Topics in Catalysis (2022) 65:6–39
https://doi.org/10.1007/s11244-021-01543-9

ORIGINAL PAPER

Autonomous Reaction Network Exploration in Homogeneous and Heterogeneous Catalysis

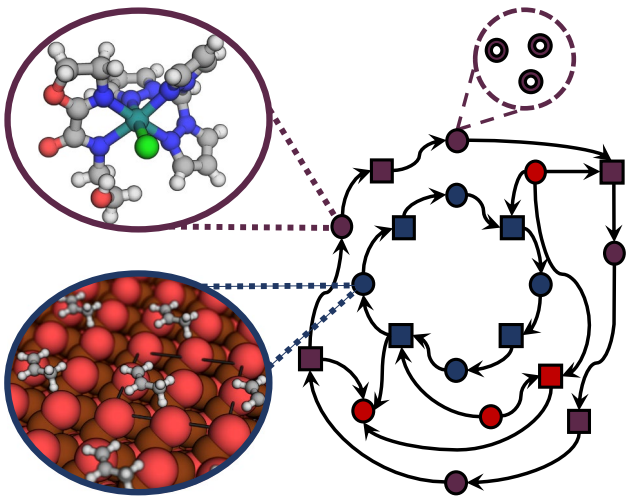
Miguel Steiner[1] · **Markus Reiher**[1]

Accepted: 17 November 2021 / Published online: 13 January 2022

Abstract

Autonomous computations that rely on automated reaction network elucidation algorithms may pave the way to make computational catalysis on a par with experimental research in the field. Several advantages of this approach are key to catalysis: (i) automation allows one to consider orders of magnitude more structures in a systematic and open-ended fashion than what would be accessible by manual inspection. Eventually, full resolution in terms of structural varieties and conformations as well as with respect to the type and number of potentially important elementary reaction steps (including decomposition reactions that determine turnover numbers) may be achieved. (ii) Fast electronic structure methods with uncertainty quantification warrant high efficiency and reliability in order to not only deliver results quickly, but also to allow for predictive work. (iii) A high degree of autonomy reduces the amount of manual human work, processing errors, and human bias. Although being inherently unbiased, it is still steerable with respect to specific regions of an emerging network and with respect to the addition of new reactant species. This allows for a high fidelity of the formalization of some catalytic process and for surprising in silico discoveries. In this work, we first review the state of the art in computational catalysis to embed autonomous explorations into the general field from which it draws its ingredients. We then elaborate on the specific conceptual issues that arise in the context of autonomous computational procedures, some of which we discuss at an example catalytic system.

Graphical Abstract



Keywords Computational catalysis · Reaction mechanism elucidation · Autonomous computational campaigns · Exhaustive quantum chemical exploration · Catalytic reactivity principles

✉ Markus Reiher
markus.reiher@phys.chem.ethz.ch

Extended author information available on the last page of the article

1 Introduction

Catalysis is a key and emergent concept in chemistry: substances are assigned a special role as they take part in a reaction but are eventually recovered unchanged after a product has been formed. It is a chemical insight that such patterns can be discovered in complex reaction mechanisms. From a quantum chemical point of view, this translates into producing and then analyzing networks of elementary steps, which map all (with respect to external conditions such as temperature) feasible chemical transformations in a sequence of structural changes across a Born-Oppenheimer potential energy surface. Understanding catalysis in terms of such reaction networks can then be a starting point for the design of processes guided by constraints such as being efficient, cheap, green, and/or sustainable.

Computational catalysis can deliver unprecedented details about catalytic reaction mechanisms [1–33]. However, a universal theoretical approach toward computational catalysis with generally applicable algorithms is not available. This can be a handicap for practical applications, especially in view of the growing field of experimental catalysis with increasingly complex catalyst structures such as metal-organic frameworks [34–37], single-atom catalysts [38–40], supported nanoparticles [41], supported organometallic catalysts [42–45], binary catalysts [46], encapsulated catalysts [47–50], self-assembling nanostructures [51, 52], nanozymes [53], protein nanocages [54], nucleic-acid catalysts [55], and artificial (metallo)enzymes [56–59].

In this work, we first provide a brief overview of the different computational approaches that have been developed for applications in catalysis and in related fields, before we then focus on the detailed first-principles modeling of vast elementary reaction networks. It is the very nature of this complex topic that requires us to touch upon many diverse subjects. While we attempt to provide a balanced overview with a focus on the most recent developments, we emphasize that a complete literature review will be impossible to achieve in the context of this work. We therefore consider the numerous references given here as a starting point for interested readers to dive deeper into the literature of a specific subject. Eventually, we will focus on automated procedures steered by autonomously working computer (meta) programs. Such approaches will have a broad future for various reasons to be discussed, but they are still in their infancy. It is for this reason that we will then consider conceptual aspects of autonomous computational explorations of catalytic systems, which we then supplement with an explicit example to highlight some of the key issues that need to be mastered.

2 Computational Catalysis and Mechanism Exploration

Considering the complexity of a catalytic process in terms of reaction steps and materials, first-principles modeling is challenging because of the structural variety and size of the atomistic systems and because of the vast amount of fine-grained transformation steps that need to be considered. Hence, it is obvious to exploit existing data first, which has already become a major strategy for the design of new materials with specific functionality [60–66]. Vast amounts of data of different origin may be utilized to understand and design catalytic processes [67, 68]. A substantial number of publicly accessible databases [69–92] has become available along with software packages encoding general workflows to interact with these databases [93–104]. Screening these data can produce valuable property predictions [105–109]. High-throughput studies can be accelerated by exploiting also surrogate models, i.e., efficient, empirical models that can produce property predictions such as adsorption energies, albeit less accurately than a first-principles-based model such as density functional theory (DFT) [110]. Surrogate models can be scaling relationships [111–113], physical descriptors [114–117], or machine learning (ML) models trained on physical or structural descriptors [118–134]. Furthermore, they can be enhanced by stability analysis to save computing time on unstable materials [135]. Such fast data-driven hypothesis generation can then be refined with uncertainty quantification by DFT calculations [110, 136–141].

The application of surrogate models of known uncertainty together with a workflow for high-throughput DFT calculations has been adapted to the evaluation of reaction networks [112, 142–145]. A small molecular size of reactants, such as the oxidation of CO or the oxidative coupling of methanol, limits the number of possible intermediates during the reaction. If, in addition, no pronounced structural changes of the catalyst occur during the reaction so that its structure may be regarded as basically stiff, small chemical reaction networks will emerge that can be considered complete [146, 147]. In such a case, a threshold for the maximum molecular size, e.g., number of carbon atoms involved, can be chosen to then define a chemical reaction network of all possible elementary steps based on reaction equations [112].

Larger reactants with increased structural degrees of freedom and/or structurally floppy catalysts require many more elementary steps for reaching a complete reaction mechanism of the catalytic process, typically much more than can be considered in manual work. Hence, automated procedures are key for the elucidation of such a complete network in order to uncover all relevant mechanistic steps [148–152]. Naturally, definitions of reaction types [153] or graph rules [154, 155] have been exploited for this purpose. The

network of all assumed reaction intermediates on a given surface can then be combined with high-throughput quantum chemical calculations and micro-kinetic modeling to compare different existing hypotheses for a reaction mechanism [156].

Constructing a reaction network simply based on viable intermediates on reactants and considering the catalyst solely as a static partner, onto which these intermediates are adsorbed, is mostly limited to flat catalytic surfaces. Most existing algorithms likely struggle with solid phases that undergo structural rearrangements during reactions on their surfaces so that the reaction intermediates significantly differ from their gas phase counterparts; examples are flexible catalysts such as nanoclusters [157], anchored organometallic complexes [158], and reactions that remove and regenerate atoms at a surface [159]. The increased degree of complexity that the direct structural involvement of such catalysts adds to the problem of the elucidation of catalytic reactions networks for large reactants with a high degree of structural flexibility highlights an even more pronounced role of automated exploration procedures, which we, given the diverse nature of potentially catalytic agents, decided to base on electronic structure information only [160–163]. This allows us to exploit general heuristic concepts based on the first principles of quantum mechanics.

Most automated reaction network generation schemes have originally been developed for molecular systems (see, e.g., Refs. [160, 164–176]). The underlying algorithms and concepts range from graph-based rules to the interpretation of the electronic wave function, and to ab initio molecular dynamics (MD). However, all these algorithms have the common goal of constructing all possible elementary steps for a given pool of reactants by locating the corresponding transition states with first-principles and semi-empirical electronic structure methods. Whereas they were developed for systems that represent a single phase (typically the gas phase or a solution), some of them have also been applied to reactions on metallic surfaces.

The latest release of the graph-based *reaction mechanism generator* (RMG) by Liu et al. [155] features additional graph rules for surfaces, in which the surface is treated as a single graph node with which every other node can form bonds with. The authors applied this approach to methane dry reforming on Ni (111) [177], for which their algorithm found many of the reactions of an established mechanism [178]. However, their approach was limited to predefined reaction types, the adsorption energies were based on literature values or group additivity for missing literature data, and the reaction energy barriers were derived from scaling relationships from the literature.

Zimmerman and co-workers have developed the software *S-ZStruct* [179] for specifically handling surface explorations. It implements an interface to the *atomistic simulation environment* (ASE) [180] to find adsorption sites and explore reaction paths of adsorbates with their *growing string method* (GSM) [168, 181]. Maeda et al. have also explored reactions of adsorbates on (111) surfaces [182–184] with their *artificial force induced reaction* (AFIR) approach [185]. While both approaches, GSM and AFIR, are versatile and general, the application studies were limited to low-index surfaces with a completely constrained slab. Moreover, the adsorption site location of ASE is implemented only for certain surfaces, while more advanced surfaces would require manual definitions [179]. Owing to the general, atomistic nature of their core algorithm, the AFIR and GSM method, both Maeda et al. [186–188] and Zimmerman et al. [189–199] have studied homogeneous catalysis more extensively, also incorporating experimental information. Their algorithms can also be applied in a semiautomatic fashion by steering the exploration into certain branches of the reaction network, either by specifying specific internal coordination transformations or fragments of the molecules that shall be combined or dissociated. However, this requires extensive knowledge of the software.

In a different approach, Liu and co-workers [200] sampled a reaction on a Cu (111) surface, namely the water gas shift reaction. They constrained two of three layers and found the reaction with their enhanced sampling technique called *stochastic surface walking* (SSW). They applied the SSW technique also for solids [201] and more complicated heterogeneous systems [202, 203] by training a neural network on MD data, which is implemented in their `LASP` software package [204]. Besides surface slabs, also first-principles-based exploration methods have been applied to cluster models of nanoparticles based on graph rules by Habershon et al. [205] and with the AFIR approach by Maeda et al. [206–208].

A common reference example, that was studied by multiple groups, is the hydroformulation of ethene by the $HCo(CO)_3$ catalyst with the goal to reproduce the mechanism by Heck and Breslow [209]. This example has been investigated with time-independent calculations by Maeda and Morokuma [210], with an MD based method by Martínez-Núñez et al. [211], and with graph rule based approaches by Habershon et al. [170] and by Kim et al. [172].

While some proof-of-principle studies on (111) metal surfaces and conformationally limited organometallic catalysts have been conducted, a general software package for autonomous studies of any catalytic system has not been established, yet. We are developing the software suite `SCINE` [212], which does not impose heuristics, reaction types, or electronic structure models that are limited to specific chemical systems. In this article, we introduce the extension of our framework toward general homo- and heterogeneous catalysis, both on a conceptual basis and in terms of first implementations.

We have set out to contribute a general approach to computational catalysis [142, 152, 160, 161] which is the mapping of chemical reactions on reaction networks in such a way that we can transcend conventional subcategories of catalysis. To achieve this goal, it is necessary that all algorithms are agnostic with respect to the type of chemical elements involved and the kind of chemical process to be considered (in solution, on a surface, in an enzyme, in a metal organic framework or zeolite, ...). Moreover, the algorithms need to be as stable as possible, requiring operator interference in an interactive manner [213–215] only in critical cases where even contiguous attempts to achieve some target with different algorithmic strategies (such as different approaches for transition state searches) have failed [216].

Here, we now focus on automated reaction network constructions for catalysis and elaborate on the specific challenges which need to be addressed in order to make such constructions feasible for routine application. For this purpose, the next section first addresses conceptual consideration in the context of catalysis. Afterwards, we discuss an explicit numerical example to highlight some of the technical challenges as well as options for their solution.

3 Conceptual Considerations

We first consider conceptual issues that are presented by problems in catalysis to automated reaction mechanism exploration.

3.1 Identifying Catalysis in Reaction Networks

A catalyst is defined as a substance that increases the rate of a chemical reaction. It is both reactant and product of a reaction and is therefore not consumed [217]. A reaction network that is constructed by automated procedures [148–152] and hence not limited with respect to the number and type of reactants (at least in principle) does not highlight catalytic or autocatalytic cycles that may be embedded within. (Auto) catalysis is an emergent chemical concept that needs to be discovered in such a network. However, the definition of catalysis given above can be turned into an algorithm for its discovery (see, for example, Ref. [161] for an autocatalytic mechanism detected in a reaction network of the formose condensation reaction).

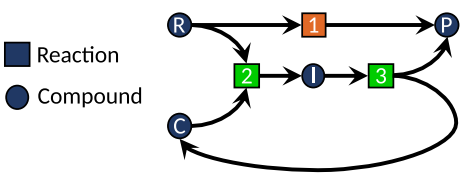


Fig. 1 A minimal reaction network including a reactant *R*, catalyst *C*, intermediate *I*, and product *P*. The orange colored reaction features a larger barrier height than the green colored reactions

Since a vast reaction network of elementary reaction steps is a priori agnostic with respect to our understanding of some of its substructures as being catalytic, their identification follows a posteriori by searching for properties given in the definition of a catalyst: (1) An individual molecule or atomistic ensemble (such as a surface) is identified to take part in a reaction, but is recovered at another position in the network. (2) The other reactants and products of this reaction are found to be connected by a set of different elementary steps somewhere else in the network. (3) Then, one may be able to extract two net reaction rates for both reactions (one with the entity that emerges unchanged from the reaction and one without such an entity). (4) If the reaction rate with the eventually recovered species is significantly faster than the one without, this species will most likely be a catalyst—obviously, the increase in rate must be significant for a catalyst in order to distinguish its role from that of a pure spectator molecule such as a solvent molecule. A minimal reaction network is depicted in Fig. 1, where the compound *R* can either react uncatalyzed to *P* in reaction 1 or via the reactions 2 and 3 enabled by the catalyst *C*.

A discovery of (auto)catalytic processes in this way is relevant mostly for exploratory studies of vast reaction networks, for which hardly any or no information is available at the start of the exploration. In practice, the problem is often simplified by the fact that one may know the (standard or a class of) catalyst structures to be investigated (and also of the chemical reaction that is to be catalyzed). A catalytic cycle can then be explored in a straightforward manner and directly compared with the reaction that lacks the catalyst as a reactant (typically in two different explorations conducted in parallel). This procedure is clearly more target-oriented and allows for catalyst design (by modification and subsequent refinement of structures in a catalytic cycle; see below) as well as for the evaluation of the catalytic potential by direct energy-based comparison with the catalyst-free reference network.

3.2 Calculation of Well-Established Diagnostics from Reaction Networks

Given a vast reaction network that includes an identified catalyst, the question remains, how the catalytic mechanism can be understood and quantified from this network.

Micro-kinetic modeling of the network, e.g., by solving a Markovian master equation based on state and transition probabilities [218–227], preferably accounting for first-principles-derived uncertainties in these probabilities [142, 228, 229], is desirable. However, this is computationally demanding for vast reaction networks, especially if several reaction networks should be compared with one another. Therefore,

some limitations of the network or approximations for the kinetic analysis are commonly introduced (see, for instance, Refs. [230, 231]).

Instead of constructing reaction networks based on heuristic rules and then conducting a kinetic analysis on the whole network, one may explore the reaction network based on quantum chemical methods with a continuously running kinetic analysis on the fly as a guide. Such a kinetics-driven steering of the exploration process can exploit the calculated barrier heights obtained so far to determine those nodes that accumulate concentration and are therefore the key nodes for further network exploration in the next step [161, 229].

Two general experimental metrics for the effectiveness of catalysts are the *turnover number* (TON) and the *turnover frequency* (TOF). However, their definitions may vary for different types of catalysis such as biocatalysis, homogeneous catalysis, and heterogeneous catalysis [232]. We take the TON to be a quantitative measure for the stability of a catalyst against deactivation reactions and the TOF as a measure for the efficiency of a catalyst.

We first define the TOF as the number of catalytic reaction cycles N^c accomplished per time t

$$\mathrm{TOF} = \frac{N^c}{t}, \tag{1}$$

which may be obtained numerically by micro-kinetic modeling of a reaction network or analytically by identifying the catalytic cycle within a network and applying the *energetic span* model [233–236].

Experimentally, this quantity must be normalized by some measure for the amount of catalyst available. One may compare experimental results and theoretical predictions based on first-principles networks based on relative theoretical TOFs rather than on absolute TOFs for reasons discussed later.

In heterogeneous catalysis, the TOF is commonly replaced with the *site time yield* [237], which is normalized with the number of active sites on the catalyst that may be approximated by the number of adsorbing gas molecules in a separate experiment. We do not need to consider such a normalization, because a complete chemical reaction network at full atomistic resolution would include a catalytic cycle for each and every individual active site. Hence, one would obtain a theoretical TOF per site and then may average over all sites afterwards, if desired.

Theoretical TOFs are often determined in the framework of transition state theory (TST) [238], which connects the reaction rate k_i with the activation free energy $\Delta G_i^{\ddagger}$

$$k_i = \frac{k_B T}{h} e^{-\beta \Delta G_i^{\ddagger}} \tag{2}$$

with Plank's constant h, temperature T and β defined as the inverse product of the Boltzmann constant k_B and T, i.e., $(k_B T)^{-1}$. In the framework of TST, Kozuch and Shaik derived the *energetic span* model [233–236], which allows one to calculate the TOF from the absolute Gibbs energies of all transition states G_i^T and intermediates G_j^I and the relative Gibbs energy ΔG_r of the catalytic cycle of N steps

$$\mathrm{TOF} = \frac{k_B T}{h} \frac{e^{-\beta \Delta G_r} - 1}{\sum_{i,j=1}^{N} e^{\beta (G_i^T - G_j^I - \delta G_{i,j})}}, \qquad \delta G_{i,j} \begin{cases} \Delta G_r, & \text{if } i > j \\ 0, & \text{if } i \leq j. \end{cases} \tag{3}$$

This general expression can be approximated in terms of two crucial concepts, the *TOF determining transition state* (TDTS) and *TOF determining intermediate* (TDI) [236]:

$$\mathrm{TOF} \approx \frac{k_B T}{h} e^{-\beta \delta E},$$
$$\delta E \begin{cases} T_{\mathrm{TDTS}} - I_{\mathrm{TDI}}, & \text{if TDTS appears after TDI} \\ T_{\mathrm{TDTS}} - I_{\mathrm{TDI}} + \Delta G_r, & \text{if TDTS appears before TDI.} \end{cases} \tag{4}$$

The two states, TDTS and TDI, maximize the *energetic span* δE of a catalytic cycle. The reliability of this approximation depends on the *degree of TOF control* [236] of TDTS and TDI.

By virtue of the *energetic span* model, the activity of a catalytic reaction cycle within a chemical reaction network can be directly estimated [239]. Additionally, crucial states and steps within a reaction mechanism can be identified. Kozuch and Shaik showed that comparisons of calculated TOFs are quantitatively reliable due to error cancellation, while absolute rate estimates are difficult to predict due to the exponential amplification of an error in the Gibbs energy [236].

The robustness of relative rate comparisons allows also for reliable estimates of the proportion of occurring catalytic reactions and degradation reactions, which allows to calculate the TON. For this, we define catalytic reactions r_i^c, i.e., a single reaction or series of reactions, for which a species has been identified to act as a catalyst and is therefore recovered after the reaction. We also define degradation reactions r_j^d, for which the catalyst is solely a reactant and not recovered, and we define degradation reactions r_k^d, which also consume the catalyst, but require an intermediate of a catalytic reaction r_i^c as reactant. Note that 'reaction' here refers to a sequence of elementary steps. In other words, if a catalytic reaction consists of multiple elementary steps, which is typically depicted as a cycle, it is solely one r_i^c in our definition.

In view of the data that are available for a reaction network of elementary steps, it would be convenient to define a 'turnover efficiency' as a measure for the TON that can be obtained as the ratio of the total probability for product

molecular production and the total probability for catalyst decomposition. Naturally, such probabilities are given by the net rate constants for sequences of elementary steps that either lead to product molecules or to catalyst decomposition. Accordingly, we may introduce such a TON as the ratio of the sum of rate constants k_i^c of all catalytic reactions and the sum of rate constants k_j^d of all degradation reactions

$$\mathrm{TON} = \frac{\sum_i k_i^c(r_k^d)}{\sum_j k_j^d}. \tag{5}$$

As indicated in the numerator, the rate constant of the catalytic reaction(s) k_i^c is, among other quantities, a function of the degradation reactions r_k^d that branch off the catalytic cycle—which only affect catalytic reaction r_i^c—while the degradation reactions r_j^d disconnected from any catalytic cycle affect all r_i^c and lower the total TON. Generally, k_i^c can be approximated by the TOF, but due to this additional consideration of N_d degradation reactions with $\Delta G_k^{\ddagger}$ Gibbs energy barriers, Eq. (3) must be slightly altered to read

$$k_i^c = \frac{k_B T}{h} \frac{e^{-\beta \Delta G_i^c} - 1}{\sum_{a,b=1}^{N} e^{\beta(G_a^T - G_b^I - \delta G_{a,b})} \sum_{k=1}^{N_d} e^{-\beta \Delta G_k^{\ddagger}}}, \qquad \delta G_{a,b} \begin{cases} \Delta G_r, & \text{if } a > b \\ 0, & \text{if } a \leq b. \end{cases} \tag{6}$$

The TON can then solely be expressed in terms of energies as

$$\mathrm{TON} = \frac{\sum_i \frac{e^{-\beta \Delta G_i^c} - 1}{\sum_{a,b=1}^{N} e^{\beta(G_a^T - G_b^I - \delta G_{a,b})} \sum_{k=1}^{N_d} e^{-\beta \Delta G_k^{\ddagger}}}, \qquad \delta G_{a,b} \begin{cases} \Delta G_r, & \text{if } a > b \\ 0, & \text{if } a \leq b. \end{cases}}{\sum_j e^{-\beta \Delta G_j^{\ddagger}}} \tag{7}$$

This allows us to calculate the stability of a catalyst against decomposition. However, this is hardly done in experimental research [240] and neither in computational research due to the complexity of finding all relevant degradation reactions. A mitigation of this problem is, in fact, the autonomous exploration of elementary steps based on automated first-principles procedures, which can deliver huge networks of complex reactions that may be considered complete after a certain exploration depth has been reached.

3.3 Autocatalysis

The simplest definition of autocatalysis is given by a (series of) elementary step(s), in which a product X catalyzes its own creation [241].

$$\mathrm{A} + \mathrm{X} \longrightarrow 2\mathrm{X} \tag{8}$$

Due to the *nonlinear chemical dynamics* [242] (such as oscillations) that autocatalysis can cause, it has attracted little interest by the chemical industry until recently [241]. Accordingly, the topic has received much attention in origin of life studies [243–246], since autocatalysis can be connected to replication, which is essential for the development of complex living organisms. It might also be the cause of homochirality of all amino acids within all living beings on Earth [247]. Recently, autocatalytic self replication has been developed and studied in synthetic chemical systems [248–251].

On a theoretical basis, autocatalytic reaction networks have been studied as a basis of the origin of life by Eigen [252], Kauffman [253], and Steel et al. [254, 255]. Steel et al. developed the *reflexively autocatalytic food generated* (RAF) network model, that was also applied in the study of metabolic pathways [256] based on slightly modified or grouped reactions stored in the UniProt database [257] to fit the RAF model. Note, however, that Andersen et al. criticized the RAF model for assuming that every reaction within a chemical reaction network is catalyzed, which is unlikely [258]. Instead, Andersen et al. developed a rigorous definition of autocatalysis in chemical reaction networks by describing the network as a directed hypergraph and the autocatalytic reaction as an integer hyperflow [259] based on reactions derived from graph rules. However, they noted that a sole definition by hyperflows is most likely not sufficient and will need complementary constraints in order to detect autocatalytic cycles in arbitrary chemical reaction networks [258].

Such algorithms, which avoid computationally expensive numerical kinetic simulations, are required and cannot be circumvented with a straightforward identification strategy solely based on thermodynamic criteria as outlined in Sect. 3.1. For example, the corresponding uncatalyzed reaction of Eq. (8), which can be formulated as

$$\mathrm{A} \longrightarrow \mathrm{X} \tag{9}$$

might simply not exist or impose such high barriers that it cannot be located with standard algorithms. Without such points of references, which are missing in experimental data of biological systems, for which most definitions and algorithms discussed in this section had been developed, autocatalysts can only be identified and distinguished from bystander molecules based on kinetic analyses.

Many theoretical models also construct chemical reaction networks solely with graph rules and do not take into account different reaction barriers and conformers. If the exploration of a network is based on first-principles calculations in such a way that all elementary steps are mapped out, the detection of autocatalysis requires micro-kinetic modeling of the reaction network. However, if one restricts the exploration by constraints that do not allow for the passing of barriers of a given height (or similarly by explicit kinetic modeling), the detection of autocatalytic paths becomes much more difficult, especially for compounds which can only be formed by an autocatalytic reaction. The issue is that

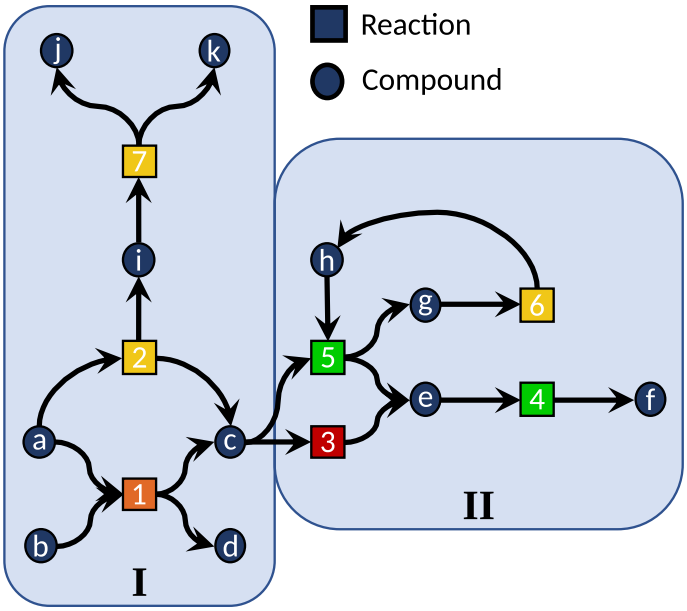


Fig. 2 A reaction network including the autocatalytic reaction 5. Light-green reactions have the lowest reaction barrier heights, followed by yellow, orange, and dark-red (indicating the largest barriers)

a product, which might act autocatalytically and, therefore, decreases the barrier(s) of the reaction(s) necessary for its own creation, might never be found. A minimal example is depicted in Fig. 2, where compound *h* acts autocatalytically. In a first-principles-based exploration of this network starting from *a* and *b*, the network would never discover the region II leading to the favored product *f*, but would, instead, stay in region I and wrongly predict the compounds *j* and *k* as the major products.

The crucial question then is how one can account for this issue in the automated exploration of a chemical reaction network? For known autocatalytic motifs, a viable option would be the systematic trial exploration of such a motif. An example is acid catalysis in the context of ester hydrolysis (see, for instance, Ref. [260] for a detailed description and further examples). If many exhaustive catalytic reaction networks become available in the future so that sufficient amounts of data are available, one may extract patterns for the onset of autocatalytic pathways with machine learning models. Unfortunately, all of this would include a heuristic bias on known chemical phenomena and further research is required to identify truly exploratory first-principles-based approaches.

3.4 Catalyst Design

Many optimization and design strategies for more stable or active catalysts have been developed for specific fields such as biocatalysis [261–270], homogeneous catalysis [271–281], or heterogeneous catalysis [282–288]. In these strategies, the activity of a catalyst is judged on various physical descriptors. For our discussion here, it is important to recall that a chemical reaction network of elementary steps is a universal means for studying a catalytic reaction: it encodes all information for understanding the catalytic process *in toto* (including deactivation processes and side reactions). Once the reaction states that are key for a catalytic process (e.g., those that determine TON and TOF) have been identified, they can become a target for catalyst optimization and even for de novo design. Note that the uncatalyzed reaction itself is already a viable starting point as its network contains those steps that require a catalyst to decrease high reaction barriers. As such, the network provides atomistic structural information about where and possibly also about how to introduce structural changes and potentially catalytic reagents. Naturally, any structural change introduced at some node of the network will then require a re-evaluation of the whole network in order to probe the viability of previously found elementary steps, to find new ones, and to assess the resulting activation (free) energies. While this is a computer time demanding task, tailored optimization strategies that target specific structure-property relationships may decrease the computational burden.

In general, it will neither be feasible nor sensible to automatically explore a complete reaction network from scratch for a large number of potential catalyst candidates, prohibiting high-throughput screening for catalysts based on networks of elementary steps. Instead, the comparison between different catalysts should happen on the basis of network inheritance in order to be efficient.

First, the chemical reaction network may be explored with one specific catalyst, e.g., the known reference catalyst that should be improved. To increase the efficiency of the exploration, this catalyst should be generic in the sense that its structure should not possess unnecessarily costly elements; i.e., those that can be expected to be spectator residues for

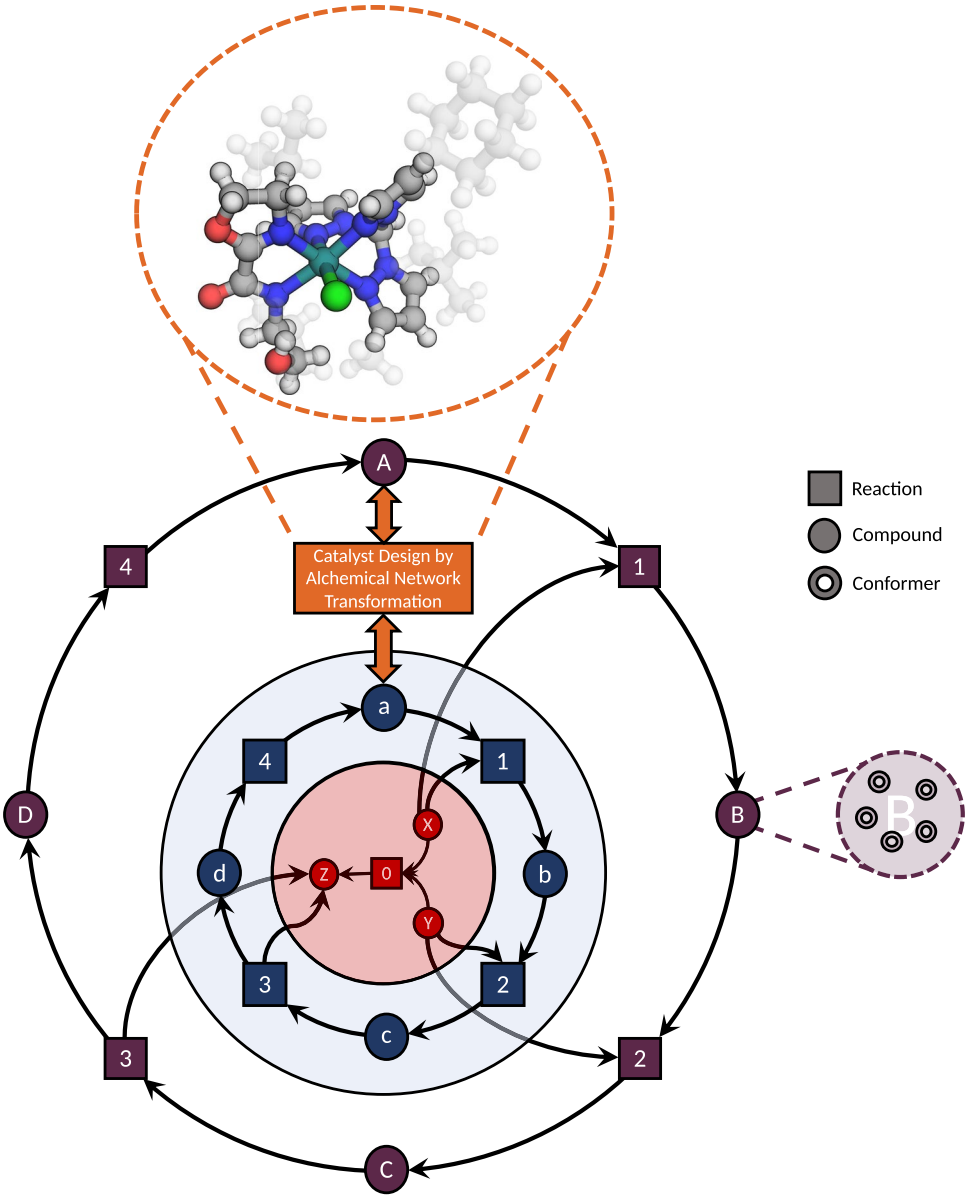


Fig. 3 A schematic reaction network depicting the uncatalyzed reaction 0 of *X* and *Y* to *Z* (red). The same chemical reaction can also be found catalyzed by a minimal catalyst *a* in the series of reactions 1–4 (in blue). This minimal cycle can then be exploited for catalyst design by systematically exchanging ligands (or substituents or central metal ions) of the catalyst, which is schematically depicted in the circle at the top. The modified reaction barriers for 1–4 based on the new catalyst *A* (in purple) can then be explored within the reaction network

the catalytic process itself, but would increase the computational time significantly. A typical example are substituents with large conformational freedom that can be expected to play hardly any role in the catalytic process itself, but are required for different purposes (e.g., solubility or preventing catalyst dimerization). Such structural elements may be discarded for the generic catalyst for which it is then much easier to generate a complete reaction network as it will not suffer from a combinatorial explosion of conformers. However, crucial misrepresentations of the catalyst, which foundamentally change the reaction mechanism, have to be avoided [289].

In a subsequent step, one may re-introduce substituents (also for the purpose of catalyst design) in a step- or shell-wise fashion, possibly aided by ML approaches [290, 291]. The generic reaction network can then serve as an efficient starting point, allowing for a fast re-evaluation of its nodes with the new catalyst structure and a search for new elementary steps.

Alternatively, the main catalytic entity—in most cases a metal or a certain structural motif—can also be substituted on a network level to study different candidates. The simplest case is that of a 'transmutation' where the metal in all structures of generic network is simply exchanged by another one, for which a homologous metal or an isoelectronic metal fragment are suitable candidates (consider, for example, replacing Ru by Fe or Co^{+}) as depicted in Fig. 3.

In this way, information about the catalytic process is inherited in such a way that computational costs are

efficiently reduced and the emerging ancestry can enhance the conceptual understanding of the catalytic system.

Since this is a direct approach, in which a molecular structure is given and its property is calculated, high-throughput virtual screening (HTVS) must be conducted to search for a better catalyst in a systematic way. However, even with an efficient HTVS approach, it is hardly possible to visit a sufficiently large fraction of the chemical space due to its sheer size [292, 293]. Therefore, a wise selection of compounds and materials of this space has to be made depending on the design target.

The key problem is that quantum and classical mechanics allow us to predict a molecular property or function for a molecular structure given. The inverse direction, i.e., from a desired function to a molecular structure that exhibits this function, is mathematically ill-defined for various reason (e.g., in quantum mechanics all dynamical degrees of freedom (such as coordinates) are integrated out when expectation values or response properties are calculated). However, one may hope to develop inverse approaches for specific goals as certain properties of these goals may be exploited to alleviate the problem.

Accordingly, *inverse design* strategies attempt to predefine a specific target property and then construct the corresponding ensemble of structures that feature this property. Many approaches for such algorithms exist and have been discussed in general reviews [294–296] and reviews focusing on ML approaches for *inverse design* [297–300].

For example, we have proposed the inverse-design approach *Gradient-driven Molecule Construction* (GdMC) [301–303], which targets design of new catalysts by sequentially constructing metal fragments that stabilize structurally activated small molecules in intermediates through reduced structure gradients on all atoms. In another approach, Hartke and co-workers have combined optimizations of minimum energy reaction paths in an electric field of point charges with global optimization techniques in their *Globally Optimized Catalyst* scheme [304, 305] and have further improved on it in a quantum-mechanical molecular-mechanical composite approach [306].

ML had a considerable impact on the field of *inverse design* in recent years as it allows for learning structure-property relationships, which can then be employed to generate structures based on a given property. Especially deep generative models have been demonstrated to be successful across multiple chemical problems ranging from drug discovery [307–309] to materials design [310–312]. A combination of such models with genetic algorithms is also possible [313]. For this endeavor to be successful, it was necessary to improve on the representation of chemical structures [314, 315] and desired properties [316]. It was also shown that the new concept of alchemical chirality [317] might allow one to draw direct energy relations across the chemical compound space to accelerate design processes.

Hence, many strategies have been developed for the design of molecules with specific properties. It can be expected that catalyst and process design by computational catalysis will continue to strive for novel as well as routinely applicable design protocols.

4 Computational Considerations

Because of the numerous elementary steps involved in catalytic processes and the fact that changes in structural composition point to new networks of elementary steps, the computational burden is truly intimidating and smart procedures are required to keep it feasible in principle, but also in view of the environmental footprint of high-performance computing campaigns. In this section, we therefore turn to a discussion of the computational resources required for autonomous first-principles-based explorations of homogeneous and heterogeneous catalysis that allow for an understanding on the basis of reaction networks. Clearly, the computational resources required will depend on the methodology chosen. Here, we rely on our methodology in order to give an idea of the magnitude of computational effort that is to be invested in autonomous first-principles-based explorations. Our computational methodology is detailed in the appendix.

4.1 Resource Estimates for Automated Explorations of Homogeneous Systems

A chemical reaction network can be constructed solely based on initial reactants as input. Starting from these structures, all elementary steps can be identified—at least in principle—by letting algorithms search for new local minima starting from the given ones on the respective Born-Oppenheimer hypersurfaces. Newly found minima, which correspond to long- or short-lived intermediates in a reaction network, become new starting points for further exploration in this rolling approach.

A key part of autonomous explorations are automated procedures that allow for the identification of elementary reaction steps with associated transition states. For instance, with our `Chemoton` exploration software, possible elementary steps are probed based on reaction coordinates defined for active sites identified within molecules. In principle, every atom (or group of atoms) in a molecule may function as an active site, an assumption that allows one to map out a reaction network that is as complete as possible. However, this will often not be feasible and so protocols are put in place that reduce the number of potentially relevant sites to those that might be active under reaction conditions. Our strategy so far has been to base this selection process on

rules that may be derived for any molecular system and that are therefore not bound to specific compound classes. Accordingly, we introduced first-principles heuristics as a way to extract conceptual information on reactivity from the electronic wave function [160–163]. Note that it is not required to make a precise prediction on what atoms may react in some intermediate. Instead, it will already be sufficient to identify with certainty those sites that will not react for diminishing the computational burden.

In a brute force approach, one possible ansatz is to define an inter- or intramolecular reaction coordinate as a push (or pull) of reactive centers, which in turn can be defined as the geometric center of one or more reactive sites. This then allows one to enumerate all possible inter- and intramolecular reactions. `Chemoton` probes potential reaction coordinates with so called *Newton trajectories*, for details see the appendix. An exploratory reaction coordinate can be defined as the vector between two geometric centers of lists of active sites. A geometric center is defined by a number a of active sites, with $a \geq 1 \wedge a \leq n_i$ and n_i being the number of nuclei in a reactant. The second center is then defined by a different list of b active sites. For intramolecular reactions the reaction coordinate is simply the vector between the centers, while intermolecular reactions require an additional vector for each combination of active sites and angle between these vectors to construct such an exploratory reaction coordinate. For each combination of a active sites there exists an infinite number of d_a possible vectors and ρ_a possible rotamers, which are reduced in `Chemoton` by discretization of the rotational angle to a finite number based on steric criteria and a fixed number of rotamers. To estimate the scaling of such a brute force approach, we limit possible intermolecular elementary steps to bimolecular reactions. In a reaction network with m compounds found at a given point in time, a number of n_{ci} structures per compound i with n_i nuclei each allows us to estimate the number of possible reaction trials r as

$$r = \underbrace{\sum_{i=1}^{m}\sum_{j=i}^{m} n_{ci}n_{cj}\sum_{a=1}^{n_i}\sum_{b=1}^{n_j} d_a d_b \rho_a \rho_b \binom{n_i}{a}\binom{n_j}{b}}_{\text{intermolecular reactions}} + \underbrace{2\sum_{i=1}^{m} n_{ci}\sum_{a=1}^{\left\lfloor \frac{n_i}{2} \right\rfloor}\sum_{b=a}^{n_i-a}\binom{n_i}{a+b}}_{\text{intramolecular reactions}}. \tag{10}$$

The factor 2 for intramolecular reactions stems from the possibility of either associative or dissociative reactions, while intermolecular reactions can only be associative, albeit they can still generate multiple products. We emphasize that the above equation solely rests on combinatorial considerations that ignores all chemistry knowledge. It is obvious that activating chemical knowledge will dramatically decrease the number of options—the question is how this can be achieved in a way that is so general that it works for any sort of atomistic system, ranging from molecules to molecular aggregates and eventually to surfaces and composite materials.

Note that r represents only the number of elementary step trials (i.e., attempts to identify an elementary step) and not the number of successful elementary steps, because chemical reactions will not be possible for every combination of nuclei. Nevertheless, r grows factorially with n_i and quadratically with m, because any intermediate or reactant can react with any other one of the network. This quickly becomes unfeasible for a large system, which is why pruning (for instance, through first-principles heuristics) will be necessary for the elementary step trials even in exhaustive reference reaction network explorations.

For our resource estimates, we introduce the assumption of maximally combining pairs of active sites ($a \leq 2 \wedge b \leq 2$) for intermolecular reactions, and only pairs of single active sites ($a = b = 1$) for intramolecular reactions, which then leads to

$$r = \underbrace{\sum_{i=1}^{m}\sum_{j=i}^{m} n_{ci}n_{cj}\sum_{a=1}^{2}\sum_{b=1}^{2} d_a d_b \rho_a \rho_b \binom{n_i}{a}\binom{n_j}{b}}_{\text{intermolecular reactions}} + \underbrace{\sum_{i=1}^{m} n_{ci}\left(n_i^2 - n_i\right)}_{\text{intramolecular reactions}}. \tag{11}$$

This reduces the scaling behavior to $\mathcal{O}(m^2 n_i^4)$. If we assume $m \gg n$—i.e., there are far more stable intermediates in the network than, on average, atoms in each of the intermediates, which is the case for most molecular networks, then the scaling will become quadratic.

Next, we impose restrictions based on graph distances δ_{AB}, which can be determined from Mayer bond orders [318] and our `Molassembler` library [319, 320], which is part of the `SCINE` project. The graph distance δ_{AB} is defined as the number of bonds that one passes when proceeding from nucleus A to nucleus B in the molecular graph. Elementary step explorations $r_{\{A,B\}-\{C,D\}}$ are defined with a reaction coordinate constructed between the active sites A and B and the active sites C and D. We limited the number of $r_{\{A,B\}-\{C,D\}}$ depending on the explored reaction type

$$\text{intermolecular association: } r_{\{A,B\}-\{C,D\}} \quad \Leftrightarrow \delta_{AB} = 1 \wedge \delta_{CD} = 1 \tag{12}$$

$$r_{\{A\}-\{C,D\}} \quad \Leftrightarrow \delta_{CD} = 1 \tag{13}$$

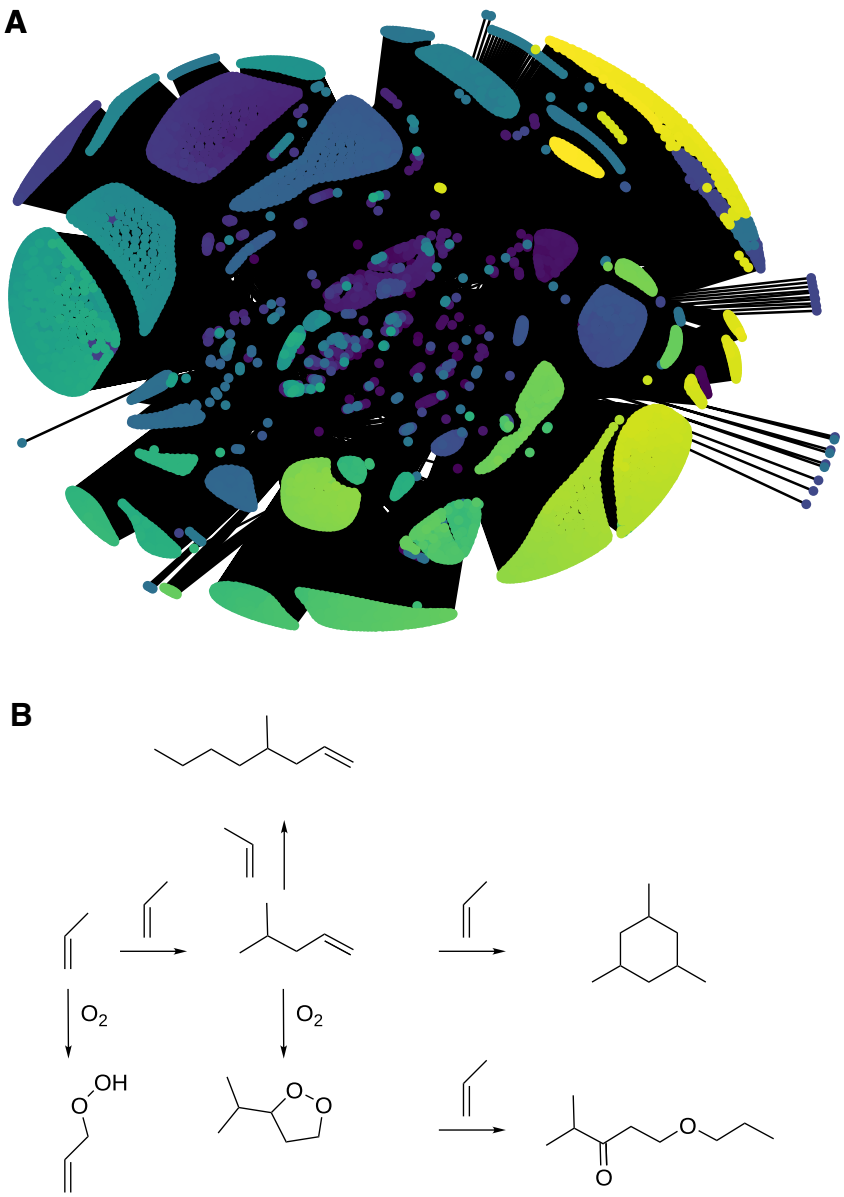


Fig. 4 **A** All compounds in our reaction network connected with lines corresponding to reactions. The compounds are colored according to their order of discovery from violet to yellow. **B** Examples of some of the first reactions in the network

$$r_{\{A,B\}-\{C\}} \quad \Leftrightarrow \delta_{AB} = 1 \tag{14}$$

$$\text{intramolecular association: } r_{\{A\}-\{B\}} \quad \Leftrightarrow \delta_{AB} = 5 \vee \delta_{AB} = 6 \tag{15}$$

$$\text{intramolecular dissociation: } r_{\{A\}-\{B\}} \quad \Leftrightarrow \delta_{AB} = 1. \tag{16}$$

Additionally, we applied a symmetry analysis to reduce the number of unique active sites and only considered further explorations for compounds, which were accessible by reactions with barrier heights below 200 kJ mol^{-1}. Moreover, to properly sample the remaining elementary steps, we considered two rotamers per reactant ($\rho_a = 2$) and multiple directions of attack ($d_a \geq 1$), where multiple local minima in steric hindrance around the active site were present.

First, we constructed from first principles a broad reference reaction network without a catalyst. Such a network allows us to estimate the scaling effects of the restrictions imposed on the explored elementary steps. As an example, we selected propylene and molecular oxygen, which already allowed us to construct a broad reaction network from first principles as shown in Fig. 4. This illustrates the potential scope of reaction networks for small systems.

For the uncatalyzed reference, we explored the reaction network starting from propylene and molecular oxygen with GFN2-xTB [321, 322]. We stopped the exploration after $\approx 3 \times 10^6$ elementary step trials carried out in a total computing time of 5775 CPU days and $\approx 1.4 \times 10^7$ elementary step trials still remaining. This resulted in 4218 compounds, 909 of which are accessible with reaction barrier heights

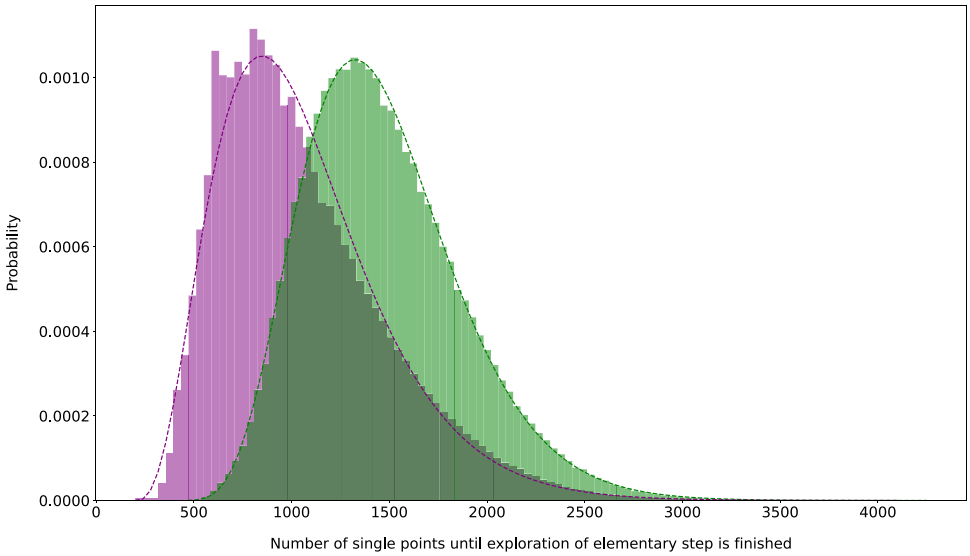


Fig. 5 Histogram of the required number of single-point calculations for an elementary step search attempt (details see Sect. 1). Green bars represent successful and purple bars represent failed attempts. The dashed lines of the same color are the fitted γ distribution

below 200 kJ mol^{-1}. The 4218 compounds include a total of 1,185,893 individual optimized minimum energy structures that are connected by 587,752 transition state structures in elementary steps, which were grouped into 6323 reactions. For the exploration we set an upper limit in terms of element composition of $C_{10}H_{22}O_7$ and the heaviest compound in our explored reaction network is $C_9H_{18}O_4$. The exploration required a total of $\approx 2.9 \times 10^9$ single-point calculations, which, for the sake of comparison, corresponds to a total runtime of $\approx 1.45\,\mu s$ of a continuous MD simulation with a timestep of 0.5 fs.

The most straightforward solution to reduce the number of elementary step trials is a pre-selection based on

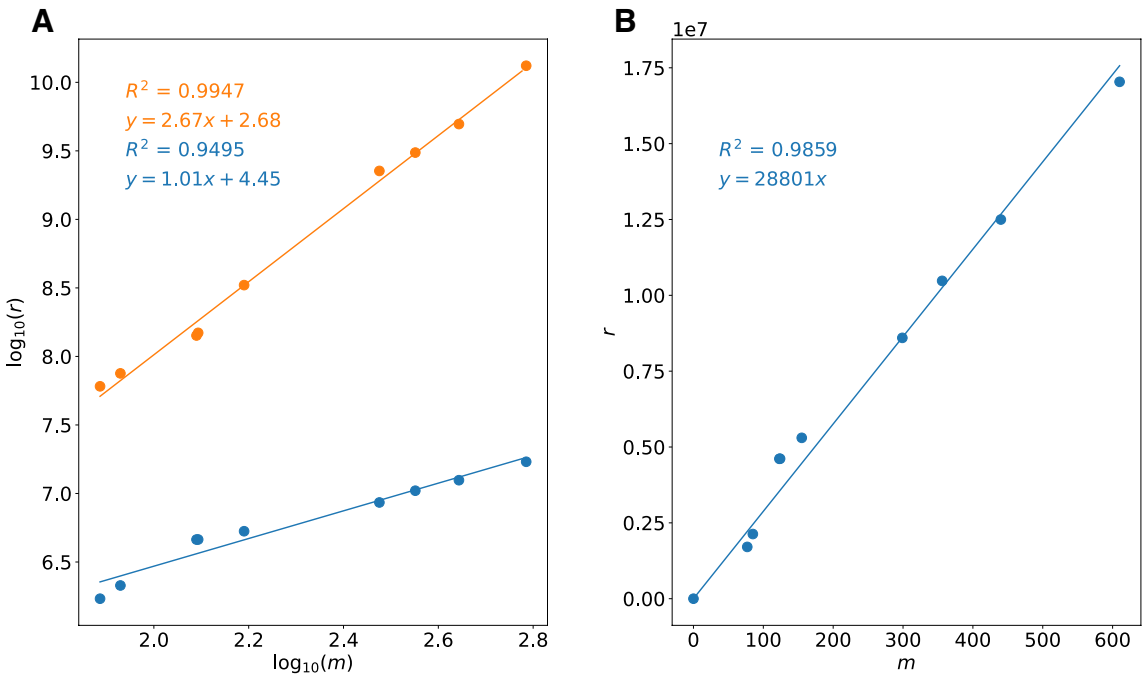


Fig. 6 **A** A logarithmic plot of the number of elementary step trials r against the number of compounds m in the reaction network within an upper limit for barrier heights of 200 kJ mol^{-1}. The orange dots were calculated from Eq. (11) based on the number of compounds in the reaction network, while the blue dots are the actual number of elementary step trials based on the constraints applied during the exploration. **B** Identical data points of the actual elementary step trials in the network, but without taking the logarithm. Lines represent linear regressions; the resulting linear equations are shown in the plot in the corresponding color

reactivity descriptors (e.g., first-principles heuristics; see above), which was deliberately *not* considered in our reference network. The fact that we did not activate such a selection/exclusion schemes for the assignment of active sites to be subjected to elementary search trials can also be observed in the low success rate σ of only 22 % in our brute force approach. Furthermore, we could have restricted the number m of intermediates to be considered as reactants by exploiting some measure for their lifetime. For instance, an intermediate connected to other low-energy intermediates by low barriers will be short-lived and may be excluded from the set of m reactants to be considered.

The average number of single-point calculations per elementary step trial is depicted in a histogram in Fig. 5, to which we fitted a γ-distribution due to the long tail towards higher numbers. This fit allows us to estimate the number of calculations for successful elementary steps to be 1473 ± 405 and of failed attempts to be 1058 ± 424 (ranges defined by the standard deviation) for the current development version of our `Chemoton` software [323]. However, a substantial number of unsuccessful attempts (11 %) already failed within the first 200 steps of the Newton trajectory set-up because structures far away from an equilibrium structure were generated so that the self-consistent-field procedure did not converge. These calculations were excluded from the fit. Upon taking them into account, the arithmetic mean of the single-point calculations required is lowered to 1050.

Structure optimizations of conformers generated with `Molassembler` [319, 320] required only 3 days of total CPU time on a single core for a total of $\approx 8.5 \times 10^6$ single-point calculations. Hence, it can be estimated that the costs of additional geometry optimizations, e.g., to refine structures based on more accurate electronic structure methods, are negligible compared to elementary step trials.

Based on this extensive network, we can now study whether our assumptions about the scaling behavior were correct and how our graph distance restrictions affect this scaling. For this numerical analysis, we plot the number of elementary step trials r against the logarithm of the number of compounds in the reaction network m that are accessible within the given barrier height limit as shown in Fig. 6A)).

It is evident that a quadratic scaling with the number of compounds can be observed. However, the total scaling is larger than quadratic, because the molecule sizes cannot be disregarded. In addition, we understand that the chosen constraints based on the graph distance have a strong effect on the scaling behavior and reduce the scaling to a linear one. Nevertheless, the slope of 28,000 of the linear scaling, shown in Fig. 6B), is still substantial, especially considering that we did not take into account the generated conformers in the reaction explorations, but probed possible elementary steps only for the first occurring conformer structure of each compound.

Note also that explicit solvation was not considered in this extensive reference network. Numerous approaches [324–329] exist that can limit the number of solvent molecules. However, they still increase the required number of calculations and may require further development to tame this increased computational burden (e.g., by transferring solvation information with machine learning models from microsolvated nodes to those for which no microsolvation had been considered).

Whereas the network structure discussed so far did not contain any catalyst, we now estimate how the addition of a homogeneous catalyst increases the computational resources required. Formally, the scaling of the reaction network still follows the same pattern as before, because the catalyst molecule is simply another compound within the network. However, because of the typical size of a catalyst of 50–150 atoms and because of the intricate relation between its structure and activity, the approximation that a single conformer is sufficient to provide a sufficiently deep and reliable overview on the reaction mechanism will, in general, no longer be valid. Moreover, organometallic catalysts often represent challenging electronic structures, which can prohibit the application of fast semi-empirical methods, but require a more accurate description of the electronic wave function based, at least, on a fast (spin-unrestricted) density functional approach. Therefore, any practical exploration of a reaction network in the context of studying catalysis benefits from further restrictions in the exploration protocol, if they can be invoked without compromising the exploration depth.

Based on the data obtained for our reference network and a representative example of an organometallic catalyst, we now show, how severe such restrictions must be and how time-consuming exhaustive explorations of a catalytic reaction network can become. We selected a ruthenium catalyst consisting of 66 atoms, which catalyzes the epoxidation of small cyclic olefins [330]; see Fig. 7A)). We assume that a minimal catalytic cycle consists of around 10 different compounds and we can further estimate that the reaction mechanism including possible side reactions may be sufficiently well explored with 100 compounds, while 1000 compounds would be a very exhaustive exploration of all reactions surrounding a catalytic cycle. Recall that our definition of a compound [152] is a set of molecular structures with the same nuclear composition and connectivity; hence, one compound consists of numerous conformers.

Our uncatalyzed reaction network of 4218 compounds starting from propylene and oxygen already covers polymerizations, cyclizations, epoxidations, various peroxides, radical reactions, and beginnings of the formose reaction network. To estimate the number of single point calculations n_{sp} that are required to find m different compounds we take the following metrics from our reference network

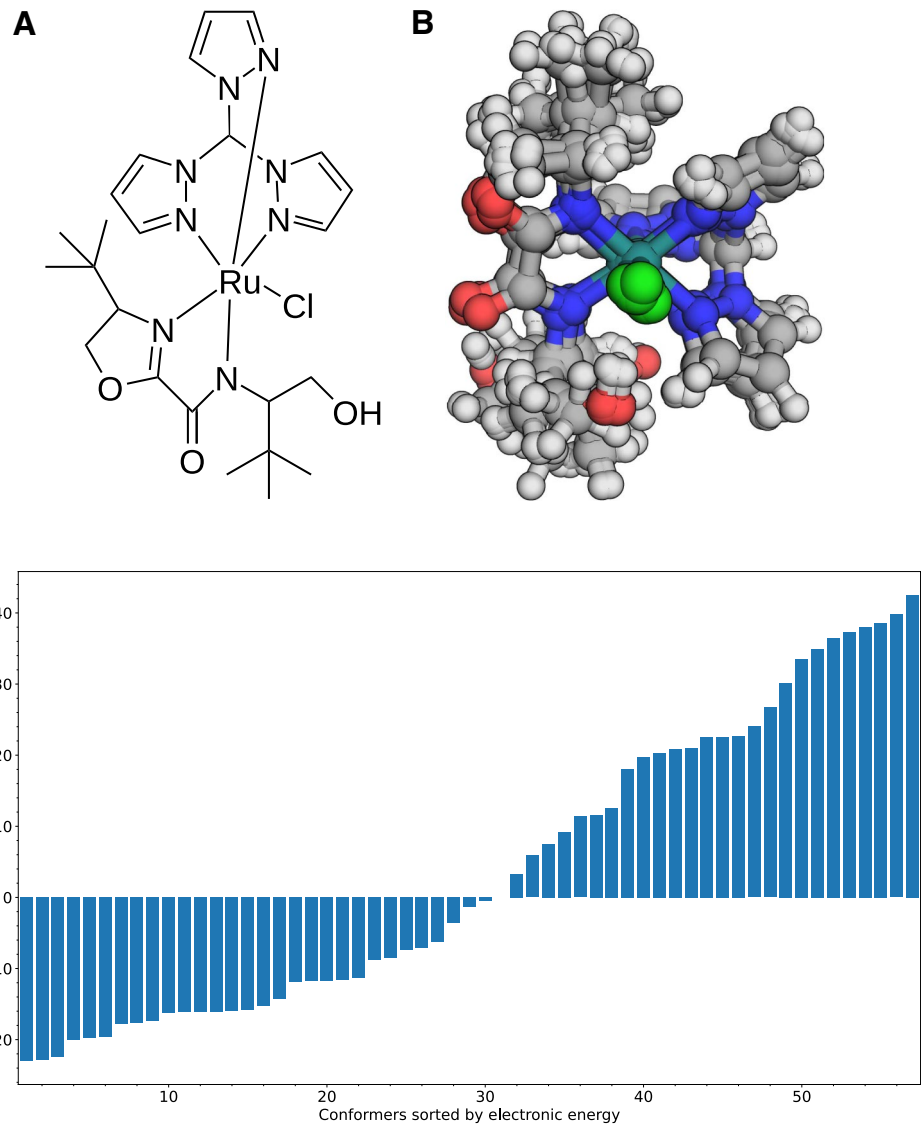


Fig. 7 Conformer analysis of an example catalyst, which we chose to be an organometallic catalyst for olefin epoxidation. **A** Lewis structure of the catalyst; **B** overlay of the optimized crystal structure and nine optimized conformers, which were the energetically lowest structures within their bin of structures after clustering; **C** electronic energies of all 57 conformers relative to the optimized crystal structure

and assume, as a starting point, that they are suitable for a network including a homogeneous catalyst:

- success rate of elementary step trials σ
- ratio between elementary steps found and reactions ε, which yields an average number of elementary steps that belong to the same reaction
- average rate of newly found compounds per reaction η, as some reactions yield more than one previously unknown compound
- single-point calculations per elementary step trial ν

Assuming that these metrics are independent of the number of compounds in the network, we arrive at Eq. (17) to estimate the number of single-point calculation for constructing a network of m compounds to be

$$n_{sp} = \frac{\varepsilon \nu}{\sigma \eta} m \tag{17}$$

$$\approx \frac{92.95 \times 1050}{0.22 \times 1.99} m \approx 2.2 \times 10^5 m. \tag{18}$$

However, all four parameters were taken from our reference reaction network and some of them will depend on the choice of our constraints in the exploration protocol. For example, ε will strongly depend on the number of conformers considered in the exploration and σ can be increased with the application of a suitable reaction descriptor, both of which were not considered in our reference network. Therefore, we assume our ε and σ to be lower bounds for unguided explorations.

Based on these data, we can estimate the number of single-point calculations to find $10^2 - 10^3$ compounds to

be approximately $10^7 - 10^8$. Any reactivity descriptors that identifies unreactive and reactive sites should manage to find all intermediates and products of the minimal catalytic cycle within these $10^2 - 10^3$ compounds, otherwise the number of required compounds and therefore calculations increases.

To estimate the number of conformers for an organometallic catalyst, we applied our conformer generation and optimization protocol implemented in `Chemoton` for our example catalyst. The crystal structure was taken from Ref. [330] and optimized with PBE-D3BJ/def2-SVP. Our graph library `Molassembler` generated 57 conformer guesses, which were then optimized and the resulting structures were clustered according to root mean square deviation (RMSD) by *average linkage agglomerative hierarchical* clustering with a distance threshold of 2.5 Å (see the appendix), which resulted in nine representative conformers. The results are shown in Fig. 7.

We expect a linear to quadratic effect of the number of considered conformers on the overall scaling, because conformers linearly increase the number of considered structures for explorations and in the worst case linearly increase the ratio of elementary steps and reactions ϵ (assuming that all conformers still lead to the identical reaction). Hence, the increase in the number of calculations for this example would be a factor of 100 in the worst case. However, this would mean a consideration of about 10 conformers per compound in the network, which might not be necessary for most substrates. Therefore, we may consider this number of conformers per compound as an upper bound requiring about $10^9 - 10^{10}$ calculations in a brute force approach without the help of any pruning algorithms. Based on the computing times for an energy and gradient of the crystal structure of the catalyst with the semi-empirical GFN2-xTB approach (i.e., 0.25 seconds per single-point in our set-up) and with the generalized-gradient-approximation density functional with density fitting PBE-D3/def2-SVP (i.e., 2 minutes per single point in our set-up), we extrapolate the required total CPU time to be 8–80 and 4000–40,000 years, respectively. In general, a reaction network exploration has the advantage of being trivially parallelizable, meaning that the use of n computing cores brings an n-fold decrease in total wall time. Therefore, the calculations for our example catalyst can be achieved with GFN2-xTB in $3 - 30$ days on 1000 cores, while a complete exploration with DFT remains basically unfeasible without further modification of the exploration protocol or without a large increase in computing power.

In this context, it can be beneficial to carry out the time-demanding exploration trials with efficient semi-empirical methods and then refine the stationary points on a more accurate potential energy surface (PES). In our reference network, the number of single-point calculations required for structure optimizations of stable intermediates was three orders of magnitude smaller than the number of single-point calculations required for elementary step trials. If we assume that $10^9 - 10^{10}$ single-point calculations are to be carried out for building a reaction network, we estimate another $10^6 - 10^7$ single-point calculations for a refinement of the network with a more accurate method, provided that the reaction mechanism or connectivity of the reaction network do not change significantly with the more accurate model. Given our set-up for DFT calculations, this results in an estimate of 3–30 years of computing time on a single core for the reaction network refinement, which again parallelizes trivially and could therefore be achieved within one day to two weeks on 1000 cores. Note that this estimate will also be about the cost for every catalyst design feedback loop (discussed in Sect. 3.4) if the design shall be based on rigorous first-principles-based reaction network information.

These estimates do not consider any restriction or constraint in the exploration process itself. Apart from the pruning options already discussed above (i.e., first-principles heuristics for reactivity descriptors [160, 162, 163] and exclusion of short-lived intermediates from further exploration), the exploration process may be kinetically driven by steering trial and search calculations to those parts of the network that can be reached under reaction conditions by exploiting barrier information [161] or explicit microkinetic modeling [229]. Hence, we may assume that broad automated reaction network explorations are within reach, provided that reliable approximate methods are available and the exploration space can be limited without excluding important reactions.

Unfortunately, resource estimates for explorations of heterogeneous catalysts cannot easily be inferred from data on homogeneous systems. For heterogeneous catalysis, we need to consider additional structures and elementary steps to bridge the phase difference between catalyst and reactant as discussed in the next section.

4.2 Special Algorithms for Heterogeneous Catalysis

Typical heterogeneous catalysts exhibit vastly different structural motifs compared to molecules in the gas phase or in solution, which need to be accounted for in the exploration. The algorithms that we implemented in `Chemoton` for this work in order to resolve these challenges are described in this section.

Any extensive exploration requires to compare individual structures in a timely manner. Root mean square deviations of Cartesian coordinates are not suitable for the process for various reasons (e.g., they depend on molecular size and will require elaborate thresholding for making reliable statements on molecular identity). Graphs are among the best options for such a metric, because (i) they can be compared efficiently and do not depend on system size, (ii) they are

chemically intuitive, and (iii) they allow for substructure/similarity searches. In the automated explorations conducted so far, we exploited graph-based comparisons that are facilitated by the `Molassembler` library [319, 320]. To construct graphs, connectivity information is required, which may be taken from simple distance information or from population analysis of electronic wave functions that yields quantum chemical bond order information. For solid-state systems such as those acting as catalyst or catalyst supports in heterogeneous catalysis, this information is not straightforward to obtain (e.g., consider the adsorption process and how an adsorbate's binding to a surface is to be characterized in terms of chemical bonding).

The seemingly easiest approach to determine bonds in a three-dimensional structure is distance criteria. Parametrized distances for each element are sufficient for molecular structure, but often fail for solid state structures. The two remaining distance-based approaches are Voronoi tessellation and nearest-neighbor criteria. Voronoi tessellation fails for surface systems without the knowledge of the corresponding crystal structure [331]; hence, it is difficult to implement within an automated exploration algorithm, where each minimum structure has to be labeled with a graph, which should ideally only be dependent on the structure's spatial coordinates and electronic structure and not be based on inheritance from other structures.

Nearest-neighbor approaches work well for crystal and surface structures, but can fail for molecular structures, because the atoms in molecules have varying elements as bonding partners with different bond lengths. Therefore, an approach to detect bonds only between the closest distances would either overlook valid bonds or require an elaborate inclusion threshold. Hence, an algorithm solely based on distances must know which nuclei are part of a solid state structure and which are part of an adsorbate. Additionally, the algorithm must then select the distance criterion based on this categorization of nuclei within one structure and also be able to handle chemical and physical adsorption. Such elaborate tracking of nuclei and categorization can introduce many system-dependent heuristics and possible points of failure within an automated exploration.

Alternatively, bonds may better be derived directly from the electronic structure, which avoids system-dependent heuristics. We implemented Mayer bond orders [318] in `SCINE` for molecular and periodic structures, which allows us to directly compare the different approaches. Alternatively, DDEC6 bond orders based on a so-called *dressed exchange hole* determined by the electron density distribution, which has been tested for a wide array of chemical structures [332], may provide more reliable bond estimates.

Adsorption is a key feature of heterogeneous catalysis that is absent in homogeneous catalysis. However, a selection of every nucleus and bond as a potential active site would make an automated exploration unfeasible. In some cases, active sites are likely to be found on high symmetry sites of the surface [333]. Accordingly, Persson et al. applied a Delaunay triangulation on the top layer of the surface slab to retrieve *top*, *bridge*, and *hollow* sites from the corners, edges, and centers of the triangles [334, 335]. The number of these sites can then be reduced based on the symmetry of the surface structure. Boes et al. [331] improved the algorithm by first constructing a graph of the corresponding crystal structure with Voronoi tessellation. This allows one to identify the top layer nuclei of any surface resulting from that crystal structure and to construct an adsorption direction based on the normal vector of a plane spanned by all neighboring atoms in the surface graph. Deshpande et al. [336] directly inferred the adsorption sites from the surface graph, but constructed the graph with a nearest-neighbors approach. This procedure allowed them to deduplicate the relaxed surface structures according to their local-graph information. Recently, Marti et al. [337] released the software *DockOnSurf*, which was specifically developed to generate structures for complex adsorbates and surfaces based on pre-screening of conformers, adsorbing them based on geometric centers of nuclei, and screening conformers on the surface according to dihedral angles. The resulting structures were then deduplicated following an energy criterion.

For this work, we adopted the already existing general algorithms in `Chemoton`, which were developed for intermolecular reactions [323], to establish a new adsorption algorithm which can handle surface slabs and nanoparticles, multidentate adsorption and any adsorbate, while also minimizing the number of screened structures. This new workflow is illustrated in Fig. 8.

In the case of surface slabs, `Chemoton` first detects the high symmetry sites based on Delaunay triangulation as shown in Fig. 8A) and implemented in *pymatgen* [93]. Then, `Chemoton` determines an adsorption vector based on steric hindrance, which allows the program to determine the optimal angle of adsorption, while requiring no graph information of the surface structure. The vectors corresponding to the detected sites are illustrated in Fig. 8B). The adsorbate is then treated as an intermolecular reaction partner and the directions of attack can be formulated for any combination of active sites within the molecule as shown in Fig. 8C) for nuclei and in D) for bonds, which can be extended to any complex combination of multiple nuclei.

The adsorption guess structure is then simply generated by alignment of the direction vectors and can additionally be diversified by considering multiple rotamers defined by a rotation around the direction vectors. The generated guess structure can be optimized with any of the available quantum chemistry programs within `SCINE` (see appendix) and an example result is shown in E). This workflow allows us to reduce the number of explored structures based on symmetry

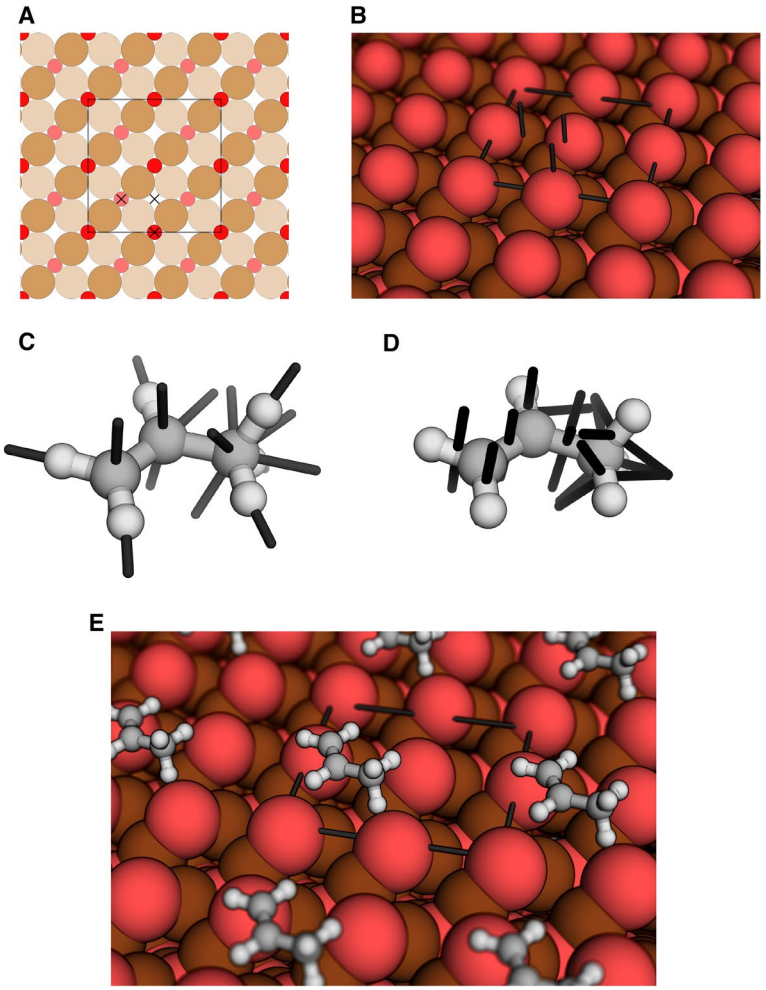


Fig. 8 Representation of the adsorption workflow implemented in `Chemoton`: **A** two-dimensional view of a Cu_2O (001) slab with detected adsorption sites marked by black crosses and the unit cell by black lines; **B** three-dimensional view of the slab with the unit cell and the adsorption vectors marked by black sticks; **C** directions of attack indicated by black sticks for each nucleus in propylene and in **D** for each bond in propylene; **E** example for an adsorbed structure after structure optimization with PBE-D3BJ/DZVP-MOLOPT-GTH

while also being able to treat any chemical system. If no significant symmetry is present, e.g., for nanoparticles, we apply the standard intermolecular approach implemented in `Chemoton`.

For systematic autonomous explorations, the generation of multiple adsorption structures of a single compound is not sufficient, but requires two more steps. First, the number of subsequently explored structures must be reduced by deduplication analysis after structure optimization, because different guess structure may lead to the same minimum. Since an energy criterion for deduplication does not directly relate to structural equality, it may lead to false positives and may hide crucial branches of the reaction network, we opt for a graph-based approach, which is required for large scale explorations in any case. Second, a first-principles-based exploration requires to sample different reactions, which, in the context of heterogeneous catalysis, often requires to adsorb multiple different reactants onto the same surface slab. This is an algorithmic problem, which has hardly been discussed in the literature.

The existence of an already adsorbed molecule causes three main issues in the context of automated adsorption protocols. First, the algorithm must be able to distinguish the existing adsorbate from the remaining surface slab, otherwise it would be detected as a surface site, which may lead to the generation of inaccessible high-energy structures. Second, the existing adsorbate breaks the symmetry of the surface slab in most cases and the number of different second adsorption structures is therefore significantly larger. Finally, the surface may have changed after the first adsorption step, which may prohibit to infer second adsorption positions from the structure of a clean slab.

Therefore, we implemented an algorithm within our automated exploration that tracks which nuclei are part of the surface and which are not. It is able to execute a modified Delaunay triangulation without symmetry exclusion, but with steric exclusion of sites too close to the first adsorbate. This leads to a plethora of possible sites, especially for larger slab models. Hence, it is often wanted to minimize the second adsorption step to sites that are within a reasonable distance to the first adsorbate, especially since the exploration

should sample potential reactions of the adsorbed molecules. Additionally, the exploration should also consider that the second molecule may directly react with the adsorbate from the gas phase, which is why one must screen for such possibilities.

The adsorption algorithm discussed here now allows us to generalize our resource requirements analysis from a homogeneous reaction network to a heterogeneous one.

4.3 Resource Estimates for Automated Exploration of Heterogeneous Catalysts

Because a heterogeneous catalyst is per se only another compound in the network, we know that our reference reaction network consisting of molecules only would be formed identically if we did not enforce any limitations or favored heterogeneous reactions. As in the case of homogeneous catalysis, we cannot consider a single structure of the catalyst only. However, the definition of 'conformers' is, of course, very different for solid state structures, which we discuss in the following. We will also see that new definitions for elementary steps are required which are elaborated on afterwards.

Conformers of a heterogeneous catalyst are not necessarily formed from already active structures, but rather stem directly from the crystal structure. For regular surfaces, these are usually discussed in terms of their Miller indices, including defects and different terminations of the surface. A consideration of all possible surfaces is impossible due to their infinite number and the consideration of many, e.g., ≥ 10, is hardly considered in manually guided studies, which can afford only fewer calculations per discovered compound and need to exploit preexisting knowledge.

For well characterizable surfaces, we may roughly categorize automated heterogeneous explorations in terms of the number of the surfaces (plus decoration) considered per solid state catalyst. A minimal exploration would consider only a single surface without defects. An extensive exploration would consider the (100), (110), and (111) surfaces, usually termed *low-index surfaces*, with different surface terminations, as clean surfaces and a point vacancy and adatom each to include effects of the most-common defects. An exhaustive exploration would consider every surface up to a maximum Miller index of four (≈ 30 surfaces), every possible surface termination as clean surfaces and with ≈ 5 different defects each. Before estimating the scaling of the number of surfaces, we first introduce the term of the number of unique elemental species e. This shall be defined as the number of types of atoms existing in a solid state structure, if all atoms are categorized based on their element, local coordination, and electronic properties. We can roughly estimate that e linearly increases the number of possible surface terminations and possible point defects each. The number of surfaces to be considered, n_{surf}, is then given by

$$n_{\text{surf}} = n_{\text{indices}} \times (n_{\text{termination}}(e) \times n_{\text{defects}} \times e + n_{\text{termination}}(e)). \tag{19}$$

For a bielemental crystal and $n_{\text{termination}}(e) \approx e \approx 2$, we estimate n_{surf} in the three exploration protocols termed above as 'minimal', 'extensive', and 'exhaustive' to be 2, 30, and 600, whereas $e = 3$ would increase n_{surf} to 3, 63, and 1500. Of course, the number of considered Miller indices n_{indices}, surface terminations $n_{\text{termination}}$, and defects n_{defects} are completely independent of each other and explorations can be envisioned that only focus on one of these aspects to decrease the computational costs.

For a given number of surfaces considered, n_{surf}, which do not mutually affect the exploration of one another, we can estimate the scaling of the elementary steps for each of them so that the total scaling will be linear in n_{surf}. While the purely molecular part of the exploration is not changed by the addition of a heterogeneous compound, new types of elementary step trials r_{surf}, which scale differently when compared to purely molecular elementary step trials r_m, must be introduced into the network exploration. Furthermore, the number of possible compounds varies for this part of the network, which is why we split the total number of compounds m into molecular compounds m_m and compounds adsorbed on surfaces m_s for our scaling estimates. Moreover, we can split any additional elementary step trials involving the solid phase into adsorption trials r_a, trials between surface species r_s, and desorption trials r_d, which yields the total number of elementary step trials r as

$$r = r_m + n_{\text{surf}} \times r_{\text{surf}} = r_m + n_{\text{surf}} \times \left(r_a + r_s + r_d\right). \tag{20}$$

The scaling of r_m was already evaluated and discussed in Sect. 4.1 and shown in Eq. (11). The elementary step trials for adsorption, r_a, can be considered as special cases of intermolecular reactions with identical scaling to Eq. (11) for the molecules, whereas each considered surface is only a multiplicative value based on its available first adsorption sites n_{sites_1}, which gives r_a as

$$r_a = n_{\text{sites}_1} \sum_{i=1}^{m_m} \sum_{a=1}^{n_i} n_{ci} d_a \rho_a \binom{n_i}{a} \tag{21}$$

with n_{ci} for the number of conformers of compound i considered for elementary steps. Similarly, the trials for elementary steps on surfaces r_s can also be considered as intermolecular reactions with the number of second adsorption sites n_{sites_2} in place of the different directions of attack d_a and rotamers ρ_a, which leads to

$$r_s = \sum_{i=1}^{m_s} \sum_{j=i}^{m_s} \sum_{a=1}^{n_i} \sum_{b=1}^{n_j} n_{\text{sites}_2} n_{ci} n_{cj} \binom{n_i}{a} \binom{n_j}{b}. \tag{22}$$

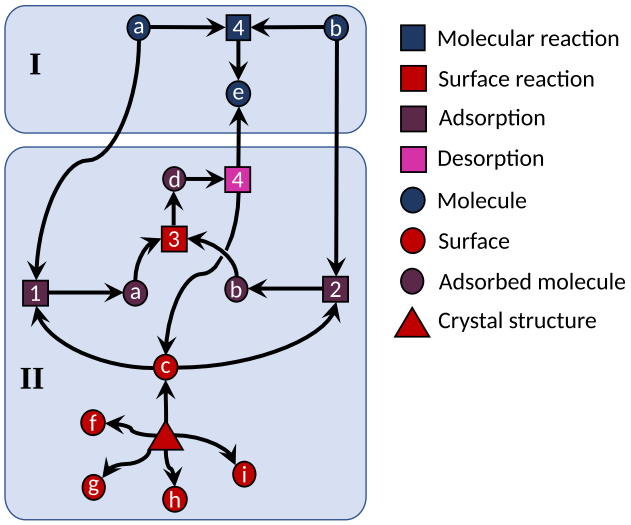


Fig. 9 A minimal reaction network is shown consisting of a molecular part **I** and a solid state interaction part **II**. It includes two compounds a and b, and one surface c. The two compounds can react to e uncatalyzed via the blue reaction 4 or catalyzed by c through the series of reactions 1-4

In a brute-force approach, every adsorbed compound must also be probed to be desorbed or dissociated. The number of elementary step trials for complete dissociation of an adsorbed compound is equal to the number of compounds, while dissociations of adsorbed compounds can be viewed as intramolecular dissociations, which gives

$$r_d = \sum_{i=1}^{m_s} n_{ci} + \sum_{a=1}^{\lfloor \frac{n_i}{2} \rfloor} \sum_{b=a}^{n_i-a} n_{ci} \binom{n_i}{a+b}. \tag{23}$$

If we now apply identical constraints on the number of possible active sites in a molecule as in our reference network (such as limiting intramolecular dissociation trials to repulsion of bonded nuclei or maximally combining bonds and bonds in intermolecular reaction trials), we arrive at

$$r_{surf} = \sum_{i=1}^{m_m} \sum_{a=1}^{2} n_{\text{sites}_1} n_{ci} d_a \rho_a \binom{n_i}{a} + \sum_{i=1}^{m_s} \sum_{j=i}^{m_s} \sum_{a=1}^{2} \sum_{b=1}^{2} n_{\text{sites}_2} n_{ci} n_{cj} \binom{n_i}{a} \binom{n_j}{b} + \sum_{i=1}^{m_s} n_{ci} + n_{ci} \frac{n_i^2 - n_i}{2}. \tag{24}$$

These additional types of reactions, which are typical for reactions on surfaces, are highlighted in a minimal reaction network in Fig. 9. In that figure, the molecular reaction network **I** in blue is enhanced by the reaction network **II**, which consists of interactions with solid state structures.

In Fig. 9, r_a corresponds to reactions 1 and 2, reaction 3 resembles r_s, the pink reaction 4 (as well as the reverse reactions of 1 and 2) corresponds to r_d, and the blue reaction 4 is an example for r_m and is in general the uncatalyzed variant of the series of reactions shown in network **II**.

In general, r_a scales quadratically with n_i, linearly with m_m, and linearly with the number of possible first adsorption sites n_{sites_1}. However, the scaling with n_i can again be reduced by exploiting graph constraints as shown in Sect. 4.1. We may also assume that r_s scales similar to our reference network, although the slope of the linear scaling might be larger due to the factor of n_{sites_2}. *A priori* n_{sites_2} is considerably larger than n_{sites_1} due to missing symmetry as discussed in Sect. 4.2. However, close-proximity constraints can limit this to an approximately constant number of sites on the order of 10. By contrast, n_{sites_1} depends on the complexity of the surface slab and it can be estimated to scale linearly with the number of unique surface species e.

The exact scaling of the elementary step trials r_s and r_d is difficult to estimate, because they only apply for compounds that include adsorbed species. As shown in Eq. (25), m_s depends on the success rate σ of the adsorption elementary steps r_a and the number of elementary steps ε that are found for the same reaction

$$m_s = \frac{\sigma}{\varepsilon} r_a. \tag{25}$$

However, the assumption that σ and ε are similar in value compared to our uncatalyzed molecular reaction network is not valid. While the screening algorithms within `Chemoton` are very similar, the underlying chemical processes are too different to expect similar numbers and they will, in general, also vary between different surfaces. Due to this dependence on the chemical structure, we cannot provide valid general estimates of the number of elementary steps and therefore on the number of single-point calculations required for a heterogeneous network. However, based on the fact that a single surface does not require conformer generation and its different adsorption sites can be viewed as similar to directions of attack in a molecular structure (albeit with a slightly different scaling), the order of magnitude of required single-point calculations for a purely heterogeneous network should be similar to a homogeneous one.

The largest cost factor is, instead, the number of considered surfaces n_{surf}, which linearly increases the number of all calculations. Therefore, the computational costs of an exploration would be increased by a factor of 1000, if various surfaces and defects would be considered as shown earlier in this section. Although it can be assumed that 100 would already cover most relevant reactions and 10 may be enough based on restrictions that may be deduced from experimental data. If we therefore assume a factor of 100, the required number of single-point calculations in a

brute-force approach may be similar to those required in the exploration of a homogeneous catalyst, which we deduced to be $10^9 - 10^{10}$.

Due to the inherently larger number of atoms in solid state structures (and imposed periodic boundary conditions), the calculation times are usually longer compared to molecular systems. If we again take the example of propylene epoxidation, for which Cu_2O is a potential catalyst [338–340], we can estimate the total computing time based on the time required for a single calculation of a (001) slab with an extension of $2 \times 2 \times 3$, which can be taken as the minimal slab size for the exploration of such a reaction. This then leads to a total computing time of $10^6 - 10^7$ years on a single core.

4.4 Requirements for Predictive Computational Catalysis

To make reliable and accurate in silico predictions about catalytic processes, the following requirements need to be fulfilled.

First, it must be guaranteed that all accessible reactions under some specified ambient conditions are explored. Since there is no way to know that everything has been found in an exploration, this can never be guaranteed. However, it needs to be shown in computer experiments that the exploration algorithms chosen can reproduce the relevant parts of a reference network. Clearly, such reference networks for diverse catalytic systems must first be developed, which will require a community effort. While the heuristic nature of this approach cannot be circumvented, it is clear that the exploration algorithms must be general (i.e., agnostic with respect to all sorts of chemical constraints) and cover all relevant reaction types.

Second, the uncertainty of predictions must be accessible, which will require error estimates for all key quantities in the exploration process. Since it is impossible to derive accurate errors for many-particle problems in quantum mechanics (otherwise, an accurate quantum mechanical solution would have been found and the approximations would no longer be needed, which is impossible for any relevant catalytic system), a Bayesian approach is required that transfers error estimates obtained for some nodes after investment of additional computational resources to nodes for which such information is not available [142, 144, 229, 341].

Third, structural fidelity, i.e., the fact that the nuclear scaffolds that define the external potential in the quantum chemical calculations sufficiently well represent the chemical system in terms of molecular structure, surface, and solvent, needs to be ensured for all predictions. Only if the structural model adequately resembles the experimental situation, reliable predictions can be made.

Finally, it should be possible to use electronic structure methods applied interchangeably in order to find the best compromise between accuracy and speed by switching from fast-approximate to expensive-accurate methods. Such switches can either be driven in an automated fashion, if a suitable descriptor (such as confidence intervals from machine learning models [144]) is available, or the software issues a warning and requires manual intervention [151]. These approaches must be combined into general workflows, some of which will be discussed in the following section.

We emphasize that the diversity of all reaction steps that can occur is so vast, even if one restricts the exploration to the known ingredients (i.e., ignoring the unknown ones such as impurities in solution or at a surface), that achieving completeness is formally impossible. This is not a key problem of an autonomous approach that targets orders of magnitudes more detail (measured, e.g., in terms of the number of elementary steps or the number of potentially important impurities such as traces of oxygen or water in a reaction liquor) than what could be inspected manually. However, manual intervention is, of course, possible and can be used to steer an exploration into specific regions of chemical reaction space by letting the search algorithms probe reactants that are potentially and unintentionally present in the experiment. It is for this reason that we have begun to estabish interactive quantum mechanics [213–216, 342–345] for an easy and simple interference of an operator with an autonomously running exploration protocol.

4.5 Workflows for Efficient Computation Protocols

As shown in Sects. 4.1 and 4.3, even if a single calculation may be efficient, the amount of data generated in an exhaustive exploration is immense and on the scale of 10^9 single-point calculations in brute force approaches. Therefore, smart automated protocols must be established to steer the exploration and reduce the number of calculations in order to maintain efficiency through all stages of the exploration process. A general paradigm for these workflows should be the automated selection of the minimally required algorithm for each specific task, while still being transparent, so that the applied approximations and their limitations can be understood. This requirement inherently requires flexible and modular workflows.

A prime challenge, which demands such an approach, is conformer generation. The generation of conformers of a chemical structure is necessary to reflect the structural ensemble accessible at a given temperature. This has been of major importance in the design of new pharmaceuticals [346]; hence, most conformer generation algorithms have been tested and compared on drug-like molecules [347, 348]. However, the importance of conformers in the

elucidation of reaction mechanisms and the calculation of reaction barriers has also been emphasized recently [161, 349, 350].

The most efficient methods for sampling the phase space of a chemical structure apply prior chemical knowledge to systematically generate conformers with rotations around rotatable bonds. Such algorithms can be developed based on heuristic rules [351–357], distance geometry [320, 358, 359], machine learning [360–367], or methods beneficial for quantum computing [368]. However, due to the combinatorial increase of possible conformers with the number of rotatable bonds, all these algorithms become unfeasible at a certain system size and stochastic sampling will be required. Hence, at this point, the conformer generation method must switch to algorithms that do not aim at covering the complete phase space, but sample most relevant regions of the PES within reasonable time. The most common examples in this regard are MD simulations with enhanced sampling techniques. Due to the plethora of different enhanced sampling techniques developed, we refer the reader to recent reviews [369–372] for discussions of their differences and advantages, and to Refs. [373–375] for different applications in the context of conformational sampling.

Since MD simulations are inherently expensive in terms of computing time, it is beneficial to additionally apply a multilevel approach for the evaluation of the PES. Larger systems can first be evaluated with faster, less accurate models and the most relevant conformers can later be studied with more accurate methods [376, 377]. However, within some finite computing time given also these algorithms will eventually fail for increasingly larger systems. The situation will then be similar to that of the prediction of a most stable protein fold, which is a conformational sampling problem at its core and for which specific knowledge-based approaches are advantageous [378, 379].

Should the end of a chain of available algorithmic switches be reached, the (meta)algorithm that implements the switching must recognize that the problem cannot be solved in an autonomous fashion. Structural fidelity can no longer be guarantedd and manual intervention might be required; e.g., the algorithm may warn the operator as already discussed in Ref. [213]. Such cases could then be approached within an interactive setting [161, 213, 216, 380].

Another well-known problem, which requires a sequential switching approach with maximum automation and minimal, but intuitive interaction, is the search for transition states (TS), i.e., first-order saddle points on a PES. Numerous stable and reliable TS optimization algorithms have been developed in the last fifty years [381–383]. However, due to the difficult nature of the optimization problem, a universally successful algorithm that is able to find all relevant TS from a limited number of start conformations will most likely never exist. Therefore, the software must be able to recognize that two or more structures should be connected via a TS, although a series of attempts of various algorithms has failed to locate a TS. This recognition can be based on physical or structural descriptors such as RMSD or graph comparisons. In such a case, the software can present the issue to the operator in an interactive manner, who can decide, whether this possible reaction is relevant and may even provide another educated guess for the TS based on real-time quantum chemistry [215, 216, 345].

Finally, we discuss required workflows to model heterogeneous processes based on the algorithms outlined in Sect. 4.2. In general, heterogeneous reactions can be explored with two different approaches. On the one hand, the chemical reactions can first be explored for molecules in the gas phase and in a second step all possible intermediates can be transferred onto the heterogeneous catalyst in an adsorption step. This saves computing time by minimizing the exploration trials in the solid state, which requires longer computing times, and enables double-ended searches for the reactions in the adsorbed state. However, this approach may fail if the intermediates of the heterogeneously catalyzed reaction are significantly different to those in the gas phase.

On the other hand, the exploration can proceed by screening for potential reactions directly in the solid state, which was outlined in our resource estimates in Sect. 4.3. In this approach, the reactants are adsorbed on a minimal surface slab directly and screened for conformations, either on the surface directly or by adsorbing various conformers. Then, the ranked adsorbed conformations of multiple reactants can be combined on a surface slab, which may require an extension of the surface. There, the second adsorption sites must be limited based on distance constraints and preferable adsorption sites already screened in the first adsorption as discussed in Sect. 4.2. The various possible extensions of the solid state structure must also be carefully stored and evaluated in reaction energy analysis across the reaction network.

5 Conclusions

Autonomous reaction network exploration presents an innovative, unbiased, and expansive approach of studying chemical reactivity. In this work, we discussed the potential of understanding catalytic processes in terms of automatically generated reaction networks from first-principles calculations and elaborated on required concepts and workflows. First-principles-based approaches are expensive in terms of computer time, but they are indispensable if detailed mechanistic insight is sought for. High throughput experimentation and data mining are complementary and may even deliver results for catalyst design purposes much faster than first-principles calculations. However, first-principles modeling

is also appropriate in cases where experiments are difficult to conduct (e.g., in high-throughput settings) or where data are incomplete.

As an example, we estimated the computational costs associated with exhaustive first-principles explorations in brute force approaches for a reference reaction network of 10^6 structures constructed by starting with two reactants. Our resource estimates showed that truly extensive explorations based on density functional theory calculations without activated pruning schemes (to cut deadwood in the exploration) are not feasible because of the sheer number of exploratory calculations to be carried out. This can be alleviated by suitable first-principles reactivity descriptors [160–163] which not only can suggest potentially reactive sites to be prioritized in the exploration process, but which can also determine those sites that are likely to be unreactive and that can therefore be given a very low priority in the exploration process.

The efficiency of building reaction networks with time-independent calculations is also increased by exploiting the fact that it parallelizes in a trivial manner because many elementary reaction trials can be carried out in parallel. Moreover, fast-but-very-approximate semi-empirical calculations can be employed for acquiring quickly a broad overview on a network. The key property of a suitable semi-empirical method must be structural fidelity since an energy refinement can be done in a subsequent step. In an autonomous setting, this is most efficiently accomplished by automated determination of those structures (based on uncertainty quantification) that should be subjected to reference calculations. Hence, computational costs are significantly reduced by such selective local refinement of the network data [142, 144, 229, 341].

If properly set up by tailored meta-algorithms that control efficient workflows, the autonomous exploration and design of catalytic processes based on reaction networks can be made routinely applicable. Its advantages, compared to standard manual exploration with standard quantum chemical techniques, are that orders of magnitude more reaction steps can be inspected, which is key for predictive work that must not miss out on important reaction steps. Obviously, no guarantee of completeness can be given, but there is no alternative other than autonomous procedures if huge sections of a reaction network shall be mapped, rather than focusing on a few steps that were considered relevant for some reason (e.g., based on prior experimental knowledge).

While this already holds true for a given set of reactants, catalytic processes should be described in open-ended and rolling reaction network explorations because minute amounts of impurities may interfere in a decisive way. This requires an interactive option for adding new reactants at any time of an autonomous exploration process, which can then benefit from human insight that can be exploited as a steering element in the exploration process.

To conclude, autonomous reaction network exploration presents a bright avenue for future computational catalysis as the depth of understanding acquired through the wealth of data are unprecedented and increases the probability of unexpected discoveries made in silico.

Appendix

Computational Methodology

All data management, quantum chemical calculations, and structure manipulations were conducted within our general software framework `SCINE` [212]. Its module `Chemoton` [161, 323] finds new elementary steps with single ended searches of geometrically aligned structures based on reaction coordinates. The reaction coordinates are based on reactive sites and directions of attack. The sites are determined by first-principles-based descriptors or a combinatorial geometric criterion in an exhaustive search as applied in this work. The directions of attack are derived from least steric hindrance. Our algorithm then extracts a potential transition state (TS) structure from a given reaction coordinate by pushing together (or pulling apart) two predefined lists of reactive sites with a constant force given as an input parameter. The force parameter controls the length of individual steps in the trajectory. The push or pull is stopped when a stop criterion, such as colliding nuclei or a change in bonding, has been reached. Upon pushing together (or pulling apart) the reactive centers, all atoms besides the reactive sites are continuously relaxed. This approach allows us to start screenings for potential elementary steps from anywhere on the PES, not necessarily starting at a minimum, the single force parameter does not control the allowed energy barriers, but rather allows to balance the computational costs and efficiency of finding a suitable TS guess because it solely controls the step length. Smaller step lengths allow for a more accurate location of a potential TS, but require more energy calculations. The potential TS structure is then refined with an optimization algorithm [384–387] and then automatically verified by *intrinsic reaction coordinate* (IRC) optimizations [388].

The elementary steps between structures are categorized into reactions, which connect compounds. A compound consists of multiple structures, which share an identical connectivity graph. The graphs are constructed by our library `Molassembler` [319] which provides the functionality for generating graphs and guess structures of conformers based on distance geometry for both organic and inorganic structures [320].

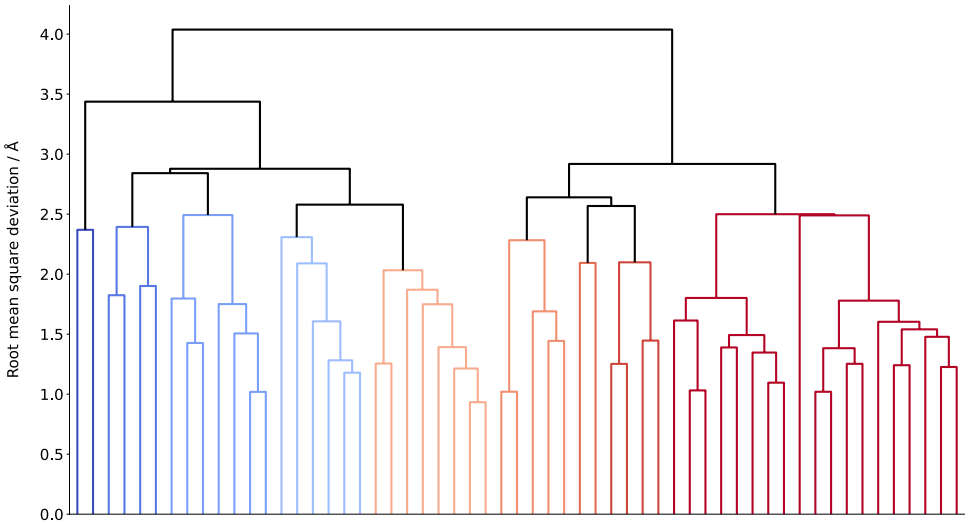


Fig. 10 A dendrogram of all 57 optimized conformers generated with *average linkage agglomerative hierarchical* clustering based on the RMSD. Clusters resulting from a cutoff value of 2.5 Å are colored

All calculations were performed by external programs, which can be controlled by the SCINE interface [389, 390] that allows to freely select and substitute the underlying physical model. The available methods range from system-focused parametrization [391], fast semi-empirical methods [322, 392], DFT [393–395], up to highly accurate multi reference calculations [396], possibly applying multiscale models [397, 398].

The uncatalyzed reference network of this work was explored with GFN2 as implemented in the xTB program [322, 399]. Molecular oxygen was calculated in its triplet state. For all bimolecular combinations of molecules within the exploration, the spin multiplicity was chosen as the sum of the individual multiplicities minus one. After one or more products were found, the smallest possible multiplicity, i.e., singlet states for molecules with an even number of electrons and doublet states otherwise, was assumed. Throughout this study electronic energies without zero-point vibrational corrections are considered. During the exploration, the Hessian was calculated for all newly found structures to confirm them as true minima before making them available for further elementary step trials.

All DFT calculations were carried out with the Perdew–Burke–Ernzerhof (PBE) exchange-correlation functional [400] with D3 dispersion correction [401] and Becke–Johnson damping [402]. The calculations of the organometallic catalyst were carried out with TURBOMOLE 7.4.1 [395] with the def2-SVP basis set [403] and density-fitting resolution of the identity through the def2/J auxiliary basis set [404]. The periodic DFT calculations in the Gaussian Plane Wave (GPW) formalism [405] were carried out with CP2K 8.1 [406] with the MOLOPT-DZVP basis set [407] and GTH pseudopotential [408], for which we implemented an interface in SCINE.

The crystal structure of Cu_2O was retrieved from the Materialsproject database [73] and the (001) surface was generated with pymatgen [93, 409]. The calculations were carried out on a $2 \times 2 \times 3$ supercell of the surface slab consisting of 72 atoms with 15 Å vacuum added in the z direction to avoid unphysical interactions of images in this direction.

Conformational Clustering

All 57 conformer structures were optimized as outlined above. The RMSD was calculated for every pair after an optimal alignment. We then constructed the dendrogram depicted in Fig. 10 based on *average linkage agglomerative hierarchical* clustering. The cutoff value was chosen to be 2.5 Å based on inspection of the dendrogram and the resulting centroids of the clusters, which were determined as the structures with the smallest sum of RMSDs to all other structures within the cluster. The nine representative structures shown in Fig. 7B) were, however, not the centroids, but those with the lowest electronic energy.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11244-021-01543-9.

Acknowledgements This publication was created as part of NCCR Catalysis, a National Centre of Competence in Research funded by the Swiss National Science Foundation, and the Swiss Government Excellence Scholarship for Foreign Scholars and Artists.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich.

References

1. Nørskov JK, Scheffler M, Toulhoat H (2006) Density functional theory in surface science and heterogeneous catalysis. MRS Bull 31:669–674
2. Balcells D, Clot E, Eisenstein O (2010) C-H bond activation in transition metal species from a computational perspective. Chem Rev 110:749–823
3. Lin Z (2010) Interplay between theory and experiment: computational organometallic and transition metal chemistry. Acc Chem Res 43:602–611
4. Sautet P, Delbecq F (2010) Catalysis and surface organometallic chemistry: a view from theory and simulations. Chem Rev 110:1788–1806
5. Nørskov JK, Abild-Pedersen F, Studt F, Bligaard T (2011) Density functional theory in surface chemistry and catalysis. Proc Natl Acad Sci USA 108:937–943
6. van der Kamp MW, Mulholland AJ (2013) Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. Biochemistry 52:2708–2728
7. Yang Q, Liu D, Zhong C, Li J-R (2013) Development of computational methodologies for metal-organic frameworks and their application in gas separations. Chem Rev 113:8261–8323
8. Thiel W (2014) Computational catalysis—past, present, and future. Angew Chem Int Ed 53:8605–8613
9. Speybroeck VV, Hemelsoet K, Joos L, Waroquier M, Bell RG, Catlow CRA (2015) Advances in theory and their application within the field of zeolite chemistry. Chem Soc Rev 44:7044–7111
10. Balcells D, Clot E, Eisenstein O, Nova A, Perrin L (2016) Deciphering selectivity in organic reactions: a multifaceted problem. Acc Chem Res 49:1070–1078
11. Lam Y-H, Grayson MN, Holland MC, Simon A, Houk KN (2016) Theory and modeling of asymmetric catalytic reactions. Acc Chem Res 49:750–762
12. Sperger T, Sanhueza IA, Schoenebeck F (2016) Computation and experiment: a powerful combination to understand and predict reactivities. Acc Chem Res 49:1311–1319
13. Vidossich P, Lledós A, Ujaque G (2016) First-principles molecular dynamics studies of organometallic complexes and homogeneous catalytic processes. Acc Chem Res 49:1271–1278
14. Zhang X, Chung LW, Wu Y-D (2016) New mechanistic insights on the selectivity of transition-metal-catalyzed organic reactions: the role of computational chemistry. Acc Chem Res 49:1302–1310
15. Romero-Rivera A, Garcia-Borràs M, Osuna S (2017) Computational tools for the evaluation of laboratory-engineered biocatalysts. Chem Commun 53:284–297
16. Seh ZW, Kibsgaard J, Dickens CF, Chorkendorff I, Nørskov JK, Jaramillo TF (2017) Combining theory and experiment in electrocatalysis: insights into materials design. Science 2017:355
17. Grajciar L, Heard CJ, Bondarenko AA, Polynski MV, Meeprasert J, Pidko EA, Nachtigall P (2018) Towards operando computational modeling in heterogeneous catalysis. Chem Soc Rev 47:8307–8348
18. Kulkarni A, Siahrostami S, Patel A, Nørskov JK (2018) Understanding catalytic activity trends in the oxygen reduction reaction. Chem Rev 118:2302–2312
19. Bruix A, Margraf JT, Andersen M, Reuter K (2019) First-principles-based multiscale modelling of heterogeneous catalysis. Nat Catal 2:659–670
20. Dubey KD, Shaik S (2019) Cytochrome P450—the wonderful nanomachine revealed through dynamic simulations of the catalytic cycle. Acc Chem Res 52:389–399
21. Vogiatzis KD, Polynski MV, Kirkland JK, Townsend J, Hashemi A, Liu C, Pidko EA (2019) Computational approach to molecular catalysis by 3d transition metals: challenges and opportunities. Chem Rev 119:2453–2523
22. Cui C-X, Chen H, Li S-J, Zhang T, Qu L-B, Lan Y (2020) Mechanism of Ir-catalyzed hydrogenation: a theoretical view. Coord Chem Rev 412:213251
23. Li J, Stephanopoulos MF, Xia Y (2020) Introduction: heterogeneous single-atom catalysis. Chem Rev 120:11699–11702
24. Funes-Ardoiz I, Schoenebeck F (2020) Established and emerging computational tools to study homogeneous catalysis-from quantum mechanics to machine learning. Chemistry 6:1904–1913
25. Reuter K, Metiu H (2020) Handbook of materials modeling. Springer International Publishing, Berlin, pp 1309–1319
26. Chen H, Li Y, Liu S, Xiong Q, Bai R, Wei D, Lan Y (2021) On the mechanism of homogeneous Pt-catalysis: a theoretical view. Coord Chem Rev 437:213863
27. Chen S, Peterson CW, Parker JA, Rice SA, Ferguson AL, Scherer NF (2021) Data-driven reaction coordinate discovery in overdamped and non-conservative systems: application to optical matter structural isomerization. Nat Commun 12:2548
28. Durand DJ, Fey N (2021) Building a toolbox for the analysis and prediction of ligand and catalyst effects in organometallic catalysis. Acc Chem Res 54:837–848
29. Wodrich MD, Sawatlon B, Busch M, Corminboeuf C (2021) The genesis of molecular volcano plots. Acc Chem Res 54:1107–1117
30. Hutchings GJ (2021) Spiers memorial lecture: understanding reaction mechanisms in heterogeneously catalysed reactions. Faraday Discuss 229:9–34
31. Catlow CRA (2021) Concluding remarks: reaction mechanisms in catalysis: perspectives and prospects. Faraday Discuss 229:502–513
32. Lledós A (2021) Computational organometallic catalysis: Where we are, where we are going. Eur J Inorg Chem 2021:n/a
33. Morales-García Á, Viñes F, Gomes JRB, Illas F (2021) Concepts, models, and methods in computational heterogeneous catalysis illustrated through CO_2 conversion. WIREs Comput Mol Sci 11:e1530
34. Rogge SMJ, Bavykina A, Hajek J, Garcia H, Olivos-Suarez AI, Sepúlveda-Escribano A, Vimont A, Clet G, Bazin P, Kapteijn

F, Daturi M, Ramos-Fernandez EV, Llabrés i Xamena FX, Speybroeck VV, Gascon J (2017) Metal-organic and covalent organic frameworks as single-site catalysts. Chem Soc Rev 46:3134–3184
35. Zhu L, Liu X-Q, Jiang H-L, Sun L-B (2017) Metal-organic frameworks for heterogeneous basic catalysis. Chem Rev 117:8129–8176
36. Bavykina A, Kolobov N, Khan IS, Bau JA, Ramirez A, Gascon J (2020) Metal-organic frameworks in heterogeneous catalysis: recent progress, new trends, and future perspectives. Chem Rev 120:8468–8535
37. Freund R et al (2021) 25 Years of reticular chemistry. Angew Chem Int Ed 60:23946-23974
38. Yang X-F, Wang A, Qiao B, Li J, Liu J, Zhang T (2013) Single-atom catalysts: a new frontier in heterogeneous catalysis. Acc Chem Res 46:1740–1748
39. Kaiser SK, Chen Z, Faust Akl D, Mitchell S, Pérez-Ramírez J (2020) Single-atom catalysts across the periodic table. Chem Rev 120:11703–11809
40. Samantaray MK, D'Elia V, Pump E, Falivene L, Harb M, Chikh SO, Cavallo L, Basset J-M (2020) The comparison between single atom catalysis and surface organometallic catalysis. Chem Rev 120:734–813
41. Li Z, Ji S, Liu Y, Cao X, Tian S, Chen Y, Niu Z, Li Y (2020) Well-defined materials for heterogeneous catalysis: from nanoparticles to isolated single-atom sites. Chem Rev 120:623–682
42. Wegener SL, Marks TJ, Stair PC (2012) Design strategies for the molecular level synthesis of supported catalysts. Acc Chem Res 45:206–214
43. Copéret C, Comas-Vives A, Conley MP, Estes DP, Fedorov A, Mougel V, Nagae H, Núñez-Zarur F, Zhizhko PA (2016) Surface organometallic and coordination chemistry toward single-site heterogeneous catalysts: strategies, methods, structures, and activities. Chem Rev 116:323–421
44. Ye R, Zhao J, Wickemeyer BB, Toste FD, Somorjai GA (2018) Foundations and strategies of the construction of hybrid catalysts for optimized performances. Nat Catal 1:318–325
45. Copéret C (2019) Fuels and energy carriers from single-site catalysts prepared via surface organometallic chemistry. Nat Energy 4:1018–1024
46. Chen D-F, Han Z-Y, Zhou X-L, Gong L-Z (2014) Asymmetric organocatalysis combined with metal catalysis: concept, proof of concept, and beyond. Acc Chem Res 47:2365–2377
47. Wörsdörfer B, Woycechowsky KJ, Hilvert D (2011) Directed evolution of a protein container. Science 331:589–592
48. Leenders SHAM, Gramage-Doria R, de Bruin B, Reek JNH (2014) Transition metal catalysis in confined spaces. Chem Soc Rev 44:433–448
49. Tetter S, Hilvert D (2017) Enzyme encapsulation by a ferritin cage. Angew Chem Int Ed 56:14933–14936
50. Jongkind LJ, Caumes X, Hartendorp APT, Reek JNH (2018) Ligand template strategies for catalyst encapsulation. Acc Chem Res 51:2115–2128
51. Azuma Y, Edwardson TGW, Hilvert D (2018) Tailoring lumazine synthase assemblies for bionanotechnology. Chem Soc Rev 47:3543–3557
52. Palmiero UC, Küffner AM, Krumeich F, Faltova L, Arosio P (2020) Adaptive chemoenzymatic microreactors composed of inorganic nanoparticles and bioinspired intrinsically disordered proteins. Angew Chem Int Ed 59:8138–8142
53. Wu J, Wang X, Wang Q, Lou Z, Li S, Zhu Y, Qin L, Wei H (2019) Nanomaterials with enzyme-like characteristics (nanozymes): next-generation artificial enzymes (II). Chem Soc Rev 48:1004–1076
54. Lv C, Zhang X, Liu Y, Zhang T, Chen H, Zang J, Zheng B, Zhao G (2021) Redesign of protein nanocages: the way from 0D, 1D, 2D to 3D assembly. Chem Soc Rev 50:3957–3989
55. Micura R, Höbartner C (2020) Fundamental studies of functional nucleic acids: aptamers, riboswitches, ribozymes and DNAzymes. Chem Soc Rev 49:7331–7353
56. Davis HJ, Ward TR (2019) Artificial metalloenzymes: challenges and opportunities. ACS Cent Sci 5:1120–1136
57. Arnold FH (2019) Innovation by evolution: bringing new chemistry to life (Nobel lecture). Angew Chem Int Ed 58:14420–14426
58. Hofmann R, Akimoto G, Wucherpfennig TG, Zeymer C, Bode JW (2020) Lysine acylation using conjugating enzymes for site-specific modification and ubiquitination of recombinant proteins. Nat Chem 12:1008–1015
59. Chen K, Arnold FH (2020) Engineering new catalytic activities in enzymes. Nat Catal 3:203–213
60. Armiento R, Kozinsky B, Fornari M, Ceder G (2011) Screening for high-performance piezoelectrics using high-throughput density functional theory. Phys Rev B 84:014103
61. Agrawal A, Choudhary A (2016) Perspective: materials informatics and big data: realization of the fourth paradigm of science in materials science. APL Mater 4:053208
62. Himanen L, Geurts A, Foster AS, Rinke P (2019) Data-driven materials science: status, challenges, and perspectives. Adv Sci 6:1900808
63. Armiento R (2020) Machine learning meets quantum physics; lecture notes in physics. Springer International Publishing, Berlin, pp 377–395
64. Yu Y-X, Yang J, Zhu K-K, Sui Z-J, Chen D, Zhu Y-A, Zhou X-G (2021) High-throughput screening of alloy catalysts for dry methane reforming. ACS Catal 11:8881–8894
65. Blau SM, Patel HD, Spotte-Smith EWC, Xie X, Dwaraknath S, Persson KA (2021) A chemically consistent graph architecture for massive reaction networks applied to solid-electrolyte interphase formation. Chem Sci 12:4931–4939
66. McDermott MJ, Dwaraknath SS, Persson KA (2021) A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. Nat Commun 12:3097
67. Vaucher AC, Schwaller P, Geluykens J, Nair VH, Iuliano A, Laino T (2021) Inferring experimental procedures from text-based representations of chemical reactions. Nat Commun 12:2573
68. Schwaller P, Hoover B, Reymond J-L, Strobelt H, Laino T (2021) Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. Sci Adv 7:eabe4166
69. Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, Amador-Bedolla C, Sánchez-Carrera RS, Gold-Parker A, Vogt L, Brockway AM, Aspuru-Guzik A (2011) The harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. J Phys Chem Lett 2:2241–2251
70. Hummelshøj JS, Abild-Pedersen F, Studt F, Bligaard T, Nørskov JK (2012) CatApp: a web application for surface chemistry and heterogeneous catalysis. Angew Chem Int Ed 51:272–274
71. Curtarolo S, Setyawan W, Wang S, Xue J, Yang K, Taylor RH, Nelson LJ, Hart GLW, Sanvito S, Buongiorno-Nardelli M, Mingo N, Levy O (2012) AFLOWLIB.ORG: a distributed materials properties repository from high-throughput ab initio calculations. Comput Mater Sci 58:227–235
72. Landis DD, Hummelshøj JS, Nestorov S, Greeley J, Dułak M, Bligaard T, Nørskov JK, Jacobsen KW (2012) The computational materials repository. Comput Sci Eng 14:51–57
73. Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson KA (2013) Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater 1:011002

74. Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C (2013) Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). JOM 65:1501–1509
75. Chung YG, Camp J, Haranczyk M, Sikora BJ, Bury W, Krungleviciute V, Yildirim T, Farha OK, Sholl DS, Snurr RQ (2014) Computation-ready, experimental metal-organic frameworks: a tool to enable high-throughput screening of nanoporous crystals. Chem Mater 26:6185–6192
76. Álvarez-Moreno M, de Graaf C, López N, Maseras F, Poblet JM, Bo C (2015) Managing the computational chemistry big data problem: the ioChem-BD platform. J Chem Inf Model 55:95–103
77. Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, Aykol M, Rühl S, Wolverton C (2015) The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. npj Comput Mater 1:1–15
78. Dima A et al (2016) Informatics infrastructure for the materials genome initiative. JOM 68:2053–2064
79. O'Mara J, Meredig B, Michel K (2016) Materials data infrastructure: a case study of the citrination platform to examine data import, storage, and access. JOM 68:2031–2034
80. Borysov SS, Geilhufe RM, Balatsky AV (2017) Organic materials database: an open-access online database for data mining. PLoS ONE 12:e0171501
81. Draxl C, Scheffler M (2018) NOMAD: the FAIR concept for big data-driven materials science. MRS Bull 43:676–682
82. Zakutayev A, Wunder N, Schwarting M, Perkins JD, White R, Munch K, Tumas W, Phillips C (2018) An open experimental database for exploring inorganic materials. Sci Data 5:180053
83. Winther KT, Hoffmann MJ, Boes JR, Mamun O, Bajdich M, Bligaard T (2019) Catalysis-Hub.Org, an open electronic structure database for surface reactions. Sci Data 6:75
84. Mamun O, Winther KT, Boes JR, Bligaard T (2019) High-throughput calculations of catalytic properties of bimetallic alloy surfaces. Sci Data 6:76
85. Blokhin E, Villars P (2020) Handbook of materials modeling: methods: theory and modeling. Springer, Berlin, pp 1837–1861
86. Choudhary K et al (2020) JARVIS: an integrated infrastructure for data-driven materials design. npj Comput Mater 6:173
87. Talirz L et al (2020) Materials cloud, a platform for open computational science. Sci Data 7:299
88. Gimadiev T, Nugmanov R, Batyrshin D, Madzhidov T, Maeda S, Sidorov P, Varnek A (2021) Combined graph/relational database management system for calculated chemical reaction pathway data. J Chem Inf Model 61:554–559
89. Pablo-García S, Álvarez-Moreno M, López N (2021) Turning chemistry into information for heterogeneous catalysis. Int J Quantum Chem 121:e26382
90. Nakata M, Shimazaki T (2017) PubChemQC project: a large-scale first-principles electronic structure database for data-driven chemistry. J Chem Inf Model 57:1300–1308
91. Smith DGA, Altarawy D, Burns LA, Welborn M, Naden LN, Ward L, Ellis S, Pritchard BP, Crawford TD (2021) The MolSSI QCArchive project: an open-source platform to compute, organize, and share quantum chemistry data. WIREs Comput Mol Sci 11:e1491
92. Andersen CW et al (2021) OPTIMADE, an API for exchanging materials data. Sci Data 8:217
93. Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier VL, Persson KA, Ceder G (2013) Python materials genomics (Pymatgen): a robust, open-source python library for materials analysis. Comput Mater Sci 68:314–319
94. Jain A, Ong SP, Chen W, Medasani B, Qu X, Kocher M, Brafman M, Petretto G, Rignanese G-M, Hautier G, Gunter D, Persson KA (2015) FireWorks: a dynamic workflow system designed for high-throughput applications. Concurr Comput 27:5037–5059
95. Pizzi G, Cepellotti A, Sabatini R, Marzari N, Kozinsky B (2016) AiiDA: automated interactive infrastructure and database for computational science. Comput Mater Sci 111:218–230
96. Mathew K et al (2017) Atomate: a high-level interface to generate, execute, and analyze computational materials science workflows. Comput Mater Sci 139:140–152
97. Aagesen LK et al (2018) PRISMS: an integrated, open-source framework for accelerating predictive structural materials science. JOM 70:2298–2314
98. Schleder GR, Padilha ACM, Acosta CM, Costa M, Fazzio A (2019) From DFT to machine learning: recent approaches to materials science—a review. J Phys 2:032001
99. Wheeler D, Keller T, DeWitt SJ, Jokisaari AM, Schwen D, Guyer JE, Aagesen LK, Heinonen OG, Tonks MR, Voorhees PW, Warren JA (2019) PFHub: the phase-field community hub. J Open Res Software 7:29
100. Yang S, Bier I, Wen W, Zhan J, Moayedpour S, Marom N (2020) Ogre: a python package for molecular crystal surface generation with applications to surface energy and crystal habit prediction. J Chem Phys 152:244122
101. Youn Y, Lee M, Hong C, Kim D, Kim S, Jung J, Yim K, Han S (2020) AMP2: a fully automated program for ab initio calculations of crystalline materials. Comput Phys Commun 256:107450
102. Huber SP et al (2021) Common workflows for computing material properties using different quantum engines. npj Comput Mater 7:1–12
103. Brlec K, Davies D, Scanlon D (2021) Surfaxe: systematic surface calculations. J Open Source Softw 6:3171
104. Wang G, Peng L, Li K, Zhu L, Zhou J, Miao N, Sun Z (2021) ALKEMIE: an intelligent computational platform for accelerating materials discovery and design. Comput Mater Sci 186:110064
105. Curtarolo S, Hart GLW, Nardelli MB, Mingo N, Sanvito S, Levy O (2013) The high-throughput highway to computational materials design. Nat Mater 12:191–201
106. Hachmann J, Olivares-Amaya R, Jinich A, Appleton AL, Blood-Forsythe MA, Seress LR, Román-Salgado C, Trepte K, Atahan-Evrenk S, Er S, Shrestha S, Mondal R, Sokolov A, Bao Z, Aspuru-Guzik A (2014) Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry—the harvard clean energy project. Energy Environ Sci 7:698–704
107. Pyzer-Knapp EO, Suh C, Gómez-Bombarelli R, Aguilera-Iparraguirre J, Aspuru-Guzik A (2015) What is high-throughput virtual screening? A perspective from organic materials discovery. Annu Rev Mater Res 45:195–216
108. Takahashi K, Takahashi L, Miyazato I, Fujima J, Tanaka Y, Uno T, Satoh H, Ohno K, Nishida M, Hirai K, Ohyama J, Nguyen TN, Nishimura S, Taniike T (2019) The rise of catalyst informatics: towards catalyst genomics. ChemCatChem 11:1146–1152
109. Luo S, Li T, Wang X, Faizan M, Zhang L (2021) High-throughput computational materials screening and discovery of optoelectronic semiconductors. WIREs Comput Mol Sci 11:e1489
110. Tran K, Palizhati A, Back S, Ulissi ZW (2018) Dynamic workflows for routine materials discovery in surface science. J Chem Inf Model 58:2392–2400
111. Bligaard T, Nørskov JK, Dahl S, Matthiesen J, Christensen CH, Sehested J (2004) The Brønsted-Evans-Polanyi relation and the volcano curve in heterogeneous catalysis. J Catal 224:206–217
112. Ulissi ZW, Medford AJ, Bligaard T, Nørskov JK (2017) To address surface reaction network complexity using scaling relations machine learning and DFT calculations. Nat Commun 8:14621
113. Mazeau EJ, Satpute P, Blöndal K, Goldsmith CF, West RH (2021) Automated mechanism generation using linear scaling

relationships and sensitivity analyses applied to catalytic partial oxidation of methane. ACS Catal 11:7114–7125
114. Xin H, Holewinski A, Linic S (2012) Predictive structure-reactivity models for rapid screening of Pt-based multimetallic electrocatalysts for the oxygen reduction reaction. ACS Catal 2:12–16
115. Zhao Z-J, Liu S, Zha S, Cheng D, Studt F, Henkelman G, Gong J (2019) Theory-guided design of catalytic materials using scaling relationships and reactivity descriptors. Nat Rev Mater 4:792–804
116. Gao W, Chen Y, Li B, Liu S-P, Liu X, Jiang Q (2020) Determining the adsorption energies of small molecules with the intrinsic properties of adsorbates and substrates. Nat Commun 11:1196
117. Xu W, Andersen M, Reuter K (2021) Data-driven descriptor engineering and refined scaling relations for predicting transition metal oxide reactivity. ACS Catal 11:734–742
118. Ulissi ZW, Tang MT, Xiao J, Liu X, Torelli DA, Karamad M, Cummins K, Hahn C, Lewis NS, Jaramillo TF, Chan K, Nørskov JK (2017) Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for CO_2 reduction. ACS Catal 7:6600–6608
119. Takahashi K, Miyazato I (2018) Rapid estimation of activation energy in heterogeneous catalytic reactions via machine learning. J Comput Chem 39:2405–2408
120. Takahashi K, Miyazato I, Nishimura S, Ohyama J (2018) Unveiling hidden catalysts for the oxidative coupling of methane based on combining machine learning with literature data. ChemCatChem 10:3223–3228
121. Tran K, Ulissi ZW (2018) Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. Nat Catal 1:696–703
122. Andersen M, Levchenko SV, Scheffler M, Reuter K (2019) Beyond scaling relations for the description of catalytic materials. ACS Catal 9:2752–2759
123. Palizhati A, Zhong W, Tran K, Back S, Ulissi ZW (2019) Towards predicting intermetallics surface properties with high-throughput DFT and convolutional neural networks. J Chem Inf Model 59:4742–4749
124. Back S, Tran K, Ulissi ZW (2019) Toward a design of active oxygen evolution catalysts: insights from automated density functional theory calculations and machine learning. ACS Catal 9:7651–7659
125. Deimel M, Reuter K, Andersen M (2020) Active site representation in first-principles microkinetic models: data-enhanced computational screening for improved methanation catalysts. ACS Catal 10:13729–13736
126. Praveen CS, Comas-Vives A (2020) Design of an accurate machine learning algorithm to predict the binding energies of several adsorbates on multiple sites of metal surfaces. ChemCatChem 12:4611–4617
127. Xu J, Cao X-M, Hu P (2021) Perspective on computational reaction prediction using machine learning methods in heterogeneous catalysis. Phys Chem Chem Phys 23:11155–11179
128. Friederich P, Häse F, Proppe J, Aspuru-Guzik A (2021) Machine-learned potentials for next-generation matter simulations. Nat Mater 20:750–761
129. Li X, Chiong R, Page AJ (2021) Group and period-based representations for improved machine learning prediction of heterogeneous alloy catalysts. J Phys Chem Lett 12:5156–5162
130. Li S, Liu Y, Chen D, Jiang Y, Nie Z, Pan F (2021) Encoding the atomic structure for machine learning in materials science. WIREs Comput Mol Sci n/a:e1558
131. Rosen AS, Iyer SM, Ray D, Yao Z, Aspuru-Guzik A, Gagliardi L, Notestein JM, Snurr RQ (2021) Machine learning the quantum-chemical properties of metal-organic frameworks for accelerated materials discovery. Matter 4:1578–1597
132. Andersen M, Reuter K (2021) Adsorption enthalpies for catalysis modeling through machine-learned descriptors. Acc Chem Res 54(12):2741–2749
133. Pablo-García S, García-Muelas R, Sabadell-Rendón A, López N (2021) Dimensionality reduction of complex reaction networks in heterogeneous catalysis: from linear-scaling relationships to statistical learning techniques. WIREs Comput Mol Sci 11:e1540
134. Esterhuizen JA, Goldsmith BR, Linic S (2021) Uncovering electronic and geometric descriptors of chemical activity for metal alloys and oxides using unsupervised machine learning. Chem Catal 107:2411–2502
135. Back S, Na J, Tran K, Ulissi ZW (2020) In silico discovery of active, stable, CO-tolerant and cost-effective electrocatalysts for hydrogen evolution and oxidation. Phys Chem Chem Phys 22:19454–19458
136. Mortensen JJ, Kaasbjerg K, Frederiksen SL, Nørskov JK, Sethna JP, Jacobsen KW (2005) Bayesian error estimation in density-functional theory. Phys Rev Lett 95:216401
137. Hellman A et al (2006) Predicting catalysis: understanding ammonia synthesis from first-principles calculations. J Phys Chem B 110:17719–17735
138. Wellendorff J, Lundgaard KT, Møgelhøj A, Petzold V, Landis DD, Nørskov JK, Bligaard T, Jacobsen KW (2012) Density functionals for surface science: exchange-correlation model development with bayesian error estimation. Phys Rev B 85:235149
139. Medford AJ, Wellendorff J, Vojvodic A, Studt F, Abild-Pedersen F, Jacobsen KW, Bligaard T, Nørskov JK (2014) Assessing the reliability of calculated catalytic ammonia synthesis rates. Science 345:197–200
140. Simm GN, Reiher M (2016) Systematic error estimation for chemical reaction energies. J Chem Theory Comput 12:2762–2773
141. Tran K, Neiswanger W, Yoon J, Zhang Q, Xing E, Ulissi ZW (2020) Methods for comparing uncertainty quantifications for material property predictions. Mach Learn 1:025006
142. Proppe J, Husch T, Simm GN, Reiher M (2017) Uncertainty quantification for quantum chemical models of complex reaction networks. Faraday Discuss 195:497–520
143. Li Q, García-Muelas R, López N (2018) Microkinetics of alcohol reforming for H2 production from a FAIR density functional theory database. Nat Commun 9:526
144. Simm GN, Reiher M (2018) Error-controlled exploration of chemical reaction networks with gaussian processes. J Chem Theory Comput 14:5238–5248
145. Stocker S, Csányi G, Reuter K, Margraf JT (2020) Machine learning in chemical reaction space. Nat Commun 11:5505
146. Freund H-J, Meijer G, Scheffler M, Schlögl R, Wolf M (2011) CO oxidation as a prototypical reaction for heterogeneous processes. Angew Chem Int Ed 50:10064–10094
147. Schlögl R (2015) Heterogeneous catalysis. Angew Chem Int Ed 54:3465–3520
148. Sameera WMC, Maeda S, Morokuma K (2016) Computational catalysis using the artificial force induced reaction method. Acc Chem Res 49:763–773
149. Vázquez SA, Otero XL, Martinez-Nunez E (2018) A trajectory-based method to explore reaction mechanisms. Molecules 23:3156
150. Dewyer AL, Argüelles AJ, Zimmerman PM (2018) Methods for exploring reaction space in molecular systems. WIREs Comput Mol Sci 8:e1354
151. Simm GN, Vaucher AC, Reiher M (2019) Exploration of reaction pathways and chemical transformation networks. J Phys Chem A 123:385–399
152. Unsleber JP, Reiher M (2020) The exploration of chemical reaction networks. Annu Rev Phys Chem 71:121–142

153. Gu T, Wang B, Chen S, Yang B (2020) Automated generation and analysis of the complex catalytic reaction network of ethanol synthesis from syngas on Rh(111). ACS Catal 10:6346–6355
154. Margraf JT, Reuter K (2019) Systematic enumeration of elementary reaction steps in surface catalysis. ACS Omega 4:3370–3379
155. Liu M, Dana AG, Johnson M, Goldman M, Jocher A, Payne AM, Grambow C, Han K, Yee NW-W, Mazeau E, Blondal K, West R, Goldsmith F, Green WH (2020) Reaction mechanism generator v3.0: advances in automatic mechanism generation. J Chem Inf Model 61(6):2686–2696
156. Wang B, Chen S, Zhang J, Li S, Yang B (2019) Propagating DFT uncertainty to mechanism determination, degree of rate control, and coverage analysis: the kinetics of dry reforming of methane. J Phys Chem C 123:30389–30397
157. Zhai H, Alexandrova AN (2017) Fluxionality of catalytic clusters: when it matters and how to address it. ACS Catal 7:1905–1911
158. Copéret C (2019) Single-sites and nanoparticles at tailored interfaces prepared via surface organometallic chemistry from thermolytic molecular precursors. Acc Chem Res 52:1697–1708
159. Mars P, Krevelen DWV (1954) Oxidations carried out by means of vanadium oxide catalysts. Chem Eng Sci 3:41–59
160. Bergeler M, Simm GN, Proppe J, Reiher M (2015) Heuristics-guided exploration of reaction mechanisms. J Chem Theory Comput 11:5712–5722
161. Simm GN, Reiher M (2017) Context-driven exploration of complex chemical reaction networks. J Chem Theory Comput 13:6108–6119
162. Grimmel SA, Reiher M (2019) The electrostatic potential as a descriptor for the protonation propensity in automated exploration of reaction mechanisms. Faraday Discuss 220:443–463
163. Grimmel SA, Reiher M (2021) On the predictive power of chemical concepts. CHIMIA 75:311–318
164. Maeda S, Ohno K, Morokuma K (2013) Systematic exploration of the mechanism of chemical reactions: the global reaction route mapping (GRRM) strategy using the ADDF and AFIR methods. Phys Chem Chem Phys 15:3683–3701
165. Rappoport D, Galvin CJ, Zubarev DY, Aspuru-Guzik A (2014) Complex chemical reaction networks from heuristics-aided quantum chemistry. J Chem Theory Comput 10:897–907
166. Kim Y, Choi S, Kim WY (2014) Efficient Basin-Hopping sampling of reaction intermediates through molecular fragmentation and graph theory. J Chem Theory Comput 10:2419–2426
167. Wang L-P, Titov A, McGibbon R, Liu F, Pande VS, Martínez TJ (2014) Discovering chemistry with an ab initio nanoreactor. Nat Chem 6:1044
168. Zimmerman PM (2015) Single-ended transition state finding with the growing string method. J Comput Chem 36:601–611
169. Gao CW, Allen JW, Green WH, West RH (2016) Reaction mechanism generator: automatic construction of chemical kinetic mechanisms. Comput Phys Commun 203:212–225
170. Habershon S (2016) Automated prediction of catalytic mechanism and rate law using graph-based reaction path sampling. J Chem Theory Comput 12:1786–1798
171. Guan Y, Ingman VM, Rooks BJ, Wheeler SE (2018) AARON: an automated reaction optimizer for new catalysts. J Chem Theory Comput 14:5249–5261
172. Kim Y, Kim JW, Kim Z, Kim WY (2018) Efficient prediction of reaction paths through molecular graph and reaction network analysis. Chem Sci 9:825–835
173. Grimme S (2019) Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations. J Chem Theory Comput 15:2847–2862
174. Rizzi V, Mendels D, Sicilia E, Parrinello M (2019) Blind search for complex chemical pathways using harmonic linear discriminant analysis. J Chem Theory Comput. 15:4507–4515
175. Jara-Toro RA, Pino GA, Glowacki DR, Shannon RJ, Martínez-Núñez E (2020) Enhancing automated reaction discovery with boxed molecular dynamics in energy space. ChemSystemsChem 2:e1900024
176. Zhao Q, Savoie BM (2021) Simultaneously improving reaction coverage and computational cost in automated reaction prediction tasks. Nat Comput Sci 1:479–490
177. Goldsmith CF, West RH (2017) Automatic generation of microkinetic mechanisms for heterogeneous catalysis. J Phys Chem C 121:9970–9981
178. Delgado KH, Maier L, Tischer S, Zellner A, Stotz H, Deutschmann O (2015) Surface reaction kinetics of steam- and CO_2-reforming as well as oxidation of methane over nickel-based catalysts. Catalysts 5:871–904
179. Jafari M, Zimmerman PM (2018) Uncovering reaction sequences on surfaces through graphical methods. Phys Chem Chem Phys 20:7721–7729
180. Larsen AH et al (2017) The atomic simulation environment—a python library for working with atoms. J Phys 29:273002
181. Jafari M, Zimmerman PM (2017) Reliable and efficient reaction path and transition state finding for surface reactions with the growing string method. J Comput Chem 38:645–658
182. Maeda S, Sugiyama K, Sumiya Y, Takagi M, Saita K (2018) Global reaction route mapping for surface adsorbed molecules: a case study for H_2O on Cu(111) surface. Chem Lett 47:396–399
183. Sugiyama K, Sumiya Y, Takagi M, Saita K, Maeda S (2019) Understanding CO oxidation on the Pt(111) surface based on a reaction route network. Phys Chem Chem Phys 21:14366–14375
184. Sugiyama K, Saita K, Maeda S (2021) A reaction route network for methanol decomposition on a Pt(111) surface. J Comput Chem 42:2163–2169
185. Maeda S, Harabuchi Y (2021) Exploring paths of chemical transformations in molecular and periodic systems: an approach utilizing force. WIREs Comput Mol Sci 11:e1538
186. Hatanaka M, Maeda S, Morokuma K (2013) Sampling of transition states for predicting diastereoselectivity using automated search method-aqueous lanthanide-catalyzed mukaiyama aldol reaction. J Chem Theory Comput 9:2882–2886
187. Yoshimura T, Maeda S, Taketsugu T, Sawamura M, Morokuma K, Mori S (2017) Exploring the full catalytic cycle of rhodium (I)-BINAP-catalysed isomerisation of allylic amines: a graph theory approach for path optimisation. Chem Sci 8:4475–4488
188. Reyes RL, Sato M, Iwai T, Suzuki K, Maeda S, Sawamura M (2020) Asymmetric remote C-H borylation of aliphatic amides and esters with a modular iridium catalyst. Science 369:970–974
189. Nett AJ, Zhao W, Zimmerman PM, Montgomery J (2015) Highly active nickel catalysts for C-H functionalization identified through analysis of off-cycle intermediates. J Am Chem Soc 137:7636–7639
190. Ludwig JR, Zimmerman PM, Gianino JB, Schindler CS (2016) Iron(III)-catalysed carbonyl-olefin metathesis. Nature 533:374–379
191. Smith ML, Leone AK, Zimmerman PM, McNeil AJ (2016) Impact of preferential π-binding in catalyst-transfer polycondensation of thiazole derivatives. ACS Macro Lett 5:1411–1415
192. Zhao Y, Nett AJ, McNeil AJ, Zimmerman PM (2016) Computational mechanism for initiation and growth of poly (3-hexylthiophene) using palladium N-heterocyclic carbene precatalysts. Macromolecules 49:7632–7641
193. Ludwig JR, Phan S, McAtee CC, Zimmerman PM, III JJD, Schindler CS (2017) Mechanistic investigations of the iron (III)-catalyzed carbonyl-olefin metathesis reaction. J Am Chem Soc 139:10832–10842

194. Dewyer AL, Zimmerman PM (2017) Simulated mechanism for palladium-catalyzed, directed γ-arylation of piperidine. ACS Catal 7:5466–5477
195. Ludwig JR, Watson RB, Nasrallah DJ, Gianino JB, Zimmerman PM, Wiscons RA, Schindler CS (2018) Interrupted carbonyl-olefin metathesis via oxygen atom transfer. Science 361:1363–1369
196. Rudenko AE, Clayman NE, Walker KL, Maclaren JK, Zimmerman PM, Waymouth RM (2018) Ligand-induced reductive elimination of ethane from azopyridine palladium dimethyl complexes. J Am Chem Soc 140:11408–11415
197. Lipinski BM, Walker KL, Clayman NE, Morris LS, Jugovic TME, Roessler AG, Getzler YDYL, MacMillan SN, Zare RN, Zimmerman PM, Waymouth RM, Coates GW (2020) Mechanistic study of isotactic poly(propylene oxide) synthesis using a tethered bimetallic chromium salen catalyst. ACS Catal 10:8960–8967
198. Malakar T, Zimmerman PM (2021) Brønsted-acid-catalyzed intramolecular carbonyl-olefin reactions: interrupted metathesis vs carbonyl-Ene reaction. J Org Chem 86:3008–3016
199. Malakar T, Hanson CS, Devery JJ, Zimmerman PM (2021) Combined theoretical and experimental investigation of Lewis acid-carbonyl interactions for metathesis. ACS Catal 11:4381–4394
200. Zhang X-J, Shang C, Liu Z-P (2017) Stochastic surface walking reaction sampling for resolving heterogeneous catalytic reaction network: a revisit to the mechanism of water-gas shift reaction on Cu. J Chem Phys 147:152706
201. Guan S-H, Zhang X-J, Liu Z-P (2015) Energy landscape of zirconia phase transitions. J Am Chem Soc 137:8010–8013
202. Ma S, Huang S-D, Liu Z-P (2019) Dynamic coordination of cations and catalytic selectivity on zinc-chromium oxide alloys during syngas conversion. Nat Catal 2:671–677
203. Ma S, Shang C, Liu Z-P (2019) Heterogeneous catalysis from structure to activity via SSW-NN method. J Chem Phys 151:050901
204. Huang S-D, Shang C, Kang P-L, Zhang X-J, Liu Z-P (2019) LASP: fast global potential energy surface exploration. WIREs Comput Mol Sci 9:e1415
205. Ismail I, Stuttaford-Fowler HBVA, Ochan Ashok C, Robertson C, Habershon S (2019) Automatic proposal of multistep reaction mechanisms using a graph-driven search. J Phys Chem A 123:3407–3417
206. Song X, Fagiani MR, Debnath S, Gao M, Maeda S, Taketsugu T, Gewinner S, Schöllkopf W, Asmis KR, Lyalin A (2017) Excess charge driven dissociative hydrogen adsorption on Ti_2 O_4^-. Phys Chem Chem Phys 19:23154–23161
207. Iwasa T, Sato T, Takagi M, Gao M, Lyalin A, Kobayashi M, ichi Shimizu K, Maeda S, Taketsugu T (2018) Combined automated reaction pathway searches and sparse modeling analysis for catalytic properties of lowest energy twins of Cu_{13}. J Phys Chem A 123:210–217
208. Ichino T, Takagi M, Maeda S (2019) A systematic study on bond activation energies of NO, N_2, and O_2 on hexamers of eight transition metals. ChemCatChem 11:1346–1353
209. Heck RF, Breslow DS (1961) The reaction of cobalt hydrotetracarbonyl with olefins. J Am Chem Soc 83:4023–4027
210. Maeda S, Morokuma K (2012) Toward predicting full catalytic cycle using automatic reaction path search method: a case study on $HCo(CO)_3$-catalyzed hydroformylation. J Chem Theory Comput 8:380–385
211. Varela JA, Vázquez SA, Martínez-Núñez E (2017) An automated method to find reaction mechanisms and solve the kinetics in organometallic catalysis. Chem Sci 8:3843–3851
212. Software for Chemical Interaction and Networks (SCINE). https://scine.ethz.ch/. Accessed (June 2021)
213. Haag MP, Reiher M (2014) Studying chemical reactivity in a virtual environment. Faraday Discuss 169:89–118
214. Vaucher AC, Haag MP, Reiher M (2016) Real-time feedback from iterative electronic structure calculations. J Comput Chem 37:805–812
215. Heuer MA, Vaucher AC, Haag MP, Reiher M (2018) Integrated reaction path processing from sampled structure sequences. J Chem Theory Comput 14:2052–2062
216. Haag MP, Vaucher AC, Bosson M, Redon S, Reiher M (2014) Interactive chemical reactivity exploration. ChemPhysChem 15:3301–3319
217. Compiled by A. D. McNaught and A. Wilkinson, catalyst. https://goldbook.iupac.org/terms/view/C00876. Accessed (June 2021)
218. Froment GF (2005) Single event kinetic modeling of complex catalytic processes. Catal Rev Sci Eng 47:83–124
219. Glowacki DR, Liang C-H, Morley C, Pilling MJ, Robertson SH (2012) MESMER: an open-source master equation solver for multi-energy well reactions. J Phys Chem A 116:9545–9560
220. Sabbe MK, Reyniers M-F, Reuter K (2012) First-principles kinetic modeling in heterogeneous catalysis: an industrial perspective on best-practice, gaps and needs. Catal Sci Technol 2:2010–2024
221. Stamatakis M, Vlachos DG (2012) Unraveling the complexity of catalytic reactions via kinetic Monte Carlo simulation: current status and frontiers. ACS Catal 2:2648–2663
222. Stamatakis M (2014) Kinetic modelling of heterogeneous catalytic systems. J Phys 27:013001
223. Gusmão GS, Christopher P (2015) A general and robust approach for defining and solving microkinetic catalytic systems. AlChE J 61:188–199
224. de Oliveira LP, Hudebine D, Guillaume D, Verstraete JJ (2016) A review of kinetic modeling methodologies for complex processes. Oil Gas Sci Technol 71:45
225. Reuter K (2016) Ab initio thermodynamics and first-principles microkinetics for surface catalysis. Catal Lett 146:541–563
226. Park GB, Kitsopoulos TN, Borodin D, Golibrzuch K, Neugebohren J, Auerbach DJ, Campbell CT, Wodtke AM (2019) The kinetics of elementary thermal reactions in heterogeneous catalysis. Nat Rev Chem 3:723–732
227. Motagamwala AH, Dumesic JA (2021) Microkinetic modeling: a tool for rational catalyst design. Chem Rev 121:1049–1076
228. Sutton JE, Guo W, Katsoulakis MA, Vlachos DG (2016) Effects of correlated parameters and uncertainty in electronic-structure-based chemical kinetic modelling. Nat Chem 8:331–337
229. Proppe J, Reiher M (2019) Mechanism deduction from noisy chemical reaction networks. J Chem Theory Comput 15:357–370
230. Campbell CT (2017) The degree of rate control: a powerful tool for catalysis research. ACS Catal 7:2770–2779
231. Maffei LP, Pelucchi M, Cavallotti C, Bertolino A, Faravelli T (2021) Master equation lumping for multi-well potential energy surfaces: a bridge between ab initio based rate constant calculations and large kinetic mechanisms. Chem Eng J 422:129954
232. Bligaard T, Bullock RM, Campbell CT, Chen JG, Gates BC, Gorte RJ, Jones CW, Jones WD, Kitchin JR, Scott SL (2016) Toward benchmarking in catalysis science: best practices, challenges, and opportunities. ACS Catal 6:2590–2602
233. Kozuch S, Shaik S (2006) A combined kinetic-quantum mechanical model for assessment of catalytic cycles: application to cross-coupling and heck reactions. J Am Chem Soc 128:3355–3365
234. Kozuch S, Shaik S (2008) Kinetic-quantum chemical model for catalytic cycles: the Haber-Bosch process and the effect of reagent concentration. J Phys Chem A 112:6032–6041
235. Kozuch S, Shaik S (2010) Defining the optimal inductive and steric requirements for a cross-coupling catalyst using the energetic span model. J Mol Catal A 324:120–126
236. Kozuch S, Shaik S (2011) How to conceptualize catalytic cycles? The energetic span model. Acc Chem Res 44:101–110

237. Boudart M (1995) Turnover rates in heterogeneous catalysis. Chem Rev 95:661–666
238. Eyring H (1935) The activated complex in chemical reactions. J Chem Phys 3:107–115
239. Kozuch S (2015) Steady state kinetics of any catalytic network: graph theory, the energy span model, the analogy between catalysis and electrical circuits, and the meaning of mechanism. ACS Catal 5:5242–5255
240. Jones CW (2010) On the stability and recyclability of supported metal-ligand complex catalysts: myths, misconceptions and critical research needs. Top Catal 53:942–952
241. Schuster P (2019) What is special about autocatalysis? Oil Gas Sci Technol 150:763–775
242. Sagués F, Epstein IR (2003) Nonlinear chemical dynamics. Dalton Trans 2003:1201–1217
243. Blackmond DG (2009) An examination of the role of autocatalytic cycles in the chemistry of proposed primordial reactions. Angew Chem Int Ed 48:386–390
244. Weissbuch I, Lahav M (2011) Crystalline architectures as templates of relevance to the origins of homochirality. Chem Rev 111:3236–3267
245. Meyer AJ, Ellefson JW, Ellington AD (2012) Abiotic self-replication. Acc Chem Res 45:2097–2105
246. Vaidya N, Manapat ML, Chen IA, Xulvi-Brunet R, Hayden EJ, Lehman N (2012) Spontaneous network formation among cooperative RNA replicators. Nature 491:72–77
247. Hein JE, Blackmond DG (2012) On the origin of single chirality of amino acids and sugars in biogenesis. Acc Chem Res 45:2045–2054
248. Mondloch JE, Bayram E, Finke RG (2012) A review of the kinetics and mechanisms of formation of supported-nanoparticle heterogeneous catalysts. J Mol Catal A 355:1–38
249. Virgo N, Ikegami T, McGregor S (2016) Complex autocatalysis in simple chemistries. Artif Life 22:138–152
250. Semenov SN, Kraft LJ, Ainla A, Zhao M, Baghbanzadeh M, Campbell VE, Kang K, Fox JM, Whitesides GM (2016) Autocatalytic, bistable, oscillatory networks of biologically relevant organic reactions. Nature 537:656–660
251. Kosikova T, Philp D (2017) Exploring the emergence of complexity using synthetic replicators. Chem Soc Rev 46:7274–7305
252. Eigen M (1971) Selforganization of matter and the evolution of biological macromolecules. Sci Nat 58:465–523
253. Kauffman SA (1986) Autocatalytic sets of proteins. J Theor Biol 119:1–24
254. Steel M (2000) The emergence of a self-catalysing structure in abstract origin-of-life models. Appl Math Lett 13:91–95
255. Hordijk W, Steel M (2004) Detecting autocatalytic, self-sustaining sets in chemical reaction systems. J Theor Biol 227:451–461
256. Sousa FL, Hordijk W, Steel M, Martin WF (2015) Autocatalytic sets in E. Coli metabolism. J Syst Chem 6:4
257. The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47:D506–D515
258. Andersen JL, Flamm C, Merkle D, Stadler PF (2021) Defining autocatalysis in chemical reaction networks. arXiv:2107.03086 [cs, q-bio]
259. Andersen JL, Flamm C, Merkle D, Stadler PF (2019) Chemical transformation motifs—modelling pathways as integer hyperflows. IEEE/ACM Trans Comput Biol Bioinf 16:510–523
260. Bissette AJ, Fletcher SP (2013) Mechanisms of autocatalysis. Angew Chem Int Ed 52:12800–12826
261. Arnold FH (2001) Combinatorial and computational challenges for biocatalyst design. Nature 409:253–257
262. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, Hilvert D, Houk KN, Stoddard BL, De Baker D (2008) Novo computational design of retro-aldol enzymes. Science 319:1387–1391
263. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, St.Clair JL, Gallaher JL, Hilvert D, Gelb MH, Stoddard BL, Houk KN, Michael FE, Baker D (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. Science 329:309–313
264. Hilvert D (2013) Design of protein catalysts. Annu Rev Biochem 82:447–470
265. Kiss G, Çelebi-Ölçüm N, Moretti R, Baker D, Houk KN (2013) Computational enzyme design. Angew Chem Int Ed 52:5700–5725
266. Zastrow ML, Pecoraro VL (2013) Designing functional metalloproteins: from structural to catalytic metal sites. Coord Chem Rev 257:2565–2588
267. Muñoz Robles V, Ortega-Carrasco E, Alonso-Cotchico L, Rodriguez-Guerra J, Lledós A, Maréchal J-D (2015) Toward the computational design of artificial metalloenzymes: from protein-ligand docking to multiscale approaches. ACS Catal 5:2469–2480
268. Zhang L, Lua LHL, Middelberg APJ, Sun Y, Connors NK (2015) Biomolecular engineering of virus-like particles aided by computational chemistry methods. Chem Soc Rev 44:8608–8618
269. Alonso-Cotchico L, Rodríguez-Guerra J, Lledós A, Maréchal J-D (2020) Molecular modeling for artificial metalloenzyme design and optimization. Acc Chem Res 53:896–905
270. Bunzel HA, Anderson JLR, Mulholland AJ (2021) Designing better enzymes: insights from directed evolution. Curr Opin Struct Biol 67:212–218
271. Maldonado AG, Rothenberg G (2010) Predictive modeling in homogeneous catalysis: a tutorial. Chem Soc Rev 39:1891–1902
272. Robbins DW, Hartwig JF (2011) A simple, multidimensional approach to high-throughput discovery of catalytic reactions. Science 333:1423–1427
273. Raugei S, DuBois DL, Rousseau R, Chen S, Ho M-H, Bullock RM, Dupuis M (2015) Toward molecular catalysts by computer. Acc Chem Res 48:248–255
274. Doney AC, Rooks BJ, Lu T, Wheeler SE (2016) Design of organocatalysts for asymmetric propargylations through computational screening. ACS Catal 6:7948–7955
275. Wheeler SE, Seguin TJ, Guan Y, Doney AC (2016) Noncovalent interactions in organocatalysis and the prospect of computational catalyst design. Acc Chem Res 49:1061–1069
276. Poree C, Schoenebeck F (2017) A holy grail in chemistry: computational catalyst design: feasible or fiction? Acc Chem Res 50:605–608
277. Lu Z, Hammond GB, Xu B (2019) Improving homogeneous cationic gold catalysis through a mechanism-based approach. Acc Chem Res 52:1275–1288
278. Foscato M, Jensen VR (2020) Automated in silico design of homogeneous catalysts. ACS Catal 10:2354–2377
279. Rinehart NI, Zahrt AF, Henle JJ, Denmark SE (2021) 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 54:2041–2054
280. dos Passos Gomes G, Pollice R, Aspuru-Guzik A (2021) Navigating through the maze of homogeneous catalyst design with machine learning. Trends Chem 3:96–110
281. Nandy A, Duan C, Taylor MG, Liu F, Steeves AH, Kulik HJ (2021) Computational discovery of transition-metal complexes: from high-throughput screening to machine learning. Chem Rev 121:9927–10000
282. Nørskov JK, Bligaard T, Rossmeisl J, Christensen CH (2009) Towards the computational design of solid catalysts. Nat Chem 1:37–46

283. Greeley J (2016) Theoretical heterogeneous catalysis: scaling relationships and computational catalyst design. Annu Rev Chem Biomol Eng 7:605–635
284. Personick ML, Montemore MM, Kaxiras E, Madix RJ, Biener J, Friend CM (2016) Catalyst design for enhanced sustainability through fundamental surface chemistry. Philos Trans R Soc London Ser A 374:20150077
285. Jimenez-Izal E, Alexandrova AN (2018) Computational design of clusters for catalysis. Annu Rev Phys Chem 69:377–400
286. Zhao C et al (2020) Rational design of layered oxide materials for sodium-ion batteries. Science 370:708–711
287. Wang Y, Hu P, Yang J, Zhu Y-A, Chen D (2021) C-H bond activation in light alkanes: a theoretical perspective. Chem Soc Rev 50:4299–4358
288. Guo C, Fu X, Long J, Li H, Qin G, Cao A, Jing H, Xiao J (2021) Toward computational design of chemical reactions with reaction phase diagram. WIREs Comput Mol Sci 11:e1514
289. Harvey JN, Himo F, Maseras F, Perrin L (2019) Scope and challenge of computational methods for studying mechanism and reactivity in homogeneous catalysis. ACS Catal 9:6803–6813
290. Cordova M, Wodrich MD, Meyer B, Sawatlon B, Corminboeuf C (2020) Data-driven advancement of homogeneous nickel catalyst activity for aryl ether cleavage. ACS Catal 10:7021–7031
291. Chen S, Nielson T, Zalit E, Skjelstad BB, Borough B, Hirschi WJ, Yu S, Balcells D, Ess DH (2021) Automated construction and optimization combined with machine learning to generate Pt(II) methane C-H activation transition states. Top Catal
292. Kirkpatrick P, Ellis C (2004) Chemical space. Nature 432:823–823
293. Reymond J-L (2015) The chemical space project. Acc Chem Res 48:722–730
294. Weymuth T, Reiher M (2014) Inverse quantum chemistry: concepts and strategies for rational compound design. Int J Quantum Chem 114:823–837
295. Zunger A (2018) Inverse design in search of materials with target functionalities. Nat Rev Chem 2:1–16
296. Freeze JG, Kelly HR, Batista VS (2019) Search for catalysts by inverse design: artificial intelligence, mountain climbers, and alchemists. Chem Rev 119:6595–6612
297. Sanchez-Lengeling B, Aspuru-Guzik A (2018) Inverse molecular design using machine learning: generative models for matter engineering. Science 361:360–365
298. von Lilienfeld OA, Müller K-R, Tkatchenko A (2020) Exploring chemical compound space with quantum-based machine learning. Nat Rev Chem 4:347–358
299. Lu Z (2021) Computational discovery of energy materials in the era of big data and machine learning: a critical review. Energy Mater Rep 1:100047
300. Pollice R, dos Passos Gomes G, Aldeghi M, Hickman RJ, Krenn M, Lavigne C, Lindner-D'Addario M, Nigam A, Ser CT, Yao Z, Aspuru-Guzik A (2021) Data-driven strategies for accelerated materials design. Acc Chem Res 54:849–860
301. Weymuth T, Reiher M (2013) Toward an inverse approach for the design of small-molecule fixating catalysts. MRS Online Proc Library 1524:601
302. Weymuth T, Reiher M (2014) Gradient-driven molecule construction: an inverse approach applied to the design of small-molecule fixating catalysts. Int J Quantum Chem 114:838–850
303. Krausbeck F, Sobez J-G, Reiher M (2017) Stabilization of activated fragments by shell-wise construction of an embedding environment. J Comput Chem 38:1023–1038
304. Dittner M, Hartke B (2018) Globally optimal catalytic fields—inverse design of abstract embeddings for maximum reaction rate acceleration. J Chem Theory Comput 14:3547–3564
305. Dittner M, Hartke B (2020) Globally optimal catalytic fields for a Diels-Alder reaction. J Chem Phys 152:114106
306. Behrens DM, Hartke B (2021) Globally optimized molecular embeddings for dynamic reaction solvate shell optimization and active site design. Top Catal. https://doi.org/10.1007/s11244-021-01486-1
307. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A (2018) Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 4:268–276
308. Boitreaud J, Mallet V, Oliver C, Waldispühl J (2020) OptiMol: optimization of binding affinities in chemical space for drug discovery. J Chem Inf Model 60:5658–5666
309. Lim J, Hwang S-Y, Moon S, Kim S, Youn Kim W (2020) Scaffold-based molecular design with a graph generative model. Chem Sci 11:1153–1164
310. Yao Z, Sánchez-Lengeling B, Bobbitt NS, Bucior BJ, Kumar SGH, Collins SP, Burns T, Woo TK, Farha OK, Snurr RQ, Aspuru-Guzik A (2021) Inverse design of nanoporous crystalline reticular materials with deep generative models. Nat Mach Intell 3:76–86
311. Pathak Y, Singh Juneja K, Varma G, Ehara M, Deva Priyakumar U (2020) Deep learning enabled inorganic material generator. Phys Chem Chem Phys 22:26935–26943
312. Kim B, Lee S, Kim J (2020) Inverse design of porous materials using artificial neural networks. Sci Adv 6:eaax9324
313. Nigam A, Pollice R, Aspuru-Guzik A (2021) JANUS: parallel tempered genetic algorithm guided by deep neural networks for inverse molecular design. arXiv:2106.04011 [cs]
314. Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A (2020) Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. Mach Learn 1:045024
315. Nigam A, Pollice R, Krenn M, dos Passos Gomes G, Aspuru-Guzik A (2021) Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES. Chem Sci 12:7079–7090
316. Meyer B, Sawatlon B, Heinen S, von Lilienfeld OA, Corminboeuf C (2018) Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. Chem Sci 9:7069–7077
317. von Rudorff GF, von Lilienfeld OA (2021) Simplifying inverse materials design problems for fixed lattices with alchemical chirality. Sci Adv 7:eabf1173
318. Mayer I (1983) Charge, bond order and valence in the ab initio SCF theory. Chem Phys Lett 97:270–274
319. Sobez J-G, Reiher M (2020) qcscine/molassembler: Release 1.0.0. https://zenodo.org/record/4293555#.YKacWCaxVH4
320. Sobez J-G, Reiher M (2020) Molassembler: molecular graph construction, modification, and conformer generation for inorganic and organic molecules. J Chem Inf Model 60:3884–3900
321. Bannwarth C, Ehlert S, Grimme S (2019) 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 15:1652–1671
322. Bannwarth C, Caldeweyher E, Ehlert S, Hansen A, Pracht P, Seibert J, Spicher S, Grimme S (2021) Extended tight-binding quantum chemistry methods. WIREs Comput Mol Sci 11:e1493
323. Unsleber JP, Grimmel SA, Reiher M. Unpublished
324. Sunoj RB, Anand M (2012) Microsolvated transition state models for improved insight into chemical properties and reaction mechanisms. Phys Chem Chem Phys 14:12715–12736

325. Varghese JJ, Mushrif SH (2019) Origins of complex solvent effects on chemical reactivity and computational tools to investigate them: a review. React Chem Eng 4:165–206
326. Pliego JR, Riveros JM (2020) Hybrid discrete-continuum solvation methods. WIREs Comput Mol Sci 10:e1440
327. Simm GN, Türtscher PL, Reiher M (2020) Systematic microsolvation approach with a cluster-continuum scheme and conformational sampling. J Comput Chem 41:1144–1155
328. Steiner M, Holzknecht T, Schauperl M, Podewitz M (2021) Quantum chemical microsolvation by automated water placement. Molecules 26:1793
329. Bensberg M, Türtscher PL, Unsleber JP, Reiher M, Neugebauer J (2021) Solvation free energies in subsystem density functional theory. arXiv:2108.11228 [cond-mat, physics:physics]
330. Serrano I, López MI, Ferrer Í, Poater A, Parella T, Fontrodona X, Solà M, Llobet A, Rodríguez M, Romero I (2011) New Ru(II) complexes containing oxazoline ligands as epoxidation catalysts. Influence of the substituents on the catalytic performance. Inorg Chem 50:6044–6054
331. Boes JR, Mamun O, Winther K, Bligaard T (2019) Graph theory approach to high-throughput surface adsorption structure generation. J Phys Chem A 123:2281–2285
332. Manz TA (2017) Introducing DDEC6 atomic population analysis: part 3. Comprehensive method to compute bond orders. RSC Adv 7:45552–45581
333. Ertl G, Knözinger H, Schüth F, Weitkamp J (2008) Handbook of heterogeneous catalysis, vol 8. Wiley, New York
334. Montoya JH, Persson KA (2017) A high-throughput framework for determining adsorption energies on solid surfaces. npj Comput Mater 3:1–4
335. Andriuc O, Siron M, Montoya JH, Horton M, Persson KA (2021) Automated adsorption workflow for semiconductor surfaces and the application to zinc telluride. J Chem Inf Model 61:8
336. Deshpande S, Maxson T, Greeley J (2020) Graph theory approach to determine configurations of multidentate and high coverage adsorbates for heterogeneous catalysis. npj Comput Mater 6:1–6
337. Martí C, Blanck S, Staub R, Loehlé S, Michel C, Steinmann SN (2021) DockOnSurf: a python code for the high-throughput screening of flexible molecules adsorbed on surfaces. J Chem Inf Model 61:7
338. Khatib SJ, Oyama ST (2015) Direct oxidation of propylene to propylene oxide with molecular oxygen: a review. Catal Rev Sci Eng 57:306–344
339. Düzenli D, Atmaca DO, Gezer MG, Onal I (2015) A density functional theory study of partial oxidation of propylene on Cu_2 O(001) and CuO(001) surfaces. Appl Surf Sci 355:660–666
340. Porter WN, Lin Z, Chen JG (2021) Experimental and theoretical studies of reaction pathways of direct propylene epoxidation on model catalyst surfaces. Surf Sci Rep. https://doi.org/10.1016/J.SURFREP.2021.100524
341. Proppe J, Reiher M (2017) Reliable estimation of prediction uncertainty for physicochemical property models. J Chem Theory Comput 13:3297–3317
342. Haag MP, Marti KH, Reiher M (2011) Generation of potential energy surfaces in high dimensions and their haptic exploration. ChemPhysChem 12:3204–3213
343. Mühlbach AH, Vaucher AC, Reiher M (2016) Accelerating wave function convergence in interactive quantum chemical reactivity studies. J Chem Theory Comput 12:1228–1235
344. Vaucher AC, Reiher M (2016) Molecular propensity as a driver for explorative reactivity studies. J Chem Inf Model 56:1470–1478
345. Vaucher AC, Reiher M (2018) Minimum energy paths and transition states by curve optimization. J Chem Theory Comput 14:3091–3099
346. Hawkins PC (2017) Conformation generation: the state of the art. J Chem Inf Model 57:1747–1756
347. Ebejer J-P, Morris GM, Deane CM (2012) Freely available conformer generation methods: how good are they? J Chem Inf Model 52:1146–1158
348. Friedrich N-O, de Bruyn Kops C, Flachsenberg F, Sommer K, Rarey M, Kirchmair J (2017) Benchmarking commercial conformer ensemble generators. J Chem Inf Model 57:2719–2728
349. Vitek AK, Jugovic TME, Zimmerman PM (2020) Revealing the strong relationships between ligand conformers and activation barriers: a case study of bisphosphine reductive elimination. ACS Catal 10:7136–7145
350. Viegas LP (2021) Simplified protocol for the calculation of multiconformer transition state theory rate constants applied to tropospheric OH-initiated oxidation reactions. J Phys Chem A 125:4499–4512
351. Leite TB, Gomes D, Miteva M, Chomilier J, Villoutreix B, Tufféry P (2007) Frog: a FRee Online druG 3D conformation generator. Nucleic Acids Res 35:W568–W572
352. Miteva MA, Guyon F, Tufféry P (2010) Frog2: efficient 3D conformation ensemble generator for small compounds. Nucleic Acids Res 38:W622–W627
353. Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT (2010) Conformer generation with OMEGA: algorithm and validation using high quality structures from the protein databank and Cambridge structural database. J Chem Inf Model 50:572–584
354. O'Boyle N, Vandermeersch T, Hutchison G (2011) Confab—generation of diverse low energy conformers. J Cheminformatics 3:P32
355. Poli G, Seidel T, Langer T (2018) Conformational sampling of small molecules with iCon: performance assessment in comparison with OMEGA. Front Chem 6:229
356. Gavane V, Koulgi S, Jani V, Uppuladinne MVN, Sonavane U, Joshi R (2019) TANGO: a high through-put conformation generation and semiempirical method-based optimization tool for ligand molecules. J Comput Chem 40:900–909
357. Friedrich N-O, Flachsenberg F, Meyder A, Sommer K, Kirchmair J, Rarey M (2019) Conformator: a novel method for the generation of conformer ensembles. J Chem Inf Model 59:731–742
358. Vainio MJ, Johnson MS (2007) Generating conformer ensembles using a multiobjective genetic algorithm. J Chem Inf Model 47:2462–2474
359. Riniker S, Landrum GA (2015) Better informed distance geometry: using what we know to improve conformation generation. J Chem Inf Model 55:2562–2574
360. Gebauer NWA, Gastegger M, Schütt KT (2018) Generating equilibrium molecules with deep neural networks. arXiv:1810.11347 [physics, stat]
361. Mansimov E, Mahmood O, Kang S, Cho K (2019) Molecular geometry prediction using a deep generative graph neural network. Sci Rep 9:20381
362. Chan L, Hutchison GR, Morris GM (2019) Bayesian optimization for conformer generation. J Cheminformatics 11:32
363. Chan L, Hutchison GR, Morris GM (2020) BOKEI: Bayesian optimization using knowledge of correlated torsions and expected improvement for conformer generation. Phys Chem Chem Phys 22:5211–5219
364. Gogineni T, Xu Z, Punzalan E, Jiang R, Kammeraad J, Tewari A, Zimmerman P (2020) TorsionNet: a reinforcement learning

approach to sequential conformer search. arXiv:2006.07078 [cs, stat]
365. Simm GNC, Hernández-Lobato JM (2020) A generative model for molecular distance geometry. arXiv:1909.11459 [cs, stat]
366. Fang L, Makkonen E, Todorović M, Rinke P, Chen X (2021) Efficient amino acid conformer search with Bayesian optimization. J Chem Theory Comput 17:1955–1966
367. Ganea O-E, Pattanaik L, Coley CW, Barzilay R, Jensen KF, Green WH, Jaakkola TS (2021) GeoMol: torsional geometric generation of molecular 3D conformer ensembles. arXiv:2106.07802 [physics]
368. Marchand DJJ, Noori M, Roberts A, Rosenberg G, Woods B, Yildiz U, Coons M, Devore D, Margl P (2019) A variable neighbourhood descent heuristic for conformational search using a quantum annealer. Sci Rep 9:13708
369. Abrams C, Bussi G (2014) Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. Entropy 16:163–199
370. Bernardi RC, Melo MCR, Schulten K (2015) Enhanced sampling techniques in molecular dynamics simulations of biological systems. Biochim Biophys Acta Gen Subj 1850:872–877
371. Tiwary P, van de Walle A (2016) Multiscale materials modeling for nanomechanics. Springer series in materials science. Springer International Publishing, Berlin, pp 195–221
372. Yang YI, Shao Q, Zhang J, Yang L, Gao YQ (2019) Enhanced sampling in molecular dynamics. J Chem Phys 151:070902
373. Kamenik AS, Lessel U, Fuchs JE, Fox T, Liedl KR (2018) Peptidic macrocycles—conformational sampling and thermodynamic characterization. J Chem Inf Model 58:982–992
374. Zivanovic S, Bayarri G, Colizzi F, Moreno D, Gelpí JL, Soliva R, Hospital A, Orozco M (2020) Bioactive conformational ensemble server and database. a public framework to speed up in silico drug discovery. J Chem Theory Comput 16:6586–6597
375. Pracht P, Bohle F, Grimme S (2020) Automated exploration of the low-energy chemical space with fast quantum chemical methods. Phys Chem Chem Phys 22:7169–7192
376. Chandramouli B, Galdo SD, Fusè M, Barone V, Mancini G (2019) Two-level stochastic search of low-energy conformers for molecular spectroscopy: implementation and validation of MM and QM models. Phys Chem Chem Phys 21:19921–19934
377. Grimme S, Bohle F, Hansen A, Pracht P, Spicher S, Stahn M (2021) Efficient quantum chemical calculation of structure ensembles and free energies for nonrigid molecules. J Phys Chem A 125:19
378. Senior AW et al (2020) Improved protein structure prediction using potentials from deep learning. Nature 577:706–710
379. Baek M et al (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science 373(6557):871–876
380. O'Connor M, Deeks HM, Dawn E, Metatla O, Roudaut A, Sutton M, Thomas LM, Glowacki BR, Sage R, Tew P, Wonnacott M, Bates P, Mulholland AJ, Glowacki DR (2018) Sampling molecular conformations and dynamics in a multiuser virtual reality framework. Sci Adv 4:eaat2731
381. Schlegel HB (2011) Geometry optimization. WIREs Comput Mol Sci 1:790–809
382. Henkelman G (2017) Atomistic simulations of activated processes in materials. Annu Rev Mater Res 47:199–216
383. Bofill JM, Quapp W (2020) Calculus of variations as a basic tool for modelling of reaction paths and localisation of stationary points on potential energy surfaces. Mol Phys 118:e1667035
384. Banerjee A, Adams N, Simons J, Shepard R (1985) Search for stationary points on surfaces. J Phys Chem 89:52–57
385. Baker J (1986) An algorithm for the location of transition states. J Comput Chem 7:385–395
386. Bofill JM (1994) Updated Hessian matrix and the restricted step method for locating transition structures. J Comput Chem 15:1–11
387. Brunken C, Steiner M, Unsleber JP, Vaucher AC, Weymuth T, Reiher M (2020) qcscine/readuct: Release 2.0.0. https://zenodo.org/record/3768539#.YKabpCaxVH6
388. Fukui K (1970) Formulation of the reaction coordinate. J Phys Chem 74:4161–4163
389. Bosia F, Brunken C, Sobez J-G, Unsleber JP, Reiher M (2020) qcscine/core: Release 3.0.1. https://zenodo.org/record/4293507
390. Bosia F, Brunken C, Grimmel SA Haag MP, Heuer MA, Simm GN, Sobez J-G, Steiner M, Türtscher PL, Unsleber JP, Vaucher AC, Weymuth T, Reiher M (2020) qcscine/utilities: release 3.0.1. https://zenodo.org/record/4293510#.YKKD0aFCRhE
391. Brunken C, Reiher M (2020) Self-parametrizing system-focused atomistic models. J Chem Theory Comput 16:1646–1665
392. Bosia F, Husch T, Vaucher AC, Reiher M (2020) qcscine/sparrow: Release 2.0.1. https://zenodo.org/record/3907313#.YKab3iaxVH4
393. Unsleber JP, Dresselhaus T, Klahr K, Schnieders D, Böckers M, Barton D, Neugebauer J (2018) Serenity: a subsystem quantum chemistry program. J Comput Chem 39:788–798
394. Neese F (2018) Software update: the ORCA program system, version 4.0. WIREs Comput Mol Sci 8:e1327
395. Balasubramani SG et al (2020) TURBOMOLE: modular program suite for ab initio quantum-chemical and condensed-matter simulations. J Chem Phys 152:184107
396. Baiardi A, Reiher M (2020) The density matrix renormalization group in chemistry and molecular physics: recent developments and new challenges. J Chem Phys 152:040903
397. Mühlbach AH, Reiher M (2018) Quantum system partitioning at the single-particle level. J Chem Phys 149:184104
398. Brunken C, Reiher M (2021) Automated construction of quantum-classical hybrid models. J Chem Theory Comput 17(6):3797–3813
399. https://github.com/grimme-lab/xtb. Accessed August 2021; commit for energy calculations was 0245411f5b8595c8ac7655d72c105c055e1da837
400. Perdew JP, Burke K, Wang Y (1996) Generalized gradient approximation for the exchange-correlation hole of a many-electron system. Phys Rev B 54:16533–16539
401. Grimme S, Antony J, Ehrlich S, Krieg H (2010) A consistent and accurate Ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. J Chem Phys 132:154104
402. Grimme S, Ehrlich S, Goerigk L (2011) Effect of the damping function in dispersion corrected density functional theory. J Comput Chem 32:1456–1465
403. Weigend F, Ahlrichs R (2005) 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 7:3297–3305
404. Weigend F (2006) Accurate coulomb-fitting basis sets for H to Rn. Phys Chem Chem Phys 8:1057–1065
405. Lippert G, Hutter J, Parrinello M (1997) A hybrid gaussian and plane wave density functional scheme. Mol Phys 92:477–488

406. Kühne TD et al (2020) CP2K: an electronic structure and molecular dynamics software package—quickstep: efficient and accurate electronic structure calculations. J Chem Phys 152:194103
407. VandeVondele J, Hutter J (2007) Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases. J Chem Phys 127:114105
408. Goedecker S, Teter M, Hutter J (1996) Separable dual-space gaussian pseudopotentials. Phys Rev B 54:1703–1710
409. Tran R, Xu Z, Radhakrishnan B, Winston D, Sun W, Persson KA, Ong SP (2016) Surface energies of elemental crystals. Sci Data 3:60080

Authors and Affiliations

Miguel Steiner[1] · **Markus Reiher**[1]

[1] Laboratory of Physical Chemistry, ETH Zurich, Vladimir-Prelog-Weg 2, 8093 Zurich, Switzerland