Paving the road towards automated homogeneous catalyst design

Kalikadien, Adarsh V.; Mirza, Adrian; Hossaini, Aydin Najl; Sreenithya, Avadakkam; Pidko, Evgeny A.

DOI
10.1002/cplu.202300702

Publication date
2024

Document Version
Final published version

Published in
ChemPlusChem

ChemPlusChem

Perspective
doi.org/10.1002/cplu.202300702

Chemistry Europe
European Chemical Societies Publishing

www.chempluschem.org

Paving the road towards automated homogeneous catalyst design

Adarsh V. Kalikadien,[a] Adrian Mirza,[a] Aydin Najl Hossaini,[a] Avadakkam Sreenithya,[a] and Evgeny A. Pidko*[a]

In the past decade, computational tools have become integral to catalyst design. They continue to offer significant support to experimental organic synthesis and catalysis researchers aiming for optimal reaction outcomes. More recently, data-driven approaches utilizing machine learning have garnered considerable attention for their expansive capabilities. This Perspective provides an overview of diverse initiatives in the realm of computational catalyst design and introduces our automated tools tailored for high-throughput *in silico* exploration of the chemical space. While valuable insights are gained through methods for high-throughput *in silico* exploration and analysis of chemical space, their degree of automation and modularity are key. We argue that the integration of data-driven, automated and modular workflows is key to enhancing homogeneous catalyst design on an unprecedented scale, contributing to the advancement of catalysis research.

1. Introduction

Numerous vital industrial processes rely on homogeneous catalysts. Their efficiency in steering a wide array of chemical transformations gives them a distinct status.[1] They are employed in the synthesis of pharmaceuticals, agrochemicals, bulk chemicals and fine chemicals.[1–6] Metal-ligand complexes are integral to modern chemistry, forming the cornerstone of homogeneous catalysis.[1,6] Despite their ubiquity and versatility, the field of homogeneous catalysis confronts an inherent challenge: the quest for the optimal catalyst.

The vast chemical and reaction space in catalysis poses a challenge to exploration.[7,8] It becomes evident that there are no singular candidates exhibiting unique catalytic performance for our applications. How to find the best performing homogeneous catalyst? The opportunity to perform brute-force exploration of potential candidates is always open. Fortunately, guidance by simple models such as the Bronsted–Evans–Polanyi (BEP) relationship, Hammett parameters and linear scaling relationships were established.[9–13] Together with chemical intuition and heuristics, these principles are often used to guide the screening process. They were originally developed to elucidate the correlation between the rate of a chemical reaction and the thermodynamic properties of the reaction constituents.[14,15] However, catalytic activity/selectivity is not straightforward and origins of high or low performance are often not easily explainable by conventional chemical principles.

In contrast to heterogeneous catalysts, homogeneous catalysts have a better defined structure that can be optimized for performance more easily. For example, a wide range of ligands that induce enantioselectivity have been developed for organometallic metal-ligand complexes, enabling high rates and selectivities.[16] Ligand engineering is the common strategy to optimize performance of the catalyst.[17] The modular architecture of transition-metal (TM) coordination complexes paves the way for larger-scale screening, achieved through methods such as fragment-based library construction and subsequent performance optimization.[18–20] Although often guided by mechanistic hypotheses and expert knowledge, ligand engineering has been a primary driver of reaction discovery and catalyst design. Identification of an optimal ligand and subsequent catalyst design is essential to achieve high performance for a desired reaction. However, beyond a specific application, the usability of ligand engineering approaches becomes contentious. Can they be employed on out-of-sample datasets, e.g. on a new chemical reaction?

Despite the potential of the many automated tools for catalyst design,[21] most use cases have been limited to retrospective analyses of experimental results.[22] Recently, successful examples of computational design directly contributing to experimentally validated catalyst discoveries started emerging. Relevant examples include: selective oligo-/polymerization, cross-coupling catalysis, and enantioselective Pauson-Khand reactions.[22–24] Generally, the aim is to optimize experimental targets condensed into a single metric, such as turnover frequency, turnover number, regioselectivity, product selectivity, yield, or enantioselectivity.[24] Rooted in classic principles, computational catalyst design approaches usually involve featurizing the catalyst structure using chemical descriptors. In a reactivity model it is assumed that an experimental objective (e.g., yield or enantioselectivity) is a function of both the experimental parameters and computational descriptors of the

[a] A. V. Kalikadien, A. Mirza, A. N. Hossaini, Dr. A. Sreenithya, Prof. Dr. E. A. Pidko
Inorganic Systems Engineering, Department of Chemical Engineering, Faculty of Applied Sciences, Delft University of Technology, Van der Maasweg 9, 2629 HZ, Delft (The Netherlands)
E-mail: e.a.pidko@tudelft.nl

catalyst structure. This function can then be learned by a statistical model to enable predictions. The success of this approach relies strongly on accurate representations of the catalytic structure.[22,25]

Descriptors are chemically intuitive features of the catalyst structures that are known to be relevant for the catalytic activity. Two classical examples from organometallic chemistry are Tolman's electronic parameter (TEP) and the Tolman cone angle.[26] This cone angle was further adapted into White's solid angle which also takes the ligand's flexbility into account. Later, the derived concept of percent buried volume was introduced.[27,28] Many descriptors can be envisioned and the development of new descriptors, as well as derivatives of classical approaches and rapid calculation methods remains an ongoing endeavor.[29–31] Individual descriptor classes are usually categorized as being electronic, steric, geometric or thermodynamic. For a comprehensive overview of descriptors used in catalyst design, we refer the reader to a review by Durand et al.[29] These descriptors have played a pivotal role in homogeneous catalyst design since its inception. Remarkably, the buried volume, a fundamental descriptor, were already employed back in the 80s to gain insights and enable the prediction of enantiomeric excesses.[32] Another noteworthy example is the bite angle, used to describe the angle between two donor atoms and the metal (L-M-L, in the case of bidentate ligands). It was reported that the bite angle has a large impact on metal-centered reactivity in 1999.[30,33,34] More and more, individual descriptors of the chemical structures have progressed into refined representations of chemical properties, which can be used to optimize particular objective(s).[25,35] For example, several libraries such as ReaLigands and the Ligand Knowledge Bases have been developed to elucidate ligand effects across a range of representative coordination environments.[29,36] The mapping of these descriptors provides an overview of the ligand space and a direction for more design, possibly within different ligand classes.[37] Additionally, less chemically intuitive descriptors such as graph-based representations[38] or derivations thereof[39] have also been applied in TM-based catalysis.

Present day statistical methods used in catalyst design range in complexity from linear explainable models to advanced natural language processing (NLP) models for chemistry.[40,41] The former category is the traditional way automated catalyst design was tackled, while the latter emerged as a powerful tool only in the recent years. NLP models became feasible by the introduction of the transformer architecture for neural networks, which allowed processing of inputs of different sizes and interpretation of chemical languages (e.g. SMILES,[42] DeepSMILES[43] or SELFIES[44]) in a similar way to human languages.

In the pursuit of understanding complex phenomena, human intuition has often led to the development of simplified and interpretable representations. In the domain of chemistry and cheminformatics, descriptors serve as static and compressed representations of specific chemical structures. Within the realm of catalysis however, every stage involved in constructing such a digital representation of a catalyst complex is susceptible to introducing significant deviations.[22,45] This process typically encompasses various steps, such as the extraction or creation of the initial complex, density functional theory (DFT) optimization, and descriptor calculation. Together, these constitute the workflow utilized for the creation of a computational and condensed structure representation. Thus, in predictive approaches, the catalyst structure, computational representation, and modeling space are deeply intertwined.[19] It is important to acknowledge that these representations are often still influenced by expert bias, mainly due to the manual generation of the initial chemical structure. This inherent bias can limit the generalizability of published approaches. In addition, this enhances the streetlight effect. This phenomenon refers to the tendency to focus on areas that are well-illuminated, or well-understood, while neglecting less-explored regions, potentially hindering a comprehensive understanding of the chemical space.

To address and mitigate the biases and constraints inherent in the manual structure generation process, the integration of automated structure generation tools is critical in advancing the field of rational catalyst design. Numerous tools have emerged, facilitating the reliable generation of 3D structures. Notable examples include DENOPTIM, Aarontools, MolSimplify, MolAssembler, and the more recent addition of Architector.[46–50] However, the pursuit of an universally applicable computational approach that streamlines all aspects, ranging from structure generation to descriptor computation for organometallic complexes, remains a highly coveted goal within the research community. Such a tool would significantly enhance the efficiency and effectiveness of catalyst design endeavors. This is the philosophy behind our in-development Python package called Open Bidentate Ligand eXplorer (OBeLiX).

In this Perspective, we aim to critically discuss approaches for automated catalyst design and highlight the path that we have followed. We will start by introducing a historic timeline of several fields that majorly contributed to modern catalyst design approaches. Further, we include a brief review of the current frameworks for catalyst design and present several challenges accompanying it. We will conclude by proposing a workflow for automating insight extraction, both about chemistry and mechanistic pathways, and how it can be coupled with machine learning for a full picture of a catalyst's behaviour. We believe that high-throughput automated knowledge extraction is a major step for propelling future endeavours of the catalysis community and that first principles of chemistry and catalysis should be incorporated into modern workflows for successful cross-disciplinary integration.

2. The foundation of the road

The current advances in computational homogeneous catalyst design primarily stem from the integration of four scientific disciplines: experimental organometallic chemistry and catalysis, quantum chemistry (QC), artificial intelligence (AI), and cheminformatics. These are at the core of current state-of-the-art approaches. Their historical evolution has significantly

influenced and shaped the modern landscape of this field. Figure 1 presents a timeline of selected seminal works and tools across this multidisciplinary field, highlighting the parallel development of key methodologies and tools alongside experimental discoveries in homogeneous catalysis. Within this graphical representation, the progress in structure optimization methods is denoted by a red star, while experimental works are represented by a blue square. The integration of cheminformatics, which is crucial for data analysis and modeling, is symbolized by a yellow circle. Lastly, the emergence and growing influence of artificial intelligence and machine learning (AI/ML) techniques in catalyst design are depicted by a purple hexagon. In this section, we will delve into the progress and advancements made in each field, shedding light on their respective developmental journeys.

Electronic structure calculations play an important role in computational materials identification, characterization and optimization. For calculating properties of systems from first principles, DFT provides a powerful compromise between predictive power and computational cost.[52] Theoretical methods for studying catalysis have undergone significant development, with computational chemistry now regarded as an essential tool in the catalysis toolbox alongside laboratory techniques.[53–55] The origins of QC can be traced back to the pioneering work of Slater in 1951, marked by a red star on the left side of Figure 1. Slater's development of the Hartree-Fock method[56] marked the beginning of computational quantum mechanical (QM) methods by enabling feasible calculations for determining the energy minima of molecules.[51] The subsequent Kohn-Sham framework for approximating the electronic kinetic energy contribution proved especially useful.[57] A plethora of exchange-correlation potentials are currently advancing the frontier in accurate simulations.[58–64] For larger organometallic complexes, these methods became particularly powerful after the introduction of Grimme's dispersion corrections.[65]

Theoretical frameworks and computational tools, must meet several criteria: (a) yielding reasonable outcomes, (b) operating efficiently within short timeframes, and (c) being applicable to a wide range of systems and physical-chemical properties.[66] The traditional DFT calculations are known to exhibit cubic scaling in computational time due to the diagonalization of the 3D Hamiltonian. This renders them inefficient for large molecular systems with a high number of electrons. Conventional force-field (FF) methods are frequently employed as a starting point, mainly for initial conformation searches.[67] These methods are not generally applicable since they lack parameterization for numerous elements, especially metals.[68,69] This has impeded the progression of the field.[66] In addressing this issue, low-level QC methods step in, offering an alternative to FFs, especially for systems of modest size, typically ranging from 500 to 1,000 atoms. For example, the GFNn-xTB methods are parameterized for applications to a wide range of chemical systems, including (organo)metallic systems[66,67,70] and polymers.[71] Grimme's Conformer-Rotamer Ensemble Sampling Tool (CREST), utilizes the GFNn-xTB methods for the creation and analysis of structure

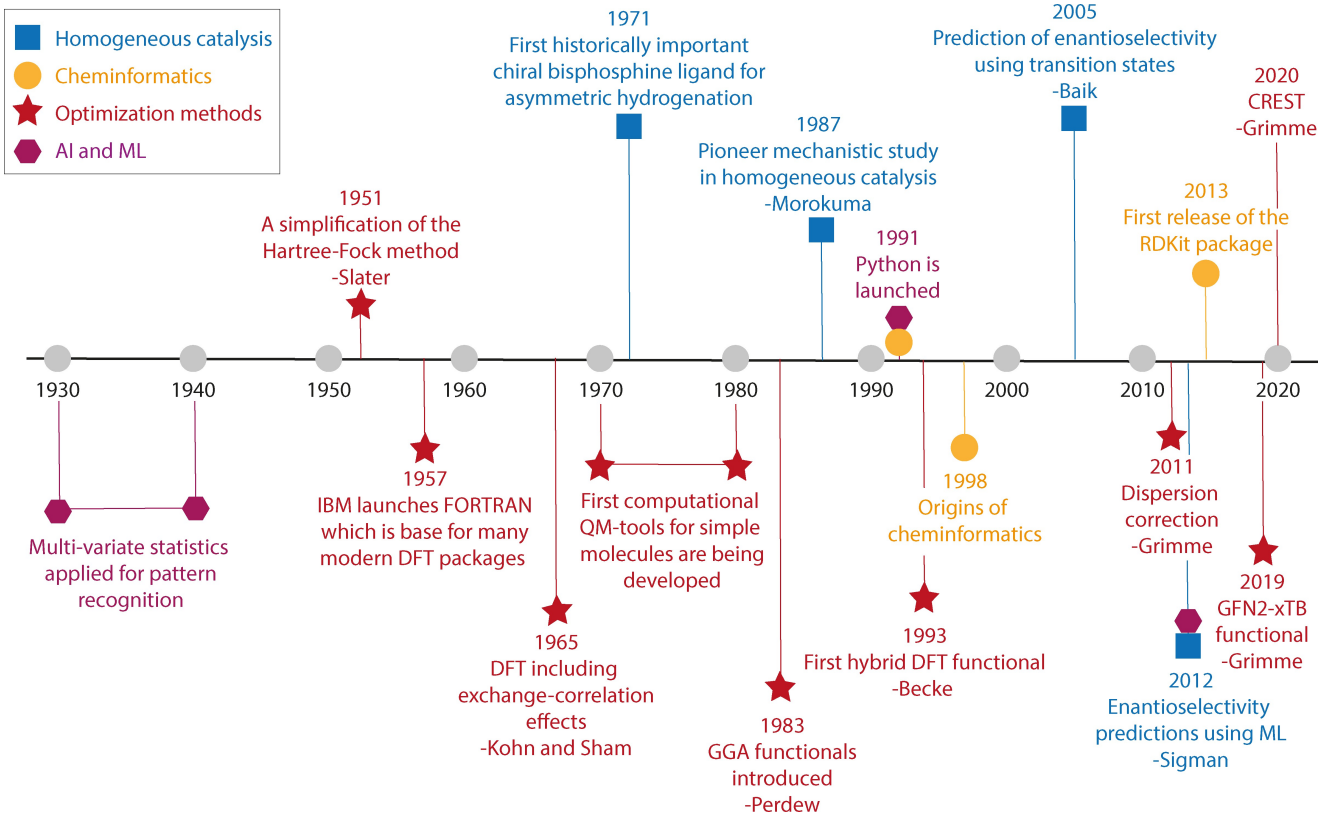


Figure 1. A timeline showing the evolution of major fields contributing to the modern multidisciplinary research in homogeneous catalyst design.[51]

ensembles.[72] Conformational sampling via meta-dynamics simulations, regular MD simulations and genetic Z-matrix crossing have been implemented.[73] While exploring avenues to address the challenges in modeling organometallic systems, it is worth noting that the focus of this Perspective does not encompass machine learning potentials for electronic structure calculations, which is discussed extensively by others.[74–78]

After the modelling step and eventual conformer search, the discrete chemical structures need to be transformed into continuous representations for usage in statistical methods. This transformation is done by the calculation of chemical descriptors, as outlined in the introduction. These aim to capture the essential features of the catalyst for further analysis and design. Calculation of descriptors in a high-throughput manner was made possible by the invention of cheminformatics. Established in 1998,[79] cheminformatics is an emerging domain of information technology. It focuses on the acquisition, organization, analysis, and management of chemical data. This discipline plays a crucial role in facilitating data-driven research and decision-making processes in chemistry. The advancement of cheminformatics is represented by a yellow circle in Figure 1. It has progressed in parallel with the field of machine learning in catalysis. Over the past two decades, continuous advancement of cheminformatics has significantly contributed to the progress achieved in the design and screening of homogeneous catalysts.[21,80] By leveraging the theoretical interpretation of chemical structures rather than relying solely on empirical measures, cheminformatics has enabled the derivation of meaningful relationships and the exploration of the vast chemical space.[81] This progress was mainly fueled by the invention of the Python programming language.[82] It allowed the creation of the OpenBabel[83] and RDKit[84] packages, which are the backbone of many modern cheminformatics workflows. Additionally, the invention and widespread sharing of code via version control platforms such as Github, has catalyzed the development of numerous innovative tools and workflows. These newly developed tools have empowered chemists with the capability to gather, analyze, and interpret chemical data in an efficient and systematic manner. Researchers have harnessed this powerful combination to build sophisticated algorithms for molecular descriptor calculation,[17,85,86] virtual screening,[87] reaction prediction[88–90] and many other aspects of catalyst design. The availability and integration of AI methods, represented by the purple hexagon in Figure 1, has further propelled the field by substantially enhancing the predictive power of these approaches. In its broadest definition, AI encompasses the theory and development of computer systems capable of performing tasks that traditionally require human intellect, such as speech recognition. As a prominent subset of computer science, AI has found significant applications in catalyst design, leveraging numerical methods and machine learning techniques to drive advancements in the field. While this Perspective will primarily focus on machine learning, it is important to acknowledge the vital role that numerical methods have played in enabling the development of DFT in the 1970s. These combined advancements have revolutionized catalyst design by augmenting the capabilities of computational models and enabling more sophisticated analyses and predictions.

The field of homogeneous catalysis has experienced a remarkable increase in the utilization and integration of AI techniques, driven by advancements in multi-variate statistics, quantitative structure-activity relationships (QSAR), and data science methodologies.[89,91,92] In recent years, there has been a notable transition within these modern machine learning approaches, as they have evolved from traditional *white-box* models to more sophisticated *black-box* models, where the emphasis is placed on the quality and size of the training data. White-box models are based on traditional statistics where causal effects are sought after and finding the most "correct" model is the goal. On the other hand, black-box models prioritize predictive accuracy, aiming to find a highly performing model. Explainable models, such as QSARs, exemplify white-box models, where the model's performance is determined by the accuracy of physico-chemical parameters. Classic examples of explainable models include the Hammett equation, the Bronsted–Evans–Polanyi relationship, molecular volcano plots,[93,94] and other linear free scaling relationships (LFERs). In contrast, black-box models often employ deep learning techniques, where descriptors derived from the molecular graph are utilized.[41,90,95] These black-box models focus on prediction quality and may lack explicit interpretability due to their complex internal representations. The exploration of both white-box and black-box models in automated catalyst design demonstrates the diverse strategies employed within the field, encompassing various approaches to predictive capability and performance. While white-box models provide interpretability and insights into causal relationships, black-box models offer greater predictive capabilities, leveraging vast amounts of data to make accurate predictions. The balance between white-box and black-box models in automated catalyst design represents a spectrum rather than a strict dichotomy.

Figure 2 provides a visual representation of the data science continuum, showcasing various modeling techniques employed in catalyst design. These models encompass a range of methodologies, from explainable white-box models that prioritize interpretability, to more complex black-box models focused on predictability. Skilled scientists are capable of extracting valuable insights even from models traditionally classified as black box, such as when employing non-linear dimensionality reduction techniques. This continuous nature of modeling

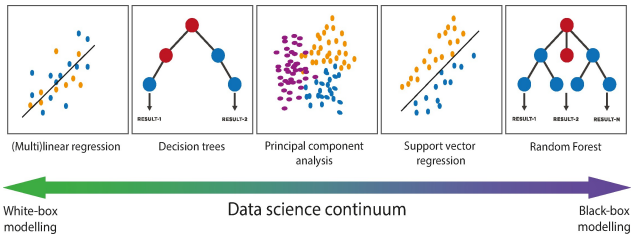


Figure 2. The spectrum of data science techniques ranging from traditional white-box models that allow for explainability to black-box models that do not provide estimations on the importance of each feature or feature interactions.[96]

approaches highlights the interconnectedness and complementarity of different methodologies for catalyst design.

3. Computational catalyst design

We described the four integrated fields forming the foundation of modern computational catalyst design: experimental homogeneous catalysis, QC, cheminformatics, and AI/ML. In this section, we discuss the state-of-the-art computational catalyst design frameworks and their inherent challenges. It is known that at the core of catalyst design, a relationship between the catalytic system and the experimental properties of interest must be established. But how do computational design endeavors work? And what are the challenges in automating them?

Foscato et al. categorized catalyst design into two primary classifications: direct and inverse design.[21] In direct catalyst design, a direct causal relationship is established between a defined catalytic system and the observed experimental performance. The catalyst performance typically consists of measures related to reactivity or selectivity. Since catalyst design is predominantly viewed as a nonlinear optimization problem, this endeavor often employs a diverse array of (non-) linear statistical methods.[21,97] On the other hand, inverse design starts from a known optimal performance and searches for systems with properties that match this performance.[97–99] Most inverse design strategies are still aimed at small organic molecules.[99,100] Only recently has the inverse strategy been applied to a subset of TM-based catalysts.[101] Since the focus of this Perspective lies on TM-based catalysts, our primary focus is on exploring direct catalyst design strategies. These can generally be classified into two categories: mechanism-based and mechanism-agnostic approaches.

The distinguishing factor among these approaches lies in their reliance on mechanistic understanding. While our objective is to understand the reactivity and selectivity of the catalysts, the necessity of the mechanistic understanding remains a matter of ongoing inquiry.

This difference is best illustrated by the example of an enantioselective reaction modeling as shown in Figure 3. In mechanism-agnostic approaches, a form of the precatalyst structure as shown in (a) ,which does not carry any mechanistic information, can be utilized. The correlation of 3D descriptors calculated on this structure with selectivity has been utilized for the design and optimization of chiral ligands.[102] For mechanism-based approaches, TS structures of the selectivity determining step (b) or (c) are used. Small energy differences of 1—3 kcal/mol can significantly impact the preferred reaction pathway in enantioselective reactions, introducing additional complexities.[103] Achieving mechanistic insights thus entails the calculation of complex transition states from competing reaction pathways, followed by rigorous analysis. This makes the mechanism-based approach extremely problem-specific. In addition to these mechanistic insights, targeted DFT calculations are necessary for each new catalyst-substrate combination. On the contrary, the mechanism-agnostic approach is aimed to be more general. However, a deep understanding of the dataset and selected chemical descriptors for predictive modeling is necessary.[104] Despite these inherent disadvantages, both approaches have had successful applications in TM-based homogeneous catalyst design.[21,105]

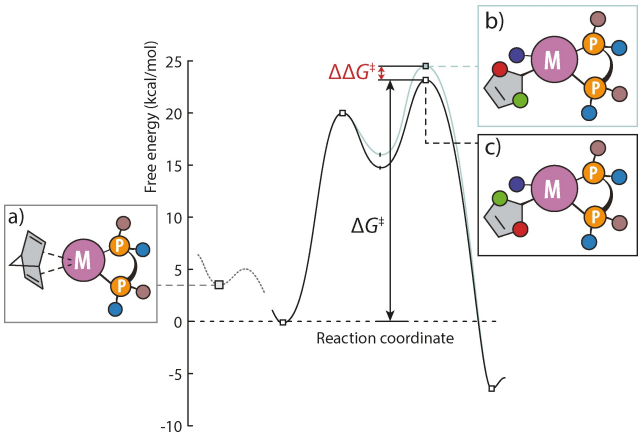


Figure 3. A representative reaction profile diagram for an enantioselective reaction showing the structures used in mechanism-agnostic and mechanism-based computational catalyst design approaches. (a) Represents the precatalyst structure utilized in mechanism-agnostic approaches, while (b) and (c) depict competing prochiral transition state (TS) structures employed in the mechanism-based approach.

In the subsequent discussion, we provide a summary of a selection of state-of-the-art computational catalyst design approaches. These are presented in Figure 4.

3.1. Mechanism-based approaches

As mentioned, mechanism-based catalyst design approaches rely on mechanistic insights and are computationally intensive. However, they have been proven to predict the experimental enantioselectivity of TM-based catalysts well. As a first step in these approaches, the transition state structures connecting two energy minima need to be found. Generally, most approaches first generate an approximate TS structure and optimize towards a saddle point on the potential energy surface.[47] Automated and high-throughput localization of transition state structures has been enabled by tools such as AutoTST and AutoTS.[110–112] The need to sample for configurational and conformational freedom is particularly important in asymmetric catalysis. This sampling is, as far as we know, not implemented yet in these methods.[47] Separate tools enabling the sampling of transition state conformers are available, e.g. AARON,[47,113] Mason[114] and MolAssembler.[49] These can be used to expedite and streamline the automated *in silico* TS screening. These modules facilitate the conformational sampling, and transition state optimization for new catalyst-substrate variants. Combined with FF methods for transition states,[115,116] these automated workflows contribute to faster and more efficient sampling of transition states. This is exemplified in several studies. A graph-based HT screening study was conducted by Laplaza et al.[107] An automated workflow was created to

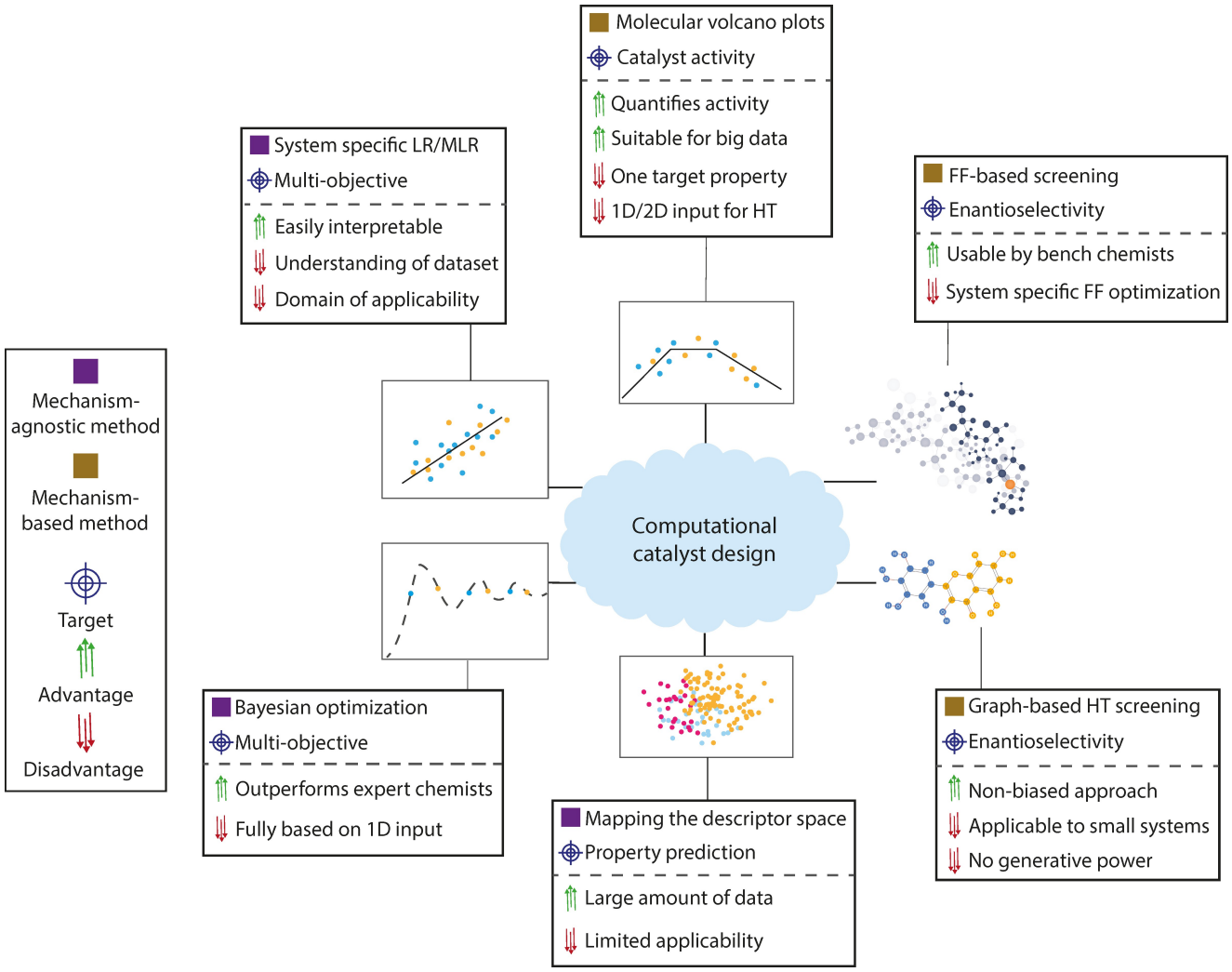


Figure 4. A summary of the methodology, target applications, inherent advantages and disadvantages of several computational workflows for catalyst design. Mechanism-based approaches are indicated by a brown square. Selected examples are: FF-based[106] and graph-based high-throughput (HT) screening.[107] Mechanism-agnostic approaches are indicated by a purple square, selected examples are: Molecular volcano plots,[108] system-specific linear regression/multi-linear regression (LR/MLR),[35] mapping the descriptor space[17] and bayesian optimization (BO).[109]

investigate multiple reaction pathways in a Rh-catalyzed asymmetric C–H functionalization and predict enantioselectivity. This was done on a set of 12 catalytic systems by sampling around 20 transition states per catalyst through unbiased conformational exploration with minimal human intervention.[107] The comparison of these computational predictions with experimental results shows that this workflow might be beneficial in the screening of new selective catalysts. Unfortunately, the computational cost of such an approach can run high due to the amount of transition states and DFT refinements needed. Alternatively, FF-based screening approaches aim to combine a QM treatment of a small core of atoms involved in the reaction, e.g. metal centers and donors of the ligands, with a force field treatment of the remaining molecule.[116–119] Virtual chemist, an approach by Patrascu et al., was specifically designed to empower experimental chemists with minimal computational chemistry knowledge.[106] This method combines Quantum-guided Molecular Mechanics (Q2MM) and molecular mechanics force field (MM3 FF) methods to model the transition states.[120,121] This approach enables bench chemists to virtually screen asymmetric reactions and make predictions about potential catalysts before conducting laboratory experiments. The study presents a significant advancement in the field of computational catalyst design through the development of a comprehensive virtual laboratory framework. This framework incorporates several modules, including Finders, React2D, Quemist, and Ace, each serving a specific function in the virtual design and evaluation of catalysts. However, Virtual Chemist does rely on a parametrized FF per reaction type. For example, the MM3 force field does not include parameterization for metals. To overcome this limitation, the force field parameters are automated using Hartree-Fock methods.[122,123] These Hartree-Fock methods are suboptimal for exploring TM-based catalysts.[106] Additionally, based on the flexibility of the catalyst, results can deteriorate since TSs are approximated as an energy minimum in Q2MM.[116] Shifts in

the position of the transition state along the reaction coordinate, such as while facing significant steric hindrance, remain unaccounted for in an energy minimum model.[116]

To study the activity of homogeneous catalysts, a well-established technique known as the volcano plot has been adopted from heterogeneous catalysis. This approach originates from Sabatier's principle, which states that an ideal catalyst should exhibit an optimal level of bond strength with the substrate: neither too weak nor too strong.[93] The energies of reaction intermediates binding to the catalyst are interconnected through scaling relations, creating empirical mathematical relationships. These relationships allow the energies of all reaction intermediate and transition states to be expressed in terms of one or a few specific intermediates, forming linear free energy scaling relationships (LFSERs) based on a descriptor intermediate.[124] By analyzing the computationally calculated energy of the descriptor intermediate, the reaction rate can be estimated, resulting in the characteristic volcano shape. These plots have been used to define thermodynamic and kinetic profiles, aligning with experimental trends, both for smaller and large datasets.[94,125–128] The goal of quickly assessing the performance of prospective catalysts makes volcano plots well suited for big data analytics.[128] Meyer et al. employed a kernel ridge regression-based machine learning model to screen over 25,000 catalyst structures for the Suzuki-Miyaura C–C cross-coupling reaction.[108] They relied on a simplified thermodynamic profile, by using ML to learn the DFT-based reaction energy associated with oxidative addition which had been proven to be a descriptor variable in this catalytic cycle.[108] This approach allowed for rapid discrimination between catalysts with promising or inadequate energy profiles. Although this approach utilizing volcano plots yielded valuable insights into broad trends in catalyst behavior, it was only limited to screening of the catalyst activity. The application of volcano plots to a computational screening of enantioselectivity for TM-based homogeneous catalysts is still limited.[128]

3.2. Mechanism-agnostic approaches

Mechanism-agnostic approaches do not necessitate an understanding of the mechanism or the stereodetermining step for making reactivity or selectivity predictions.[129] For example, a promising approach in homogeneous catalyst design lies in various applications of quantitative structure-selectivity/activity relationships (QSSR/QSAR). A combination of quantum mechanical (QM) and statistical methods is the modern version of the QSAR approaches. In essence, QM-derived descriptors of the molecules are used to identify relationships between the catalyst structure and the observed experimental performance. Often, a general simplified catalyst structure with a "dummy" substrate is used to derive these descriptors. Non-linear black-box models such as random forest, support vector machines, neural networks etc. have found to be successful in the prediction of target values such as reaction yield or enantioselectivity.[130–132] More interpretable white-box univariate or multivariate linear regression have been successfully used in these applications as well.[17,102,129,133–137] A recently reported approach by Dotson et al. showcased an extensive workflow combining ML and HTE for multi-objective optimization.[35] The study focused on catalyst design and optimization, specifically targeting the yield and regioselectivity of chiral bisphosphine ligands. An extensive computational database consisting of 550 ligands was established, where diverse descriptors were computed for each ligand. The study was conducted on a Pd-catalyzed Hayashi–Heck reaction and a Rh-catalyzed alkene hydroformylation reaction. Their methodology was shown to identify ligands with improved regioselectivity by ~1 kcal/mol compared to the previous best ligand. This novel methodology demonstrates the application of ML in addressing the simultaneous improvement of both yield and selectivity.[35] Unfortunately, although the results from the predictive model are readily interpretable, the construction of such a model requires a deep understanding of the calculated descriptors in relation to the reaction at hand. These descriptors depend on the computational catalyst structures, yet a detailed description of the process involved in their selection is often lacking. In addition, if the domain of applicability is limited and automation is minimal, the whole approach needs to be repeated for every addition of new catalysts.

It is well known that the chemical space is too large to be explored without automation. New avenues for understanding the catalytic chemical space were opened by the development of Kraken, a comprehensive platform for mapping and predicting ligand properties.[17] Kraken contains a collection of 300,000 monodentate organophosphorus ligands, accompanied by 190 chemical descriptors that capture their conformational dependence. This mapping endeavor covered a broad range of conceivable structures relevant to organo(transition)metal reactions, providing valuable insights for catalyst design and optimization. The Kraken platform offers researchers access to computed data at different theoretical levels: semi-empirical QM, DFT, and ML. The database includes detailed information on 1,558 organophosphorus compounds, featuring semi-empirical QM and DFT data, computed descriptors and properties, as well as coordinates information for the conformers. Two versions of the compound were simulated, the free ligand and the ligand coordinated to the metal. To digitally represent structures, molecular descriptors are used. Additionally, the platform incorporates ML data, comprising 331,776 entries generated through combinatorial exploration of organophosphorus ligands with up to two distinct substituents. ML models are trained on the DFT dataset, enabling the on-the-fly prediction of properties for an extensive dataset of approximately 191 million distinct organophosphorus compounds. By utilizing the dataset and computational tools provided by Kraken, researchers can optimize reaction process parameters, inspire new ligand choices, and drive the synthesis of novel organophosphorus compounds.[87,134–136] The open-source nature of the Kraken platform and the accessibility of its extensive database facilitate collaboration and encourage contributions from the scientific community. Although this platform fosters ongoing advancements in the field of fully automated homoge-

neous catalyst design, it is currently limited to only monodentate organophosphorus ligands.

3.3. Reaction conditions & real-world data

In the realm of automated homogeneous catalyst design, it is crucial to optimize towards reaction conditions which allow for maximum experimental productivity and efficiency of the catalyst.[138,139] This endeavor can be influenced by several factors such as the experimental error, the number of measured metrics, dataset size and data resolution.[140] Showcased by Shields et al, bayesian optimization (BO) in combination with a mechanism-agnostic approach enables optimization of reaction conditions.[109] The objective was to optimize the yield of the desired product by exploring a combinatorial space of reaction conditions. The performance of their open-source BO framework could then be compared to a selected group of expert chemist. The framework employed different representations of reaction components, such as chemically descriptive fingerprint encodings based on quantum chemical properties computed via DFT, cheminformatics descriptors, and binary one-hot-encoded (OHE) representations generated using the Mordred package.[85] These reaction components were represented as a SMILES string and transformed into different representations using the Auto-Qchem Python package.[141] Remarkably, the BO framework incorporating DFT-derived features outperformed the chemists' expertise. Within the first 15 experiments, the framework consistently achieved higher average performance, yielding over 99% in all cases. The chemists, on the other hand, either prematurely terminated the optimization process or failed to identify the conditions that yielded the highest product yield.

While all the aforementioned approaches serve as exemplary methods, the utilization of non-structured real-world data, e.g. from electronic lab notebooks, for predictive endeavors raises concerns.[142] Determining whether the predictive value obtained is attributed to an inherent structure within the dataset presents a challenge. Additionally, biases can inadvertently manifest during the initial stages of data generation, such as when drawing catalyst structures for subsequent feature extraction or when making assumptions regarding reaction mechanisms. The discussed state-of-the-art approaches are automated to some extent. However, automating all steps from structure representation to prediction could make our work faster, more reproducible, and less prone to human error. The concept of modularity from the field of computer science can be useful here. The modular design of the workflows would mean that there is a logical partitioning of the steps that allows the separate parts to be integrated with easier implementation and maintenance. Though such an integrated workflow is more efficient, it is unfortunately not widely implemented yet in TM-based homogeneous catalyst design.

4. Roadblocks

As introduced in the previous section, a direct catalyst design workflow usually consists of four components: structure generation, QC optimization, descriptor calculation, and a statistical method to relate these descriptors to properties of interest. Integrating these steps into a universally applicable computational workflow that streamlines all aspects, ranging from structure generation to descriptor computation, seems trivial. Nevertheless, the challenge lies in dealing with the interdisciplinary nature of these steps and modeling the variables influencing experimental catalytic performance.

This section will delve into some challenges that encompass key aspects of automation tasks in TM-based homogeneous catalyst design: the representation of catalyst structures in computational workflows, the generation of reliable and diverse descriptors, and the inherent complexity of dynamics in catalysis. We attempt to address these challenges with our in-development tool focused on monodentate and bidentate ligand-containing structures, Open Bidentate Ligand eXplorer (OBeLiX). Various of our in-house developed Python tools are currently integrated into OBeLiX, including stand-alone modules for automated structure generation and subsequent descriptor calculation. With this Python package, we aim to automate and streamline the direct catalyst design workflow.

4.1. Structure representation

Regardless of the QSAR/QSSR approaches being implemented, there are three key parameters central to these workflows, as depicted in Figure 5. These are: 1) the amount of data that is available, both computationally and experimentally, for the objective to be predicted, 2) the interpretability of the prediction model, and the associated computational cost and expertise required, 3) the dimensionality of the computational representation of the catalyst structure.[143]

These components are coupled and ever-changing, but more importantly, they can be a limiting factor. As an example, consider representation dimensionality. A 1D representation can be a SMILES string or one-hot encoded vector, a 2D representation is usually topology-based, in a 3D representation (QM-based) descriptors are derived from a 3D structure, while a 4D representation would also take the conformer ensembles into account. In TM-based complexes, spin, oxidation state, coordinative bonds, and chirality can also be of importance to catalytic performance. Depending on the experimental objective to model, precise structural information from DFT-optimized 3D structures is required in a computational catalyst design workflow.[144] Since the catalyst structure, computational representation, and modeling space are deeply intertwined, it is critical to address and mitigate the biases that could be introduced in a manual structure generation approach.[19,145]

The automation of structure generation from string representations for TM complexes remains an active area of research, highlighting the ongoing efforts to overcome the challenges specific to these systems.[146] Computational packages have

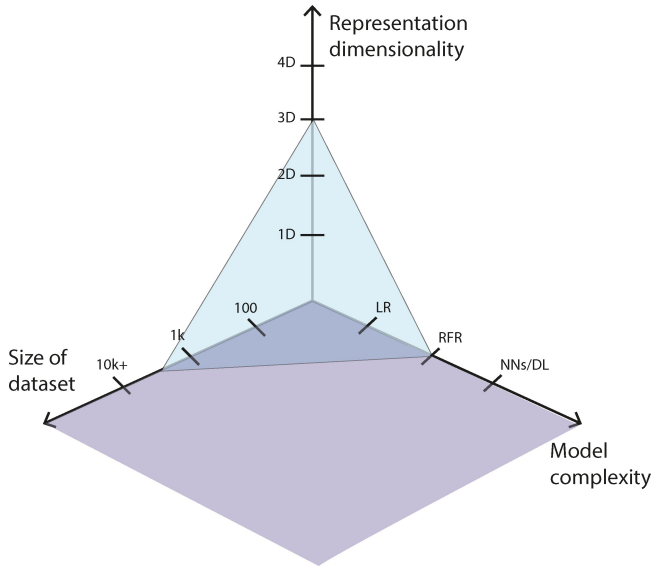


Figure 5. An illustration of three critical parameters in computational catalyst design: the size of the dataset, quantified by the number of available data points; the model complexity, encompassing variations from simple linear regression (LR) to more complex non-linear models such as random forest regression (RFR), neural networks (NN), and other deep learning (DL) approaches; and lastly, the dimensionality of the computational structure representation, indicative of the level of detail captured by the representation. The shaded blue region signifies the size and complexity within which current computational catalyst design studies are mainly conducted.[143]

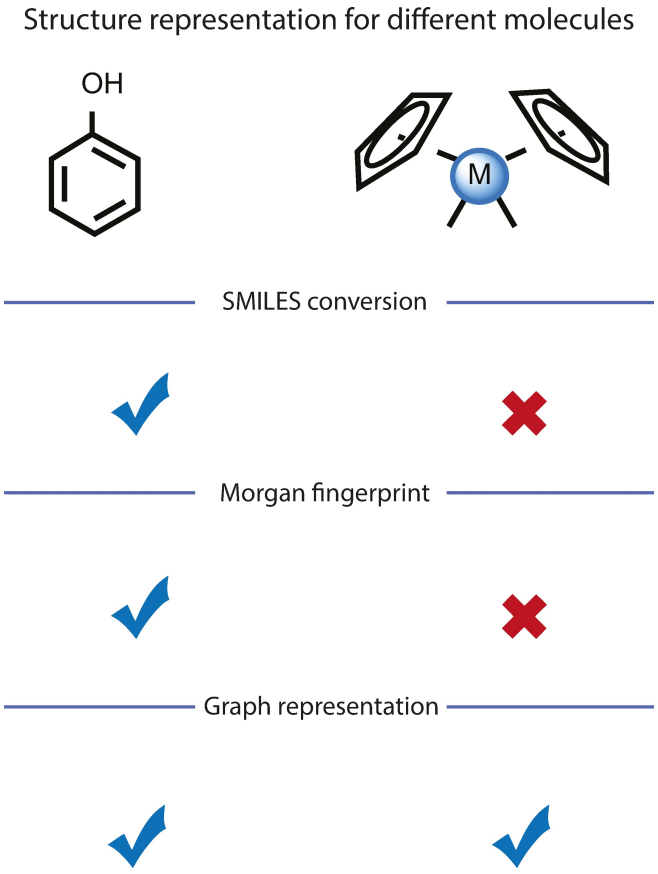


Figure 6. Comparison of structure representations for TM Complexes. The Figure illustrates the challenges of representing TM complexes using different file formats. While SMILES conversion and Morgan fingerprint representations are inadequate for capturing the geometric complexity, the graph representation accurately encodes the 3D structure.

played a pivotal role in enabling various cheminformatics functions, including format conversion and other useful operations. Two widely recognized and extensively utilized tools in this domain are RDKit[84] and OpenBabel,[83] which were introduced in 2013 and 2011 respectively. It is important to note that these tools primarily cater to organic molecules, reflecting the current focus of cheminformatics. However, as the field evolves, it is anticipated that future cheminformatics packages will expand their capabilities to handle more complex molecules and incorporate coordination bonds to cover the broader inorganic and organometallic chemistry landscape. Figure 6 visually presents the challenges of representing transition-metal (TM) complexes in three distinct formats: SMILES conversion, Morgan fingerprint, and graph representations. SMILES, or Simplified Molecular Input Line Entry System, is a concise notation for expressing chemical structures as text strings, offering a human-readable format. Morgan fingerprints, known as circular fingerprints, are a cheminformatics technique encoding molecular features based on substructures within a defined radius, producing a fixed-length binary vector. Graph representations in cheminformatics involve depicting molecules as graphs, portraying atoms as nodes and bonds as edges, capturing connectivity and topology.

These representations have varying degrees of accuracy in capturing the intricate structure of TM complexes, as they have been successfully applied to organic molecules but often fail when applied to coordination complexes. When two coordinative bonds are formed with the metal center as shown in Figure 6, SMILES conversion and Morgan fingerprint representations fail to adequately represent the complex structure (indicated by crosses in both rows). The work of Sobez et al.[49] has demonstrated the effectiveness of graph representations in accurately encoding the 3D structure of TM complexes. While string representations have become commonplace in cheminformatics, they are not yet well-suited for predicting a delicate objective such as enantioselectivity in TM complexes, which is highly sensitive to structural variations.[107] Therefore, utilizing 3D representations for predictive models is desirable.

Building 3D representations of TM complexes is not always straightforward due to the possibility of multiple geometrical isomers, and coordination environment around the metal. Manual generation of these structures can introduce expert-bias by considering the chemical space partly. Utilizing the SMILES representation of components such as substrate, and ligand of the TM-based catalyst complex, a 3D structure can be built and configurationally explored. Two approaches are commonly employed for the automated generation of catalyst structures: 1) an exhaustive search aided by heuristics and 2) searching algorithms aided by computational intelligence. However, neither method can generate perfect structures, and each has its own limitations. While machine learning algorithms may not achieve perfect or near-perfect accuracy, they are

usually well-suited for designing small-scale systems, offering the advantage of speed. The molSimplify package is a notable example of a tool implementing an ML-based optimization tailored towards larger metal-ligand complexes. It employs a DFT-based pre-trained model to determine the skeleton structure, followed by selective force field optimization.[48] This approach allows for user-defined or program-determined ligand positioning on the metal center. DENOPTIM is an additional illustration of a computational intelligence-guided approach that combines fragment-building and genetic algorithms to construct hypothetical complexes with optimized fitness functions.[46] Examples of exhaustive algorithms are the Molassembler and Architector code.[49,50] Molassembler is a software tool that utilizes graph enumeration, stereopermuters, and the distance geometry algorithm to analyze and explore molecular structures, providing insights into connectivity, stereochemistry, and spatial arrangement.[49] It facilitates the generation of isomers and conformers, considers stereoisomeric configurations, and ensures physically realistic structures based on tabulated bond lengths. Architector leverages metal-center symmetry analysis, distance geometry, fragment assembly, and ranking of conformers based on GFN2-xTB energies to capture the diversity of known experimental chemical space and design new complexes.[50]

Our approach to developing a platform for direct design of homogeneous catalysts is centered around enabling chemists to provide drawings of ligands and substrates in a chemically intuitive manner, from which the chemical space is automatically explored. The combined use of our in-house developed tools, MACE and ChemSpaX, facilitates this process as depicted in Figure 7.

MACE is used for generating 3D structures of TM-based metal-ligand scaffolds, starting from a 2D input. It is specifically aimed at conducting exhaustive searches of configurations and stereoisomers in square planar and octahedral complexes. MACE offers the advantage of generating a diverse range of stereoisomers, including those involving configurational isomers where ligands are swapped, while also providing computed energies through force-field calculations to rank these isomers. By incorporating the MACE protocol into computational workflows for organometallic complexes, we enable the expert-bias-free exploration of stereoisomers.[147–149] The introduction of structural variations at an early stage aligns with the complexity observed in real systems, allowing for the identification of likely stereoisomers in a high-throughput manner. When 3D scaffold generation is done, the local chemical space of this structure can be explored by systematically placing substituent groups on the ligand. This approach creates close variations of the ligand structure. In OBeLiX, this is done by utilizing the ChemSpaX package.[150] Overall, these methodologies facilitate the exploration of an extensive chemical space for the screening of potential catalyst structures. Subsequent structural refinement can be achieved through geometry optimization, utilizing methods such as QC at any level of precision.

It is essential to note that current automated 3D structure generation methods do not inherently consider synthesizability or automatically adhere to chemical rules. Typically, manual error checking based on a random sampling is performed. Examples of automated error-checking workflows exist for small organic molecules cheminformatics, where SMILES and synthesizability scores can easily be generated. Such methodologies

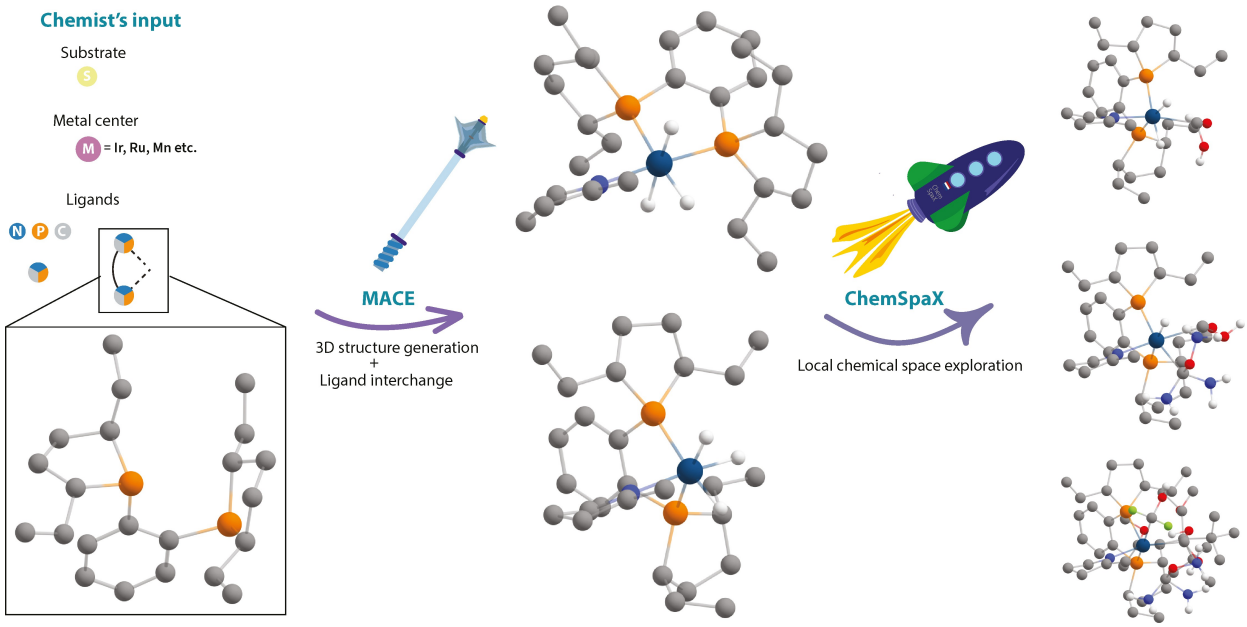


Figure 7. An illustrative example of chemical space exploration of a TM-complex containing a bidentate ligand using MACE and ChemSpaX. The process begins with chemists providing 2D drawings of ligands and substrates. MACE generates 3D structures of TM-based metal-ligand scaffolds, conducting exhaustive searches of stereoisomers. This method, integrated into OBeLiX, enables expert-bias-free exploration of stereoisomers. ChemSpaX further explores the local chemical space by systematically placing substituent groups on the ligand. Together, these tools facilitate 3D structure generation and subsequent high-throughput screening for potential catalyst structures.

could involve 1) a blend of de novo design and synthesis planning or 2) a combination of biased generation with a synthesizability heuristic.[151] In the first approach, the structure generation method undergoes training on the existing data to incorporate knowledge of the synthetic steps involved in compound creation. This approach relies on reactivity rules encoded in a discrete action space of reaction templates, trained on artificial pathways generated from a pool of purchasable compounds and a list of expert-curated templates.[152] The second approach proposes the use of heuristics based on synthesizability to effectively bias generation towards synthetically tractable chemical space.[153] However, it has been observed that this may divert the generative model from its primary optimization objective.[153] In both cases, it is important to note that these automated error-checking methods are still in the development phase and have seen limited application even for small organic molecules. Looking forward, as our understanding of the inorganic chemical space advances, coupled with the availability of more experimental data and enhanced computational representations, automated error-checking approaches may find application in the domain of TM-based homogeneous catalyst design.

4.2. Generation of descriptors

After geometry optimization, relevant electronic, steric, geometric, thermodynamic or combined descriptors can be extracted from the results of computations. Some classic examples such as the TEP, Tolman cone angle, and buried volume were mentioned in the introduction. It is possible that descriptors require manual input, e.g. when defining quadrants/octants of the buried volume for the prediction of enantioselectivity.[154]

An example of a typical buried volume orientation is shown in Figure 8. The direction of axes here is defined with respect to the two donor phosphorus (P) atoms attached to the metal center. The relevance of this directional definition becomes apparent when there is a need to distinguish the quadrant occupied by "P1" from the quadrant occupied by "P2". In the mechanism-agnostic multi-objective optimization approach discussed in the previous section, all atoms of the generated scaffold are manually mapped to define the orientation of this descriptor.[35] The indices of "P1", "P2" and the metal center, among others, are saved in an Excel file. If a new ligand is added to the dataset, this mapping has to be redone. To address this challenge, we employ a graph-based method to identify the ligand donor atoms using interatomic distances in the OBeLiX platform. These donor atoms are then numbered based on their charge and this definition is used in subsequent descriptor calculation. By leveraging this automated approach, we aim to reduce manual input in the calculation of descriptors, ensuring a more objective and efficient exploration of structures in catalyst design. This becomes especially relevant for smaller datasets, where the sensitivity of descriptors can impact the performance of predictive models.[25,39,155]

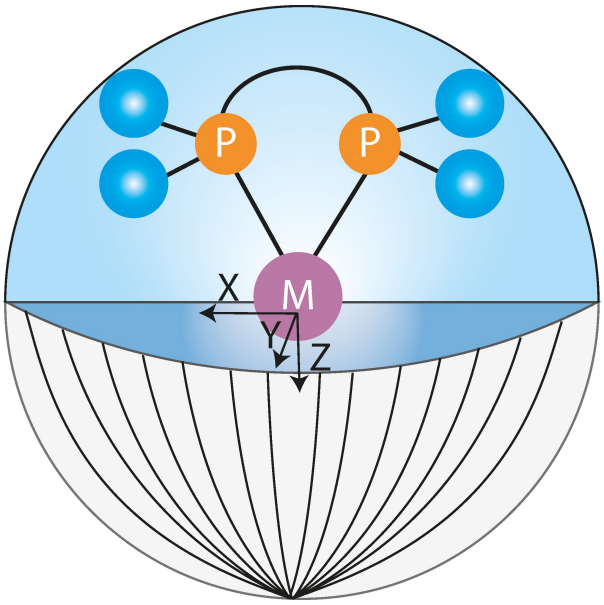


Figure 8. Illustration depicting a representative orientation of a buried volume, defined with respect to the metal center and bidentate ligand donors. Typically, the buried volume is quantified at a designated radius and expressed as a percentage relative to the ligand occupying the encompassing sphere. Furthermore, the contributions of distinct quadrants and octants can be assessed by establishing a 3D axis and subdividing the sphere into discrete sections.

In addition to addressing biases within the descriptor generation part of catalyst design, it is important to acknowledge the challenges faced in the experimental domain which influence any predictive capabilities. While the use of AI has gained significant attention in various fields, its implementation in chemistry is still evolving.[156] Glorius et al. have highlighted the impact of systematic errors on dataset balance and completeness, which can severely limit the reliability of ML-based predictions.[157] In the context of direct design of catalysts, the objective is to create structures with desired properties. However, the accuracy of predictive models heavily relies on the quality of the training data.[158] If the experimentally measured target property is prone to significant errors or bias, it can introduce difficulties.[158] To address these limitations, it has been suggested that systematic reporting, including the documentation of underperforming reactions, can mitigate errors and improve dataset quality.[158,159] Additionally, due to the rapid advancements in ML, it may be more efficient to develop new algorithms that can overcome the challenges associated with unbalanced and missing data, facilitating accurate predictions of quantitative properties.[160]

4.3. The complexity of catalysis

A computer sees the catalytic system through the prism of the models and methods designed by humans, which are usually far from mimicking the complex chemical interactions in a real system. Based on the chosen settings and methods, DFT-based modeling might lack certain crucial aspects that are relevant at

both micro and macro scales.[53] Real catalytic systems exhibit complexity arising from various factors, including solvent effects, conformational variations, and catalyst deactivation.[139] Modelling these complex phenomena might necessitate a combination of mechanism-agnostic and mechanism-based approaches. For example, in addition to the descriptors calculated on a general simplified catalyst structure, as is often done in the mechanism-agnostic approaches, those derived from specific TSs or reaction intermediates would make the model more realistic.[116,161] Mechanistic studies in catalysis are usually conducted using DFT methods,[162] but the computational cost of exploring all possible reaction paths considering the aforementioned factors can be prohibitively high. Hence, a form of conformer searching or reaction network exploration based on semi-empirical methods might prove useful.[67,72,163–165] In such a conformer search, the presence of multiple conformations may itself be an important descriptor relating to the catalytic performance.[25,166] To enable such a high-throughput 4D-QSAR/QSSR approach, automated descriptor calculation should be facilitated on a conformer ensemble. In that context, OBeLiX workflow uses the cclib and Morfeus packages to calculate descriptors on DFT outputs, CREST conformer ensembles or XYZ files of TM complexes with monodentate and bidentate ligands.[17,167]

5. Summary

The scientific landscape has undergone a paradigmatic transformation with the emergence of powerful large language models such as GPT-3. These models showcase unprecedented capabilities, demonstrating proficiency in tasks ranging from crafting poetry to programming, rivaling and even surpassing human performance. In chemistry and catalysis however, AI approaches are not as successful for the understanding of the principles underlying molecular design. Computational homogeneous catalyst design is limited by the scarcity of high-quality data, the complexity of catalytic reactions and minimal automation. Despite the faced challenges, there are vast opportunities for catalyst discovery by combining computational chemistry, automation and AI.

The definition of descriptors in catalytic reactions is complex, requiring a thorough understanding of the involved dynamics and mechanisms. While high-throughput *in silico* chemical space exploration and analysis provides valuable insights, the key lies in automated and modular workflows. Through the creation of OBeLiX, our aim is to democratize the endeavors of the data-driven catalysis community, paving the way for a future marked by *in silico* high-throughput exploration of the catalytic chemical space, particularly in the realm of TM-based homogeneous catalysis.

Author Contributions

A.V. Kalikadien: Investigation, Conceptualization, Visualization, Writing – Original Draft, Writing – Review & Editing, Project administration **A. Mirza**: Investigation, Conceptualization, Visualization, Writing - Original Draft **A. Najl Hossaini**: Visualization, Writing - Original Draft **A. Sreenithya**: Conceptualization, Writing – Original Draft, Writing – Review & Editing **E.A. Pidko**: Supervision, Conceptualization, Resources, Funding acquisition, Writing – Review & Editing, Project administration

Acknowledgements

The authors acknowledge the financial support provided by Janssen Pharmaceutica. A.S. and E.A.P. thank Advanced Research Center Chemical Building Blocks Consortium (ARC CBBC) for support under the project number 2021.038.C. The authors thank the NWO Domein Exacte en Natuurwetenschappen for the use of the national supercomputer, Snellius. The authors thank Ivan Yu. Chernyshov for his pivotal role in developing MACE. Finally, the authors are grateful for the insightful discussions with Dr. Mikko Muuronen.

Conflict of Interests

The authors have no conflict of interest to declare.

Data Availability Statement

The OBeLiX Python package will soon be accessible through our GitHub repository (https://github.com/epics-group). Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Keywords: automation · catalysis · cheminformatics · machine learning · quantum chemistry

[1] M. L. Crawley, B. M. Trost, *Applications of Transition Metal Catalysis in Drug Discovery and Development: An Industrial Perspective*, John Wiley and Sons **2012**.
[2] R. A. Fernandes, A. K. Jha, P. Kumar, *Catal. Sci. Technol.* **2020**, *10*, 7448.
[3] A. I. Green, C. P. Tinworth, S. Warriner, A. Nelson, N. Fey, *Chemistry* **2021**, *27*, 2402.
[4] B. L. Tran, S. I. Johnson, K. P. Brooks, S. T. Autrey, *ACS Sustainable Chem. Eng.* **2021**, *9*, 7130.
[5] W. Kuriyama, T. Matsumoto, O. Ogata, Y. Ino, K. Aoki, S. Tanaka, K. Ishida, T. Kobayashi, N. Sayo, T. Saito, *Org. Process Res. Dev.* **2012**, *16*, 166.
[6] W. Keim, *Concepts for the Use of Transition Metals in Industrial Fine Chemical Synthesis*, Wiley-VCH Verlag GmbH **2008**.
[7] P. Kirkpatrick, C. Ellis, *Nature* **2004**, *432*, 823.
[8] F. I. Saldívar-González, B. A. Pilón-Jiménez, J. L. Medina-Franco, *Phys. Sci. Rev.* **2019**, *4*, 20180103.
[9] L. P. Hammett, *Chem. Rev.* **1935**, *17*, 125.
[10] L. P. Hammett, *J. Am. Chem. Soc.* **1937**, *59*, 96.
[11] L. P. Hammett, *Trans. Faraday Soc.* **1938**, *34*, 156.
[12] R. P. Bell, C. N. Hinshelwood, *Proc. Roy. Soc. A* **1997**, *154*, 414.
[13] M. G. Evans, M. Polanyi, *Trans. Faraday Soc.* **1938**, *34*, 11.
[14] H. Gerischer, *Bull. Soc. Chim. Belg.* **1958**, *67*, 506.
[15] R. Parsons, *Trans. Faraday Soc.* **1958**, *54*, 1053.
[16] H. U. Blaser, B. Pugin, F. Spindler, L. A. Saudan, *Hydrogenation*, Wiley Online Books **2017**.

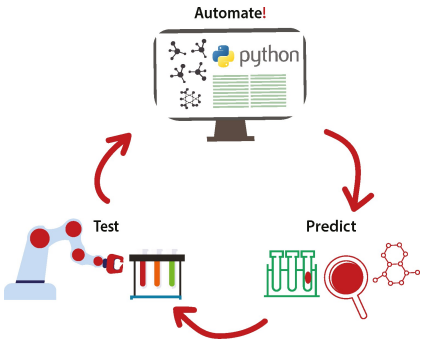
[17] T. Gensch, G. D. P. Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario, M. S. Sigman, A. Aspuru-Guzik, *J. Am. Chem. Soc.* **2022**, *144*, 1205.
[18] A. G. Maldonado, G. Rothenberg, *Chem. Soc. Rev.* **2010**, *39*, 1891.
[19] J. Hageman, J. Westerhuis, H.-W. Frejhauf, G. Rothenberg, *Adv. Synth. Catal.* **2006**, *348*, 361.
[20] M. Foscato, G. Occhipinti, V. Venkatraman, B. K. Alsberg, V. R. Jensen, *J. Chem. Inf. Model.* **2014**, *54*, 767.
[21] M. Foscato, V. R. Jensen, *ACS Catal.* **2020**, *10*, 2354.
[22] L. C. Burrows, L. T. Jesikiewicz, G. Lu, S. J. Geib, P. Liu, K. M. Brummond, *J. Am. Chem. Soc.* **2017**, *139*, 15022.
[23] D.-H. Kwon, J. T. Fuller, U. J. Kilgore, O. L. Sydora, S. M. Bischof, D. H. Ess, *ACS Catal.* **2018**, *8*, 1138.
[24] A. Nandy, C. Duan, M. G. Taylor, F. Liu, A. H. Steeves, H. J. Kulik, *Chem. Rev.* **2021**, *121*, 9927.
[25] L. C. Gallegos, G. Luchini, P. C. S. John, S. Kim, R. S. Paton, *Acc. Chem. Res.* **2021**, *54*, 827.
[26] C. A. Tolman, *Chem. Rev.* **1977**, *77*, 313.
[27] D. White, B. C. Taverner, P. G. Leach, N. J. Coville, *J. Comput. Chem.* **1993**, *14*, 1042.
[28] H. Clavier, S. P. Nolan, *Chem. Commun.* **2010**, *46*, 841.
[29] D. J. Durand, N. Fey, *Chem. Rev.* **2019**, *119*, 6561.
[30] R. J. Lundgren, M. Stradiotto, *Key Concepts in Ligand Design*, John Wiley Sons, Ltd **2016**.
[31] G. Skoraczyăśki, P. DIttwald, B. Miasojedow, S. Szymkuc, E. P. Gajewska, B. A. Grzybowski, A. Gambin, *Sci. Rep.* **2017**, *7*, 1.
[32] K. E. Koenig, M. J. Sabacky, G. L. Bachman, W. C. Christopfel, H. D. Bamstorff, R. B. Friedman, W. S. Knowles, B. R. Stults, B. D. Vineyard, D. J. Weinkauff, *Ann. N. Y. Acad. Sci.* **1980**, *333*, 16.
[33] P. W. V. Leeuwen, P. C. Kamer, J. N. Reek, P. Dierkes, *Chem. Rev.* **2000**, *100*, 2741.
[34] P. W. V. Leeuwen, P. C. Kamer, J. N. Reek, *Pure Appl. Chem.* **1999**, *71*, 1443.
[35] J. J. Dotson, L. van Dijk, J. C. Timmerman, S. Grosslight, R. C. Walroth, F. Gosselin, K. Pāijntener, K. A. Mack, M. S. Sigman, *J. Am. Chem. Soc.* **2022**, *145*, 110.
[36] S.-S. Chen, Z. Meyer, B. Jensen, A. Kraus, A. Lambert, D. H. Ess, *J. Chem. Inf. Model.* **2023**, *63*, 7412.
[37] N. Fey, A. Koumi, A. V. Malkov, J. D. Moseley, B. N. Nguyen, S. N. Tyler, C. E. Willans, *Dalton Trans.* **2020**, *49*, 8169.
[38] P. Friederich, G. dos Passos Gomes, R. D. Bin, A. Aspuru-Guzik, D. Balcells, *Chem. Sci.* **2020**, *11*, 4584.
[39] J. P. Janet, H. J. Kulik, *J. Phys. Chem. A* **2017**, *121*, 8939.
[40] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, A. A. Lee, *ACS Cent. Sci.* **2019**, *5*, 1572.
[41] S. Singh, R. B. Sunoj, *Digital Discovery* **2022**, *1*, 303.
[42] D. Weininger, *J. Chem. Inf. Comput.* **1988**, *28*, 31.
[43] N. O'Boyle, A. Dalke, *ChemRxiv* **2018**.
[44] M. Krenn, Q. Ai, S. Barthel, N. Carson, A. Frei, N. C. Frey, P. Friederich, T. Gaudin, A. A. Gayle, K. M. Jablonka, R. F. Lameiro, D. Lemm, A. Lo, S. M. Moosavi, J. M. Nāpoles-Duarte, A. Nigam, R. Pollice, K. Rajan, U. Schatzschneider, P. Schwaller, M. Skreta, B. Smit, F. Strieth-Kalthoff, C. Sun, G. Tom, G. Falk von Rudorff, A. Wang, A. D. White, A. Young, R. Yu, A. Aspuru-Guzik, *Patterns* **2022**, *3*, 100588.
[45] A. V. Brethomé, S. P. Fletcher, R. S. Paton, *ACS Catal.* **2019**, *9*, 2313.
[46] M. Foscato, V. Venkatraman, V. R. Jensen, *J. Chem. Inf. Model.* **2019**, *59*, 32.
[47] Y. Guan, V. M. Ingman, B. J. Rooks, S. E. Wheeler, *J. Chem. Theory Comput.* **2018**, *14*, 5249.
[48] E. I. Ioannidis, T. Z. H. Gani, H. J. Kulik, *J. Comput. Chem.* **2016**, *37*, 2106.
[49] J. G. Sobez, M. Reiher, *J. Chem. Inf. Model.* **2020**, *60*, 3884.
[50] M. G. Taylor, D. J. Burrill, J. Janssen, E. R. Batista, D. Perez, P. Yang, *Nat. Commun.* **2023**, *14*, 1.
[51] R. Haunschild, A. Barth, B. French, *J. Cheminf.* **2019**, *11*, 72.
[52] B. Huang, G. F. von Rudorff, O. A. von Lilienfeld, *Science* **2023**, *381*, 170.
[53] E. A. Pidko, *ACS Catal.* **2017**, *7*, 4230.
[54] M. Besora, F. Maseras, *WIREs Comput. Mol. Sci.* **2018**, *8*, e1372.
[55] W. M. Sameera, F. Maseras, *WIREs Comput. Mol. Sci.* **2012**, *2*, 375.
[56] J. Slater, *Phys. Rev.* **1951**, *81*, 385.
[57] W. Kohn, L. J. Sham, *Phys. Rev.* **1965**, *140*, A1133.
[58] J. P. Perdew, *Phys. Rev. B* **1986**, *33*, 8822.
[59] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865.
[60] A. D. Becke, *Phys. Rev. A* **1988**, *38*, 3098.
[61] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785.
[62] J. Tao, J. P. Perdew, V. N. Staroverov, G. E. Scuseria, *Phys. Rev. Lett.* **2003**, *91*, 146401.
[63] J. P. Perdew, S. Kurth, A. Zupan, P. Blaha, *Phys. Rev. Lett.* **1999**, *82*, 2544.
[64] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 1372.
[65] S. Grimme, *WIREs Comput. Mol. Sci.* **2011**, *1*, 211.
[66] M. Bursch, H. Neugebauer, S. Grimme, *Angew. Chem.* **2019**, *131*, 11195.
[67] M. Bursch, A. Hansen, P. Pracht, J. T. Kohn, S. Grimme, *Phys. Chem. Chem. Phys.* **2021**, *23*, 287.
[68] S. Spicher, S. Grimme, *Angew. Chem. Int. Ed.* **2020**, *59*, 15665.
[69] L. Hu, U. Ryde, *J. Chem. Theory Comput.* **2011**, *7*, 2452.
[70] I. Iribarren, C. Trujillo, *J. Chem. Inf. Model.* **2022**, *62*, 5568.
[71] L. Wilbraham, E. Berardo, L. Turcani, K. E. Jelfs, M. A. Zwijnenburg, *J. Chem. Inf. Model.* **2018**, *58*, 2450.
[72] P. Pracht, F. Bohle, S. Grimme, *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169.
[73] S. Grimme, *J. Chem. Theory Comput.* **2019**, *15*, 2847.
[74] J. Behler, *J. Chem. Phys.* **2016**, *145*, 170901.
[75] J. Behler, G. Csányi, *Eur. Phys. J. B* **2021**, *94*, 1.
[76] Y. Mishin, *Acta Mater.* **2021**, *214*, 116980.
[77] R. Nagai, R. Akashi, O. Sugino, *npj Comput. Mater.* **2020**, *6*, 1.
[78] M. Eckhoff, M. Reiher, *J. Chem. Theory Comput.* **2023**, *19*, 3509.
[79] F. K. Brown, Chapter 35 – Chemoinformatics: What is it and How does it Impact Drug Discovery., vol. 33 of *Annu. Rep. Med. Chem.*, pages 375–384, Academic Press **1998**.
[80] G. dos Passos Gomes, R. Pollice, A. Aspuru-Guzik, *Trends Chem.* **2021**, *3*, 96.
[81] P. Polishchuk, *J. Chem. Inf. Model.* **2017**, *57*, 2618.
[82] G. Van Rossum, F. L. Drake Jr, *Python reference manual*, Centrum voor Wiskunde en Informatica Amsterdam **1995**.
[83] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, *J. Cheminf.* **2011**, *3*, 33.
[84] G. Landrum, RDKit: Open-source cheminformatics **2020**.
[85] H. Moriwaki, Y.-S. Tian, N. Kawashita, T. Takagi, *J. Cheminf.* **2018**, *10*, 4.
[86] G. Luchini, T. Patterson, R. Paton, *Zenodo* **2023**.
[87] T. Gensch, S. R. Smith, T. J. Colacot, Y. N. Timsina, G. Xu, B. W. Glasspoole, M. S. Sigman, *ACS Catal.* **2022**, *12*, 7773.
[88] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **2018**, *360*, 186.
[89] A. M. Źurański, J. I. M. Alvarado, B. J. Shields, A. G. Doyle, *Acc. Chem. Res.* **2021**, *54*, 1856.
[90] P. Schwaller, A. C. Vaucher, T. Laino, J. L. Reymond, *Mach. Learn.: Sci. Technol.* **2021**, *2*, 015016.
[91] E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov, A. Tropsha, *Chem. Soc. Rev.* **2020**, *49*, 3525.
[92] C. Hansch, T. Fujita, *J. Biol. Chem.* **1964**, *39*, 284.
[93] P. Sabatier, *La catalyse en chimie organique, Librairie Polytechnique*, Ch. Béranger Paris et Liège **1913**.
[94] M. Busch, M. D. Wodrich, C. Corminboeuf, *ACS Catal.* **2017**, *7*, 5643.
[95] A. Hoque, R. B. Sunoj, *Digital Discovery* **2022**, *1*, 926.
[96] O. Loyola-Gonzalez, *IEEE Access* **2019**, *7*, 154096.
[97] C. Kuhn, D. N. Beratan, *J. Phys. Chem.* **1996**, *100*, 10595.
[98] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360.
[99] J. G. Freeze, H. R. Kelly, V. S. Batista, *Chem. Rev.* **2019**, *119*, 6595.
[100] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268.
[101] O. Schilter, A. Vaucher, P. Schwaller, T. Laino, *Digital Discovery* **2023**, *2*, 728.
[102] M. S. Sigman, K. C. Harper, E. N. Bess, A. Milo, *Acc. Chem. Res.* **2016**, *49*, 1292.
[103] A. Hamza, G. Schubert, T. Soós, I. Pápai, *J. Am. Chem. Soc.* **2006**, *128*, 13151.
[104] J. G. Estrada, D. T. Ahneman, R. P. Sheridan, S. D. Dreher, A. G. Doyle, *Science* **2018**, *362*, eaat8763.
[105] M. P. Maloney, B. A. Stenfors, P. Helquist, P.-O. Norrby, O. Wiest, *ACS Catal.* **2023**, *13*, 14285.
[106] M. B. Patrascu, J. Pottel, S. Pinus, M. Bezanson, P.-O. Norrby, N. Moitessier, *Nat. Catal.* **2020**, *3*, 574.
[107] R. Laplaza, J. G. Sobez, M. D. Wodrich, M. Reiher, C. Corminboeuf, *Chem. Sci.* **2022**, *13*, 6858.
[108] B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld, C. Corminboeuf, *Chem. Sci.* **2018**, *9*, 7069.
[109] B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, A. G. Doyle, *Nature* **2021**, *590*, 89.

[110] P. L. Bhoorasingh, R. H. West, *Phys. Chem. Chem. Phys.* **2015**, *17*, 32173.
[111] P. L. Bhoorasingh, B. L. Slakman, F. S. Khanshan, J. Y. Cain, R. H. West, *J. Phys. Chem. A* **2017**, *121*, 6896.
[112] L. D. Jacobson, A. D. Bochevarov, M. A. Watson, T. F. Hughes, D. Rinaldo, S. Ehrlich, T. B. Steinbrecher, S. Vaitheeswaran, D. M. Philipp, M. D. Halls, R. A. Friesner, *J. Chem. Theory Comput.* **2017**, *13*, 5780.
[113] V. M. Ingman, A. J. Schaefer, L. R. Andreola, S. E. Wheeler, *WIREs Comput. Mol. Sci.* **2021**, *11*, e1510.
[114] S. Chen, T. Nielson, E. Zalit, B. B. Skjelstad, B. Borough, W. J. Hirschi, S. Yu, D. Balcells, D. H. Ess, *Top. Catal.* **2022**, *65*, 312.
[115] F. Jensen, P.-O. Norrby, *Theor. Chem. Acc.* **2003**, *109*, 1.
[116] A. R. Rosales, T. R. Quinn, J. Wahlers, A. Tomberg, X. Zhang, P. Helquist, O. Wiest, P.-O. Norrby, *Chem. Commun.* **2018**, *54*, 8294.
[117] L. W. Chung, W. M. C. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. Ke, F. Liu, H.-B. Li, L. Ding, K. Morokuma, *Chem. Rev.* **2015**, *115*, 5678.
[118] M. T. Reetz, A. Meiswinkel, G. Mehler, K. Angermund, M. Graf, W. Thiel, R. Mynott, D. G. Blackmond, *J. Am. Chem. Soc.* **2005**, *127*, 10305.
[119] S. Feldgus, C. R. Landis, *J. Am. Chem. Soc.* **2000**, *122*, 12714.
[120] N. L. Allinger, Y. H. Yuh, J. H. Lii, *J. Am. Chem. Soc.* **1989**, *111*, 8551.
[121] J. Wahlers, A. R. Rosales, N. Berkel, A. Forbes, P. Helquist, P.-O. Norrby, O. Wiest, *J. Org. Chem.* **2022**, *87*, 12334.
[122] J. M. Seminario, *Int. J. Quantum Chem.* **1996**, *60*, 1271.
[123] A. E. A. Allen, M. C. Payne, D. J. Cole, *J. Chem. Theory Comput.* **2018**, *14*, 274.
[124] M. Anand, J. K. Nørskov, *ACS Catal.* **2020**, *10*, 336.
[125] M. Busch, M. D. Wodrich, C. Corminboeuf, *Chem. Sci.* **2015**, *6*, 6754.
[126] M. D. Wodrich, M. Busch, C. Corminboeuf, *Chem. Sci.* **2016**, *7*, 5723.
[127] M. D. Wodrich, B. Sawatlon, E. Solel, S. Kozuch, C. Corminboeuf, *ACS Catal.* **2019**, *9*, 5716.
[128] M. D. Wodrich, B. Sawatlon, M. Busch, C. Corminboeuf, *Acc. Chem. Res.* **2021**, *54*, 1107.
[129] J. P. Reid, M. S. Sigman, *Nature* **2019**, *571*, 343.
[130] S. Singh, M. Pareek, A. Changotra, S. Banerjee, B. Bhaskararao, P. Balamurugan, R. B. Sunoj, *Proc. Nat. Acad. Sci.* **2020**, *117*, 1339.
[131] L. C. Xu, S. Q. Zhang, X. Li, M. J. Tang, P. P. Xie, X. Hong, *Angew. Chem. Int. Ed.* **2021**, *60*, 22804.
[132] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science* **2019**, *363*, eaau5631.
[133] S. H. Newman-Stonebraker, S. R. Smith, J. E. Borowski, E. Peters, T. Gensch, H. C. Johnson, M. S. Sigman, A. G. Doyle, *Science* **2021**, *374*, 301.
[134] D. Zell, C. Kingston, J. Jermaks, S. R. Smith, N. Seeger, J. Wassmer, L. E. Sirois, C. Han, H. Zhang, M. S. Sigman, F. Gosselin, *J. Am. Chem. Soc.* **2021**, *143*, 19078.
[135] M. Christensen, L. P. E. Yunker, F. Adedeji, F. Häse, L. M. Roch, T. Gensch, G. dos Passos Gomes, T. Zepel, M. S. Sigman, A. Aspuru-Guzik, J. E. Hein, *Commun. Chem.* **2021**, *4*, 112.
[136] J. M. Crawford, T. Gensch, M. S. Sigman, J. M. Elward, J. E. Steves, *Org. Process Res. Dev.* **2022**, *26*, 1115.
[137] W. Matsuoka, Y. Harabuchi, S. Maeda, *ACS Catal.* **2022**, *12*, 3752.
[138] P. O. Kuliaev, E. A. Pidko, *ChemCatChem* **2020**, *12*, 795.
[139] W. Yang, G. A. Filonenko, E. A. Pidko, *Chem. Commun.* **2023**, *59*, 1757.
[140] N. Jiscoot, E. A. Uslamin, E. A. Pidko, *Digital Discovery* **2023**, *2*, 994.
[141] A. M. Żurański, J. Y. Wang, B. J. Shields, A. G. Doyle, *React. Chem. Eng.* **2022**, *7*, 1276.
[142] M. Saebi, B. Nan, J. E. Herr, J. Wahlers, Z. Guo, A. M. Zurański, T. Kogej, P.-O. Norrby, A. G. Doyle, N. V. Chawla, O. Wiest, *Chem. Sci.* **2023**, *14*, 4997.
[143] H. J. Kulik, *WIREs Comput. Mol. Sci.* **2020**, *10*, e1439.
[144] D. S. Wigh, J. M. Goodman, A. A. Lapkin, *WIREs Comput. Mol. Sci.* **2022**, *12*, e1603.
[145] J. P. Janet, H. J. Kulik, *Chem. Sci.* **2017**, *8*, 5137.
[146] J. Jensen, GitHub – jensengroup/xyz2 mol: Converts an xyz file to an RDKit mol object.
[147] W. Yang, I. Y. Chernyshov, R. K. A. van Schendel, M. Weber, C. Müller, G. A. Filonenko, E. A. Pidko, *Nat. Commun.* **2021**, *12*, 1.
[148] W. Yang, T. Y. Kalavalapalli, A. M. Krieger, T. A. Khvorost, I. Y. Chernyshov, M. Weber, E. A. Uslamin, E. A. Pidko, G. A. Filonenko, *J. Am. Chem. Soc.* **2022**, *144*, 8129.
[149] W. Yang, I. Y. Chernyshov, M. Weber, E. A. Pidko, G. A. Filonenko, *ACS Catal.* **2022**, *12*, 10818.
[150] A. V. Kalikadien, E. A. Pidko, V. Sinha, *Digital Discovery* **2022**, *1*, 8.
[151] B. Goldman, S. Kearnes, T. Kramer, P. Riley, W. P. Walters, *J. Med. Chem.* **2022**, *65*, 7073.
[152] W. Gao, R. Mercado, C. W. Coley, *arXiv* **2022**.
[153] W. Gao, C. W. Coley, *J. Chem. Inf. Model.* **2020**, *60*, 5714.
[154] G. Antinucci, B. Dereli, A. Vittoria, P. H. Budzelaar, R. Cipullo, G. P. Goryunov, P. S. Kulyabin, D. V. Uborsky, L. Cavallo, C. Ehm, A. Z. Voskoboynikov, V. Busico, *ACS Catal.* **2022**, *12*, 6934.
[155] X. Jia, A. Lynch, Y. Huang, M. Danielson, I. Langât, A. Milder, A. E. Ruby, H. Wang, S. A. Friedler, A. J. Norquist, J. Schrier, *Nature* **2019**, *573*, 251.
[156] J. M. Cole, *Nature* **2023**, *617*, 438.
[157] F. Strieth-Kalthoff, F. Sandfort, M. Kühnemund, F. R. Schäfer, H. Kuchen, F. Glorius, *Angew. Chem. Int. Ed.* **2022**, *61*, e202204647.
[158] T. Taniike, K. Takahashi, *Nat. Catal.* **2023**, *6*, 108.
[159] S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen, C. W. Coley, *J. Am. Chem. Soc.* **2021**, *143*, 18820.
[160] J. M. Cole, *Nat. Chem.* **2022**, *14*, 973.
[161] J. D. Oslob, B. Åkermark, P. Helquist, P.-O. Norrby, *Organometallics* **1997**, *16*, 3015.
[162] H. Ryu, J. Park, H. K. Kim, J. Y. Park, S.-T. Kim, M.-H. Baik, *Organometallics* **2018**, *37*, 3228.
[163] S. Grimme, *J. Chem. Theory Comput.* **2019**, *15*, 2847.
[164] A. Hashemi, S. Bougueroua, M.-P. Gaigeot, E. A. Pidko, *J. Chem. Theory Comput.* **2022**, *18*, 7470.
[165] A. Hashemi, S. Bougueroua, M.-P. Gaigeot, E. A. Pidko, *J. Chem. Inf. Model.* **2023**, *63*, 6081.
[166] J. Crawford, M. Sigman, *Synthesis* **2019**, *51*, 1021.
[167] N. M. O'boyle, A. L. Tenderholt, K. M. Langner, *J. Comput. Chem.* **2008**, *29*, 839.

Manuscript received: November 29, 2023
Revised manuscript received: December 20, 2023
Version of record online: ■■, ■■

PERSPECTIVE

Computational tools have become integral to catalyst design, providing crucial support for experimental synthesis. This Perspective explores diverse initiatives in computational catalyst design, emphasizing the rise of data-driven methods and machine learning. Additionally, we introduce our automated tools tailored for high-throughput *in silico* exploration, highlighting the pivotal role of integrated data-driven, automated workflows in advancing homogeneous catalyst design and catalysis research.



*A. V. Kalikadien, A. Mirza, A. N. Hossaini, Dr. A. Sreenithya, Prof. Dr. E. A. Pidko**

1 – 15

Paving the road towards automated homogeneous catalyst design