.*, electronic energies plus solvation free energies) at this level were found to be more robust to reproduce experimental results for this reaction.[37,340] A volcano plot is constructed for the propargylation of benzaldehyde with allenyltrichlorosilane (Figure 5.5a), leading to the identification of the descriptor variable $\Delta E(\mathbf{1})$ and of the region of maximum activity ($\Delta E(\mathbf{1}) \approx 3$ kcal/mol). Enantioselectivity is calculated as a function of $\Delta\Delta E^{\ddagger}$, which is defined as the difference between the (*R*)- and (*S*)-Boltzmann-weighted activation energies of the **2** → **TS1** reaction step, relative to the lowest-lying (*R*)- or (*S*)-ligand arrangement of **2** (see Supporting Information for further details).

Two Multivariate Linear Regression (MLR) expressions are then parametrized to predict $\Delta E(\mathbf{1})$ and the $\Delta\Delta E^{\ddagger}$ from the unoptimized 3D structure of intermediate **1** assembled by the genetic algorithm (Figure 5.5), using as parameters five dihedral angles, the Sterimol B5 and L values of the 6,6′-substituents, and ε_{LUMO} (see Supporting Information for details). The parametrized MLR expressions lead to RMSE values of 1.65 kcal/mol and 0.25 kcal/mol for $\Delta E(\mathbf{1})$ and $\Delta\Delta E^{\ddagger}$, respectively. Details and cross-validation of the MLR models are given in the Supporting Information.

Using the two MLR models, activity is gauged by the proximity of $\Delta E(\mathbf{1}) \approx 3$ kcal/mol (plateau of maximum activity, see Figure 5.5b) and selectivity is defined as proportional to $\Delta\Delta E^{\ddagger}$. While the explicit MLR equations give a rough idea of the balance between different parameters, an explicit criterion has to be used to narrow down the Pareto front. Several options are explored (see below) to showcase the importance of proper multi-objective criteria.

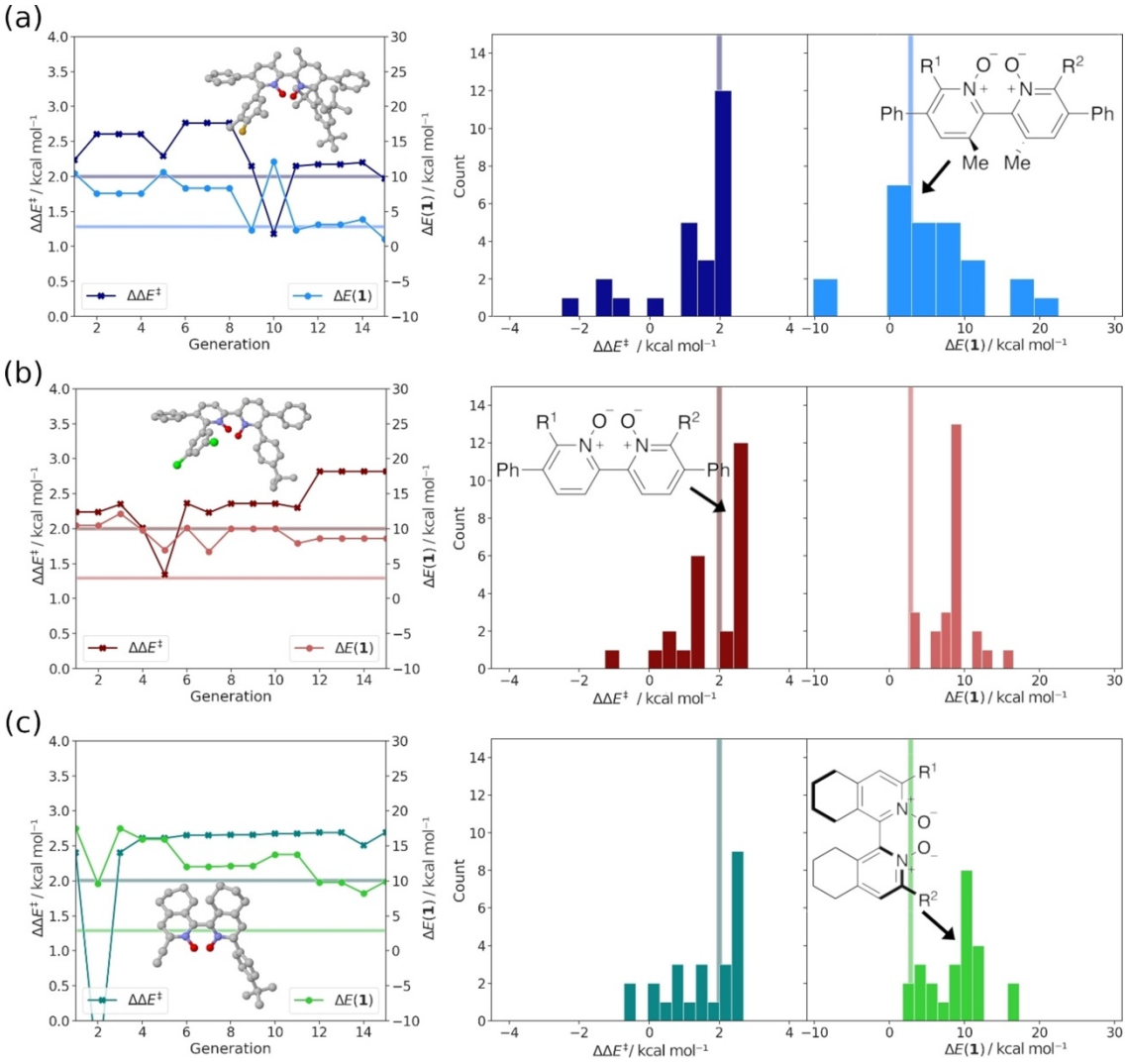


Figure 5.6 Evolution of maximum selectivity and activity over optimization runs with three different scalarization setups. The best catalyst candidate from each run is highlighted (H atoms omitted for clarity). Shaded lines indicate the optimal activity region of $\Delta E(\mathbf{1})$ (light hue) and a minimum $\Delta\Delta E^{\ddagger}$ threshold for guaranteed enantioselectivity (dark hue). The distribution of $\Delta\Delta E^{\ddagger}$ and $\Delta E(\mathbf{1})$ for the final populations of each run are shown right. (a) First setup with minimum $\Delta\Delta E^{\ddagger}$ = 1.5 kcal/mol and 25% compromise on minimizing $\Delta E(\mathbf{1})$. (b) Second setup with maximum $\Delta E(\mathbf{1})$ = 10 kcal/mol and 25% compromise on maximizing $\Delta\Delta E^{\ddagger}$. (c) Third setup with minimum $\Delta\Delta E^{\ddagger}$ = 2.5 kcal/mol and 50% compromise on minimizing $\Delta E(\mathbf{1})$.

Optimization

Three GA runs are started with an initial population of 25 randomized individuals, consisting of a bipyridine *N,N'*-dioxide scaffold and two R^1 and R^2 substituents each, a mutation rate of 5% and a selection rate of 25%. All optimizations are run for 15 generations leading to a maximum of 375 evaluations out of the > 10^4 combinatorial possibilities. The fitness functions are all based on the aforementioned MLR expressions but scalarized differently using Chimera: for the first

run, a minimum absolute $\Delta\Delta E^{\ddagger}$ = 1.5 kcal/mol is imposed while ΔE(**1**) is minimized with a 25% degradation threshold, due to the flatness of the activity plateau around ΔE(**1**) = 0 kcal/mol. This exemplifies a standard situation in which enantioselectivity is to be guaranteed and only subsequently activity has to be optimized. After the optimization procedure (Figure 5.6a), several good candidates are found with predicted $\Delta\Delta E^{\ddagger}$ of ≈ 2 kcal/mol and ΔE(**1**) of 1 kcal/mol, with the top candidate having the (*S*)-2,2′-bipyridine *N,N*′-dioxide scaffold with Ph substituents at the 5,5′-positions, R = Me, and R^1 = 3,5–Me–4–F–Ph and R^2 = 2,4,6–^{t}Bu–Ph. NaviCatGA, driven by the scalarizer, is able to explore activity and selectivity and find a good compromise between both. The distribution of values in the final population shows how it is enriched with high $\Delta\Delta E^{\ddagger}$ and low ΔE(**1**) candidates after 15 generations: a rightmost bump in the distribution of $\Delta\Delta E^{\ddagger}$ and a bump in the region between 0 and 10 in the distribution of ΔE(**1**) (Figure 5.6a).

For the second run, a maximum value of ΔE(**1**) = 10 kcal/ mol is imposed, while $\Delta\Delta E^{\ddagger}$ is maximized with a 25% degradation threshold. This represents the opposite setting, in which good activity is guaranteed (the estimated TOF for ΔE(**1**) = 10 is ≈ 50 000 s^{-1}, see Figure 5.6b) and selectivity comes as a second priority. By inverting the priorities, the optimization problem becomes noticeably more difficult. For the first 10 generations, the top candidate found with this setup is stuck at the ΔE(**1**) = 10 kcal/mol mark, having $\Delta\Delta E^{\ddagger}$ slightly over 2 kcal/mol (scaffold = (*S*)-1,1′-disubstituted 3,3′-biisoquinoline *N,N*′-dioxide, R = H, R^1 = I, R^2 = 4–^{t}Bu–Ph). In this case, the scalarizer setup leads to a very steep local optimum after a few exploratory generations, and evolution is hindered due to the relatively tight 25% degradation margin. The final population thus shows a very large percentage of nearly identical candidates. However, through mutation, the optimizer finds an optimal candidate with high selectivity and acceptable activity in the last four generations, depicted in Figure 5.6b. Here, the scaffold is (*S*)-2,2′-bipyridine *N,N*′-dioxide with 5,5′-Ph substituents, R = H, R^1 = 3,5–Cl–Ph, and R^2 = 4–^{t}Bu–Ph. Some common trends are evident comparing the top performer of this optimization run with the results from the previous one, particularly the presence of a ^{t}Bu-substituted phenyl group in the R^2 position and of halogen-containing groups as R^1, as well as the similar (*S*)-2,2′-bipyridine *N,N*′-dioxide scaffold. The

small change in scaffold (R = Me in the first run, R = H in the second one), which is associated to reduced activity in the second evolutionary experiment, exemplifies the difficulties associated with activity cliffs in catalyst design.

In the third run, we exemplify a more flexible setup requiring a minimum $\Delta\Delta E^{\ddagger}$ value of 2.5 kcal/mol while attempting to reach the top of the volcano as before, but accepting a 50% degradation of the latter to enforce the former, which provides much more flexibility than in the previous examples. In this case, the top candidates quickly present significant selectivity, but no compromise is achieved with respect to activity, and thus ΔE(**1**) is barely improved over the run and remains over 10 kcal/mol, in spite of the noticeable trade-off exploration in the early generations (with even a generation exploring structures that would lead to (*S*)-product formation in search of improved activity), which is afforded by the increased degradation margin. The final population of the run, shown in Figure 5.6c, excels in selectivity but is worse than the first two runs in terms of activity, with the distribution heavily centered around the 10 kcal/mol mark. The top candidate has a (*S*)-H_8-[1,1′-biisoquinoline] 2,2′-dioxide backbone with R^1 = CCH, R^2 = 4–^{t}Bu–Ph; this scaffold is shown to be associated with improved selectivity because of its dominating presence in the final population.

The comparison between the three runs highlights how the same optimization setup, guided by slightly different human input, ends up exploring very different areas of the combinatorial space and finds diverse solutions in the Pareto front. Hence, the use of scalarization and careful problem definition is recommended in order to navigate multi-objective optimization. For typical bipyridine *N,N*′-dioxide-derived organocatalysts, selectivity is believed to arise from favorable electrostatic interactions between the formyl C–H of benzaldehyde and the nearby Cl ligand in the lowest-lying transition state structure leading to the (*R*)-alcohol.[37] Activity is largely a function of the organocatalyst's Lewis basicity, with better electron-donors able to more efficiently activate the allenyltrichlorosilane substrate (and hence being located closer to the volcano plateau), while catalysts bearing strongly electron-withdrawing substituents are less active and found lower on the right slope of the volcano. The evolutionary experiments highlight

how changes in the scaffold and in the nature of R^1 and R^2 affect this selectivity-activity interplay and reveal the unique role played by aromatic substituents. Ph groups at the 6 or 6′-position with electron-donating alkyl substituents are clearly important for enhanced activity, although additional ^{t}Bu substituents (at the *ortho*-positions) cause unfavorable steric interactions with the formyl C–H, overwhelming the stabilizing effect from favorable C–H...Cl interactions and hence reducing selectivity (this is the case of the first run, Figure 5.6a). When placed at the 5,5′-positions, the Ph groups lead to additional π-stacking interactions favoring the (*R*)-pathway (benzene trimer-like interactions involving benzaldehyde and two Ph substituents)[341] and offsetting otherwise unfavorable π-stacking and CH/π interactions that stabilize the (*S*)-pathway.[37,342] Thus, in the second run (Figure 5.6b), the presence of less electron-rich substituents (including hydrogen atoms instead of methyl groups at the R position) results in a slight loss of activity, but ensures favorable noncovalent interactions that yield very high selectivity. In line with recent experimental results,[342] the presence of aliphatic substituents (instead of aromatic ones) is associated with reduced activity (as in the third run, Figure 5.6c), however the ethynyl group as R^1 helps improve selectivity, since it leads to a more favorable electrostatic environment for the formyl C–H in the (*R*)-pathway (partially positively charged C–H interacting with the π-bonds in CCH).[37]

5.4 Conclusions

We presented NaviCatGA, a tool capable of optimizing the structure of homogeneous catalysts to find top candidates with tailored properties for a given reaction. Using evolutionary techniques, it is possible to perform the optimization task with the possibility of tracing the origin of favorable catalyst components (*e.g.*, ligand substituents, catalyst scaffolds or side groups) during the evolutionary experiments and pinpoint their influence on different aspects of a catalyst's performance (*e.g.*, activity, selectivity).

From a technical perspective, NaviCatGA is versatile, flexible and thus applicable to a variety of catalytic problems. Thanks to its hierarchical structure, it is compatible with diverse structural

representations (*e.g.*, SMILES, 3D structures), genetic operations and fitness functions. Additional functionalities, including ML-based acceleration,[150,343–348] can also be conveniently deployed for fitness evaluation. While NaviCatGA, as presented here, is a core component of inverse design efforts in catalysis, it also constitutes a powerful stand-alone program for general optimization problems. In order to further streamline the inverse design workflow, it is desirable to automate the elucidation of the fitness function as well as of other eventual quantum chemical tasks. Within this context, NaviCatGA is integrated into the broader NaviCat platform (https://github.com/lcmd-epfl/NaviCat), collecting an ensemble of tools for computational catalysis. This set of utility tools, which include, for instance, automated construction of volcano plots (https://github.com/lcmd-epfl/volcanic), can be used independently and/or in combination with each other. Overall, these efforts represent a complementary addition to alternative programs such as those addressing automated mechanistic studies[79,349–352] and structure generation.[78,97,331]

5.5 Author Contributions

S.G., R.L., and C.C. conceptualized the project. R.L. designed and coded NaviCatGA and implemented it in Example 1 and 2. S.G. performed computations, curated the data, and analyzed results of Example 2. S.G. and R.L. wrote the manuscript with help and feedback from C.C., who provided supervision throughout.

5.6 Supporting Information

The Supporting Information for this Chapter may be found at https://chemistry-europe.onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Fcmtd.202100107&file=cmtd202100107-sup-0001-misc_information.pdf

6 Reaction-Based Machine Learning Representations for Predicting the Enantioselectivity of Organocatalysts

This chapter is based on following publication:
Gallarati S., Fabregat R., Laplaza R., Bhattacharjee S., Wodrich M. D. and Corminboeuf C., Reaction-Based Machine Learning Representations for Predicting the Enantioselectivity of Organocatalysts. *Chem. Sci.* **2021**, *12*, 6879.

6.1 Introduction

The bottleneck of closed-loop reaction optimization pipelines (including genetic algorithms)[36,38] is evaluating a catalyst candidate's fitness, which is typically exemplified in terms of catalytic activity and/or stereo/regio/chemoselectivity. Obtaining such measures experimentally is time- and resource-intensive, and only tractable with robotized high-throughput experimentation methods and self-driving laboratories.[353,354] Selectivity prediction based on high-level quantum chemical methods is complex, even for simple molecules. That is because *e.e.* (enantiomeric excess) values, estimated as the ratio between the competitive reaction rates leading to the two enantiomeric products,[22] are computationally expensive and challenging to predict accurately. The energy difference between the transition states (TSs) leading to the major and minor enantiomers can be quite small (< 2 kcal mol^{-1}) and multiple diastereomeric transition states, stemming from the large conformational space of flexible catalysts, can yield the same enantiomer.[7,83] While the intrinsic error of the quantum chemical level is often addressed in comprehensive benchmark studies,[8,22,340,355,356] automated toolkits,[27,79] such as AARON[78] and CatVS,[85] have been developed to streamline the tedious and error-prone task of optimizing hundreds of thermodynamically accessible stereocontrolling transition states. Although such

accelerated prediction of selectivity is enticing for the prospect of computational catalyst design,[23] the applicability of QM-based tools such as AARON remains limited either by the cost of quantum mechanical computations, which quickly becomes prohibitive, or by the inherent difficulty of locating all transition state structures. On the other hand, tools using QM-derived molecular mechanics force fields (Q2MM), like CatVS, require the development of an MM force field for each new reaction type considered, a major limitation to their widespread application.[83]

To accelerate fitness evaluation, statistical models are used to predict catalyst's selectivity. They may be trained using either experimental or computational data.[17,334,357–359] The first approach is often limited by the small size of the experimental datasets available and by their inherent noise,[129] while the second suffers from the difficulties associated with reproducing difficult-to-compute targets *e.g.*, *e.e.* values.[360] Sigman and co-workers have popularized the use of multivariate linear regression to fit experimental reaction outcomes to physical organic molecular descriptors.[13,28,361] However, such models, which depend on DFT computations of relatively expensive properties (*e.g.*, vibrational frequencies and intensities, polarizabilities)[362] are not adapted to the purpose of fast (*e.g.*, GA)[36,38] optimization for which bypassing the DFT bottleneck is key. Similarly, nonlinear regression models (*e.g.*, support-vector machine,[68,135] random forest,[67,280,363–366] neural networks,[63,347,367–370] Gaussian process regression[371–374]) often use system-specific and expensive features (*e.g.*, physical organic descriptors like Charton or Sterimol values, NBO charges, NMR chemical shifts, bond distances and angles, HOMO–LUMO gaps, local electro/nucleophilicity) as input.[375–377]

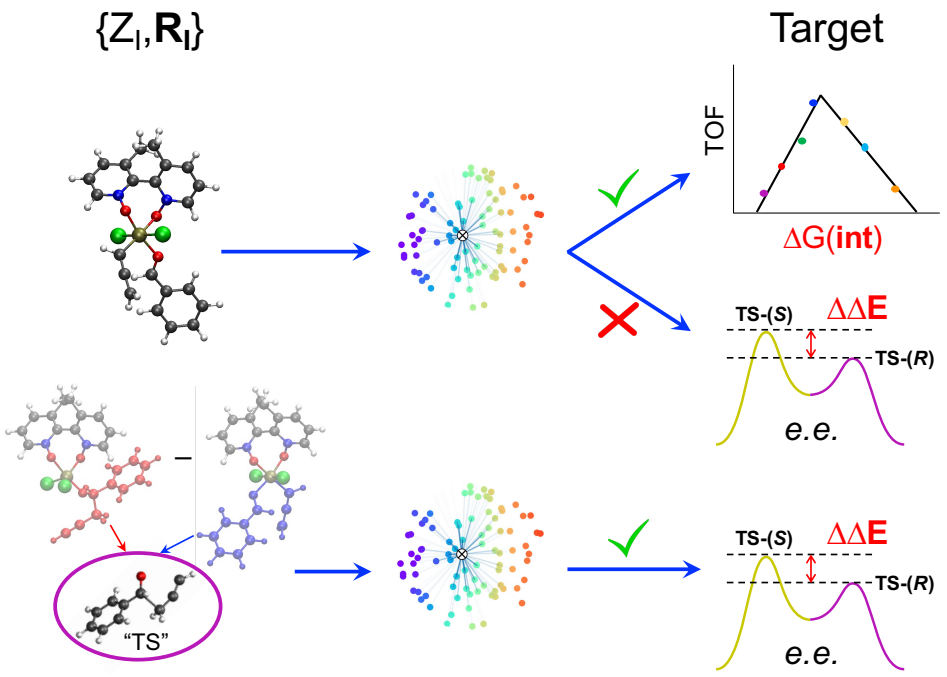
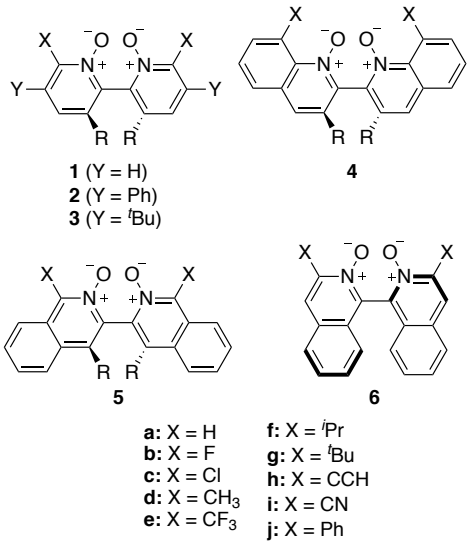


Figure 6.1 Schematic illustration of the importance of reaction-based ML representations for challenging targets. (Top) Standard QML representations, built using one structure (*e.g.*, that of a catalytic cycle intermediate), are successful at regressing "simple" targets, such as thermodynamic quantities (*e.g.*, the volcano plot descriptor), but struggle with reaction properties (*e.g.*, the enantioselectivity). (Bottom) A "reaction-based" representation, built as the difference between the representations of two structures (*e.g.*, two sequential catalytic cycle intermediates), is a more faithful fingerprint of the TS geometry, and thus is successful at predicting enantioselectivity.

The so-called quantum (or atomistic) ML models, which map a molecular representations *e.g.* CM,[378] SLATM,[337] SOAP[379] obtained from a set of 3D atomic coordinates, to a representative target (typically) computed with quantum chemistry, constitute an appealing complementary strategy owing to its broad applicability and dependence on the laws of physics.[337,380,381] Since accurate geometries are not necessarily needed as input,[16,333] their use is compatible with fast closed-loop optimization.[38] While these approaches provide a favorable combination of efficiency, scalability, accuracy, and transferability for predicting energetic and more complex molecular properties,[380] identifying enantioselective catalysts requires precise predictions of the relative energy barriers for the stereocontrolling transition states, a target currently beyond their accuracy (Figure 6.1).

Here, we provide a stepwise route to improve such QML approaches to reach sufficient accuracy for subtle properties such as those associated with asymmetric catalysis (*i.e.*, *e.e.*). This objective is achieved by rationally designing a reaction-based representation (Figure 6.1) that is a more faithful fingerprint of the enantiodetermining TS energy. The performance of the approach is demonstrated through accurately predicting the DFT-computed enantiomeric excess of Lewis base-catalyzed propargylation reactions directly from the structure of the catalytic cycle intermediates. Unlike other ML models trained on (absolute) experimental *e.e.*'s,[121,122] our model is able to predict the absolute configuration of the excess product, because it is trained on the activation energy of the enantiodetermining step for each pair of enantiomers (pro-(*R*) and pro-(*S*) intermediates) independently.



Scheme 6.5 Library of axially chiral bipyridine *N,N'*-dioxide organocatalysts. R = H or Me. Adapted from ref. 37.

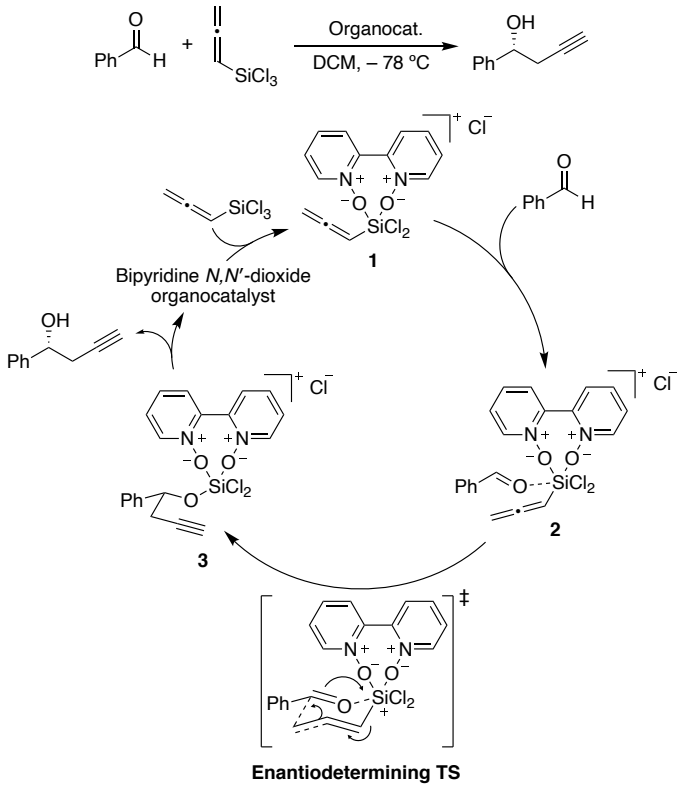
6.2 Methods

6.2.1 Reaction and Organocatalysts Database

Asymmetric allylations[382–385] and propargylations[339] of aromatic aldehydes are key C–C bond forming transformations, providing access to optically enriched homoallylic and homopropargylic alcohols, respectively, which serve as valuable building blocks for the synthesis of complex chiral molecules.[386] Catalysts that are selective for allylations are generally not highly

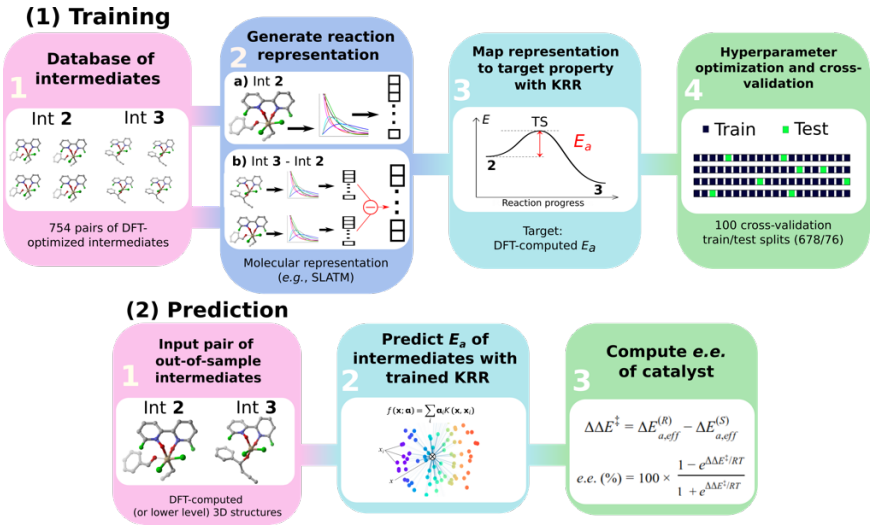
stereoselective for propargylations, which has led to a dearth of stereoselective propargylation catalysts.[52,387–390] Tools to screen dozens of allylation catalysts to find promising candidates for propargylation reactions are therefore highly valuable.[23] To this end, Wheeler and co-workers have investigated 76 Lewis base organocatalysts (Scheme 6.5)[37] and used the computational toolkit AARON[78] to build a database of 760 stereocontrolling transition states to predict their enantioselectivity in the propargylation of benzaldehyde (Scheme 6.6).[37,340,391] Large databases of kinetic data for asymmetric catalysis generated *in silico* are scarce.[17] Therefore, this library constitutes an ideal training and validation set for the development of an atomistic ML model with reaction-based representations capable of predicting the *e.e.* of organocatalysts readily from the structures of intermediates. Note that the workflow presented below would improve the ML performance independently of the size of the training data. The target of the ML model is the DFT-computed relative forward activation energy (E_a, *i.e.*, the energy difference between the TS and the preceding intermediate) associated with each of the 10 (*R*)- or (*S*)-ligand arrangements (see Figure S1) of the enantiodetermining TS in Scheme 6.6 for the 76 catalysts in Scheme 6.5 (11 catalysts of type **1**, 16 of type **2**, 15 **3**, 11 **4**, 13 **5**, and 10 catalysts of type **6**), yielding a total

of 754 E_a values.[392] *e.e.* values are computed from E_a (*vide infra*), thus accurate predictions of E_a lead to accurate *e.e.* predictions.



Scheme 6.6 Catalytic cycle for the propargylation of benzaldehyde with allenyltrichlorosilane, showing the rate-limiting and stereocontrolling transition state. Adapted from ref. 52.

6.2.2 General ML Workflow



Scheme 6.7 Graphical overview of the workflow used to build an atomistic ML model for *e.e.* prediction.

The general workflow exploited and improved herein relies on a physics-based ML model for the prediction of the *e.e.* of the asymmetric catalytic reactions, as illustrated in Scheme 6.7 and described hereafter. It comprises two parts: part (1) is a training procedure that relies on the following steps:

(1) Database construction: a library of 3D geometries and energies of catalytic cycle intermediates is curated. Here, the structures of 754 pairs of intermediates **2** and **3** are optimized with DFT (see the next section) and used to train the ML model. As shown in our previous work,[333] accurate geometries are not necessarily needed as inputs for atomistic ML models; thus, rough-coordinate estimates (*e.g.*, obtained directly from SMILES strings) or low-cost xTB structures could potentially be used to generate suitable molecular representations.

(2) Generation of molecular representations: information intrinsically contained within the 3D structure of each intermediate is transformed into a suitable molecular representation. Here we build different variants based on the Spectral London and Axilrod-Teller-Muto (SLATM)[337] representation. SLATM is composed of two- and three-body potentials, which are derived from

the atomic coordinates, and contain most of the relevant information to predict molecular properties.[379,393–399]

(3) Training of the model: input representations are mapped onto the corresponding target values (E_a, computed at the DFT level, see the next section) using Kernel Ridge Regression (KRR)[400] with a Gaussian kernel. Note that even if target values based on DFT are used here to train the ML model, the strategy proposed hereafter is expected to perform equally well on experimental or more accurate quantum chemical data.

(4) Hyperparameter optimization and cross-validation: the full dataset is split randomly 100 times into 90/10 training/test sets (678/76 datapoints) to optimize the KRR hyperparameters and obtain the learning curves.

In part (2), the trained ML model is used to predict the activation energy of out-of-sample organocatalysts. The model requires as input the 3D structures of **2** and **3** and delivers the corresponding E_a value. Using the energy of **2** as reference, the relative energies of the enantiodetermining (*R*)- and (*S*)-TSs can be calculated, and the *e.e.* of the catalyst under investigation computed (*vide infra*).

6.3 Computational Details

6.3.1 Quantum Chemistry

Catalytic cycle intermediates **2** and **3** were optimized at the B97-D/TZV(2p,2d) level of theory,[250,251,401] accounting for solvent effects (dichloromethane, ε = 8.93) using the polarizable continuum model (PCM)[291,402,403] with Gaussian16,[249,404] in analogy with the study by Wheeler and co-workers.[37] Density fitting techniques were used throughout. The structures of 1508 intermediates were obtained *via* intrinsic reaction coordinate calculations (IRC)[292] from the TS database curated by Wheeler *et al.*[37] 754 target E_a values (11 catalysts of type **1**, 16 type **2**, 15 **3**, 11 **4**, 13 **5**, and 10 of type **6**, each in 5 distinct pro-(*R*) and pro-(*S*) ligand arrangements)[392] were computed (relative to the lowest-lying intermediate **2** ligand arrangement) at the same level,

which was shown to provide the best compromise between accurate predictions of low-lying TS energies and stereoselectivities for allylation and propargylation reactions.[340] *e.e.* values were not predicted from Gibbs free energy barriers, but rather from relative energy barriers (*i.e.*, electronic energies plus solvation free energies), since they have been found to be more reliable than those based on either relative enthalpies or free energy barriers for this reaction.[340] The symbol E_a was therefore used to indicate the energy (electronic plus solvation) difference between the TS and the preceding intermediate. For each C_2-symmetric catalyst (Scheme 6.5), 10 distinct ligand arrangements around a hexacoordinate Si center are possible (**BP1**–**5**, (*R*)- and (*S*)-, Figure S1).[52,390,391] Since each of these can lead to thermodynamically accessible reaction pathways, and the stereoselectivity is largely a consequence of which ligand arrangement is low-lying for a particular catalyst, all diastereomeric TSs were considered viable and the *e.e.* calculated from a Boltzmann weighting of the relative energy barriers.[37] In equations **1**–**3**, $\Delta E_{a,eff}$ is the relative Boltzmann-weighted activation energy of each (*R*)- or (*S*)-species, $\Delta\Delta E^{\ddagger}$ is the difference between the (*R*)- and (*S*)-Boltzmann-weighted activation energies, R is the ideal gas constant, and T is the propargylation reaction temperature (195 K).

$$\Delta E_{\mathrm{a,eff}} = -\,\mathrm{RT}\ln\left(\sum_{\mathbf{i}}^{\mathbf{BPi}} \mathrm{e}^{-\left(E_{\mathrm{a}}^{\mathbf{BPi}}/\mathrm{RT}\right)}\right) \quad (\mathbf{1})$$

$$\Delta\Delta E^{\ddagger} = \Delta E_{\mathrm{a,eff}}^{(R)} - \Delta E_{\mathrm{a,eff}}^{(S)} \quad (\mathbf{2})$$

$$\textit{e.e.}\ (\%) = 100 \times \left(1 - e^{\Delta\Delta E^{\ddagger}/RT}\right) / \left(1 + e^{\Delta\Delta E^{\ddagger}/RT}\right) \quad (\mathbf{3})$$

6.3.2 Machine Learning

The Python package QML[405] was used to construct standard SLATM representations. Feature selection and the construction of the reaction-based representations $\mathrm{SLATM_{DIFF}}$ and $\mathrm{SLATM_{DIFF+}}$ were done using the Python package Scikit-learn.[406] To generate the learning curves and the *e.e.* predictions, a cross-validation scheme was used with 100 different 90/10 training/test sets (678/76). The KRR hyperparameters (the width of the Gaussian kernel σ and the ridge

regularization λ) were optimized for each train/test split, systematically obtaining essentially the same results for each split (see the SI). From the 100 train/test splits, the E_a of each intermediate pair (**2** and **3**) was predicted approximately 10 times; these test predictions were then averaged to obtain one final prediction. The standard deviation from the ~10 test predictions were used to generate the error bars. The final average prediction of the E_a value was used to calculate the Boltzmann-weighted $\Delta E_{a,eff}$ values (eq. **1**) and the $\Delta\Delta E^{\ddagger}$ of each (*R*)- and (*S*)-pair (eq. **2**), and so the *e.e.* value of each organocatalyst (eq. **3**). The out-of-sample predictions were done with the same $SLATM_{DIFF+}$ models trained in the cross-validation scheme. Additionally, out-of-sample predictions were done re-training the model on the entire dataset (see Figure S6), although this did not lead to noticeable improvement. While simpler representations (*e.g.*, CM,[378] BoB[407]) were tested, SLATM performs significantly better (see Figure S2).

6.4 Results and Discussion

6.4.1 Molecular Representations

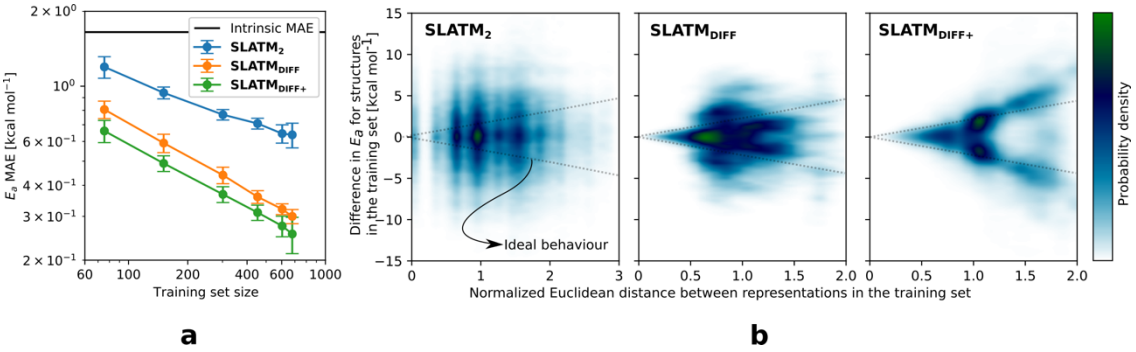


Figure 6.2 (a) Learning curves with MAE in test sets predictions of E_a for the three approaches discussed. The error bars correspond to the standard deviations and are computed from the results of 100 different random train/test splits. (b) Dissimilarity plots *i.e.*, difference in target values (E_a) *vs*. Euclidean distance between representations for each pair of points in the dataset (the Euclidean distances have been divided by the average distance between points). When the difference in E_a values tends to zero, the corresponding points should lie in the area delimited by the two dotted straight lines (ideal behavior).

The key step of the workflow presented above is generating a molecular representation, which is mapped onto the target value (*i.e.*, the activation energy E_a) and used as a fingerprint of the enantiodetermining TS. Representations can be constructed from single molecules and more

recently as "ensemble representations": instead of associating one fixed configuration of atoms to a single-point geometry energetic target value, information from multiple structures can be combined to generate a representation for an ensemble property, such as the free energy of solvation (ΔG_{sol}).[408] This has recently been achieved by calculating the ensemble average of the FCHL19 representations[208,409] of a set of configurational snapshots obtained through MD sampling.[408] Here, we propose an alternative approach that goes beyond standard QML representations (*i.e.*, KRR using one given gas-phase geometry)[408] by describing the chemical transformation occurring during the enantiodetermining step of an asymmetric reaction through the comparison of the representations of the two catalytic cycle intermediates preceding and following the stereocontrolling TS. This allows us to generate a "reaction-based" representation, which can be closely mapped to the activation energy of the enantiodetermining step, as discussed later. We rely on "dissimilarity" plots as a diagnostic tool to determine whether a particular representation can adequately characterize the stereocontrolling step. By dissimilarity plots, we refer to histograms of the Euclidean distance between any two representations *vs*. the difference in their target property, which in this case is E_a. For a particular representation to be effective, small distances between structures must correspond to small differences between target properties, as the Euclidean distance is used to measure the similarity of two molecular representations. Similar plots have previously been exploited to analyze the behavior of molecular representations,[379,410] but only parenthetically. Here we highlight their importance as a fundamental analytical tool to understand the performance of molecular representations in

kernel methods for asymmetric catalysis and demonstrate their utility for constructing reliable ML models.

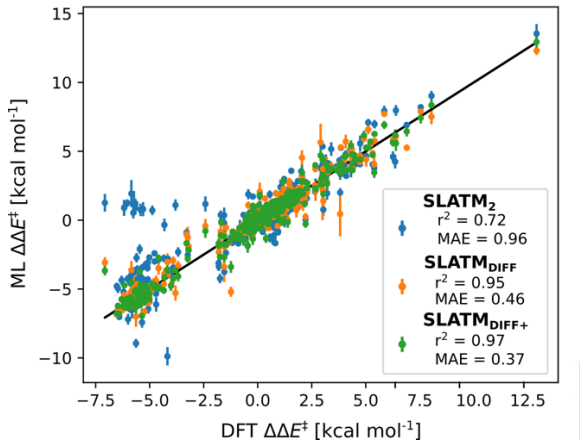


Figure 6.3 Predictions of $\Delta\Delta E^{\ddagger}$ *vs.* DFT reference for the three approaches discussed. Mean Absolute Errors (MAE) are reported in kcal mol^{-1}. These predictions are obtained by averaging the predictions obtained from the cross-validation scheme with 100 different random train/test splits. The error bars indicate the standard deviation of ML $\Delta\Delta E^{\ddagger}$, derived from the standard deviations in the E_a prediction of the 100 different random train/test splits.

Before discussing our proposed representation variants, we report in Figure 6.2a the performance of the standard SLATM representation using the structure of a single intermediate (*e.g.*, **2**). Due to the structural similarities between **2** and the enantiodetermining TS (in both, the Si atom has 6 coordination sites occupied, whereas only the coordination number is only 5 or 4 in intermediate **3**), intermediate **2** was first chosen to construct the input representation. The learning curve for the prediction of E_a using SLATM (blue) of intermediate **2** (denoted SLATM$_2$) reaches a Mean Absolute Error (MAE) of 0.54 ± 0.06 kcal mol^{-1} for the prediction of E_a with 90% of the data used for training (*i.e.*, 680 structures). Considering the exponential relationship between relative activation energies and *e.e.* values, which implies a dramatic propagation of errors, the accuracy of this approach is insufficient. This is further demonstrated in Figure 6.3, which shows the correlation between the predicted and reference $\Delta\Delta E^{\ddagger}$ values (MAE = 0.96 kcal mol^{-1}), and in Figure 6.4, where the *e.e.* values obtained from SLATM$_2$ are compared to the reference quantities: the large number of red-colored cells indicates large deviations between ML-predicted and DFT-computed *e.e.* values. The rather poor mapping between SLATM$_2$ and

the E_a of the stereocontrolling step (associated with the key **2** → **3** transition state) is evident from the visual inspection of Figure 6.4, where the large number of red-colored cells associated with catalysts bearing substituents **a**, **d**, **e**, **g**, **f** and **j** indicates inaccurate predictions of *e.e.* values, and from the analysis of the corresponding dissimilarity plot in Figure 6.2b (left). In the latter, the large scattering of points lying outside the area delimited by the dotted lines, particularly when the Euclidean distance tends to zero, means that two different structures might be considered equal by the kernel (distance ≈ 0) albeit leading to very different E_a values. Thus, the shape of the dissimilarity plot of SLATM$_2$ deviates considerably from ideal one, indicated by the dotted straight lines.[379] Note that the MAE for E_a increases up to 0.77 ± 0.05 kcal mol^{-1} (see Figure S2) if starting from the SLATM representation of **3**, the intermediate following the enantiodetermining step in the catalytic cycle (Scheme 6.6). The higher accuracy achieved using the representation of **2** *vs*. **3** could be attributed to the reaction step being exergonic and, according to the Hammond Postulate,[411] the enantiodetermining TS resembling **2** more closely.

In any case, neither the structure of **2** or **3** provide sufficiently good fingerprints of E_a on their own.

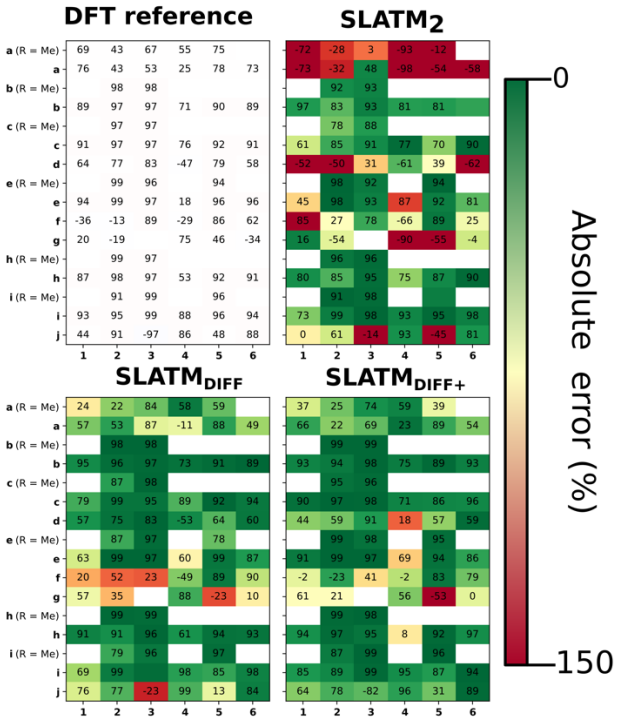


Figure 6.4 *e.e.* values obtained from DFT computations (top left) and from the ML predictions of E_a using the three approaches discussed. These predictions are obtained by averaging the predictions obtained from the cross-validation scheme with 100 different random train/test splits. Cells are colored according to their accuracy with respect to the reference, ranging from dark green (best) to dark red (worst). Positive *e.e.* values correspond to excess (*R*)-alcohol formation, negative values to excess (*S*)-alcohol formation.

Unlike other intrinsic molecular properties that depend on the structure of a single molecule,[408] enantioselectivity is determined by electronic and/or steric effects stabilizing or destabilizing one enantiomeric TS to a greater or lesser degree than the other. In that sense, it is to be expected that our target accuracy for E_a, well below 1 kcal mol^{-1}, cannot be reached using only one structure that does not adequately describe the stereocontrolling transition state as an input. To improve the model performance, an alternative representation is constructed by comparing the representations of both intermediates. Knowing that neither the structure of **2** or **3** are uniquely related to the corresponding activation energies, we can generate such a "reaction-based" representation that draws information from both structures, subtracting the global SLATM of **2**

from **3**. This is reminiscent of binary reaction fingerprints (obtained by subtracting the products' from reactants' RDKit[94] fingerprints), which reflect changes in molecular features over reaction processes.[359] The resulting representation (denoted $SLATM_{DIFF}$) accounts for the differences between the two intermediates and is thus more sensitive to the structural changes occurring during the enantiodetermining step. By subtracting "reactant" from "product", the global features that do not change during the catalytic cycle step are eliminated from the representation, and the structural changes between intermediates are highlighted. In this way, we obtain a more faithful representation of the reaction step, which corresponds to a more unique fingerprint of E_a. Although the construction of $SLATM_{DIFF}$ requires the SLATM representations of both intermediates (**2** and **3**), the computational cost associated with its generation is negligible.

As depicted in the dissimilarity plot (Figure 6.2b, middle), the reaction-based representation ($SLATM_{DIFF}$) is significantly better than $SLATM_2$: the difference in E_a values tends to zero as the Euclidean distance between their representations tends to zero. In line with this observation, the learning curve (shown by the orange line in Figure 6.2a) is significantly improved. The MAE of $SLATM_{DIFF}$ is reduced to 0.31 ± 0.2 kcal mol^{-1}, roughly 50% better than $SLATM_2$ and up to 60% better than that of $SLATM_3$ using 90% of the data for training (*i.e.*, 680 structures) in the train/test splits of the cross-validation scheme. Given the rationality of the approach leading to the construction of $SLATM_{DIFF}$, its gain in accuracy is encouraging. As shown in Figure 6.3 and Figure 6.4, the halved MAE leads to a very notable improvement in the prediction of *e.e.* values. Nevertheless, we note again that very small errors in E_a are amplified when *e.e.* values are calculated, and therefore even a small accuracy gain can be significant.

The high probability density of normalized Euclidean distances between 0.5 and 0.75 seen in Figure 6.2b (middle, $SLATM_{DIFF}$) indicates that the shape adopted by the dissimilarity histogram of $SLATM_{DIFF}$ is not yet ideal, and that further improvement is possible. To achieve higher accuracy, we focus on improving the shape of this dissimilarity plot. Notice that in our ML model, the Euclidean distance is used as a measure of similarity between representations. This means

that features with high variance (*i.e.*, that change the most between molecules) dominate the notion of similarity, as they contribute the most to the Euclidean distance between representations. By feature, we mean each of the terms in the molecular representation, which, for SLATM, consist of two- (London dispersion) and three- (Axilrod-Teller-Muto) body potentials computed on groups of atoms closer than a certain cut-off (here, 4.8 Å). The results of these potentials are averaged over their atom-type sets (*e.g.*, all C–C interactions for the two-body terms, all the C–C–C for the three-body terms), which are then concatenated to generate the SLATM vector. The size of the SLATM representation depends on the existing atom-type sets in the database. Given that our dataset contains the elements C, H, O, N, F, Cl and Si, the total number of features of the SLATM representations is 27 827.

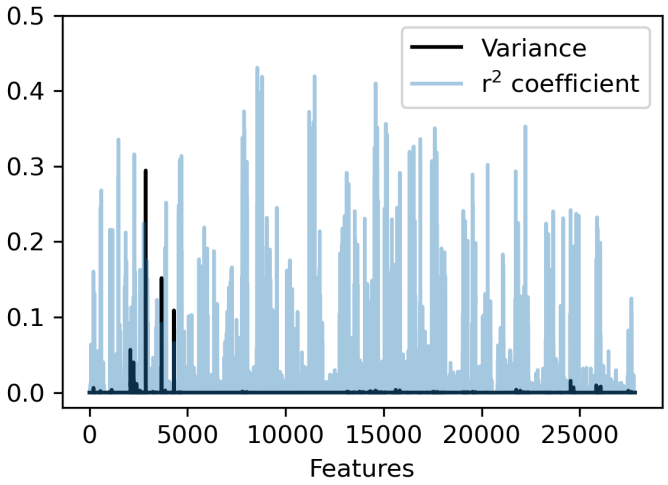


Figure 6.5 Variance and correlation coefficient with the target value for each of the 27 827 features of the $SLATM_{DIFF}$ representation in the dataset.

In $SLATM_{DIFF}$, features with high variance dominate the notion of similarity, measured through the Euclidean distance. However, we are using SLATM to predict a property that is very different from the single-molecule properties for which it was originally designed. Consequently, features with high variance in SLATM are not necessarily the most important fingerprints of E_a. In pursuit of the best possible fingerprint of the activation energy, we assign importance scores to each feature and attempt to focus on the most relevant ones. The linear correlation coefficient (r^2) between each feature and the target property is used as an estimate of the importance of the

different terms in the representation. The results, presented in Figure 6.5, show that in $SLATM_{DIFF}$ there are only a few high-variance features, while the computed importance scores are spread over many other features that have relatively small variances. Simply put, the variances in the features of the $SLATM_{DIFF}$ representation are not well correlated with their real importance in this application.

Based on this observation, an improved representation, labelled $SLATM_{DIFF+}$, is generated by selecting only the N_f most important features of $SLATM_{DIFF}$ (specifically, N_f = 500) and discarding the rest. This feature selection was done using only the training data at each train/test split of the cross-validation step, as otherwise it could lead to severe overfitting. Nevertheless, the importance scores were consistent across the cross-validation splits thanks to the robustness of the linear regressions. An improved relationship between representation and target distances (Figure 6.2b, right) is obtained with the $SLATM_{DIFF+}$ representation, in spite of its reduced size. This simple feature selection leads to a noticeable improvement in accuracy, with a cross-validated MAE of 0.25 ± 0.4 kcal mol^{-1} (see the green curve in Figure 6.2a). Using the $SLATM_{DIFF+}$ representation, the resulting cross-validated correlation coefficients for the difference between (*R*)- and (*S*)-activation energies ($\Delta\Delta E^{\ddagger}$, Figure 6.3) in the test set are greatly improved ($r^2 > 0.95$). The quality of our fitted model far supersedes previously reported approaches. Good qualitative and even quantitative agreement is achieved between predicted and reference *e.e.* values computed using the test data splits from the cross-validation runs (Figure 6.4).

Since linear correlation constitutes a very limited notion of relevance, other methods, such as nonlinear mutual information criteria,[412] were tested as feature importance estimators, but the resulting models showed similar or even worse performance (see the SI). Similarly, methods based on metric learning[410,413] did not lead to any improvement, as the high dimensionality of the problem led to severe overfitting. Ceriotti *et al.*[414] suggested the use of principal covariates regression (PCovR) to solve similar issues. PCovR is a supervised feature selection method that

interpolates between principal component analysis (PCA) and linear regression. Herein, because the variance of the features is completely unrelated to the importance scores, the addition of PCA would not offer any advantage. Nevertheless, these findings highlight the importance of adapting molecular representations to the application at hand, while still preserving the overall generality of the approach.

6.4.2 Chemical Insight on Asymmetric Propargylation Catalysts

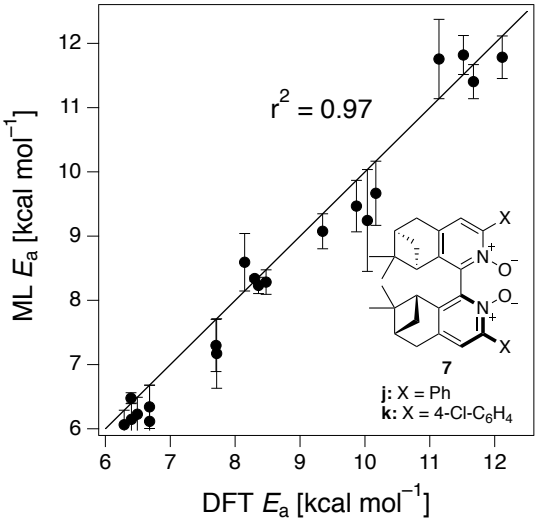


Figure 6.6 Out-of-sample predictions on terpene-derived atropisomeric organocatalysts **7j** and **7k**. 10 distinct TSs were computed for each catalyst (**BP1**–**5**, (*R*)- and (*S*)-). The error bars are the standard deviation of the 100 predictions from each trained model from the cross-validation scheme.

The ML model is able to reproduce the main trends in *e.e.* observed across the different catalysts from the 100 different random train/test splits (Figure 6.4, top left table). For example, using $SLATM_{DIFF+}$ (Figure 6.4, bottom right table), which gives the best predictions with respect to the reference data, catalysts built on scaffold **4** (Scheme 6.5) are revealed to be outliers, yielding *e.e.*'s that are significantly different to those obtained with other scaffolds, for a given substituent **a**–**j**. This is due to the different placement of the substituent X on the organocatalysts' scaffold. Excluding **4**, the effect of different substituents on the *e.e.* is qualitatively the same across all scaffolds, with the exception of **f** (^{i}Pr) and **j** (Ph). The introduction of a phenyl group on the organocatalysts' scaffold leads to highly varied *e.e.* values, from -97 (*S*) to 91 (*R*). This variation,

which is due to the presence of favorable π-stacking and CH/π interactions stabilizing some (*S*)-TSs and degrading the enantioselectivity,[37] is nicely captured by $SLATM_{DIFF+}$. Overall, the high enantioselectivity displayed by (most) catalysts in the library can be attributed to the favorable electrostatic interaction between the formyl C–H of benzaldehyde and one of the chlorines bound to Si, which is present in the lowest-lying (*R*)-ligand arrangement, and absent in the (*S*)-structures.[37]

In their computational screening with AARON,[37] Wheeler and co-workers identified derivatives of **6** as promising candidates for propargylation reactions. However, these catalysts are difficult to synthesize stereoselectively.[387,415] Recently, Malkov *et al*. reported the synthesis of a set of terpene-derived atropisomeric bipyridine *N,N'*-dioxides **7** (Figure 6.6) as easily-separated diastereoisomers.[342] Aromatically-substituted catalysts **7j** and **7k** were shown to be highly active and selective (*e.e*. of 96 and 97, respectively); additionally, the TS structures for **7** were computationally shown to be nearly identical to the corresponding substituted forms of **6**.[342] Prompted by these results, we decided to test the ML model with $SLATM_{DIFF+}$ to predict the activation energy of the 10 distinct ligand arrangements afforded by **7j** and **7k**. The out-of-sample results are shown in Figure 6.6. Despite scaffold **7** and substituent **k** not being in the original training set, excellent correlation between predicted and reference E_a values is obtained (r^2 = 0.97). Thus, the enantioselectivity of these out-of-sample catalysts is qualitatively reproduced, despite not achieving exact quantitative agreement between DFT and ML predicted $\Delta\Delta E^{\ddagger}$ values (1.2 and 1.3 for **7j** and **7k**, respectively, *vs*. 0.2 and 0.5 kcal mol^{-1}).

In summary, we provide a logical route to improve atomistic ML methods for enantioselectivity prediction of asymmetric catalytic reactions, which are limited by both the required accuracy and the small amount of data generally available. Firstly, the intermediates associated with the enantiodetermining step (**2** and **3** in Scheme 6.6) must be identified, and their SLATM representations generated. Secondly, using the difference between the two SLATM representations ($SLATM_{DIFF}$) as input, a set of features that map the activation energy accurately

can be obtained. Finally, feature engineering can be used to improve $SLATM_{DIFF}$, keeping only the most relevant features that relate to the target property. The results show that the ML workflow presented herein is able to accurately predict enantioselectivity from the molecular structures of catalytic cycle intermediates.

6.5 Conclusions

In this work, we have developed an atomistic machine learning model to predict the DFT-computed *e.e.* of Lewis base-catalyzed propargylation reactions (Scheme 6.6). The use of dissimilarity plots allowed us to rationally develop and progressively improve a reaction-based representation that can be adequately mapped onto the activation energy of the stereocontrolling step. We identified two fundamental limitations of many standard physics-based molecular representations for subtle catalytic properties. First, we have shown that neither the structure of the preceding nor that of the following catalytic cycle intermediate is a fine fingerprint of the energy of the stereocontrolling transition state. This issue can be circumvented by using a reaction-based molecular representation derived from both structures. Finally, we have demonstrated how feature selection can be used to fine-tune this representation.

The resulting model can accurately predict the DFT-computed enantioselectivity of asymmetric propargylations from the structure of catalytic cycle intermediates. Thus, it constitutes a valuable tool to quickly identify potentially selective propargylation organocatalysts. By design, the model is well-balanced between computational cost, generality and accuracy. It is easy to implement for a wide region of chemical space and seamlessly compatible with experimental (*e.g.*, X-ray structures of stable intermediates) and computational data alike. Our results prove that semi-quantitative predictions of *e.e.* values in asymmetric catalysis can be achieved by accurately predicting E_a.

We conclude that atomistic ML models with adequately tailored molecular representations can be a practical and accurate alternative to both traditional quantum chemical computations of relative rate constants and multivariate linear regression with physical organic molecular

descriptors. The stepwise improvement to the model described in this work opens the door to more complex reaction-based and catalytic cycle-based representations. Indeed, ensemble representations, which were recently introduced for properties very sensitive to conformational freedom, such as the free energy of solvation ΔG_{sol},[108] are a promising path to go beyond the single structure-to-property paradigm and allow for further generalization, once combined with the approach discussed herein. Such methodologies will be explored in future work for the accurate screening of enantioselective catalysts in asymmetric reactions.

6.6 Author Contributions

S.G. and R.F. contributed equally to this work. S.G. performed DFT computations and analyzed the results. R.F. trained and improved the ML models. S.G. and R.F. jointly wrote the manuscript with help from R.L. MD. W. and R.L. provided feedback on the DFT and ML components, respectively. S.B. and M.D.W. ran preliminary computations initiating this work. All authors discussed the results and commented on the manuscript. C.C. conceived the project with M.D.W., provided supervision and wrote the final version of the manuscript.

6.7 Supporting Information

The Supporting Information for this Chapter may be found at

https://www.rsc.org/suppdata/d1/sc/d1sc00482d/d1sc00482d1.pdf

7 Optimizing Generality in Asymmetric Organocatalysis with Evolutionary Experiments

This chapter is based on following publication:
Gallarati S., van Gerwen P., Laplaza R., Brey L., and Corminboeuf C., Optimizing Generality in Asymmetric Organocatalysis with Evolutionary Experiments. **2023**, *in preparation*.

7.1 Introduction

Developing catalytic methods that are tolerant to many functional groups exerting different steric and electronic influences on the reaction center without significant reduction in yield or product selectivity of is a long-standing goal of organic chemistry. Despite being a highly desired feature, such "generality" *i.e.*, breadth of substrate scope,[164] is rare and only few transformations become routinely incorporated into the synthetic chemist's toolbox.[416,417] This is due to reaction development usually beginning with the examination of a simple, readily available model substrate, with subsequent re-optimization on more complex systems guided by empirical trial-and-error.[418] Discovering more general conditions requires evaluating wider regions of chemical space derived from a large matrix of diverse catalysts crossed with a panel of substrates that effectively represent the whole target molecule class (Figure 7.1). Today, "one-pot-multisubstrate" screening[419–421] is tractable with high-throughput experimentation techniques,[42,137,273,422] but has found limited applicability due to issues associated with chemical compatibility and product analysis. Perhaps worse, the most general conditions or catalysts might be excluded from the original screening set, biasing the results.[423]

In the last decade, data-driven computational methods, in tandem with supervised and unsupervised machine learning algorithms, have been applied to address numerous challenges in organic chemistry,[424–426] such as prediction of reaction outcomes,[427–429] multistep synthetic planning,[430–432] and catalyst discovery.[16,132,133,150,433] In particular, Bayesian optimization[371,434] has been combined with robotic experimentation to find general conditions for heteroaryl Suzuki-Miyaura coupling.[435] More recently, Reid *et al*. have proposed a workflow for assigning and predicting generality through clustering of reaction sets, but manually curated literature databases and a user-defined success value were required.[134] Overall, existing data-driven tools are still aimed at accelerating the evaluation of a pre-defined set of catalysts/conditions,[436] rather than suggesting entirely new species exhibiting high performance across the whole substrate scope.

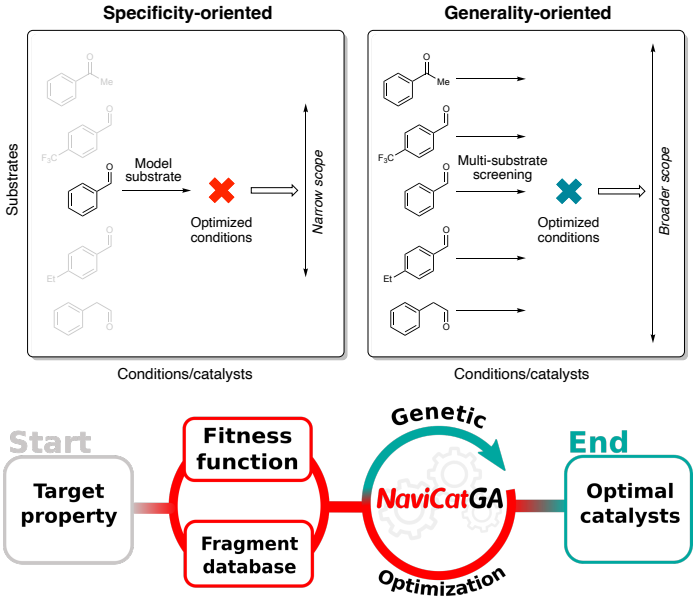


Figure 7.1 (Top) Reaction optimization tactics for the development of catalytic methods: traditional specificity-oriented approach *vs*. data-driven multi-substrate screening. (Bottom) Schematic inverse design pipeline powered by NaviCatGA.

Generative models[18] are an attractive alternative to direct screening by enabling the inverse design of functional molecules and materials.[35,437] In this paradigm, the desired functionality is first defined, and chemical structures tailored to that property are suggested (Figure 7.1). Although applications of generative models, such as genetic algorithms,[36] to homogeneous catalysis are increasingly being reported,[19,138,160,321,438,439] only specificity-oriented catalyst design

has been addressed. Optimizing generality as primary target requires adapting existing tools and pipelines to tackle this multi-dimensional problem.

Here, we show how evolutionary experiments performed with a genetic algorithm, NaviCatGA,[38] are designed to simultaneously probe the catalyst and substrate space and find organocatalysts predicted to exhibit both high turnover and enantioselectivity. We discuss the nature of fitness function used to estimate how close candidate species are to achieving optimal performance, the surrogate models that accelerate fitness evaluation, the database of molecular building blocks to generate millions of prospective catalysts on-the-fly, and the strategy followed to choose an unbiased and diverse substrate scope. We select the Pictet–Spengler cyclization as a synthetically relevant case study to illustrate how multi-objective genetic optimization across an expansive substrate space affords organocatalysts with good *average* activity and selectivity, while simultaneously providing information rich data on the areas of chemical space where even the best candidates are under-performing. Analysis of the challenging substrates gives insights into the structural features that limit generality, validating evolutionary experiments as a means to extract structure–activity–selectivity relationships and probe the existence of "privileged" catalysts.

7.2 Methods: The NaviCatGA Components

Below, we describe the components of the NaviCatGA pipeline for performing genetic optimization (Figure 7.1), and discuss the results of the evolutionary experiments, along with the chemical conclusions, in the Results and Discussion section.

7.2.1 Target property and reaction database

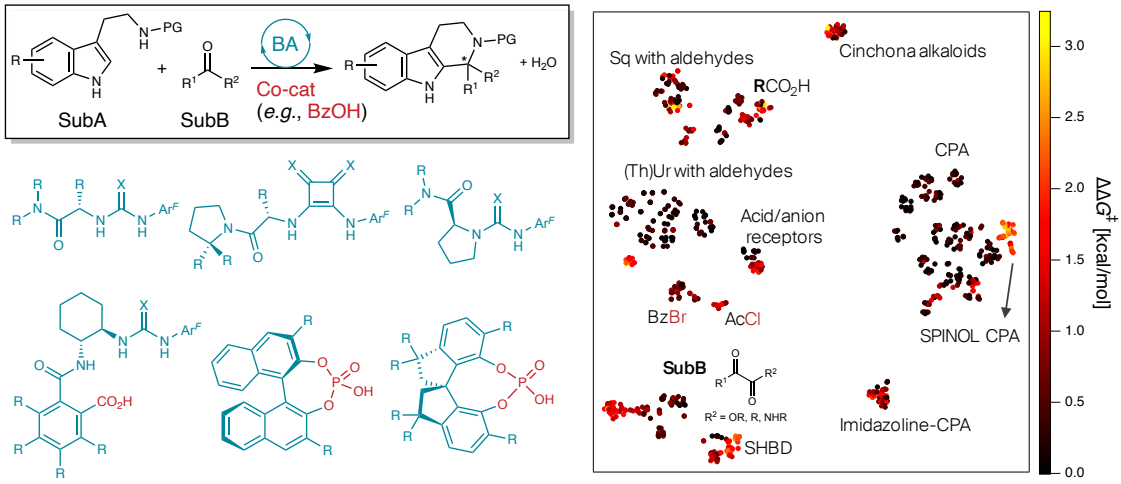


Figure 7.2 (Left) Pictet–Spengler cyclization of tryptamine derivatives (SubA, PG = protecting group, H, or OH) and carbonyls (SubB) in the presence of Brønsted acid organocatalysts and weak acid co-catalysts. Examples of hydrogen-bond donors, acid/anion receptor catalysts, and chiral phosphoric acids are shown. Ar^F = 3,5-CF_3-C_6H_3. (Right) 2D t-SNE map[207] of the reaction space on the basis of the concatenated Morgan fingerprints of the substrates and catalysts with experimental selectivity ($\Delta\Delta G^{\ddagger}$).

Herein, we define "generality" *i.e.*, the property targeted in the inverse design pipeline, as high enantioselectivity *and* activity across a wide and diverse substrate scope. Inspired by recent work by Jacobsen *et al.*,[42] we investigate the asymmetric Pictet–Spengler reaction[440–442] of tryptamine derivatives and carbonyl compounds (Figure 7.2), one of the most important methods for the synthesis of privileged pharmacophores such as tetrahydro-*β*-carbolines, due to the diversity of catalyst chemotypes capable of inducing high enantioselectivity. Although dozens of systems have been reported,[443] employing a variety of Brønsted acids such as chiral phosphoric acids (CPAs)[178] or single-[444] and dual-hydrogen-bond donors (S-/D-HBD)[445] used cooperatively with weak acids or bearing an acidic functional group internally,[446] no method has found widespread application, since each study was focused on a limited number of substrates. This reaction thus constitutes an ideal case study to develop an optimization strategy with generality as primary target.

At the onset of our investigation, we curated a database of 820 Pictet–Spengler condensations from the literature.[42,444,447–461] For simplicity, we constrain ourselves to protected or unprotected

tryptamines (as shown in Figure 7.2), excluding isotryptamines,[462] aryl ethanols,[463,464] phenethylamines,[465] and other substrates involved in more complex cascade reactions.[466–473] The database contains 240 unique transformations (*i.e.*, tetrahydro-*β*-carboline products) of 33 SubA and 164 SubB (aldehydes, ketones, *α*-ketoacids/esters/amides, and *α*-diones), catalyzed by 160 different organocatalysts and 30 co-catalysts (carboxylic acids, acyl and benzoyl chlorides and bromides). It is visualized in Figure 7.2 with a 2D t-SNE map[207] based on the concatenated Morgan FingerPrints[474,475] (MFPs) of the catalyst, co-catalyst, and substrates, where each point representing a reaction is colored according to its selectivity ($\Delta\Delta G^{\ddagger} = -RT\ln|e.r.|$, with *e.r.* being the experimentally measured enantiomeric ratio). The map is essentially divided into two regions, the right-hand side containing cyclizations catalyzed by CPAs, and the left-hand side those with single and dual-HBDs. Interestingly, despite "islands" of high enantioselectivity associated with catalysts being tested on a selected and limited class of carbonyl compounds (*e.g.*, SPINOL CPAs with aldehydes,[451] or SHBDs with ketoamides[444]), nearly 50% of the transformations display exceedingly low $\Delta\Delta G^{\ddagger}$ values (< 0.5 kcal/mol, and 70% < 1 kcal/mol). It is clear that choosing the conditions for carrying out an enantioselective Pictet–Spengler reaction on a never-before-tested substrate or estimating what the most general catalyst would be simply based on literature precedence is a nearly impossible task, further supporting the need for predictive and generative models.

7.2.2 Fitness function: evaluation of catalyst activity and selectivity

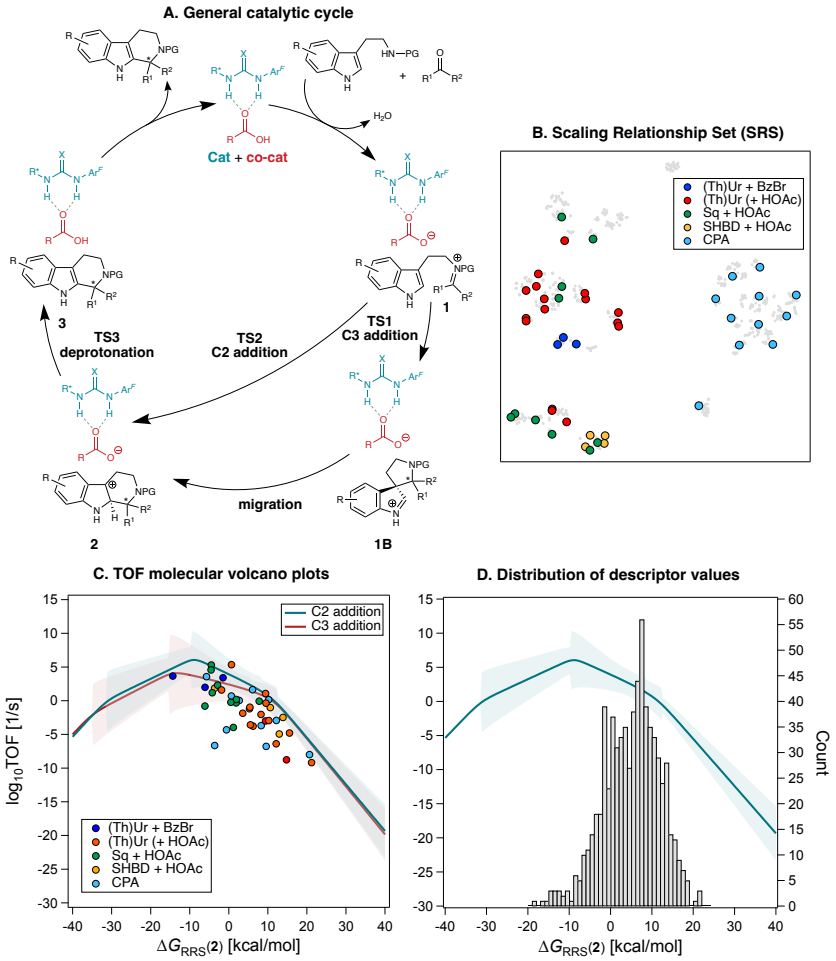


Figure 7.3 (A) General mechanism for the Pictet–Spengler reaction *via* anion-binding catalysis. (Thio)urea catalysts (X = O/S) are shown as an example. (B) The reactions used to construct a molecular volcano plot (SRS) are plotted on the t-SNE map, coloured according to the nature of the organocatalyst. (C) Molecular volcano plots based on the C2 and C3 addition mechanism. The shaded areas denote the 95% confidence interval based on the linear free energy scaling relationships. (D) Distribution of descriptor values and their location on the volcano plot.

The role of the fitness function in the inverse design pipeline (Figure 7.1) is evaluating how close a candidate organocatalyst is to achieving optimal performance. According to our definition of generality (*vide supra*), we are looking for species whose activity approaches the maximum theoretically achievable one. Molecular volcano plots[30] are therefore ideally suited for this task as they provide a way of connecting a descriptor variable, typically the energy change associated with a step in a catalytic cycle (*x*-axis), to the overall catalytic performance (*y*-axis, expressed in terms of energy span or TOF),[287,476] while simultaneously giving knowledge of the descriptor

value corresponding to the volcano peak or plateau (maximum performance *i.e.*, the target for genetic optimization).[38] Volcano plots are built from Linear Free Energy Scaling Relationships (LFESRs) that connect the value of the descriptor to the relative energies of the other cycle intermediates and transition states. While extensive details on the construction of these plots are given in ref. 32 and in the Computational Details, Figure 7.3A shows the mechanism of the Pictet–Spengler reaction,[477] whose knowledge is fundamental for building the volcanos. Following condensation of the *β*–arylethylamine (SubA) with the carbonyl compound (SubB) and formation of iminium ion **1**, nucleophilic attack by the aryl group and cyclization can occur either directly at position C2 of the indole *via* **TS2**, or at C3 *via* **TS1** to form the five-membered aza-spiroindolenine **1B**, which undergoes C–C migration to yield **2**. Deprotonation of **2** by the conjugate base of the acid co-catalyst, or of the CPA catalyst, is then necessary to form the tetrahydro-*β*-carboline product.

Constructing molecular volcanos requires computing the potential energy profiles of a medium-sized pool of sterically and electronically diverse systems.[32] 44 reactions from the Pictet–Spengler database are selected *via* farthest point sampling of the 2D t-SNE map. This Scaling Relationships Set (SRS, Figure 7.3B) comprises 39 unique transformations (*i.e.*, products) of 11 SubA and 31 SubB, catalyzed by 33 different Brønsted acids. Because the mechanism must be the same for all systems investigated, reactions catalyzed by cinchona alkaloid DHBDs (corresponding to upper cluster in the t-SNE map, Figure 7.2) are excluded, as these bifunctional organocatalysts have been shown to operate *via* a different mechanism.[453] In analogy with computational studies by Jacobsen *et al.*, who found no clear trend relating the benzoic acid electronic properties to the reaction rate,[477] the carboxylic acid co-catalyst, which sometimes contains large and bulky groups like triphenylmethyl, 9-anthracentyl, or 1-adamantyl,[456] is modelled with acetic acid to simplify the conformational complexity and reduce the computational cost of the system.

Using the SRS, TOF molecular volcanos for concerted C2 and stepwise C3 addition are constructed using the relative energy of intermediate **2** as descriptor (Figure 7.3C). Mechanistic

aspects of the Pictet–Spengler reaction, including the preferred pathway and the nature of the rate- and enantiodetermining step, have been a topic of intensive research:[478] Jacobsen *et al*. found a strong energetic preference for C2 over C3 addition in reactions catalyzed by chiral thioureas,[477] while You and co-workers showed that the spiroindolenine **1B** acts as either a productive or non-productive intermediate depending on the shape of the potential energy surface.[479] Evaluating the mechanism over a broad and diverse catalyst and substrate scope, as afforded by the SRS, reveals that, although the concerted pathways is generally preferred, the difference between the barriers for spiroindolization at C3 and electrophilic aromatic substitution at C2 is on average quite small (the volcanos are close to each other). Additionally, analysis of the LFESRs (see the SI) shows that there is often not one single rate- and enantiodetermining step, as rearomatization *via* deprotonation (**TS3**) and C–C bond formation (**TS1** or **TS2**) are almost isoenergetic: indeed, reactions are found for which **TS2** and **TS3** have similar degree of TOF-control.[301] The location of the SRS on the volcano plots indicates that cyclizations of hydroxylamines in the presence of benzoyl bromide co-catalyst (blue points),[457] as well as reactions of aldehydes catalyzed by squaramides (green points) display the highest TOFs. This observation is in line with the higher reactivity of ketonitrones[480] and the stronger H-bonding ability of squaramides, which has been found to correlate with faster turnover.[33] Conversely, the performance of CPAs and other DHBDs is strongly dependent on the nature of the substrates, as evinced by the bigger spread of TOF values. Among the poorest performing organocatalysts, sulfinamido urea derivatives[481] and carboxylic acids equipped with anion-recognition sites[452] are found lower on the volcano.

Having constructed the volcano plots and established the identity of the descriptor variable, we compute ΔG_{RRS}(**2**) for all the reactions in the Pictet–Spengler dataset (703 datapoints *i.e.*, excluding reactions catalyzed by cinchona alkaloids and those where only the carboxylic acid co-catalyst is varied). Structures are generated and optimized according to the pipeline described in the Computational Details. Figure 7.3D shows the Gaussian-type distribution of ΔG_{RRS}(**2**) superimposed on the TOF volcano for C2 addition, centered around 7 kcal/mol. Most Pictet–

Spengler reactions are found on the right slopes of the volcano (*i.e.*, weak-binding side), and their turnover is limited by iminium ion formation and deprotonation of the tetrahydro-*β*-carboline intermediate (or C–C bond formation). Overall, only few condensations have TOF close to the theoretical maximum. We then use this dataset to train a XGBoost machine learning model[482] to predict ΔG_{RRS}(**2**) using the concatenated Morgan fingerprints of the substrates, catalyst, and co-catalyst (acetic acid, BzBr, or none) as reaction representation (Figure 7.4A). A similar model is also trained on the whole Pictet–Spengler database (Figure 7.2 *i.e.*, 820 datapoints, using the real identity of the carboxylic acid co-catalysts rather than acetic acid) to predict the experimental $\Delta\Delta G^{\ddagger}$ values. Together, these models are used to accelerate fitness evaluation during genetic optimization (*vide infra*).[36]

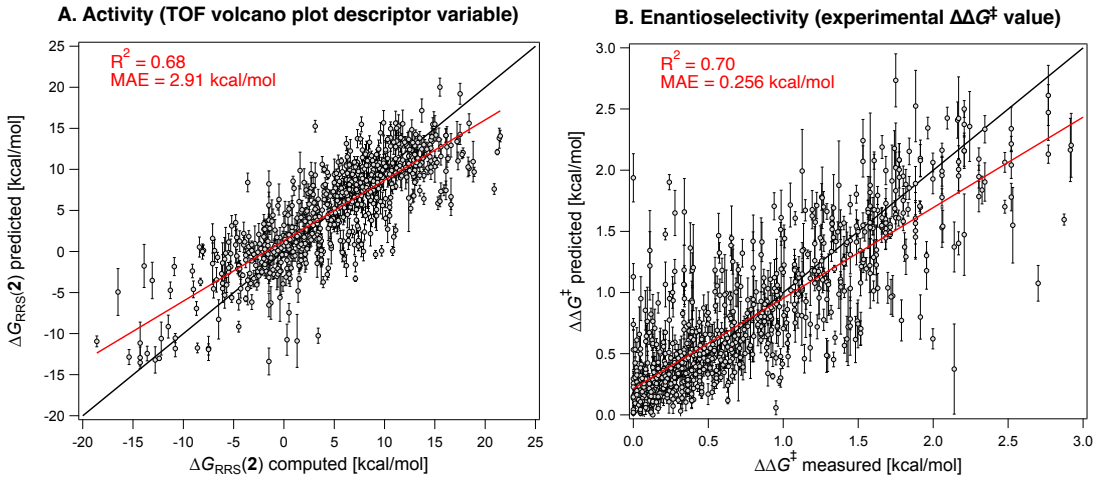


Figure 7.4 XGBoost models predicting the (A) descriptor variable [ΔG_{RRS}(**2**)] of the TOF molecular volcano plots and (B) the experimentally measured enantioselectivity (expressed as $\Delta\Delta G^{\ddagger}$) of the Pictet–Spengler reactions.

7.2.3 Interlude: reaction-inspired molecular representations for experimental enantioselectivity predictions

In the previous Chapter, we introduced physics-based or quantum machine learning (QML) models constructed using only atomic positions and nuclear charges as a generally applicable, efficient, and accurate framework for predicting molecular properties.[39] We showed how reaction-inspired representations are constructed to better describe the transformational nature of

chemical reactions.[41] In Chapter 6, representations were constructed as the difference between the SLATM[337] of the intermediates preceding and following the enantiodetermining transition state (*i.e.*, Int–(*R*) and P–(*R*) in Figure 7.5). This was consistent with the target of the kernel ridge regression model being the *computed* activation energy of the stereocontrolling step in the propargylation reaction (corresponding to $\Delta G^{\ddagger}$–(*R*) in Figure 7.5). Each enantiomeric pathway could be treated independently and enantioselectivity calculated from the difference between $\Delta G^{\ddagger}$–(*R*) and $\Delta G^{\ddagger}$–(*S*). However, experimental selectivity measures only provide information on $\Delta\Delta G^{\ddagger}$ (assuming Transition State Theory and Curtin–Hammett conditions),[12,483] with $\Delta G^{\ddagger}$–(*R*/*S*) being individually inaccessible. Therefore, our reaction-inspired representations must be reformulated to accommodate the different nature of the target, in this case the experimental $\Delta\Delta G^{\ddagger}$ values.

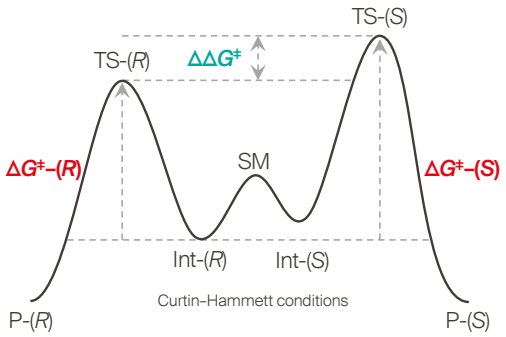


Figure 7.5 Illustrative Gibbs energy profile for a reaction under Curtin–Hammett conditions. SM = starting material.

Herein, we train kernel ridge regression (KRR) models and construct reaction-inspired representations as the difference between the SLATM of the enantiomers of **2** in the catalytic cycle of the Pictet–Spengler cyclization (Figure 7.3A, corresponding to Int–(*R*) and Int–(*S*) in Figure 7.5). Intermediate **2** lies between the transition states for C–C bond formation (**TS2**) and proton abstraction (**TS3**) on the potential energy surface, either of which could be rate- and enantiodetermining (*vide supra*), and should therefore be a good "fingerprint" of the key structural rearrangements occurring during the reaction; it connects the concerted and stepwise cyclization pathways and is identified as the best descriptor for constructing the molecular

volcanos. Bearing in mind that surrogate models used to accelerate fitness evaluation during closed-loop optimization with generative models must be fast and affordable, xTB-optimized structures of **2** (prior to conformational sampling, see the Computational Details) are used to generate the SLATM fingerprints. For comparison, a KRR model using as input the representation of only one enantiomer is also tested. The results are shown in Figure 7.6. Compared to standard molecular representations (A, $SLATM_2$), $SLATM_{DIFF}$ (B) is associated with a lower mean absolute error (0.29 kcal/mol) and higher R^2 (0.61). By subtracting the SLATM of **2**–(*R*) and **2**–(*S*), the global features that are common to the two enantiomeric pathways are eliminated, and the structural elements that are important for enantioselectivity are highlighted. In fact, while $SLATM_2$ contains (on average) 27,678 features that are effectively non-zero (*i.e.*, $> 10^{-10}$), $SLATM_{DIFF}$ only has 17,482: a more compact representation with a better "signal-to-noise ratio" is thus a better fingerprint of $\Delta\Delta G^{\ddagger}$.

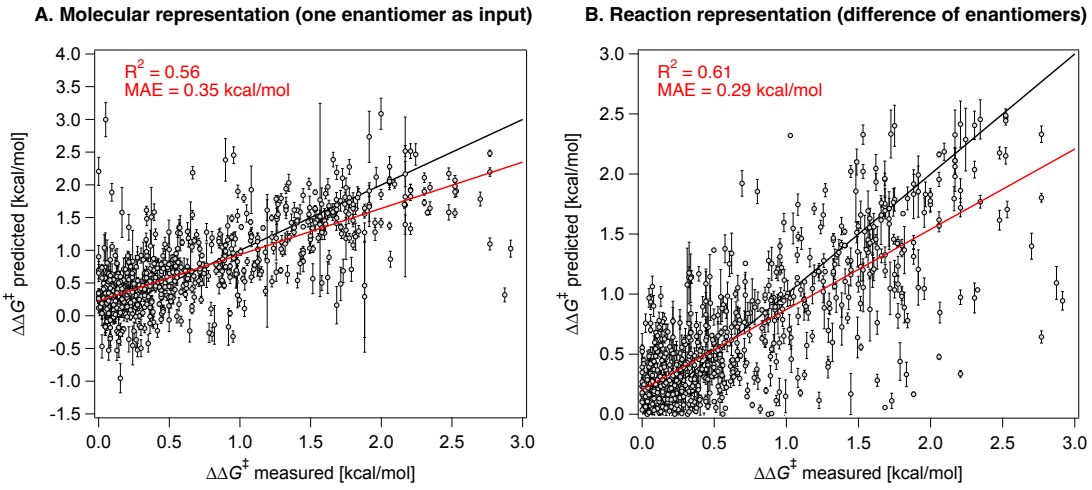


Figure 7.6 Kernel ridge regression models of $\Delta\Delta G^{\ddagger}$ using as input the SLATM representation of one of the enantiomers of intermediate **2** in the Pictet–Spengler catalytic cycle (A) or the difference between the SLATM representations of **2**–(*R*) and **2**–(*S*) *i.e.*, $SLATM_{DIFF}$.

Despite these promising results, and the rationality behind the design of $SLATM_{DIFF}$, the KRR model is outperformed by XGBoost with the concatenated Morgan fingerprints (Figure 7.4B). We attribute this performance to the "noisy" nature of the database, which has been collected from over 15 publications spanning nearly two decades. In fact, reactions have been performed under extremely different conditions with varying degrees of reproducibility.[484] The

conformational complexity of the system also poses a significant challenge: **2** consists of three non-covalently bound species, two of which are charged, and the lowest energy structures of each enantiomer might be significantly dissimilar. Efforts towards improving the accuracy of the KRR models and the design of reaction representations compatible with experimental targets are ongoing; however, since fingerprints derived from SMILES strings[328] are easier and faster to implement in the genetic optimization pipeline (Figure 7.1), fitness evaluation is performed with the XGBoost models.

7.2.4 Fragment database: the catalyst and substrate scope

The total combinatorial space explored during the evolutionary experiments is determined by the extent of the library of catalyst components and the scheme chosen to fragment them into building blocks. Here, we leverage the recently reported Organic Structures for CAtalysis Repository (OSCAR),[29] which contains 4,000 organocatalysts mined from the literature and CSD along with their corresponding molecular fragments. From OSCAR, we select 17 catalyst templates and 553 possible substituents (grouped into 7 categories $R^{1–7}$ depending on which template they may substitute, see the SI for a full list). The templates include 12 single and dual-HBDs (ureas, thioureas, squaramides, thiosquaramides, and prolyl-ureas) and 5 CPAs as shown in Figure 7.2 (and SI), which have been screened in the asymmetric Pictet–Spengler reaction. They are represented as flexible SMILES strings, written in such a way that that different $R^{1–7}$ can easily be introduced and exchanged, yielding valid SMILES. This results in a total combinatorial space of 2.85×10^8 HBDs and 1428 CPAs. Note that only CPAs with equal substituents at the 6 and 6′ positions of the BINOL/SPINOL scaffold are considered: although this significantly reduces the size of the combinatorial space, it ensures synthetic accessibility, a common problem of generative models.[332]

Having established the catalyst scope, we turn our attention to the substrate scope. Since our previous experiments with NaviCatGA were specificity-oriented,[38] we have to implement a different workflow for selecting a representative subset of substrates for generality-driven genetic

optimization. Inspired by recent work by Doyle *et al.*[485] and Sigman *et al.*,[486,487] we use the web platform Reaxys® to identify a list of 743 distinct Pictet–Spengler reactions (selective and non-, catalytic and non-) following the scheme in Figure 7.2. Additionally, 197 unprotected *β*–arylethylamines (SubA), filtered according to molecular weight (< 300 g/mol), commercial availability, and functional group compatibility, are included. Combined with the 240 unique organocatalytic reactions from the original Pictet–Spengler database, we obtain 258 distinct tryptamine derivatives (SubA) and 379 carbonyl compounds (SubB). Thus, the total combinatorial substrate space, shown in Figure 7.7A, encompasses 97,782 possible tetrahydro-*β*-carboline products.

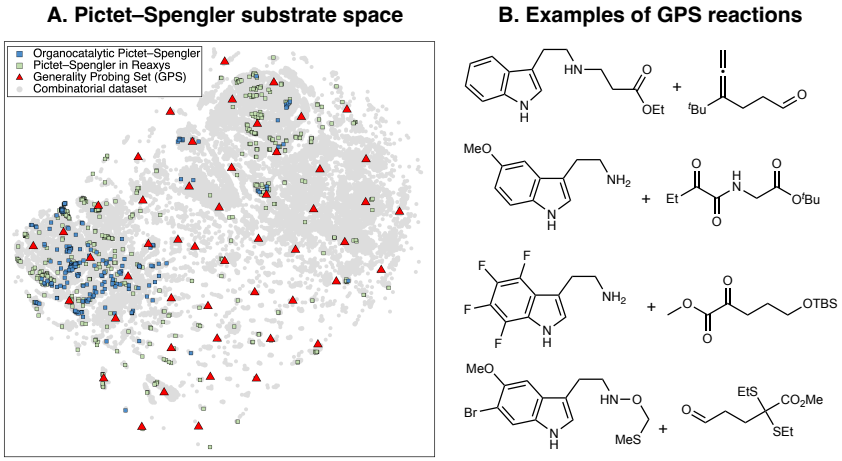


Figure 7.7 2D t-SNE map of the substrate scope on the basis of the concatenated MFPs of SubA and SubB. Blue squares indicate organocatalytic reactions, green squares reactions reported in Reaxys®, red triangles the Generality Probing Set (GPS).

Broadly speaking, examples from the literature cover the left half of the chemical space (which corresponds to unsubstituted tryptamines), while the right and bottom areas are sparsely covered. To generate a diverse and unbiased substrate scope for evolutionary experiments, we perform farthest point sampling and select 50 reactions aimed at covering the whole chemical space. Examples of this Generality Probing Set (GPS) are shown in Figure 7.7B (the full list is given in the SI). Carbonyls (SubB) include predominantly aromatic and aliphatic aldehydes, as reflected by the popularity of these substrates in the Pictet–Spengler reaction,[42] but also less explored *α*-diones, *α*-ketoamides, esters, and acids. Substituents on the tryptamine derivative are present on

all positions of the indole ring through mono-, di-, tri-, and even tetrasubstitution patterns, encompassing both electron-donating (*e.g.*, hydroxyl, methoxy, alkyl) and electron-withdrawing (*e.g.*, nitro, halide, ester) functional groups. This significantly contrasts the previously reported scope (*i.e.*, organocatalytic reactions from the literature or those mined from Reaxys®), dominated by monosubstituted *β*–arylethylamines. Approximately 60% of SubA in the GPS are unprotected, although a variety of protecting groups (*e.g.*, benzyl, 4-NO_2-benzyl, methylthiomethyl ether,[488] allyl[489]) are present.

7.3 Results and Discussion

7.3.1 Evolutionary experiments

With the different components of the inverse design pipeline at hand (Figure 7.1), we perform evolutionary experiments looking for organocatalysts displaying high enantioselectivity and activity (*i.e.*, TOF) across the whole substrate scope. The optimization targets are the median $\Delta\Delta G^{\ddagger}$ and $\Delta G_{RRS}(\mathbf{2})$ (the molecular volcano plot descriptor) of the 50 reactions in the Generality Probing Set. To solve this multi-objective problem and find trade-offs between activity and selectivity, we use the achievement scalarizing function Chimera.[338] In the first experiment, performed on the HBD catalyst space, a minimum $\Delta\Delta G^{\ddagger}_{med}$ = 2.0 kcal/mol value is imposed, the activity fitness score f_i of candidate i is maximized with a 10% degradation threshold, and the standard deviations of $\Delta\Delta G^{\ddagger}_{med}$ and f_i are reduced with a 25% compromise. The fitness score f_i is obtained by evaluating the corresponding $\Delta G_{RRS}(\mathbf{2})_{med}$ value compared to a normalized gaussian distribution centered on the target x (–9 kcal/mol, the volcano peak): $f_i = \exp\left(-\frac{1}{2}\left(\frac{\Delta G_{RRS}(\mathbf{2})_{med} - x}{\sigma}\right)^2\right)$ where $\sigma = \frac{|x|}{2}$. This experiment exemplifies a typical optimization campaign, where enantioselectivity is to be guaranteed and only subsequently catalyst activity is to be optimized. It is initiated with 10 randomized individuals per population, a mutation rate of 10%, a selection rate of 25%, and run for 50 generations. The results are shown in Figure 7.8. For simplicity, we focus on the best individual in each generation and show how $\Delta\Delta G^{\ddagger}_{med}$ and

$\Delta G_{RRS}(\mathbf{2})_{med}$ of the GPS evolve. Therefore, if the composition of the population varies but the top-ranking candidate remains the same, the quantities plotted in Figure 7.8 do not change. For example, despite the experiment being run for 50 iterations, after generation 32 the identity of the best organocatalyst remains unchanged, and no further variation of $\Delta\Delta G^{\ddagger}_{med}$ and $\Delta G_{RRS}(\mathbf{2})_{med}$ is observed.

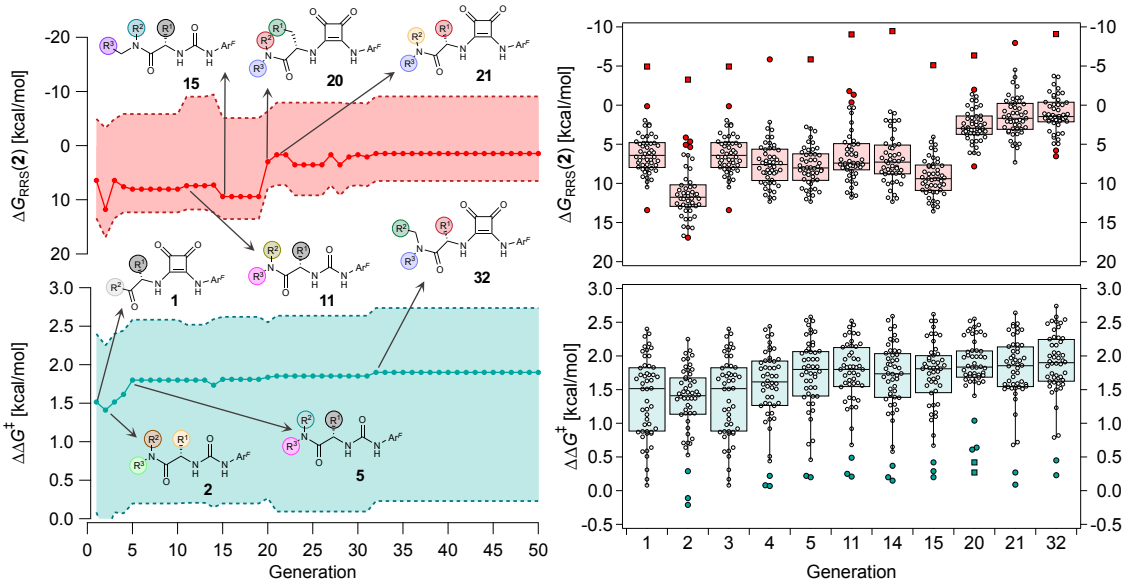


Figure 7.8 (Left) Evolution of $\Delta\Delta G^{\ddagger}$ and $\Delta G_{RRS}(\mathbf{2})$ of the top individual in the population over 50 generations. The solid lines indicate the median across the GPS, and the shaded areas represent the upper and lower values. Selected catalysts are shown, with different colored spheres representing different R^{1-3} substituents. (Right) Box and whisker chart of $\Delta\Delta G^{\ddagger}$ and $\Delta G_{RRS}(\mathbf{2})$ for selected generations *i.e.*, only when the structure of the best-performing catalyst changes. Each datapoint corresponds to a reaction in the GPS. Outliers and far outliers are indicated with filled circles and squares, respectively.

Over the first 5 generations, $\Delta\Delta G^{\ddagger}_{med}$ increases from 1.5 kcal/mol to 1.8 kcal/mol while the interquartile range (IQR) decreases, indicating that the top candidate is generally more selective across the GPS. At the onset of the evolutionary experiment, NaviCatGA locates DHBDs with the amide-based template [–C(=O)NR_2] as important for selectivity. Indeed, computational studies[477] have shown that the amide O engages the substrate through an H-bonding interaction with the indoline N–H. This template[490] is preserved throughout the GA run and preferred over catalysts containing the pyrrolidino-moiety:[123,164] Jacobsen *et al.* similarly found that aryl pyrrolidine substituted thioureas had lower generality metric than acyclic amides in the Pictet–

Spengler condensation of aldehydes.[42] Regarding the identity of the hydrogen-bonding unit, for the first 20 generations ureas are selected over squaramides to increase $\Delta\Delta G^{\ddagger}_{med}$, but, in accordance with trends extracted from the volcano plots and the lower acidity/H-bonding ability of ureas *vs*. squaramides,[33,289] this results in diminished activity ($\Delta G_{RRS}(\mathbf{2})_{med}$ values farther away from the volcano peak). This situation exemplifies a typical problem in reaction optimization, where improving one objective is sometimes only possible at the expense of another.[230,373] The same amino acid substituent (R^1) is also maintained until generation 20, with NaviCatGA favoring the diphenyl group (black spheres in Figure 7.8). At this particular iteration of the optimization procedure, the squaramide HBD unit is "rediscovered", which leads to a noticeable improvement in activity ($\Delta G_{RRS}(\mathbf{2})_{med}$ from 9.4 to 3.0 kcal/mol). Although this is associated with only marginal increase in $\Delta\Delta G^{\ddagger}_{med}$ (1.81 to 1.84 kcal/mol), the IQR significantly decreases, and most reactions in the GPS have $\Delta\Delta G^{\ddagger} \geq 1.7$ kcal/mol. Different R^{1-3} substituents are also selected, and in the remaining generations NaviCatGA explores different substitution patterns to achieve further activity and selectivity enhancements. In particular, $\Delta G_{RRS}(\mathbf{2})_{med}$ is decreased to 1.5 kcal/mol with small IQR (generation 32), while $\Delta\Delta G^{\ddagger}_{med}$ reaches the value of 1.9 kcal/mol. The most general organocatalyst found at the end of the evolutionary experiment exhibits the 2,4,6-^{i}Pr-C_6H_2 substituent as R^1, 3,5-CF_3-C_6H_3 as R^2, and the CH(2-^{t}Bu-C_6H_4)$_2$ group in place of R^3. Clearly, bulky substituents are privileged in inducing high enantioselectivity and activity across the GPS.

7.3.2 Chemical insight into generality

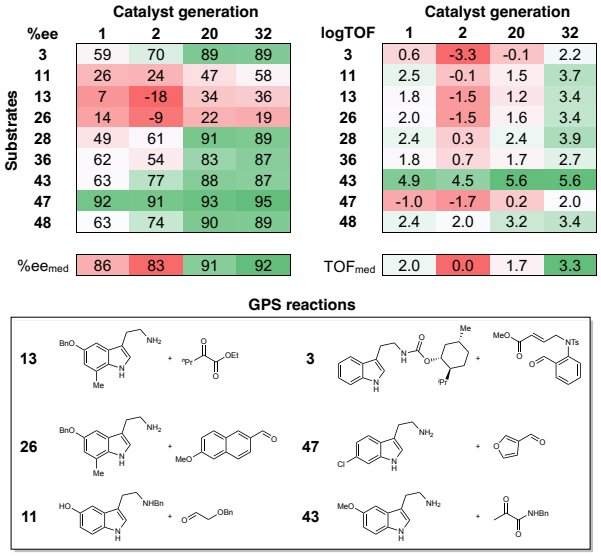


Figure 7.9 Calculated *ee* and logTOF values from the predicted $\Delta\Delta G^{\ddagger}$ and ΔG_{RRS}(**2**). Results are shown for selected catalyst generations (*x*-axis) and reactions in the GPS (*y*-axis), while *ee* and logTOF median values (bottom) consider all 50 reactions. Selected SubA and SubB combinations are shown.

Tabulation of the results of the evolutionary experiment on the HBD space as a heatmap, converted to *ee* and logTOF values (Figure 7.9) shows that, although a catalyst with good *median* selectivity and activity may be found (%ee_{med} = 92, $\log TOF_{med}$ = 3.3), some reactions in the GPS are always associated with poor performance *i.e.*, no matter how the structure of the catalyst evolves during the optimization, certain tetrahydro-*β*-carboline products may not be obtained in high *ee* or TOF. For example, the best performing HBD organocatalyst (*vide supra*) is predicted to achieve *ee* values of only 36% and 19% in reactions **13** and **26**, respectively. Both condensations involve an unprotected *β*–arylethylamine (SubA) substituted at the 7-position of the indole ring; similarly, Suzuki and co-workers found that 7-methyltryptamine and ethyl 2-oxopentanoate could only be converted in 45% *ee*.[458] These results can be explained in terms of steric effects of the methyl group on the substrate disrupting key non-covalent interactions between the catalyst's amide O and the indole N–H, which are evidently essential for inducing high enantioselectivity.[477] Considering activity, throughout the NaviCatGA run reactions **3** and **47** are underperforming: according to the volcano plot (Figure 7.C), the formation of the corresponding protonated tetrahydro-*β*-carboline **2** is energetically unfavorable, in line with the

electron-deficient nature of SubA and the electron-withdrawing character of the aldehyde substituent, which hinders the rate-determining deprotonation step. Regardless of the specific substitution patterns the GA may explore during the optimization, finding organocatalysts that non-covalently stabilize such unstable intermediates is clearly a challenge. Reaction **47** also exemplifies a situation where high selectivity and activity are incompatible: while most HBD organocatalysts explored during the evolutionary experiment are predicted to exhibit large $\Delta\Delta G^{\ddagger}$ values, the TOF always remains far from the theoretical maximum indicated by the volcano plot. Conversely, reaction **43**, which features an electron-rich indole and an α-ketoamide (essentially an activated carbonyl compound),[491] has predicted TOF always close to the volcano peak, while selectivity is more challenging to optimize,[444] and *ee* values considerably improve during the GA run (from 63% to 87%).

7.4 Conclusions

Given the synthetic utility of catalytic methods that provide high enantioselectivities and activities across a wide assortment of substrates, we have developed an optimization workflow centered on the open-source genetic algorithm NaviCatGA[38] with the aim of demonstrating how generative models[18] are an enticing alternative to experimental[42] or computational[436] high-throughput screening, provided that the various component of the pipeline for *de novo* catalyst design are adapted to optimize generality as primary target. We have adopted a hybrid approach for scoring candidate organocatalysts that combines a mechanistic-guided strategy (*i.e.*, activity estimations through TOF molecular volcano plots[30]) with enantioselectivity predictions based on training on experimental data. Catalysts were generated from molecular building blocks extracted from the OSCAR database.[29] We have tested our approach on the asymmetric Pictet–Spengler reaction[443] because of the large amount of data available in the literature and the many catalyst chemotypes that have been tested on individual substrate classes, resulting in system-specific islands of high performance.[42] We selected a broad and diverse substrate scope guided by mapping the chemical space of commercially and synthetically available tryptamine derivatives

and carbonyl compounds tested in the Pictet–Spengler cyclization, and performed evolutionary experiments on this Generality Probing Set (GPS). Through multi-objective optimization, we have explored activity/selectivity trade-offs and located solutions in the Pareto front with good *average* performance. However, we found that even the top organocatalysts are underperforming in certain areas of substrate space, while other areas are less sensitive to the identity of the HBD catalyst. Analysis of these outliers provided support to hypotheses on the principle of stereoinduction[477] and activity trends extracted from molecular volcanos, demonstrating how genetic optimization also yields mechanistic understanding and reveals structure–property relationships, as long as an unbiased substrate scope is chosen.[485] Given the encouraging results obtained here, we believe generality-oriented evolutionary experiments, coupled with experimental verification, will accelerate the discovery of broadly applicable catalyst systems for other interesting transformations. Ongoing and further investigations (currently omitted from this Chapter) will focus on evaluating the CPA space, performing NaviCatGA runs with different Chimera settings, and comparing the outcome of specificity- *vs*. generality-oriented experiments.

7.5 Computational Details

7.5.1 Quantum chemistry

The structure of both enantiomers of intermediate **2** in the catalytic cycle of the Pictet–Spengler reaction (Figure 7.3A, labeled as "Big group pointing Up", "BU", or "Big group pointing Down", "BD", depending on the relative position of R^1 and R^2 in **2**) were generated by substituting 3D fragments on 20 pre-optimized templates based on work by Jacobsen *et al*.[477] using AaronTools[78,79] and optimizing them with the semiempirical GFN2-xTB Hamiltonian[257] in the gas phase. Conformational sampling of the resulting 703 complexes was carried out using the Conformer-Rotamer Ensemble Sampling Tool[93,492,493] (CREST) at the GFN2-xTB//GFN-FF level of theory,[257] constraining positions of the bond-forming atoms. The lowest-energy conformer was selected and optimized at the PCM(Toluene)/M06-2X-D3/Def2-TZVP//M06-2X-D3/Def2-SVP level.[252,290,291,494–496] The other intermediates and TSs in the SRS were located

using scans and IRC computations.[292] The potential energy profile of only one enantiomeric pathway (corresponding to "BD"–labeled structures) was generated to construct volcano plots (*vide infra*). Stationary points were characterized on the basis of their vibrational frequencies (minima with zero imaginary frequencies, TSs with one imaginary frequency). Thermal and entropic corrections were calculated using Grimme's quasi-RRHO approximation[497] from frequencies computed at 298 K using the GoodVibes program[293] with a frequency cut-off value of 100 wavenumbers. All DFT computations were carried out using Gaussian16 (revision C.01).[249] The relative Gibbs free energies were automatically post-processed using the toolkit volcanic[32] to establish LFESRs, determine the choice of the descriptor variable [the relative energy of intermediate **2**, ΔG_{RRS}(**2**)], and construct TOF-volcano plots. Extensive instructions on how volcano plots are constructed are given elsewhere,[32] while the input for volcanic is provided in the SI.

7.5.2 Machine learning

MFPs of catalysts, co-catalysts, substrates, and solvents with a fingerprint size of 1024 were generated using RDKit[94] from their SMILES strings.[328] Chemical space maps were generated using Scikit-learn[406] on the basis of the concatenated MFPs with dimensions reduced to 100 using Principal Component Analysis, followed by t-SNE embedding[207] with perplexity of 30 to further reduce the featurization to two dimensions for visualization. The Python package QML[405] was used to construct standard SLATM representations,[337] while reaction-inspired $\text{SLATM}_{\text{DIFF}}$ representations were constructed in analogy to our previous work.[40,41] Random forest models from the XGBoost library were used with default hyperparameters. The input of the XGBoost models were the concatenated MPFs of Cat, Co-cat, SubA, SubB, and Solvent for $\Delta\Delta G^{\ddagger}$, and of Cat, Co-Cat (*i.e.*, AcOH, BzBr, or none), SubA, and SubB for ΔG_{RRS}(**2**). Not that, during the evolutionary experiments on the HBD space, toluene was fixed as solvent, while benzoic and acetic acid were fixed as co-catalysts and used in the input of the ML models for selectivity and activity, respectively. A cross-validation scheme was used with 100 different 90/10 training/test

splits [738/82 for $\Delta\Delta G^{\ddagger}$, 633/70 for ΔG_{RRS}(**2**)]. For the KRR models with the SLATM representations, hyperparameters were optimized for each train/test split. From the 100 different train/test splits, the target [$\Delta\Delta G^{\ddagger}$ or ΔG_{RRS}(**2**)] was predicted approximately 10 times; these test predictions were then averaged to obtain one final prediction. The standard deviation from the test predictions were used to generate the error bars.

7.6 Supporting Information

The Supporting Information for this Chapter will be made available prior to its submission for publication.

8 General Conclusions and Outlook

Catalyst design has played a pivotal role in optimizing organocatalytic reactions by improving chemical efficiency, expanding the number of amenable transformations, and diversifying the breath of possible substrate activation modes, leading to applications of organocatalysts in the asymmetric total synthesis of compounds of biological and pharmaceutical interest.[6] Since the field's infancy,[43] the increasing implementation of automation and computational techniques, primarily DFT methods to create potential energy profiles, has facilitated the discovery of new catalytically competent motifs and the screening of reactions for next generation catalysts. As recognized by the community, a paradigm shift is underway,[498] whereby the introduction of artificial intelligence-based strategies, fueled by "Big Data" availability and more sophisticated machine learning algorithms, is overcoming some of the previous limits in catalyst design and synthetic planning.[426,499] This thesis emphasizes the development and use of data-driven tools and concepts, such as molecular volcano plots, (un)supervised ML techniques, and generative models going beyond the state-of-the-art,[28] to predict the performance of catalyst and optimize reaction properties. A brief summary of the work presented herein is found below, following the three objectives stated in the Introduction.

Firstly, we have introduced the OSCAR repository and a fragment-based strategy for database curation. With its thousands of structures mined from the literature and corresponding building blocks to re-assemble in a combinatorial fashion, OSCAR represents the first steps towards an extensive mapping of organocatalyst space with large chemical diversity, aiding in the implementation of generative and predictive models of catalyst performance. We then used the

kind of molecular fragments found in OSCAR to tailor the structure of bifunctional hydrogen-bond donor/amines for improved turnover. Our approach relies on curating functionally diverse libraries of catalytic motifs and evaluating activity in terms of the individual fragment contributions through the use of activity maps and statistical modelling. Altogether, this work shows how the under-exploited modularity of organocatalysts and bottom-up protocols may be leveraged to streamline activity-based screening and chemical space exploration.

Secondly, we have established the use of volcano plots/activity maps as fitness function in closed-loop genetic optimization of homogeneous catalysts, in combination with the aforementioned fragment libraries. The versatile GA package, NaviCatGA, was developed for this purpose and we showcased its ability to efficiently explore large combinatorial spaces and optimize multiple targets simultaneously *i.e.*, find solutions in the activity–selectivity Pareto front.

Finally, we have addressed issues regarding accurate predictions of difficult-to-learn properties, such as enantioselectivity and catalyst generality. While physics-based ML models hold great potential to accelerate fitness evaluation during genetic optimization owing to their transferability and efficiency, their use for catalytic properties is less routine.[40] Reaction-inspired representations are a chemically intuitive strategy to improve accuracy for subtle targets such as DFT-computed *e.e.* values. Going beyond specificity-oriented optimization, we have shown how statistical models for enantioselectivity and activity prediction (trained on experimental data mined from the literature and/or on the volcano plot's descriptor), tailored fragments databases, and genetic optimization are combined to design evolutionary experiments that may address the existence of "general" catalysts.

Data-driven tools and concepts are undoubtedly invaluable to streamline the discovery of prospective organocatalysts and identifying trends surrounding catalytic behavior. However, we believe that the capability of this toolbox has not yet been fully exhausted. There are still possible extensions and refinements, some of which are listed below:

- *Alternative chemical descriptors to explore new design principles in non-covalent organocatalysis*

  Over the years our group has developed an extensive toolbox based on molecular volcanos.[30] Broadly speaking, they have been geared towards quickly estimating catalytic cycle energetics with the aim of identifying prospective new catalysts or better understand why particular species possess certain reactivity. However, the study of organocatalytic reactions prompts the creation of "next generation" plots for further generalization. In particular, analysis of the molecular volcanos shown in this thesis (Figure 4.2, 5.5, and 7.3) reveals one common feature: despite representing different reactions and mechanisms, no species was found lying on the left slope, corresponding to the organocatalyst binding intermediates too strongly and turnover being limited by product release. Although product inhibition is a known, common problem in catalysis (*e.g.*, Claisen rearrangement[500]), it is unclear at this stage whether this aspect is a general feature of (non-covalent) organocatalytic reactions and whether organocatalysts lying on the strong-binding slope can be found. Having access to this information would disclose new design principles and allow maximum activity (*i.e.*, the volcano peak) to be reached "from the other side": not by modulating the strength of starting material–catalyst interactions (right slope), but between product and catalyst (left slope). To achieve this, we envision that new fragment-based approaches and families of molecular volcanos will have to be created, for example employing non-energy-based descriptors (pK_a, polarizability, quadrupole moment, *etc*.)[501] to account for the strength of NCIs.

- *Improving fitness evaluation, multi-objective design, and catalyst fragmentation in the context of genetic optimization*

  While implementing genetic and other generative algorithms for catalyst discovery is becoming routine, developing affordable models to accurately predict complex catalytic properties on-the-fly remains a challenge. This is partly due to the scarcity of high-quality HTE datasets, and to the difficulty associated with using *ab initio* methods to generate

reliable reactivity data. In this thesis, we have shown that physics-based models with approximate (*i.e.*, computationally inexpensive) geometries are compatible with genetic optimization; however, training them on "noisy" experimental data mined from different publications limits the applicability of reaction-based representations. Evaluating the performance of different atomistic models using "clean" HTE datasets would reveal the requirements and compatibility of molecular representations with experimental targets (*e.g.*, yields) and help in designing more accurate reaction fingerprints.

Catalyst discovery is a multi-objective task and improvements in the decision-making process will make evolutionary experiments more efficient. One approach would be to implement more sophisticated acquisition functions, such as non-dominated sorting (*i.e.*, Pareto ranking)[502] to select sampled catalysts to include in subsequent populations. Alternatively, the definition of performance, rather than being fixed prior to the experiment, can be dynamic and able to respond to new knowledge generated on-the-fly, such as the unforeseen stability of new catalysts.[27]

Finally, to overcome the bias connected with user-defined libraries of molecular building blocks, diversity quantification[503] of the fragments database and active learning approaches,[371] which balance the exploitation of familiar chemical spaces with the exploration of areas of high uncertainty, could be implemented into our pipeline.

- *Extending other NaviCat platform tools to organocatalytic reactions*

  Over the past four years, our laboratory has been developing and collecting a number of data-driven tools for digital chemistry and catalyst discovery under the NaviCat platform (https://github.com/lcmd-epfl/NaviCat). Some of NaviCat's modules include database and structure generation utilities,[29] ML advancements and optimization pipelines,[38] and an automated volcano plots builder.[32] While much of this thesis' work has revolved around extending these modules to organocatalysis, further integration is envisioned. For example, one tool from the Reiher group that has become part of NaviCat for transition-metal catalyst

screening[360,504] is the SCINE Molassembler module.[97] Molassembler couples automated functionalization with conformer generation on-the-fly for *e.e.* estimation: given the challenges associated with enantioselectivity prediction in (non-covalent) organocatalysis, extending the applicability of this module to organocatalytic reactions would greatly enrich our pipeline for catalyst optimization, facilitating high-throughput mechanistic investigations and helping prioritize experimental testing of promising candidates.

In closing, we believe that data-driven tools can streamline the process of reaction optimization by enabling the discovery of key catalyst structure–activity and structure–selectivity relationships. This thesis demonstrates how organocatalysis benefits from the application of tailored yet transferable fragment-oriented (inverse) design pipelines, powered by (un)supervised machine learning algorithms, serving as a powerful driving force for the development of new sustainable transformations.

Bibliography

(1) Xiang, S.-H.; Tan, B. Advances in Asymmetric Organocatalysis over the Last 10 Years. *Nat. Commun.* **2020**, *11* (1), 3786.

(2) Gomollón-Bel, F. Ten Chemical Innovations That Will Change Our World: IUPAC Identifies Emerging Technologies in Chemistry with Potential to Make Our Planet More Sustainable. *Chem. Int.* **2019**, *41*, 12–17.

(3) MacMillan, D. W. C. The Advent and Development of Organocatalysis. *Nature* **2008**, *455* (7211), 304–308.

(4) Aukland, M. H.; List, B. Organocatalysis Emerging as a Technology. *Pure Appl. Chem.* **2021**, *93* (12), 1371–1381.

(5) Zhou, Q.-L. Transition-Metal Catalysis and Organocatalysis: Where Can Progress Be Expected? *Angew. Chem. Int. Ed.* **2016**, *55* (18), 5352–5353.

(6) García Mancheño, O.; Waser, M. Recent Developments and Trends in Asymmetric Organocatalysis. *Eur. J. Org. Chem.* **2023**, *26* (1), e202200950.

(7) Poree, C.; Schoenebeck, F. A Holy Grail in Chemistry: Computational Catalyst Design: Feasible or Fiction? *Acc. Chem. Res.* **2017**, *50* (3), 605–608.

(8) 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* (8), 5042–5137.

(9) Iribarren, I.; Garcia, M. R.; Trujillo, C. Catalyst Design within Asymmetric Organocatalysis. *WIREs Comput. Mol. Sci.* e1616.

(10) Tsang, A. S.-K.; Sanhueza, I. A.; Schoenebeck, F. Combining Experimental and Computational Studies to Understand and Predict Reactivities of Relevance to Homogeneous Catalysis. *Chem. Eur. J.* **2014**, *20* (50), 16432–16441.

(11) Melnyk, N.; Iribarren, I.; Mates-Torres, E.; Trujillo, C. Theoretical Perspectives in Organocatalysis. *Chem. Eur. J.* **2022**, *28* (58), e202201570.

(12) Harper, K. C.; Sigman, M. S. Using Physical Organic Parameters To Correlate Asymmetric Catalyst Performance. *J. Org. Chem.* **2013**, *78* (7), 2813–2818.

(13) 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* (6), 1292–1301.

(14) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and Mechanistic Multivariate Linear Regression Models for Reaction Development. *Chem. Sci.* **2018**, *9* (9), 2398–2412.

(15) Pollice, R.; dos Passos Gomes, G.; Aldeghi, M.; Hickman, R. J.; Krenn, M.; Lavigne, C.; Lindner-D'Addario, M.; Nigam, A.; Ser, C. T.; Yao, Z.; Aspuru-Guzik, A. Data-Driven Strategies for Accelerated Materials Design. *Acc. Chem. Res.* **2021**, *54* (4), 849–860.

(16) Cordova, M.; Wodrich, M. D.; Meyer, B.; Sawatlon, B.; Corminboeuf, C. Data-Driven Advancement of Homogeneous Nickel Catalyst Activity for Aryl Ether Cleavage. *ACS Catal.* **2020**, *10* (13), 7021–7031.

(17) Friederich, P.; dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine Learning Dihydrogen Activation in the Chemical Space Surrounding Vaska's Complex. *Chem. Sci.* **2020**, *11* (18), 4584–4601.

(18) Anstine, D. M.; Isayev, O. Generative Models as an Emerging Paradigm in the Chemical Sciences. *J. Am. Chem. Soc.* **2023**, *145* (16), 8736–8750.
(19) Schilter, O.; Vaucher, A.; Schwaller, P.; Laino, T. Designing Catalysts with Deep Generative Models and Computational Data. A Case Study for Suzuki Cross Coupling Reactions. *Digital Discovery* **2023**, *2* (3), 728–735.
(20) Funes-Ardoiz, I.; Schoenebeck, F. Established and Emerging Computational Tools to Study Homogeneous Catalysis—From Quantum Mechanics to Machine Learning. *Chem* **2020**, *6* (8), 1904–1913.
(21) Fabrizio, A.; Meyer, B.; Fabregat, R.; Corminboeuf, C. Quantum Chemistry Meets Machine Learning. *CHIMIA* **2019**, *73* (12), 983.
(22) 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.
(23) Wheeler, S. E.; Seguin, T. J.; Guan, Y.; Doney, A. C. Noncovalent Interactions in Organocatalysis and the Prospect of Computational Catalyst Design. *Acc. Chem. Res.* **2016**, *49* (5), 1061–1069.
(24) Neel, A. J.; Hilton, M. J.; Sigman, M. S.; Toste, F. D. Exploiting Non-Covalent π Interactions for Catalyst Design. *Nature* **2017**, *543* (7647), 637–646.
(25) Seguin, T. J.; Lu, T.; Wheeler, S. E. Enantioselectivity in Catalytic Asymmetric Fischer Indolizations Hinges on the Competition of π-Stacking and CH/π Interactions. *Org. Lett.* **2015**, *17* (12), 3066–3069.
(26) Hageman, J. A.; Westerhuis, J. A.; Frühauf, H.-W.; Rothenberg, G. Design and Assembly of Virtual Homogeneous Catalyst Libraries –Towards in Silico Catalyst Optimisation. *Adv. Synth. Catal.* **2006**, *348* (3), 361–369.
(27) Foscato, M.; Jensen, V. R. Automated in Silico Design of Homogeneous Catalysts. *ACS Catal.* **2020**, *10* (3), 2354–2377.
(28) Reid, J. P.; Sigman, M. S. Comparing Quantitative Prediction Methods for the Discovery of Small-Molecule Chiral Catalysts. *Nat. Rev. Chem.* **2018**, *2* (10), 290–305.
(29) Gallarati, S.; van Gerwen, P.; Laplaza, R.; Fabrizio, A.; Vela, S.; Corminboeuf, C. OSCAR: An Extensive Repository of Chemically and Functionally Diverse Organocatalysts. *Chem. Sci.* **2022**, *13*, 13782–13794.
(30) Wodrich, M. D.; Sawatlon, B.; Busch, M.; Corminboeuf, C. The Genesis of Molecular Volcano Plots. *Acc. Chem. Res.* **2021**, *54* (5), 1107–1117.
(31) Busch, M.; Wodrich, M. D.; Corminboeuf, C. Linear Scaling Relationships and Volcano Plots in Homogeneous Catalysis – Revisiting the Suzuki Reaction. *Chem. Sci.* **2015**, *6* (12), 6754–6761.
(32) Laplaza, R.; Das, S.; Wodrich, M. D.; Corminboeuf, C. Constructing and Interpreting Volcano Plots and Activity Maps to Navigate Homogeneous Catalyst Landscapes. *Nat. Protoc.* **2022**, *17*, 2550–2569.
(33) Gallarati, S.; Laplaza, R.; Corminboeuf, C. Harvesting the Fragment-Based Nature of Bifunctional Organocatalysts to Enhance Their Activity. *Org. Chem. Front.* **2022**, *9* (15), 4041–4051.
(34) dos Passos Gomes, G.; Pollice, R.; Aspuru-Guzik, A. Navigating through the Maze of Homogeneous Catalyst Design with Machine Learning. *Trends Chem.* **2021**, *3* (2), 96–110.
(35) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360–365.
(36) Gallarati, S.; van Gerwen, P.; Schoepfer, A. A.; Laplaza, R.; Corminboeuf, C. Genetic Algorithms for the Discovery of Homogeneous Catalysts. *CHIMIA* **2023**, *77* (1/2), 39.
(37) Doney, A. C.; Rooks, B. J.; Lu, T.; Wheeler, S. E. Design of Organocatalysts for Asymmetric Propargylations through Computational Screening. *ACS Catal.* **2016**, *6* (11), 7948–7955.
(38) Laplaza, R.; Gallarati, S.; Corminboeuf, C. Genetic Optimization of Homogeneous Catalysts. *Chem. Methods* **2022**, e202100107.

(39) von Lilienfeld, O. A. Quantum Machine Learning in Chemical Compound Space. *Angew. Chem. Int. Ed.* **2018**, *57* (16), 4164–4169.
(40) van Gerwen, P.; Fabrizio, A.; Wodrich, M. D.; Corminboeuf, C. Physics-Based Representations for Machine Learning Properties of Chemical Reactions. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045005.
(41) Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C. Reaction-Based Machine Learning Representations for Predicting the Enantioselectivity of Organocatalysts. *Chem. Sci.* **2021**, *12* (20), 6879–6889.
(42) Wagen, C. C.; McMinn, S. E.; Kwan, E. E.; Jacobsen, E. N. Screening for Generality in Asymmetric Catalysis. *Nature* **2022**, *610*, 680–686.
(43) Dondoni, A.; Massi, A. Asymmetric Organocatalysis: From Infancy to Adolescence. *Angew. Chem. Int. Ed.* **2008**, *47* (25), 4638–4660.
(44) Melnyk, N.; Garcia, M. R.; Iribarren, I.; Trujillo, C. Evolution of Design Approaches in Asymmetric Organocatalysis over the Last Decade. *Tetrahedron Chem.* **2023**, *5*, 100035.
(45) List, B.; Lerner, R. A.; Barbas, C. F. Proline-Catalyzed Direct Asymmetric Aldol Reactions. *J. Am. Chem. Soc.* **2000**, *122* (10), 2395–2396.
(46) Ahrendt, K. A.; Borths, C. J.; MacMillan, D. W. C. New Strategies for Organic Catalysis: The First Highly Enantioselective Organocatalytic Diels−Alder Reaction. *J. Am. Chem. Soc.* **2000**, *122* (17), 4243–4244.
(47) Cheong, P. H.-Y.; Houk, K. N.; Warrier, J. S.; Hanessian, S. Catalysis of the Hajos-Parrish-Eder-Sauer-Wiechert Reaction by Cis- and Trans-4,5-Methanoprolines: Sensitivity of Proline Catalysis to Pyrrolidine Ring Conformation. *Adv. Synth. Catal.* **2004**, *346* (9–10), 1111–1115.
(48) Houk, K. N.; Cheong, P. H.-Y. Computational Prediction of Small-Molecule Catalysts. *Nature* **2008**, *455* (7211), 309–313.
(49) Mitsumori, S.; Zhang, H.; Ha-Yeon Cheong, P.; Houk, K. N.; Tanaka, F.; Barbas, C. F. Direct Asymmetric Anti-Mannich-Type Reactions Catalyzed by a Designed Amino Acid. *J. Am. Chem. Soc.* **2006**, *128* (4), 1040–1041.
(50) Fleming, E. M.; Quigley, C.; Rozas, I.; Connon, S. J. Computational Study-Led Organocatalyst Design: A Novel, Highly Active Urea-Based Catalyst for Addition Reactions to Epoxides. *J. Org. Chem.* **2008**, *73* (3), 948–956.
(51) Gammack Yamagata, A. D.; Datta, S.; Jackson, K. E.; Stegbauer, L.; Paton, R. S.; Dixon, D. J. Enantioselective Desymmetrization of Prochiral Cyclohexanones by Organocatalytic Intramolecular Michael Additions to α,β-Unsaturated Esters. *Angew. Chem. Int. Ed.* **2015**, *54* (16), 4899–4903.
(52) Rooks, B. J.; Haas, M. R.; Sepúlveda, D.; Lu, T.; Wheeler, S. E. Prospects for the Computational Design of Bipyridine N,N′-Dioxide Catalysts for Asymmetric Propargylation Reactions. *ACS Catal.* **2015**, *5* (1), 272–280.
(53) Milo, A.; Neel, A. J.; Toste, F. D.; Sigman, M. S. A Data-Intensive Approach to Mechanistic Elucidation Applied to Chiral Anion Catalysis. *Science* **2015**, *347* (6223), 737–743.
(54) Crawford, J. M.; Kingston, C.; Toste, F. D.; Sigman, M. S. Data Science Meets Physical Organic Chemistry. *Acc. Chem. Res.* **2021**, *54* (16), 3136–3148.
(55) 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* (11), 3863–3875.
(56) Orlandi, M.; Coelho, J. A. S.; Hilton, M. J.; Toste, F. D.; Sigman, M. S. Parametrization of Non-Covalent Interactions for Transition State Interrogation Applied to Asymmetric Catalysis. *J. Am. Chem. Soc.* **2017**, *139* (20), 6803–6806.
(57) Orlandi, M.; Toste, F. D.; Sigman, M. S. Multidimensional Correlations in Asymmetric Catalysis through Parameterization of Uncatalyzed Transition States. *Angew. Chem. Int. Ed.* **2017**, *56* (45), 14080–14084.
(58) Biswas, S.; Kubota, K.; Orlandi, M.; Turberg, M.; Miles, D. H.; Sigman, M. S.; Toste, F. D. Enantioselective Synthesis of N,S-Acetals by an Oxidative Pummerer-Type

Transformation Using Phase-Transfer Catalysis. *Angew. Chem. Int. Ed.* **2018**, *57* (2), 589–593.

(59) Coelho, J. A. S.; Matsumoto, A.; Orlandi, M.; Hilton, M. J.; Sigman, M. S.; Toste, F. D. Enantioselective Fluorination of Homoallylic Alcohols Enabled by the Tuning of Non-Covalent Interactions. *Chem. Sci.* **2018**, *9* (35), 7153–7158.

(60) Miró, J.; Gensch, T.; Ellwart, M.; Han, S.-J.; Lin, H.-H.; Sigman, M. S.; Toste, F. D. Enantioselective Allenoate-Claisen Rearrangement Using Chiral Phosphate Catalysts. *J. Am. Chem. Soc.* **2020**, *142* (13), 6390–6399.

(61) 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. *J. Am. Chem. Soc.* **2020**, *142* (26), 11578–11592.

(62) 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. *Acc. Chem. Res.* **2021**, *54* (9), 2041–2054.

(63) 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.

(64) 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.

(65) Zahrt, A. F.; Henle, J. J.; Denmark, S. E. Cautionary Guidelines for Machine Learning Studies with Combinatorial Datasets. *ACS Comb. Sci.* **2020**, *22* (11), 586–591.

(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. *React. Chem. Eng.* **2021**, *6* (4), 694–708.

(67) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem* **2020**, *6* (6), 1379–1390.

(68) Tsuji, N.; Sidorov, P.; Zhu, C.; Nagata, Y.; Gimadiev, T.; Varnek, A.; List, B. Predicting Highly Enantioselective Catalysts Using Tunable Fragment Descriptors. *Angew. Chem. Int. Ed.* **2023**, *62* (11), e202218659.

(69) 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* (11), 6509–6560.

(70) Cheng, G.-J.; Zhang, X.; Chung, L. W.; Xu, L.; Wu, Y.-D. Computational Organic Chemistry: Bridging Theory and Experiment in Establishing the Mechanisms of Chemical Reactions. *J. Am. Chem. Soc.* **2015**, *137* (5), 1706–1725.

(71) Walden, D. M.; Ogba, O. M.; Johnston, R. C.; Cheong, P. H.-Y. Computational Insights into the Central Role of Nonbonding Interactions in Modern Covalent Organocatalysis. *Acc. Chem. Res.* **2016**, *49* (6), 1279–1291.

(72) Liu, Z.; Patel, C.; Harvey, J. N.; Sunoj, R. B. Mechanism and Reactivity in the Morita–Baylis–Hillman Reaction: The Challenge of Accurate Computations. *Phys. Chem. Chem. Phys.* **2017**, *19* (45), 30647–30657.

(73) Yepes, D.; Neese, F.; List, B.; Bistoni, G. Unveiling the Delicate Balance of Steric and Dispersion Interactions in Organocatalysis Using High-Level Computational Methods. *J. Am. Chem. Soc.* **2020**, *142* (7), 3613–3625.

(74) Harden, I.; Neese, F.; Bistoni, G. An Induced-Fit Model for Asymmetric Organocatalytic Reactions: A Case Study of the Activation of Olefins via Chiral Brønsted Acid Catalysts. *Chem. Sci.* **2022**, *13* (30), 8848–8859.

(75) Chin, Y. P.; Krenske, E. H. Nazarov Cyclizations Catalyzed by BINOL Phosphoric Acid Derivatives: Quantum Chemistry Struggles To Predict the Enantioselectivity. *J. Org. Chem.* **2022**, *87* (3), 1710–1722.

(76) 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.
(77) Gerosa, G. G.; Marcarino, M. O.; Spanevello, R. A.; Suárez, A. G.; Sarotti, A. M. Re-Engineering Organocatalysts for Asymmetric Friedel–Crafts Alkylation of Indoles through Computational Studies. *J. Org. Chem.* **2020**, *85* (15), 9969–9978.
(78) 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* (10), 5249–5261.
(79) Ingman, V. M.; Schaefer, A. J.; Andreola, L. R.; Wheeler, S. E. QChASM: Quantum Chemistry Automation and Structure Manipulation. *WIREs Comput. Mol. Sci.* **2021**, *11* (4), e1510.
(80) Schaefer, A. J.; Ingman, V. M.; Wheeler, S. E. SEQCROW: A ChimeraX Bundle to Facilitate Quantum Chemical Applications to Complex Molecular Systems. *J. Comput. Chem.* **2021**, *42* (24), 1750–1754.
(81) Drudis-Solé, G.; Ujaque, G.; Maseras, F.; Lledós, A. A QM/MM Study of the Asymmetric Dihydroxylation of Terminal Aliphatic n-Alkenes with OsO4·(DHQD)2PYDZ: Enantioselectivity as a Function of Chain Length. *Chem. Eur. J* **2005**, *11* (3), 1017–1029.
(82) Corbeil, C. R.; Moitessier, N. Theory and Application of Medium to High Throughput Prediction Method Techniques for Asymmetric Catalyst Design. *J. Mol. Catal. A: Chem.* **2010**, *324* (1), 146–155.
(83) Hansen, E.; Rosales, A. R.; Tutkowski, B.; Norrby, P.-O.; Wiest, O. Prediction of Stereochemistry Using Q2MM. *Acc. Chem. Res.* **2016**, *49* (5), 996–1005.
(84) 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. *J. Comput. Chem.* **2011**, *32* (13), 2878–2889.
(85) 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. *Nat. Catal.* **2019**, *2* (1), 41–45.
(86) 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.
(87) Harriman, D. J.; Deslongchamps, G. Reverse-Docking as a Computational Tool for the Study of Asymmetric Organocatalysis. *J. Comput. Aided Mol. Des.* **2004**, *18* (5), 303–308.
(88) Harriman, D. J.; Deslongchamps, G. Reverse-Docking Study of the TADDOL-Catalyzed Asymmetric Hetero-Diels–Alder Reaction. *J. Mol. Model.* **2006**, *12* (6), 793–797.
(89) Harriman, D. J.; Lambropoulos, A.; Deslongchamps, G. In Silico Correlation of Enantioselectivity for the TADDOL Catalyzed Asymmetric Hetero-Diels–Alder Reaction. *Tetrahedron Lett.* **2007**, *48* (4), 689–692.
(90) Joseph Harriman, D.; Deleavey, G. F.; Lambropoulos, A.; Deslongchamps, G. Reverse-Docking Study of the Organocatalyzed Asymmetric Strecker Hydrocyanation of Aldimines and Ketimines. *Tetrahedron* **2007**, *63* (52), 13032–13038.
(91) Sterling, A. J.; Zavitsanou, S.; Ford, J.; Duarte, F. Selectivity in Organocatalysis—From Qualitative to Quantitative Predictive Models. *WIREs Comput. Mol. Sci.* **2021**, *11* (5), e1518.
(92) Grimme, S.; Bohle, F.; Hansen, A.; Pracht, P.; Spicher, S.; Stahn, M. Efficient Quantum Chemical Calculation of Structure Ensembles and Free Energies for Nonrigid Molecules. *J. Phys. Chem. A* **2021**, *125* (19), 4039–4054.
(93) Pracht, P.; Bohle, F.; Grimme, S. Automated Exploration of the Low-Energy Chemical Space with Fast Quantum Chemical Methods. *Phys. Chem. Chem. Phys.* **2020**, *22* (14), 7169–7192.
(94) RDKit: Open-Source Chemoinformatics and Machine Learning. https://www.rdkit.org.

(95) Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47* (6), 2462–2474.
(96) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catal.* **2019**, *9* (3), 2313–2323.
(97) Sobez, J.-G.; Reiher, M. Molassembler: Molecular Graph Construction, Modification, and Conformer Generation for Inorganic and Organic Molecules. *J. Chem. Inf. Model.* **2020**, *60* (8), 3884–3900.
(98) Iribarren, I.; Trujillo, C. Efficiency and Suitability When Exploring the Conformational Space of Phase-Transfer Catalysts. *J. Chem. Inf. Model.* **2022**, *62* (22), 5568–5580.
(99) Plata, R. E.; Singleton, D. A. A Case Study of the Mechanism of Alcohol-Mediated Morita Baylis–Hillman Reactions. The Importance of Experimental Observations. *J. Am. Chem. Soc.* **2015**, *137* (11), 3811–3826.
(100) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314* (1–2), 141–151.
(101) Petraglia, R.; Nicolaï, A.; Wodrich, M. D.; Ceriotti, M.; Corminboeuf, C. Beyond Static Structures: Putting Forth REMD as a Tool to Solve Problems in Computational Organic Chemistry. *J. Comput. Chem.* **2016**, *37* (1), 83–92.
(102) Gallarati, S.; Fabregat, R.; Juraskova, V.; Inizan, T. J.; Corminboeuf, C. How Robust Is the Reversible Steric Shielding Strategy for Photoswitchable Organocatalysts? *J. Org. Chem.* **2022**, *87* (14), 8849–8857.
(103) 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.
(104) Miller, E.; Mai, B. K.; Read, J. A.; Bell, W. C.; Derrick, J. S.; Liu, P.; Toste, F. D. A Combined DFT, Energy Decomposition, and Data Analysis Approach to Investigate the Relationship Between Noncovalent Interactions and Selectivity in a Flexible DABCOnium/Chiral Anion Catalyst System. *ACS Catal.* **2022**, *12* (19), 12369–12385.
(105) Reid, J. P.; Simón, L.; Goodman, J. M. A Practical Guide for Predicting the Stereochemistry of Bifunctional Phosphoric Acid Catalyzed Reactions of Imines. *Acc. Chem. Res.* **2016**, *49* (5), 1029–1041.
(106) Reid, J. P.; Goodman, J. M. Selecting Chiral BINOL-Derived Phosphoric Acid Catalysts: General Model To Identify Steric Features Essential for Enantioselectivity. *Chem. Eur. J.* **2017**, *23* (57), 14248–14260.
(107) Reid, J. P.; Ermanis, K.; Goodman, J. M. BINOPtimal: A Web Tool for Optimal Chiral Phosphoric Acid Catalyst Selection. *Chem. Commun.* **2019**, *55* (12), 1778–1781.
(108) Li, X.; Xu, J.; Li, S.-J.; Qu, L.-B.; Li, Z.; Chi, Y. R.; Wei, D.; Lan, Y. Prediction of NHC-Catalyzed Chemoselective Functionalizations of Carbonyl Compounds: A General Mechanistic Map. *Chem. Sci.* **2020**, *11* (27), 7214–7225.
(109) Shi, Q.; Wang, W.; Wang, Y.; Lan, Y.; Yao, C.; Wei, D. Prediction on the Origin of Chemoselectivity in Lewis Base-Mediated Competition Cyclizations between Allenoates and Chalcones: A Computational Study. *Org. Chem. Front.* **2019**, *6* (15), 2692–2700.
(110) Chakraborty, D.; Chattaraj, P. K. Conceptual Density Functional Theory Based Electronic Structure Principles. *Chem. Sci.* **2021**, *12* (18), 6264–6279.
(111) Maji, B.; Breugst, M.; Mayr, H. N-Heterocyclic Carbenes: Organocatalysts with Moderate Nucleophilicity but Extraordinarily High Lewis Basicity. *Angew. Chem. Int. Ed.* **2011**, *50* (30), 6915–6919.
(112) Mayr, H.; Lakhdar, S.; Maji, B.; Ofial, A. R. A Quantitative Approach to Nucleophilic Organocatalysis. *Beilstein J. Org. Chem.* **2012**, *8*, 1458–1478.
(113) An, F.; Maji, B.; Min, E.; Ofial, A. R.; Mayr, H. Basicities and Nucleophilicities of Pyrrolidines and Imidazolidinones Used as Organocatalysts. *J. Am. Chem. Soc.* **2020**, *142* (3), 1526–1547.
(114) Maji, B.; Joannesse, C.; Nigst, T. A.; Smith, A. D.; Mayr, H. Nucleophilicities and Lewis Basicities of Isothiourea Derivatives. *J. Org. Chem.* **2011**, *76* (12), 5104–5112.

(115) Mayr, H.; Schneider, R.; Grabis, U. Linear Free Energy and Reactivity-Selectivity Relationships in Reactions of Diarylcarbenium Ions with π-Nucleophiles. *J. Am. Chem. Soc.* **1990**, *112* (11), 4460–4467.
(116) Li, X.; Deng, H.; Zhang, B.; Li, J.; Zhang, L.; Luo, S.; Cheng, J.-P. Physical Organic Study of Structure-Activity-Enantioselectivity Relationships in Asymmetric Bifunctional Thiourea Catalysis: Hints for the Design of New Organocatalysts. *Chem. Eur. J* **2010**, *16* (2), 450–455.
(117) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59* (1), 96–103.
(118) 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 Cent. Sci.* **2021**, *7* (10), 1622–1637.
(119) Yang, C.; Zhang, E.-G.; Li, X.; Cheng, 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* (22), 6506–6510.
(120) 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* (23), 5488–5497.
(121) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571* (7765), 343–348.
(122) Werth, J.; Sigman, M. S. Connecting and Analyzing Enantioselective Bifunctional Hydrogen Bond Donor Catalysis Using Data Science Tools. *J. Am. Chem. Soc.* **2020**, *142* (38), 16382–16391.
(123) Samha, M. H.; Wahlman, J. L. H.; Read, J. A.; Werth, J.; Jacobsen, E. N.; Sigman, M. S. Exploring Structure–Function Relationships of Aryl Pyrrolidine-Based Hydrogen-Bond Donors in Asymmetric Catalysis Using Data-Driven Techniques. *ACS Catal.* **2022**, *12* (24), 14836–14845.
(124) Melville, J. L.; Andrews, B. I.; Lygo, B.; Hirst, J. D. Computational Screening of Combinatorial Catalyst Libraries. *Chem. Commun.* **2004**, No. 12, 1410–1411.
(125) 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* (4), 971–981.
(126) 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* (11), 4260–4336.
(127) 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* (11), 4337–4357.
(128) Yamaguchi, S. Molecular Field Analysis for Data-Driven Molecular Design in Asymmetric Catalysis. *Org. Biomol. Chem.* **2022**, *20* (31), 6057–6071.
(129) Lustosa, D. M.; Milo, A. Mechanistic Inference from Statistical Models at Different Data-Size Regimes. *ACS Catal.* **2022**, *12* (13), 7886–7906.
(130) Lexa, K. W.; Belyk, K. M.; Henle, J.; Xiang, B.; Sheridan, R. P.; Denmark, S. E.; Ruck, R. T.; Sherer, E. C. Application of Machine Learning and Reaction Optimization for the Iterative Improvement of Enantioselectivity of Cinchona-Derived Phase Transfer Catalysts. *Org. Process Res. Dev.* **2022**, *26* (3), 670–682.
(131) 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). *Chem. Sci.* **2018**, *9* (34), 6922–6927.

(132) Rose, B. T.; Timmerman, J. C.; Bawel, S. A.; Chin, S.; Zhang, H.; Denmark, S. E. High-Level Data Fusion Enables the Chemoinformatically Guided Discovery of Chiral Disulfonimide Catalysts for Atropselective Iodination of 2-Amino-6-Arylpyridines. *J. Am. Chem. Soc.* **2022**, *144* (50), 22950–22964.
(133) Liles, J. P.; Rouget-Virbel, C.; Wahlman, J. L. H.; Rahimoff, R.; Crawford, J. M.; Medlin, A.; O'Connor, V. S.; Li, J.; Roytman, V. A.; Toste, F. D.; Sigman, M. S. Data Science Enables the Development of a New Class of Chiral Phosphoric Acid Catalysts. *Chem* **2023**.
(134) Betinol, I. O.; Lai, J.; Thakur, S.; Reid, J. P. A Data-Driven Workflow for Assigning and Predicting Generality in Asymmetric Catalysis. *J. Am. Chem. Soc.* **2023**, *145* (23), 12870–12883.
(135) 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. *J. Mol. Catal. A: Chem.* **2010**, *319* (1), 114–118.
(136) 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.
(137) Rein, J.; Rozema, S. D.; Langner, O. C.; Zacate, S. B.; Hardy, M. A.; Siu, J. C.; Mercado, B. Q.; Sigman, M. S.; Miller, S. J.; Lin, S. Generality-Oriented Optimization of Enantioselective Aminoxyl Radical Catalysis. *Science* **2023**, *380* (6646), 706–712.
(138) Seumer, J.; Kirschner Solberg Hansen, J.; Brøndsted Nielsen, M.; Jensen, J. H. Computational Evolution Of New Catalysts For The Morita–Baylis–Hillman Reaction. *Angew. Chem. Int. Ed.* **2023**, *62* (18), e202218565.
(139) Bo, C.; Maseras, F.; López, N. The Role of Computational Results Databases in Accelerating the Discovery of Catalysts. *Nat. Catal.* **2018**, *1* (11), 809–810.
(140) Nandy, A.; Duan, C.; Kulik, H. J. Audacity of Huge: Overcoming Challenges of Data Scarcity and Data Quality for Machine Learning in Computational Materials Discovery. *Curr. Opin. Chem. Eng.* **2022**, *36*, 100778.
(141) McNally, A.; Prier, C. K.; MacMillan, D. W. C. Discovery of an Alpha-Amino C-H Arylation Reaction Using the Strategy of Accelerated Serendipity. *Science* **2011**, *334* (6059), 1114–1117.
(142) Iribarren, I.; Trujillo, C. Improving Phase-Transfer Catalysis by Enhancing Non-Covalent Interactions. *Phys. Chem. Chem. Phys.* **2020**, *22* (37), 21015–21021.
(143) Houk, K. N.; Liu, F. Holy Grails for Computational Organic Chemistry and Biochemistry. *Acc. Chem. Res.* **2017**, *50* (3), 539–543.
(144) 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* (10), 872–879.
(145) Nandy, A.; Duan, C.; Taylor, M. G.; Liu, F.; Steeves, A. H.; Kulik, H. J. Computational Discovery of Transition-Metal Complexes: From High-Throughput Screening to Machine Learning. *Chem. Rev.* **2021**, *121* (16), 9927–10000.
(146) Janet, J. P.; Duan, C.; Nandy, A.; Liu, F.; Kulik, H. J. Navigating Transition-Metal Chemical Space: Artificial Intelligence for First-Principles Design. *Acc. Chem. Res.* **2021**, *54* (3), 532–545.
(147) Nandy, A.; Zhu, J.; Janet, J. P.; Duan, C.; Getman, R. B.; Kulik, H. J. Machine Learning Accelerates the Discovery of Design Rules and Exceptions in Stable Metal–Oxo Intermediate Formation. *ACS Catal.* **2019**, *9* (9), 8243–8255.
(148) Gugler, S.; Janet, J. P.; Kulik, H. J. Enumeration of de Novo Inorganic Complexes for Chemical Discovery and Machine Learning. *Mol. Syst. Des. Eng.* **2020**, *5* (1), 139–152.
(149) Liu, F.; Duan, C.; Kulik, H. J. Rapid Detection of Strong Correlation with Machine Learning for Transition-Metal Complex High-Throughput Screening. *J. Phys. Chem. Lett.* **2020**, *11* (19), 8067–8076.

(150) Hueffel, J. A.; Sperger, T.; Funes-Ardoiz, I.; Ward, J. S.; Rissanen, K.; Schoenebeck, F. Accelerated Dinuclear Palladium Catalyst Identification through Unsupervised Machine Learning. *Science* **2021**, *374* (6571), 1134–1140.
(151) Fey, N.; Tsipis, A. C.; Harris, S. E.; Harvey, J. N.; Orpen, A. G.; Mansson, R. A. Development of a Ligand Knowledge Base, Part 1: Computational Descriptors for Phosphorus Donor Ligands. *Chem. Eur. J.* **2006**, *12* (1), 291–302.
(152) Jover, J.; Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Orpen, A. G.; Owen-Smith, G. J. J.; Murray, P.; Hose, D. R. J.; Osborne, R.; Purdie, M. Expansion of the Ligand Knowledge Base for Monodentate P-Donor Ligands (LKB-P). *Organometallics* **2010**, *29* (23), 6245–6258.
(153) Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Murray, P.; Orpen, A. G.; Osborne, R.; Purdie, M. Computational Descriptors for Chelating P,P- and P,N-Donor Ligands1. *Organometallics* **2008**, *27* (7), 1372–1383.
(154) Durand, D. J.; Fey, N. Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019**, *119* (11), 6561–6594.
(155) Durand, D. J.; Fey, N. Building a Toolbox for the Analysis and Prediction of Ligand and Catalyst Effects in Organometallic Catalysis. *Acc. Chem. Res.* **2021**, *54* (4), 837–848.
(156) Fey, N.; Koumi, A.; Malkov, A. V.; Moseley, J. D.; Nguyen, B. N.; Tyler, S. N. G.; Willans, C. E. Mapping the Properties of Bidentate Ligands with Calculated Descriptors (LKB-Bid). *Dalton Trans.* **2020**, *49* (24), 8169–8178.
(157) 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.; Aspuru-Guzik, A. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J. Am. Chem. Soc.* **2022**, *144* (3), 1205–1217.
(158) Foscato, M.; Venkatraman, V.; Occhipinti, G.; Alsberg, B. K.; Jensen, V. R. Automated Building of Organometallic Complexes from 3D Fragments. *J. Chem. Inf. Model.* **2014**, *54* (7), 1919–1931.
(159) Foscato, M.; Occhipinti, G.; Venkatraman, V.; Alsberg, B. K.; Jensen, V. R. Automated Design of Realistic Organometallic Molecules from Fragments. *J. Chem. Inf. Model.* **2014**, *54* (3), 767–780.
(160) Chu, Y.; Heyndrickx, W.; Occhipinti, G.; Jensen, V. R.; Alsberg, B. K. An Evolutionary Algorithm for de Novo Optimization of Functional Transition Metal Compounds. *J. Am. Chem. Soc.* **2012**, *134* (21), 8885–8895.
(161) Thorpe, T. W.; Marshall, J. R.; Harawa, V.; Ruscoe, R. E.; Cuetos, A.; Finnigan, J. D.; Angelastro, A.; Heath, R. S.; Parmeggiani, F.; Charnock, S. J.; Howard, R. M.; Kumar, R.; Daniels, D. S. B.; Grogan, G.; Turner, N. J. Multifunctional Biocatalyst for Conjugate Reduction and Reductive Amination. *Nature* **2022**, *604* (7904), 86–91.
(162) Lapidoth, G.; Khersonsky, O.; Lipsh, R.; Dym, O.; Albeck, S.; Rogotner, S.; Fleishman, S. J. Highly Active Enzymes by Automated Combinatorial Backbone Assembly and Sequence Design. *Nat. Commun.* **2018**, *9* (1), 2780.
(163) Yoon, T. P.; Jacobsen, E. N. Privileged Chiral Catalysts. *Science* **2003**, *299* (5613), 1691–1693.
(164) Strassfeld, D. A.; Algera, R. F.; Wickens, Z. K.; Jacobsen, E. N. A Case Study in Catalyst Generality: Simultaneous, Highly-Enantioselective Brønsted- and Lewis-Acid Mechanisms in Hydrogen-Bond-Donor Catalyzed Oxetane Openings. *J. Am. Chem. Soc.* **2021**, *143* (25), 9585–9594.
(165) Seayad, J.; List, B. Asymmetric Organocatalysis. *Org. Biomol. Chem.* **2005**, *3* (5), 719–724.
(166) Kenny, R.; Liu, F. Trifunctional Organocatalysts: Catalytic Proficiency by Cooperative Activation. *Eur. J. Org. Chem.* **2015**, *2015* (24), 5304–5319.
(167) Kirkpatrick, P.; Ellis, C. Chemical Space. *Nature* **2004**, *432* (7019), 823–823.
(168) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48* (3), 722–730.
(169) Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat. Rev. Drug Discov.* **2005**, *4* (8), 649–663.

(170) Liu, T.; Naderi, M.; Alvin, C.; Mukhopadhyay, S.; Brylinski, M. Break Down in Order To Build Up: Decomposing Small Molecules for Fragment-Based Drug Design with *e*MolFrag. *J. Chem. Inf. Model.* **2017**, *57* (4), 627–631.
(171) Betinol, I. O.; Kuang, Y.; Reid, J. P. Guiding Target Synthesis with Statistical Modeling Tools: A Case Study in Organocatalysis. *Org. Lett.* **2022**, *24* (7), 1429–1433.
(172) Shoja, A.; Zhai, J.; Reid, J. P. Comprehensive Stereochemical Models for Selectivity Prediction in Diverse Chiral Phosphate-Catalyzed Reaction Space. *ACS Catal.* **2021**, *11* (19), 11897–11905.
(173) Tsuji, N.; Sidorov, P.; Zhu, C.; Nagata, Y.; Gimadiev, T.; Varnek, A.; List, B. Predicting Highly Enantioselective Catalysts Using Tunable Fragment Descriptors. *ChemRxiv* **2022**.
(174) Fraux, G.; Cersonsky, R. K.; Ceriotti, M. Chemiscope: Interactive Structure-Property Explorer for Materials and Molecules. *J. Open Source Softw.* **2020**, *5* (51), 2117.
(175) Holland, M. C.; Gilmour, R. Deconstructing Covalent Organocatalysis. *Angew. Chem. Int. Ed.* **2015**, *54* (13), 3862–3871.
(176) Dalko, P. I.; Moisan, L. In the Golden Age of Organocatalysis. *Angew. Chem. Int. Ed.* **2004**, *43* (39), 5138–5175.
(177) Dalko, P. I.; Moisan, L. Enantioselective Organocatalysis. *Angew. Chem. Int. Ed.* **2001**, *40* (20), 3726–3748.
(178) Maji, R.; Mallojjala, S. C.; Wheeler, S. E. Chiral Phosphoric Acid Catalysis: From Numbers to Insights. *Chem. Soc. Rev.* **2018**, *47* (4), 1142–1158.
(179) Enders, D.; Niemeier, O.; Henseler, A. Organocatalysis by N-Heterocyclic Carbenes. *Chem. Rev.* **2007**, *107* (12), 5606–5655.
(180) Wong, O. A.; Shi, Y. Organocatalytic Oxidation. Asymmetric Epoxidation of Olefins Catalyzed by Chiral Ketones and Iminium Salts. *Chem. Rev.* **2008**, *108* (9), 3958–3987.
(181) McGarrigle, E. M.; Myers, E. L.; Illa, O.; Shaw, M. A.; Riches, S. L.; Aggarwal, V. K. Chalcogenides as Organocatalysts. *Chem. Rev.* **2007**, *107* (12), 5841–5883.
(182) Marcelli, T.; Hiemstra, H. Cinchona Alkaloids in Asymmetric Organocatalysis. *Synthesis* **2010**, No. 8, 1229–1279.
(183) Benaglia, M.; Rossi, S. Chiral Phosphine Oxides in Present-Day Organocatalysis. *Org. Biomol. Chem.* **2010**, *8* (17), 3824–3830.
(184) Liu, X.; Lin, L.; Feng, X. Chiral N,N′-Dioxides: New Ligands and Organocatalysts for Catalytic Asymmetric Reactions. *Acc. Chem. Res.* **2011**, *44* (8), 574–587.
(185) Wei, Y.; Shi, M. Applications of Chiral Phosphine-Based Organocatalysts in Catalytic Asymmetric Reactions. *Chem. Asian J.* **2014**, *9* (10), 2720–2734.
(186) Yang, Q.; Li, Y.; Yang, J.-D.; Liu, Y.; Zhang, L.; Luo, S.; Cheng, J.-P. Holistic Prediction of the PKa in Diverse Solvents Based on a Machine-Learning Approach. *Angew. Chem. Int. Ed.* **2020**, *59* (43), 19282–19291.
(187) Yang, C.; Xue, X.-S.; Li, X.; Cheng, J.-P. Computational Study on the Acidic Constants of Chiral Brønsted Acids in Dimethyl Sulfoxide. *J. Org. Chem.* **2014**, *79* (10), 4340–4351.
(188) Christ, P.; Lindsay, A. G.; Vormittag, S. S.; Neudörfl, J.-M.; Berkessel, A.; O'Donoghue, A. C. PKa Values of Chiral Brønsted Acid Catalysts: Phosphoric Acids/Amides, Sulfonyl/Sulfuryl Imides, and Perfluorinated TADDOLs (TEFDDOLs). *Chem. Eur. J.* **2011**, *17* (31), 8524–8528.
(189) Walvoord, R. R.; Huynh, P. N. H.; Kozlowski, M. C. Quantification of Electrophilic Activation by Hydrogen-Bonding Organocatalysts. *J. Am. Chem. Soc.* **2014**, *136* (45), 16055–16065.
(190) Moyano, A. Activation Modes In Asymmetric Organocatalysis. In *Stereoselective Organocatalysis*; John Wiley & Sons, Ltd, 2013; pp 11–80.
(191) Dalko, P. I. Asymmetric Organocatalysis: A New Stream in Organic Synthesis. In *Enantioselective Organocatalysis*; John Wiley & Sons, Ltd, 2007; pp 1–17.
(192) Chi, Y. R. Comprehensive Enantioselective Organocatalysis. Edited by Peter I. Dalko. *Angew. Chem. Int. Ed.* **2014**, *53* (27), 6858–6858.
(193) List, B. *Asymmetric Organocatalysis*; Top. Curr. Chem., volume 291; Springer, 2009.

(194) Pihko, P. Introduction. In *Hydrogen Bonding in Organic Synthesis*; John Wiley & Sons, Ltd, 2009; pp 1–4.
(195) Asymmetric Organocatalysis. *ChemFiles, Sigma-Aldrich* **2006**, *6* (4), 1–16.
(196) Organocatalysis. *ChemFiles, Sigma-Aldrich* **2007**, *7* (9), 1–24.
(197) Groom, C. R.; Allen, F. H. The Cambridge Structural Database in Retrospect and Prospect. *Angew. Chem. Int. Ed.* **2014**, *53* (3), 662–671.
(198) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. B* **2016**, *72* (2), 171–179.
(199) Gražulis, S.; Chateigner, D.; Downs, R. T.; Yokochi, A. F. T.; Quirós, M.; Lutterotti, L.; Manakova, E.; Butkus, J.; Moeck, P.; Le Bail, A. Crystallography Open Database – an Open-Access Collection of Crystal Structures. *J. Appl. Crystallogr.* **2009**, *42* (4), 726–729.
(200) Vela, S.; Laplaza, R.; Cho, Y.; Corminboeuf, C. Cell2mol: Encoding Chemistry to Interpret Crystallographic Data. *npj Comput. Mater.* **2022**, *8* (1), 188.
(201) Garuti, L.; Roberti, M.; Bottegoni, G.; Ferraro, M. Diaryl Urea: A Privileged Structure in Anticancer Agents. *Curr. Med. Chem.* **2016**, *23* (15), 1528–1548.
(202) Anil, S. M.; Rajeev, N.; Kiran, K. R.; Swaroop, T. R.; Mallesha, N.; Shobith, R.; Sadashiva, M. P. Multi-Pharmacophore Approach to Bio-Therapeutics: Piperazine Bridged Pseudo-Peptidic Urea/Thiourea Derivatives as Anti-Oxidant Agents. *Int. J. Pept. Res. Ther.* **2020**, *26* (1), 151–158.
(203) Azeem, S.; Ataf Ali, A.; Ashfaq Mahmood, Q.; Amin, B. Thiourea Derivatives in Drug Design and Medicinal Chemistry: A Short Review. *J. Drug Des. Med. Chem.* **2016**, *2* (1), 10–20.
(204) Bregović, V. B.; Basarić, N.; Mlinarić-Majerski, K. Anion Binding with Urea and Thiourea Derivatives. *Coord. Chem. Rev.* **2015**, *295*, 80–124.
(205) Xu, L.-W.; Luo, J.; Lu, Y. Asymmetric Catalysis with Chiral Primary Amine-Based Organocatalysts. *Chem. Commun.* **2009**, No. 14, 1807–1821.
(206) Li Petri, G.; Raimondi, M. V.; Spanò, V.; Holl, R.; Barraja, P.; Montalbano, A. Pyrrolidine in Drug Discovery: A Versatile Scaffold for Novel Biologically Active Compounds. *Top. Curr. Chem.* **2021**, *379* (5), 34.
(207) Maaten, L. van der; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9* (86), 2579–2605.
(208) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole von Lilienfeld, O. FCHL Revisited: Faster and More Accurate Quantum Machine Learning. *J. Chem. Phys.* **2020**, *152* (4), 044107.
(209) Nguyen, T. N.; Chen, P.-A.; Setthakarn, K.; May, J. A. Chiral Diol-Based Organocatalysts in Enantioselective Reactions. *Molecules* **2018**, *23* (9), 2317.
(210) Parmar, D.; Sugiono, E.; Raja, S.; Rueping, M. Complete Field Guide to Asymmetric BINOL-Phosphate Derived Brønsted Acid and Metal Catalysis: History and Classification by Mode of Activation; Brønsted Acidity, Hydrogen Bonding, Ion Pairing, and Metal Phosphates. *Chem. Rev.* **2014**, *114* (18), 9047–9153.
(211) Shu, L.; Shi, Y. An Efficient Ketone-Catalyzed Epoxidation Using Hydrogen Peroxide as Oxidant. *J. Org. Chem.* **2000**, *65* (25), 8807–8810.
(212) Denmark, S. E.; Wu, Z. The Development of Chiral, Nonracemic Dioxiranes for the Catalytic, Enantioselective Epoxidation of Alkenes. *Synlett* **2000**, *1999* (Sup. 1), 847–859.
(213) Formica, M.; Rozsar, D.; Su, G.; Farley, A. J. M.; Dixon, D. J. Bifunctional Iminophosphorane Superbase Catalysis: Applications in Organic Synthesis. *Acc. Chem. Res.* **2020**, *53* (10), 2235–2247.
(214) Lee, B.; Yoo, J.; Kang, K. Predicting the Chemical Reactivity of Organic Materials Using a Machine-Learning Approach. *Chem. Sci.* **2020**, *11* (30), 7813–7822.
(215) Patureau, F. W.; Worch, C.; Siegler, M. A.; Spek, A. L.; Bolm, C.; Reek, J. N. H. SIAPhos: Phosphorylated Sulfonimidamides and Their Use in Iridium-Catalyzed

Asymmetric Hydrogenations of Sterically Hindered Cyclic Enamides. *Adv. Synth. Catal.* **2012**, *354* (1), 59–64.
(216) Xiang, B.; Belyk, K. M.; Reamer, R. A.; Yasuda, N. Discovery and Application of Doubly Quaternized Cinchona-Alkaloid-Based Phase-Transfer Catalysts. *Angew. Chem. Int. Ed.* **2014**, *53* (32), 8375–8378.
(217) Flanigan, D. M.; Romanov-Michailidis, F.; White, N. A.; Rovis, T. Organocatalytic Reactions Enabled by N-Heterocyclic Carbenes. *Chem. Rev.* **2015**, *115* (17), 9307–9387.
(218) Mood, A.; Tavakoli, M.; Gutman, E.; Kadish, D.; Baldi, P.; Van Vranken, D. L. Methyl Anion Affinities of the Canonical Organic Functional Groups. *J. Org. Chem.* **2020**, *85* (6), 4096–4102.
(219) Kadish, D.; Mood, A. D.; Tavakoli, M.; Gutman, E. S.; Baldi, P.; Van Vranken, D. L. Methyl Cation Affinities of Canonical Organic Functional Groups. *J. Org. Chem.* **2021**, *86* (5), 3721–3729.
(220) Kaupmees, K.; Tolstoluzhsky, N.; Raja, S.; Rueping, M.; Leito, I. On the Acidity and Reactivity of Highly Effective Chiral Brønsted Acid Catalysts: Establishment of an Acidity Scale. *Angew. Chem. Int. Ed.* **2013**, *52* (44), 11569–11572.
(221) Jakab, G.; Tancon, C.; Zhang, Z.; Lippert, K. M.; Schreiner, P. R. (Thio)Urea Organocatalyst Equilibrium Acidities in DMSO. *Org. Lett.* **2012**, *14* (7), 1724–1727.
(222) Ni, X.; Li, X.; Wang, Z.; Cheng, J.-P. Squaramide Equilibrium Acidities in DMSO. *Org. Lett.* **2014**, *16* (6), 1786–1789.
(223) Li, Z.; Li, X.; Ni, X.; Cheng, J.-P. Equilibrium Acidities of Proline Derived Organocatalysts in DMSO. *Org. Lett.* **2015**, *17* (5), 1196–1199.
(224) Li, Y.; Zhang, L.; Luo, S. Bond Energies of Enamines. *ACS Omega* **2022**, *7* (7), 6354–6374.
(225) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.* **2015**, *55* (2), 460–473.
(226) Enders, D.; Balensiefer, T. Nucleophilic Carbenes in Asymmetric Organocatalysis. *Acc. Chem. Res.* **2004**, *37* (8), 534–541.
(227) Hopkinson, M. N.; Richter, C.; Schedler, M.; Glorius, F. An Overview of N-Heterocyclic Carbenes. *Nature* **2014**, *510* (7506), 485–496.
(228) Biju, A. T.; Kuhl, N.; Glorius, F. Extending NHC-Catalysis: Coupling Aldehydes with Unconventional Reaction Partners. *Acc. Chem. Res.* **2011**, *44* (11), 1182–1195.
(229) Ryan, S. J.; Candish, L.; Lupton, D. W. Acyl Anion Free N-Heterocyclic Carbene Organocatalysis. *Chem. Soc. Rev.* **2013**, *42* (12), 4906–4917.
(230) 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. *J. Am. Chem. Soc.* **2023**, *145* (1), 110–121.
(231) Clavier, H.; Nolan, S. P. Percent Buried Volume for Phosphine and N-Heterocyclic Carbene Ligands: Steric Properties in Organometallic Chemistry. *Chem. Commun.* **2010**, *46* (6), 841–861.
(232) Gómez-Suárez, A.; Nelson, D. J.; Nolan, S. P. Quantifying and Understanding the Steric Properties of N-Heterocyclic Carbenes. *Chem. Commun.* **2017**, *53* (18), 2650–2660.
(233) Li, Z.; Li, X.; Cheng, J.-P. An Acidity Scale of Triazolium-Based NHC Precursors in DMSO. *J. Org. Chem.* **2017**, *82* (18), 9675–9681.
(234) Chen, X.-Y.; Gao, Z.-H.; Ye, S. Bifunctional N-Heterocyclic Carbenes Derived from l-Pyroglutamic Acid and Their Applications in Enantioselective Organocatalysis. *Acc. Chem. Res.* **2020**, *53* (3), 690–702.
(235) Zhang, Y.-R.; He, L.; Wu, X.; Shao, P.-L.; Ye, S. Chiral N-Heterocyclic Carbene Catalyzed Staudinger Reaction of Ketenes with Imines: Highly Enantioselective Synthesis of N-Boc β-Lactams. *Org. Lett.* **2008**, *10* (2), 277–280.

(236) He, L.; Zhang, Y.-R.; Huang, X.-L.; Ye, S. Chiral Bifunctional N-Heterocyclic Carbenes: Synthesis and Application in the Aza-Morita-Baylis-Hillman Reaction. *Synthesis* **2008**, *2008* (17), 2825–2829.
(237) Kerr, M. S.; Read de Alaniz, J.; Rovis, T. A Highly Enantioselective Catalytic Intramolecular Stetter Reaction. *J. Am. Chem. Soc.* **2002**, *124* (35), 10298–10299.
(238) He, M.; Struble, J. R.; Bode, J. W. Highly Enantioselective Azadiene Diels−Alder Reactions Catalyzed by Chiral N-Heterocyclic Carbenes. *J. Am. Chem. Soc.* **2006**, *128* (26), 8418–8420.
(239) Wang, N.; Xu, J.; Lee, J. K. The Importance of N-Heterocyclic Carbene Basicity in Organocatalysis. *Org. Biomol. Chem.* **2018**, *16* (37), 8230–8244.
(240) Li, Z.; Li, X.; Cheng, J.-P. Recent Progress in Equilibrium Acidity Studies of Organocatalysts. *Synlett* **2019**, *30* (17), 1940–1949.
(241) Gadekar, S. C.; Dhayalan, V.; Nandi, A.; Zak, I. L.; Mizrachi, M. S.; Kozuch, S.; Milo, A. Rerouting the Organocatalytic Benzoin Reaction toward Aldehyde Deuteration. *ACS Catal.* **2021**, *11* (23), 14561–14569.
(242) Niu, Y.; Wang, N.; Muñoz, A.; Xu, J.; Zeng, H.; Rovis, T.; Lee, J. K. Experimental and Computational Gas Phase Acidities of Conjugate Acids of Triazolylidene Carbenes: Rationalizing Subtle Electronic Effects. *J. Am. Chem. Soc.* **2017**, *139* (42), 14917–14930.
(243) Ho, J.; Zwicker, V. E.; Yuen, K. K. Y.; Jolliffe, K. A. Quantum Chemical Prediction of Equilibrium Acidities of Ureas, Deltamides, Squaramides, and Croconamides. *J. Org. Chem.* **2017**, *82* (19), 10732–10736.
(244) Zwicker, V. E.; Yuen, K. K. Y.; Smith, D. G.; Ho, J.; Qin, L.; Turner, P.; Jolliffe, K. A. Deltamides and Croconamides: Expanding the Range of Dual H-Bond Donors for Selective Anion Recognition. *Chem. Eur. J* **2018**, *24* (5), 1140–1150.
(245) Zhang, X.; Liu, S.; Li, X.; Yan, M.; Chan, A. S. C. Highly Enantioselective Conjugate Addition of Aldehydes to Nitroolefins Catalyzed by Chiral Bifunctional Sulfamides. *Chem. Commun.* **2009**, No. 7, 833–835.
(246) Sandler, I.; Larik, F. A.; Mallo, N.; Beves, J. E.; Ho, J. Anion Binding Affinity: Acidity versus Conformational Effects. *J. Org. Chem.* **2020**, *85* (12), 8074–8084.
(247) Wittkopp, A.; Schreiner, P. R. Metal-Free, Noncovalent Catalysis of Diels–Alder Reactions by Neutral Hydrogen Bond Donors in Organic Solvents and in Water. *Chem. Eur. J.* **2003**, *9* (2), 407–414.
(248) Luchini, G.; Ascough, D. M. H.; Alegre-Requena, J. V.; Gouverneur, V.; Paton, R. S. Data-Mining the Diaryl(Thio)Urea Conformational Landscape: Understanding the Contrasting Behavior of Ureas and Thioureas with Quantum Chemistry. *Tetrahedron* **2019**, *75* (6), 697–702.
(249) Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Montgomery, J.; Vreven, T.; Kudin, K.; Burant, J.; Millam, J.; Iyengar, S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J.; Hratchian, H.; Cross, J.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R.; Yazyev, O.; Austin, A.; Cammi, R.; Pomelli, C.; Ochterski, J.; Ayala, P.; Morokuma, K.; Voth, G.; Salvador, P.; Dannenberg, J.; Zakrzewski, V.; Dapprich, S.; Daniels, A.; Strain, M.; Farkas, O.; Malick, D.; Rabuck, A.; Raghavachari, K.; Foresman, J.; Ortiz, J.; Cui, Q.; Baboul, A.; Clifford, S.; Cioslowski, J.; Stefanov, B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R.; Fox, D.; Keith, T.; Laham, A.; Peng, C.; Nanayakkara, A.; Challacombe, M.; Gill, P.; Johnson, B.; Chen, W.; Wong, M.; Gonzalez, C.; Pople, J. Gaussian 16, Revision C.01. **2016**.
(250) Becke, A. D. Density-Functional Thermochemistry. V. Systematic Optimization of Exchange-Correlation Functionals. *J. Chem. Phys.* **1997**, *107* (20), 8554–8560.
(251) Grimme, S. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**, *27* (15), 1787–1799.

(252) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7* (18), 3297–3305.
(253) Chai, J.-D.; Head-Gordon, M. Systematic Optimization of Long-Range Corrected Hybrid Density Functionals. *J. Chem. Phys.* **2008**, *128* (8), 084106.
(254) Haldar, S.; Riplinger, C.; Demoulin, B.; Neese, F.; Izsak, R.; Dutta, A. K. Multilayer Approach to the IP-EOM-DLPNO-CCSD Method: Theory, Implementation, and Application. *J. Chem. Theory Comput.* **2019**, *15* (4), 2265–2277.
(255) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA Quantum Chemistry Program Package. *J. Chem. Phys.* **2020**, *152* (22), 224108.
(256) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. Efficient, Approximate and Parallel Hartree–Fock and Hybrid DFT Calculations. A 'Chain-of-Spheres' Algorithm for the Hartree–Fock Exchange. *Chem. Phys.* **2009**, *356* (1), 98–109.
(257) 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. *J. Chem. Theory Comput.* **2019**, *15* (3), 1652–1671.
(258) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (2), 97–101.
(259) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3* (1), 33.
(260) Gupta, K.; Roy, D. R.; Subramanian, V.; Chattaraj, P. K. Are Strong Brønsted Acids Necessarily Strong Lewis Acids? *J. Mol. Struct. - THEOCHEM* **2007**, *812* (1), 13–24.
(261) Koopmans, T. Über Die Zuordnung von Wellenfunktionen Und Eigenwerten Zu Den Einzelnen Elektronen Eines Atoms. *Physica* **1934**, *1* (1), 104–113.
(262) Domingo, L. R.; Ríos-Gutiérrez, M.; Pérez, P. Applications of the Conceptual Density Functional Theory Indices to Organic Chemistry Reactivity. *Molecules* **2016**, *21* (6), 748.
(263) Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual Density Functional Theory. *Chem. Rev.* **2003**, *103* (5), 1793–1874.
(264) Domingo, L. R.; Pérez, P. The Nucleophilicity N Index in Organic Chemistry. *Org. Biomol. Chem.* **2011**, *9* (20), 7168–7175.
(265) Domingo, L. R.; Chamorro, E.; Pérez, P. Understanding the Reactivity of Captodative Ethylenes in Polar Cycloaddition Reactions. A Theoretical Study. *J. Org. Chem.* **2008**, *73* (12), 4615–4624.
(266) Chauhan, P.; Mahajan, S.; Kaya, U.; Hack, D.; Enders, D. Bifunctional Amine-Squaramides: Powerful Hydrogen-Bonding Organocatalysts for Asymmetric Domino/Cascade Reactions. *Adv. Synth. Catal.* **2015**, *357* (2–3), 253–281.
(267) Fang, X.; Wang, C.-J. Recent Advances in Asymmetric Organocatalysis Mediated by Bifunctional Amine–Thioureas Bearing Multiple Hydrogen-Bonding Donors. *Chem. Commun.* **2015**, *51* (7), 1185–1197.
(268) Serdyuk, O. V.; Heckel, C. M.; Tsogoeva, S. B. Bifunctional Primary Amine-Thioureas in Asymmetric Organocatalysis. *Org. Biomol. Chem.* **2013**, *11* (41), 7051–7071.
(269) Roca-López, D.; Uria, U.; Reyes, E.; Carrillo, L.; Jørgensen, K. A.; Vicario, J. L.; Merino, P. Mechanistic Insights into the Mode of Action of Bifunctional Pyrrolidine-Squaramide-Derived Organocatalysts. *Chem. Eur. J.* **2016**, *22* (3), 884–889.
(270) Limnios, D.; Kokotos, C. G. Chapter 19 Ureas and Thioureas as Asymmetric Organocatalysts. In *Sustainable Catalysis: Without Metals or Other Endangered Elements, Part 2*; The Royal Society of Chemistry, 2016; pp 196–255.
(271) Phillips, A. M. F.; Prechtl, M. H. G.; Pombeiro, A. J. L. Non-Covalent Interactions in Enantioselective Organocatalysis: Theoretical and Mechanistic Studies of Reactions Mediated by Dual H-Bond Donors, Bifunctional Squaramides, Thioureas and Related Catalysts. *Catalysts* **2021**, *11* (5), 569.
(272) Fonseca, M. H.; List, B. Combinatorial Chemistry and High-Throughput Screening for the Discovery of Organocatalysts. *Curr. Opin. Chem. Biol.* **2004**, *8* (3), 319–326.

(273) Kim, H.; Gerosa, G.; Aronow, J.; Kasaplar, P.; Ouyang, J.; Lingnau, J. B.; Guerry, P.; Farès, C.; List, B. A Multi-Substrate Screening Approach for the Identification of a Broadly Applicable Diels–Alder Catalyst. *Nat. Commun.* **2019**, *10* (1), 770.
(274) 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* (6), 543–551.
(275) Kuang, Y.; Lai, J.; Reid, J. P. Transferrable Selectivity Profiles Enable Prediction in Synergistic Catalyst Space. *Chem. Sci.* **2023**, *14* (7), 1885–1895.
(276) 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* (23), 5488–5497.
(277) Golec, J. C.; Carter, E. M.; Ward, J. W.; Whittingham, W. G.; Simón, L.; Paton, R. S.; Dixon, D. J. BIMP-Catalyzed 1,3-Prototropic Shift for the Highly Enantioselective Synthesis of Conjugated Cyclohexenones. *Angew. Chem. Int. Ed.* **2020**, *59* (40), 17417–17422.
(278) Crawford, J. M.; 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* (3), 868–871.
(279) 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.
(280) 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.
(281) Mougel, V.; Santiago, C. B.; Zhizhko, P. A.; Bess, E. N.; Varga, J.; Frater, G.; Sigman, M. S.; Copéret, C. Quantitatively Analyzing Metathesis Catalyst Activity and Structural Features in Silica-Supported Tungsten Imido–Alkylidene Complexes. *J. Am. Chem. Soc.* **2015**, *137* (20), 6699–6704.
(282) Niemeyer, Z. L.; Milo, A.; Hickey, D. P.; Sigman, M. S. Parameterization of Phosphine Ligands Reveals Mechanistic Pathways and Predicts Reaction Outcomes. *Nature Chem.* **2016**, *8* (6), 610–617.
(283) Wheeler, S. E. Understanding Substituent Effects in Noncovalent Interactions Involving Aromatic Rings. *Acc. Chem. Res.* **2013**, *46* (4), 1029–1038.
(284) Díaz-Salazar, H.; Jiménez, E. I.; Vallejo Narváez, W. E.; Rocha-Rinza, T.; Hernández-Rodríguez, M. Bifunctional Squaramides with Benzyl-like Fragments: Analysis of CH···π Interactions by a Multivariate Linear Regression Model and Quantum Chemical Topology. *Org. Chem. Front.* **2021**, *8* (13), 3217–3227.
(285) Seguin, T. J.; Wheeler, S. E. Electrostatic Basis for Enantioselective Brønsted-Acid-Catalyzed Asymmetric Ring Openings of Meso-Epoxides. *ACS Catal.* **2016**, *6* (4), 2681–2688.
(286) Werth, J.; Sigman, M. S. Connecting and Analyzing Enantioselective Bifunctional Hydrogen Bond Donor Catalysis Using Data Science Tools. *J. Am. Chem. Soc.* **2020**.
(287) Wodrich, M. D.; Sawatlon, B.; Solel, E.; Kozuch, S.; Corminboeuf, C. Activity-Based Screening of Homogeneous Catalysts through the Rapid Assessment of Theoretically Derived Turnover Frequencies. *ACS Catal.* **2019**, *9* (6), 5716–5725.
(288) Jiang, H.; Rodríguez-Escrich, C.; Johansen, T. K.; Davis, R. L.; Jørgensen, K. A. Organocatalytic Activation of Polycyclic Aromatic Compounds for Asymmetric Diels–Alder Reactions. *Angew. Chem. Int. Ed.* **2012**, *51* (41), 10271–10274.
(289) Lu, T.; Wheeler, S. E. Origin of the Superior Performance of (Thio)Squaramides over (Thio)Ureas in Organocatalysis. *Chem. Eur. J.* **2013**, *19* (45), 15141–15147.
(290) Miertuš, S.; Scrocco, E.; Tomasi, J. Electrostatic Interaction of a Solute with a Continuum. A Direct Utilizaion of AB Initio Molecular Potentials for the Prevision of Solvent Effects. *Chem. Phys.* **1981**, *55* (1), 117–129.

(291) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105* (8), 2999–3094.
(292) Fukui, K. The Path of Chemical Reactions - the IRC Approach. *Acc. Chem. Res.* **1981**, *14* (12), 363–368.
(293) Luchini, G.; Alegre-Requena, J. V.; Funes-Ardoiz, I.; Paton, R. S. GoodVibes: Automated Thermochemistry for Heterogeneous Computational Chemistry Data. *F1000Research* **2020**, *291* (9).
(294) Albrecht, Ł.; Dickmeiss, G.; Acosta, F. C.; Rodríguez-Escrich, C.; Davis, R. L.; Jørgensen, K. A. Asymmetric Organocatalytic Formal [2 + 2]-Cycloadditions via Bifunctional H-Bond Directing Dienamine Catalysis. *J. Am. Chem. Soc.* **2012**, *134* (5), 2543–2546.
(295) Albrecht, Ł.; Dickmeiss, G.; Weise, C. F.; Rodríguez-Escrich, C.; Jørgensen, K. A. Dienamine-Mediated Inverse-Electron-Demand Hetero-Diels–Alder Reaction by Using an Enantioselective H-Bond-Directing Strategy. *Angew. Chem. Int. Ed.* **2012**, *51* (52), 13109–13113.
(296) Albrecht, Ł.; Cruz Acosta, F.; Fraile, A.; Albrecht, A.; Christensen, J.; Jørgensen, K. A. Enantioselective H-Bond-Directing Approach for Trienamine-Mediated Reactions in Asymmetric Synthesis. *Angew. Chem. Int. Ed.* **2012**, *51* (36), 9088–9092.
(297) Albrecht, Ł.; Gómez, C. V.; Jacobsen, C. B.; Jørgensen, K. A. 1,4-Naphthoquinones in H-Bond-Directed Trienamine-Mediated Strategies. *Org. Lett.* **2013**, *15* (12), 3010–3013.
(298) Weise, C. F.; Lauridsen, V. H.; Rambo, R. S.; Iversen, E. H.; Olsen, M.-L.; Jørgensen, K. A. Organocatalytic Access to Enantioenriched Dihydropyran Phosphonates via an Inverse-Electron-Demand Hetero-Diels–Alder Reaction. *J. Org. Chem.* **2014**, *79* (8), 3537–3546.
(299) Orue, A.; Uria, U.; Reyes, E.; Carrillo, L.; Vicario, J. L. Catalytic Enantioselective [5+2] Cycloaddition between Oxidopyrylium Ylides and Enals under Dienamine Activation. *Angew. Chem. Int. Ed.* **2015**, *54* (10), 3043–3046.
(300) Sawatlon, B.; Wodrich, M. D.; Corminboeuf, C. Probing Substrate Scope with Molecular Volcanoes. *Org. Lett.* **2020**, *22*, 7936–7941.
(301) Kozuch, S.; Shaik, S. How to Conceptualize Catalytic Cycles? The Energetic Span Model. *Acc. Chem. Res.* **2011**, *44* (2), 101–110.
(302) Schnitzer, T.; Möhler, J. S.; Wennemers, H. Effect of the Enamine Pyramidalization Direction on the Reactivity of Secondary Amine Organocatalysts. *Chem. Sci.* **2020**, *11* (7), 1943–1947.
(303) Pihko, P. M.; Majander, I.; Erkkila, A. Enamine Catalysis. In *Asymmetric Organocatalysis*; List, B, Ed.; Topics in Current Chemistry-Series; 2009; Vol. 291, pp 29–75.
(304) Jeppesen, A.; Nielsen, B. E.; Larsen, D.; Akselsen, O. M.; Sølling, T. I.; Brock-Nannestad, T.; Pittelkow, M. Croconamides: A New Dual Hydrogen Bond Donating Motif for Anion Recognition and Organocatalysis. *Org. Biomol. Chem.* **2017**, *15* (13), 2784–2790.
(305) Aleman, J.; Parra, A.; Jiang, H.; Jorgensen, K. A. Squaramides: Bridging from Molecular Recognition to Bifunctional Organocatalysis. *Chem. Eur. J* **2011**, *17* (25), 6890–6899.
(306) Mukherjee, S.; Yang, J. W.; Hoffmann, S.; List, B. Asymmetric Enamine Catalysis. *Chem. Rev.* **2007**, *107* (12), 5471–5569.
(307) Castro-Alvarez, A.; Carneros, H.; Costa, A. M.; Vilarrasa, J. Computer-Aided Insight into the Relative Stability of Enamines. *Synthesis* **2017**, *49* (24), 5285–5306.
(308) Lippert, K. M.; Hof, K.; Gerbig, D.; Ley, D.; Hausmann, H.; Guenther, S.; Schreiner, P. R. Hydrogen-Bonding Thiourea Organocatalysts: The Privileged 3,5-Bis(Trifluoromethyl)Phenyl Group. *Eur. J. Org. Chem.* **2012**, *2012* (30), 5919–5927.
(309) Auvil, T. J.; Schafer, A. G.; Mattson, A. E. Design Strategies for Enhanced Hydrogen-Bond Donor Catalysts. *Eur. J. Org. Chem.* **2014**, *2014* (13), 2633–2646.
(310) Wåhlander, J.; Amedjkouh, M.; Balcells, D. A DFT Perspective on Diels–Alder Organocatalysts Based on Substituted Phosphoramides. *Eur. J. Org. Chem.* **2019**, *2019* (2–3), 442–450.

(311) Klare, H.; Neudoerfl, J. M.; Goldfuss, B. New Hydrogen-Bonding Organocatalysts: Chiral Cyclophosphazanes and Phosphorus Amides as Catalysts for Asymmetric Michael Additions. *Beilstein J. Org. Chem.* **2014**, *10*, 224–236.
(312) Inokuma, T.; Furukawa, M.; Uno, T.; Suzuki, Y.; Yoshida, K.; Yano, Y.; Matsuzaki, K.; Takemoto, Y. Bifunctional Hydrogen-Bond Donors That Bear a Quinazoline or Benzothiadiazine Skeleton for Asymmetric Organocatalysis. *Chem. Eur. J* **2011**, *17* (37), 10470–10477.
(313) Zielińska-Błajet, M.; Najdek, J. Novel Selenoureas Based on Cinchona Alkaloid Skeleton: Synthesis and Catalytic Investigations. *Materials* **2021**, *14* (3).
(314) Takaishi, K.; Okuyama, T.; Kadosaki, S.; Uchiyama, M.; Ema, T. Hemisquaramide Tweezers as Organocatalysts: Synthesis of Cyclic Carbonates from Epoxides and CO2. *Org. Lett.* **2019**, *21* (5), 1397–1401.
(315) Lin, Y.; Hirschi, W. J.; Kunadia, A.; Paul, A.; Ghiviriga, I.; Abboud, K. A.; Karugu, R. W.; Vetticatt, M. J.; Hirschi, J. S.; Seidel, D. A Selenourea-Thiourea Brønsted Acid Catalyst Facilitates Asymmetric Conjugate Additions of Amines to α,β-Unsaturated Esters. *J. Am. Chem. Soc.* **2020**, *142* (12), 5627–5635.
(316) Bian, G.; Yang, S.; Huang, H.; Zong, H.; Song, L.; Fan, H.; Sun, X. Chirality Sensing of Tertiary Alcohols by a Novel Strong Hydrogen-Bonding Donor – Selenourea. *Chem. Sci.* **2016**, *7* (2), 932–938.
(317) Kobayashi, Y.; Taniguchi, Y.; Hayama, N.; Inokuma, T.; Takemoto, Y. A Powerful Hydrogen-Bond-Donating Organocatalyst for the Enantioselective Intramolecular Oxa-Michael Reaction of α,β-Unsaturated Amides and Esters. *Angew. Chem. Int. Ed.* **2013**, *52* (42), 11114–11118.
(318) Nanjo, T.; Zhang, X.; Tokuhiro, Y.; Takemoto, Y. Divergent and Scalable Synthesis of α-Hydroxy/Keto-β-Amino Acid Analogues by the Catalytic Enantioselective Addition of Glyoxylate Cyanohydrin to Imines. *ACS Catalysis* **2019**, *9* (11), 10087–10092.
(319) Christensen, M.; Yunker, L. P. E.; Adedeji, F.; Häse, F.; Roch, L. M.; Gensch, T.; dos Passos Gomes, G.; Zepel, T.; Sigman, M. S.; Aspuru-Guzik, A.; Hein, J. E. Data-Science Driven Autonomous Process Optimization. *Commun. Chem.* **2021**, *4* (1), 112.
(320) Clerc, F.; Lengliz, M.; Farrusseng, D.; Mirodatos, C.; Pereira, S. R. M.; Rakotomalala, R. Library Design Using Genetic Algorithms for Catalyst Discovery and Optimization. *Rev. Sci. Instrum.* **2005**, *76* (6), 062208.
(321) Foscato, M.; Venkatraman, V.; Jensen, V. R. DENOPTIM: Software for Computational de Novo Design of Organic and Inorganic Molecules. *J. Chem. Inf. Model.* **2019**, *59* (10), 4077–4082.
(322) Renom-Carrasco, M.; Lefort, L. Ligand Libraries for High Throughput Screening of Homogeneous Catalysts. *Chem. Soc. Rev.* **2018**, *47* (13), 5038–5060.
(323) Gensch, T.; Smith, S. R.; Colacot, T. J.; Timsina, Y. N.; Xu, G.; Glasspoole, B. W.; Sigman, M. S. Design and Application of a Screening Set for Monophosphine Ligands in Cross-Coupling. *ACS Catal.* **2022**, *12* (13), 7773–7780.
(324) Soyemi, A.; Szilvási, T. Trends in Computational Molecular Catalyst Design. *Dalton Trans.* **2021**, *50* (30), 10325–10339.
(325) Jensen, J. H. A Graph-Based Genetic Algorithm and Generative Model/Monte Carlo Tree Search for the Exploration of Chemical Space. *Chem. Sci.* **2019**, *10* (12), 3567–3572.
(326) Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *J. Phys. Chem. Lett.* **2018**, *9* (5), 1064–1071.
(327) Stumpfe, D.; Hu, H.; Bajorath, J. Evolving Concept of Activity Cliffs. *ACS Omega* **2019**, *4* (11), 14360–14368.
(328) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36.
(329) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach. Learn.: Sci. Technol.* **2020**, *1* (4), 045024.

(330) Nigam, A.; Pollice, R.; Krenn, M.; Gomes, G. dos P.; Aspuru-Guzik, A. Beyond Generative Models: Superfast Traversal, Optimization, Novelty, Exploration and Discovery (STONED) Algorithm for Molecules Using SELFIES. *Chem. Sci.* **2021**, *12* (20), 7079–7090.
(331) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. MolSimplify: A Toolkit for Automating Discovery in Inorganic Chemistry. *J. Comput. Chem.* **2016**, *37* (22), 2106–2117.
(332) Gao, W.; Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* **2020**, *60* (12), 5714–5723.
(333) Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chem. Sci.* **2018**, *9* (35), 7069–7077.
(334) Heinen, S.; von Rudorff, G. F.; von Lilienfeld, O. A. Toward the Design of Chemical Reactions: Machine Learning Barriers of Competing Mechanisms in Reactant Space. *J. Chem. Phys.* **2021**, *155* (6), 064105.
(335) Sun, Q.; Berkelbach, T. C.; Blunt, N. S.; Booth, G. H.; Guo, S.; Li, Z.; Liu, J.; McClain, J. D.; Sayfutyarova, E. R.; Sharma, S.; Wouters, S.; Chan, G. K.-L. PySCF: The Python-Based Simulations of Chemistry Framework. *WIREs Comput. Mol. Sci.* **2018**, *8* (1), e1340.
(336) Sun, Q.; Zhang, X.; Banerjee, S.; Bao, P.; Barbry, M.; Blunt, N. S.; Bogdanov, N. A.; Booth, G. H.; Chen, J.; Cui, Z.-H.; Eriksen, J. J.; Gao, Y.; Guo, S.; Hermann, J.; Hermes, M. R.; Koh, K.; Koval, P.; Lehtola, S.; Li, Z.; Liu, J.; Mardirossian, N.; McClain, J. D.; Motta, M.; Mussard, B.; Pham, H. Q.; Pulkin, A.; Purwanto, W.; Robinson, P. J.; Ronca, E.; Sayfutyarova, E. R.; Scheurer, M.; Schurkus, H. F.; Smith, J. E. T.; Sun, C.; Sun, S.-N.; Upadhyay, S.; Wagner, L. K.; Wang, X.; White, A.; Whitfield, J. D.; Williamson, M. J.; Wouters, S.; Yang, J.; Yu, J. M.; Zhu, T.; Berkelbach, T. C.; Sharma, S.; Sokolov, A. Yu.; Chan, G. K.-L. Recent Developments in the PySCF Program Package. *J. Chem. Phys.* **2020**, *153* (2), 024109.
(337) Huang, B.; von Lilienfeld, O. A. Quantum Machine Learning Using Atom-in-Molecule-Based Fragments Selected on the Fly. *Nat. Chem.* **2020**, *12* (10), 945–951.
(338) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Chimera: Enabling Hierarchy Based Multi-Objective Optimization for Self-Driving Laboratories. *Chem. Sci.* **2018**, *9* (39), 7642–7655.
(339) Ding, C.-H.; Hou, X.-L. Catalytic Asymmetric Propargylation. *Chem. Rev.* **2011**, *111* (3), 1914–1937.
(340) Sepúlveda, D.; Lu, T.; Wheeler, S. E. Performance of DFT Methods and Origin of Stereoselectivity in Bipyridine N,N′-Dioxide Catalyzed Allylation and Propargylation Reactions. *Org. Biomol. Chem.* **2014**, *12* (41), 8346–8353.
(341) Tauer, T. P.; Sherrill, C. D. Beyond the Benzene Dimer: An Investigation of the Additivity of Π−π Interactions. *J. Phys. Chem. A* **2005**, *109* (46), 10475–10478.
(342) Vaganov, V. Yu.; Fukazawa, Y.; Kondratyev, N. S.; Shipilovskikh, S. A.; Wheeler, S. E.; Rubtsov, A. E.; Malkov, A. V. Optimization of Catalyst Structure for Asymmetric Propargylation of Aldehydes with Allenyltrichlorosilane. *Adv. Synth. Catal.* **2020**, *362* (23), 5467–5474.
(343) Thiede, L. A.; Krenn, M.; Nigam, A.; Aspuru-Guzik, A. Curiosity in Exploring Chemical Space: Intrinsic Rewards for Deep Molecular Reinforcement Learning. *arXiv:1909.11655v4* **2020**.
(344) Nigam, A.; Friederich, P.; Krenn, M.; Aspuru-Guzik, A. Augmenting Genetic Algorithms with Deep Neural Networks for Exploring the Chemical Space. *arXiv:1909.11655* **2020**.
(345) Vanhaelen, Q.; Lin, Y.-C.; Zhavoronkov, A. The Advent of Generative Chemistry. *ACS Med. Chem. Lett.* **2020**, *11* (8), 1496–1505.
(346) Le, T. T.; Fu, W.; Moore, J. H. Scaling Tree-Based Automated Machine Learning to Biomedical Big Data with a Feature Set Selector. *Bioinformatics* **2020**, *36* (1), 250–256.

(347) Xu, L.-C.; Zhang, S.-Q.; Li, X.; Tang, M.-J.; Xie, P.-P.; Hong, X. Towards Data-Driven Design of Asymmetric Hydrogenation of Olefins: Database and Hierarchical Learning. *Angew. Chem. Int. Ed.* **2021**, *60* (42), 22804–22811.
(348) Jorner, K.; Tomberg, A.; Bauer, C.; Sköld, C.; Norrby, P.-O. Organic Reactivity from Mechanism to Machine Learning. *Nat. Rev. Chem.* **2021**, *5* (4), 240–255.
(349) Suleimanov, Y. V.; Green, W. H. Automated Discovery of Elementary Chemical Reaction Steps Using Freezing String and Berny Optimization Methods. *J. Chem. Theory Comput.* **2015**, *11* (9), 4248–4259.
(350) Simm, G. N.; Reiher, M. Context-Driven Exploration of Complex Chemical Reaction Networks. *J. Chem. Theory Comput.* **2017**, *13* (12), 6108–6119.
(351) Simm, G. N.; Vaucher, A. C.; Reiher, M. Exploration of Reaction Pathways and Chemical Transformation Networks. *J. Phys. Chem. A* **2019**, *123* (2), 385–399.
(352) Young, T. A.; Silcock, J. J.; Sterling, A. J.; Duarte, F. AutodE: Automated Calculation of Reaction Energy Profiles— Application to Organic and Organometallic Reactions. *Angew. Chem. Int. Ed.* **2021**, *60* (8), 4266–4274.
(353) Shen, Y.; Borowski, J. E.; Hardy, M. A.; Sarpong, R.; Doyle, A. G.; Cernak, T. Automation and Computer-Assisted Planning for Chemical Synthesis. *Nat. Rev. Methods Primers* **2021**, *1* (1), 23.
(354) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, *365* (6453), eaax1566.
(355) Balcells, D.; Clot, E.; Eisenstein, O.; Nova, A.; Perrin, L. Deciphering Selectivity in Organic Reactions: A Multifaceted Problem. *Acc. Chem. Res.* **2016**, *49* (5), 1070–1078.
(356) Sperger, T.; Sanhueza, I. A.; Kalvet, I.; Schoenebeck, F. Computational Studies of Synthetically Relevant Homogeneous Organometallic Catalysis Involving Ni, Pd, Ir, and Rh: An Overview of Commonly Employed DFT Methods and Mechanistic Insights. *Chem. Rev.* **2015**, *115* (17), 9532–9586.
(357) Von Rudorff, G.; Heinen, S.; Bragato, M.; Von Lilienfeld, O. Thousands of Reactants and Transition States for Competing E2 and S(N)2 Reactions. *Mach. Learn.: Sci. Technol.* **2020**, *1* (4).
(358) Bragato, M.; von Rudorff, G. F.; von Lilienfeld, O. A. Data Enhanced Hammett-Equation: Reaction Barriers in Chemical Space. *Chem. Sci.* **2020**, *11* (43), 11859–11868.
(359) Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E. P.; Grzybowski, B. A.; Gambin, A. Predicting the Outcomes of Organic Reactions via Machine Learning: Are Current Descriptors Sufficient? *Sci. Rep.* **2017**, *7* (1), 3582.
(360) Laplaza, R.; Sobez, J.-G.; Wodrich, M. D.; Reiher, M.; Corminboeuf, C. The (Not so) Simple Prediction of Enantioselectivity – a Pipeline for High-Fidelity Computations. *Chem. Sci.* **2022**, *13* (23), 6858–6864.
(361) Harper, K. C.; Sigman, M. S. Three-Dimensional Correlation of Steric and Electronic Free Energy Relationships Guides Asymmetric Propargylation. *Science* **2011**, *333* (6051), 1875–1878.
(362) Milo, A.; Bess, E. N.; Sigman, M. S. Interrogating Selectivity in Catalysis Using Molecular Vibrations. *Nature* **2014**, *507* (7491), 210–214.
(363) Singh, S.; Pareek, M.; Changotra, A.; Banerjee, S.; Bhaskararao, B.; Balamurugan, P.; Sunoj, R. B. A Unified Machine-Learning Protocol for Asymmetric Catalysis as a Proof of Concept Demonstration Using Asymmetric Hydrogenation. *Proc. Natl. Acad. Sci. USA* **2020**, *117* (3), 1339–1345.
(364) Li, X.; Zhang, S.-Q.; Xu, L.-C.; Hong, X. Predicting Regioselectivity in Radical C−H Functionalization of Heterocycles through Machine Learning. *Angew. Chem. Int. Ed.* **2020**, *59* (32), 13253–13259.
(365) Maley, S. M.; Kwon, D.-H.; Rollins, N.; Stanley, J. C.; Sydora, O. L.; Bischof, S. M.; Ess, D. H. Quantum-Mechanical Transition-State Model Combined with Machine Learning

Provides Catalyst Design Features for Selective Cr Olefin Oligomerization. *Chem. Sci.* **2020**, *11* (35), 9665–9674.
(366) Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **2018**, *140* (15), 5004–5008.
(367) Chen, J.; Jiwu, W.; Mingzong, L.; You, T. Calculation on Enantiomeric Excess of Catalytic Asymmetric Reactions of Diethylzinc Addition to Aldehydes with Topological Indices and Artificial Neural Network. *J. Mol. Catal. A: Chem.* **2006**, *258* (1), 191–197.
(368) Qiu, J.; Xie, J.; Su, S.; Gao, Y.; Meng, H.; Yang, Y.; Liao, K. Selective Functionalization of Hindered Meta-C–H Bond of o-Alkylaryl Ketones Promoted by Automation and Deep Learning. *Chem* **2022**.
(369) 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. *Chem. Sci.* **2021**, *12* (6), 2198–2208.
(370) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10* (2), 370–377.
(371) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* **2021**, *590* (7844), 89–96.
(372) Amar, Y.; Schweidtmann, A. M.; Deutsch, P.; Cao, L.; Lapkin, A. Machine Learning and Molecular Descriptors Enable Rational Solvent Selection in Asymmetric Catalysis. *Chem. Sci.* **2019**, *10* (27), 6697–6706.
(373) Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G. A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. *J. Am. Chem. Soc.* **2022**, *144* (43), 19999–20007.
(374) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. Machine Learning Meets Mechanistic Modelling for Accurate Prediction of Experimental Activation Energies. *Chem. Sci.* **2021**, *12* (3), 1163–1175.
(375) Tomberg, A.; Johansson, M. J.; Norrby, P.-O. A Predictive Tool for Electrophilic Aromatic Substitutions Using Machine Learning. *J. Org. Chem.* **2019**, *84* (8), 4695–4703.
(376) Banerjee, S.; Sreenithya, A.; Sunoj, R. B. Machine Learning for Predicting Product Distributions in Catalytic Regioselective Reactions. *Phys. Chem. Chem. Phys.* **2018**, *20* (27), 18311–18318.
(377) Beker, W.; Gajewska, E. P.; Badowski, T.; Grzybowski, B. A. Prediction of Major Regio-, Site-, and Diastereoisomers in Diels–Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors. *Angew. Chem. Int. Ed.* **2019**, *58* (14), 4515–4519.
(378) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108* (5), 058301.
(379) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18* (20), 13754–13769.
(380) von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring Chemical Compound Space with Quantum-Based Machine Learning. *Nat. Rev. Chem.* **2020**, *4* (7), 347–358.
(381) von Lilienfeld, O. A.; Burke, K. Retrospective on a Decade of Machine Learning for Chemical Discovery. *Nat. Commun.* **2020**, *11* (1), 4895.
(382) Denmark, S. E.; Coe, D.; Pratt, N.; Griedel, B. Asymmetric Allylation of Aldehydes with Chiral Lewis-Bases. *J. Org. Chem.* **1994**, *59* (21), 6161–6163.
(383) Denmark, S. E.; Fu, J. P. On the Mechanism of Catalytic, Enantioselective Allylation of Aldehydes with Chlorosilanes and Chiral Lewis Bases. *J. Am. Chem. Soc.* **2000**, *122* (48), 12021–12022.

(384) Denmark, S. E.; Wynn, T. Lewis Base Activation of Lewis Acids: Catalytic Enantioselective Allylation and Propargylation of Aldehydes. *J. Am. Chem. Soc.* **2001**, *123* (25), 6199–6200.
(385) Denmark, S. E.; Beutner, G. L. Lewis Base Catalysis in Organic Synthesis. *Angew. Chem. Int. Ed.* **2008**, *47* (9), 1560–1638.
(386) Marshall, J. A. Chiral Allylic and Allenic Metal Reagents for Organic Synthesis. *J. Org. Chem.* **2007**, *72* (22), 8153–8166.
(387) Nakajima, M.; Saito, M.; Shiro, M.; Hashimoto, S. (S)-3,3‘-Dimethyl-2,2‘-Biquinoline N,N‘-Dioxide as an Efficient Catalyst for Enantioselective Addition of Allyltrichlorosilanes to Aldehydes. *J. Am. Chem. Soc.* **1998**, *120* (25), 6419–6420.
(388) Nakajima, M.; Saito, M.; Hashimoto, S. Selective Synthesis of Optically Active Allenic and Homopropargylic Alcohols from Propargyl Chloride. *Tetrahedron: Asymmetry* **2002**, *13* (22), 2449–2452.
(389) Chen, J.; Captain, B.; Takenaka, N. Helical Chiral 2,2 ’-Bipyridine N- Monoxides as Catalysts in the Enantioselective Propargylation of Aldehydes with Allenyltrichlorosilane. *Org. Lett.* **2011**, *13* (7), 1654–1657.
(390) Lu, T.; Porterfield, M. A.; Wheeler, S. E. Explaining the Disparate Stereoselectivities of N-Oxide Catalyzed Allylations and Propargylations of Aldehydes. *Org. Lett.* **2012**, *14* (20), 5310–5313.
(391) Lu, T.; Zhu, R.; An, Y.; Wheeler, S. E. Origin of Enantioselectivity in the Propargylation of Aromatic Aldehydes Catalyzed by Helical N-Oxides. *J. Am. Chem. Soc.* **2012**, *134* (6), 3095–3102.
(392) 4 intermediates (1f_S_bp2 Int2, 3e_R_bp1 Int3, 3e_S_bp1 Int3, and 3j_S_bp2 Int3) could not be converged, therefore the corresponding enantiomeric TS structures were removed from the original database of 760 TSs.
(393) Vu, K.; Snyder, J. C.; Li, L.; Rupp, M.; Chen, B. F.; Khelif, T.; Müller, K.-R.; Burke, K. Understanding Kernel Ridge Regression: Common Behaviors from Simple Functions to Density Functionals. *Int. J. Quantum Chem.* **2015**, *115* (16), 1115–1128.
(394) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9* (8), 3404–3419.
(395) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11* (5), 2087–2096.
(396) Hu, D.; Xie, Y.; Li, X.; Li, L.; Lan, Z. Inclusion of Machine Learning Kernel Ridge Regression Potential Energy Surfaces in On-the-Fly Nonadiabatic Molecular Dynamics Simulation. *J. Phys. Chem. Lett.* **2018**, *9* (11), 2725–2732.
(397) Westermayr, J.; Faber, F. A.; Christensen, A. S.; Lilienfeld, O. A. von; Marquetand, P. Neural Networks and Kernel Ridge Regression for Excited States Dynamics of CH2NH2+: From Single-State to Multi-State Representations and Multi-Property Machine Learning Models. *Mach. Learn.: Sci. Technol.* **2020**, *1* (2), 025009.
(398) Nguyen, Q. V.; De, S.; Lin, J.; Cevher, V. Chemical Machine Learning with Kernels: The Impact of Loss Functions. *Int. J. Quantum Chem.* **2019**, *119* (9), e25872.
(399) Pozdnyakov, S. N.; Willatt, M. J.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Incompleteness of Atomic Structure Representations. *Phys. Rev. Lett.* **2020**, *125* (16), 166001.
(400) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer New York: New York, NY, 2009.
(401) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully Optimized Contracted Gaussian Basis Sets of Triple Zeta Valence Quality for Atoms Li to Kr. *J. Chem. Phys.* **1994**, *100* (8), 5829–5835.

(402) Cancès, E.; Mennucci, B. New Applications of Integral Equations Methods for Solvation Continuum Models: Ionic Solutions and Liquid Crystals. *J. Math. Chem.* **1998**, *23* (3), 309–326.
(403) Cancès, E.; Mennucci, B.; Tomasi, J. A New Integral Equation Formalism for the Polarizable Continuum Model: Theoretical Background and Applications to Isotropic and Anisotropic Dielectrics. *J. Chem. Phys.* **1997**, *107* (8), 3032–3041.
(404) Because the original TS database was computed with Gaussian09, the Fine (75,302) integration grid (default of Gaussian09) was used instead of the UltraFine (99,590) grid (default of Gaussian16).
(405) Christensen, A. S.; Faber, F. A.; Huang, B.; Bratholm, L. A.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. QML: A Python Toolkit for Quantum Machine Learning, 2017.
(406) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
(407) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6* (12), 2326–2331.
(408) Weinreich, J.; Browning, N. J.; von Lilienfeld, O. A. Machine Learning of Free Energies in Chemical Compound Space Using Ensemble Representations: Reaching Experimental Uncertainty for Solvation. *J. Chem. Phys.* **2021**, *154* (13), 134113.
(409) Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and Structural Distribution Based Representation for Universal Quantum Machine Learning. *J. Chem. Phys.* **2018**, *148* (24), 241717.
(410) Na, G. S.; Chang, H.; Kim, H. W. Machine-Guided Representation for Accurate Graph-Based Molecular Machine Learning. *Phys. Chem. Chem. Phys.* **2020**, *22* (33), 18526–18535.
(411) Hammond, G. S. A Correlation of Reaction Rates. *J. Am. Chem. Soc.* **1955**, *77* (2), 334–338.
(412) Ross, B. C. Mutual Information between Discrete and Continuous Data Sets. *PLOS ONE* **2014**, *9* (2), 1–5.
(413) Kilian Q. Weinberger; Gerald Tesauro. Metric Learning for Kernel Regression. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*; Marina Meila, Xiaotong Shen, Eds.; PMLR, 2007; pp 612–619.
(414) Helfrecht, B. A.; Cersonsky, R. K.; Fraux, G.; Ceriotti, M. Structure-Property Maps with Kernel Principal Covariates Regression. *Mach. Learn.: Sci. Technol.* **2020**, *1* (4), 045021.
(415) Malkov, A. V.; Westwater, M.-M.; Gutnov, A.; Ramírez-López, P.; Friscourt, F.; Kadlčíková, A.; Hodačová, J.; Rankovic, Z.; Kotora, M.; Kočovský, P. New Pyridine N-Oxides as Chiral Organocatalysts in the Asymmetric Allylation of Aromatic Aldehydes. *Tetrahedron* **2008**, *64* (49), 11335–11348.
(416) Collins, K. D.; Glorius, F. A Robustness Screen for the Rapid Assessment of Chemical Reactions. *Nat. Chem.* **2013**, *5* (7), 597–601.
(417) Brown, D. G.; Boström, J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **2016**, *59* (10), 4443–4458.
(418) Brethomé, A. V.; Paton, R. S.; Fletcher, S. P. Retooling Asymmetric Conjugate Additions for Sterically Demanding Substrates with an Iterative Data-Driven Approach. *ACS Catal.* **2019**, *9* (8), 7179–7187.
(419) Gao, X.; Kagan, H. B. One-Pot Multi-Substrate Screening in Asymmetric Catalysis. *Chirality* **1998**, *10* (1–2), 120–124.
(420) Satyanarayana, T.; Kagan, H. B. The Multi-Substrate Screening of Asymmetric Catalysts. *Adv. Synth. Catal.* **2005**, *347* (6), 737–748.

(421) Burgess, K.; Lim, H.-J.; Porte, A. M.; Sulikowski, G. A. New Catalysts and Conditions for a C-H Insertion Reaction Identified by High Throughput Catalyst Screening. *Angew. Chem. Int. Ed.* **1996**, *35* (2), 220–222.
(422) Prieto Kullmer, C. N.; Kautzky, J. A.; Krska, S. W.; Nowak, T.; Dreher, S. D.; MacMillan, D. W. C. Accelerating Reaction Generality and Mechanistic Insight through Additive Mapping. *Science* **2022**, *376* (6592), 532–539.
(423) Beker, W.; Roszak, R.; Wołos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. *J. Am. Chem. Soc.* **2022**, *144* (11), 4819–4827.
(424) Tu, Z.; Stuyver, T.; Coley, C. W. Predictive Chemistry: Machine Learning for Reaction Deployment, Reaction Development, and Reaction Discovery. *Chem. Sci.* **2023**, *14* (2), 226–244.
(425) Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H. S.; Glorius, F. Machine Learning the Ropes: Principles, Applications and Directions in Synthetic Chemistry. *Chem. Soc. Rev.* **2020**, *49* (17), 6154–6168.
(426) Oliveira, J.; Frey, J.; Zhang, S.; Xu, L.; Li, X.; Li, S.; Hong, X.; Ackermann, L. When Machine Learning Meets Molecular Synthesis. *Trends Chem.* **2022**, *4* (10), 863–885.
(427) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572–1583.
(428) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3* (5), 434–443.
(429) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2* (10), 725–732.
(430) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51* (5), 1281–1289.
(431) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem. Int. Ed.* **2016**, *55* (20), 5904–5937.
(432) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604–610.
(433) Karl, T. M.; Bouayad-Gervais, S.; Hueffel, J. A.; Sperger, T.; Wellig, S.; Kaldas, S. J.; Dabranskaya, U.; Ward, J. S.; Rissanen, K.; Tizzard, G. J.; Schoenebeck, F. Machine Learning-Guided Development of Trialkylphosphine Ni(I) Dimers and Applications in Site-Selective Catalysis. *J. Am. Chem. Soc.* **2023**, *145* (28), 15414–15424.
(434) Guo, J.; Ranković, B.; Schwaller, P. Bayesian Optimization for Chemical Reactions. *CHIMIA* **2023**, *77* (1/2), 31.
(435) Angello, N. H.; Rathore, V.; Beker, W.; Wołos, A.; Jira, E. R.; Roszak, R.; Wu, T. C.; Schroeder, C. M.; Aspuru-Guzik, A.; Grzybowski, B. A.; Burke, M. D. Closed-Loop Optimization of General Reaction Conditions for Heteroaryl Suzuki-Miyaura Coupling. *Science* **2022**, *378* (6618), 399–405.
(436) Lai, J.; Li, J.; Betinol, I. O.; Kuang, Y.; Reid, J. P. A Statistical Modeling Approach to Catalyst Generality Assessment in Enantioselective Synthesis. *ChemRxiv* **2022**.
(437) Freeze, J. G.; Kelly, H. R.; Batista, V. S. Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists. *Chem. Rev.* **2019**, *119* (11), 6595–6612.
(438) 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* (25), 6267–6278.
(439) Strandgaard, M.; Seumer, J.; Benediktsson, B.; Bhowmik, A.; Vegge, T.; Jensen, J. H. Genetic Algorithm-Based Re-Optimization of the Schrock Catalyst for Dinitrogen Fixation. *ChemRxiv* **2023**.

(440) Pictet, A.; Spengler, Theod. Über Die Bildung von Isochinolin-Derivaten Durch Einwirkung von Methylal Auf Phenyl-Äthylamin, Phenyl-Alanin Und Tyrosin. *Ber. Dtsch. Chem. Ges.* **1911**, *44* (3), 2030–2036.
(441) Calcaterra, A.; Mangiardi, L.; Delle Monache, G.; Quaglio, D.; Balducci, S.; Berardozzi, S.; Iazzetti, A.; Franzini, R.; Botta, B.; Ghirga, F. The Pictet-Spengler Reaction Updates Its Habits. *Molecules* **2020**, *25* (2).
(442) Stöckigt, J.; Antonchick, A. P.; Wu, F.; Waldmann, H. The Pictet–Spengler Reaction in Nature and in Organic Chemistry. *Angew. Chem. Int. Ed.* **2011**, *50* (37), 8538–8564.
(443) Biswas, A. Organocatalyzed Asymmetric Pictet-Spengler Reactions. *ChemistrySelect* **2023**, *8* (3), e202203368.
(444) Andres, R.; Wang, Q.; Zhu, J. Catalytic Enantioselective Pictet–Spengler Reaction of α-Ketoamides Catalyzed by a Single H-Bond Donor Organocatalyst. *Angew. Chem. Int. Ed.* **2022**, *61* (19), e202201788.
(445) Zhang, Z.; Schreiner, P. R. (Thio)Urea Organocatalysis—What Can Be Learnt from Anion Recognition? *Chem. Soc. Rev.* **2009**, *38* (4), 1187–1198.
(446) Min, C.; Mittal, N.; Sun, D. X.; Seidel, D. Conjugate-Base-Stabilized Brønsted Acids as Asymmetric Catalysts: Enantioselective Povarov Reactions with Secondary Aromatic Amines. *Angew. Chem. Int. Ed.* **2013**, *52* (52), 14084–14088.
(447) Taylor, M. S.; Jacobsen, E. N. Highly Enantioselective Catalytic Acyl-Pictet−Spengler Reactions. *J. Am. Chem. Soc.* **2004**, *126* (34), 10558–10559.
(448) Wanner, M. J.; van der Haas, R. N. S.; de Cuba, K. R.; van Maarseveen, J. H.; Hiemstra, H. Catalytic Asymmetric Pictet–Spengler Reactions via Sulfenyliminium Ions. *Angew. Chem. Int. Ed.* **2007**, *46* (39), 7485–7487.
(449) Sewgobind, N. V.; Wanner, M. J.; Ingemann, S.; de Gelder, R.; van Maarseveen, J. H.; Hiemstra, H. Enantioselective BINOL-Phosphoric Acid Catalyzed Pictet−Spengler Reactions of N-Benzyltryptamine. *J. Org. Chem.* **2008**, *73* (16), 6405–6408.
(450) Klausen, R. S.; Jacobsen, E. N. Weak Brønsted Acid−Thiourea Co-Catalysis: Enantioselective, Catalytic Protio-Pictet−Spengler Reactions. *Org. Lett.* **2009**, *11* (4), 887–890.
(451) Huang, D.; Xu, F.; Lin, X.; Wang, Y. Highly Enantioselective Pictet–Spengler Reaction Catalyzed by SPINOL-Phosphoric Acids. *Chem. Eur. J* **2012**, *18* (11), 3148–3152.
(452) Mittal, N.; Sun, D. X.; Seidel, D. Conjugate-Base-Stabilized Brønsted Acids: Catalytic Enantioselective Pictet–Spengler Reactions with Unmodified Tryptamine. *Org. Lett.* **2014**, *16* (3), 1012–1015.
(453) Qi, L.; Hou, H.; Ling, F.; Zhong, W. The Cinchona Alkaloid Squaramide Catalyzed Asymmetric Pictet–Spengler Reaction and Related Theoretical Studies. *Org. Biomol. Chem.* **2018**, *16* (4), 566–574.
(454) Odagi, M.; Araki, H.; Min, C.; Yamamoto, E.; Emge, T. J.; Yamanaka, M.; Seidel, D. Insights into the Structure and Function of a Chiral Conjugate-Base-Stabilized Brønsted Acid Catalyst. *Eur. J. Org. Chem.* **2019**, *2019* (2–3), 486–492.
(455) Andres, R.; Wang, Q.; Zhu, J. Asymmetric Total Synthesis of (−)-Arborisidine and (−)-19-Epi-Arborisidine Enabled by a Catalytic Enantioselective Pictet–Spengler Reaction. *J. Am. Chem. Soc.* **2020**, *142* (33), 14276–14285.
(456) Chan, Y.-C.; Sak, M. H.; Frank, S. A.; Miller, S. J. Tunable and Cooperative Catalysis for Enantioselective Pictet-Spengler Reaction with Varied Nitrogen-Containing Heterocyclic Carboxaldehydes. *Angew. Chem. Int. Ed.* **2021**, *60* (46), 24573–24581.
(457) Lynch-Colameta, T.; Greta, S.; Snyder, S. A. Synthesis of Aza-Quaternary Centers via Pictet–Spengler Reactions of Ketonitrones. *Chem. Sci.* **2021**, *12* (17), 6181–6187.
(458) Nakamura, S.; Matsuda, Y.; Takehara, T.; Suzuki, T. Enantioselective Pictet–Spengler Reaction of Acyclic α-Ketoesters Using Chiral Imidazoline-Phosphoric Acid Catalysts. *Org. Lett.* **2022**, *24* (4), 1072–1076.
(459) Andres, R.; Sun, F.; Wang, Q.; Zhu, J. Organocatalytic Enantioselective Pictet–Spengler Reaction of α-Ketoesters: Development and Application to the Total Synthesis of (+)-Alstratine A. *Angew. Chem. Int. Ed.* **2023**, *62* (1), e202213831.

(460) Andres, R.; Wang, Q.; Zhu, J. Divergent Asymmetric Total Synthesis of (−)-Voacafricines A and B. *Angew. Chem. Int. Ed.* **2023**, *62* (16), e202301517.
(461) Mauger, A.; Jarret, M.; Tap, A.; Perrin, R.; Guillot, R.; Kouklovsky, C.; Gandon, V.; Vincent, G. Collective Total Synthesis of Mavacuran Alkaloids through Intermolecular 1,4-Addition of an Organolithium Reagent. *Angew. Chem. Int. Ed.* **2023**, *62* (21), e202302461.
(462) Lee, Y.; Klausen, R. S.; Jacobsen, E. N. Thiourea-Catalyzed Enantioselective Iso-Pictet–Spengler Reactions. *Org. Lett.* **2011**, *13* (20), 5564–5567.
(463) Das, S.; Liu, L.; Zheng, Y.; Alachraf, M. W.; Thiel, W.; De, C. K.; List, B. Nitrated Confined Imidodiphosphates Enable a Catalytic Asymmetric Oxa-Pictet–Spengler Reaction. *J. Am. Chem. Soc.* **2016**, *138* (30), 9429–9432.
(464) Adili, A.; Webster, J.-P.; Zhao, C.; Mallojjala, S. C.; Romero-Reyes, M. A.; Ghiviriga, I.; Abboud, K. A.; Vetticatt, M. J.; Seidel, D. Mechanism of a Dually Catalyzed Enantioselective Oxa-Pictet–Spengler Reaction and the Development of a Stereodivergent Variant. *ACS Catal.* **2023**, *13* (4), 2240–2249.
(465) Scharf, M. J.; List, B. A Catalytic Asymmetric Pictet–Spengler Platform as a Biomimetic Diversification Strategy toward Naturally Occurring Alkaloids. *J. Am. Chem. Soc.* **2022**, *144* (34), 15451–15456.
(466) Raheem, I. T.; Thiara, P. S.; Peterson, E. A.; Jacobsen, E. N. Enantioselective Pictet−Spengler-Type Cyclizations of Hydroxylactams: H-Bond Donor Catalysis by Anion Binding. *J. Am. Chem. Soc.* **2007**, *129* (44), 13404–13405.
(467) Muratore, M. E.; Holloway, C. A.; Pilling, A. W.; Storer, R. I.; Trevitt, G.; Dixon, D. J. Enantioselective Brønsted Acid-Catalyzed N-Acyliminium Cyclization Cascades. *J. Am. Chem. Soc.* **2009**, *131* (31), 10796–10797.
(468) Holloway, C. A.; Muratore, M. E.; Storer, R. Ian; Dixon, D. J. Direct Enantioselective Brønsted Acid Catalyzed N-Acyliminium Cyclization Cascades of Tryptamines and Ketoacids. *Org. Lett.* **2010**, *12* (21), 4720–4723.
(469) Aillaud, I.; Barber, D. M.; Thompson, A. L.; Dixon, D. J. Enantioselective Michael Addition/Iminium Ion Cyclization Cascades of Tryptamine-Derived Ureas. *Org. Lett.* **2013**, *15* (12), 2946–2949.
(470) Gregory, A. W.; Jakubec, P.; Turner, P.; Dixon, D. J. Gold and BINOL-Phosphoric Acid Catalyzed Enantioselective Hydroamination/N-Sulfonyliminium Cyclization Cascade. *Org. Lett.* **2013**, *15* (17), 4330–4333.
(471) Cai, Q.; Liang, X.-W.; Wang, S.-G.; You, S.-L. An Olefin Isomerization/Asymmetric Pictet–Spengler Cascade via Sequential Catalysis of Ruthenium Alkylidene and Chiral Phosphoric Acid. *Org. Biomol. Chem.* **2013**, *11* (10), 1602–1605.
(472) Wang, S.-G.; Xia, Z.-L.; Xu, R.-Q.; Liu, X.-J.; Zheng, C.; You, S.-L. Construction of Chiral Tetrahydro-β-Carbolines: Asymmetric Pictet–Spengler Reaction of Indolyl Dihydropyridines. *Angew. Chem. Int. Ed.* **2017**, *56* (26), 7440–7443.
(473) Long, D.; Zhao, G.; Liu, Z.; Chen, P.; Ma, S.; Xie, X.; She, X. Enantioselective Pictet–Spengler Condensation to Access the Total Synthesis of (+)-Tabertinggine. *Eur. J. Org. Chem.* **2022**, *2022* (10), e202200088.
(474) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5* (2), 107–113.
(475) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.
(476) Solel, E.; Tarannam, N.; Kozuch, S. Catalysis: Energy Is the Measure of All Things. *Chem. Commun.* **2019**, *55* (37), 5306–5322.
(477) Klausen, R. S.; Kennedy, C. R.; Hyde, A. M.; Jacobsen, E. N. Chiral Thioureas Promote Enantioselective Pictet–Spengler Cyclization by Stabilizing Every Intermediate and Transition State in the Carboxylic Acid-Catalyzed Reaction. *J. Am. Chem. Soc.* **2017**, *139* (35), 12299–12309.

(478) Zheng, C.; You, S.-L. Exploring the Chemistry of Spiroindolenines by Mechanistically-Driven Reaction Development: Asymmetric Pictet–Spengler-Type Reactions and Beyond. *Acc. Chem. Res.* **2020**, *53* (4), 974–987.
(479) Zheng, C.; Xia, Z.-L.; You, S.-L. Unified Mechanistic Understandings of Pictet-Spengler Reactions. *Chem* **2018**, *4* (8), 1952–1966.
(480) Lisnyak, V. G.; Lynch-Colameta, T.; Snyder, S. A. Mannich-Type Reactions of Cyclic Nitrones: Effective Methods for the Enantioselective Synthesis of Piperidine-Containing Alkaloids. *Angew. Chem. Int. Ed.* **2018**, *57* (46), 15162–15166.
(481) Xu, H.; Zuend, S. J.; Woll, M. G.; Tao, Y.; Jacobsen, E. N. Asymmetric Cooperative Catalysis of Strong Brønsted Acid–Promoted Reactions Using Chiral Ureas. *Science* **2010**, *327* (5968), 986–990.
(482) Association for Computing Machinery Special Interest Group on Management of Data; ACM Special Interest Group on Knowledge Discovery in Data. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM New York, NY: New York, NY, 2016.
(483) Burés, J.; Armstrong, A.; Blackmond, D. G. Curtin–Hammett Paradigm for Stereocontrol in Organocatalysis by Diarylprolinol Ether Catalysts. *J. Am. Chem. Soc.* **2012**, *134* (15), 6741–6750.
(484) Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. Machine Learning for Chemical Reactivity: The Importance of Failed Experiments. *Angew. Chem. Int. Ed.* **2022**, *61* (29), e202204647.
(485) Kariofillis, S. K.; Jiang, S.; Żurański, A. M.; Gandhi, S. S.; Martinez Alvarado, J. I.; Doyle, A. G. Using Data Science To Guide Aryl Bromide Substrate Scope Analysis in a Ni/Photoredox-Catalyzed Cross-Coupling with Acetals as Alcohol-Derived Radical Sources. *J. Am. Chem. Soc.* **2022**, *144* (2), 1045–1055.
(486) Haas, B. C.; Goetz, A. E.; Bahamonde, A.; McWilliams, J. C.; Sigman, M. S. Predicting Relative Efficiency of Amide Bond Formation Using Multivariate Linear Regression. *Proc. Natl. Acad. Sci.* **2022**, *119* (16), e2118451119.
(487) Tang, T.; Hazra, A.; Min, D. S.; Williams, W. L.; Jones, E.; Doyle, A. G.; Sigman, M. S. Interrogating the Mechanistic Features of Ni(I)-Mediated Aryl Iodide Oxidative Addition Using Electroanalytical and Statistical Modeling Techniques. *J. Am. Chem. Soc.* **2023**, *145* (15), 8689–8699.
(488) Yamashita, T.; Kawai, N.; Tokuyama, H.; Fukuyama, T. Stereocontrolled Total Synthesis of (−)-Eudistomin C. *J. Am. Chem. Soc.* **2005**, *127* (43), 15038–15039.
(489) Gobé, V.; Guinchard, X. Stereoselective Synthesis of Chiral Polycyclic Indolic Architectures through Pd0-Catalyzed Tandem Deprotection/Cyclization of Tetrahydro-β-Carbolines on Allenes. *Chem. Eur. J.* **2015**, *21* (23), 8511–8520.
(490) Wenzel, A. G. L., Mathieu P; Jacobsen, Eric N. Divergent Stereoinduction Mechanisms in Urea-Catalyzed Additions to Imines. *Synlett* **2003**, *2003* (12), 1919–1922.
(491) Muthukumar, A.; Sangeetha, S.; Sekar, G. Recent Developments in Functionalization of Acyclic α-Keto Amides. *Org. Biomol. Chem.* **2018**, *16* (39), 7068–7083.
(492) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2019**, *15* (5), 2847–2862.
(493) Pracht, P.; Grimme, S. Calculation of Absolute Molecular Entropies and Heat Capacities Made Simple. *Chem. Sci.* **2021**, *12* (19), 6551–6568.
(494) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120* (1), 215–241.
(495) Zhao, Y.; Truhlar, D. G. Density Functionals with Broad Applicability in Chemistry. *Acc. Chem. Res.* **2008**, *41* (2), 157–167.

(496) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104.
(497) Grimme, S. Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory. *Chem. Eur. J* **2012**, *18* (32), 9955–9964.
(498) Kulik, H. J.; Sigman, M. S. Advancing Discovery in Chemistry with Artificial Intelligence: From Reaction Outcomes to New Materials and Catalysts. *Acc. Chem. Res.* **2021**, *54* (10), 2335–2336.
(499) Alektiar, S. N.; Wickens, Z. K. Rethinking Catalyst Design by Using Data Science. *Chem* **2023**.
(500) Nonoshita, K.; Banno, H.; Maruoka, K.; Yamamoto, H. Organoaluminum-Promoted Claisen Rearrangement of Allyl Vinyl Ethers. *J. Am. Chem. Soc.* **1990**, *112* (1), 316–322.
(501) Knowles, R. R.; Lin, S.; Jacobsen, E. N. Enantioselective Thiourea-Catalyzed Cationic Polycyclizations. *J. Am. Chem. Soc.* **2010**, *132* (14), 5030–5032.
(502) Fromer, J.; Coley, C. Computer-Aided Multi-Objective Optimization in Small Molecule Discovery. *Patterns* **2023**, *4* (2).
(503) Leguy, J.; Glavatskikh, M.; Cauchy, T.; Da Mota, B. Scalable Estimator of the Diversity for de Novo Molecular Generation Resulting in a More Robust QM Dataset (OD9) and a More Efficient Molecular Optimization. *J. Cheminform.* **2021**, *13* (1), 76.
(504) Wodrich, M.; Laplaza, R.; Cramer, N.; Reiher, M.; Corminboeuf, C. Toward in Silico Catalyst Optimization. *CHIMIA* **2023**, *77* (3), 139–143.

Simone Gallarati

EPFL SB ISIC LCMD, CH-1015 Lausanne
✉ simone.gallarati@epfl.ch
0000-0002-2349-1944
Place and date of birth: Milan, 27th September 1995

EDUCATION

Sep. 2019 – Present	**École Polytechnique Fédérale de Lausanne, PhD in Chemistry and Chemical Engineering, Laboratory for Computational Molecular Design (LCMD)** Doctoral thesis: Enriching the computational toolbox for organocatalysis. Supervisor: Prof. Clémence Corminboeuf. Oral exam: 14.09.2023, public defence: 12.10.2023
Sep. 2014 – June 2019	**University of St Andrews, Master in Chemistry (Honours) Materials Chemistry with External Placement (year)** Master's thesis: Understanding catalyst-structure performance relationships in Pd catalysed enantioselective carbonylation of alkenes. Supervisors: Prof. Michael Bühl, Prof. Matthew L. Clarke

SELECTED PRIZES, AWARDS, AND FELLOWSHIPS

- **GRC Physical Organic Chemistry Poster Prize** (June 2023, Holderness School, NH, US; short presentation titled "Genetic Optimization of Homogeneous Catalysts: from Specificity to Generality" in the closing session)
- **WATOC 2020 Poster Prize** (12th Triennial Congress of the World Association of Theoretical and Computational Chemists, Vancouver, BC, Canada, July 2023)
- **The Principal's Scholarship for Academic Excellence** (2019, awarded to the fifty final year students from across the University whose grades are the highest in their faculties)
- **Charles Horrex Prize** (2019, awarded to the best Honours Research Project in Physical Chemistry)
- **Forrester Prize** (2019, awarded to the best finishing students in the Fourth and Fifth year classes)
- **Irvine Jubilee Prize and Medal** (2019, awarded to the most distinguished finishing MChem student in Chemistry)
- **Gray Prize** (2019, awarded to the best essay on a prescribed topic in Chemistry)
- **Elizabeth Soutar Prize** Joint (2017, awarded to the "Best Students in Third Year Chemistry") and Medal in 3rd Year Chemistry (2017, awarded for best performance in 3rd Level Chemistry)

EMPLOYMENT HISTORY

Sep. 2019 – Present	**Laboratory for Computational Molecular Design, EPFL** Doctoral thesis under the supervision of Prof. Clémence Corminboeuf
1 June – 27 July 2018	**Summer Program at the Centre for Computational Quantum Chemistry** University of Georgia, Athens, GA, US I worked in the group of Prof Steven E. Wheeler on the prediction of catalytic activity and selectivity in asymmetric homogeneous catalysis using modern DFT methods. The summer program also involved the attendance of a series of lectures on important topics in quantum chemistry and the completion of programming projects in Python
26 June 2017 – 18 May 2018	**Undergraduate Year in Industry Placement at Diamond Light Source, UK** In the "Chirally modified Catalyst Nanoparticles" project, I was involved in the synthesis and study of the catalytic activity and enantioselectivity of Ni nanoparticles in the chemical laboratory facilities of Diamond Light Source and of the UK Catalysis Hub in the Research Complex at Harwell, along with their investigations at the Versatile Soft X-ray (VERSOX) beamline, where I conducted Near-Ambient Pressure XPS and NEXAFS experiments. I presented this work at the Faraday Discussion "Designing Nanoparticle Systems for Catalysis" (London, 16–18 May 2018)
June – August 2016	**Forensic Chemistry internship at Nottingham Trent University, UK** Synthesis of dye-impregnated fingerprint lifting gels and their characterization *via* fluorescence spectroscopy
Summer 2015	**Private chemistry and physics tutor** for high school students in Bergamo, Italy
Summer – Autumn 2013	**BergamoScienza**, presenter–guide to school laboratories (Bergamo, Italy) Presenter and guide of various school laboratories. I oversaw the school experiments, which the public could participate in, while planning and working on my personal project (building a microbial fuel cell), which was then exhibited to the public
Summer 2013	**SIAD S.p.A., Production plant of Osio Sopra** (Bergamo, Italy) Visited and assisted in daily laboratory activities at the local chemical company of the SIAD group

POSITIONS OF RESPONSIBILITY

July – December 2023	**Supervisor of junior researchers** Lucien Brey, Alexander Makaveev (Master students). Project: Optimizing generality in asymmetric organocatalysis with evolutionary experiments
2018 – 2019	**President of the University of St Andrews Chemical Society** My role was to promote and provide a social forum for the propagation of Chemistry within the community of the University and the town of St Andrews. I was ultimately responsible for the organisation of all events, the delegation of duties within the Society and the administration of ChemSoc's affairs and finances
2016 – 2019	**University of St Andrews Chemistry Class Rep and Library Rep** This position allowed me to be a valuable connection between other students and the staff members, to gather and organise feedback from my peers, act on it or report it to others within the School or University

TEACHING ACTIVITIES

Advanced General Chemistry I, Bachelor's course. Teaching assistant for two semesters between 2020–2021 (69 hours overall)

Organic Chemistry, Bachelor's course. Teaching assistant for three semesters between 2020–2022 (156 hours overall)

Physical and Computational Organic Chemistry, Master's course. Teaching assistant for two semesters between 2021–2022 (32 hours overall)

Project of Computational Chemistry, Bachelor's course. Supervision of the project "Accelerating the screening of organocatalysts through fragmentation" (56 hours overall)

SELECTED PRESENTATIONS, CONFERENCES, OR SEMINARS

24 – 30 June 2023	**Gordon Research Seminar** and **Conference**, Physical Organic Chemistry. Holderness School in New Hampshire, US. Talk: Optimizing generality in organocatalysis with evolutionary experiments
3 – 8 July 2022	**12th World Association of Theoretical and Computational Chemistry (WATOC) 2020.** Vancouver, BC, Canada. Poster: OSCAR, an extensive repository of functionally diverse organocatalysts
9 December 2021	**Theoretical Physical Organic Chemistry (TPOC) Meeting**. Talk: Data-driven tools for organocatalysis (online)
23 August 2021	**ACS Fall Meeting 2021**, Accelerating Catalysis Research with Machine Learning. Talk: Data-driven advancement of computational tools for organocatalysis (ID: 3595899, online)
5 – 9 June 2021	**EPFL–ETHZ Summer School Big Data and Machine Learning for Chemistry**. Hybrid live/online event, EPFL, Switzerland. Member of the Organising Committee.

PERSONAL SKILLS

Programming Languages, Software, Programs	Bash, Python, MS Office, LaTeX, Gaussian, ADF, DFTB+, VMD, Molden, CYLView, ChemDraw, ConQuest, Reaxys, Igor Pro
Languages	Italian (native speaker), English (fluent), French (basic knowledge, A1/A2)

SELECTED SCIENTIFIC PUBBLICATIONS

1. S. Gallarati, P. van Gerwen, R. Laplaza, S. Vela, A. Fabrizio, and C. Corminboeuf, *Chem. Sci.*, 2022, **13**, 13782.
2. S. Gallarati, R. Laplaza, and C. Corminboeuf, *Org. Chem. Front.*, 2022, **9**, 4041.
3. S. Gallarati, R. Fabregat, V. Juraskova, T. J. Inizan, and C. Corminboeuf, *J. Org. Chem.*, 2022, **87**, 8849.
4. R. Laplaza, S. Gallarati, and C. Corminboeuf, *Chem. Methods*, 2022, e202100107.
5. M. D. Wodrich, M. Chang, S. Gallarati, Ł. Woźniak, N. Cramer, and C. Corminboeuf, *Chem. Eur. J.*, 2022, **28**, e202200399.
6. S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich, and C. Corminboeuf, *Chem. Sci.*, 2021, **12**, 6879.
7. S. Gallarati, P. Dingwall, J. A. Fuentes, M. Bühl, and M. L. Clarke, *Organometallics*, 2020, **39**, 4544.
8. R. Arrigo, S. Gallarati, M. E. Schuster, J. M. Seymour, D. Gianolio, I. da Silva, J. Callison, H. Feng, J. E. Proctor, P. Ferrer, F. Venturini, D. Grinter, and G. Held, *ChemCatChem*, 2020, **12**, 1491–1503.