DISS ETH NO. 25343

Error-Controlled Quantum Chemical Exploration of Reaction Networks

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
Gregor Nils Christoph Simm

MSc ETH in Interdisciplinary Sciences, ETH Zurich

born on 20.12.1991
citizen of Germany

accepted on the recommendation of

Prof. Dr. Markus Reiher, examiner
Prof. Dr. Gunnar Jeschke, co-examiner

ETH Zurich
Zurich, Switzerland
2018

Gregor Nils Christoph Simm:
Error-Controlled Quantum Chemical Exploration of Reaction Networks
Dissertation ETH Zurich No. 25343, 2018.

Contents

Abstract

For a detailed analysis of a chemical system, all relevant intermediates and elementary reactions on the potential energy surface (PES) connecting them need to be known. An in-depth understanding of all reaction pathways would allow one to study the evolution of a system over time, given a set of initial conditions (e.g., reactants and their concentrations, temperature, and pressure) and propose derivatives of the original reactants to avoid undesired side reactions. Manual explorations of complex reaction mechanisms employing quantum-chemical methods are slow and error-prone. In addition, due to the high dimensionality of PESs exhaustive exploration is generally unfeasible. However, to rationalize, for instance, the formation of undesired side products or decomposition reactions, unexpected reaction pathways need to be uncovered.

In this thesis, we present a computational protocol that constructs reaction networks, consisting of intermediates and transition states, in a fully automated fashion. Starting from a set of initial reagents new intermediates are explored through intra- and intermolecular reactions of already explored ones. This is done by assembling reactive complexes based on heuristic rules derived from conceptual electronic-structure theory and exploring the corresponding approximate reaction path. A subsequent path refinement leads to a minimum-energy path which connects the new intermediate to the existing ones to form a connected reaction network. Tree traversal algorithms are then employed to detect reaction channels and catalytic cycles. We apply our protocol to the formose reaction to study different pathways of sugar formation and to rationalize its autocatalytic nature. Furthermore, we investigate the Schrock dinitrogen-fixation catalyst and discover alternative pathways of catalytic ammonia production.

To be able to draw reliable conclusions from the generated reaction networks, accurate relative energies between intermediates and transition states are required. To date, density functional theory (DFT) is the only method that is computationally feasible for the *ab initio* exploration in this detail. However, DFT often fails to provide sufficiently accurate results, especially for systems containing transition metals. In this thesis, we apply a framework based on Bayesian statistics that allows for error estimation of properties calculated with DFT. Instead of considering only the best-fit parameters of an approximate density functional, we assign a conditional probability distribution to the continuous set of parameters from which a confidence interval can be calculated for any observable. We assess our approach at two challenging chemical systems: catalytic nitrogen fixation and the formose reaction.

Finally, to overcome the lack of systematic improvability of approximate quantum chemical methods we apply Bayesian statistical learning. This new approach allows for the systematic, problem-oriented, and rolling improvement of quantum chemical results through the application of Gaussian processes. Due to its Bayesian nature, reliable error estimates are provided for each prediction. A reference method of high accuracy will be employed to provide a new data point if the uncertainty associated with a particular calculation is above a given threshold. This data point is then added to a growing data set in order to continuously improve the model, and as a result, all subsequent predictions. Previous predictions are validated by the updated model to ensure that uncertainties remain within the given confidence bound, which we call backtracking. We demonstrate our approach with the example of a complex chemical reaction network.

Zusammenfassung

Um die Reaktivität eines chemischen Systems detailliert verstehen zu können, ist eine ausführliche Analyse der Potenzialhyperfläche unabdingbar. Für die vollständige Aufklärung eines chemischen Prozesses muss diese Analyse alle relevanten Intermediate und Übergangszustände enthalten. Damit wäre es möglich (gegeben die Anfangskonzentration aller chemischen Spezies, die Temperatur und den Druck), den zeitlichen Verlauf einer Reaktion vorherzusagen. Zudem wäre man in der Lage, alternative Reaktanden und Reaktionsbedingungen vorzuschlagen, um unerwünschte Nebenreaktionen zu vermeiden. Die manuelle Untersuchung komplexer Reaktionsmechanismen mit Hilfe von quantenchemischen Methoden ist fehleranfällig und sehr zeitaufwendig. Außerdem ist eine vollständige Analyse in der Regel nicht möglich, da die zu untersuchenden Hyperflächen meist hochdimensional sind. Um jedoch zum Beispiel Nebenreaktionen vorhersagen zu können, müssen auch unerwartete Reaktionspfade untersucht werden.

In dieser Doktorarbeit erarbeiten wir ein Verfahren, das die vollautomatische Exploration eines Reaktionsnetzwerks, bestehend aus Intermediaten und Übergangszuständen, erlaubt. Durch intra- und intermolekulare Reaktionen zwischen bereits entdeckten Intermediaten werden neue Intermediate dem Netzwerk hinzugefügt. Mit Hilfe von heuristischen Regeln, welche auf Konzepten der Elektronenstrukturtheorie basieren, werden sogenannte *reaktive Komplexe* erstellt. Wird diesen Komplexen Energie (beispielsweise in Form von kinetischer Energie) zugeführt, können genäherte Reaktionspfade erforscht werden. Die Verfeinerung dieser Pfade führt zu Elementarreaktionen, welche neue Intermediate mit bereits Abgebildeten verbinden, sodass sich ein zusammenhängendes Reaktionsnetzwerk ergibt. Algorithmen zur Analyse von Netzwerken erlauben einem dann, Reaktionskanäle und katalytische Zyklen zu entdecken. Am Beispiel einer präbiotischen Polymerisierungsreaktion, der Formosereaktion, zeigen wir, dass es mit unserem Verfahren möglich ist, auf automatische Weise große Reaktionsnetzwerke zu erstellen. Insbesondere werden auch autokatalytische Eigenschaften dieser Reaktion reproduziert. Zudem untersuchen wir die Reaktivität eines Stickstoff fixierenden Katalysators, dem Yandulov–Schrock-Katalysators, und finden dabei Nebenreaktionen, die seine niedrige Wechselzahl erklären.

Genaue (freie) Energien sind nötig, um zuverlässige Schlüsse aus der Analyse von Reaktionsnetzwerken ziehen zu können. Momentan ist Dichtefunktionaltheorie (DFT) die einzige Elektronenstrukturmethode, die eine Untersuchung solch detaillierter Netzwerke erlaubt. Mittels DFT ist es jedoch nicht immer möglich, genaue thermodynamische Grö-

ßen zu berechnen, insbesondere bei Systemen, die Übergangsmetalle enthalten. Deshalb entwickeln wir in dieser Dissertation einen Ansatz, der auf bayesscher Statistik beruht und für Observablen, die mit DFT berechnet wurden, eine Fehlerabschätzung ermöglicht. Typischerweise werden für empirische Parameter in approximativen Dichtefunktionalen lediglich vordefinierte optimierte Werte verwendet. Im Gegensatz dazu wird in unserem statistischen Ansatz eine Wahrscheinlichkeitsverteilung für die Parameter bestimmt, mit der Vertrauensintervalle für beliebige Observablen ermittelt werden können. Wir wenden die neu entwickelte Methode auf die Formosereaktion und den Katalysezyklus des Yandulov–Schrock-Katalysators an.

Im letzten Kapitel dieser Dissertation nutzen wir bayessches statistisches Lernen, um approximative quantenchemische Methoden systematisch zu verbessern. Mit Hilfe von Gaussprozessen können nun zuverlässige Abschätzungen zu statistischen Unsicherheiten von berechneten Resultaten gemacht werden. Ist diese Unsicherheit zu groß, wird eine hochgenaue Referenzberechnung durchgeführt. Das Resultat dieser Berechnung wird einem wachsenden Referenzdatensatz hinzugefügt, um Vorhersagen des Gaussprozesses zu verbessern. Anschließend werden bisherige Vorhersagen überprüft, um sicherzustellen, dass die Unsicherheiten innerhalb des gegebenen Konfidenzintervalls liegen. Wir demonstrieren die Nützlichkeit dieses Ansatzes zur Fehlerabschätzung an einem komplexen Reaktionsnetzwerk.

1 Introduction

Complex reaction mechanisms are ubiquitous in chemistry. They are, for instance, the basis of transition-metal catalysis,[1] polymerizations,[2] cell metabolism,[3] flames, and environmental processes[4] and are the objective of systems chemistry.[5] Knowing all chemical compounds and elementary reactions of a specific chemical process is essential for its understanding in atomistic detail. Even though many chemical reactions result in the selective formation of one main product,[6] in general, multiple reaction paths compete with each other leading to a variety of side products. In such cases, a reactive species (such as a radical, a valence-unsaturated species, a charged particle, a strong acid, or base) can be involved or high-energy states (e.g., vibrational states) are populated (due to a high reaction temperature, for example).

For a detailed analysis of a chemical system, all relevant intermediates and elementary reactions connecting them need to be known. Given a set of initial conditions (e.g., reactants and their concentrations, temperature, and pressure), a detailed understanding of all reaction pathways would allow one to study the evolution of a system over time and propose derivatives of the original reactants to avoid undesired side reactions. Manual searches for all components of complex reaction mechanisms employing quantum chemical methods are slow, tedious, and error-prone. In addition, due to the high dimensionality of PESs, exhaustive explorations are generally unfeasible. However, to rationalize, for instance, the formation of undesired side products or decomposition reactions, alternative reaction pathways need to be uncovered. Therefore, it is desirable to develop a fully automated protocol for an efficient and accurate exploration of configuration spaces involving both minimum-energy structures and transition states (TSs).

Clearly, it would be too much to demand from this protocol to be universally applicable and, therefore, we focus on thermal reactions in the condensed and gas-phase.

Access to accurate thermodynamic properties (e.g., standard Gibbs free energy) of all intermediates and elementary reactions is mandatory for a reliable description of a chemical process. Quantum electrodynamics (QED) allows for the description of all electromagnetic processes occurring between the elementary particles of chemical systems (e.g., molecules).[7] It is the fundamental theory of chemistry (focusing on the dominant electromagnetic interactions and ignoring the other fundamental forces). If we were able to solve its equations for chemical systems with arbitrary accuracy, truly predictive results would be obtained. However, for all but the simplest systems, calculations based on QED are unfeasible. As a result, quantum chemical methods employed to describe chemical reactions rely on a number of approximations. However, the effect of such approximations on observables derived from them is often unpredictable. Therefore, it is challenging to quantify the uncertainty of a computational result, which, however, is necessary to assess the suitability of a computational model. Moreover, in practice, multiple approximations are made for the calculation of an observable of interest so that they are available in reasonable time and with reasonable effort. Eventually, the number and types of approximations necessary for a feasible description of molecular systems are vast and diverse such that it is difficult to attribute errors to certain approximations. In addition, the precise effect of such approximations (computational models) on observables derived from them is generally unknown and difficult to estimate for arbitrary molecules,[8] let alone entire reaction networks consisting of a multitude of intermediates and transition states.

While the procedure of uncertainty quantification for physical measurements is well established,[9] this is not the case for results of computational models (virtual measurements[10]). By the very nature of a deterministic (or fully converged stochastic) calculation, the repetition of such a calculation does not lead to an oscillation around the true result (if the calculation is fully reproducible, as it should be) and, therefore, there is no obvious approach of reliably estimating prediction uncertainty of the computational model employed. However, the result of a computational model is incomplete without an accurate uncertainty associated with it.[10] Given a reliable uncertainty measure for a computational result, one could not only estimate the effects on observables derived from that result (through uncertainty propagation) but also directly assess the quality of approximations in the model development stage. Finally, availability of prediction uncertainties would help select an appropriate computational model of sufficient accuracy for a problem at hand.

The challenges mentioned above are elaborated on in this thesis which is organized as follows: In Chapter 2, an overview is given over the plethora of existing approaches

for the exploration of chemical reaction networks. In addition, the limitations of current methods are discussed. These limitations are addressed in a new computational protocol which we present in Chapter 3. In contrast to existing approaches, this protocol constructs complex chemical reaction networks in a fully automated fashion and is applicable to any molecular system. Throughout this thesis, we demonstrate our developments with two challenging chemical systems: the oligomerization reaction of formaldehyde[11–13] and catalytic nitrogen fixation under ambient conditions with the Yandulov–Schrock catalyst.[14,15] The former reaction results in a highly complex mixture of linear and branched compounds, the latter features a plethora of possible intermediates leading to a very low turnover number. We employ our new approach to study these chemical systems by constructing reaction networks of unprecedented depth. In Chapter 4, we discuss problems and solutions for performance assessment of computational models based on several examples from the quantum chemistry literature. For this purpose, we elucidate the different sources of uncertainty, the elimination of systematic errors, and the combination of individual uncertainty components to the uncertainty of a prediction. To obtain reliable uncertainty predictions for the exploration of chemical reaction networks, we introduce Bayesian statistics for system-focused DFT in Chapter 5. Two case studies then demonstrate how important reliable error estimates are for meaningful conclusions drawn from quantum chemical results. In Chapter 6, we address the lack of systematic improvability of approximate density functionals and the limitations of our previous approach by applying Gaussian process (GP) regression. We present an algorithm that allows for the on-the-fly construction of a reference data set adapted to the system to be explored and the required confidence level. The thesis concludes with a summary of the advances that were achieved and an outlook on future work.

2

Exploration of Reaction Paths and Chemical Transformation Networks

The construction of a reaction network containing all relevant intermediates and elementary reactions is necessary for the accurate description of chemical processes. In the case of a complex chemical reaction (involving, for instance, many reactants or highly reactive species), the size of such network may grow rapidly. Manual search for intermediates and TSs is not feasible in these cases. Therefore, there is a need for efficient and reliable methods that require minimal human intervention or intuition. In this Chapter, we review existing approaches for the effective exploration of complex PESs. In Section 2.1, two classes of exploration strategies are introduced. Current implementations of these strategies are discussed in Sections 2.2 and 2.3.

2.1 Strategies for the Exploration of Complex PESs

Some excellent reviews on the exploration of reaction paths exist (for recent one see Ref. 16). This overview has a different focus that also extends the literature covered in previous work. We group the plethora of strategies developed for the exploration of PESs in two classes as illustrated in Fig. 2.1:

Strategy 1: Starting from a minimum energy or approximate TS structure (indicated by the green region in Fig. 2.1, left) local curvature information is exploited to climb up a PES towards TSs, and then, by following the minimum-energy path (MEP), towards new intermediates (indicated by blue regions). This process is repeated for

all minimum-energy structures (possibly in multiple directions) until all extrema of the PES (below some energy cutoff) are explored.

Strategy 2: Starting from a minimum energy structure, new intermediates are explored through the application of heuristics (guided by chemical intuition, indicated by dashed lines in the Fig. 2.1, right). This includes, for example, the formulation of graph-based transformation rules or the application of an artificial force pushing reactive moieties together. Once a new intermediate is found, the MEP connecting it to the starting structure is searched for.

In practice, many modern exploration strategies lie in between these idealized classes or employ a combination of the two. In the following, an overview of current implementations of these strategies is given.

For the accurate description of a chemical process, not only bond-breaking and bond-forming transformations need to be considered but also the conformational space of each intermediate needs to be explored. This is particularly critical for an accurate description of thermodynamic properties (e.g., Gibbs free energy) and catalytic reactions in which multiple MEPs connecting the same configurational isomers can exist. In general, both exploration strategies mentioned above can be employed to locate conformational isomers. It should be noted, however, that in contrast to bond-breaking or bond-forming transformations, in most practical applications, TSs between conformers are not of interest as the timescale of a reaction can be assumed to be longer than the time the conformers require for equilibration. For a recent review on conformer generation see Ref. 17.

In the following, successful examples of the two strategies are detailed. Overall, the approaches differ in the degree of automation possible (an aspect which is particularly critical for large reaction networks), the amount of heuristics required, and the thoroughness of the exploration. Approaches involving little or no heuristics tend to explore the PES in a more systematic fashion, however, they are often limited by computational effort, and hence, are often applicable to only small chemical systems.

2.2 Exploration through Exploitation of PES Curvature

In many theoretical studies, approximate TS candidates (obtained, for example, through manual construction) are refined by utilizing Hessian information – the second-order derivatives of the potential energy with respect to the nuclear coordinates. The Hessian of a TS guess structure is required to obtain the vibrational mode representing the reaction coordinate to follow. Eigenvector following (EVF)[18–24] is a prominent example of such an approach, that will be extremely reliable if the TS guess structure is close to the true TS. For large molecules, a full Hessian calculation will become computationally

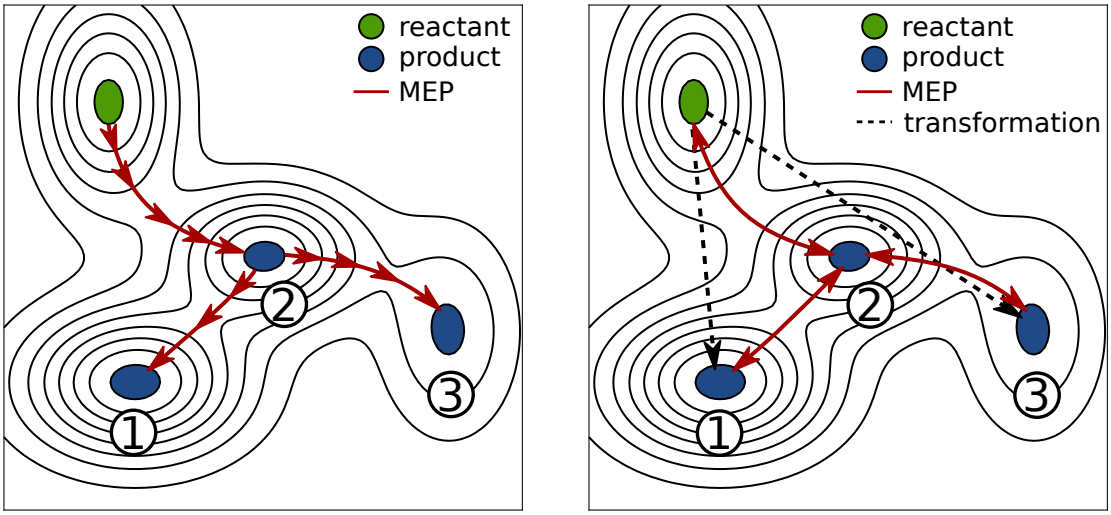

Figure 2.1: Two general strategies for the exploration of PESs. Left: local curvature information of the PES is exploited to identify TSs and products. Right: through the application of heuristics new intermediates are identified. Starting from two intermediates an MEP is searched for.

demanding or even unfeasible even if the Hessian is approximated. Therefore, several algorithms have been developed to circumvent the calculation of the full Hessian. A quasi-Newton–Raphson method was introduced by Broyden,[25] in which an approximate Hessian is built from gradients only and then updated by the gradients of intermediate points obtained during the optimization. Other approaches include schemes proposed by Munro and Wales[26] that avoid the full diagonalization of the Hessian, Lanczos subspace iteration methods,[27] and Davidson subspace iteration algorithms.[28–30]

Nevertheless, the manual construction of approximate TS candidates is slow, cumbersome, and highly unsystematic. Recent developments focused on providing more streamlined and systematic means for the location of TSs. Maeda and coworkers exploit curvature information of the PES in a strategy called anharmonic downward distortion following (ADDF).[31–35] By distorting a minimum structure orthogonal to the potential energy contours, MEPs can be found. As a result, a path explored through ADDF is very close to the one obtained from an intrinsic reaction coordinate (IRC) calculation. Repeated application of ADDF on newly explored intermediates yields all relevant reaction paths. ADDF was successfully applied to small systems such as formaldehyde, propyne, and formic acid reactions.

To address the limitations of ADDF, Maeda and coworkers developed the artificial force-induced reaction (AFIR) method.[36–41] AFIR overcomes intermolecular activation energies by applying an artificial force pushing the reactants together, hence distorting the original PES. When optimizing under this biasing force, the maximum energy point lies close to the true TS. AFIR has been successfully applied to the Claisen rearrange-

ment,[38,39] Biginelli reactions,[42] and cobalt-catalyzed hydroformylations.[43] A downside of AFIR is that many random initial orientations of the two reactants need to be sampled before all relevant reactions paths are found. In addition, human input is often required to select pairs of reacting molecules to circumvent a combinatorial explosion.

Martínez-Núñez and coworkers developed an approach termed TS search using chemical dynamics simulations (TSSCDS),[44–47] in which high-energy dynamics employing semi-empirical quantum chemical methods are performed to induce reactions to occur at high rates. Vibrational modes are populated to increase the rate at which TSs can be overcome. For large systems, due to their large number of vibrational modes, manual intervention is required to steer the simulation in directions of interest. The trajectory generated through the simulation is subsequently post-processed, and bond-forming and bond-breaking events are identified. TS guesses are extracted from the trajectory and refined employing semi-empirical and density-functional methods. The TSSCDS method was successfully applied to reactions involving formaldehyde, formic acid,[45] vinyl cyanide,[44] and cobalt catalysis.[47]

In reactive molecular dynamics (MD) simulations, the nuclear equations of motion are solved to explore and sample the part of configuration space that is accessible under the constraints imposed by a predefined thermodynamic ensemble. The capability of reactive *ab initio* molecular dynamics for studying complex chemical reactions was shown with the example of the prebiotic Urey–Miller experiment.[48–50] As the configuration space can become very large, comprising multiple copies of all chemical species involved in the reaction, computational costs of carrying out first-principles calculations grow rapidly. This issue can be overcome by the application of a reactive force-field.[51,52] Unfortunately, next to the reduced accuracy, force-field parameters will, in general, not be available for any type of system which limits their applicability. Therefore, hybrid quantum-mechanical–molecular-mechanical approaches have been frequently applied to explore different reaction paths of complex systems with many degrees of freedom such as enzymatic reactions (for examples see Refs. 53–58 and reviews by Senn and Thiel[59–61]). However, to increase the possibility of a reaction to occur, the temperature and pressure of the simulation need to be increased which in turn leads to the frequent occurrence of unphysical transformations.

Naturally, MD simulations employing classical and *ab initio* force fields can be applied to sample conformational degrees of freedom. Recent progress in this field has been reviewed in Refs. 62 and 63. These approaches are among the most complex and time-consuming for conformational sampling.[64] Stochastic methods based on Monte Carlo-simulated annealing (MC) are often faster than MD methods.[65–67] By sampling low-lying eigenmodes they require less computational effort than MD simulations. Both MD and MC approaches are computationally too expensive for a high-throughput setting.

Recently, Satoh et al.[68] systematically explored conformational transitions of D-glucose by employing ADDF to trace only low TS barriers.

2.3 Locating Minimum-Energy Paths by Connecting Intermediates

Despite being highly systematic, exploration strategies solely based on *ab initio* curvature information of the PES are often unsuitable for large chemical systems with many degrees of freedom. Starting from a minimum structure, it can be more effective to apply heuristics to rapidly identify potential products, and subsequently, search for an MEP connecting them. If both endpoints of an elementary reaction are known, interpolation methods (see e.g., Ref. 69) and string methods[70–78] can be applied to locate the MEP connecting them.

Conceptual knowledge of chemistry can be applied to rapidly identify potential candidates for intermediates connected to the starting structure through an elementary reaction, in particular, if the types of reaction mechanisms relevant to the study at hand are known. For example, from reaction databases or chemical heuristics, transformation rules can be formulated and applied to graph representations of the reacting molecules. These rules originate from concepts of bond order and valence and, therefore, these approaches are popular in organic chemistry. In 1994, Broadbelt and coworkers pioneered this approach with a method called Netgen.[79] The three-dimensional arrangement of atoms in molecules is transformed into a graph structure in which atoms and bonds are represented by nodes and edges, respectively. This gives rise to adjacency matrices which can be manipulated by matrix operations representing chemical transformations.[80,81] With the derived adjacency matrices, new three-dimensional arrangements of atoms are generated. Through repeated application of these transformation rules, new molecules are added to the list of intermediates involved in the global mechanism. In Broadbelt's original work, elementary steps are not identified, and hence, activation barriers are crudely estimated with the Evans–Polanyi principle.[82] One shortcoming of Netgen is that many intermediates are proposed that can only be reached through high-energy paths.

In a similar spirit, Green and coworkers developed a software package called Reaction Mechanism Generator (RMG).[83,84] Multiple, significant steps were taken to overcome challenges demonstrated by the approach of Broadbelt. In particular, kinetic parameters were estimated employing quantum chemical calculations to discard products that are likely to be reachable only by overcoming TSs featuring high activation barriers. RMG was employed to automatically map the mechanisms of the pyrolysis of *n*-butanol[85] and methane.[83] The Green and West groups have shown that RMG can be applied

to a variety of complex systems with good success.[86–91] Similar to Netgen, RMG is ultimately limited by the application of concepts of bond order and valence.

Aspuru-Guzik and coworkers developed a methodology based on formal bond orders[92,93] to model prebiotic reactions such as the *formose* reaction.[11] Instead of specifically encoding elementary reactions, transformation rules are based on a concept popular in organic chemistry, commonly denoted as "arrow pushing".[94] Similar to the approach above, activation barriers are estimated (in Ref. 92, employing Hammond's postulate). While the exhaustiveness of the exploration is encouraging, the resulting reaction networks may contain intermediates that can be considered not viable (e.g., three-membered rings).

Recently, Kim and coworkers utilized chemical heuristics to rapidly search reaction paths.[95] Through the application of molecular graphs and reaction network analyses, they explored a so-called minimal reaction network consisting of intermediates that can be reached from the starting structures within a fixed number of bond dissociation and formation reactions. The minimal network is subjected to quantum chemical calculations to determine kinetically the most favorable reaction path. They applied their method to recover the accepted mechanisms of the Claisen ester condensation and of cobalt-catalyzed hydroformylation reactions.[95]

The ZStruct approach developed by Zimmerman and coworkers utilizes connectivity graphs to identify potential intermediates that could form when connections are formed or broken.[96,97] Intermediates are then subjected to a double-ended reaction path search employing the growing string method.[78,98–100] A limitation of ZStruct is the requirement that reactants are prealigned, which restricted this approach to intramolecular reactions. In addition, if two intermediates where connected by two elementary reactions, many will struggle to find a TS. Despite its shortcomings, ZStruct uncovered an unexpected side-reaction that was hampering a Ni-based C–H functionalization catalyst.[101] Furthermore, several other (catalytic) reactions have been studied with ZStruct.[102–104] Recently, Zimmerman addressed the limitations of ZStruct in ZStruct2.[105] In ZStruct2, reactants are prealigned to sample so-called driving coordinates that describe the expected elementary reactions. ZStruct2 has been successfully applied to study transition metal catalysts.[106–109]

Green and coworkers developed a graph-based approach to find reaction pathways.[110,111] Very recently, they explored the reaction network of the simplest γ-ketohydroperoxide, 3-hydroperoxypropanal, by applying the Berny algorithm[112–114] coupled with the freezing string method,[77] single- and double-ended growing string methods, and the AFIR method.

In the approach of Habershon,[115,116] connectivity graphs are employed to describe intermediates. Reaction pathways are examined by dynamics simulations over a Hamil-

tonian that can be updated to suit a change in the connectivity graph. The trajectories are processed and unique reaction pathways are refined. Compared to the approach developed by Green and coworkers, Habershon's method explores the potential energy surface more extensively, at the cost of running dynamics simulations.

There exist several efficient approaches for the exploration of conformational intermediates that are related to this class of exploration strategy. Distance geometry (DG) methods stochastically generate sets of atomic coordinates which are refined against a set of interatomic distance constraints. Generated conformers are usually optimized employing a molecular mechanics force field or quantum-chemical methods to afford a candidate conformer. Implementations of DG can be found in DG-AMMOS[117] and RDKit.[118] To reduce the conformational space that needs to be explored, the so-called rigid-rotor approximation is often introduced in which bond lengths and bond angles are kept fixed so that only torsional degrees of freedom are sampled. Genetic algorithms are a prominent class of methods for stochastic sampling of vast torsion space. A popular implementation of this method is Balloon_GA.[119] Aside from genetic algorithms, Monte Carlo methods have been used for the stochastic sampling of torsional angles.[120,121] A general problem with stochastic methods is the possibility of missing relevant (i.e., accessible) intermediates. As a result, the amount of sampling required is *a priori* not known. Systematic conformer generation methods, which rely explicitly on the rigid-rotor approximation, attempt to enumerate all possible torsional degrees of freedom of a molecule. Systematic enumeration of all possible torsion angles based on a starting conformation in a brute-force fashion will result in a combinatorial explosion of candidate conformers. Rule-based conformation generators limit the conformational space they explore. These rules are usually derived from analyses of torsional angles in solid-state structures found in databases such as the Protein Database[122] or the Cambridge Structural Database.[123,124]

3

Automated Exploration of Chemical Reaction Networks*

The many different exploration strategies elucidated in Chapter 2 were successfully applied to a plethora of chemical systems. However, they all suffer from limitations. Approaches exploiting local curvature information of the PES are highly systematic but limited by computational effort, and hence, applicable to only small chemical systems (see Section 2.2). In contrast, current approaches making use of heuristics (described in Section 2.3) employ graph-based transformation rules to discover potential intermediates. While being computationally efficient, they rely on the concept of valence which may perform well for many organic molecules, and thus, will fail for systems containing species with complex electronic structures such as transition-metal clusters. Furthermore, to ensure an exhaustive exploration with such an approach, completeness of the set of transformation rules is required. However, for an arbitrary, unknown chemical system this cannot be guaranteed if heuristics rules need to be formulated. One will then be restricted to known chemical transformations, which may hamper the discovery of new chemical processes. In this Chapter, we present a new computational protocol that constructs complex chemical reaction networks in a fully automated fashion. It addresses the limitations of the many approaches reviewed in Chapter 2 and is implemented our protocol in a software called CHEMOTON (named after a theory for the functioning of living systems proposed by Gánti[125]). In Sections 3.7 and 3.8, the func-

* This Chapter is reproduced in part with permission from M. Bergeler, G. N. Simm, J. Proppe, M. Reiher, *J. Chem. Theory Comput.* **2015**, *11*, 5712–5722 and G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119. Copyright 2017 American Chemical Society.

tionality of Chemoton is illustrated with the example of two challenging chemical systems. To illustrate the functionality of our machinery, we apply Chemoton to two chemical systems featuring a different set of challenges: the formose reaction, which is an oligomerization reaction of formaldehyde,[11–13] and the Yandulov–Schrock catalyst, a nitrogen-fixating molybdenum complex developed by Schrock and co-workers.[14,15]

3.1 Exploration Strategy of Chemoton

Starting from a set of initial conditions, our exploration protocol is applied repeatedly to expand a reaction network in a rolling fashion. These conditions comprise the reactants and their concentrations, solvents, and standard thermodynamic ensemble parameters such as temperature and pressure. In addition, the timescale of the reaction is relevant as it allows one to define the slowest reaction which still affects the concentration of all species in a significant manner (for details see Refs. 126 and 127). Reactions slower than that one can be safely discarded, whereas reactions which are much faster can be considered to be in quasi-equilibrium.[127]

Unlike reactive molecular-dynamics simulations, a concept-driven exploration does not take place on a single PES but on multiple low-dimensional PESs consisting of rather few nuclear coordinates that can represent specific elementary reactions as will be discussed in Section 3.3. Exploring many low-dimensional PESs instead of one of very high dimension bears several advantages. Calculations are, in general, faster due to the reduced number of atoms. Geometry optimizations and TS searches converge more quickly and the exploration can be more easily steered into regions of interest.

The exploration strategy in Chemoton belongs to the second class of exploration strategies (see Section 2.3), but (in contrast to previous approaches) is designed to be applicable to arbitrary molecular systems.

3.2 Generation of Conformers

A PES is explored starting from one minimum-energy structure which may be a molecule or a cluster of molecules (such as a microsolvated solute). Usually, there exist many minima which can be reached from this minimum through a series of elementary reactions featuring a sufficiently low reaction barrier. Often, these minima will turn out to be conformers of the same molecular configuration. With the introduction of some electronic-structure measure for molecular bonds, these minima can be explored very efficiently employing conformer generators.[118–121,128] We determine the molecular bonding by calculating Mayer bond orders from an electronic wave function.[129] The geometries of the generated conformers are subsequently optimized with quantum chemical methods to obtain sufficiently reliable minimum structures.

If the timescale of a reaction can be assumed to be longer than the time the conformers require for equilibration, TSs between the conformers do not need to be optimized, which reduces the computational effort significantly. In this case, at a given temperature, only a fraction of the conformers can be assumed to be significantly populated, and hence, only this fraction needs to be considered in the subsequent steps of the exploration protocol. Otherwise, TSs need to be located (see Section 3.5).

3.3 Assembly of Reactive Complexes and Induction of Reactions

Searching for minima that can only be reached by overcoming a non-negligible barrier is not straightforward, as, in general, this requires a chemical transformation (i.e., breaking or forming bonds). In the following steps of our protocol, we distinguish between *intermolecular* and *intramolecular* reactions.

To exhaustively explore the intermolecular reaction between two intermediates (i.e., all possible products and their reaction paths), the following steps are carried out. Firstly, to explore the reaction between any pair of atoms from the intermediates they need to be positioned relative to one another with the aim of obtaining a *reactive complex* (see Fig. 3.1). Here, their relative orientation (three rotational degrees of freedom) must be considered. This is particularly important for reactions in which non-covalent bonding is important or where no single pair of reacting atoms can be defined (as, for example, in a Diels-Alder reaction). Note also that our restriction to two intermediates does not exclude reactions with a molecularity higher than two as an intermediate may consist of more than one molecule, which is also important when considering microsolvated structures.

For an exhaustive exploration, reactive complexes must then be generated for every pair of atoms. However, it is obvious that this is, in general, not feasible (see Section 3.4 for a viable solution to this problem). Finally, through a constrained optimization along a shrinking distance between the reacting atoms, an approximate reaction path is constructed. A full geometry optimization is carried out afterward so that either new species are formed or the reactants are recovered, which is automatically detected. For an intramolecular reaction, the minimum structure acts as a starting point for the constrained optimization. The relative orientation of the reacting atoms in the intermediate will determine the reaction path and product. For both, intra- and intermolecular reactions, the conformational diversity of the intermediates needs to be taken into account to ensure the exploration of all reaction paths and products.

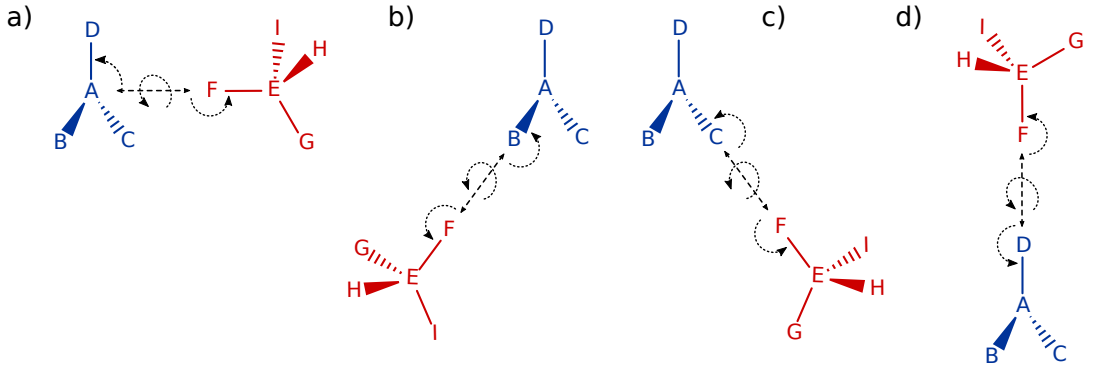

Figure 3.1: Assembly of reactive complexes between two intermediates colored blue and red. Three rotational degrees of freedom are indicated by curly arrows. For clarity, reactive complexes constructed from pairs containing atoms E, G, H, and I are omitted. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119. Copyright 2017 American Chemical Society.

3.4 Identification of Reactive Atoms

Fig. 3.1 implies that it is, in general, unfeasible to consider the reaction between every pair of atoms of two intermediates under full orientational freedom as the complete pairing would lead to a myriad of reactions most of which potentially featuring high reaction barriers. Due to the exponential growth of the number of possible reactions highly systematic exploration algorithms such as ADDF[35] reach their limits of feasibility. Therefore, a descriptor is required that allows one to identify pairs of atomic centers that, when brought together in close proximity, are likely to react. At the same time, the choice of descriptor must not compromise the exhaustiveness of the exploration, that is it must not confine the exploration to known, expected reaction paths. Hence, the descriptor should be based on fundamental physical quantities evaluated in a quantum mechanical framework such as the electron density.

When considering the reactivity of spatially extended reactants, descriptors are appropriate that are based on first principles such as the electron localization function (ELF) by Becke and Edgecombe,[130] the Laplacian of the electron density[131] (see also Ref. 132), Fukui functions,[133] partial atomic charges,[134–137] atomic polarizabilities,[138–140] or dual descriptors[141–144] (see also Refs. 145–147 for reviews). However, all of them suffer from the limitation that it is difficult to assess the height of reaction barriers from information at the reactants' minimum structures, for which they are evaluated. Nonetheless, it is usually a very fruitful assumption in chemistry that the minimum structure holds some information on the system's reactivity, which is reflected in the considerable success of chemical concepts and of expert systems applied to synthesis planning.[148–157]

In this work, we pursue an approach that combines basic chemical knowledge and physical principles (such as attraction of oppositely charged residues) with information

extracted from quantum mechanical quantities. When formulating such heuristic rules one faces a trade-off between efficiency and transferability. Our descriptors then determine the location of *reactive sites* situated around atoms (depicted as discs in Fig. 3.2). To restrict the exploration to reactions that are likely to feature surmountable reaction barriers under the reaction conditions given, only pairs of atoms with reactive sites of opposite reactivity are considered. In the case of an intermolecular reaction, reactants are oriented so that reactive sites are facing each other (see Fig. 3.2). Clearly, these restrictions can be easily lifted in our algorithm to guarantee a successively expanding exploration. This is very much in the spirit of a rolling exploration of reaction networks that also allows for changing reaction conditions.

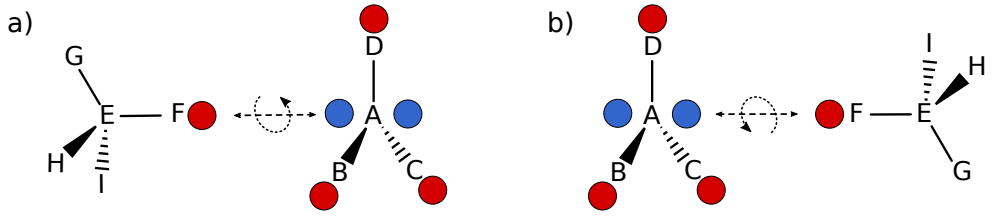

Figure 3.2: Assembly of reactive complexes after identification of reactive sites. Atoms A and F are arranged so that reactive sites of opposite reactivity (discs colored blue and red) are facing each other. The rotational degrees of freedom of the reactive complexes are indicated by curly arrows. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119. Copyright 2017 American Chemical Society.

3.5 Exploration of Minimum-Energy Paths and Transition States

From the reactive complex and the reaction product, a minimum-energy path connecting them is to be found. Double-ended TS search methods such as nudged elastic band[72,73,158] and string methods[75,77,78,99,100,159] are efficient in suggesting an initial guess for a TS. A single-ended search method, such as EVF,[18,19,21–23,30,160] is then employed to optimize the TS candidate so that a stationary point with exactly one negative eigenvalue of the Hessian matrix is found. The corresponding eigenvector is followed in the forward and backward directions (by a steepest-descent method) to connect to two local minima.

Employing the Mayer bond-order criterion, minimum-energy structures consisting of more than one molecule are split into separate molecules. Here, the charge to be assigned to each molecule is determined by calculating the atomic partial charges in the minimum-energy structure. Finally, through the application of the bond-order criterion, it is determined whether the molecules have been encountered before in the exploration (Fig. 3.3 illustrates the entire protocol).

This protocol is applied repeatedly until no new structures are explored or the exploration reaches some specified bound (e.g., determined by a maximum molar mass for an

intermediate). In an advanced setup, this bound is given by thermodynamic ensemble parameters such as temperature and pressure. Then, a kinetic simulation would allow one to identify intermediates which are not significantly populated and to exclude those from subsequent steps of the exploration.

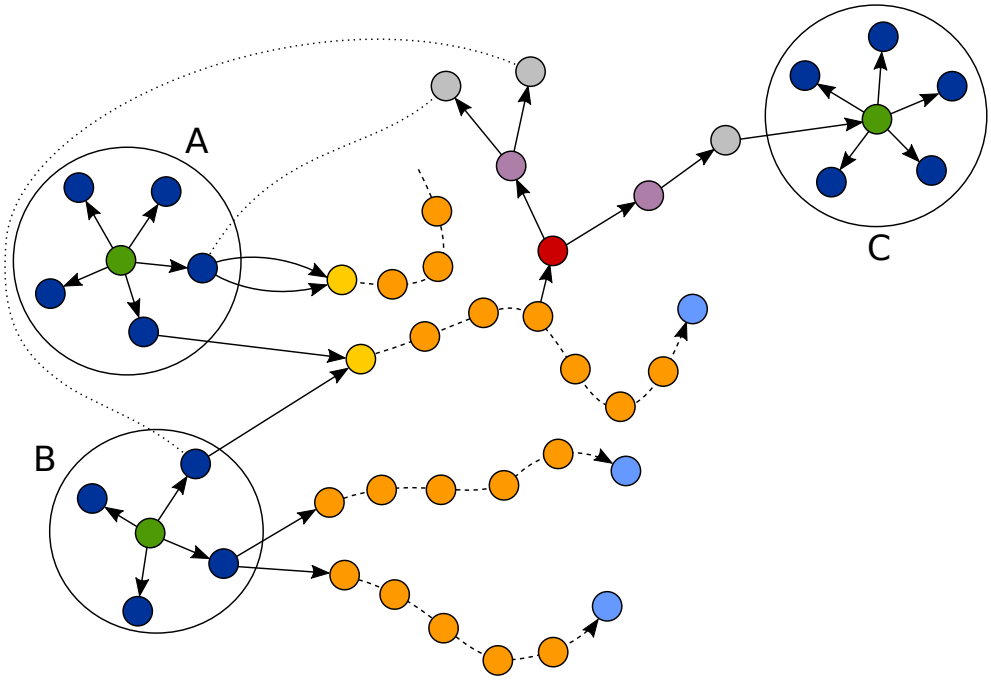

Figure 3.3: Illustration of the exploration protocol. Nodes (discs) represent molecular structures. Conformers of the same configurational isomer (A, B, and C) are enclosed in a circle. Conformers are generated (dark blue) from an initial conformer (green). When two conformers react a reactive complex (yellow) is formed. For both, inter- and intramolecular reactions, an approximate reaction path (orange nodes, dashed line) is explored. The last point of the approximate reaction path is optimized to yield a reaction product (light blue). If the product is different from the reactants, a TS (red) will be searched for. An IRC calculation is performed to obtain the two ends of the minimum-energy path (purple). Minimum energy structures are split into individual molecular structures (gray). Finally, it is determined whether these gray structures are new (solid arrow) or whether they are part of the existing network (dotted line). Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119. Copyright 2017 American Chemical Society.

3.6 Construction of Reaction Networks

From the explored intermediates and TSs a focused network consisting of nodes (representing molecular configurations) and edges (representing reaction channels) needs to be constructed. This step is critical for the understanding of the underlying chemical processes and for carrying out additional analyses such as molecular property calculations and kinetic studies.

In Fig. 3.4, the compression of the raw network to a compact and accessible format

is illustrated. Of the raw network (shown in Fig. 3.3) the molecular configurations (labeled A, B, and C) including their conformers (blue nodes), the TSs (red nodes), the minimum structures of the minimum-energy paths (purple nodes), and the molecular structures they consist of (gray nodes) are of interest (see Fig. 3.4, top). There usually exist multiple reaction paths with different barrier heights for the same chemical transformation (in Fig. 3.4, $A + B \rightleftarrows C$). This multitude of reaction paths arises from the consideration of conformational degrees of freedom of the reactants (see Section 3.2) and the rotational degrees of freedom of the reactive complexes (see Section 3.3). This situation is shown in Fig. 3.4 (top) by three reaction paths.

In Fig. 3.4 (bottom), a compact representation of the raw network is given. Molecular configurations A and B are placed in a *virtual flask* (diamond, left in Fig. 3.4) and react to form a different virtual flask (diamond, right in Fig. 3.4) which consists of one molecular configuration C. The thickness of the arrow between two virtual flasks is proportional to the effective rate constant of the reaction. The calculation of the effective rate constant from multiple reaction paths is not straightforward and will be discussed in detail in a forthcoming study (see also our recent work in Ref. 127). In this study, the thickness of the arrow between two virtual flasks is determined from the height of the lowest activation barrier of the reaction paths: high barriers are represented by thin arrows, low barriers by thick arrows.

3.7 Application to Formose Reaction

The *formose* reaction is a well-studied prebiotic oligomerization reaction of formaldehyde resulting in a highly complex mixture of linear and branched compounds, including monosaccharides.[11–13] The identification of all products poses a major experimental challenge and the exact composition has not been elucidated yet, although over 50 products have already been characterized.[161,162] While some major reaction pathways are known,[163,164] many mechanistic details are not.[165] Due to the formation of biologically important monosaccharides, the formose reaction may constitute a plausible prebiotic source of sugars. The first step in this reaction is the dimerization of formaldehyde to glycolaldehyde which is extremely slow. It was shown that pure formaldehyde in water is unreactive and that small amounts of some contamination are required to initiate the reaction.[166] For example, by addition of glycolaldehyde to the reaction mixture, the formation of glycolaldehyde and higher sugars is greatly accelerated, suggesting an autocatalytic mechanism.[164]

To explore the formose reaction, we applied our exploration protocol described in Section 3.1 to an initial state consisting of formaldehyde, glycolaldehyde, and water. Since the formose reaction results in an intractable polymeric mixture, we restricted the

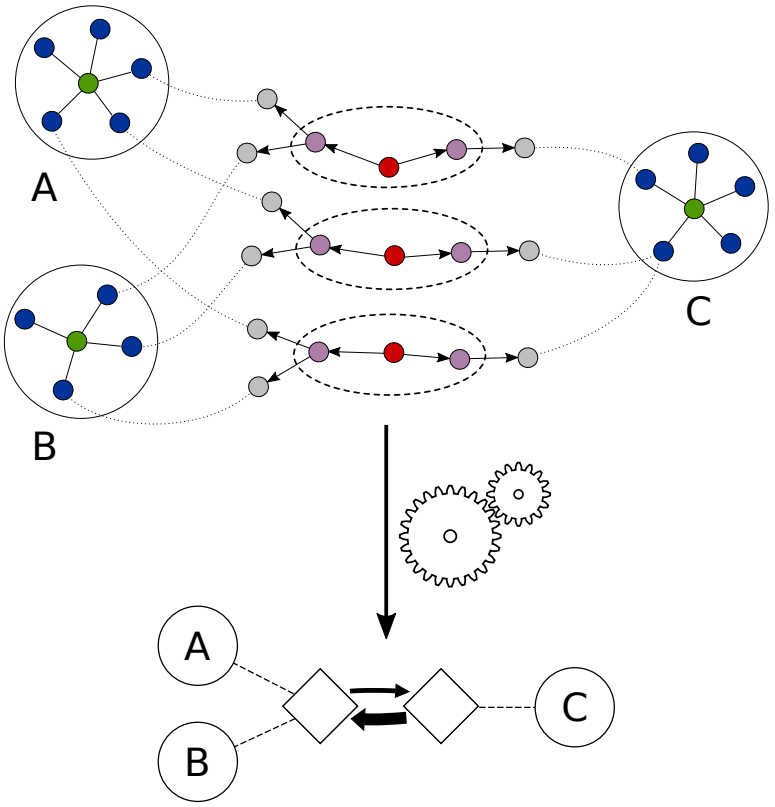

Figure 3.4: Construction of a reaction network by reduction of the raw exploration network. *Top*: molecular configurations (circles) A and B react to form molecular configuration C. Three minimum-energy paths (dashed ovals) are shown, each consisting of a TS (red node) and two minimum-energy structures (purple nodes). The minimum-energy structures are split into their molecular structures (gray nodes) which correspond to conformers (blue nodes) of the molecular configurations. *Bottom*: molecular configurations A and B are placed in a virtual flask (diamond, left), react and form a different virtual flask (diamond, right) consisting of a molecular configuration C. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119. Copyright 2017 American Chemical Society.

exploration to a volume in chemical space that does not exceed the chemical formula of tetrose, i.e., $C_4H_8O_4$.

Employing RDKit,[118,167] conformers were generated for each molecular configuration (see Section 3.2) according to the protocol described in Ref. 168. To reduce the number of quantum chemical calculations, only the most stable conformer was considered in the subsequent steps of the exploration.

From Mayer bond orders extracted from the electronic wave function[129], we constructed a molecular graph consisting of atoms (vertices) connected by bonds (edges). Based on arguments of electronegativity, we considered heteroatoms (i.e., oxygen in this system) to be electron-rich. Hydrogen atoms were considered to have the opposite reactivity if they were found next to a heteroatom within the distance of three edges. Carbon atoms were considered to feature the reactivity of both unless they were

a neighbor of a heteroatom in which case they were automatically labeled electron-poor. We found that these simple rules work well for the system of organic reactions under consideration here (see below). For future work, it will be interesting to compare a multitude of descriptors (various concepts evaluated with the electronic wave function such as partial charges, hardness and softness, electronegativity, the dual descriptor and so forth) in order to assess their general reliability and transferability for other types of reaction networks, involving also transition metals.

To form a reactive complex, two reactants were positioned so that atom i of one reactant with a reactive site of i and atom j of the other reactant with a reactive site of j formed one axis. Reactive sites of an atom i were located on a sphere centered on i with a radius equal to the van der Waals radius of i by maximizing the distance to all neighboring atoms. Two additional reactive complexes were generated by rotating one reactant around this axis by 120° and 240° (see Fig. 3.2). In principle, the value and number of angles as a means for orientational screening can, however, be chosen freely.

Two molecular configurations were compared by finding the maximum common subgraph (MCS) of their graph representations and stereochemical information was considered. The MCS was determined by `RDKit`.[167]

The exploration comprised 82990 geometry optimizations, 23690 constrained PES scans, 7657 freezing-string, 13675 EVF, and 10458 IRC calculations. Details on the computational methodology are provided in Appendix A.1.1. Note that the TS guess obtained from the freezing-string calculation may contain more than one imaginary frequency. As a result, the number of EVF calculations is larger than the number of freezing-string calculations. In total, 934 unique molecular configurations were identified and 6871 minimum-energy paths connecting them were explored. The reaction network comprising all structures is given in the supporting information of Ref. 169.

3.7.1 Reaction Network

As described in Section 3.6, the raw exploration network was processed to generate a reaction network. In Fig. 3.5, the resulting network is shown. In this network, reactions with barriers above 50 kJ/mol are omitted. The nodes representing the starting materials formaldehyde and glycolaldehyde are colored light blue. Water is not shown explicitly, but virtual flasks (diamonds) containing at least one water molecule are colored dark blue. If not stated otherwise the fill color of disc-shaped nodes indicates the number of carbon atoms in the corresponding molecule.

It can be seen that starting from formaldehyde, glycolaldehyde, and water two trioses (purple nodes) and three tetroses (orange nodes) can be formed through multiple cascades of reactions. In addition, multiple three- to six-membered rings can be identified and even a seven-membered ring is among the products. It can also be seen that most

of the polymerization reactions are irreversible which is in accord with experimental findings.[11,163]

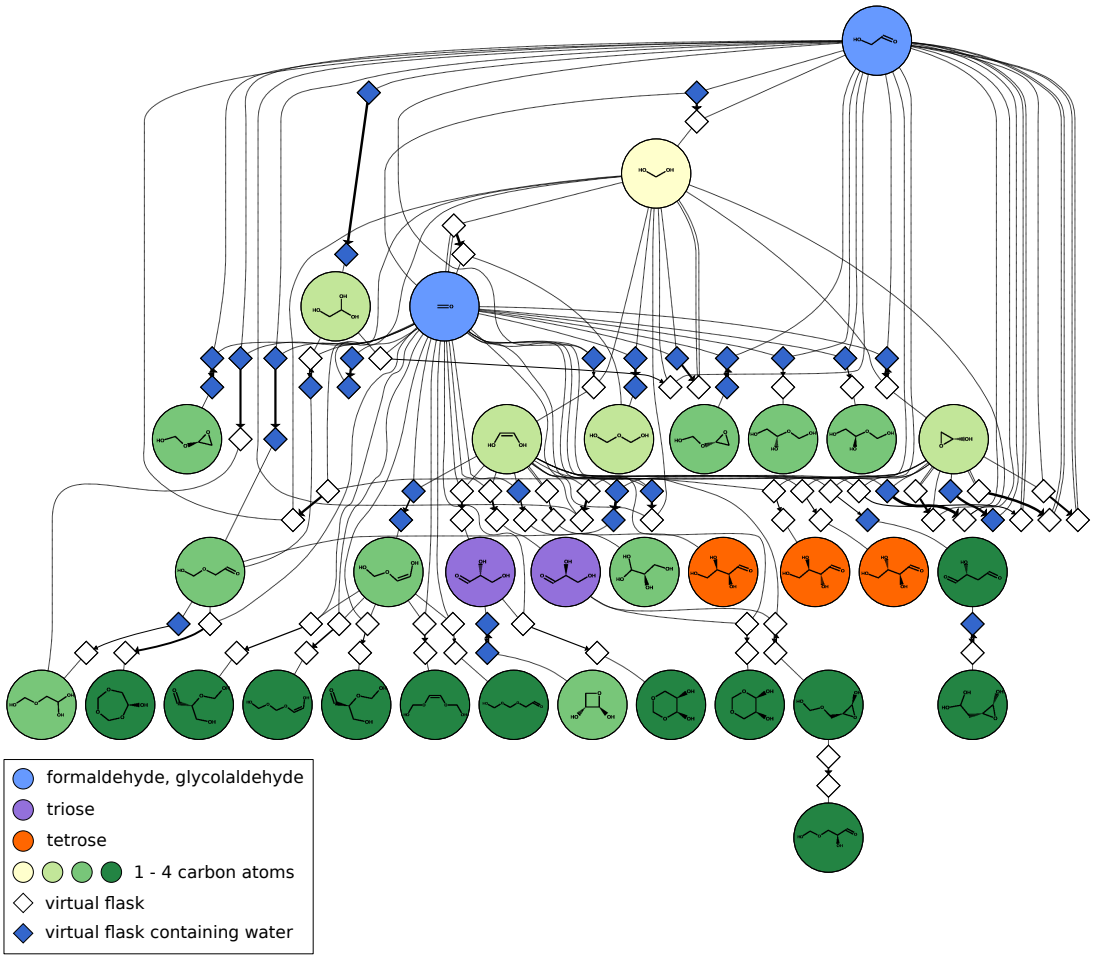

Figure 3.5: Reaction network generated from formaldehyde, glycolaldehyde, and water (the last not shown explicitly) consisting of reactions with activation barriers below 50 kJ/mol. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119. Copyright 2017 American Chemical Society.

Furthermore, one can observe that for some intermediates (e.g., (2R)-oxiran-2-ol) the corresponding enantiomer is missing in the network. However, given achiral starting materials, the product should be racemic. This bias can be explained by the selection of only one conformer for each molecular configuration which was considered in intra- and intermolecular reactions. This issue can be easily resolved by considering sufficiently many conformers for each molecular configuration.

In Fig. 3.6, the reaction network was further expanded to reactions with barriers between 50 and 85 kJ/mol. It can be clearly seen that the reaction network becomes increasingly complex when considering higher activation barriers. It can also be seen that there are multiple different pathways to form a species and it is not obvious from

the barrier heights which of the pathways are the most dominant. Kinetic simulations and additional data analysis are therefore necessary to reach a deeper understanding of the underlying reaction process.

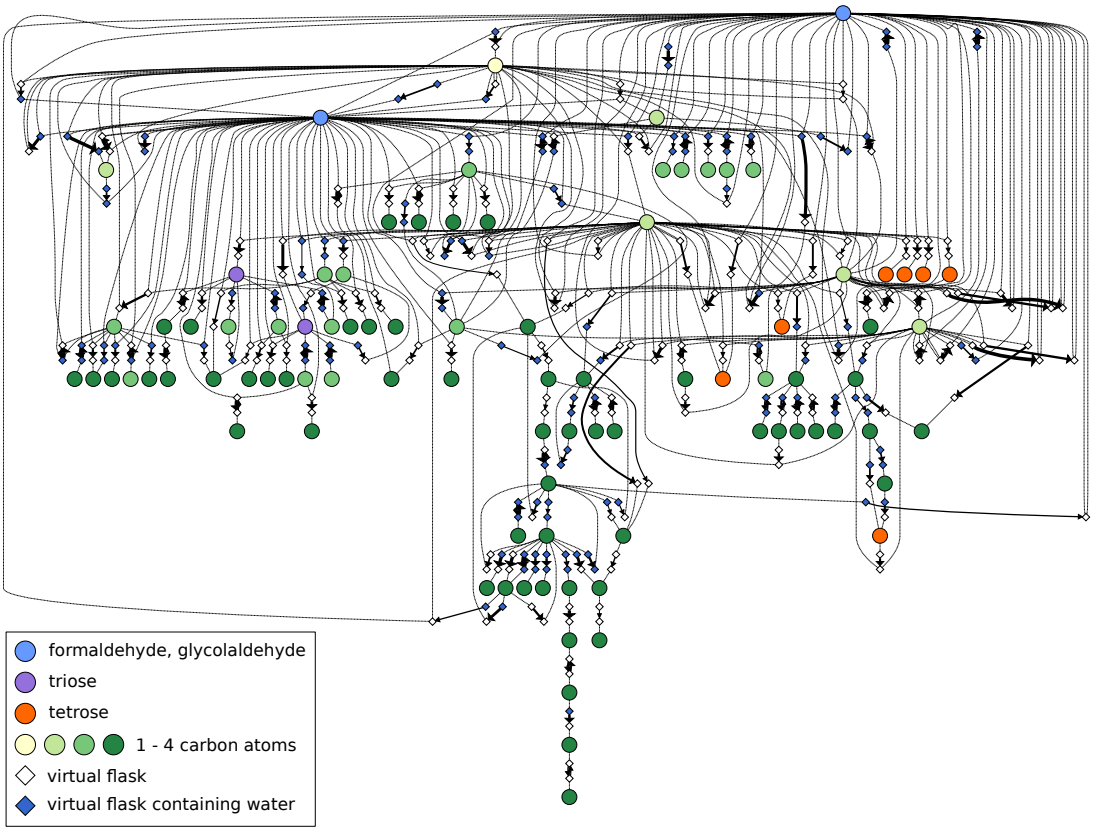

Figure 3.6: Reaction network generated from formaldehyde, glycolaldehyde, and water (the last not shown explicitly) consisting of reactions with activation barriers below 85 kJ/mol. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119. Copyright 2017 American Chemical Society.

Finally, to assess the extensiveness of our protocol we compared our reaction network with a much smaller one obtained from a manual exploration.[170] Within the bounds preset for the present exploration, each intermediate and reaction path identified in a limited manual exploration by Kua et al.[170] can be found in our reaction network.

3.7.2 Alternative Reaction Paths

Through the consideration of conformational diversity and orientational degrees of freedom in the assembly of reactive complexes, our exploration protocol aims to explore all potential reaction paths between two intermediates. The multitude of reaction paths is discovered through the assembly of multiple reactive complexes (as sketched in Fig. 3.2). The following example shall demonstrate the importance of a thorough exploration of reaction paths.

In Fig. 3.7, a selection of minimum-energy paths (from the raw exploration network) for the reaction between ethene-1,2-diol and formaldehyde forming 2-(hydroxymethoxy)ethen-1-ol is shown. The barrier heights of the paths in the forward direction range from 80.8 to 125.9 kJ/mol (neglecting solvation effects). In conventional TS theory,[171] a difference of ≈ 45 kJ/mol in the barrier height results in a reaction rate that is different by a factor on the order of 10^8 at room temperature (assuming that the difference in electronic energy solely determines the free energy difference). Therefore, for kinetic analyses the exploration of all reaction paths is crucial. It can also be seen that despite the small number of atoms involved in this reaction, the structures of the TSs differ significantly. For example, in Fig. 3.7 a), the linear arrangement prevents the stabilizing interactions present in the cyclic TS shown in Fig. 3.7 d). It is the explicit consideration of rotational degrees of freedom when constructing the reactive complexes that leads to the uncovering of these reaction paths.

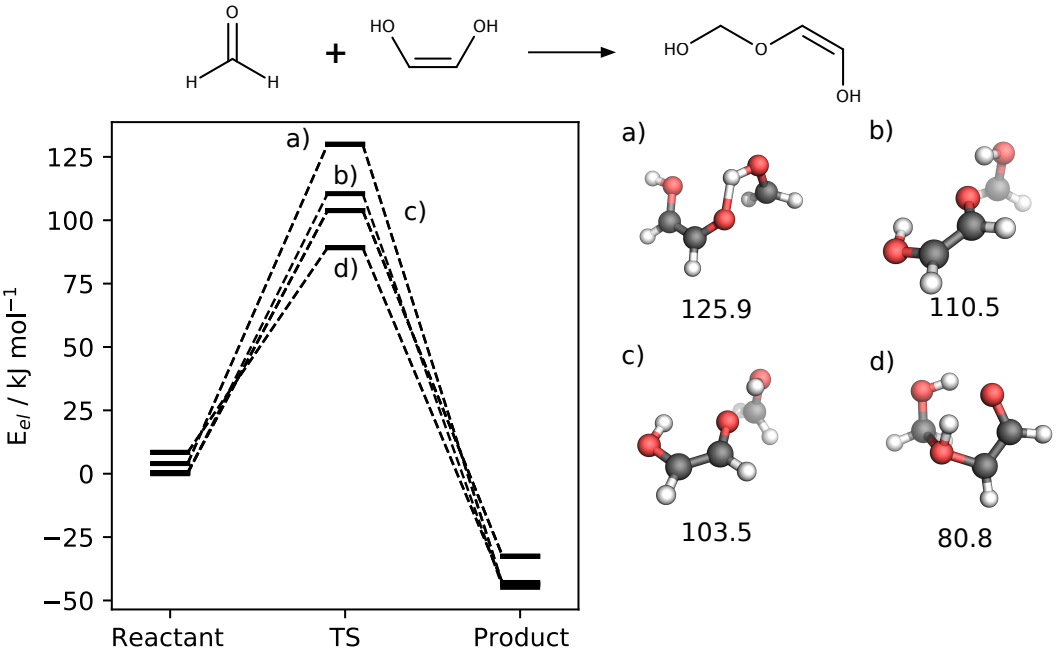

Figure 3.7: *Left*: reaction profiles of four minimum-energy paths for the reaction between ethene-1,2-diol and formaldehyde forming 2-(hydroxymethoxy)ethen-1-ol. *Right*: molecular structures of the TSs. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119. Copyright 2017 American Chemical Society.

In the following, the effect of microsolvation on the exploration of reaction paths is investigated. In Fig. 3.8, seven paths from the exploration network are shown. The chemical transformation is the same as in the previous example but this time in the presence of one water molecule.

It can be seen that compared to Fig. 3.7, both the number of different reaction

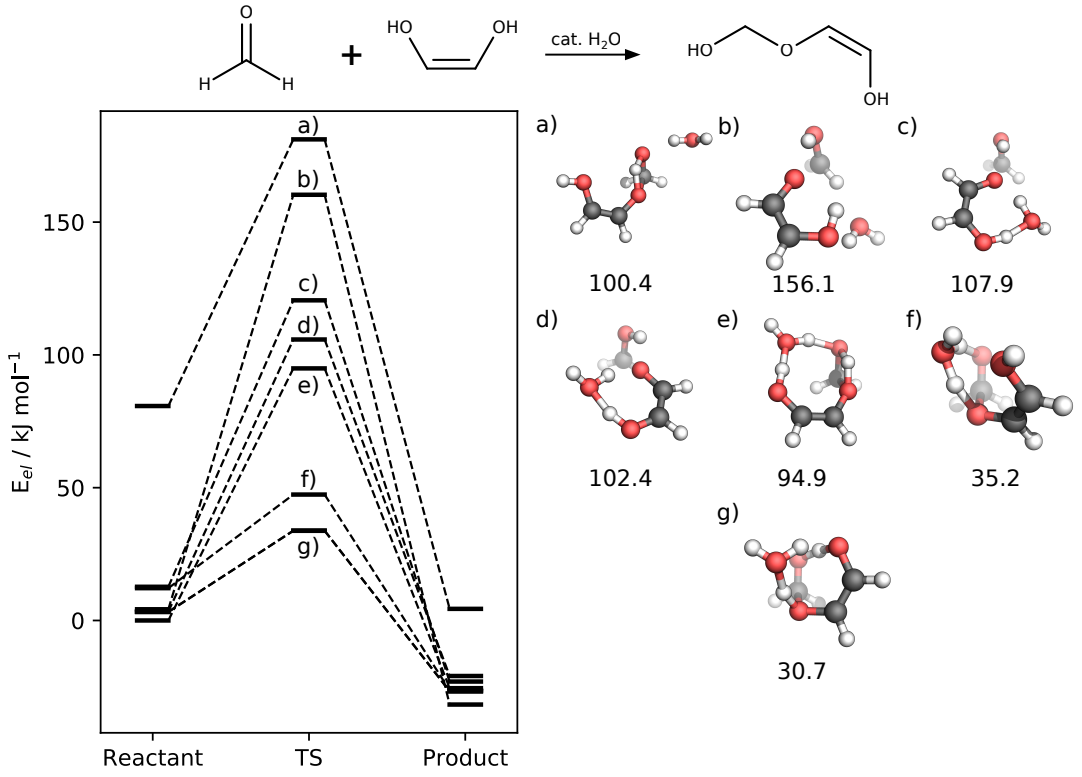

Figure 3.8: *Left*: reaction profiles of seven minimum-energy paths for the reaction between ethene-1,2-diol and formaldehyde forming 2-(hydroxymethoxy)ethen-1-ol catalyzed by a water molecule. *Right*: molecular structures of the TSs. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119. Copyright 2017 American Chemical Society.

paths and the spread of barrier heights (ranging from 30.7 to 156.1 kJ/mol) increased. Moreover, Fig. 3.7, e) shows that our exploration protocol is clearly capable of finding reaction paths that involve more than two reacting atoms, although all reactive complexes started from the pairing-of-atoms concept. The plethora of possible transition paths due to the added degrees of freedom of the solvent molecules renders explorations very challenging. While the application of a continuum model may suffice for unreactive, apolar solvents such as hexane,[172] for polar solvents exhibiting directional bonding such as water which may actively participate in the reaction through hydrogen bonding and transfer (as can be seen in Fig. 3.8, f) this is not a viable solution. A hybrid approach in which microsolvated solutes are embedded into a continuum model is a convenient compromise as long as explicit sampling by molecular dynamics or Monte Carlo methods can be avoided for certain parts of the network.

3.7.3 Graph Analysis of Reaction Network

To study the process of sugar formation in the formose reaction we applied a tree traversal algorithm to the reaction network to find all paths that start from glycolaldehyde and lead to the naturally occurring tetrose D-erythrose. To take into account the low concentration of all products at the beginning of the reaction, we took paths only into consideration if in all elementary reactions there was not more than one reactant that was not a starting material. In addition, edges representing reactions with barriers above 250 kJ/mol were removed.

We were able to identify 40 distinct paths comprising up to five elementary reactions. Fig. 3.9 shows a subnetwork in which each molecular configuration and reaction is present in at least one of these paths. The elementary reactions of the path with the lowest barrier heights are indicated by the numbers 1 to 5. With an activation barrier of 190 kJ/mol, the third reaction of this path features the highest barrier.

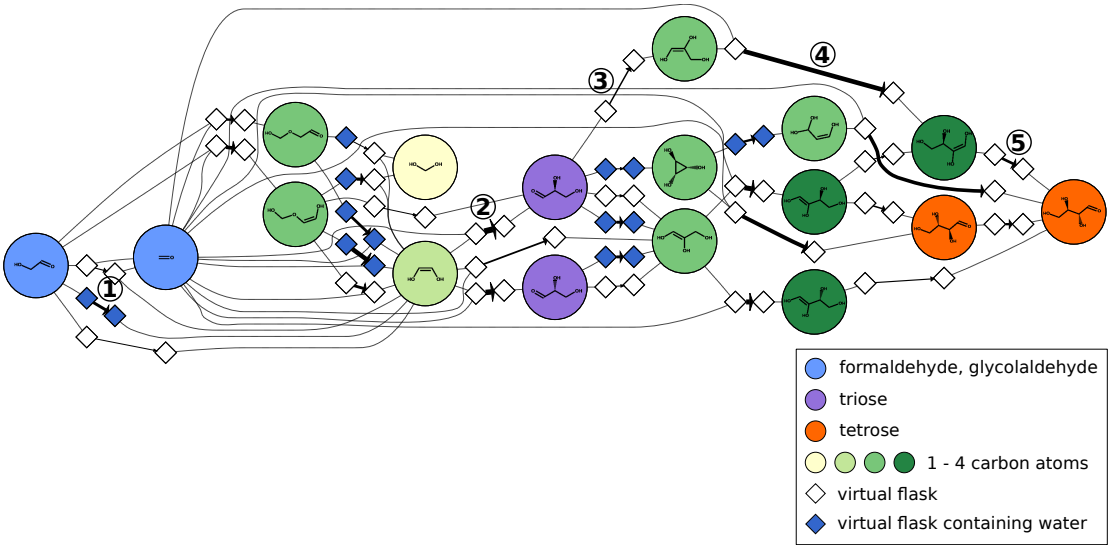

Figure 3.9: Reaction pathways starting from glycolaldehyde (far left node) leading to the formation of d-erythrose (far right node). Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119. Copyright 2017 American Chemical Society.

Employing tree traversal algorithms we also searched for autocatalytic processes in the reaction network. Fig. 3.10 shows a subnetwork consisting of ten cycles (colored solid lines) consisting of up to four elementary reactions with the lowest reaction barriers starting from ethene-1,2-diol that lead to the formation of glycolaldehyde through the consecutive addition of formaldehyde. Ethene-1,2-diol can readily be formed from the starting material via an enolization reaction (see Fig. 3.5). For clarity, formaldehyde is not shown in this network.

The cycle with the lowest reaction barriers (dark blue path) is depicted in Fig. 3.11

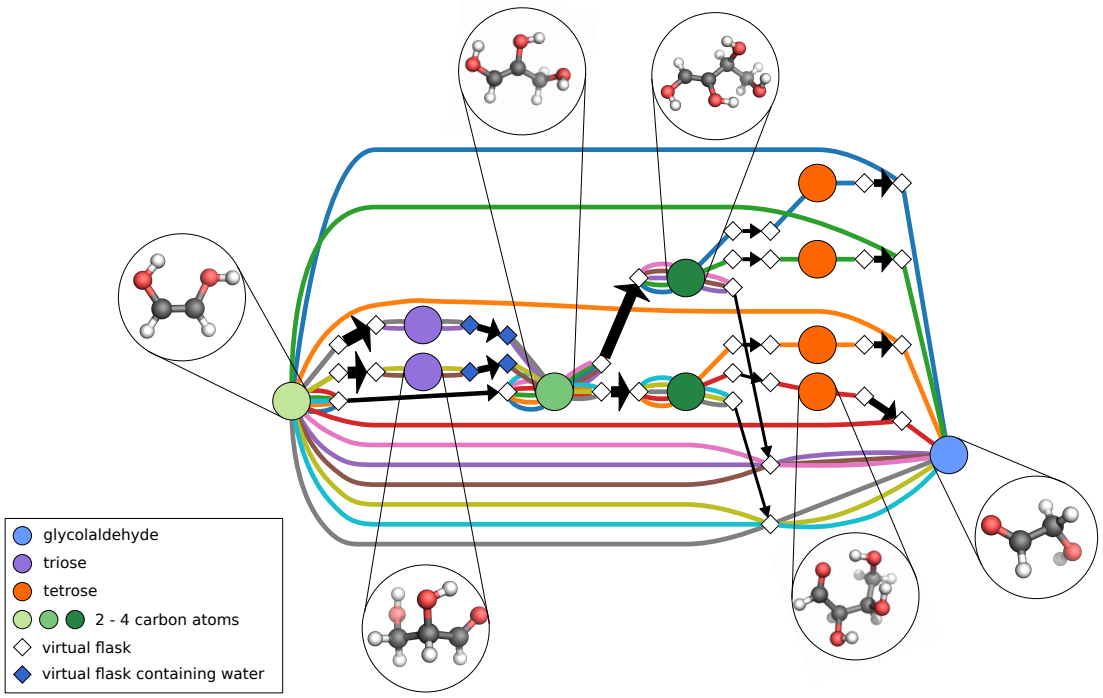

Figure 3.10: Autocatalytic cycles starting from ethene-1,2-diol (far left node) leading to the formation of glycolaldehyde (far right node). For clarity, formaldehyde is not shown. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119. Copyright 2017 American Chemical Society.

and involves the formation of prop-1-ene-1,2,3-triol from ethene-1,2-diol and formaldehyde, followed by the formation of a compound consisting of four carbons (dark green node). A subsequent enolization reaction yields a tetrose which undergoes a fragmentation reaction in which glycolaldehyde is produced and ethene-1,2-diol is recovered. It can also be seen that both trioses and all four tetroses are formed in multiple autocatalytic cycles.

3.8 Application to Yandulov–Schrock Catalyst

In this Section, we apply Chemoton to an important and still not sufficiently well-understood problem in chemistry: catalytic nitrogen fixation under ambient conditions in the homogeneous phase. Specifically, we investigate the chemical reactivity of the molybdenum complex developed by Schrock and co-workers (shown in Fig. 3.12).[14,15,173] This catalyst, like all others developed for this purpose,[174–176] is plagued by a very low turnover number. By applying our protocol to a simplified model system, we aim to better understand the low efficiency of the catalyst. One of the main challenges of this system is the adequate choice of a reactivity descriptor to effectively tackle the combinatorial exploration of possible intermediates.

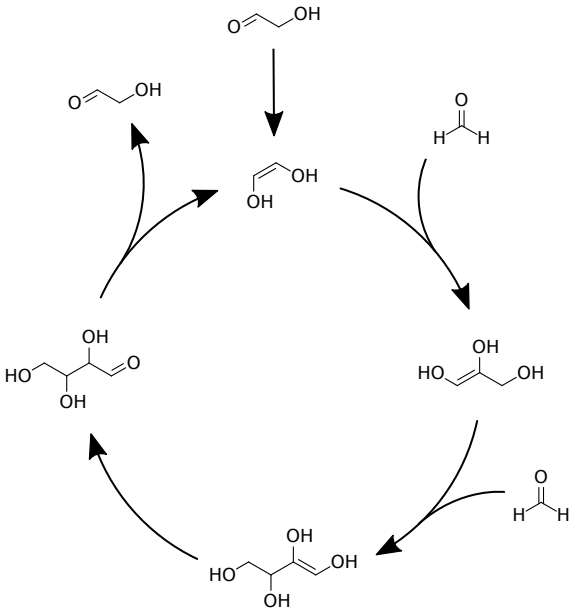

Figure 3.11: Explored autocatalytic cycle with the lowest activation barriers. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119. Copyright 2017 American Chemical Society.

The (generic) Chatt–Schrock cycle[14,15] (Fig. 3.13) is the prominent example of catalytic nitrogen fixation.[174] Its intermediates (referred to as *Schrock intermediates* hereafter) are formed by an alternating sequence of single protonation and single electron-reduction steps of Schrock's nitrogen-ligated molybdenum complex.[14,15] The sources of protons and electrons are 2,6-lutidinium (2,6-LutH) and decamethylchromocene ($CrCp_2^*$), respectively. This mechanism, however, does not explain the small turnover number of the catalyst. To demonstrate our heuristic network-exploration algorithm described above, we aim at identifying competing reaction paths of the Chatt–Schrock cycle. For details on the computational methodology, see Appendix A.1.2.

3.8.1 Heuristic Guidance for Explorations of Transition-Metal Catalyzed Reactions

Crucial for the construction of such heuristic rules is the choice of molecular descriptors. While graph-based descriptors perform well for many organic molecules, they may fail for transition-metal complexes, where the chemical bond is not always well defined.[177] In contrast to the previously studied system, we aim at a less context-driven method to be applied to an example of transition-metal catalysis. Such an approach should be based on information directly extracted from the electronic wave function so that no additional (ad hoc) assumptions on a particular class of molecules are required. A

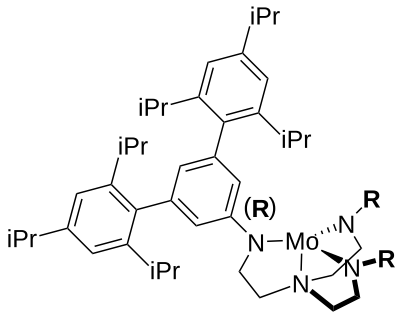

Figure 3.12: Schematical drawing of the [HIPTN$_3$N]Mo complex ([Mo]) published by Yandulov and Schrock.

simple example of the first-principles identification of reactive sites is the localization of Lewis-base centers in a molecule as attractors for Lewis acids. Lone pairs are an example for such Lewis-base centers and can be detected by inspection of an electron localization measure such as the ELF by Becke and Edgecombe[130] or the Laplacian of the electron density as a measure of charge concentration[131] (see also Ref. 132). Other quantum chemical reactivity indices can also be employed, such as Fukui functions,[133] partial atomic charges,[134,135,178] or atomic polarizabilities.[138,139] With these descriptors, reactive sites can be discriminated, i.e., not every reactive site may be a candidate for every reactive species (indicated by the coloring in Fig. 3.2). For example, an electron-poor site is more likely to react with a nucleophile rather than with an electrophile. Moreover, reactive species consisting of more than one atom may have distinct reactive sites. Naturally, the spatial orientation of a reactive species toward a reactive site is important.

Even though our heuristics-guided approach aims at restricting the number of possible minimum-energy structures, the number of generated intermediates may still be exhaustively large as the following example illustrates. For a protonation reaction, we may assume that the number of different protonated intermediates can be determined from the unprotonated target species by identifying all reactive sites (RS) which a proton, the reactive species, can attack. This number is given by a sum of binomial coefficients,

$$N = \sum_{p=1}^{n_{\mathrm{RS}}} \binom{n_{\mathrm{RS}}}{p} = 2^{n_{\mathrm{RS}}} - 1, \tag{3.1}$$

where n_{RS} is the number of reactive sites and p is the number of protons added to the target species. Even for such a simple example, the number of possible intermediates increases exponentially. For example, for a target species with ten reactive sites, $N = 1023$ intermediates will be generated. Obviously, the transfer of several protons to a single target species is not very likely from a physical point of view as the charge will

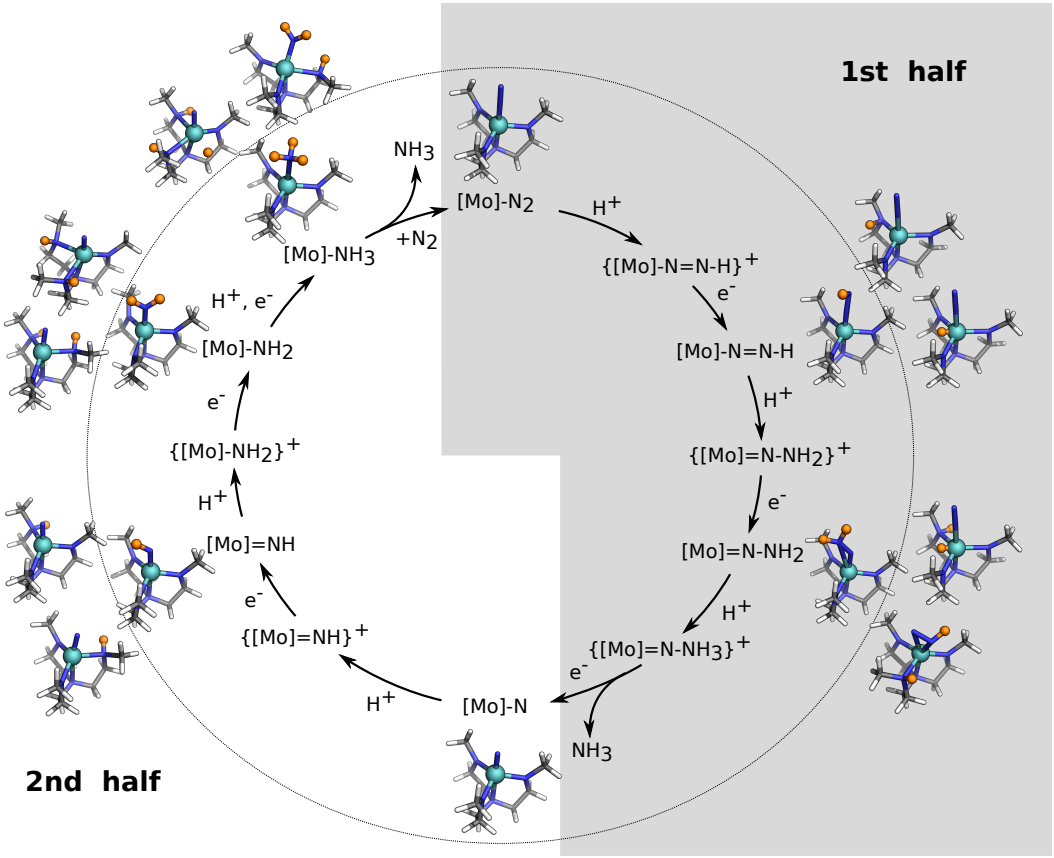

Figure 3.13: The Chatt–Schrock nitrogen-fixation cycle: The first and second half of the catalytic cycle are based on the [Mo]–N_2 and [Mo]–N scaffolds, respectively. [Mo] refers to the Yandulov–Schrock complex[14,15] where the hexa-*iso*-propyl terphenyl (HIPT) substituents are replaced by methyl groups to reduce the computational cost. Molecular structures within the circle represent Schrock intermediates. A selection of isomers of these Schrock intermediates is shown outside the circle. Element color code: gray, C; blue, N; turquoise, Mo; white, H; orange, H added to reactive sites. Reprinted with permission from M. Bergeler, G. N. Simm, J. Proppe, M. Reiher, *J. Chem. Theory Comput.* **2015**, *11*, 5712–5722. Copyright 2015 American Chemical Society.

increase such that the acidity of the protonated species might not allow for further protonation. In the presence of a reducing agent, however, these species can become accessible in reduced form.

3.8.2 Heuristics-Guided Structure Search

For the first and second half of the catalytic cycle (Fig. 3.13) [Mo]–N_2 and [Mo]–N (see Fig. 3.15) are taken as zeroth-generation structures, respectively. Here, [Mo] refers to the Yandulov–Schrock complex[14,15] where the HIPT substituents are replaced by methyl groups to reduce the computational cost. The bulky HIPT substituents can be reintroduced once the network has been established.

In this study, we only consider protons as reactive species since protonations of the amide nitrogen atoms are likely to deactivate the catalyst.[173,179] Additionally, we take different charges of the protonated complexes into account (single electron-reduction steps from $y+$ to neutral, with y being the number of protons added). However, for a more extensive exploration, H_2, N_2, NH_3, N_xH_y, and intermediates themselves must also be considered as reactive species.

To determine the reactive sites of the substrates, we exploit knowledge about negative charge concentrations extracted from the electronic wave function. As an example in Fig. 3.14, we present the isosurface of the ELF colored with the value of the electrostatic potential for the two parent species, [Mo]-N_2 and [Mo]-N, of the two halves of the Chatt–Schrock cycle in Fig. 3.13. Whereas the ELF highlights regions in space where the electron density is localized, the electrostatic potential allows us to pick those regions that can function as a Lewis base (highlighted in blue in Fig. 3.14) by contrast to the other regions that are electron deficient and feature hardly any Lewis basicity (highlighted in orange in Fig. 3.14 and showing, e.g., C–H σ-bonds). The blue regions, therefore, define spatial areas that function as reactive sites to which protons should be added as reactive species.

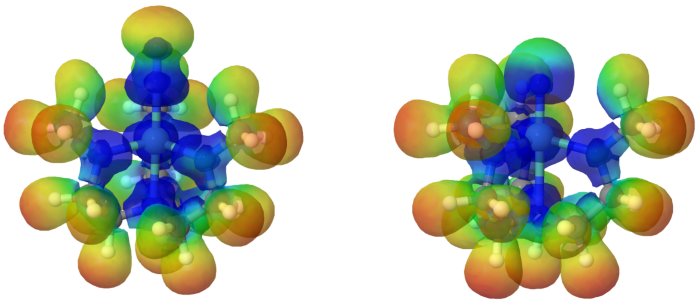

Figure 3.14: Electron localization function (ELF) for [Mo]-N_2 (left) and [Mo]-N (right) colored according to the electrostatic potential (an isosurface value for ELF of 0.6 a.u. was chosen). Reprinted with permission from M. Bergeler, G. N. Simm, J. Proppe, M. Reiher, *J. Chem. Theory Comput.* **2015**, *11*, 5712–5722. Copyright 2015 American Chemical Society.

Note that this procedure is solely based on the first principles of quantum mechanics, but that it is also in line with conventional chemical wisdom that, for carbon and nitrogen atoms, the formation of a tetrahedral surrounding including both bonding neighbors and reactive sites for incoming reactants (protons) represents a valence-saturated electronic situation. At the molybdenum center, three reactive sites are introduced in the plane spanned by the three amide-nitrogen atoms. We refrain from adding protons to the coordinating amine nitrogen atom in *trans*-position to the N_2 ligand as this would produce a decomposition pathway of the catalyst that is not likely to lead to an alterna-

tive catalytic cycle. Clearly, these decomposition reactions are important to track for a complete understanding of Schrock-type dinitrogen fixation catalysis, but we devote this aspect to future work. Instead, we are less restrictive with respect to the possible protonation sites at the metal center and at the terminal nitrogen atom of N_2 in [Mo]-N_2 — density-functional-theory calculations are fast for the size of system under study and can be carried out in parallel so that one should not limit the number of possible reactive sites too much in order not to risk overlooking of important intermediates. All reactive sites considered as proton-acceptor sites in this work are shown in Fig. 3.15. Up to four protons are added to the zeroth-generation structures. This number may be considered a chemically reasonable upper limit.

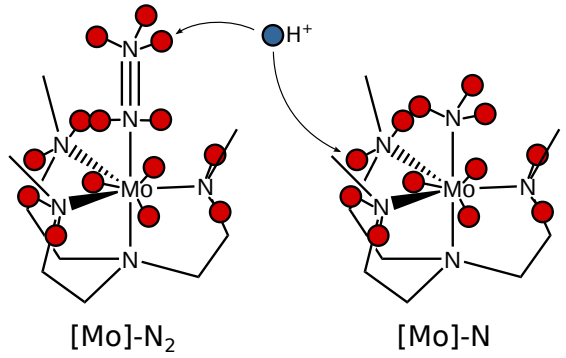

Figure 3.15: The reactive species (H^+) attacking the reactive sites (proton-accepting red circles) of [Mo]-N_2 and [Mo]-N. Lines are drawn between an atom and its reactive sites to highlight their spatial arrangement. Each amide nitrogen atom exposes two protonation sites, although we occupy at most one to produce an amine nitrogen atom. Reprinted with permission from M. Bergeler, G. N. Simm, J. Proppe, M. Reiher, *J. Chem. Theory Comput.* **2015**, *11*, 5712–5722. Copyright 2015 American Chemical Society.

Since the Yandulov–Schrock catalyst operates in the presence of a strong reducing agent ($CrCp_2^*$), protonated species can be readily reduced. Therefore, for a p-fold protonated reactive complex, we consider the charges $0 \leq c \leq p$. This results in a total number of

$$N = \sum_{p=1}^{n_{\mathrm{RS}}} \left\{ \binom{n_{\mathrm{RS}}}{p} (p+1) \right\} \tag{3.2}$$

structures. For [Mo]-N_2 N=6762 and for [Mo]-N N=3577 structures are obtained. However, for subsequent structure optimizations, these numbers are slightly reduced as the two protonation sites exposed by each amide nitrogen atom are occupied by at most one proton to yield an amine nitrogen atom. Decomposed reactive complexes such as those from which molecular hydrogen dissociated or in which the chelating ligand (partially) dissociated from the metal center, are automatically removed from the network.

In this study, conformational degrees of freedom of intermediates were not explored. In fact, the chelating ligand is designed to feature few degrees of freedom (apart from

the bulky HIPT moieties which aim to prevent dimerization and were replaced with methyl groups in this study). Therefore, we believe that this approximation does not compromise the conclusions drawn from the reaction networks.

In this study, TS searches are performed only for intramolecular elementary reactions for which the energy difference between the intermediates is below $E_c = 25$ kcal/mol. Note that this threshold does not refer to a Gibbs free reaction energy but to an electronic-energy difference. Nevertheless, since only reactions of the same type are compared, one can expect only small deviations of electronic energies from Gibbs free energies for intra-subnetwork reactions (proton shifts) and reduction steps. One can assume that this simplification is also a good approximation for protonation reactions. For intermolecular reactions (i.e., proton transfers from 2,6-LutH), no TSs were calculated, but a predefined root-mean-square deviation (RMSD) cutoff (0.5 Å for the first half and 0.65 Å for the second half of the Chatt–Schrock cycle) was chosen to determine the shared identity of two molecular structures (apart from an added proton in case of a protonation reaction).

3.8.3 Network Superstructure

In Fig. 3.16, subnetworks are arranged according to the number of protons and electrons added. Here, the subnetworks are denoted as $(x\mathrm{H}, c)_i$, where x is the number of hydrogen atoms added to the substrate i (1 = [Mo]-N_2, 2 = [Mo]-N) and c is the charge of the subnetwork. An arrow pointing from subnetwork a to subnetwork b will be crossed out if all intermediates of subnetwork b are at least by E_C energetically higher than all intermediates of subnetwork a.

Due to the energy cutoff E_c, entire subnetworks can be pruned and excluded from further analysis. For instance, starting from $(0\mathrm{H},0)_1$, $(2\mathrm{H},2+)_1$ cannot be reached via any other subnetwork without having to overcome a TS that is above E_C. Therefore, $(2\mathrm{H},2+)_1$ can be removed from the network. In both halves of the Chatt–Schrock cycle, all subnetworks with a total charge larger than one can be neglected. The pruning of these networks largely reduces the complexity of the network since now every subnetwork can only be reached from one other subnetwork.

The energy profiles of the first and second half of the catalytic cycle are shown in Figs. 3.17 and 3.18, respectively. In both figures, energy levels of Schrock intermediates are connected by dashed lines. An additional energy level will be shown if an intermediate lower in energy than the Schrock intermediate is part of that subnetwork. Moreover, if intermediates from which H_2 dissociated are observed, the intermediate with the lowest energy will be shown in red.

It can be seen that most reactions of the first half of the Schrock cycle are exothermic. Especially the reductions of the Schrock intermediates of $(1\mathrm{H},1+)_1$ and $(3\mathrm{H},1+)_1$ are

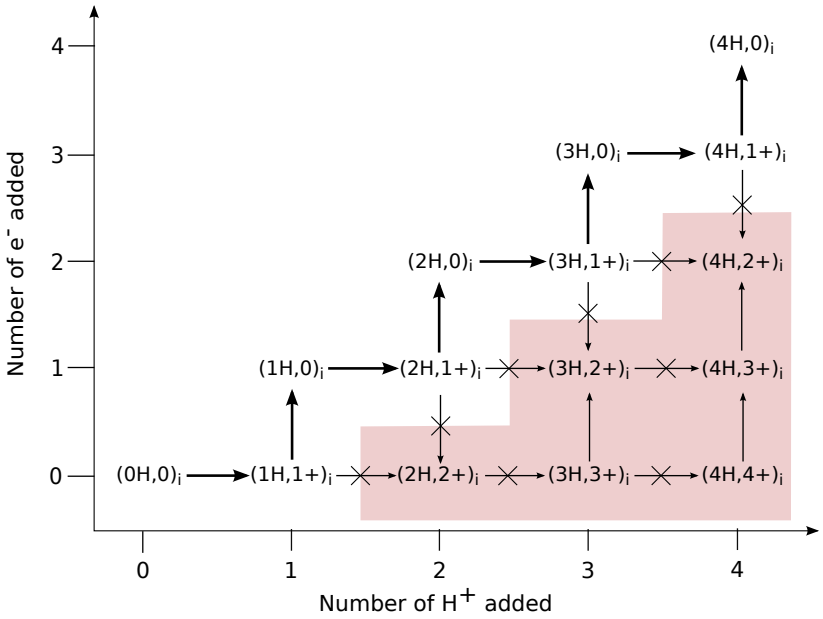

Figure 3.16: Energy analysis of Schrock's catalytic nitrogen fixation. Subnetworks in the red shaded area cannot be reached from their respective substrate i (1 = [Mo]-N_2, 2 = [Mo]-N) without exceeding a TS energy above E_C. Reprinted with permission from M. Bergeler, G. N. Simm, J. Proppe, M. Reiher, *J. Chem. Theory Comput.* **2015**, *11*, 5712–5722. Copyright 2015 American Chemical Society.

thermodynamically favorable. In addition, the dissociation of NH_3 from $(3H, 0)_1$ has a highly negative reaction energy. Note, however, that this dissociation energy holds for a specific choice of acid and reductant so that the assignment of this reaction energy solely to the breaking of the N–N bond would be misleading as discussed in our earlier work.[180,181] There are also endothermic reactions. For example, the protonation of the Schrock intermediate of $(0H,0)_1$, was calculated to have a positive reaction energy of +3.1 kcal/mol. A thermodynamically favorable alternative to the protonation of N_2 is the protonation of the amide of the chelate ligand. This intermediate is lower in energy ($\Delta E = -13.8$ kcal/mol) than the Schrock intermediate.

In addition, most reactions in the second half are exothermic. The protonation of the Schrock intermediate in $(1H,0)_2$ is particularly exothermic with a reaction energy of −32.2 kcal/mol. Nonetheless, there are subnetworks in which the Schrock intermediate is not the most stable species. In $(3H,0)_2$, for instance, there is an intermediate which is more stable ($\Delta E = 4.7$ kcal/mol) than the respective Schrock intermediate. Furthermore, it can be seen that the dissociation of H_2 is thermodynamically favorable in several subnetworks. For example, in $(2H,0)_1$ the dissociation of H_2 even results in the most stable intermediate.

However, we should emphasize that those structures which are very similar in terms

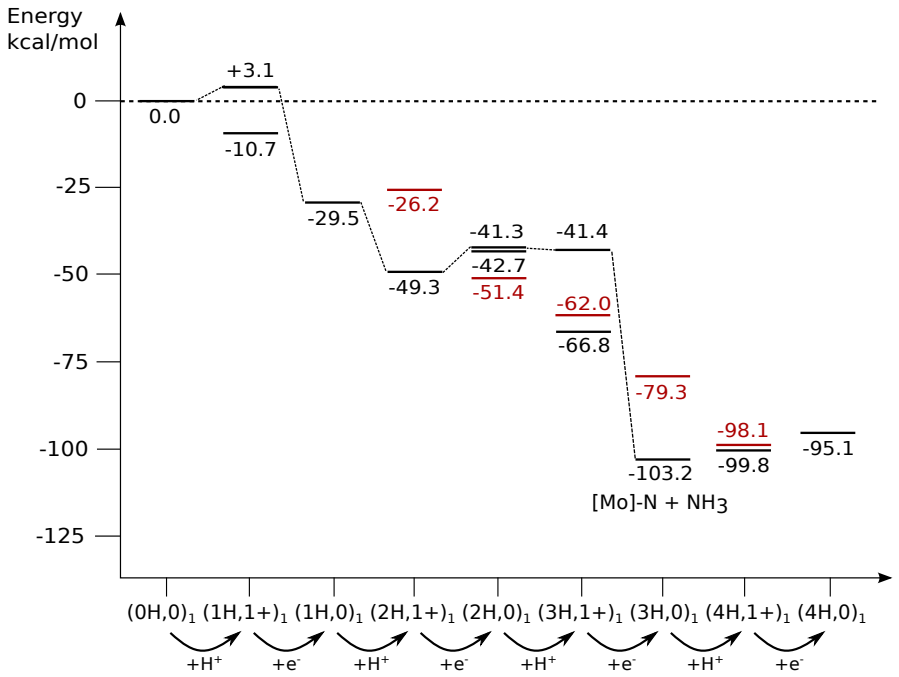

Figure 3.17: Energy profile of the first half of the Chatt–Schrock cycle including intermediates lower in energy than the Schrock intermediates of the inner circle in Fig. 3.13. Schrock intermediates are connected by dashed lines. Intermediates from which H_2 dissociated are given in red. The oxidation potential of $CrCp_2^*$ and the dissociation energy of LutH were calculated (BP86/def2-SV(P)) to be $+103.7$ and -237.7 kcal/mol, respectively. For further details see Appendix A.1.2. Reprinted with permission from M. Bergeler, G. N. Simm, J. Proppe, M. Reiher, *J. Chem. Theory Comput.* **2015**, *11*, 5712–5722. Copyright 2015 American Chemical Society.

of energy may be considered equally stable, especially when viewed in the light of the quantum chemical methodology chosen here. Moreover, one must keep in mind that the network exploration was carried out for a small model complex of the Schrock catalyst with a double-zeta basis set.

While the dissociation of NH_3 is energetically favorable in the first half, it is very unfavorable in the second. Therefore, a four-coordinate [Mo] intermediate appears unlikely, and an associative exchange mechanism might be favored over the dissociative one as has already been discussed in the literature.[179,181,182] Since further protonation of $(3H,0)_2$ results in low-energy intermediates, we can identify this subnetwork as a possible starting point of degradation.

The results for the Schrock intermediates reported here are in qualitative agreement with those reported earlier by the Reiher group[179–183] and by Tuczek and coworkers.[184,185] Numerical deviations for the Schrock intermediates are mostly due to the choice of a small model structure and the smaller basis set employed in this work.

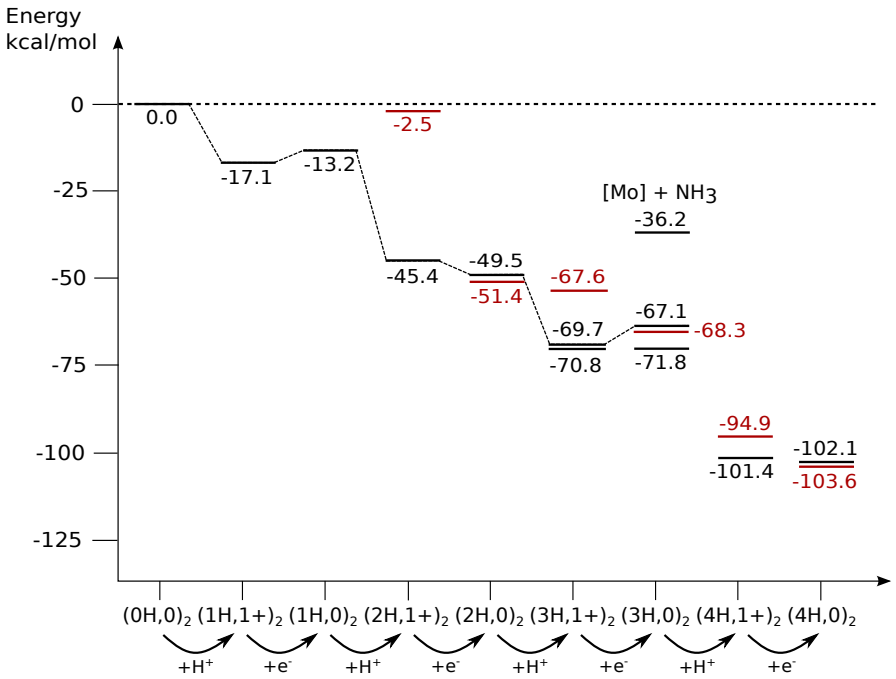

Figure 3.18: Energy profile of the second half of the Chatt–Schrock cycle including intermediates lower in energy than the Schrock intermediates of the inner circle in Fig. 3.13. Schrock intermediates are connected by dashed lines. Intermediates from which H_2 dissociated are given in red. The oxidation potential of $CrCp_2^*$ and the dissociation energy of LutH were calculated (BP86/def2-SV(P)) to be $+103.7$ and -237.7 kcal/mol, respectively. For further details see Appendix A.1.2. Reprinted with permission from M. Bergeler, G. N. Simm, J. Proppe, M. Reiher, *J. Chem. Theory Comput.* **2015**, *11*, 5712–5722. Copyright 2015 American Chemical Society.

3.8.4 Reaction Network

To rationalize the low efficiency and stability of Schrock's nitrogen fixation catalyst, not only all possible intermediates but also the TSs connecting them need to be analyzed. The reaction network of the Chatt–Schrock cycle automatically generated by Chemoton is shown in Fig. 3.19. Each vertex represents an intermediate, whereby the color encodes the energy difference with respect to the lowest intermediate of the subnetwork. Vertices representing a Schrock intermediate are enlarged. A collection of vertices belonging to the same subnetwork is enclosed by a solid black line. Two vertices of the same subnetwork are connected by an undirected edge if a TS was found between them. The gray scale of such an edge serves as a visual cue indicating the height of the transition barrier with respect to the lower-energy intermediate. Light-gray edges represent high-energy barriers, dark-gray edges represent low-energy barriers. It is important to note that according to our exclusion rule, many TSs in this network need not be optimized and can, therefore, be omitted. However, to illustrate the complexity of such reaction networks, TSs with an energy above E_C were not removed. Vertices

of different subnetworks are connected by undirected edges (dashed lines) for which no TSs were calculated. In addition, the molecular structures of selected intermediates are shown in Fig. 3.19 a) – g). It should be noted that due to the network's structure its visualization is different from the one described in Section 3.6.

Starting from the [Mo]-N_2 complex in $(0H, 0)_1$, a proton is added to reach $(1H,1+)_1$. As can be seen from the dashed lines, this reaction can result in four different intermediates. The Schrock intermediate, the [Mo]-N_2 complex protonated at the molybdenum atom (Fig. 3.19 a)), the [Mo]-N_2 complex with a proton at one of the amido groups (Fig. 3.19 b)), and the enantiomer of that intermediate (Fig. 3.19 c)). From there, each intermediate can either undergo a reduction to form an intermediate in $(1H,0)_1$ or—through an intramolecular reaction—transform into another intermediate of the same subnetwork. The subsequent protonation of intermediates in $(1H,0)_1$ leads to the subnetwork $(2H,1+)_1$.

The inspection of the first four subnetworks already suggests a feasible alternative to the Chatt–Schrock mechanism: The [Mo]-N_2 complex is protonated at the amido group; this intermediate undergoes reduction, protonation (of the axial N_2), and finally, a proton shift to reach the energetically most favorable intermediate, the Schrock intermediate of $(2H,1+)_1$.

The reduction of the intermediates in $(2H,1+)_1$ leads to a subnetwork in which not the Schrock intermediate but intermediate d) is the most stable intermediate. This intermediate can be reached through several different cascades of transformations, which however all contain at least one that is comparatively high in energy. It can be seen in Fig. 3.19 that once an intermediate of $(2H,0)_1$ other than the Schrock intermediate is protonated, no rearrangement reaction within $(3H,1+)_1$ was found which leads to the Schrock intermediate. This also suggests that the Schrock intermediate in $(3H,1+)_1$, which is relatively high in energy, does not easily transform into a more stable intermediate of the same subnetwork. Likewise, $(3H,1+)_1$ and $(3H,0)_1$ (not shown in Fig. 3.19) can be considered relevant for the process of degradation of this catalyst.

After reduction of the Schrock intermediate of $(3H,1+)_1$, NH_3 dissociates and the [Mo]-N complex is formed. Similar to the first half, the protonation of the [Mo]-N complex can lead to two different intermediates: the Schrock intermediate and the [Mo]-N complex with a proton at one of the amido groups. These two structures could give rise to two different reaction paths. Furthermore, two other subnetworks appear to be particularly prone to initiating degradation: $(3H,1+)_2$ and $(3H,0)_2$. In both subnetworks, there are intermediates that are more stable than the Schrock intermediate, which can be reached via low-energy TSs. In $(3H,1+)_2$, it is the shift of one of the three protons from one of the axial nitrogen atoms to an amido group (see structures e) and f)). Either of these structures can undergo an additional transformation where the proton bound

to the amido group shifts to the molybdenum center. After reduction, this structure forms intermediate g)—the most stable conformation of $(3H,0)_2$. Likewise, the Schrock intermediate of $(3H,0)_2$ can undergo a proton shift to form intermediate g). As mentioned earlier, the dissociation of NH_3 from [Mo]-NH_3 is highly endothermic,[179,181,182] and therefore, the exchange of NH_3 and N_2 via a six-coordinated complex is likely to occur. Therefore, the intermediate g) in $(3H,0)_2$ can be considered particularly relevant to understanding the low turnover number of the catalyst.

It should be evident from the presentation above that our automated visualization strategy generates a presentation of chemical reaction networks that directly unveils its essence to the reader. Thereby, even complex reaction mechanisms involving many side reactions become lucid and, hence, will be comprehensible.

3.8.5 Alternative Pathways to the Chatt–Schrock Cycle

By applying the energy cutoff $E_C = 25$ kcal/mol, many intermediates in the reaction network can be removed. The resulting reaction network allows for the identification of reaction pathways, other than the Chatt–Schrock cycle, that are likely to occur at ambient conditions. These pathways are shown in Fig. 3.20.

It can be seen that multiple pathways next to the Chatt–Schrock cycle are indeed possible. For example, two pathways running parallel to the Chatt–Schrock cycle can be identified. It is important to note that pathways which do not form a cycle and thus lead to the degradation of the catalyst, are not shown, but will be investigated in future studies.

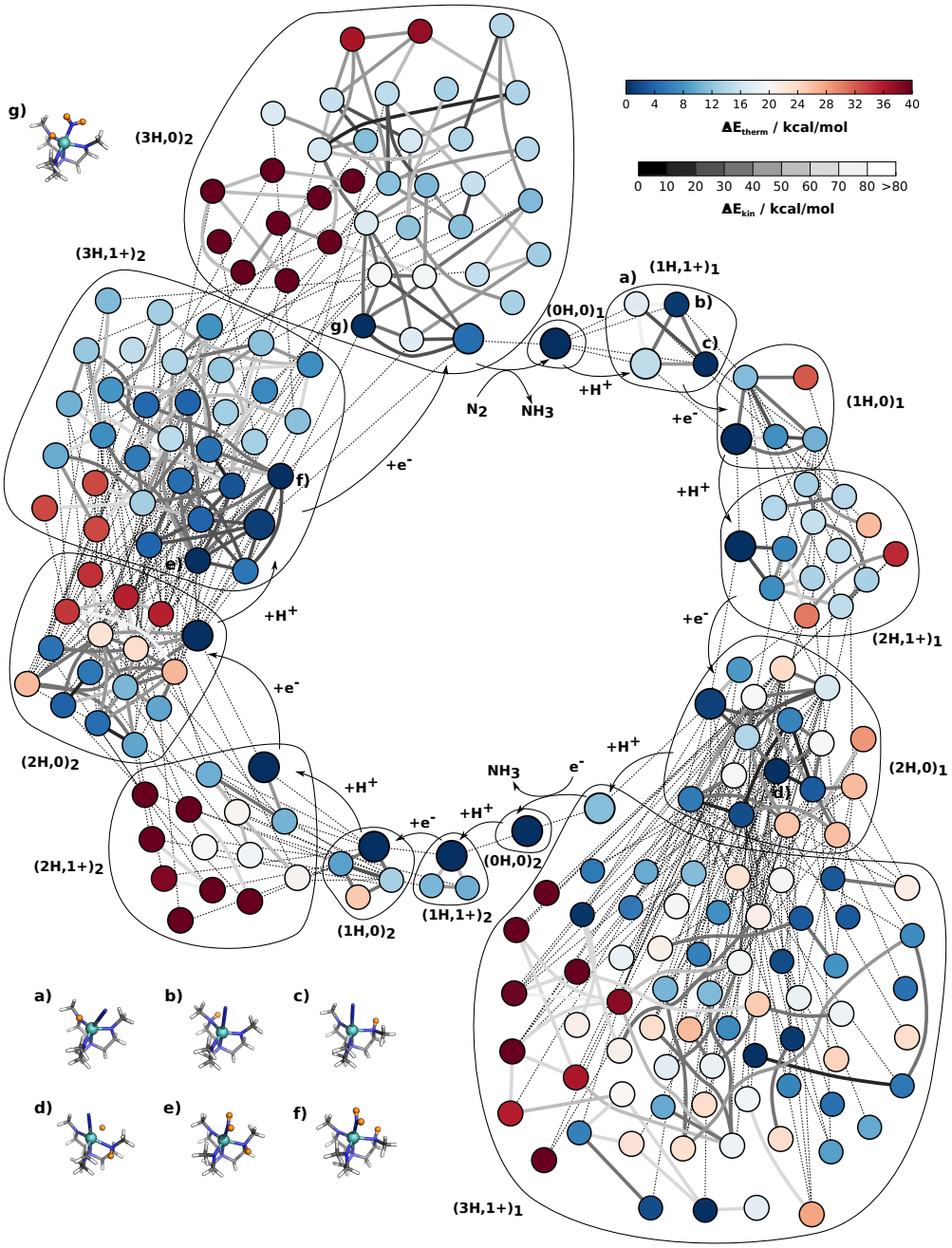

Figure 3.19: The Chatt–Schrock network of catalytic nitrogen fixation. Dark-blue vertices refer to the lowest-energy intermediates of a subnetwork, dark-red vertices to the corresponding highest-energy intermediates. Vertices representing Schrock intermediates are enlarged. Low-energy transition barriers between intermediates of the same subnetwork are indicated by dark-gray edges, high-energy transition barriers by light-gray edges. Inter-subnetwork connections are indicated by dashed lines. In a)–g) a selection of intermediates is shown. Element color code: gray, C; blue, N; turquoise, Mo; white, H; orange, H added to reactive sites. Reprinted with permission from M. Bergeler, G. N. Simm, J. Proppe, M. Reiher, *J. Chem. Theory Comput.* **2015**, *11*, 5712–5722. Copyright 2015 American Chemical Society.

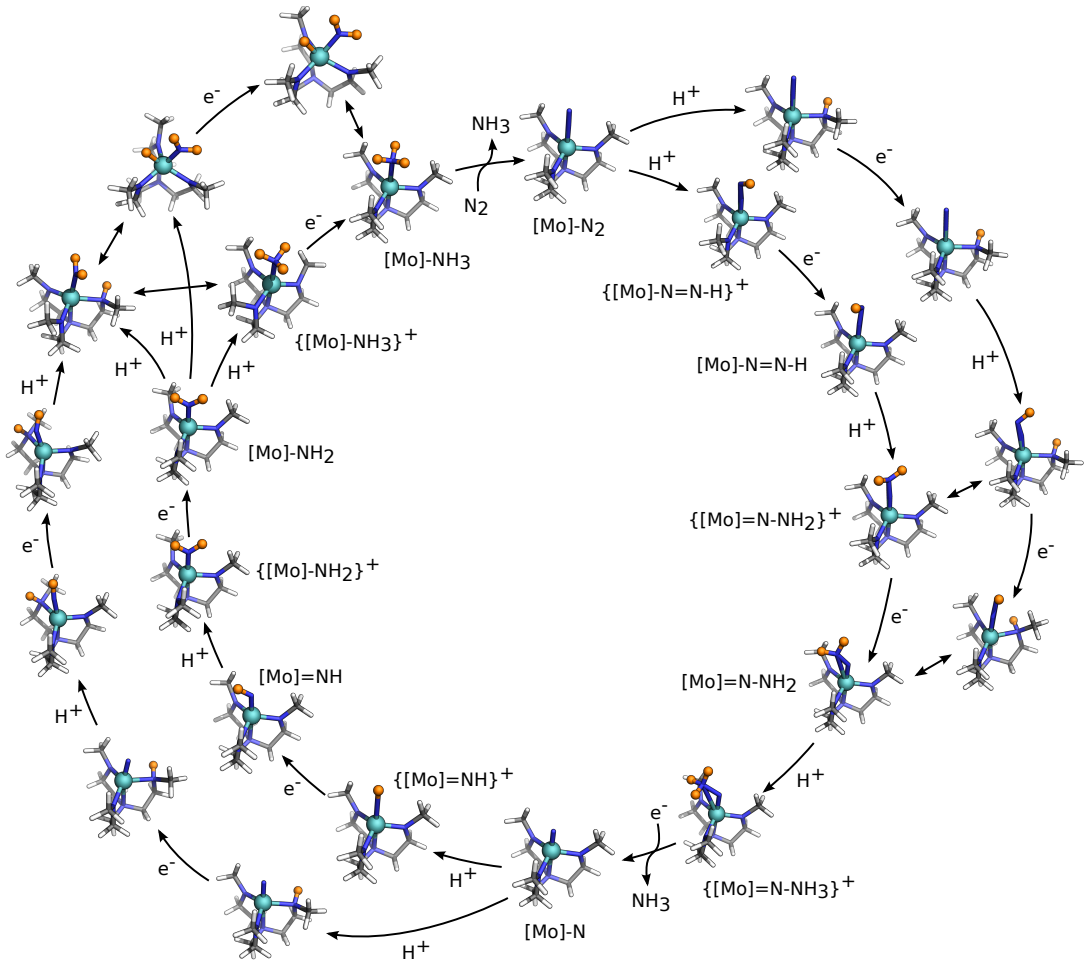

Figure 3.20: Alternative pathways (outer circle) to the Chatt–Schrock cycle (inner circle). Proton shifts are indicated by double-headed arrows. Element color code: gray, C; blue, N; turquoise, Mo; white, H; orange, H added to reactive sites. Reprinted with permission from M. Bergeler, G. N. Simm, J. Proppe, M. Reiher, *J. Chem. Theory Comput.* **2015**, *11*, 5712–5722. Copyright 2015 American Chemical Society.

4

Error Assessment of Computational Models in Chemistry*

In general, for the study of all but the smallest chemical systems, the application of so-called "gold standard" quantum chemical methods (e.g., coupled cluster) is not computationally feasible. In particular, when investigating large chemical systems approximate quantum chemical methods are indispensable. However, the accuracy of such approximate methods is often difficult to determine. In this Chapter, we discuss the issue of performance assessment of computational models based on several examples from the quantum chemistry literature. For this purpose, we elucidate the different sources of uncertainty, the elimination of systematic errors, and the combination of individual uncertainty components to the uncertainty of a prediction.

4.1 Role of Benchmark Studies in Uncertainty Quantification

It is generally assumed that performance statistics based on benchmark systems are good estimates for the prediction uncertainty of a quantum chemical method. Due to the availability of large amounts of experimental and computational reference data (for

*This Chapter is reproduced in part with permission from G. N. Simm, J. Proppe, M. Reiher, *CHIMIA* **2017**, 71, 202–208. Copyright 2017 Swiss Chemical Society.

a recent review see Ref. 186), benchmark studies are carried out to provide statistical quantities such as the mean absolute error (MAE),

$$\mathrm{MAE}_m = \frac{1}{N}\sum_{s=1}^{N}|e_{m,s}|, \tag{4.1}$$

and the largest absolute error (LAE),

$$\mathrm{LAE}_m = \max\{|e_{m,1}|, |e_{m,2}|, ..., |e_{m,N}|\}, \tag{4.2}$$

with $e_{m,s} = c_{m,s} - o_s$ and N being the size of the data set. Here, the error $e_{m,s}$ of model m with respect to system s (typically a molecule) is defined as the difference between the calculated result $c_{m,s}$ and the experimental or computational reference o_s. These summarizing statistics are then applied to estimate the prediction uncertainty of a method of choice for a system of interest.

However, there is a major caveat associated with this approach: the assumption that such statistics are transferable to a system not represented in the reference data set is generally invalid. In Table 4.1 the MAE of common density functionals with respect to ligand dissociation energies of transition metal complexes from three previous studies are compared. The WCCR10 data set[187] consists of 10 ligand dissociation energies of large cationic transition metal complexes. The 3dBE70 database[188] contains average bond energies of 70 transition metal compounds. The data set by Furche and Perdew[189] containing 18 dissociation energies of transition metal compounds is herein abbreviated as FP06. The comparison of the different benchmark studies shows that the MAEs are

Table 4.1: Mean absolute error (MAE_m) of ligand dissociation energies (in kJ/mol) calculated with a selection of common density functionals m taken from the literature.

Model m	WCCR10[187]	3dBE70[188]	FP06[189]
B3LYP[190–192]	39.1	20.9	50.2
PBE[193–195]	31.8	25.5	45.2
TPSSh[196]	32.0	17.6	40.6

strongly data set dependent. For instance, the spread of MAEs ranges from 17.6 to 40.6 kJ/mol in the case of the TPSSh density functional.

Even for small systems such as metal dimers, the reported statistics can vary. For example, for the dissociation energy of metal dimers the study by Furche and Perdew[189] and Schultz et al.[197] report MAEs of 50.6 and 69.9 kJ/mol, respectively. This finding is in accordance with many studies demonstrating that the accuracy of density functionals varies strongly with the chemical system,[187,198–203] and therefore, undermining the transferability of such performance statistics. In the case of density functional theory,

this lack of transferability is particularly critical to studies on transition metals since most of the benchmark data sets include only small (unsaturated and therefore atypical) compounds (e.g., transition metal hydrides such as FeH).

In addition, it can be seen from Table 4.1 that all MAEs are considerably large (a result is said to be within chemical accuracy if the expected error is within $\approx$ 4.2 kJ/mol). For the WCCR10 and FP06 data sets LAEs are reported as well (e.g., 83.4 and 157.3 kJ/mol, respectively, for the B3LYP functional). MAE and LAE of this size are unacceptable for studies in which accurate reaction energy are of high importance. In the framework of conventional transition state theory, an error of 30 kJ/mol in the barrier height of an elementary reaction step results in a reaction rate that is off by a factor of 10^5 at room temperature.

Lastly, it should be noted that the uncertainty within the (experimental and computational) reference data is generally not accounted for.[204]

In Fig. 4.1, we illustrate the system dependency of an arbitrary observable given an adequate computational model (see Section 4.3.2 for a definition of model adequacy). The transferability of statistical measures such as the MAE would only be valid in the ideal case of homoscedasticity (Fig. 4.1, left), where the prediction uncertainty is independent of the input, here, chemical space (the space of all chemical compounds, e.g., molecules, where small distances indicate high structural similarity).

So far, there exists no strategy to develop approximate quantum chemical methods with system-invariant uncertainty (homoscedasticity), which is not to be confused with strategies to develop systematically improvable methods (such as the coupled cluster expansion, which still reveals systematic errors due to the truncation of the degree of excitation — even if the degree is taken to be rather high). Consequently, we are generally faced with approximations yielding heteroscedastic results (Fig. 4.1, right), where the prediction uncertainty somehow depends on the nature of the chemical system.

This dependency is generally unknown (not as indicated in the right frame of Fig. 4.1), which also implies that estimation of prediction uncertainty for data lying in the same region of the chemical space employed for model training can be unreliable. Noteworthy, the Hohenberg–Kohn functional would, in principle, yield results with system-invariant accuracy (for chemical systems in their electronic ground states), however, this is not the case in practice due to the approximations of the exchange–correlation density functional.

Due to the continuous advancement of accurate and efficient black-box methods (such as explicitly correlated coupled cluster theory, for a review see Ref. 205) and the increase of computational power, it is believed that gold standard methods will, eventually, become the standard method of choice. In this case, uncertainty estimation will be less important if chemical accuracy is reached and considered sufficient. For higher accuracy

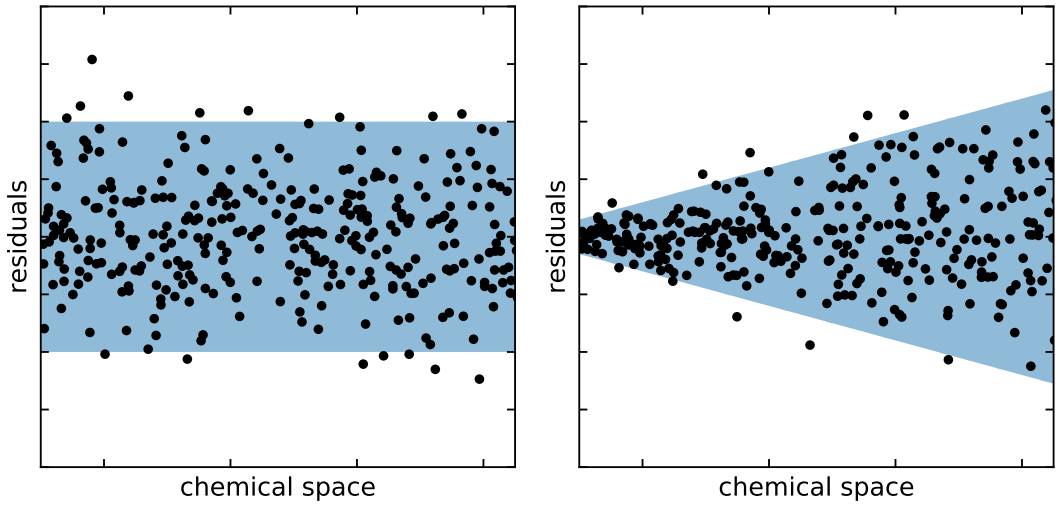

Figure 4.1: Illustration of homoscedasticity (left) and heteroscedasticity (right) for synthetic data. In the former case (left), the uncertainty (blue 95% confidence band) associated with an observable of interest is independent of the chemical system studied. In the latter case (right), which is the more general case, the uncertainty associated with the observable of interest is a function of the chemical space. The distance between two data points along the abscissa is thought to be inversely proportional to the similarity of the corresponding molecular structures. Hence, if a prediction method is trained on a small hypervolume of the chemical space, it will not be possible to transfer the associated uncertainty to a larger hypervolume. Moreover, since the variance function is generally unknown, also internal predictions (in the same hypervolume where the method has been trained) are unreliable.

also standard coupled cluster models will require rigorous error estimation. Although the system size for which these methods are feasible increases due to constant method-development efforts, less accurate methods are usually chosen for feasibility reasons when a large number of calculations must be carried out. This is the case for extensive explorations of vast reaction networks,[78,92,93,99] screening studies,[206,207] and reactive molecular dynamics simulations.[48,51,52]

4.2 Error Assignment for Approximate Models

The identification and separation of sources of uncertainty are difficult since multiple approximations of unequal accuracy are made during method development. For example, in density functional theory, the exact density functional is approximated in a rather involved way. In standard coupled cluster theory, the wave function is based on a single reference (Slater determinant). On the one hand, these and other sources of uncertainty may combine in an arbitrary manner and even lead to counter-intuitive total errors.[208] For example, coincidental error compensation can lead to overestimation of prediction accuracy. This is an effect often encountered in density functional theory. For instance, the success of the B3LYP[190–192] functional together with the poor 6-31G* basis set[209]

is often attributed to error cancellation.[210] Error compensation was also reported for coupled cluster methods, for instance, CCSD(T) was found to provide more accurate results than CCSDT in combination with certain one-electron basis sets.[211] On the other hand, there are approximations (e.g., considering the atomic nucleus as a point charge rather than as an extended charge distribution, ignoring certain relativistic effects) that are local (atomic) and cancel out for reaction energies (or valence properties).

4.3 Uncertainty Classification

In general, one distinguishes between three main sources of uncertainty: parameter uncertainty, numerical uncertainty, and systematic errors due to inconsistent data and inadequate model approximations (here, to the fundamental theory of chemistry, QED).[212] Except for stochastic models (e.g., Monte Carlo simulations), numerical uncertainty is expected to be negligible and will not be discussed in the following. The remaining sources of uncertainty are elaborated on and approaches for their remedy are elucidated.

4.3.1 Parameter Uncertainty

The uncertainty of a model's parameters needs to be considered when making predictions on chemical systems not included in the fitting (training) of the model. Solely considering the "optimal" values (e.g., obtained by minimizing the sum of squared residuals) is not sufficient, as one would neglect a potentially essential component of the prediction uncertainty of a model. Parameter uncertainty results from random and systematic errors in both the reference data and the model under consideration (see Section 4.3.2), in particular, if the number of reference data is small. Only for large data sets and small domains of application, parameter uncertainty becomes negligible.

Parameter uncertainty can be estimated, for example, through Bayesian inference[213] or through sampling methods such as bootstrapping.[214] In the latter case, the reference data set itself replaces the assumption of a parametric population distribution (e.g., a Gaussian distribution) underlying the data. With bootstrapping, one draws multiple samples from the data set with replacement. Every such bootstrap sample will yield different parameter values compared to the original sample, the ensemble of which allows estimation of parameter uncertainty.

Assuming that systematic errors in the computational model have been eliminated (for instance, by *a posteriori* corrections of its results[204]), the effect of the reference set employed on the parameter distributions remains to be examined. If the reference data contain systematic errors, small changes in its composition (e.g., removal or addition of a few data points) will have a large effect on the parameter distributions. The jackknife[214]

is a well-established method for the detection of inconsistencies, In this method, a high dependence of the parameter distributions on certain items in the data set is identified by randomly removing data points. With a data set containing N data points, one obtains N jackknife estimates of the parameter distributions, each of them derived from a modified data set in which the s-th data point is removed ($s = 1, ..., N$).

4.3.2 Model Inadequacy

An inadequate computational model is not able to reproduce reference data within their uncertainty range,[212] i.e., the model under- or overestimates the uncertainty of the reference data. Underestimating prediction uncertainty is a result of overfitting, where the computational model is too flexible (features too many parameters) such that it does not only fit the explainable part of the reference data (the underlying physics), but also its unexplainable part (noise). By contrast, underfitting is caused by models which are too rigid (possess too few parameters) to fit the explainable part of the reference data, leading to overestimation of prediction uncertainty. Moreover, model inadequacy can be divided into an explainable (systematic) and an unexplainable (random) part, which is illustrated in Fig. 4.2.

For instance, most quantum chemical methods (with the exception of multi-configurational methods) struggle to correctly describe two hydrogen atoms at large distance. In fact, all density functionals fail to describe stretched H_2^+ and H_2.[215] The smoothness of the corresponding energy–distance plots (see, for instance, Figure 2 in Ref. 215) reveals that random model inadequacy plays a negligible role in this "simple" case of two nuclei. However, the fact that all of these energy–distance plots reveal a non-constant deviation from those obtained with accurate multi-configurational methods shows the large significance of systematic model inadequacy. While in this special case, model inadequacy could be easily eliminated by fitting a reasonable function linking data from benchmark and approximate calculations, the situation will become much more complicated if a larger fraction of chemical space is considered. For instance, due to their complex electronic structure, molecular structures containing transition metals are challenging targets for current quantum chemical methods. Despite containing adjustable empirical parameters, many density functionals fail to achieve a statistically valid description of these systems.[216] For example, Reiher and coworkers showed that the parameters of a standard functional are flexible enough to be chosen to exactly reproduce each coordination energies of a data set containing large organometallic complexes.[216] However, due to model inadequacy, there exists no unique parameter set that is equally accurate for all coordination energies in this data set at the same time.

Note that model inadequacy is difficult to distinguish from data inconsistency. If the

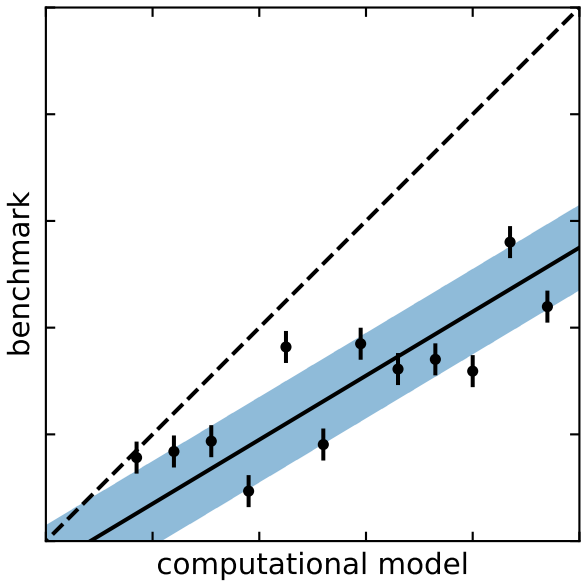

Figure 4.2: Illustration of systematic and random model inadequacy for synthetic data. Data points of an adequate model would scatter around the line going through the origin (dashed line). However, it can be seen that the (inadequate) model deviates from the benchmark results (black points). By fitting a linear function to the data set (solid line), an *a posteriori* correction of the model can be achieved. The scattering of the data points around the solid line appears to be random, however, for the most part, the residuals are significantly larger the uncertainty in the benchmark results (indicated by error bars representing two standard deviations). This effect is referred to as random model inadequacy and implies that the uncertainty of the model (represented by the blue 95% prediction band) exceeds the uncertainty of the benchmark.[204]

reference data contain systematic errors, even high-accuracy models would not be able to reproduce the reference data. In that case, it would be the wrong decision to improve on the computational model (high overfitting tendency). Given the reference data is corrected for inconsistencies, there are several tools at hand to tackle model inadequacy: one can improve the underlying model, reduce the domain of application, or correct predictions through a statistical calibration approach.[212,217]

Model Improvement

If the computational model at hand is systematically improvable (as, for instance, in the case of a coupled cluster expansion) reduction of model inadequacy is, in principle, straightforward. However, such methods are currently limited to relatively small system sizes and a few structures to be considered.

In density functional theory, model improvement is often referred to as climbing up Jacob's ladder.[218] Higher rungs incorporate increasingly complex ingredients constructed from the density or the Kohn–Sham orbitals (e.g., gradient and Laplacian of

the electron density, kinetic energy density). The original proposition of a ladder is that each rung satisfies certain exact constraints (there exist 17 of them, see the Supplementary Material of Ref. 219) and the next higher rung should be based on the previous rungs.[215] Since the exact density functional is not known and the number of known exact constraints is severely limited, systematic model improvement is not trivial.

In fact, a very recent study has shown that current developments steer away from systematic model improvement and towards functionals of empirical nature lacking physical rigor.[220] Most density functional development is focused on energies, implicitly assuming that functionals producing better energies become better approximations of the exact functional. The exact functional will produce the correct energy only if the input electron density is exact as well. By contrast, Peverati and Truhlar[186] argued that exact constraints can be neglected for the sake of greater flexibility in the energy fitting. However, such flexibility comes at the cost of reduced transferability (due to overfitting) to both other observables and chemical systems not included in the training of the computational model. To avoid loss of model transferability, Mardirossian and Head-Gordon suggest a validation approach in which the performance of a certain density functional is assessed for a data set not involved in the training of that density functional.[221,222] This way, one can successively increase model flexibility until the validation indicates a decrease of transferability (due to an increase in the performance statistics chosen).

Composite methods such as Gaussian-n (G-n),[223–226] Weizmann (W-n),[227–229] and HEAT[230] aim for high accuracy by combining the results of several calculations. They build a hierarchy of computational thermochemistry methods which allows the calculation of molecular properties such as total atomization energies and heats of formation to a high accuracy. The W-4 method calculated atomization energies of a set of small molecules with an MAE below 1 kJ/mol.[229] Similarly, the HEAT protocol predicted enthalpies of formation with an accuracy below 1 kJ/mol for 31 atoms and small molecules.[230] These protocols rely on computationally expensive coupled cluster calculations including high excitations. The HEAT method applies additional calculations (e.g., the diagonal Born–Oppenheimer correction) to be able to reproduce experimental results to higher accuracy. While the results from such methods are promising, the computational cost is far too high for large-scale applications mentioned above.

Errors in estimating prediction uncertainty due to model inadequacy can be eliminated not only by *internal* correction of a computational model (see the examples above) but also through *external* correction of the results produced with a computational model.[204] The simplest external corrections are linear functions, which are applied in the prediction of, for example, vibrational frequencies[231–233] or Mössbauer isomer shifts.[234–239] In such cases, parameter inference (calibration) can be much more efficient than internal calibration of the result-generating model. A drawback is the loss

of transferability to other observables since the external calibration model corrects an expectation value of a certain observable and not its underlying wave function, which is the unique common physical ground of all observables.

Reduction of Domain of Application

Another way of reducing model inadequacy is by training a computational model on a smaller domain of chemical space,[240] i.e., a small set of similar molecules such as sugars or amino acids. For example, due to the strong approximations made during method development (to gain efficiency), semi-empirical methods exhibit model inadequacy, which they attempt to remedy by introducing parameters which are then fitted to a specific data set (for a recent review see Ref. 241). This data set comprises a selection of molecules for which the resulting method is tailored. In fact, semi-empirical methods have been reparameterized to improve their description of a single molecule.[242] Similarly, density functionals were developed for specific applications, e.g., for kinetic studies.[127,243] In Fig. 4.3, the effect of the domain of application on model inadequacy is illustrated with a toy model.

Increase of Parameter Uncertainty

One can attempt to compensate model inadequacy by a controlled increase in parameter uncertainty. This way, one can build a statistical method with prediction uncertainty representative of the model residuals (deviation of benchmark data from model predictions).

In 2005, Nørskov, Sethna, Jacobsen, and co-workers implemented this approach for error estimation of results from density functionals[244] (see also Refs. 245–247). Instead of considering only the best-fit parameters of a density functional, they assigned a conditional probability distribution to them so that a mean and a variance can be assigned to each computational result. While promising general-purpose non-hybrid density functionals were designed within this framework (e.g., BEEF-vdW[248] and mBEEF[249,250]), the accuracy of uncertainty predictions remains unsatisfying.[212] This limitation can be attributed to model inadequacy and the heteroscedasticity of the large domain of chemical space to which they applied the functionals.

Compared to improving the computational model itself, increasing parameter uncertainty is straightforward as it only requires modification of the unknown part (parameter distributions) of an otherwise known model. Compared to external calibration (*a posteriori* correction of results obtained from a computational model), increasing parameter uncertainty in the corresponding prediction model preserves its transferability to other observables than the reference observable (for which model inadequacy has been

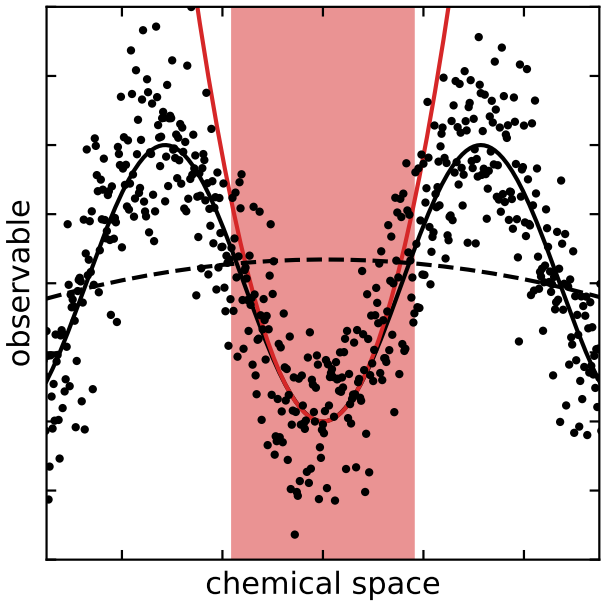

Figure 4.3: Illustration of model inadequacy for synthetic data. The black solid curve is set to be the underlying model. The distance between two data points along the abscissa is thought to be inversely proportional to the similarity of the corresponding molecular structures. Here, the approximate model is a quadratic function. If the (noisy) reference data (dots) are spread across the entire domain of chemical space shown, we will observe a systematic deviation of the observable from our approximate model (dashed line). However, if we choose a specific domain of application (red shaded area), our approximate model (red curve) will be a good approximation to the underlying model. To avoid model inadequacy, in this case, we can either improve our model by increasing its complexity or reduce the domain of application (to the red region).

corrected). While increased parameter uncertainty seems to be clearly favorable over model improvement when it comes to reliably estimating prediction uncertainty for any observable obtained on the basis of a given computational model, it does not resolve the issue of model inadequacy per se. For instance, in multiscale modeling where the target observable is built on a hierarchy of other observables, all uncertainties inferred at low levels will propagate to the final prediction uncertainty. Consequently, increasing parameter uncertainty at low levels can lead to a prediction uncertainty so large that no sensible conclusions can be drawn from it.

Uncertainty in the electronic energy propagates to all energy contributions based on nuclear motion, to any kind of free energy, to rate constants, and to concentration fluxes of chemical species (an incomplete but lucid list). The dependencies between these observables are partially exponential, which requires the minimization of systematic errors in the low-level observables (instead of hiding them in increased parameter uncertainty). In such cases, the only possible way to obtain reasonably small prediction uncertainties is the systematic improvement of the models.

5

Systematic Error Estimation for Chemical Reaction Networks*

For the theoretical understanding of the reactivity of complex chemical systems, accurate relative energies between intermediates and transition states are required. Despite its popularity, DFT often fails to provide sufficiently accurate results, especially for molecules containing transition metals. In Chapter 3, vast reaction networks were generated and a large number of intermediates needed to be studied. To date, DFT is the only method that is computationally feasible for explorations of this depth. In this Chapter, we introduce a Bayesian framework for DFT that allows for system-specific error estimation of calculated properties. We demonstrate our approach with systems already studied in this thesis: catalytic nitrogen fixation and the formose reaction.

5.1 Canonical Approach to Density Functional Assessment

Most approximate exchange–correlation (XC) density functionals are constructed by fitting their parameters to benchmark data sets. While many extensive data sets exist, such as the ones proposed by Pople,[198,199,223,224] Truhlar,[197,201,251–259] and Grimme,[260–262] studies have shown that the accuracy of XC functionals can be strongly system-dependent,[187,198–203] which, naturally, will become more severe for

* This Chapter is reproduced in part with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2016**, *12*, 2762–2773 and J. Proppe, T. Husch, G. N. Simm, M. Reiher, *Faraday Discuss.* **2016**, *195*, 497–520. Copyright 2016 American Chemical Society and Royal Society of Chemistry.

short-lived reactive intermediates. In Chapter 4, we elucidated how the accuracy reported in benchmark studies is not necessarily transferable to a specific system under consideration. It is common practice[263] (see also benchmark studies such as the one in Ref. 264) to investigate the spread of results from a selection of present-day density functionals to estimate the sensitivity of the investigated property with respect to functional form and choice of parameters. But as the selection of functionals is in parts arbitrary, this approach is highly unsystematic and the spread has no statistical significance. Therefore, a systematic framework for the assessment of the accuracy of density functionals is required.

In 2005, Nørskov, Sethna, Jacobsen, and co-workers presented a scheme for systematic error estimation of DFT results[244] based on Bayesian statistics (see also Refs. 245–247).[265,266] In their approach, an ensemble of XC functionals is generated by which a mean and a variance can be assigned to each computational result. Two types of density functionals were designed within this framework: BEEF-vdW[248] and mBEEF.[249,250] While both functionals were parameterized employing a wide range of data sets, transition metal complexes were not included and also transferability issues remain (especially for such complexes). In addition, BEEF-vdW and mBEEF are both pure functionals, whereas, it is well known that hybrid functionals tend to be more accurate than pure functionals (see, e.g., Refs. 187,203). Along these lines, Zabaras and coworkers[267] developed a new XC functional employing a Bayesian approach combined with machine learning to predict bulk properties of transition metals and monovalent semiconductors. Very recently, Vlachos and coworkers successfully applied Bayesian statistics to DFT reaction rates on surfaces.[268] However, so far the application of Bayesian statistics in DFT has been limited to solid-state and surface chemistry.[269]

Here, we develop Bayesian error estimation for molecules. It is one goal of this study to obtain a class of hybrid functionals that accurately describes the reaction energies of a specific chemical system. We advocate for a system-focused re-parameterization of our ensemble of density functionals to overcome the issue of transferability while preserving standard design principles of density functionals. Through Bayesian statistics, our class of functionals reports uncertainties for each calculated result which eliminates the arbitrariness of a system-specific parameterization.

5.2 Bayesian Error Estimation in DFT

The parameters $\mathbf{w}$ of a density functional are usually determined by parameterization to some data set $\mathcal{D} = \{(\mathbf{x}_i, \mathcal{R}(\mathbf{x}_i))\}_{i=1}^{N}$ containing molecular structures $\mathbf{x}_i$ and an observable which is determined with an (experimental or computational) reference method $\mathcal{R}$ (with the exception of those fixed by exact DFT conditions). This is accomplished

by minimizing a cost function $C(\mathbf{w})$ to obtain a best fit $\mathbf{w}_0$, which is then reported. However, information on the neighborhood of $C(\mathbf{w}_0)$ is thereby lost. For instance, it cannot be determined if the reported minimum is shallow or steep (see Ref. 216) or how perturbations in the parameter space (e.g., due to a new item in the data set) translate into variations of some observable $\mathcal{O}$.

Instead of considering only the best-fit parameters, one can assign a conditional probability distribution to the continuous set of parameters

$$p_w = p(w|\mathcal{O}, D) \propto \exp\left(-\frac{C(w)}{T}\right), \tag{5.1}$$

where the observable $\mathcal{O}$ is obtained from a single linear parameter w, and C denotes a cost function quadratic in w.[265,266] It can be shown[244] that the spread of this distribution is determined by the ensemble temperature $T = 2C(w_0)$ (see Eq. (5.14) below). A standard parameterization of density functionals can be considered a special case of this distribution where $T = 0$, so that $p(w|\mathcal{O}, D) = \delta(w - w_0)$.[244,246,247]

In practice, this distribution needs to be sampled for which a set of parameters $\{w_1, w_2, ..., w_K\}$ is generated. It can be shown that,[247] with a cost function quadratic in w, a Gaussian distribution $\mathcal{N}$,

$$p_w = \mathcal{N}(w_0, \sigma^2), \tag{5.2}$$

with mean w_0 and variance $\sigma^2 = T/(\partial^2 C(w)/\partial w^2|_{w_0})$ must be sampled. From the ensemble of parameters, a confidence interval can be calculated for any observable $\mathcal{O}$.[244]

5.3 Short Derivation of Bayesian Error Estimation

Consider some observable $\mathcal{O}^{\mathbf{w}}$ with parameters $\mathbf{w}$ to be calculated for some molecular system $\mathbf{x}_i$. In this work, the observable will be the energy difference between a pair of structural isomers. We now approximate a reference result $\mathcal{R}(\mathbf{x}_i)$ for system $\mathbf{x}_i$ by $\mathcal{O}^{\mathbf{w}}$ and therefore define

$$\Delta^{\mathbf{w}}(\mathbf{x}_i) = \mathcal{O}^{\mathbf{w}}(\mathbf{x}_i) - \mathcal{R}(\mathbf{x}_i). \tag{5.3}$$

We aim to find a probability distribution $p_{\mathbf{w}}$ so that, across the data set $\mathcal{D}$, the deviation of $\mathcal{O}^{\mathbf{w}}$ from $\mathcal{O}^{\mathbf{w}_0}$,

$$\delta^{\mathbf{w}}(\mathbf{x}_i) = \mathcal{O}^{\mathbf{w}}(\mathbf{x}_i) - \mathcal{O}^{\mathbf{w}_0}(\mathbf{x}_i), \tag{5.4}$$

is, on average, equal to the deviation of $\mathcal{O}^{\mathbf{w}}$ from $\mathcal{R}$, i.e.:

$$\sum_{i=1}^{N} \left\langle [\delta^{\mathbf{w}}(\mathbf{x}_i)]^2 \right\rangle_{\mathbf{w}} = \sum_{i=1}^{N} [\Delta^{\mathbf{w}_0}(\mathbf{x}_i)]^2, \tag{5.5}$$

where $\mathbf{w}_0$ is the parameter set that minimizes the cost function $C(\mathbf{w})$,

$$C(\mathbf{w}) = \sum_{i=1}^{N} [\Delta^{\mathbf{w}}(\mathbf{x}_i)]^2. \tag{5.6}$$

Defining the quadratic deviation of a parameter set $\mathbf{w}$ from the optimal set $\mathbf{w}_0$ as $F(\mathbf{w})$,

$$F(\mathbf{w}) = \sum_{i=1}^{N} [\delta^{\mathbf{w}}(\mathbf{x}_i)]^2, \tag{5.7}$$

we can write Eq. (5.5) in more compact form as

$$\langle F(\mathbf{w}) \rangle_{\mathbf{w}} = C(\mathbf{w}_0) \tag{5.8}$$

To obtain the probability distribution with the highest information entropy, we maximize the Shannon entropy of the distribution under the condition in Eq. (5.8). Introducing a fixed number K of parameter sets $\{\mathbf{w}_k\}$ and obeying that the sum over all probabilities equals one as an additional constraint, we have for the variation of the resulting Lagrangian function with respect to the probability $p_{\mathbf{w}_j}$ of one of these parameter sets $\mathbf{w}_j$

$$\frac{\partial}{\partial p_{\mathbf{w}_j}} \left(-\sum_{k=1}^{K} p_{\mathbf{w}_k} \ln(p_{\mathbf{w}_k}) - \lambda \left(C(\mathbf{w}_0) - \sum_{k=1}^{K} p_{\mathbf{w}_k} F(\mathbf{w}_k) \right) - \mu \left(1 - \sum_{k=1}^{K} p_{\mathbf{w}_k} \right) \right) \stackrel{!}{=} 0, \tag{5.9}$$

where λ and μ are Lagrange multipliers. Solving Eq. (5.9) yields the well-known relation

$$p_{\mathbf{w}_j} = \frac{\exp(-\lambda F(\mathbf{w}_j))}{\sum_{k=1}^{K} \exp(-\lambda F(\mathbf{w}_k))}. \tag{5.10}$$

To determine the Lagrange multiplier λ, we consider an observable $\mathcal{O}^w$ with a single linear parameter w,

$$\mathcal{O}^w(\mathbf{x}_i) = wA(\mathbf{x}_i) + B(\mathbf{x}_i), \tag{5.11}$$

where $A(\mathbf{x})$ and $B(\mathbf{x})$ are some functions of molecular system $\mathbf{x}$. Then $F(\mathbf{w})$ simplifies to

$$F(w) = \sum_{i=1}^{N} ((w - w_0) \cdot A(\mathbf{x}_i))^2. \tag{5.12}$$

The expectation value of $F(w)$ for the K parameters $\{w_k\}$ can be written as

$$\langle F(w) \rangle_{w_k} = \frac{\sum_{k=1}^{K} F(w_k) \exp(-\lambda F(w_k))}{\sum_{k=1}^{K} \exp(-\lambda F(w_k))}. \tag{5.13}$$

According to the equipartition theorem, each harmonic degree of freedom contributes $T/2$ to the cost (with the Boltzmann constant taken to be one), which implies for Eq. (5.8) in our single-parameter model that

$$\langle F(w)\rangle_{w_k} = C(w_0) = \frac{1}{2}T, \tag{5.14}$$

so that an expression for λ which corresponds to the inverse ensemble temperature T, can be derived.[244,246,247] Finally, the probability distribution p_w needs to be sampled. From the definition of $C(w)$ we have for a single linear parameter

$$C(w) = \sum_{i=1}^{N} \left[\mathcal{O}^w(\mathbf{x}_i) - R(\mathbf{x}_i)\right]^2 \tag{5.15}$$

$$= \sum_{i=1}^{N} \left[(wA(\mathbf{x}_i) + B(\mathbf{x}_i)) - R(\mathbf{x}_i)\right]^2 \tag{5.16}$$

and may expand $C(w)$ around $C(w_0)$

$$C(w) = C(w_0) + \frac{1}{2}\left.\frac{\partial^2 C(w)}{\partial w^2}\right|_{w_0} (w - w_0)^2 + \cdots . \tag{5.17}$$

The second derivative of $C(w)$ at the position $w = w_0$ is easy to evaluate

$$\left.\frac{\partial^2 C(w)}{\partial w^2}\right|_{w_0} = \sum_{i=1}^{N} 2A(\mathbf{x}_i)^2 \tag{5.18}$$

so that with Eq. (5.12) and Eq. (5.17) we find

$$F(w) = \frac{1}{2}\left.\frac{\partial^2 C(w)}{\partial w^2}\right|_{w_0} (w - w_0)^2. \tag{5.19}$$

From Eqs. (5.10) and (5.19), it can be seen that the probability distribution of w is a normal distribution:

$$p_w = \mathcal{N}\left(w_0, T \Big/ \left.\frac{\partial^2 C(w)}{\partial w^2}\right|_{w_0}\right) \tag{5.20}$$

$$= \mathcal{N}\left(w_0, \frac{C(w_0)}{\sum_{i=1}^{N} A(\mathbf{x}_i)^2}\right) \tag{5.21}$$

This distribution is then sampled by choosing the parameters $\{w_k\}$ of the K models

(the samples) so that a standard deviation σ for the observable $\mathcal{O}$ of system $\mathbf{x}_i$ can be calculated

$$\sigma(\mathcal{O}(\mathbf{x}_i)) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\mathcal{O}^{w_k}(\mathbf{x}_i) - \mathcal{O}^{w_0}(\mathbf{x}_i)\right)^2}. \tag{5.22}$$

K must be chosen so that $\sigma(\mathcal{O}(\mathbf{x}_i))$ is converged. The sets of linear parameters $\mathbf{w}_k$ (or w_k in the case of a single linear parameter) are obtained from computer-generated random numbers with the normal distribution in Eq. (5.21).

5.4 Exchange-Correlation Functional as Statistical Model

5.4.1 Range Separation in DFT

In this study, the parameters of the range-separated hybrid (RSH) version of the popular density functional PBE0[193–195] are considered for Bayesian error estimation for the following reasons: Firstly, exact exchange plays an important role in the description of transition metals.[187,200,203,270,271] Secondly, many issues of present-day density functionals, such as the underestimation of barriers of chemical reactions, can be attributed to the delocalization error.[272] Baer et al. showed that long-range corrected (LC) functionals appear to have resolved this issue.[273] Finally, it was observed[274–280] that the parameters in the RSH scheme are in fact system-dependent and that their adjustment can improve the functional's accuracy.

In RSH functionals,[281–286] the exchange functional is divided into short-range DFT exchange and long-range Hartree–Fock (HF) exchange by splitting the electron-electron interaction operator $1/r_{12}$:

$$\frac{1}{r_{12}} = \underbrace{\frac{1-[\alpha+\beta\cdot\mathrm{erf}(\gamma r_{12})]}{r_{12}}}_{\text{short-range}} + \underbrace{\frac{\alpha+\beta\cdot\mathrm{erf}(\gamma r_{12})}{r_{12}}}_{\text{long-range}} \tag{5.23}$$

This ansatz introduces three adjustable parameters: α, β, and the range-separation parameter γ. In the long-range corrected scheme, only two are independent since $\alpha+\beta = 1$ if the two operators on the right-hand side of Eq. (5.23) are evaluated by different energy expressions. LC-PBE0 is such a functional, where $\alpha = 0.25$, $\beta = 0.75$, and $\gamma = 0.3$ (if $\alpha = 0.25$, $\beta = 0.75$, and $\gamma = 0$, PBE0[194] is recovered). By contrast, in the Coulomb-attenuating method by Yanai et al.,[285] $\alpha = 0.19$, $\beta = 0.46$, and $\gamma = 0.33$, so that $\alpha + \beta = 0.65$. However, only for $\alpha + \beta = 1$ the potential shows the correct asymptotic behavior of $1/r_{12}$.[278]

5.4.2 Parameters in PBE Density Functional

In addition to the parameters in the LC scheme, we optimize parameters of the original PBE functional[195] to increase model flexibility. In Hartree atomic units, the correlation part of the PBE functional can be written as

$$E_c^{\mathrm{PBE}}[\rho_\uparrow, \rho_\downarrow] = \int \rho \left[\epsilon_c^{\mathrm{unif}}(r_s, \zeta) + H(r_s, \zeta, t)\right] \, d^3r, \tag{5.24}$$

with

$$H(r_s, \zeta, t) = \gamma_c \phi^3 \ln\left(1 + \frac{\beta_c}{\gamma_c} \frac{t^2 + At^4}{1 + At^2 + A^2 t^4}\right), \tag{5.25}$$

where $\rho = \rho_\uparrow + \rho_\downarrow$ is the electron density (obtained as a sum of spin-up and spin-down densities), $\epsilon_c^{\mathrm{unif}}(r_s, \zeta)$ the correlation energy per particle of the uniform electron gas, $r_s = [(4\pi/3)\rho]^{1/3}$ the local Wigner-Seitz radius, $t = |\nabla\rho|/(2\phi k_s \rho)$ the correlation density gradient, $\zeta = (\rho_\uparrow - \rho_\downarrow)/\rho$ the relative spin polarization, and $\phi = ((1+\zeta)^{2/3} + (1-\zeta)^{2/3})/2$ a spin scaling factor. The factor A is a function of ϕ and $\epsilon_c^{\mathrm{unif}}$.[195] The parameter $\beta_c = 0.066725$ is the second-order gradient expansion coefficient of the correlation energy in the high-density limit and the parameter $\gamma_c = (1 - \ln 2)/\pi^2$ is given by the uniform scaling to the high-density limit of the spin-unpolarized correlation energy.

The exchange part of the PBE functional is given by

$$E_x^{\mathrm{PBE}}[\rho] = \int \rho \, \epsilon_x^{\mathrm{unif}}(\rho) F_x^{\mathrm{PBE}}(s) \, d^3r, \tag{5.26}$$

where $F_x^{\mathrm{PBE}}(s) = 1 + \kappa - \kappa/(1 + \frac{\mu}{\kappa} s^2)$, $\kappa = 0.804$, and the reduced gradient $s = |\nabla\rho|/(2k_F\rho)$. The parameter κ is determined by the Lieb–Oxford bound[287] for the exchange energy, and the parameter μ is determined to satisfy the correct linear response of the spin-unpolarized uniform electron gas ($\mu = \beta_c \pi^2/3$) such that $\mu = 0.21951$.

Since its introduction, many variations of the original PBE functional were presented, such as revPBE,[288] PBEsol,[289,290] and APBE.[291] In these functionals, the functional form of PBE is kept, however, the parameters μ, β_c, and κ are varied. A study by Della Sala and coworkers[292] showed that a property-specific optimization of these parameters can lead to an increase in accuracy.

We adjust the parameters α, γ, μ, and κ to obtain a class of functionals LC*-PBE0($\mathcal{D}$) that allows us to describe a particular system of interest represented by reference data $\mathcal{D}$; for this optimization, we choose the L-BFGS-B scheme.[293] Although this system-specific parameterization is generally viewed as an illicit departure from the first-principles character of DFT toward a semi-empirical approach,[294] it is key to accurate error estimation in this work. A small number of parameters comes with the advantage that a small data set suffices for the parameterization. Being the only parameter that contributes

linearly to the total electronic energy, α is then considered in the error estimation protocol, keeping the other parameters constant at their re-optimized value. We wish to emphasize that the linearity of the energy with respect to α will only be guaranteed if the energies are calculated non-selfconsistently, i.e., employing the same electron density. We calculate the electronic energy of the ensemble non-self-consistently employing the electron density obtained from a self-consistent calculation with the best-fit parameters w_0.[248,249] Therefore, the error estimation scheme does not result in a significant computational overhead.

5.5 Construction of Appropriate Reference Data Set

For an accurate re-parameterization, the reference data set needs to be representative of the system of interest. Specifically, the data set should contain structures that are intermediates and transition states of the chemical process under consideration. Of course, one cannot expect to include every relevant structure, but the stochastic nature of our approach takes this limitation into account. Moreover, knowledge-based Bayesian statistics may even be considered in a rolling re-parameterization scheme, in which more accurate reference data are constantly added when they become available.

The observable $\mathcal{O}$ is the energy difference $\Delta E_{i,j}$ between two structural isomers $\mathbf{x}_i$ and $\mathbf{x}_j$. Then, the cost function C employed in the parameterization reads

$$C(\alpha,\gamma,\kappa,\mu) = \sum_{i=1,i<j}^{N} \left(\Delta E_{i,j}(\alpha,\gamma,\kappa,\mu) - \Delta E_{i,j}^{\text{ref}}\right)^2 = \sum_{i=1,i<j}^{N} C_{i,j}(\alpha,\gamma,\kappa,\mu), \qquad (5.27)$$

where $\Delta E_{i,j}(\alpha,\gamma,\kappa,\mu)$ and $\Delta E_{i,j}^{\text{ref}}$ are the relative energies obtained with the LC-PBE0 functional with parameters (α, γ, κ, μ) and the reference value, respectively, and $\mathbf{x}_i$ and $\mathbf{x}_j$ are structures on the same PES.

5.6 Study of Chatt–Schrock Cycle with Error Assessment

In Section 3.8, the chemical reactivity of the catalyst synthesized by Yandulov and Schrock[14,15] is investigated. A proposed catalytic cycle for this catalyst is the Chatt–Schrock cycle,[14,15,295] in which intermediates are formed by a sequence of protonation and reduction steps (see Fig. 3.13 in Section 3.8). The acid 2,6-lutidinium (LutH) and reducing agent decamethylchromocene ($CrCp_2^*$) are the sources of protons and electrons, respectively. The energetics of this cycle were subjected to many theoretical studies.[30,179–185,296] Due to different computational setups (e.g., model catalyst, density functional, and basis sets), the results of these studies varied. In the following Section,

the LC*-PBE0(D) functional is applied to study reaction energies in the Chatt–Schrock cycle.

If only little experimental reference data exists for a chosen system, highly accurate post-HF methods, such as coupled-cluster theory, can be employed. Usually, their steep scaling of computing time with system size require the restriction to rather small model systems. For the construction of the reference data set, we chose the CCSD(T) method; i.e., $\mathcal{R}$ is CCSD(T). Moreover, a model is constructed in which the HIPT substituents are replaced by methyl groups or hydrogen atoms; in this way, the computational effort is reduced, while the first coordination sphere remains intact (see Fig. 5.1). To probe the transferability of our functional optimized on data for the (*pruned*) model system to the original complex, an intermediate (*1-armed*) model is also investigated. The resulting reference data sets, referred to as D_P and D_A, accordingly, contain energy differences between structures on the same PES, i.e., structures with the same number and type of atomic nuclei, the same number of electrons, and the same electronic spin state (see Fig. 5.2 for an example of two reference values). Details on the computational methodology can be found in Appendix A.2.1.

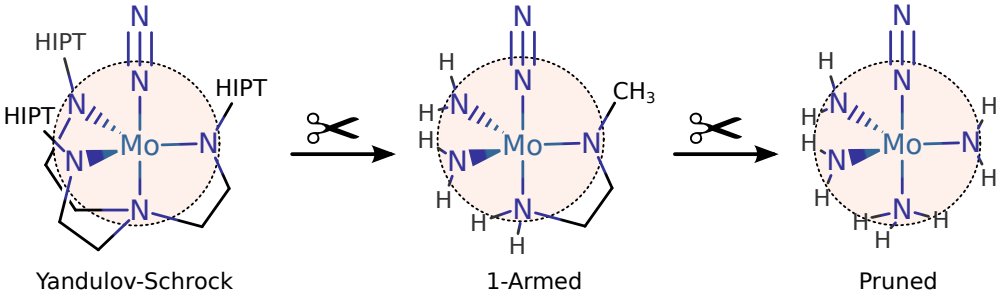

Figure 5.1: Model systems for the Yandulov–Schrock catalyst. While keeping the first coordination sphere (dashed circle) intact, carbon and hydrogen atoms are removed to reduce computational effort. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2016**, *12*, 2762–2773. Copyright 2016 American Chemical Society.

5.6.1 Parameter Selection and Optimization

Since the parameters κ and μ in the PBE functional were determined by fulfilling exact boundary conditions,[195] we first investigated whether the optimization of the parameters in the range-separation scheme, i.e., α and γ, suffices to obtain an accurate functional. Accordingly, $C_{i,j}(\alpha, \gamma, \kappa = \kappa^\mathrm{PBE}, \mu = \mu^\mathrm{PBE})$ were calculated for structures in D_P as a function of α ($\beta = 1 - \alpha$) and γ, where κ and μ were kept constant. As an example, the results for two relative energies between three isomers of [Mo]-NH_2^+ are shown in Fig. 5.3. Results for additional structures are given in the supporting information of Ref. 297. Even though the three structures are similar (differing in the position of only

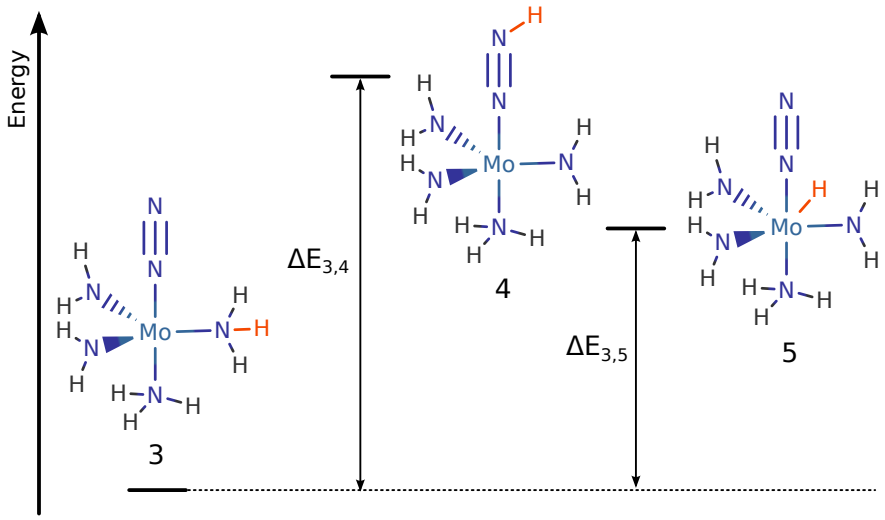

Figure 5.2: Example for relative energies $\Delta E_{3,4}$ and $\Delta E_{3,5}$ between three isomers (structures 3, 4, and 5 in D_{P}) of the pruned Yandulov–Schrock complex. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2016**, *12*, 2762–2773. Copyright 2016 American Chemical Society.

one hydrogen atom), the optimal parameters deviate significantly (as can be seen from Fig. 5.3). We note that the shape of the contour plot would not change significantly for a shifted reference energy $\Delta E_i^{\mathrm{ref}}$. A slightly different reference energy would only result in a shift of the observed pattern. Hence, it is not decisive for this study whether or not our coupled-cluster reference data is of ultimate accuracy.

Furthermore, we investigated whether incomplete LC, i.e., $\alpha + \beta < 1$, can increase model flexibility. In Fig. 5.4, the amount of LC, $\zeta = \alpha + \beta$, is varied for the cost function $C_{8,11}$. It can be seen that the form of the contour plot is hardly affected by ζ; only the curvature of the contour lines increases. This can be understood when appreciating that the effect of γ increases with ζ (see Eq. (5.23)). Therefore, we consider it unlikely that changing the amount of LC leads to an increase in accuracy worth compromising the correct asymptotic behavior. For the rest of this study, we therefore preserve complete LC, i.e., $\alpha + \beta = 1$.

To investigate whether the adjustment of κ and μ, in addition to α and γ, results in a significant increase in accuracy, the cost functions $C_{23,24}$ and $C_{23,25}$ depending on α, γ, κ, and μ are given in Figs. 5.5 and 5.6 (results for additional structures are given in the supporting information of Ref. 297). In each contour plot, the cost function depending on κ and μ is given, whereby α and γ are varied between contour plots. Note that β_c in the PBE functional depends on μ, $\beta_c = 3\mu/\pi^2$. By comparing Figs. 5.5 and 5.6, we see that for $\alpha = 0.2$ and $\gamma = 0.0$ the cost functions are similar. From this result, we conclude that the optimization of the parameters κ and μ, in addition to α and γ, is necessary to obtain a sufficiently flexible LC-PBE0 functional.

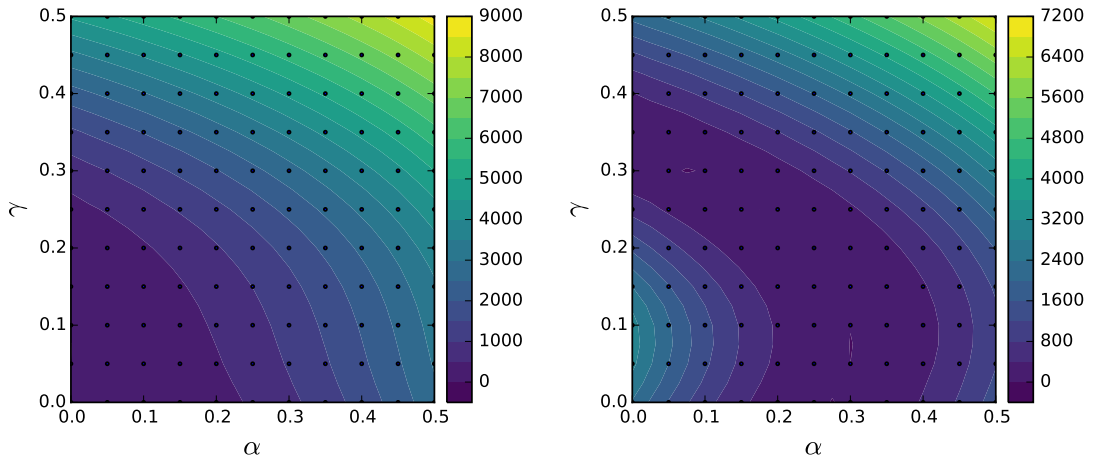

Figure 5.3: Two cost functions, $C_{26,27}$ (left) and $C_{26,28}$ (right), depending on the parameters α and γ (in (kJ/mol)^2). The cost functions were calculated from the relative energies between three isomers of [Mo]-NH_2^+. The parameters $\kappa = \kappa^{PBE}$ and $\mu = \mu^{PBE}$ were kept constant. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2016**, *12*, 2762–2773. Copyright 2016 American Chemical Society.

The four parameters were optimized employing the D_P reference set and the following parameter values were obtained: $\alpha = 0.176$ ($\sigma = 0.080$), $\gamma = 0.111$, $\kappa = 1.48$, and $\mu = 0.471$. The functional with these parameters we refer to as LC*-PBE0(D_P), where the star indicates that the original parameters were modified and 'D_P' denotes that these parameter changes were made for the D_P reference data set. All parameters clearly differ from the ones in LC-PBE0. While the parameters κ and μ were determined by fulfilling exact boundary conditions,[195] the behavior of the functional between those boundary conditions may still be incorrect. Hence, deviations from the exact parameters can lead to a functional that is more accurate for the chemical system of interest than LC-PBE0. We emphasize that our LC*-PBE0 functional is system-dependent in such a way that its optimum parameters will be different for different reference data sets. However, this is not a drawback as the reliability of this class of functionals can be assessed according to an error measure for each individual result in the error estimation procedure.

5.6.2 Assessment of Re-parameterization and Error Estimation

Before we consider the conceptually decisive error estimation step for our system-dependent functionals, we first demonstrate that they, in fact, achieve a significant improvement with respect to accuracy for the reference data set. While one might expect that this is naturally the case, it is not guaranteed because the explicit analytical form of the functional might not allow for such an improvement and the different reference data points might not be represented equally well by a common parameter set.

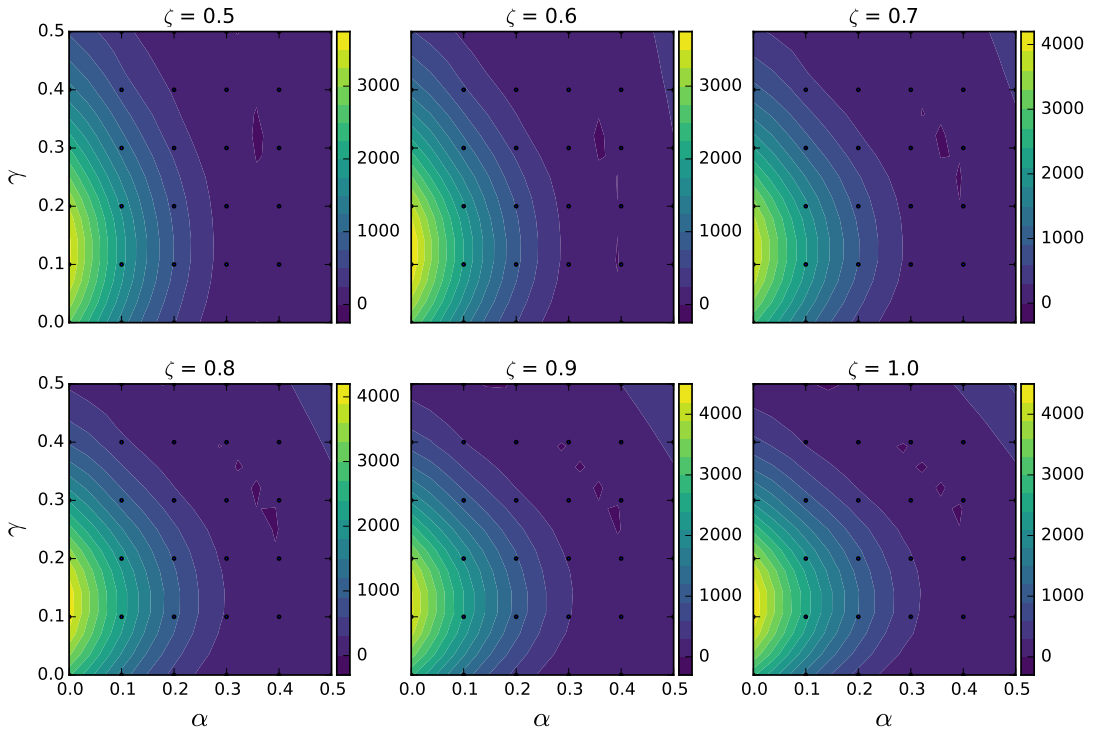

Figure 5.4: Cost function $C_{8,11}$ depending on α and γ and on the amount of long-range correction $\zeta = \alpha + \beta$ (in (kJ/mol)^2). The parameters κ and μ were kept constant at their original values in the PBE functional. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2016**, *12*, 2762–2773. Copyright 2016 American Chemical Society.

In Table 5.1, the accuracy of LC*-PBE0(D_P) is compared to that of common density functionals (including D3 dispersion corrections). LC*-PBE0(D_P) features the lowest MAE, followed by B3LYP and PBE0. As expected, GGA and meta-GGA functionals are less accurate than most hybrid functionals. Moreover, due to the small molecular size, D3 corrections have no significant effect. In addition, the MAE of no functional is within chemical accuracy and all functionals feature a high LAE of at least 25 kJ/mol. Considering LC*-PBE0(D_P) was fitted to this data set and still shows an LAE of 25.7 kJ/mol, underlines the fact that the electronic structure of transition metal complexes is difficult to reproduce by density functionals because of their restrictive functional form.

While the results in Table 5.1 confirm the well-known fact[298] that density functionals applied to transition metal complexes rarely achieve chemical accuracy of about one kcal/mol ($\approx$ 4.2 kJ/mol), it is known that DFT can be very accurate for certain cases.[264] Clearly, it is desirable to identify cases for which DFT fails and cases for which the results are reliable.

As described in Section 5.2 and 5.3, our functional allows for error estimates to be calculated. With the standard deviation σ and the best-fit parameters w_0, the

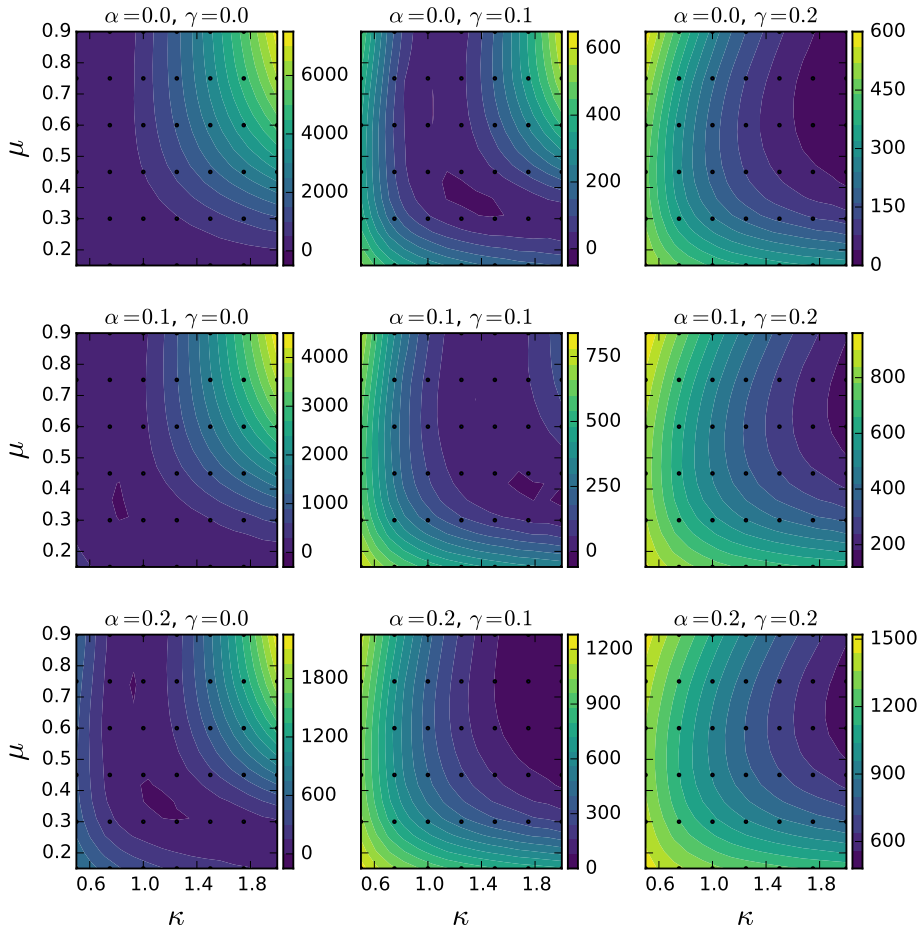

Figure 5.5: Cost function $C_{23,24}$ as a function of α, γ, κ, and μ (in (kJ/mol)^2). In each contour plot, $C_{23,24}$ is given as a function of κ and μ, whereas α and γ are kept constant. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2016**, *12*, 2762–2773. Copyright 2016 American Chemical Society.

normal distribution given in Eq. (5.2) can be sampled and a set of parameters $\vec{a} = \{w_1, w_2, \ldots, w_K\}$ can be generated (we introduce the vector notation to denote the set of parameter sets, which is a set of parameters in this special case). Employing the self-consistent electron density obtained from the functional with parameters w_0, the electronic energies for the parameters in $\vec{w}$ are calculated. The standard deviation $\sigma(\mathcal{O}(i))$ is then calculated according to Eq. (5.22). In Fig. 5.7, LC*-PBE0(D_P) (with error bars, calculated from an ensemble of $K = 25$ functionals given in Appendix A.2.1) is compared to popular density functionals with respect to D_P. It can be seen that for many elements of the data set the error with respect to the reference is within one standard deviation. For almost all reference data points the error is within two standard deviations; only for P3 and P13, the error was underestimated by LC*-PBE0(D_P).

Further, the standard deviation reported by LC*-PBE0(D_P) not always coincides

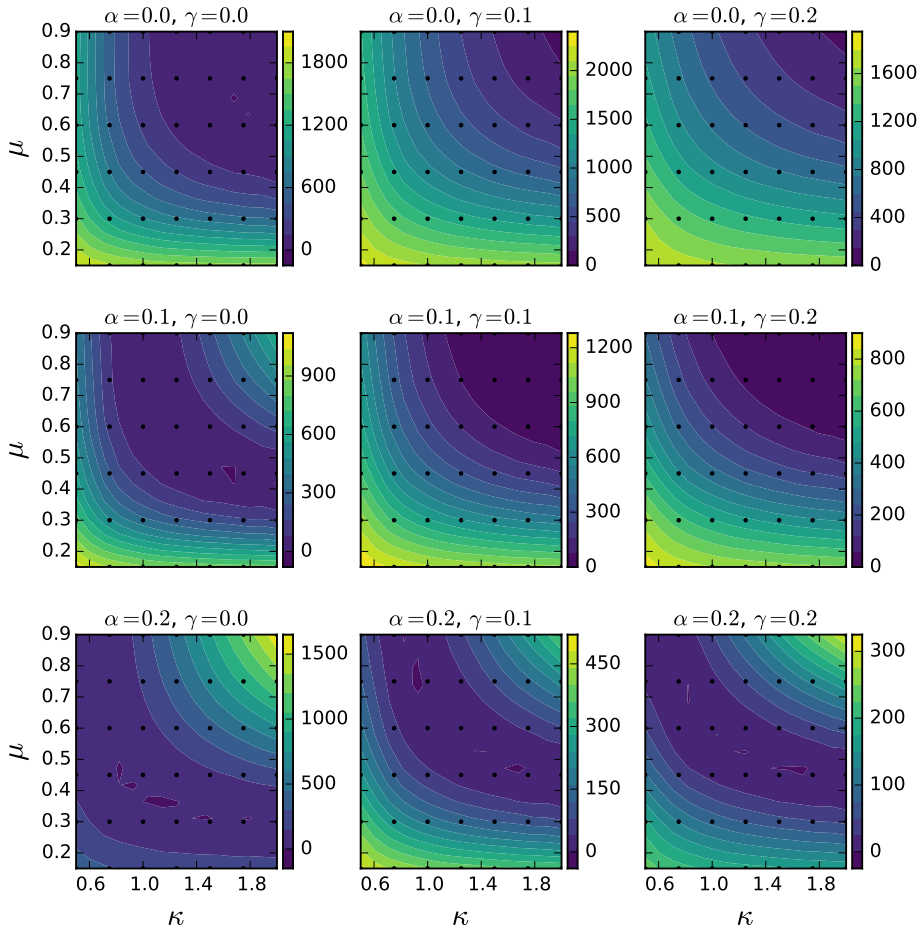

Figure 5.6: Cost function $C_{23,25}$ as a function of α, γ, κ, and μ (in (kJ/mol)^2). In each contour plot, $C_{23,25}$ is given as a function of κ and μ, whereas α and γ are kept constant. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2016**, *12*, 2762–2773. Copyright 2016 American Chemical Society.

with the spread of results from other functionals. For example, the standard deviation of P4 is comparatively small (5.2 kJ/mol), whereas the errors of the other functionals are ranging from 2–34 kJ/mol. Therefore, taking the spread of results from a set of functionals is not a stochastically meaningful indicator for the accuracy. In addition, the errors of all functionals are highly unsystematic and the spread of errors is large. This result is particularly striking when considering the fact that the structures in our data set are homologous by construction.

5.6.3 Transferability of the Model System

For the reference data set D_{P} we showed that the re-parameterization of the LC-PBE0 resulted in a significantly more accurate functional LC*-PBE0(D_{P}) that also provides reliable error estimates for each result. In this Section, we investigate the transferability

Table 5.1: Largest absolute error (LAE), mean absolute error (MAE), and mean signed error (MSE) of a selection of functionals, some with D3 dispersion corrections, for the D_P reference set (in kJ/mol).

	LAE	MAE	MSE
B3LYP	31.2	13.4	-0.1
B3LYP-D3	30.2	13.8	-0.0
BP86	65.0	33.1	-8.6
BP86-D3	66.5	35.5	-8.6
LC-PBE0	68.9	20.8	-2.3
M06-2X	69.6	28.1	4.6
M06-2X-D3	69.6	28.1	4.7
M06-L	45.7	24.7	-1.6
M06-L-D3	45.8	24.6	-1.6
PBE	66.3	32.8	-8.1
PBE0	32.3	13.6	0.1
PBE0-D3	31.6	13.8	0.3
TPSS	60.8	31.3	-7.5
TPSS-D3	62.2	32.9	-7.4
TPSSh	45.1	20.7	-4.2
TPSSh-D3	46.4	22.5	-2.7
LC*-PBE0(D_P)	25.7	10.0	-0.1

of the model system to the chemical system of interest. As shown in Fig. 5.1, the (*1-armed*) model which more closely resembles the core structure of the Yandulov–Schrock catalyst, probes the effect of the second coordination shell on the parameterization.

In Table 5.2, the accuracy of LC*-PBE0(D_P) and popular density functionals (some including D3 dispersion corrections) with respect to the data set D_A is shown. With an MAE of 8.7 kJ/mol, LC*-PBE0(D_P) is more accurate than all other standard functionals. Furthermore, due to increased system size, the contribution of the D3 corrections rose compared to D_P and has a slight positive effect on the MAE for most functionals. Finally, the strikingly high LAE of density functionals with a reasonable MAE (e.g., B3LYP-D3), highlights the need for a method with error estimation.

To investigate the effect of the model system on the parameterization, the parameters of LC*-PBE0 were optimized for D_A to yield LC*-PBE0(D_A). The obtained optimal parameters are: $\alpha = 0.128$ ($\sigma = 0.081$), $\gamma = 0.080$, $\kappa = 1.49$, and $\mu = 0.512$. In comparison to the parameters of LC*-PBE0(D_P), only α and γ changed, whereas κ and μ remained more or less the same. From Table 5.2, it can be seen that also the LAE and MAE decreased only slightly compared to LC*-PBE0(D_P). This suggests that it is the flexibility of the functional and not the choice of the model system that limits its accuracy.

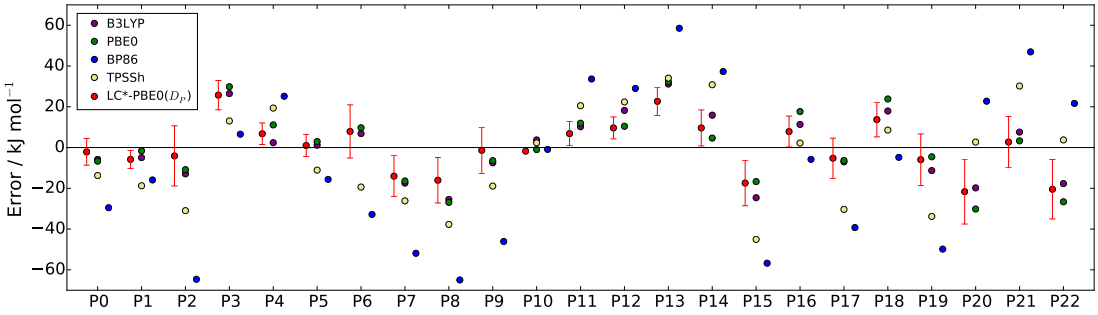

Figure 5.7: Errors of LC*-PBE0(D_P) with error bars indicating a standard deviation and standard functionals for the reference data set D_P. All data points in the set are denoted as Pi. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2016**, *12*, 2762–2773. Copyright 2016 American Chemical Society.

In Fig. 5.8, the errors of LC*-PBE0(D_P), LC*-PBE0(D_A), and standard density functionals with respect to D_A are shown. It can be seen that the error bars reported by both error estimation functionals give a reliable and consistent indication for the accuracy of a result: in nearly all cases the actual error is within two standard deviations.

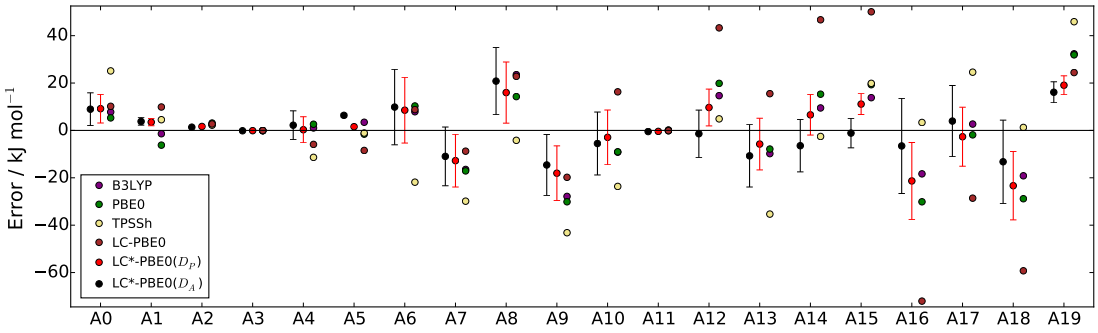

Figure 5.8: Errors of LC*-PBE0(D_P) and LC*-PBE0(D_A) (with error bars indicating ± 1 standard deviation) and standard functionals for data set D_A. All data points in the set are denoted as Ai. Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2016**, *12*, 2762–2773. Copyright 2016 American Chemical Society.

5.6.4 Error Estimation Applied to the Chatt–Schrock Cycle

In Table 5.3, the calculated reaction energies for the complete Chatt–Schrock cycle including standard deviations are given. While the majority of reactions features a small standard deviation of below 6 kJ/mol, there are reactions for which the functional predicts an unacceptably large error. For example, with a standard deviation of 18.7 kJ/mol the reaction energy of the first protonation is apparently difficult to determine, whereas LC*-PBE0(D_A) reports a low uncertainty for subsequent protonation reactions.

Table 5.2: Largest absolute error (LAE), mean absolute error (MAE), and mean signed error (MSE) of a selection of functionals, some with D3 dispersion corrections, for the D_A reference set (in kJ/mol).

	LAE	MAE	MSE
B3LYP	32.3	11.1	0.8
B3LYP-D3	28.1	10.2	1.8
BP86	70.1	24.8	-8.1
BP86-D3	68.0	25.8	-7.0
LC-PBE0	72.1	22.7	2.4
M06-2X	71.1	25.9	6.1
M06-2X-D3	71.1	25.8	6.0
M06-L	50.2	17.5	-5.0
M06-L-D3	50.1	17.6	-5.0
PBE	71.5	24.5	-8.7
PBE0	31.9	12.7	-0.6
PBE0-D3	29.7	12.0	0.0
TPSS	58.7	24.4	-5.0
TPSS-D3	56.8	25.1	-4.2
TPSSh	45.9	15.2	-2.1
TPSSh-D3	42.8	15.6	-1.3
LC*-PBE0(D_P)	23.3	8.7	0.0
LC*-PBE0(D_A)	20.8	7.2	0.1

Since the parameters in LC*-PBE0(D_A) were optimized for a data set which contains neither the reducing agent $CrCp^*_2$ nor the acid lutidinium, no error can be calculated for either the oxidation of $CrCp^*_2$ or for the abstraction of the proton from lutidinium. A more extensive data set needs to be constructed to be able to assign an uncertainty to these reactions. Therefore, we may anticipate that the errors reported here underestimate the actual errors. Since, however, the error of electron and proton abstraction would result in a constant shift for the reduction and protonation reactions, respectively, it does not affect our conclusions.

Due to the large HIPT substituents, calculations on the full Chatt–Schrock catalyst require dispersion corrections to be considered. These cannot be well described by LC*-PBE0(D_A) because D_A does not contain reference data on large model complexes for which dispersion is increasingly important. However, since no heptane solvent molecules are included in our Yandulov–Schrock structural models, dispersion corrections are not considered here as they would artificially overestimate all intra-complex dispersion. Clearly, in general, dispersion corrections must be considered. As empirical force-field-type dispersion corrections would require an extensive parameterization,

Table 5.3: LC*-PBE0(D_A) reaction energies (with standard deviations) for the first and second half of the full Chatt–Schrock cycle (in kJ/mol). $LutH^+$ and $CrCp_2^*$ are abbreviated as AH^+ and R, respectively.

Reaction	ΔE	σ
[Mo]-N_2 + AH^+ → {[Mo]-N_2H}$^+$ + A	27.8	18.7
{[Mo]-N_2H}$^+$ + R → [Mo]-N_2H + R^+	-120.9	5.9
[Mo]-N_2H + AH^+ → {[Mo]-N_2H_2}$^+$ + A	-103.4	2.6
{[Mo]-N_2H_2}$^+$ + R → [Mo]-N_2H_2 + R^+	21.8	10.6
[Mo]-N_2H_2 + AH^+ → {[Mo]-N_2H_3}$^+$ + A	-40.0	6.1
{[Mo]-N_2H_3}$^+$ + R → [Mo]-N_2H_3 + R^+	-237.7	5.3
[Mo]-N + AH^+ → {[Mo]-NH}$^+$ + A	-74.4	5.4
{[Mo]-NH}$^+$ + R → [Mo]-NH + R^+	0.2	10.5
[Mo]-NH + AH^+ → {[Mo]-NH_2}$^+$ + A	-151.8	1.7
{[Mo]-NH_2}$^+$ + R → [Mo]-NH_2 + R^+	-22.7	15.1
[Mo]-NH_2 + AH^+ → {[Mo]-NH_3}$^+$ + A	-146.7	1.1
{[Mo]-NH_3}$^+$ + R → [Mo]-NH_3 + R^+	9.4	3.0
[Mo]-NH_3 + N_2 → [Mo]-N_2 + NH_3	-7.6	13.6

we recommend density-based techniques (see, e.g., Refs. 299,300) for a system-focused density functional optimization.

In Fig. 5.9, the mean energy profile (red) together with the ensemble of LC*-PBE0(D_A) (gray) is depicted. The uncertainty associated with the energy of each intermediate with respect to the first intermediate of the cycle can be seen from the spread of the energy profiles. Similarly, a change in the spread of the energy profiles resembles the error of each reaction energy. Fig. 5.9 highlights the importance of error estimation when interpreting reaction profiles commonly found in the literature.

5.7 Error Estimation for Reaction Network of Formose Reaction

In Section 3.7, we explored the vast reaction network of the formose reaction. Furthermore, we could identify multiple pathways in the reaction network that rationalize the autocatalytic properties of this reaction. In addition, we showed that there can exist many minimum-energy paths with different reaction barriers for the same chemical transformation. In the following Section, the LC*-PBE0 functional is applied to the formose reaction to investigate how uncertainties in reaction barriers affect conclusions drawn from kinetic models.

5.7.1 Assessment of Re-parameterization and Error Estimation

For an accurate reparameterization, the reference data set $\mathcal{D}$ needs to be representative of the system to be studied. In this study, $\mathcal{D}$ contains structures of intermediates and

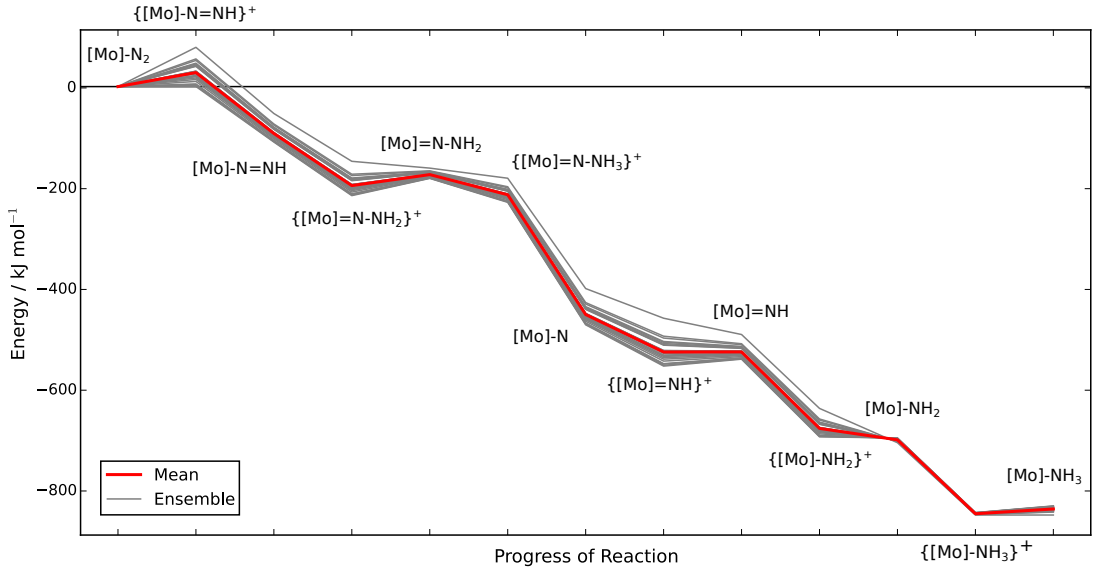

Figure 5.9: Statistical representation of energy profile of Chatt-Schrock cycle. Red: mean of LC*-PBE0(D_A); gray: ensemble of LC*-PBE0(D_A); Reprinted with permission from G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2016**, *12*, 2762–2773. Copyright 2016 American Chemical Society.

transition states of the formose reaction. Electronic energies from the DF-LCCSD(T0)-F12 method are chosen as reference $\mathcal{R}$. To assess the transferability of the reparameterized functional, the data set $\mathcal{D}$ was arbitrarily split into a training test and a test set with 25 and 17 entries, respectively. By minimizing C with respect to the training set with the L-BFGS-B algorithm,[301] a new set of parameter values for the LC*-PBE0 functional were obtained. Details on the computational methodology are provided in Appendix A.2.2.

In Tables 5.4 and 5.5, the accuracy of LC*-PBE0 (in comparison to standard functionals) with respect to the training and test sets is given. It can be seen that LC*-PBE0 is significantly more accurate than most standard functionals considered here. The optimized parameters of LC*-PBE0 are close to those of PBE0 (see Appendix A.2.2), which explains why the functionals are of similar accuracy. Due to its additional parameters and, therefore, higher flexibility, LC-PBE0 was chosen over PBE0 for the reparameterization. Nonetheless, with a largest absolute error between 8–10 kJ/mol, it is clear that error estimation is still necessary.

In Figs. 5.10 and 5.11, LC*-PBE0 is compared to contemporary density functionals with respect to the training and test sets, respectively. For both data sets, we observe that the error is at least within ±4.2 kJ/mol (≈ 1 kcal/mol), unless the error estimate reported by the functional indicates otherwise (i.e., $\sigma > 4.2$ kJ/mol). It can be seen that there are several relative energies for which the errors are underestimated (D2, D4, and D25 in the training set and D30 and D38 in the test set). This indicates that the

Table 5.4: Largest absolute error (LAE), mean absolute error (MAE), and mean signed error (MSE) of a selection of functionals, some with D3 dispersion corrections, for the training set (in kJ/mol).

	LAE	MAE	MSE
B3LYP	18.7	7.6	1.7
B3LYP-D3	22.2	7.0	1.6
BP86	28.5	7.3	1.8
BP86-D3	32.6	6.5	1.7
LC-PBE0	37.2	13.6	0.7
M06-2X	20.9	7.5	1.2
M06-2X-D3	20.8	7.4	1.2
M06-L	19.4	9.6	2.1
M06-L-D3	19.5	9.6	2.1
PBE	28.8	6.2	1.6
PBE0	13.5	5.7	1.2
PBE0-D3	16.3	5.2	1.2
TPSS	37.3	14.3	3.4
TPSS-D3	33.2	13.9	3.3
TPSSh	32.3	13.6	3.0
TPSSh-D3	29.2	13.1	2.9
LC*-PBE0	9.8	3.7	1.0

density functional severely suffers from model inadequacy (see Section 4.3.2). The poor performance of standard functionals (see Tables 5.4 and 5.5) supports this hypothesis. In addition, the domain of application is considerably larger than the one covered by the catalytic cycle of the Yandulov–Schrock catalyst (see Section 5.6.3). The data set in this study does not only contain a more diverse set of molecular structures (containing different functional groups) but also both transition states and intermediates. As a result, the error estimates provided here have to be interpreted as lower bounds on the error.

5.7.2 Kinetic Modeling

A reaction network containing all relevant intermediates and TSs of a chemical reaction allows one to study population trajectories through the network. In solution chemistry, trajectories of molar concentrations can be studied, but it remains a challenge to rationalize why certain product distributions were found. A theoretical, time-resolved model would allow one to identify elementary steps that are responsible for the observed product distribution and to develop strategies to promote the selective formation of the desired product and to suppress the formation of undesired side products. For the

Table 5.5: Largest absolute error (LAE), mean absolute error (MAE), and mean signed error (MSE) of a selection of functionals, some with D3 dispersion corrections, for the test set (in kJ/mol).

	LAE	MAE	MSE
B3LYP	14.7	6.0	-0.1
B3LYP-D3	20.0	6.4	0.8
BP86	19.6	6.7	0.4
BP86-D3	25.0	7.8	1.5
LC-PBE0	27.5	8.4	-1.1
M06-2X	12.0	4.7	-0.1
M06-2X-D3	12.0	4.7	-0.1
M06-L	20.0	7.5	1.5
M06-L-D3	20.3	7.6	1.5
PBE	19.9	6.4	0.7
PBE0	11.9	4.0	-0.1
PBE0-D3	14.7	4.0	0.5
TPSS	16.6	6.4	-1.0
TPSS-D3	17.6	7.4	-0.3
TPSSh	15.0	5.4	-1.2
TPSSh-D3	15.4	6.2	-0.4
LC*-PBE0	8.0	2.7	0.1

construction of a kinetic model, rate constants are necessary. Conventional TS theory allows one to estimate the rate constant k of an isothermal process

$$k = \frac{k_\mathrm{B} T}{h} \exp\left(-\frac{\Delta A^\ddagger}{RT}\right), \tag{5.28}$$

where k_B is the Boltzmann constant, R the ideal gas constant, h the Planck constant, $\Delta A^\ddagger$ the Helmholtz free energy difference between reactant and TS, and T the temperature. From the rate constants calculated in Eq. (5.28), differential equations describing the time propagation of the concentrations of all chemical species can be constructed. By integrating these differential equations, the underlying chemical process can be modeled. Since these differential equations are generally coupled, analytical integration becomes intractable. Therefore, numerical integration is a popular choice for solving them. However, numerical integration will become inefficient,[302] if the underlying process spans multiple timescales. For this purpose, a variety of approaches exists that simplify kinetic models.[303] Here, the kinetic simulation algorithm is based on Markov state models[304,305] and computational singular perturbation.[306,307] By separating fast from slow processes, the issue of large span of timescales can be overcome. Details on the kinetic algorithm can be found in Ref. 127.

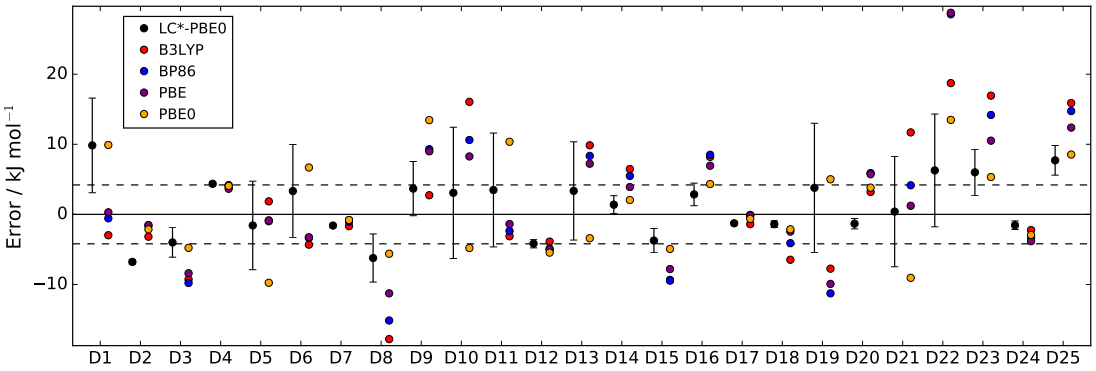

Figure 5.10: Errors of LC*-PBE0 (with error bars indicating $\pm\sigma$) and several standard functionals with respect to the training set (D1–D25). The dashed lines indicate an error of ± 4.2 kJ/mol. Reprinted with permission from J. Proppe, T. Husch, G. N. Simm, M. Reiher, *Faraday Discuss.* **2016**, *195*, 497–520. Copyright 2016 Royal Society of Chemistry.

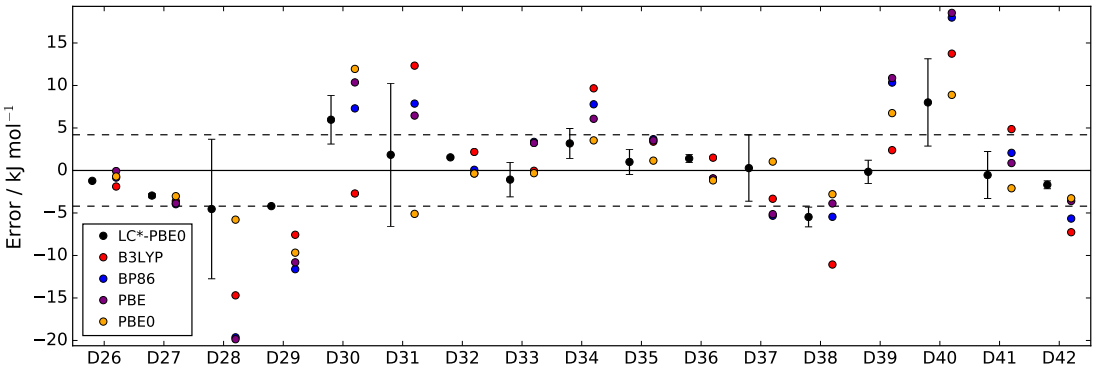

Figure 5.11: Errors of LC*-PBE0 (with error bars indicating $\pm\sigma$) and several standard functionals with respect to the test set (D26–D42). The dashed lines indicate an error of ± 4.2 kJ/mol. Reprinted with permission from J. Proppe, T. Husch, G. N. Simm, M. Reiher, *Faraday Discuss.* **2016**, *195*, 497–520. Copyright 2016 Royal Society of Chemistry.

The formose reaction is an example of a process that spans multiple timescales. The reaction network explored with Chemoton (see Section 3.7.1) would be outside the scope of this study, only a subnetwork of the formose reaction is investigated here. The structure coordinates of the intermediates and TSs are adapted from Ref. 170 (see Appendix A.2.2). This subnetwork, which already features many conceptual challenges of the entire formose reaction, is shown in Figure 5.12. This network contains the first steps of the formose reaction as described by Kua *et al.*[170] and comprises six chemical species and five reaction pairs. Free energies were obtained as described in Appendix A.2.2. In water, formaldehyde (**1**) is in equilibrium with its hydrated form (**2**). **1** dimerizes to glycolaldehyde (**3**), a high free energy of activation (see Table 5.6). The exact mechanism of the dimerization is not well-understood.[166,308–310] Experimental studies showed that this process proceeds slowly. **3** reacts with water to form 1,1,2-

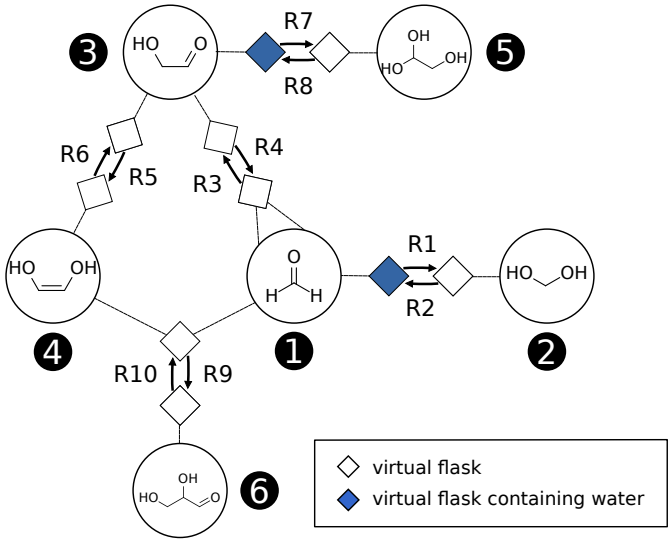

Figure 5.12: Reaction subnetwork of the formose reaction.

ethanetriol (**5**). In addition, **3** can enolize to form 1,2-ethenediol (**4**). The addition of **1** to **4** yields glyceraldehyde (**6**). This bimolecular reaction results in significant entanglement in the model network. It should be noted that this model network does not capture the autocatalytic nature of the formose reaction discussed in Section 3.7.3.

In Table 5.6, standard-state Helmholtz free activation energies, $\Delta A^{\ddagger,*}$, and the resulting rate constants k (together with error estimates calculated according to Eq. (5.22)) for the reactions in the model network are presented. For computational details see Appendix A.2.2. It can be seen that $\Delta A^{\ddagger,*}$ is large (above 100 kJ/mol) for most reac-

Table 5.6: Helmholtz free energies of activation $\Delta A^{\ddagger,*}$ (in kJ/mol, with error estimates) and rate constants k (in 1/s and 1/(s mol) for unimolecular and bimolecular reactions, respectively) for the reactions in the network.

	Reactant(s)	Product(s)	$\Delta A^{\ddagger,*}$	$\sigma_{\Delta A^{\ddagger,*}}$	k
R1	1	2	95.4	4.8	6.7×10^{-3}
R2	2	1	124.9	13.2	8.1×10^{-10}
R3	1 + 1	3	215.4	14.2	1.2×10^{-25}
R4	3	1 + 1	311.1	23.0	1.9×10^{-42}
R5	3	4	157.3	11.6	1.7×10^{-15}
R6	4	3	130.8	10.2	7.5×10^{-11}
R7	3	5	100.3	3.2	9.2×10^{-4}
R8	5	3	119.2	12.3	8.0×10^{-9}
R9	1 + 4	6	112.5	13.4	1.2×10^{-7}
R10	6	1 + 4	185.4	23.1	2.0×10^{-20}

tions, and consequently, the reaction rates are small. In addition, most reactions have estimated errors of above 10 kJ/mol, which reflects the large uncertainty of the respec-

tive reaction rates. In Section 5.7.1, we showed that the LC*-PBE0 functional provides reliable error estimates above 4.2 kJ/mol. The estimated error for reaction R7 is below that and, therefore, most likely too small. For the simulation, we selected an absolute temperature of 298.15 K and a 1 M solution of formaldehyde in water as initial feed. For technical details of the kinetic modeling employed here, see Ref. 127.

From Fig. 5.13 it can be seen that even though the uncertainty in free activation energies is large, it does not affect the qualitative flux of concentrations through the network. This finding can be explained by the distinct separation of the magnitude of the free activation energies. Furthermore, the free activation energies and their uncertainties listed in Table 5.6 show that all free energies of activation are of different orders of magnitude. This scenario does not allow for an alternative reaction mechanism. In a reaction network featuring multiple reaction barriers of the same magnitude (found in enantioselective organocatalysis, for example) large uncertainties would also lead to qualitatively different results. The qualitative validity of the kinetic simulation is also underlined by the fact that in all cases, 1,1,2-ethanetriol (**5**) is the main product at chemical equilibrium. The population dominance of **5** over **3** was also found experimentally by Kua *et al.*[311] However, their calculated free energies of activation for the corresponding reaction pair (R7, R8)[170] ($\Delta G^{\ddagger,*}_{3\to5} - \Delta G^{\ddagger,*}_{5\to3} = 2.5$ kJ mol^{-1}) are very similar to each other. Since volume changes can be neglected, their Gibbs free activation energies can be directly compared to Helmholtz free activation energies calculated in this study. Our free activation energies for the reaction pair (R7, R8) differ significantly from each other ($\Delta A^{\ddagger,*}_{3\to5} - \Delta A^{\ddagger,*}_{5\to3} = -18.9$ kJ mol^{-1}). This difference can be explained by the different choice of computational methods (i.e., different density functional and solvation model). It might seem surprising that **5** is the main product in our simulation even though glyceraldehyde (**6**) is a thermodynamic sink. However, one should keep in mind that the concentration trajectory of **6** is temporally significantly populated and that the model network considered here is only a small subnetwork of the whole reaction network.

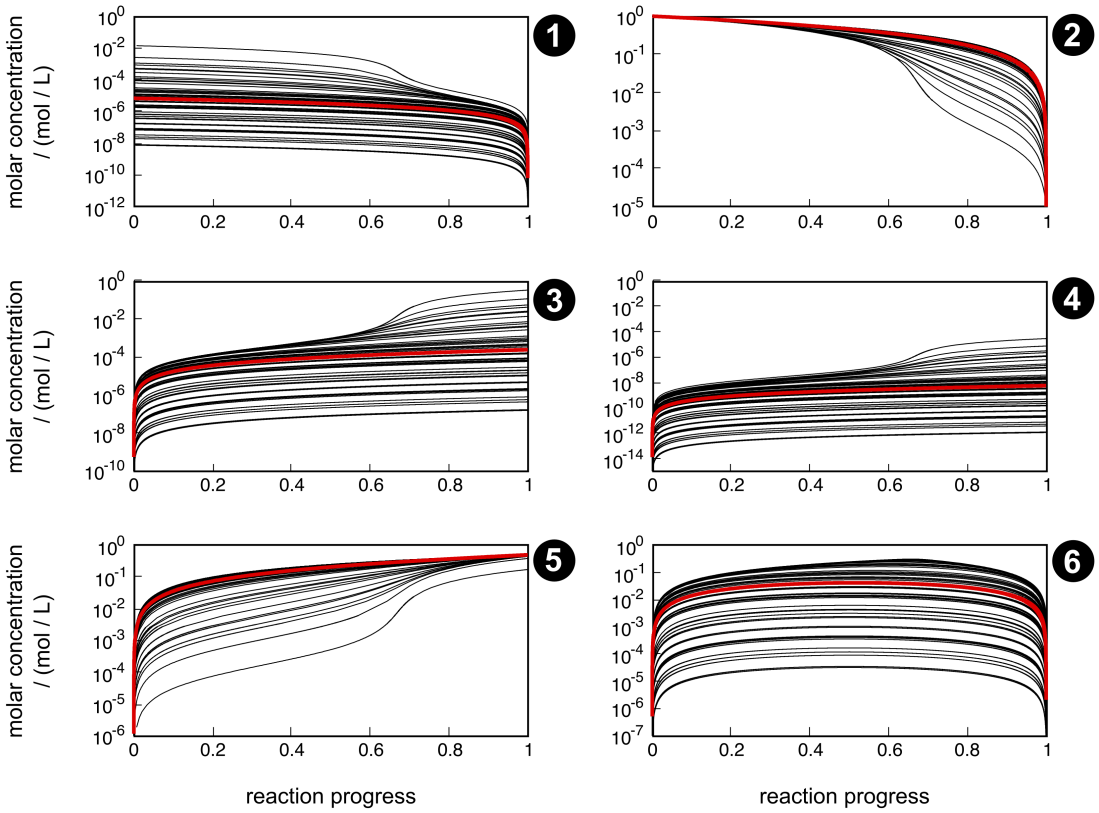

Figure 5.13: Concentration trajectories with respect to reaction progress for chemical species **1**–**6** according to the reaction network shown in Fig. 5.12. Trajectories resulting from the free activation energies of activation listed in Table 5.6 are shown in red. Trajectories resulting from the free activation energies calculated from the ensemble of density functionals generated by the Bayesian error estimation scheme are shown in black. Reprinted with permission from J. Proppe, T. Husch, G. N. Simm, M. Reiher, *Faraday Discuss.* **2016**, *195*, 497–520. Copyright 2016 Royal Society of Chemistry.

6

Error-Controlled Exploration of Chemical Reaction Networks*

In the previous Chapter, we concluded that despite the plethora of benchmark studies conducted, the accuracy of a quantum chemical method is often difficult to assess. At the same time, uncertainty quantification is absolutely mandatory for drawing meaningful conclusions from computational data. In addition, due to the limited flexibility of common density functionals, a significant improvement of a method's accuracy (e.g., through reparameterization or systematic model extension) is rarely possible. In this Chapter, we address these issues by presenting a new Bayesian approach that allows for the systematic, problem-oriented, and rolling improvement of quantum chemical results. We demonstrate our approach with the example of a complex chemical reaction network.

6.1 Application of Machine Learning in Quantum Chemistry

Over the last years, many studies on the application of statistical learning to chemistry have been published, with applications ranging from electronic structure predictions (e.g., Refs. 312–325) to applications in force-field development (e.g., Refs. 326–333), materials discovery (e.g., Refs. 334–338), and reaction prediction.[156,339–346] For recent reviews on the applications of machine learning in chemistry see Refs. 347 and 348.

De Vita, Csányi, and coworkers presented a scheme that combines *ab initio* calculation and machine-learning for molecular dynamics simulations.[349–352] Forces on atoms are either predicted by Bayesian inference or, if necessary, computed by on-the-fly quantum-

*This Chapter is reproduced in part from G. N. Simm, M. Reiher, arXiv:1805.09886.

mechanical calculations and added to a growing machine learning database.[350] However, this approach requires a considerable data set size to be accurate. So far, their approach was applied to the simulation of metal solids but not to molecular systems.

In 2017, Nørskov, Bligaard, and coworkers employed Gaussian processes (GPs) to construct a surrogate model on the fly to efficiently study surface reaction networks involving hydrocarbons.[353] The surrogate model is iteratively used to predict the rate-limiting reaction step to be calculated explicitly with DFT. In their study, extended connectivity fingerprints based on graph representations of molecules are applied to represent adsorbed species. However, if the uncertainties provided by the GP are high, then reference calculations are not automatically performed to improve the model. Therefore, the construction of the reference data set is not directly guided by the GP's predictions. Finally, their approach was applied to study surface chemistry, for which more accurate *ab initio* approaches, typically coupled-cluster methods, are not applied on a routine basis.

Despite continuous advances, most machine learning approaches are unsuitable for the study of chemical reactivity. Training data sets, which are required for the learning process of the statistical model, are commonly assembled by drawing from a predefined pool of chemical species. This approach would only be applicable to the exploration of a chemical system if the specific species had been known before (which cannot be achieved as these species are the result of the exploration process). By contrast, structure discovery through exploration requires a system-focused uncertainty quantification in order to be reliable.[297] While some machine learning methods provide error estimates for such system-focused, rolling approaches, in most studies applying statistical learning to investigate chemical systems, the focus is placed on the prediction accuracy (e.g., Refs. 312–325). In molecular applications,[354,355] confidence intervals are not exploited to define structures for which reference data should be calculated in a rolling fashion.

6.2 Gaussian Process Regression

GPs have been extensively studied by the machine learning community. They are rooted in a sophisticated and consistent theory combined with computational feasibility.[356] In chemistry, however, GPs are fairly new and, therefore, a short overview is given here. We refer the reader to Ref. 356 for a more detailed derivation.

Supervised learning is the problem of learning input to output mappings from a training data set. We define the training data set containing N observations as $\mathcal{D} = \{(\mathbf{x}_i, y_i)|i = 1, \ldots, N\}$, where $\mathbf{x}$ is the input and y the output. From $\mathcal{D}$ we aim to learn the underlying function f, to make predictions for an unseen input $\mathbf{x}_*$, i.e., input that is not in $\mathcal{D}$. Because no function that reproduces the training data is equally valid, it

is necessary to make assumptions about the characteristics of f. With a GP, which is a stochastic *process* describing distributions over functions,[356] one includes all possible functions and assigns weights to these functions depending on how likely they are to model the underlying function.

By defining a *prior* distribution we encode our prior belief on the function that we are trying to model. The prior distribution over functions includes not only the mean and point-wise variance over the functions at a certain point $\mathbf{x}$ but also how smooth these functions are. The latter is encoded in the covariance function or *kernel* which determines how rapidly the functions should change based on a change in the input $\mathbf{x}$. The task of *learning* is finding the optimal values for the parameters in the model. The *posterior* distribution is the result of combining the prior and the knowledge that we get from $\mathcal{D}$. With a trained GP, one can make predictions on unseen input. Due to its Bayesian nature, an error estimate, indicating the model's confidence in the prediction, is provided for each prediction. Finally, the GP is systematically improvable, i.e., predictions and their error estimates improve with data set size.

6.2.1 Gaussian Process Regression – Brief Derivation

Let us consider a simple linear regression model with Gaussian noise

$$f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{w}, \qquad y = f(\mathbf{x}) + \varepsilon, \tag{6.1}$$

where $\mathbf{x}$ is a D-dimensional input vector, $\mathbf{w}$ is a vector of parameters, and y is the observed target value. The function $\boldsymbol{\phi}(\mathbf{x})$ maps a D-dimensional input vector to a D'-dimensional feature space. Moreover, we assume that the observed target value y differs from f by some noise ε, which obeys an independent and identically distributed Gaussian distribution $\mathcal{N}$ with a mean and variance σ_n^2

$$\varepsilon \sim \mathcal{N}(0, \sigma_n^2). \tag{6.2}$$

Furthermore, as our prior, we place a zero-mean Gaussian with covariance matrix Σ_p on the weights

$$\mathbf{w} \sim \mathcal{N}(0, \Sigma_p). \tag{6.3}$$

Following Bayes' rule, the posterior distribution reads

$$p(\mathbf{w}|X, \mathbf{y}) = \frac{p(\mathbf{y}|X, \mathbf{w})\, p(\mathbf{w}|X)}{p(\mathbf{y}|X)}, \tag{6.4}$$

where $X = \{\mathbf{x}_i | i = 1, \ldots, N\}$ and $\mathbf{y} = [y_1, \ldots, y_N]^\intercal$. In Eq. (6.4), the marginal likelihood, $p(\mathbf{y}|X)$, is independent of the weights and can be calculated according to

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|X, \mathbf{w}) p(\mathbf{w})\, d\mathbf{w}. \tag{6.5}$$

For some unseen $\mathbf{x}_*$, the probability distribution of $f(\mathbf{x}_*)$ is given by the following expression:

$$p(f_*|\mathbf{x}_*, X, \mathbf{y}) = \int p(f_*|\mathbf{x}_*, \mathbf{w}) p(\mathbf{w}|X, \mathbf{y})\, d\mathbf{w}. \tag{6.6}$$

This can be shown to be[356]

$$\begin{aligned} p(f_*|\mathbf{x}_*, X, \mathbf{y}) = \mathcal{N}\Big(&\boldsymbol{\phi}_*^\intercal \Sigma_p \boldsymbol{\Phi} (\boldsymbol{\Phi}^\intercal \Sigma_p \boldsymbol{\Phi} + \sigma_n^2 I)^{-1} \mathbf{y}, \\ &\boldsymbol{\phi}_*^\intercal \Sigma_p \boldsymbol{\phi}_* - \boldsymbol{\phi}_*^\intercal \Sigma_p \boldsymbol{\Phi} (\boldsymbol{\Phi}^\intercal \Sigma_p \boldsymbol{\Phi} + \sigma_n^2 I)^{-1} \boldsymbol{\Phi}^\intercal \Sigma_p \boldsymbol{\phi}_* \Big), \end{aligned} \tag{6.7}$$

where $\boldsymbol{\phi}_* = \boldsymbol{\phi}(\mathbf{x}_*)$ and $\boldsymbol{\Phi} = \boldsymbol{\Phi}(X)$ is the column-wise aggregation of $\boldsymbol{\phi}(\mathbf{x})$ for all inputs in $\mathcal{D}$. In Eq. (6.7), the feature space always enters in the form of $\boldsymbol{\phi}(\mathbf{x})^\intercal \Sigma_p \boldsymbol{\phi}(\mathbf{x}')$, where $\mathbf{x}$ and $\mathbf{x}'$ are in either the training or test set. It is useful to define the *covariance function* or *kernel* $k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^\intercal \Sigma_p \boldsymbol{\phi}(\mathbf{x}')$ and the corresponding kernel matrix $K(X, X') = \boldsymbol{\Phi}(X)^\intercal \Sigma_p \boldsymbol{\Phi}(X')$. Since the covariance matrix Σ_p is positive semidefinite, we can define $\Sigma^{1/2}$ so that $\left(\Sigma_p^{1/2}\right)^2 = \Sigma_p$. Therefore, we can write $\boldsymbol{\phi}(\mathbf{x})^\intercal \Sigma_p \boldsymbol{\phi}(\mathbf{x}')$ as an inner product $\langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle$, where $\psi(\mathbf{x}) = \Sigma_p^{1/2} \boldsymbol{\phi}(\mathbf{x})$. This is also known as the *kernel trick*, which allows one to circumvent the explicit representation of the function $\boldsymbol{\phi}$ in Eq. (6.1). Conveniently, on the basis of Mercer's theorem,[357] it suffices to verify that $k(\mathbf{x}, \mathbf{x}')$ satisfies Mercer's condition. For a more elaborate explanation see Section 4.3 in Ref. 356. Finally, the key predictive equations for a GP regression are:[356]

$$\mathbf{f}_* | X, \mathbf{y}, X_* \sim \mathcal{N}\left(\bar{\mathbf{f}}_*, \mathrm{cov}(\mathbf{f}_*)\right), \tag{6.8}$$

where

$$\bar{\mathbf{f}}_* \triangleq \mathbb{E}[\mathbf{f}_* | X, \mathbf{y}, X_*] = K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} \mathbf{y} \tag{6.9}$$

and

$$\mathrm{cov}(\mathbf{f}_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*). \tag{6.10}$$

A GP trained on $\mathcal{D}$ to make predictions on f can be employed to model functions such as:

$$g(\mathbf{x}, \mathbf{x}') = f(\mathbf{x}) - f(\mathbf{x}'). \tag{6.11}$$

The prediction mean can be readily obtained from the individual prediction means

$$\overline{g}(\mathbf{x}, \mathbf{x}') = \overline{f}(\mathbf{x}) - \overline{f}(\mathbf{x}') \tag{6.12}$$

and the prediction uncertainty can be estimated employing the individual variances and covariance $\mathrm{cov}(f(\mathbf{x}), f(\mathbf{x}'))$, which can be computed with Eq. (6.10):

$$\mathrm{cov}(g(\mathbf{x}, \mathbf{x}')) = \mathrm{cov}(f(\mathbf{x})) + \mathrm{cov}(f(\mathbf{x}')) - 2\,\mathrm{cov}(f(\mathbf{x}), f(\mathbf{x}')). \tag{6.13}$$

6.3 Molecular Kernels – Distance in Chemical Space

From Eqs. (6.9) and (6.10) it can be seen that in order to be able to apply GPs to learn a molecular target $\mathcal{T}(\mathbf{x})$ (e.g., an enthalpy of atomization), the kernel $k(\mathbf{x}, \mathbf{x}')$ needs to be evaluated. Here, $\mathbf{x}$ may be some point in chemical space, i.e., the atomic configuration, charge, and spin multiplicity. The kernel should measure the similarity between two points in chemical space and satisfy invariance properties such as translations, rotations, and permutation of atoms of the same element. The search for new kernels to encode physical invariances is a subject of active research.

If the target $\mathcal{T}(\mathbf{x})$ can be approximately decomposed as a sum of local contributions the formulation of the kernel can be simplified:

$$\mathcal{T}(\mathbf{x}) = \sum_{\ell=1}^{n} t(\tilde{x}_\ell), \tag{6.14}$$

where ℓ is an atomic index, n is the total number of atoms, and $\tilde{x}_\ell$ is a local atomic environment. This approximation can be appropriate for properties such as the energy or molecular polarizability.[358] Then, we can model $t(\tilde{x}_\ell)$ as a linear combination of abstract descriptors $\tilde{\boldsymbol{\phi}}(\tilde{x}_\ell)$ (see Eq. (6.1)):

$$\hat{t}(\tilde{x}_\ell) = \tilde{\boldsymbol{\phi}}(\tilde{x}_\ell)^\intercal \mathbf{w}. \tag{6.15}$$

In analogy to equation (6.14), we obtain

$$\hat{\mathcal{T}}(\mathbf{x}) = \sum_{\ell=1}^{n} \tilde{\boldsymbol{\phi}}(\tilde{x}_\ell)^\intercal \mathbf{w} = \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{w}, \tag{6.16}$$

where $\boldsymbol{\phi}(\mathbf{x}) = \sum_{\ell=1}^{n} \tilde{\boldsymbol{\phi}}(\tilde{x}_\ell)$ so that we recover Eq. (6.1). One can see that the kernel $k(\mathbf{x}, \mathbf{x}')$ can be written as a sum of kernels acting on local atomic environments

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^\intercal \Sigma_p \boldsymbol{\phi}(\mathbf{x}') = \sum_{\ell=1}^{n} \sum_{\ell'=1}^{n'} \tilde{k}(\tilde{x}_\ell, \tilde{x}'_{\ell'}), \tag{6.17}$$

where $\tilde{k}(\tilde{x}_\ell, \tilde{x}'_{\ell'}) = \tilde{\phi}(\tilde{x}_\ell) \sum_p \tilde{\phi}(\tilde{x}'_{\ell'})$. There are many kernels developed to act on atomic environments $\tilde{k}(\tilde{x}_\ell, \tilde{x}'_{\ell'})$, such as the kernel developed by Behler and Parrinello,[312] the Smooth Overlap of Atomic Potentials (SOAP),[359] or the Graph Approximated Energy (GRAPE).[360]

6.4 Error-Controlled Exploration Algorithm

In the exploration of a chemical reaction network, the data set $\mathcal{D}$ is not known beforehand and must be generated during the exploration for a system-focused uncertainty quantification. Naturally, the size of this data set should be related to the desired level of confidence with which the target $\mathcal{T}$ needs to be determined. Our protocol starts with an initial training data set $\mathcal{D}$ of size $m > 0$ and the desired level of confidence given by the variance σ^2_{thresh}. The initial data set consists of the first m structures $s_{1:m} = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ encountered during the exploration and the corresponding targets. This is necessary to allow for reliable predictions by the learning algorithm. However, it is critical that the initial training data set does not result in the model being overly confident. Therefore, the optimal choice of m depends on the chemical system and the exploration method. For example, if $\mathcal{D}$ had consisted of m consecutive snapshots of a molecular dynamics trajectory, m should be chosen to be larger than if it had contained largely different configurational isomers. We also note that one could construct the initial data set by sampling the configuration space employing an inexpensive method and, subsequently, applying a clustering algorithm (e.g., k-means clustering) so that the $\mathcal{D}$ consists of the centroids of the m clusters.

Subsequently, new structures $s_{m+1:N}$ (given by a list of structures here but constructed in a rolling fashion in practice) are encountered. Each structure $\mathbf{x}_i$ is fed to the GP and a prediction mean $\bar{\mathcal{T}}(\mathbf{x}_i)$ and a variance σ_i^2 are obtained. If σ_i^2 is less than σ^2_{thresh}, the prediction confidence will be sufficiently high and the next structure will be attained. If σ_i^2 is larger than σ^2_{thresh}, the prediction will be discarded and the target will be explicitly calculated (e.g., with an electronic structure reference method) for that structure. The newly obtained data point is added to $\mathcal{D}$ and the GP is retrained on the extended data set. Naturally, there is a trade-off between confidence and computational effort. If σ^2_{thresh} is decreased, the prediction confidence will be required to be higher throughout the exploration. This requires a larger data set, and hence, more reference calculations. If, however, σ^2_{thresh} is increased, fewer reference calculations are needed, but the overall prediction accuracy is lower. Next, all predictions made before are repeated with the updated GP. Through this process, which we refer to as *backtracking*, we ensure that predictions on previously encountered structures are still within the given confidence interval after the GP was updated. Our

error-controlled exploration protocol with backtracking can be summarized as:

Algorithm 1 Error-controlled exploration strategy.

Input: $\mathcal{D} = \{(\mathbf{x}_i, \mathcal{T}(\mathbf{x}_i))\}_{i=1}^{m}$, $s_{m+1:N}$, σ^2_{thresh}
  for $i \leftarrow m+1, N$ **do**
    $\bar{\mathcal{T}}(\mathbf{x}_i) \leftarrow \mathbb{E}_{GP}[\mathcal{T}(\mathbf{x}_i)|\mathcal{D}, \mathbf{x}_i]$
    $\sigma_i^2 \leftarrow \mathbb{V}_{GP}[\mathcal{T}(\mathbf{x}_i)|\mathcal{D}, \mathbf{x}_i]$
    if $\sigma_i^2 > \sigma^2_{\text{thresh}}$ **then**
      add $(\mathbf{x}_i, \mathcal{T}(\mathbf{x}_i))$ to $\mathcal{D}$
      update GP and backtrack (i.e., check $x_{j<i}$)
  return $\mathcal{D}$

6.5 Application of Exploration Algorithm to Chemical Reaction Network

6.5.1 Construction of Reaction Network

We demonstrate our error-controlled exploration strategy with the example of a subset of the GDB-9 database[361] consisting of three-dimensional molecular structures of 6095 constitutional isomers of the $C_7H_{10}O_2$ stoichiometry. We chose this database in order to adhere to a publicly available data set that promotes reproducibility and comparability of new algorithms such as the one proposed in Section 6.4 above.

We constructed a graph in which nodes represent items in this data set. Edges are placed between two nodes if their molecular graphs can be interconverted by at least one rule from a set of transformation rules. These rules describe reactions commonly found in organic chemistry including nucleophilic addition and substitution, isomerization, and cycloaddition reactions (see Appendix A.3 for details). The application of these rules divided this graph into multiple strongly connected subgraphs, the largest of which contained 1494 nodes. This subgraph will serve as an artificial exploration network for the rest of this study and is provided in the supporting information of Ref. 362. The exploration network is shown in Fig. 6.1. The color of each node represents the graph distance to some randomly chosen node in the network, i.e., the number of edges in the shortest path connecting them.

We calculated the SOAP kernel[359] $k(\mathbf{x}, \mathbf{x}')$ for every pair of structures in the data set. This measure of molecular similarity is suitable for a special class of molecular structures that we consider in this work: stable intermediates. In fact, many electronic structure methods ranging from Kohn–Sham DFT to single-reference coupled cluster models have been developed for this special type of stationary points on the Born–Oppenheimer

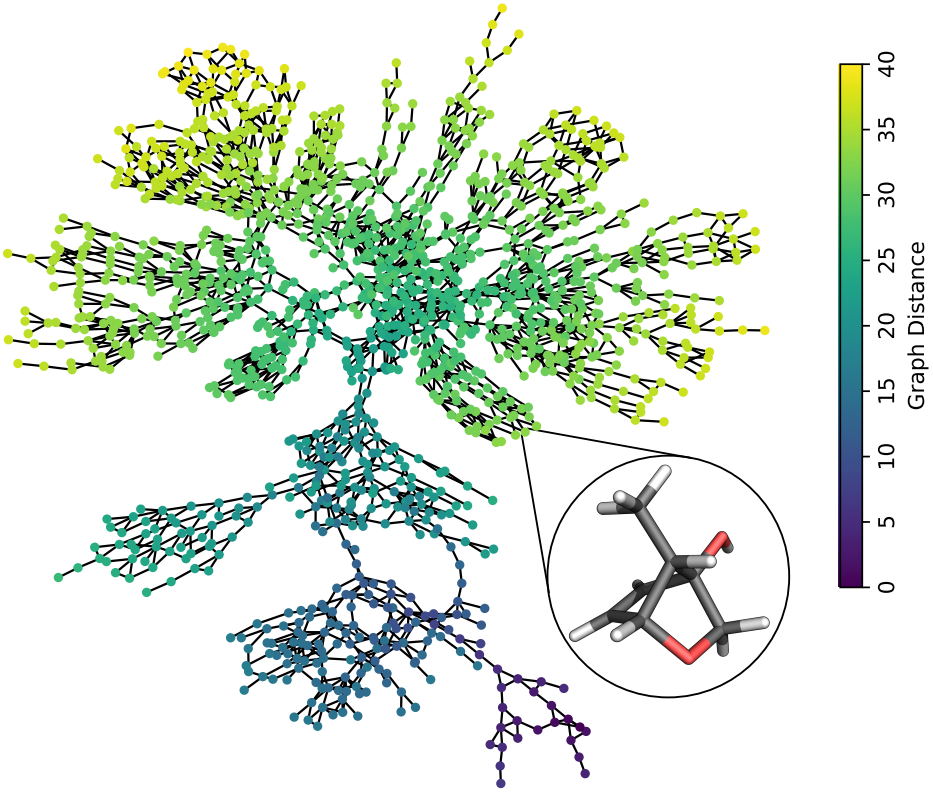

Figure 6.1: Reaction network considered in this study. Nodes represent three-dimensional molecular structures of constitutional isomers of the $C_7H_{10}O_2$ stoichiometry. Edges are drawn between two nodes if there is a transformation rule interconverting their molecular graphs. A node's color represents its graph distance to a (randomly chosen) node in the network.

potential energy hypersurface (PES). It is well-known that many of them will fail for dissociation processes (examples are wrong asymptotes of coupled cluster calculations and the Hartree–Fock dissociation error). Clearly, considering also structures away from stable intermediates would require an extension of the descriptor chosen for this work. However, such extensions are rather straightforward to define. Consider, for example, a multi-dimensional descriptor that also considers electronic structure information such as the gap between the highest occupied molecular orbital and the lowest unoccupied molecular orbital; see also the work of Kulik and coworkers.[355,363,364] Such an extension of the kernel would also improve its ability to capture long-range effects.

A special and important class of stationary points on the PES next to that of stable intermediates are transition-state structures, i.e., first-order saddle points on the PES. We would need to consider these structures in order to transgress the thermodynamic view of reaction networks and to approach kinetic modeling. Whereas this is beyond the scope of the present work, we note in passing that apart from the option to explicitly include information on the electronic structure of a given molecular structure

(which would also allow one to consider different charge and spin states), we may treat transition-state structures as a new class of structures characterized by the fact that an electronically excited state is generally closer in energy than in the case of the stable intermediate. One may, therefore, keep intermediates and transition states (and species of different charge or spin multiplicity) in separate data sets in order to best account for these different types of electronic structures (e.g., closed-shell ground-state minima, ground-state bond-activated structures with a tendency to multi-configurational nature, neutral vs. excess-charge species, and so forth).

If a set of intermediates on different PESs (but with the same charge and spin multiplicity) are encountered during the exploration, the smallest collection of atoms, from which every molecule in the set can be constructed, can be assembled. Then, upon comparison of two structures x and x' from this set with the kernel $k(\mathbf{x}, \mathbf{x}')$, the atoms that are not needed to form either of the two would still be part of the comparison, but in the form of idealized "isolated" species.[321] In this way, all comparisons between structures from this set are on equal footing.

6.5.2 Assessment of Learning and Prediction Accuracy

Calculating a thermodynamic property $P^{\text{ref}}(\mathbf{x})$ (e.g., the standard enthalpy of atomization) with accurate methods, such as G4MP2,[226] is computationally demanding. Statistical learning can be employed to improve a result of computationally (comparatively) inexpensive quantum chemical methods, $P^{\text{base}}(\mathbf{x})$, by predicting the error of a method with respect to some accurate reference result:

$$\Delta P^{\text{ref}}_{\text{base}}(\mathbf{x}) = P^{\text{ref}}(\mathbf{x}) - P^{\text{base}}(\mathbf{x}). \tag{6.18}$$

This strategy is often referred to as Δ-machine learning.[365] It is based on the idea that inexpensive quantum chemical methods are able to describe a significant portion of the underlying physics (e.g., nuclear repulsion) but fail to capture more complex phenomena such as electron correlation. It is these effects which are then learned in a Δ-machine learning approach. By design, Δ-machine learning approaches require the evaluation of the inexpensive P^{base} to arrive at the desired property.

In this work, we apply the Δ-machine learning approach by learning the difference in the calculated standard enthalpy of atomization between G4MP2 and the density-functional approach with PBE[193] ($\Delta H^{\text{G4MP2}}_{\text{PBE}}$) as well as G4MP2 and the semi-empirical model PM7[366] ($\Delta H^{\text{G4MP2}}_{\text{PM7}}$). We emphasize that the choice of the inexpensive (here, PBE and PM7) and reference (here, G4MP2) method are to a certain degree arbitrary and other choices work as well for our protocol (provided that the reference method has been demonstrated to be more accurate than the inexpensive models for the data set

under consideration). The distributions of $\Delta H_{\mathrm{PBE}}^{\mathrm{G4MP2}}$ and $\Delta H_{\mathrm{PM7}}^{\mathrm{G4MP2}}$ in the data set are shown in Fig. 6.2 (for details on the computational methodology see Appendix A.3). Due to the more approximate nature of the semi-empirical PM7 method compared to the PBE density functional, the distribution of $\Delta H_{\mathrm{PM7}}^{\mathrm{G4MP2}}$ is much wider than the one of $\Delta H_{\mathrm{PBE}}^{\mathrm{G4MP2}}$.

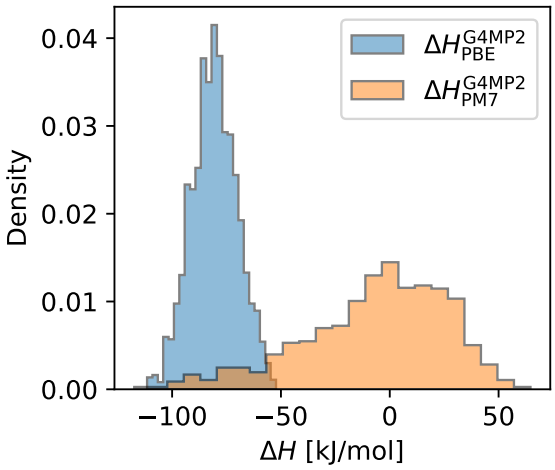

Figure 6.2: Distributions of $\Delta H_{\mathrm{PBE}}^{\mathrm{G4MP2}}$ and $\Delta H_{\mathrm{PM7}}^{\mathrm{G4MP2}}$ for the data set.

We calculate the SOAP kernel[359] $k(\mathbf{x}, \mathbf{x}')$ for every pair of structures in the data set. This kernel also provides a definition of the distance between two structures[321]

$$d(\mathbf{x}, \mathbf{x}') = \sqrt{2 - 2k(\mathbf{x}, \mathbf{x}')}. \tag{6.19}$$

To illustrate the notion of distance in a reaction network, a subnetwork of the whole reaction network is arranged according to $d(\mathbf{x}, \mathbf{x}')$ in Fig. 6.3, where $\mathbf{x}$ is some reactant and $\mathbf{x}'$ a possible product.

For both targets separately, we trained a GP on randomly selected subsets of different size and employed the remaining structures as an out-of-sample validation set. The GP's hyperparameters are optimized by maximizing the marginal likelihood. For predictions on the validation set we calculated the mean absolute error (MAE),

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\bar{\mathcal{T}}(\mathbf{x}_i) - \mathcal{T}(\mathbf{x}_i)|, \tag{6.20}$$

and root-mean-square error (RMSE),

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\bar{\mathcal{T}}(\mathbf{x}_i) - \mathcal{T}(\mathbf{x}_i)\right)^2}, \tag{6.21}$$

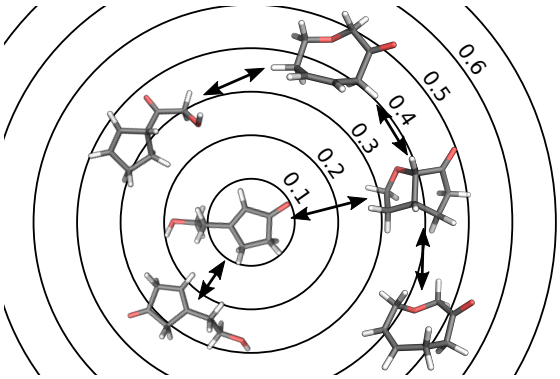

Figure 6.3: Illustration of the distance metric in Eq. (6.19) introduced by the kernel with the example of a reaction subnetwork. The contour lines represent the distance $d(\mathbf{x}, \mathbf{x}')$ between the reactant in the center ($\mathbf{x}$) and possible reaction products present in the data set ($\mathbf{x}'$). Double arrows are drawn between structures if there is a transformation rule interconverting their molecular graphs.

where N is the size of the out-of-sample validation set, $\bar{\mathcal{T}}(\mathbf{x}_i)$ the prediction mean and $\mathcal{T}(\mathbf{x}_i)$ the target value. To better assess the behavior of the GP, we also calculated the MAE (MAE_{ref}) and the RMSE (RMSE_{ref}) of a trivial statistical model that simply predicts the mean of the training data set for every test input. In addition, to guarantee the accuracy of the error estimates we calculate the percentage of predictions r_{cb} for which the target lies outside of the 95% confidence band given by $\bar{\mathcal{T}}(\mathbf{x}_i) \pm 2\sigma(\mathbf{x}_i)$. We repeated this process 25 times to ensure that the average of the above metrics converged. The average properties are summarized in Table 6.1. It can be seen that the prediction accuracy improves significantly with the size of the training data set. When comparing the MAE and the RMSE to the MAE_{ref} and the RMSE_{ref}, respectively, the benefit of employing a GP over simply predicting the average of the training data set is evident for training data set sizes of 200 and larger. It can also be seen that the prediction error of $\Delta H^{\text{G4MP2}}_{\text{PM7}}$ is larger than that of $\Delta H^{\text{G4MP2}}_{\text{PBE}}$. This can be explained by the approximate nature of the semi-empirical PM7 method (see Fig. 2). Nonetheless, the results suggest that the prediction error estimates are reliable as r_{cb} is close to 5% for all data set sizes and targets.

For the study of chemical reactivity, not enthalpies of formation but (free) enthalpy differences between intermediates are usually of interest. From a GP trained on a molecular target, predictions on differences with respect to that target between molecular structures are readily available through Eqs. (6.12) and (6.13). For both targets separately, we trained a GP on randomly selected subsets of different size and then predicted relative energies between the remaining structures. This process was repeated 25 times to obtain converged means of the MAE, RMSE, and r_{cb}. From the results shown in Table 6.2 it can be seen that the MAE and the RMSE decrease rapidly with data set

Table 6.1: Mean absolute error (MAE), reference MAE (MAE_{ref}), root-mean-square error (RMSE), reference RMSE (RMSE_{ref}) (in kJ/mol), and r_{cb} of GP predictions on $\Delta H_{\text{PBE}}^{\text{G4MP2}}$ and $\Delta H_{\text{PM7}}^{\text{G4MP2}}$ for different training data set sizes.

Size	Target	MAE	MAE_{ref}	RMSE	RMSE_{ref}	r_{cb}
50	$\Delta H_{\text{PBE}}^{\text{G4MP2}}$	7.82	8.42	9.71	10.53	5.24
	$\Delta H_{\text{PM7}}^{\text{G4MP2}}$	21.61	26.24	27.86	33.13	6.40
100	$\Delta H_{\text{PBE}}^{\text{G4MP2}}$	7.30	8.42	9.03	10.53	4.53
	$\Delta H_{\text{PM7}}^{\text{G4MP2}}$	19.15	26.16	25.01	32.99	6.03
200	$\Delta H_{\text{PBE}}^{\text{G4MP2}}$	6.37	8.40	7.84	10.50	3.52
	$\Delta H_{\text{PM7}}^{\text{G4MP2}}$	15.71	26.12	21.06	32.97	6.48
500	$\Delta H_{\text{PBE}}^{\text{G4MP2}}$	4.42	8.39	5.45	10.48	3.83
	$\Delta H_{\text{PM7}}^{\text{G4MP2}}$	8.31	26.16	11.25	32.99	6.21
1000	$\Delta H_{\text{PBE}}^{\text{G4MP2}}$	2.90	8.37	3.64	10.45	4.26
	$\Delta H_{\text{PM7}}^{\text{G4MP2}}$	4.64	26.15	6.21	32.91	4.74

size, however, the accuracy is lower than that of predictions on the standard enthalpy of atomization. Nonetheless, r_{cb} indicates, that the error estimates remain reliable.

Table 6.2: Mean absolute error (MAE), root-mean-square error (RMSE) (in kJ/mol), and r_{cb} of predictions on differences in the standard enthalpy between molecular structures from GPs trained on targets $\Delta H_{\text{PBE}}^{\text{G4MP2}}$ and $\Delta H_{\text{PM7}}^{\text{G4MP2}}$.

Size	Target	MAE	RMSE	r_{cb}
50	$\Delta H_{\text{PBE}}^{\text{G4MP2}}$	10.96	13.67	5.35
	$\Delta H_{\text{PM7}}^{\text{G4MP2}}$	30.69	39.11	6.34
100	$\Delta H_{\text{PBE}}^{\text{G4MP2}}$	10.22	12.74	4.91
	$\Delta H_{\text{PM7}}^{\text{G4MP2}}$	27.54	35.26	5.56
200	$\Delta H_{\text{PBE}}^{\text{G4MP2}}$	8.88	11.07	4.22
	$\Delta H_{\text{PM7}}^{\text{G4MP2}}$	22.95	29.75	5.81
500	$\Delta H_{\text{PBE}}^{\text{G4MP2}}$	6.17	7.70	4.37
	$\Delta H_{\text{PM7}}^{\text{G4MP2}}$	12.13	15.88	5.96
1000	$\Delta H_{\text{PBE}}^{\text{G4MP2}}$	4.09	5.15	4.53
	$\Delta H_{\text{PM7}}^{\text{G4MP2}}$	6.72	8.78	5.36

Furthermore, for the targets $\Delta H_{\text{PBE}}^{\text{G4MP2}}$ and $\Delta H_{\text{PM7}}^{\text{G4MP2}}$ separately, we trained a GP on randomly selected subsets of different size and employed the remaining structures as an out-of-sample validation set. The MSE and mean prediction variance ($\langle\sigma^2\rangle$) was calculated for GP predictions on the out-of-sample validation set. We repeated this process to ensure that the average MSE and $\langle\sigma^2\rangle$ converged for each size of the training data set. In Fig. 6.4, the average MSE and average $\langle\sigma^2\rangle$ is shown as a function of the size of the training data set. It can be seen that both the average prediction variance and the prediction error decrease with the size of the training data set. In addition,

it can be observed that the values of the MSE and $\langle \sigma^2 \rangle$ are close for a given training data set size. This indicates that the prediction uncertainties provided by the GP are strongly related to the actual prediction error.

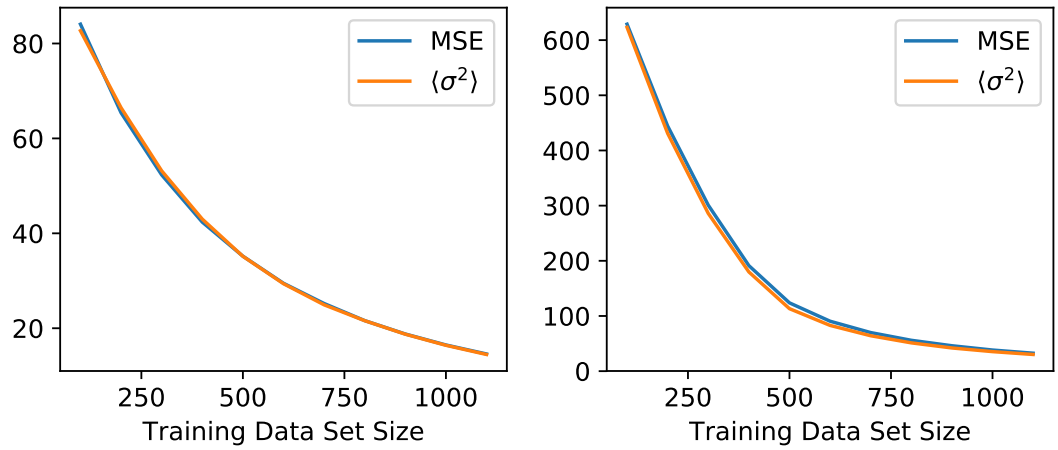

Figure 6.4: Mean squared error (MSE) and mean prediction variance ($\langle \sigma^2 \rangle$) of GP predictions on out-of-sample data sets for targets $\Delta H_{\text{PBE}}^{\text{G4MP2}}$ (left) and $\Delta H_{\text{PM7}}^{\text{G4MP2}}$ (right).

Hence, we demonstrated that GPs are capable of learning molecular properties of molecular structures with reliable error estimates. Furthermore, relative molecular properties can be predicted with sufficient accuracy employing a statistical model trained on individual molecular properties.

6.5.3 Error-Controlled Exploration

For the consecutive discovery of intermediates in the exploration of a chemical system, we generated sequences of nodes from our reaction network. Whereas all nodes are already known in our example network, an actual exploration procedure would expand the network in a continuous fashion (see Chapter 3). Starting from a random initial node in the reaction network, the remaining nodes were visited in the order of their graph distance to the initial node (see Fig. 6.1). Nodes with the same graph distance were discovered in a random order. Next, the error-controlled exploration strategy outlined in Section 6.4 was applied. Here, the initial data set consisted of the first $m = 75$ explored nodes. The explorations were separately performed for the targets $\Delta H_{\text{PBE}}^{\text{G4MP2}}$ and $\Delta H_{\text{PM7}}^{\text{G4MP2}}$. For each target, three different runs with different variance thresholds were carried out. Results for the exploration with targets $\Delta H_{\text{PBE}}^{\text{G4MP2}}$ and $\Delta H_{\text{PM7}}^{\text{G4MP2}}$ (on the same sequence) are shown in Figs. 6.5 and 6.6, respectively.

From Fig. 6.5 it can be seen that the size of the training data set initially increases. This is due to the low prediction confidence at the beginning of the exploration. The

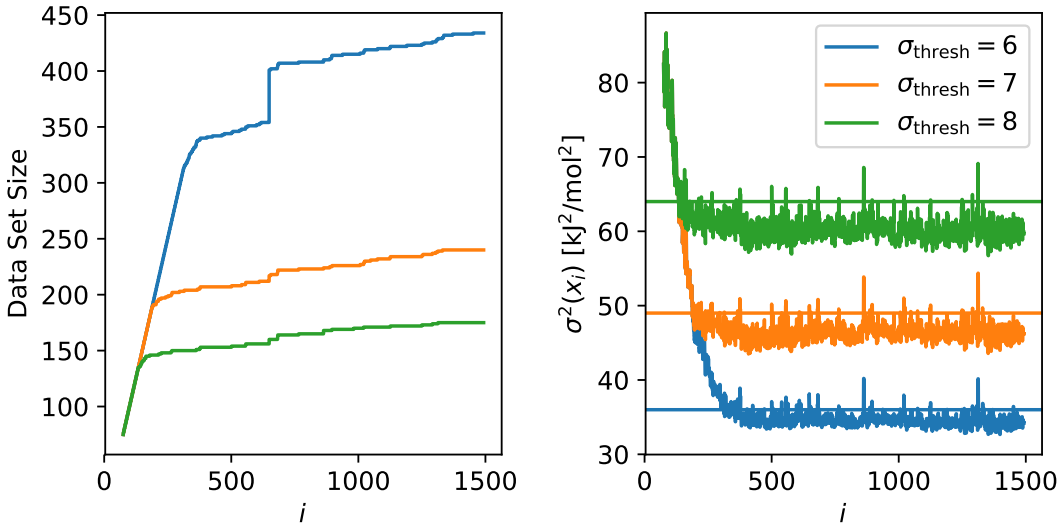

Figure 6.5: Size of the training data set (left) and prediction variance on the enthalpy of atomization (right) for the ith structure in an exploration employing the PBE functional and G4MP2 as the reference.

data set increases until the prediction uncertainty is below σ^2_{thresh} (shown as a horizontal line in Fig. 6.5, right). This is the point at which the predictions made by the GP are trusted for the first time. If, however, the exploration reaches regions of chemical space that are distant to the previously explored ones, the confidence will drop and new reference calculations will be required. This can be observed in Fig. 6.5, right, where the variance exceeds σ^2_{thresh}. Naturally, the total number of reference calculations for the entire exploration depends on the target and σ^2_{thresh}. Finally, it can be seen that the backtracking mechanism described in Section 6.4 is indeed necessary. In Fig. 6.5, for $\sigma_{\text{thresh}} = 6$ kJ/mol at $i = 651$, the GP is updated and some predictions which previously were inside the confidence bound now lie outside of it. Consequently, data points are added to the data set followed by an update of the GP until all predictions are within the confidence bound.

Fig. 6.6 shows that a larger data set is required for the target $\Delta H^{\text{G4MP2}}_{\text{PM7}}$ to reach a standard deviation of 15 kJ/mol, than that for the target $\Delta H^{\text{G4MP2}}_{\text{PBE}}$ to reach a standard deviation of 8 kJ/mol. This finding is in accordance with the results presented in Table 6.1. The calculation of the enthalpy of atomization is faster with PM7 than that with PBE by about an order of magnitude (for the systems studied in this work). However, since the exploration with PM7 as the base method requires far more computationally expensive G4MP2 reference calculations (which take more than three orders of magnitude longer than PBE calculations for the systems studied in this work), the overall exploration takes longer with PM7 than that with PBE as the base method. We

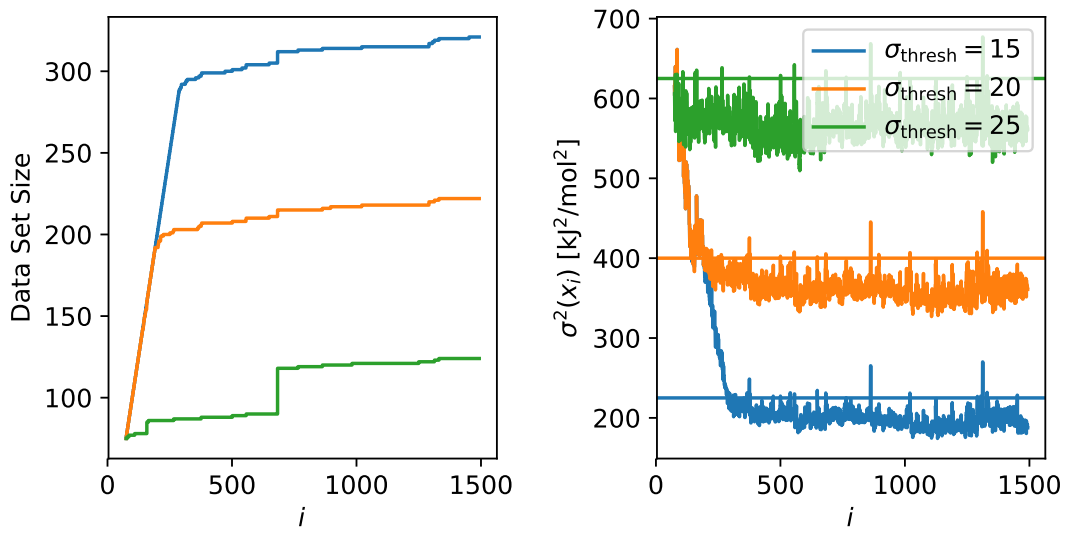

Figure 6.6: Size of the training data set (left) and prediction variance on the enthalpy of atomization (right) for the ith structure in an exploration employing PM7 and G4MP2 as the reference.

note that the time required for the evaluation of the kernel and GP predictions is negligible for data sets of this size. This illustrates the philosophy of the Δ-machine learning approach that should work more efficiently for the physically more reliable model (in our case, this is PBE). As a result, given a required confidence level, a trade-off needs to be found between the required number of reference calculations and the computational effort of the base method.

7

Conclusions and Outlook

In this thesis, a new, robust, and generally applicable protocol for the fully automated exploration of complex chemical reactions is presented. This protocol is implemented in a software package called Chemoton. Through a user-friendly graphical interface, even non-experts are now able to routinely explore the reactivity of molecular systems in an automated fashion. In addition, we devised strategies to attach error bars to properties calculated with quantum chemical methods so that meaningful conclusions can be drawn from the exploration data.

The following steps allowed us to achieve this goal: In Chapter 3, we devised a new protocol that through the application of heuristics based on quantum chemical observables, effectively reduces the myriad of possible side reactions that need to be explored for all but the simplest systems to a number that is computationally feasible. Based on electronic-structure theory the approach is applicable to arbitrary reactions and is not limited to any sort of molecule. Starting from a set of initial conditions (i.e., starting material, solvent, and temperature), an exploration network is built and extended through the repeated application of our protocol. New intermediates are explored through the construction of reactive complexes between already explored ones, new reactants, and intramolecular reactions. By applying heuristic rules based on conceptual electronic-structure theory, bond orders, and graph theoretical considerations, we were able to tame the combinatorial explosion of possible reactive complexes. Subsequently, a reaction was induced in each reactive complex and an approximate reaction path was explored. To ensure a thorough exploration of these paths, we considered rotational degrees of freedom during the construction of reactive complexes. Through transition state optimizations, the approximate paths were refined to minimum-energy paths. The

exploration network was processed to afford a compressed reaction network consisting of molecular configurations and reaction channels connecting them. The reaction network was then visualized by an automatically generated graph structure.

We applied our protocol to the formose reaction, a prebiotic oligomerization reaction resulting in a plethora of products, including monosaccharides. We explored a vast number of intermediates and minimum-energy paths to yield a reaction network featuring complex network structure and product distribution. Through the application of tree traversal algorithms, different pathways leading to the formation of the naturally occurring tetrose D-erythrose were identified. Furthermore, we discovered multiple pathways in the reaction network that rationalize the autocatalytic properties of the formose reaction. In addition, we showed that there can exist many minimum-energy paths with different reaction barriers for the same chemical transformation. Many of these reaction paths would remain undiscovered in a tedious manual exploration attempt.

In addition, we applied Chemoton to the Chatt–Schrock nitrogen-fixation cycle. Its competing reaction paths have not been studied in sufficient detail prior to this work. We explored a vast number of possible elementary reactions that describe protonation, proton-rearrangement, and reduction steps. The resulting network turned out to be highly complex and alternative routes that still sustain the catalytic cycle emerged. The application of an automated visualization strategy by which thermodynamic and kinetic network properties was crucial to facilitate the interpretation of such complex reaction mechanisms.

For an improved description of complex chemical processes, the following aspects need to be addressed. First, an appropriate solvent model is mandatory for chemical reactions studied in the liquid phase. Clearly, the application of a continuum model is not adequate for every solvent and region of the reaction network. At the same time, the addition of explicit solvent molecules to the reaction network not only increases the computational demand but could also hamper the identification of intermediates and, in particular, transition states due to the added degrees of freedom. Strategies have to be developed that allow one to examine in which regions of the reaction network the addition of explicit solvent molecules (in addition to a continuum model) is strictly necessary and where an implicit description is sufficient.

Second, the full conformational complexity of all intermediates needs to be taken into account. The current implementation in Chemoton can only explore conformational degrees of freedom of organic molecules. In our laboratory, methods are developed to extend the stochastic generation of conformers to molecules containing transition metal centers. For some intermediates with many floppy degrees of freedom (e.g., polymers), *ab initio* molecular dynamics simulations or Monte Carlo configurational sampling might be

more suitable for a sufficiently rigorous exploration of local minima and may, therefore, be employed in specific sections of the growing reaction network.

Third, bond dissociation reactions and accompanying kinetic analyses need to be incorporated in the exploration protocol. Similar to the reactivity descriptors, descriptors will need to be developed indicating the propensity for a bond to break under certain conditions (e.g., high temperature). With this one will be able to consider the possible degradation reactions that may deactivate reactive species (e.g., a catalyst).

Currently, the chemical reaction network is pruned by defining an energy cutoff, which allows for the exclusion of those intermediates which are inaccessible under a range of reasonable physical reaction conditions and on the timescale of interest. While this rather crude approach worked well for the study of the Yandulov–Schrock catalyst, a full kinetic study considering the time-solved concentration of all intermediates will allow one to reduce the complexity of the reaction network even further. As a result, less computational time will be spent on the exploration of inaccessible reaction paths, so that methods of increased accuracy with higher computational demand can be employed to study the remaining, accessible structures. To achieve this, CHEMOTON needs to be coupled to a software for kinetic modeling (e.g., KINETX[367]), in a closed-loop fashion to determine which intermediates and reaction paths are relevant to the overall process under given conditions.

In Chapter 4, we then argued that a procedure for quantifying the uncertainty associated with computational models, in particular with quantum chemical calculations, is mandatory despite their first-principles character. Otherwise, it may be difficult to draw meaningful conclusions in view of unknown uncertainties. Unfortunately, this procedure is neither well established nor straightforward. The abundance of benchmark studies reporting (potentially misleading) statistical measures such as the MAE and LAE, the hope for accurate post-Hartree–Fock methods to become applicable in a targeted way, and the difficulty of identifying the source of error largely prevented the development of novel approaches for reliable error estimation.

We also illustrated the different sources of errors and how to tackle them. We stress that a clear differentiation between the different sources of error is critical for the effective application of countermeasures. While numerical errors can often be controlled, model inadequacy and parameter uncertainty remain a major issue in quantum chemistry. Reducing model inadequacy through model improvement is a popular approach, although not straightforward for most methods. In these cases, statistical methods need to be applied in a rigorous way. While in most cases this does not improve accuracy, it allows for reliable uncertainty predictions which are critical, especially if the error is propagated to subsequent investigations such as kinetic studies.

In Chapter 5, a new approach for the construction of reliable, *system-specific* density

functionals with Bayesian error estimation is presented. By employing a system-focused re-parameterization of the RSH functional LC-PBE0, we were able to obtain a functional that allows for the accurate description of a particular system of interest. By choosing a functional based on physical principles with few parameters we also overcame the issue of transferability. Whereas a system-specific parameterization of density functionals is in general not a recommended strategy, here it is viable and useful because our functional provides confidence intervals for each result, and thus, allows one to assess whether the reported result is reliable. Clearly, our approach requires the generation of sufficiently accurate reference data for the class of molecules under consideration, but this is becoming comparatively easy with modern quantum chemistry software (see, e.g., Refs. 368,369) — even for multi-configuration cases (see, e.g., Refs. 370–372).

We applied our approach to the Yandulov–Schrock catalyst and identified that parameters in both the long-range corrected scheme and the PBE functional need to be optimized to obtain a sufficiently flexible functional. Furthermore, we were able to show that the reported error estimates are indeed reliable. Finally, we calculated the reaction energies of the Chatt–Schrock cycle. We showed that the confidence level of reaction energies can vary significantly — even if the reactions are very similar — therefore, highlighting the need for error estimation.

We applied our error estimation scheme to a simplified model network of the formose reaction. Since the rate constants depend on free activation energies $\Delta A^{\ddagger,*}$ through an exponential function, errors in $\Delta A^{\ddagger,*}$ strongly affect the kinetic simulation. Therefore, error estimates for $\Delta A^{\ddagger,*}$ are mandatory for meaningful conclusions drawn from a kinetic analysis. Nevertheless, for the simplified network, we could observe that even though the uncertainties in free activation energies were large, the qualitative flux of concentrations through the network remained qualitatively the same. It should be noted that errors of other contributions of $\Delta A^{\ddagger,*}$ were not accounted for in a systematic way, and thus, the error bars reported can be considered a lower bound for the true error.

From the two case studies, we concluded that to further increase the functionals accuracy and error estimation reliability, a functional form with greater flexibility would be beneficial. Therefore, in Chapter 6, we developed a novel approach for the rolling improvement of quantum chemical results through the application of Gaussian processes. By learning the error of an efficient quantum chemical method with respect to some reference method of higher accuracy, we obtained accurate standard enthalpies of formation for configurational isomers of the $C_7H_{10}O_2$ stoichiometry. Accurate differences in standard enthalpy between isomers are accessible as well. Furthermore, we showed that the uncertainty estimates provided by our predictive model for both the standard enthalpies of formation for molecules and differences in this standard enthalpy of different molecules are reliable. If the uncertainty associated with a particular cal-

culation is above a given threshold, the chosen reference method will be employed to produce additional reference data. In this way, reference calculations are performed only when needed, i.e., if regions of chemical space unknown to our model are approached and explored. We emphasize that our approach is independent of the chosen electronic structure models, ranging from semi-empirical and tight-binding models to multi-configurational approaches with multi-reference perturbation theory. Through *backtracking*, previous predictions are validated by the updated model to ensure that uncertainties remain within the given confidence bound.

In future work, the error-controlled exploration algorithm needs to be coupled to CHEMOTON to obtain error estimates during the exploration in an online fashion. In combination with our KINETX[367] algorithm for, reliable first-principles explorations of those portions of chemical reaction space that are relevant for a specific chemical problem become accessible. Obviously, this requires the accessibility to accurate reference calculations on demand. Exploiting, for instance, our multi-configurational diagnostic[373] allows one to decide on the single-reference vs. multi-reference nature of the molecular structure subjected to a reference calculation. For single-reference cases, explicitly correlated, local coupled-cluster calculations[368,369] are the method of choice as they can be easily launched in an automated manner and are known to be highly accurate. For multi-configurational cases, automated complete-active-space type calculations can be launched with our fully automated procedure for the selection of active orbital spaces.[372,374,375]

A

Computational Methodology

A.1 Exploration of Reaction Networks

The explorations of the formose reaction and the Yandulov–Schrock catalyst were carried out with our program package Chemoton in a fully automated fashion. The exploration protocol was implemented in C++.

All calculations and the progress of the explorations were saved to a Mongo database.[376] Automated data analyses were performed with the Python libraries `pandas`[377] and `matplotlib`.[378] The graphical representations of the reaction networks were created by the `Graphviz` program.[379]

A.1.1 Formose Reaction

All quantum chemical calculations were performed with the `Q-Chem` program package (version 4.3)[380] employing the PBE exchange-correlation functional[195] and a double-ζ basis.[381] We emphasize that our exploration protocol works with any electronic structure method and we chose a density-functional model for the sake of convenience. Also, for the raw-data generation other quantum chemistry packages can be easily interfaced.

For single point calculations, structure optimizations, and vibrational analyses default settings were kept. The maximum number of self-consistent field calculations was set to 1200 for structure optimizations, potential energy scans, freezing-string calculations, TS searches, and IRC calculations. For potential energy scans the convergence on energy change of successive optimization cycles was set to 10^{-4} Hartree. In freezing-string calculations the number of nodes was chosen to be 20 and the number of perpendicular

gradient steps was six. For TS searches the maximum allowed step size was reduced to 0.05 .

Activation barriers were approximated by the difference in electronic energy of reactant and TS, i.e., vibrational corrections were not included (this may be easily changed in standard approximations valid for the gas phase, but will increase the computational effort for the raw data generation). It is also easy to switch on a continuum solvation model in the exploration. However, we were interested in the exploration of a generic network first and therefore did not switch on dielectric continuum solvation. Solvation can then be studied in a subsequent step, where the network is copied multiple times to account for different solvation environments (also considering extensive microsolvation and configuration-space sampling as an intermediary layer between the solute and the continuum embedding) into which each isolated-structure node is then automatically embedded. Further refinement of these networks within the solvation model may be considered afterward.

A.1.2 Yandulov–Schrock Catalyst

Restricted and unrestricted density-functional-theory calculations were carried out depending on the lowest spin multiplicity of a given intermediate. For this, BP86/def2-SV(P)[192,382,383] structure optimizations of reactive complexes were performed with the program package Turbomole[384] (version 6.4.0) including the resolution-of-the-identity density-fitting technique. Single-point calculations were considered to be converged when the total electronic-energy difference between two iteration steps was less then 10^{-7} Hartree. Structure optimizations were considered converged when the norm of the electronic-energy gradient with respect to the nuclear coordinates dropped below 10^{-4} Hartree/Bohr. If a structure optimization failed because a self-consistent-field calculation did not converge, the damping parameters had been changed automatically and the optimization was restarted. In those cases where a structure optimization did not converge within 1200 iterations, the corresponding data was saved and the structure was manually inspected to decide whether it should be part of the chemical reaction network or not. 9607 structure optimizations were carried out in total.

Constrained BP86/def2-SV(P) optimizations were performed with Gaussian[385] (version 09, revision C.1) to obtain reasonable starting structures of TSs, which were refined with Turbomole's trust-radius-image-based EVF optimization choosing a trust-radius of 0.2 Å. The eigenmode to follow was obtained from a Mode-Tracking calculation.[29,30] From the converged TSs, intrinsic reaction paths were calculated with Gaussian to determine whether a desired TS was found. We employed the default convergence criteria for all Gaussian calculations. If the constrained optimization scan with a subsequent EVF calculation did not converge to the desired TS, the freezing-string method as imple-

mented in Q-Chem (version 4.0.1)[380] was employed with subsequent EVF as described above. We identified 2318 elementary reactions for which TSs were optimized.

To shed more light on the success rate of identifying TSs for these reactions by our automated search and therefore of verifying the assumption that two intermediates are truly connected by an elementary reaction, we may add some additional details. 1082 potential elementary reactions were automatically identified for the first and 1236 for the second half of the Chatt–Schrock cycle. The TS search was then conducted with three different strategies yielding a total of 6954 TS searches. This number, however, is only an upper bound as a search was stopped once one of these strategies was successful. In the first half of the cycle, our automated protocol identified 329 out of the 1082 potential elementary reactions by optimizing the TS. In the second half, it identified 613 out of the 1236 potential elementary reactions. For some steps, for which our implementation was not able to find a TS, we verified by manual inspection that a TS is not likely to exist or to be of sufficiently low energy. Hence, our algorithm produces more potential pairs that could be connected by an elementary reaction than there are. Of course, this number is determined by the structural similarity measure that we employ to relate the two structures. Obviously, our RMSD criterion produces many false positive results. However, this is actually desired as one cannot be certain to have found all relevant vertices and edges of a reaction network so that all criteria and measures should be set and selected in a conservative and therefore not too restrictive way.

The calculation of the ELF and of the electrostatic potential were performed with Molden 5.4.[386] In the color range of the electrostatic potential mapped onto the ELF isosurface, the most positive charge was omitted as otherwise the color differences between all other atoms would have been very small. The presentation of the data in Fig. 3.14 was generated with Jmol 14.0.7[387] from a cube file produced with Molden.

A.2 Bayesian Error Estimation

For both the study of the Yandulov-Schrock catalyst and the formose reaction network data analysis and visualization were carried out with the software packages `pandas`[377] and `matplotlib`.[378]

A.2.1 Yandulov–Schrock Catalyst

All BP86/RI/def2-TZVP[192,382,383] model-catalyst structures in D_P and D_A were optimized with the program package Turbomole.[384] BP86/RI/TZVP+SV(P) optimized structures of the full Yandulov–Schrock catalyst were taken from Ref. 179.

All CCSD(T) single-point calculations were carried out with the Molpro 2010.1 program package.[388] For the elements hydrogen, carbon, and nitrogen the aug-cc-pVDZ

basis set[389] was chosen. For molybdenum a double-ζ basis set together with an effective core potential (aug-cc-pVDZ-PP) was employed.[390] Clearly, for truly accurate reference data much larger one-electron basis sets or F12 basis sets are required. However, we already stress at this point that all conclusions drawn in this work will remain unchanged if the reference energies are corrected by a constant energy shift that may be different for different pairs of structures.

All subsequent DFT single-point calculations were carried out with the NWChem program package.[391] The following density functionals were employed: BP86,[192,382] B3LYP,[190–192] PBE,[195] PBE0,[194] LC-PBE0, M06-2X,[257] M06-L,[392] TPSS,[393] and TPSSh.[196] Furthermore, for BP86, B3LYP, PBE0, M06-2X, M06-L, TPSS, and TPSSh we considered Grimme's third generation dispersion correction,[264,394] denoted as BP86-D3, B3LYP-D3, PBE0-D3, M06-2X-D3, M06-L-D3, TPSS-D3, and TPSSh-D3, respectively. For all DFT calculations on structures in D_{P} and D_{A} a triple-ζ basis set (def2-TZVP) was chosen for all atoms.[383] Calculations on the Yandulov–Schrock catalyst were carried out with a triple-ζ basis set (def2-TZVP) on molybdenum and nitrogen atoms, and a double-ζ basis set (def2-SV(P)) on carbon and hydrogen atoms.[383] In all DFT calculations, scalar-relativistic effects were taken into account for the elements molybdenum and chromium by means of Stuttgart effective core potentials.[395]

The set of parameters employed for α in the LC*-PBEO(D_{P}) functional is: {0.2147, 0.1640, 0.2267, 0.2966, 0.1563, 0.1563, 0.3011, 0.2363, 0.1375, 0.2183, 0.1380, 0.1378, 0.1943, 0.0222, 0.0372, 0.1301, 0.0941, 0.2001, 0.1025, 0.0622, 0.2921, 0.1570, 0.1804, 0.0612, 0.1315}.

The set of parameters employed for α in the LC*-PBEO(D_{A}) functional is: {0.001, 0.1255, 0.0775, 0.0088, 0.2429, 0.0892, 0.0645, 0.2151, 0.0236, 0.0199, 0.1383, 0.1982, 0.1847, 0.1008, 0.0468, 0.2582, 0.3978, 0.2084, 0.1381, 0.1885, 0.0959, 0.1401, 0.0944, 0.1149, 0.1394}.

A.2.2 Formose Reaction Network

The reference set consists of relative energies between structures taken from Refs. 170 and 311. These energies were determined from density fitting local coupled cluster (DF-LCCSD(T0)-F12a/cc-pVTZ-F12)[396,397] single-point calculations which were carried out with Molpro (version 2015).[388] The cc-pVTZ/JKFIT basis[398] was employed for Fock matrix fitting and the aug-cc-pVTZ/MP2FIT basis[399] for the fitting of all other integrals. All DFT single-point calculations were carried out with the NWChem program package.[391] The following density functionals were employed: BP86,[192,382] B3LYP,[190–192] PBE,[193] PBE0,[194] LC-PBE0, M06-2X,[400] M06-L,[392] TPSS,[393] and TPSSh.[196] Furthermore, for BP86, B3LYP, PBE0, M06-2X, M06-L, TPSS, and TPSSh

we considered Grimme's third generation dispersion correction,[262,394] denoted as BP86-D3, B3LYP-D3, PBE0-D3, M06-2X-D3, M06-L-D3, TPSS-D3, and TPSSh-D3, respectively. For all DFT calculations a triple-ζ basis set (def2-TZVP) was chosen for all atoms.[383]

The set of parameters employed for α in the LC*-PBE0 functional is: {0.5590, 0.3943, 0.4642, 0.6166, 0.5715, 0.2280, 0.4607, 0.3277, 0.3335, 0.3956, 0.3634, 0.5216, 0.4379, 0.3607, 0.3996, 0.3863, 0.5264, 0.3212, 0.3838, 0.2429, 0.0377, 0.4249, 0.4504, 0.2564, 0.6201, 0.1704, 0.3515, 0.3234, 0.5311, 0.5235, 0.3647, 0.3917, 0.2388, 0.1068, 0.3040, 0.3649, 0.4946, 0.4912, 0.2992, 0.3095, 0.2194, 0.1745, 0.1399, 0.5816, 0.2845, 0.2931, 0.1947, 0.4399, 0.1511, 0.3203}.

A.3 Error-Controlled Exploration

The data set employed in this study is a subset of the GDB-17 data set.[401] All G4 geometries were taken from Ref. 361. G4MP2 enthalpies of atomization were also taken from Ref. 361. DFT enthalpies of atomization were based on electronic energies obtained with the PBE exchange-correlation functional[193] and a double-ζ basis.[381] DFT calculations were performed with the program packages `Q-Chem` (version 4.3).[380] Vibrational frequencies and rotational constants were taken from Ref. 361. Accordingly, $\Delta H_{\mathrm{PBE}}^{\mathrm{G4MP2}}$ is given by the difference in G4MP2 and PBE electronic energies of atomization as the nuclear contributions cancel in this setup. By contrast, PM7 enthalpies of atomization were calculated from enthalpies of formation obtained with the MOPAC program (version 2016).[402]

For the construction of the reaction network, transformation rules were applied to the graph representations of the constitutional isomers of the $C_7H_{10}O_2$ stoichiometry. These rules describe nucleophilic addition and substitution reactions, isomerizations of double bonds, and [2+2] cycloaddition reactions. In this study, the hydroxyl group and the α-carbon of an aldehyde or a ketone acted as nucleophiles. Double bonds and ethers were considered electrophilic. We emphasize that some reactions in this reaction network may feature high activation barriers.

The SOAP average kernel was evaluated with the `glosim` package.[321] Following previous work,[359,360] we chose an exponent of $\zeta = 4.0$. In addition, we set the Gaussian width parameter to be $\sigma = 0.3\,\text{Å}$ and the cutoff radius to be $R_{\mathrm{cut}} = 4.0\,\text{Å}$. Furthermore, we chose the number of radial and angular functions to be 12 and 10, respectively. Our model would likely benefit from an exhaustive search over hyperparameters, however, consistent with previous findings,[321] the performance of the kernel is not highly sensitive to the chosen set of parameters.

GP predictions were carried out with the library `GPy`.[403] Data analysis and visual-

ization were performed with the Python libraries `pandas`[377] and `matplotlib`,[378] respectively. The graphical representation of the reaction network was created by the `Graphviz` program.[379]

Abbreviations

ADDF	anharmonic downward distortion following
AFIR	artificial force-induced reaction
Cp*	pentamethylcyclopentadienyl
DFT	density functional theory
DG	distance geometry
ELF	electron localization function
EVF	eigenvector following
HF	Hartree–Fock
HIPT	hexa-*iso*-propyl terphenyl
GP	Gaussian process
IRC	intrinsic reaction coordinate
LAE	largest absolute error
LC	long-range corrected
Lut	lutidine
MAE	mean absolute error
MC	Monte Carlo
MCS	maximum common subgraph
MD	molecular dynamics
MEP	minimum-energy path
MSE	mean signed error
PES	potential-energy surface
QED	quantum electrodynamics
RMG	reaction mechanism generator
RMSD	root-mean-square deviation
RMSE	root-mean-square error

RSH	range-separated hybrid
SOAP	smooth overlap of atomic potentials
TS	transition state
TSSCDS	transition state search using chemical dynamics simulations
XC	exchange–correlation

Acknowledgments

First of all, I would like to thank Prof. Dr. Markus Reiher for giving me the opportunity to work on this fascinating research topic. I appreciate his patience, friendliness, and passion for science. Furthermore, I am grateful for the freedom he gave me to follow my scientific curiosity. Without his strong support, I would not be where I am right now.

I would also like to thank Prof. Dr. Gunnar Jeschke for accepting to be the co-examiner of this thesis.

Many thanks go to (former) group members of the Reiher group: Dr. Alberto Baiardi, Dr. Maike Bergeler, Francesco Bosia, Christoph Brunken, Vera von Burg, Dr. Leon Freitag, Dr. Moritz Haag, Tamara Husch, Dr. Sebastian Keller, Dr. Stefan Knecht, Dr. Arseny Kovyrshin, Dr. Florian Krausbeck, PD Dr. Hans Peter Lüthi, Andrea Muolo, Dr. Jonny Proppe, Jan-Grimo Sobez, Dr. Christopher Stein, Dr. Alain Vaucher, and Dr. Thomas Weymuth. Throughout my stay, I enjoyed the friendly, encouraging, and highly motivating atmosphere.

In addition, I would like to thank Tamara Husch for the challenging political discussions and invaluable scientific insight. Without Adrian Mühlbach's expertise in graphic design this thesis would not have the same style. I would also like to thank Andrea (Andi) Muolo for the deep conversations on science, life, and physical fitness. I want to give Dr. Jonny Proppe, the group's beer brewer, a big shout-out for the tasty beverages and the very enjoyable collaborative work. Furthermore, I am grateful for Jan-Grimo Sobez's amazing cooking and advice on C++ programming. Throughout my stay in the Reiher group, I really appreciated Dr. Christopher Stein's sense of humor and musical creativity. Finally, a warm thank you goes to Alain Vaucher for being such a good friend. I will miss the fruitful collaboration, discussions on software design, and traveling together.

I also had the great pleasure of supervising two bright and enthusiastic students: Vera von Burg and Paul Türtscher.

I consider myself lucky to have made amazing friends during my studies: Luzia Gyr, Jethro Hemmann, Nina Hentzen, Alain Jeanrenaud, Caroline Martin, Irina Ritsch, Laurent Sévery, Martin Slusarczyk, and Flurin Sturzenegger. Thank you for the legendary trips, for making lectures so much more enjoyable, and the fun we had over the last eight years.

Finally, I would like to thank my dear parents Christoph and Diemud and my beloved sisters Franziska and Marie for their constant support.

References

[1] Masters, C. *Homogeneous Transition-Metal Catalysis: A Gentle Art*, 1st ed.; Springer Netherlands, 2011.

[2] Vinu, R.; Broadbelt, L. J. Unraveling Reaction Pathways and Specifying Reaction Kinetics for Complex Systems. *Annu. Rev. Chem. Biomol. Eng.* **2012**, *3*, 29–54.

[3] Ross, J. Determination of Complex Reaction Mechanisms. Analysis of Chemical, Biological and Genetic Networks. *J. Phys. Chem. A* **2008**, *112*, 2134–2143.

[4] Vereecken, L.; Glowacki, D. R.; Pilling, M. J. Theoretical Chemical Kinetics in Tropospheric Chemistry: Methodologies and Applications. *Chem. Rev.* **2015**, *115*, 4063–4114.

[5] Ludlow, R. F.; Otto, S. Systems Chemistry. *Chem. Soc. Rev.* **2008**, *37*, 101–108.

[6] Clayden, J.; Greeves, N.; Warren, S.; Wothers, P. *Organic Chemistry*; Oxford University Press: Oxford, 2001.

[7] Greiner, W.; Reinhardt, J. *Quantum Electrodynamics*, 4th ed.; Springer: Berlin, 2009.

[8] Glotzer, S. C.; Kim, S.; Cummings, P. T.; Deshmukh, A.; Head-Gordon, M.; Karniadakis, G.; Petzold, L.; Sagui, C.; Shinozuka, M. *International Assessment of Research and Development in Simulation-Based Engineering and Science. Panel Report*; 2009.

[9] ISO, *Guide to the Expression of Uncertainty in Measurement*, 1st ed.; International Organization for Standardization: Genève, Switzerland, 1995.

[10] Irikura, K. K.; III, R. D. J.; Kacker, R. N. Uncertainty Associated with Virtual Measurements from Computational Quantum Chemistry Models. *Metrologia* **2004**, *41*, 369.

[11] Butlerow, A. Bildung Einer Zuckerartigen Substanz Durch Synthese. *Justus Liebigs Ann. Chem.* **1861**, *120*, 295–298.

[12] Eschenmoser, A.; Loewenthal, E. Chemistry of Potentially Prebiological Natural Products. *Chem. Soc. Rev.* **1992**, *21*, 1–16.

[13] Delidovich, I. V.; Simonov, A. N.; Taran, O. P.; Parmon, V. N. Catalytic Formation of Monosaccharides: From the Formose Reaction towards Selective Synthesis. *ChemSusChem* **2014**, *7*, 1833–1846.

[14] Yandulov, D. V.; Schrock, R. R. Reduction of Dinitrogen to Ammonia at a Well-Protected Reaction Site in a Molybdenum Triamidoamine Complex. *J. Am. Chem. Soc.* **2002**, *124*, 6252–6253.

[15] Yandulov, D. V.; Schrock, R. R.; Rheingold, A. L.; Ceccarelli, C.; Davis, W. M. Synthesis and Reactions of Molybdenum Triamidoamine Complexes Containing Hexaisopropylterphenyl Substituents. *Inorg. Chem.* **2003**, *42*, 796–813.

[16] Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. Methods for Exploring Reaction Space in Molecular Systems. *WIREs Comput. Mol. Sci.* **2017**, e1354.

[17] Hawkins, P. C. D. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.* **2017**, *57*, 1747–1756.

[18] Cerjan, C. J.; Miller, W. H. On Finding Transition States. *J. Chem. Phys.* **1981**, *75*, 2800–2806.

[19] Simons, J.; Joergensen, P.; Taylor, H.; Ozment, J. Walking on Potential Energy Surfaces. *J. Phys. Chem.* **1983**, *87*, 2745–2753.

[20] Davis, H. L.; Wales, D. J.; Berry, R. S. Exploring Potential Energy Surfaces with Transition State Calculations. *J. Chem. Phys.* **1990**, *92*, 4308–4319.

[21] Wales, D. J. Basins of Attraction for Stationary Points on a Potential-Energy Surface. *J. Chem. Soc., Faraday Trans.* **1992**, *88*, 653–657.

[22] Wales, D. J. Locating Stationary Points for Clusters in Cartesian Coordinates. *J. Chem. Soc., Faraday Trans.* **1993**, *89*, 1305–1313.

[23] Jensen, F. Locating Transition Structures by Mode Following: A Comparison of Six Methods on the Ar_8 Lennard-Jones Potential. *J. Chem. Phys.* **1995**, *102*, 6706–6718.

[24] Doye, J. P. K.; Wales, D. J. Surveying a Potential Energy Surface by Eigenvector-Following. *Z. Phys. D: At. Mol. Clusters* **1997**, *40*, 194–197.

[25] Broyden, C. G. Quasi-Newton Methods and Their Application to Function Minimisation. *Math. Comp.* **1967**, *21*, 368–381.

[26] Munro, L. J.; Wales, D. J. Defect Migration in Crystalline Silicon. *Phys. Rev. B* **1999**, *59*, 3969–3980.

[27] Malek, R.; Mousseau, N. Dynamics of Lennard-Jones Clusters: A Characterization of the Activation-Relaxation Technique. *Phys. Rev. E* **2000**, *62*, 7723–7728.

[28] Deglmann, P.; Furche, F. Efficient Characterization of Stationary Points on Potential Energy Surfaces. *J. Chem. Phys.* **2002**, *117*, 9535–9538.

[29] Reiher, M.; Neugebauer, J. A Mode-Selective Quantum Chemical Method for Tracking Molecular Vibrations Applied to Functionalized Carbon Nanotubes. *J. Chem. Phys.* **2003**, *118*, 1634–1641.

[30] Bergeler, M.; Herrmann, C.; Reiher, M. Mode-Tracking Based Stationary-Point Optimization. *J. Comput. Chem.* **2015**, *36*, 1429–1438.

[31] Ohno, K.; Maeda, S. A Scaled Hypersphere Search Method for the Topography of Reaction Pathways on the Potential Energy Surface. *Chem. Phys. Lett.* **2004**, *384*, 277–282.

[32] Maeda, S.; Ohno, K. Ab Initio Studies on Synthetic Routes of Glycine from Simple Molecules via Ammonolysis of Acetolactone: Applications of the Scaled Hypersphere Search Method. *Chem. Lett.* **2004**, *33*, 1372–1373.

[33] Maeda, S.; Ohno, K. Global Mapping of Equilibrium and Transition Structures on Potential Energy Surfaces by the Scaled Hypersphere Search Method: Applications to Ab Initio Surfaces of Formaldehyde and Propyne Molecules. *J. Phys. Chem. A* **2005**, *109*, 5742–5753.

[34] Ohno, K.; Maeda, S. Global Reaction Route Mapping on Potential Energy Surfaces of Formaldehyde, Formic Acid, and Their Metal-Substituted Analogues. *J. Phys. Chem. A* **2006**, *110*, 8933–8941.

[35] Maeda, S.; Ohno, K.; Morokuma, K. 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.* **2013**, *15*, 3683–3701.

[36] Maeda, S.; Morokuma, K. Communications: A Systematic Method for Locating Transition Structures of A+B→X Type Reactions. *J. Chem. Phys.* **2010**, *132*, 241102.

[37] Maeda, S.; Morokuma, K. Finding Reaction Pathways of Type A+B→X: Toward Systematic Prediction of Reaction Mechanisms. *J. Chem. Theory Comput.* **2011**, *7*, 2335–2345.

[38] Maeda, S.; Taketsugu, T.; Morokuma, K. Exploring Transition State Structures for Intramolecular Pathways by the Artificial Force Induced Reaction Method. *J. Comput. Chem.* **2013**, *35*, 166–173.

[39] Maeda, S.; Harabuchi, Y.; Takagi, M.; Taketsugu, T.; Morokuma, K. Artificial Force Induced Reaction (AFIR) Method for Exploring Quantum Chemical Potential Energy Surfaces. *Chem. Rec.* **2016**, 2232–2248.

[40] Sameera, W. M. C.; Maeda, S.; Morokuma, K. Computational Catalysis Using the Artificial Force Induced Reaction Method. *Acc. Chem. Res.* **2016**, *49*, 763–773.

[41] Yoshimura, T.; Maeda, S.; Taketsugu, T.; Sawamura, M.; Morokuma, K.; Mori, S. Exploring the Full Catalytic Cycle of Rhodium(I)–BINAP-Catalysed Isomerisation of Allylic Amines: A Graph Theory Approach for Path Optimisation. *Chem. Sci.* **2017**, *8*, 4475–4488.

[42] Puripat, M.; Ramozzi, R.; Hatanaka, M.; Parasuk, W.; Parasuk, V.; Morokuma, K. The Biginelli Reaction Is a Urea-Catalyzed Organocatalytic Multicomponent Reaction. *J. Org. Chem.* **2015**, *80*, 6959–6967.

[43] Hebrard, F.; Kalck, P. Cobalt-Catalyzed Hydroformylation of Alkenes: Generation and Recycling of the Carbonyl Species, and Catalytic Cycle. *Chem. Rev.* **2009**, *109*, 4272–4282.

[44] Vázquez, S. A.; Martínez-Núñez, E. HCN Elimination from Vinyl Cyanide: Product Energy Partitioning, the Role of Hydrogen–Deuterium Exchange Reactions and a New Pathway. *Phys. Chem. Chem. Phys.* **2015**, *17*, 6948–6955.

[45] Martínez-Núñez, E. An Automated Method to Find Transition States Using Chemical Dynamics Simulations. *J. Comput. Chem.* **2015**, *36*, 222–234.

[46] Martínez-Núñez, E. An Automated Transition State Search Using Classical Trajectories Initialized at Multiple Minima. *Phys. Chem. Chem. Phys.* **2015**, *17*, 14912–14921.

[47] Varela, J. A.; Vázquez, S. A.; Martínez-Núñez, E. An Automated Method to Find Reaction Mechanisms and Solve the Kinetics in Organometallic Catalysis. *Chem. Sci.* **2017**, *8*, 3843–3851.

[48] Saitta, A. M.; Saija, F. Miller Experiments in Atomistic Computer Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 13768–13773.

[49] Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J. Discovering Chemistry with an Ab Initio Nanoreactor. *Nat. Chem.* **2014**, *6*, 1044–1048.

[50] Wang, L.-P.; McGibbon, R. T.; Pande, V. S.; Martinez, T. J. Automated Discovery and Refinement of Reactive Molecular Dynamics Pathways. *J. Chem. Theory Comput.* **2016**, *12*, 638–649.

[51] van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A. ReaxFF: A Reactive Force Field for Hydrocarbons. *J. Phys. Chem. A* **2001**, *105*, 9396–9409.

[52] Döntgen, M.; Przybylski-Freund, M.-D.; Kröger, L. C.; Kopp, W. A.; Ismail, A. E.; Leonhard, K. Automated Discovery of Reaction Pathways, Rate Constants, and Transition States Using Reactive Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2015**, *11*, 2517–2524.

[53] Fischer, S.; Karplus, M. Conjugate Peak Refinement: An Algorithm for Finding Reaction Paths and Accurate Transition States in Systems with Many Degrees of Freedom. *Chem. Phys. Lett.* **1992**, *194*, 252–261.

[54] Florián, J.; Goodman, M. F.; Warshel, A. Computer Simulation of the Chemical Catalysis of DNA Polymerases: Discriminating between Alternative Nucleotide Insertion Mechanisms for T7 DNA Polymerase. *J. Am. Chem. Soc.* **2003**, *125*, 8163–8177.

[55] Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. How Enzymes Work: Analysis by Modern Rate Theory and Computer Simulations. *Science* **2004**, *303*, 186–195.

[56] Imhof, P.; Fischer, S.; Smith, J. C. Catalytic Mechanism of DNA Backbone Cleavage by the Restriction Enzyme EcoRV: A Quantum Mechanical/Molecular Mechanical Analysis. *Biochemistry* **2009**, *48*, 9061–9075.

[57] Reidelbach, M.; Betz, F.; Mäusle, R. M.; Imhof, P. Proton Transfer Pathways in an Aspartate-Water Cluster Sampled by a Network of Discrete States. *Chem. Phys. Lett.* **2016**, *659*, 169–175.

[58] Imhof, P. A Networks Approach to Modeling Enzymatic Reactions. *Methods Enzymol.* **2016**, *578*, 249–271.

[59] Senn, H. M.; Thiel, W. QM/MM Methods for Biological Systems. *Top. Curr. Chem.* **2007**, *268*, 173–290.

[60] Senn, H. M.; Thiel, W. QM/MM Studies of Enzymes. *Curr. Opin. Chem. Biol.* **2007**, *11*, 182–187.

[61] Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem. Int. Ed.* **2009**, *48*, 1198–1229.

[62] Jihyun Shim,; MacKerell Jr, A. D. Computational Ligand-Based Rational Design: Role of Conformational Sampling and Force Fields in Model Development. *Med. Chem. Commun.* **2011**, *2*, 356–370.

[63] De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59*, 4035–4061.

[64] Tsujishita, H.; Hirono, S. Camdas: An Automated Conformational Analysis System Using Molecular Dynamics. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 305–315.

[65] Wilson, S. R.; Cui, W.; Moskowitz, J. W.; Schmidt, K. E. Applications of Simulated Annealing to the Conformational Analysis of Flexible Molecules. *J. Comput. Chem.* **1991**, *12*, 342–349.

[66] Sperandio, O.; Souaille, M.; Delfaud, F.; Miteva, M. A.; Villoutreix, B. O. MED-3DMC: A New Tool to Generate 3D Conformation Ensembles of Small Molecules with a Monte Carlo Sampling of the Conformational Space. *Eur. J. Med. Chem.* **2009**, *44*, 1405–1409.

[67] Grebner, C.; Becker, J.; Stepanenko, S.; Engels, B. Efficiency of Tabu-Search-Based Conformational Search Algorithms. *J. Comput. Chem.* **2011**, *32*, 2245–2253.

[68] Satoh, H.; Oda, T.; Nakakoji, K.; Uno, T.; Tanaka, H.; Iwata, S.; Ohno, K. Potential Energy Surface-Based Automatic Deduction of Conformational Transition Networks and Its Application on Quantum Mechanical Landscapes of d-Glucose Conformers. *J. Chem. Theory Comput.* **2016**, *12*, 5293–5308.

[69] Halgren, T. A.; Lipscomb, W. N. The Synchronous-Transit Method for Determining Reaction Pathways and Locating Molecular Transition States. *Chem. Phys. Lett.* **1977**, *49*, 225–232.

[70] Ayala, P. Y.; Schlegel, H. B. A Combined Method for Determining Reaction Paths, Minima, and Transition State Geometries. *J. Chem. Phys.* **1997**, *107*, 375–384.

[71] Henkelman, G.; Jónsson, H. A Dimer Method for Finding Saddle Points on High Dimensional Potential Surfaces Using Only First Derivatives. *J. Chem. Phys.* **1999**, *111*, 7010–7022.

[72] Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths. *J. Chem. Phys.* **2000**, *113*, 9901–9904.

[73] Henkelman, G.; Jónsson, H. Improved Tangent Estimate in the Nudged Elastic Band Method for Finding Minimum Energy Paths and Saddle Points. *J. Chem. Phys.* **2000**, *113*, 9978–9985.

[74] Maragakis, P.; Andreev, S. A.; Brumer, Y.; Reichman, D. R.; Kaxiras, E. Adaptive Nudged Elastic Band Approach for Transition State Calculation. *J. Chem. Phys.* **2002**, *117*, 4651–4658.

[75] E, W.; Ren, W.; Vanden-Eijnden, E. String Method for the Study of Rare Events. *Phys. Rev. B* **2002**, *66*, 052301.

[76] E, W.; Ren, W.; Vanden-Eijnden, E. Finite Temperature String Method for the Study of Rare Events. *J. Phys. Chem. B* **2005**, *109*, 6688–6693.

[77] Behn, A.; Zimmerman, P. M.; Bell, A. T.; Head-Gordon, M. Efficient Exploration of Reaction Paths via a Freezing String Method. *J. Chem. Phys.* **2011**, *135*, 224108.

[78] Zimmerman, P. Reliable Transition State Searches Integrated with the Growing String Method. *J. Chem. Theory Comput.* **2013**, *9*, 3043–3050.

[79] Broadbelt, L. J.; Stark, S. M.; Klein, M. T. Computer Generated Pyrolysis Modeling: On-the-Fly Generation of Species, Reactions, and Rates. *Ind. Eng. Chem. Res.* **1994**, *33*, 790–799.

[80] Broadbelt, L. J.; Stark, S. M.; Klein, M. T. Computer Generated Reaction Modelling: Decomposition and Encoding Algorithms for Determining Species Uniqueness. *Comput. Chem. Eng.* **1996**, *20*, 113–129.

[81] Broadbelt, L. J.; Pfaendtner, J. Lexicography of Kinetic Modeling of Complex Reaction Networks. *AIChE J.* **2005**, *51*, 2112–2121.

[82] Evans, M. G.; Polanyi, M. Inertia and Driving Force of Chemical Reactions. *Trans. Faraday Soc.* **1938**, *34*, 11–24.

[83] Matheu, D. M.; Dean, A. M.; Grenda, J. M.; Green, W. H. Mechanism Generation with Integrated Pressure Dependence: A New Model for Methane Pyrolysis. *J. Phys. Chem. A* **2003**, *107*, 8552–8565.

[84] Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

[85] Harper, M. R.; Van Geem, K. M.; Pyl, S. P.; Marin, G. B.; Green, W. H. Comprehensive Reaction Mechanism for N-Butanol Pyrolysis and Combustion. *Combust. Flame* **2011**, *158*, 16–41.

[86] Geem, K. M. V.; Reyniers, M.-F.; Marin, G. B.; Song, J.; Green, W. H.; Matheu, D. M. Automatic Reaction Network Generation Using RMG for Steam Cracking of N-hexane. *AIChE J.* **2005**, *52*, 718–730.

[87] Petway, S. V.; Ismail, H.; Green, W. H.; Estupiñán, E. G.; Jusinski, L. E.; Taatjes, C. A. Measurements and Automated Mechanism Generation Modeling of OH Production in Photolytically Initiated Oxidation of the Neopentyl Radical. *J. Phys. Chem. A* **2007**, *111*, 3891–3900.

[88] Hansen, N.; Merchant, S. S.; Harper, M. R.; Green, W. H. The Predictive Capability of an Automatically Generated Combustion Chemistry Mechanism: Chemical Structures of Premixed Iso-Butanol Flames. *Combust. Flame* **2013**, *160*, 2343–2351.

[89] Slakman, B. L.; Simka, H.; Reddy, H.; West, R. H. Extending Reaction Mechanism Generator to Silicon Hydride Chemistry. *Ind. Eng. Chem. Res.* **2016**, *55*, 12507–12515.

[90] Seyedzadeh Khanshan, F.; West, R. H. Developing Detailed Kinetic Models of Syngas Production from Bio-Oil Gasification Using Reaction Mechanism Generator (RMG). *Fuel* **2016**, *163*, 25–33.

[91] Dana, A. G.; Buesser, B.; Merchant, S. S.; Green, W. H. Automated Reaction Mechanism Generation Including Nitrogen as a Heteroatom. *Int. J. Chem. Kinet.* **2018**, *50*, 243–258.

[92] Rappoport, D.; Galvin, C. J.; Zubarev, D. Y.; Aspuru-Guzik, A. Complex Chemical Reaction Networks from Heuristics-Aided Quantum Chemistry. *J. Chem. Theory Comput.* **2014**, *10*, 897–907.

[93] Zubarev, D. Y.; Rappoport, D.; Aspuru-Guzik, A. Uncertainty of Prebiotic Scenarios: The Case of the Non-Enzymatic Reverse Tricarboxylic Acid Cycle. *Sci. Rep.* **2015**, *5*, 1–7.

[94] Levy, D. E. *Arrow-Pushing in Organic Chemistry: An Easy Approach to Understanding Reaction Mechanisms*, 2nd ed.; Wiley, 2017.

[95] Kim, Y.; Kim, J. W.; Kim, Z.; Kim, W. Y. Efficient Prediction of Reaction Paths through Molecular Graph and Reaction Network Analysis. *Chem. Sci.* **2018**, *9*, 825–835.

[96] Zimmerman, P. M. Automated Discovery of Chemically Reasonable Elementary Reaction Steps. *J. Comput. Chem.* **2013**, *34*, 1385–1392.

[97] Zimmerman, P. M. Navigating Molecular Space for Reaction Mechanisms: An Efficient, Automated Procedure. *Mol. Simul.* **2015**, *41*, 43–54.

[98] Zimmerman, P. M. Growing String Method with Interpolation and Optimization in Internal Coordinates: Method and Examples. *J. Chem. Phys.* **2013**, *138*, 184102.

[99] Zimmerman, P. M. Single-Ended Transition State Finding with the Growing String Method. *J. Comput. Chem.* **2015**, *36*, 601–611.

[100] Jafari, M.; Zimmerman, P. M. Reliable and Efficient Reaction Path and Transition State Finding for Surface Reactions with the Growing String Method. *J. Comput. Chem.* **2017**, *38*, 645–658.

[101] Nett, A. J.; Zhao, W.; Zimmerman, P. M.; Montgomery, J. Highly Active Nickel Catalysts for C–H Functionalization Identified through Analysis of Off-Cycle Intermediates. *J. Am. Chem. Soc.* **2015**, *137*, 7636–7639.

[102] Li, M. W.; Pendleton, I. M.; Nett, A. J.; Zimmerman, P. M. Mechanism for Forming B,C,N,O Rings from NH_3BH_3 and CO_2 via Reaction Discovery Computations. *J. Phys. Chem. A* **2016**, *120*, 1135–1144.

[103] Pendleton, I. M.; Pérez-Temprano, M. H.; Sanford, M. S.; Zimmerman, P. M. Experimental and Computational Assessment of Reactivity and Mechanism in C(sp^3)–N Bond-Forming Reductive Elimination from Palladium(IV). *J. Am. Chem. Soc.* **2016**, *138*, 6049–6060.

[104] Zhao, Y.; Nett, A. J.; McNeil, A. J.; Zimmerman, P. M. Computational Mechanism for Initiation and Growth of Poly(3-Hexylthiophene) Using Palladium N-Heterocyclic Carbene Precatalysts. *Macromolecules* **2016**, *49*, 7632–7641.

[105] Dewyer, A. L.; Zimmerman, P. M. Finding Reaction Mechanisms, Intuitive or Otherwise. *Org. Biomol. Chem.* **2017**, *15*, 501–504.

[106] Ludwig, J. R.; Zimmerman, P. M.; Gianino, J. B.; Schindler, C. S. Iron(III)-Catalysed Carbonyl–Olefin Metathesis. *Nature* **2016**, *533*, 374–379.

[107] Smith, M. L.; Leone, A. K.; Zimmerman, P. M.; McNeil, A. J. Impact of Preferential π-Binding in Catalyst-Transfer Polycondensation of Thiazole Derivatives. *ACS Macro Lett.* **2016**, *5*, 1411–1415.

[108] Ludwig, J. R.; Phan, S.; McAtee, C. C.; Zimmerman, P. M.; Devery, J. J.; Schindler, C. S. Mechanistic Investigations of the Iron(III)-Catalyzed Carbonyl-Olefin Metathesis Reaction. *J. Am. Chem. Soc.* **2017**, *139*, 10832–10842.

[109] Dewyer, A. L.; Zimmerman, P. M. Simulated Mechanism for Palladium-Catalyzed, Directed γ-Arylation of Piperidine. *ACS Catal.* **2017**, *7*, 5466–5477.

[110] Suleimanov, Y. V.; Green, W. H. Automated Discovery of Elementary Chemical Reaction Steps Using Freezing String and Berny Optimization Methods. *J. Chem. Theory Comput.* **2015**,

[111] Grambow, C. A.; Jamal, A.; Li, Y.-P.; Green, W. H.; Zádor, J.; Suleimanov, Y. V. Unimolecular Reaction Pathways of a γ-Ketohydroperoxide from Combined Application of Automated Reaction Discovery Methods. *J. Am. Chem. Soc.* **2018**, *140*, 1035–1048.

[112] Schlegel, H. B. Optimization of Equilibrium Geometries and Transition Structures. *J. Comput. Chem.* **2004**, *3*, 214–218.

[113] Schlegel, H. B. Estimating the Hessian for Gradient-Type Geometry Optimizations. *Theoret. Chim. Acta* **1984**, *66*, 333–340.

[114] Peng, C.; Ayala, P. Y.; Schlegel, H. B.; Frisch, M. J. Using Redundant Internal Coordinates to Optimize Equilibrium Geometries and Transition States. *J. Comput. Chem.* **1998**, *17*, 49–56.

[115] Habershon, S. Automated Prediction of Catalytic Mechanism and Rate Law Using Graph-Based Reaction Path Sampling. *J. Chem. Theory Comput.* **2016**, *12*, 1786–1798.

[116] Habershon, S. Sampling Reactive Pathways with Random Walks in Chemical Space: Applications to Molecular Dissociation and Catalysis. *J. Chem. Phys.* **2015**, *143*, 094106.

[117] Lagorce, D.; Pencheva, T.; Villoutreix, B. O.; Miteva, M. A. DG-AMMOS: A New Tool to Generate 3D Conformation of Small Molecules Using Distance Geometry and Automated Molecular Mechanics Optimization for in Silico Screening. *BMC Chem. Biol.* **2009**, *9*, 6.

[118] Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

[119] Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462–2474.

[120] Leite, T. B.; Gomes, D.; Miteva, M. A.; Chomilier, J.; Villoutreix, B. O.; Tufféry, P. Frog: A FRee Online druG 3D Conformation Generator. *Nucleic Acids Res.* **2007**, *35*, W568–W572.

[121] Miteva, M. A.; Guyon, F.; Tufféry, P. Frog2: Efficient 3D Conformation Ensemble Generator for Small Compounds. *Nucleic Acids Res.* **2010**, *38*, W622–W627.

[122] Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572–584.

[123] Guba, W.; Meyder, A.; Rarey, M.; Hert, J. Torsion Library Reloaded: A New Version of Expert-Derived SMARTS Rules for Assessing Conformations of Small Molecules. *J. Chem. Inf. Model.* **2016**, *56*, 1–5.

[124] Schärfer, C.; Schulz-Gasch, T.; Hert, J.; Heinzerling, L.; Schulz, B.; Inhester, T.; Stahl, M.; Rarey, M. CONFECT: Conformations from an Expert Collection of Torsion Patterns. *ChemMedChem* **2013**, *8*, 1690–1700.

[125] Gánti, T. Organization of Chemical Reactions into Dividing and Metabolizing Units: The Chemotons. *BioSystems* **1975**, *7*, 15–21.

[126] Bergeler, M.; Simm, G. N.; Proppe, J.; Reiher, M. Heuristics-Guided Exploration of Reaction Mechanisms. *J. Chem. Theory Comput.* **2015**, *11*, 5712–5722.

[127] Proppe, J.; Husch, T.; Simm, G. N.; Reiher, M. Uncertainty Quantification for Quantum Chemical Models of Complex Reaction Networks. *Faraday Discuss.* **2016**, *195*, 497–520.

[128] O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab - Systematic Generation of Diverse Low-Energy Conformers. *J. Cheminf.* **2011**, *3*, 8.

[129] Mayer, I. Charge, Bond Order and Valence in the AB Initio SCF Theory. *Chem. Phys. Lett.* **1983**, *97*, 270–274.

[130] Becke, A. D.; Edgecombe, K. E. A Simple Measure of Electron Localization in Atomic and Molecular Systems. *J. Chem. Phys.* **1990**, *92*, 5397–5403.

[131] Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Clarendon Press: Oxford, 1994.

[132] Ayers, P. W. Electron Localization Functions and Local Measures of the Covariance. *J. Chem. Sci.* **2005**, *117*, 441–454.

[133] Fukui, K. Role of Frontier Orbitals in Chemical Reactions. *Science* **1982**, *218*, 747–754.

[134] Mulliken, R. S. Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I. *J. Chem. Phys.* **1955**, *23*, 1833–1840.

[135] Mulliken, R. S. Electronic Population Analysis on LCAO–MO Molecular Wave Functions. II. Overlap Populations, Bond Orders, and Covalent Bond Energies. *J. Chem. Phys.* **1955**, *23*, 1841–1846.

[136] Meister, J.; Schwarz, W. H. E. Principal Components of Ionicity. *J. Phys. Chem.* **1994**, *98*, 8245–8252.

[137] Bultinck, P.; Van Alsenoy, C.; Ayers, P. W.; Carbó-Dorca, R. Critical Analysis and Extension of the Hirshfeld Atoms in Molecules. *J. Chem. Phys.* **2007**, *126*, 144111.

[138] Brown, R. E.; Simas, A. M. On the Applicability of CNDO Indices for the Prediction of Chemical Reactivity. *Theoret. Chim. Acta* **1982**, *62*, 1–16.

[139] Kang, Y. K.; Jhon, M. S. Additivity of Atomic Static Polarizabilities and Dispersion Coefficients. *Theoret. Chim. Acta* **1982**, *61*, 41–48.

[140] Kunz, C. F.; Hättig, C.; Hess, B. A. Ab Initio Study of the Individual Interaction Energy Components in the Ground State of the Mercury Dimer. *Mol. Phys.* **1996**, *89*, 139–156.

[141] Morell, C.; Grand, A.; Toro-Labbé, A. New Dual Descriptor for Chemical Reactivity. *J. Phys. Chem. A* **2005**, *109*, 205–212.

[142] Morell, C.; Grand, A.; Toro-Labbé, A. Theoretical Support for Using the Δf(r) Descriptor. *Chem. Phys. Lett.* **2006**, *425*, 342–346.

[143] Ayers, P. W.; Morell, C.; De Proft, F.; Geerlings, P. Understanding the Woodward–Hoffmann Rules by Using Changes in Electron Density. *Chem. Eur. J.* **2007**, *13*, 8240–8247.

[144] Cárdenas, C.; Rabi, N.; Ayers, P. W.; Morell, C.; Jaramillo, P.; Fuentealba, P. Chemical Reactivity Descriptors for Ambiphilic Reagents: Dual Descriptor, Local Hypersoftness, and Electrostatic Potential. *J. Phys. Chem. A* **2009**, *113*, 8660–8667.

[145] Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual Density Functional Theory. *Chem. Rev.* **2003**, *103*, 1793–1874.

[146] Geerlings, P.; Proft, F. D. Conceptual DFT: The Chemical Relevance of Higher Response Functions. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3028–3042.

[147] Johnson, P. A.; Bartolotti, L. J.; Ayers, P. W.; Fievez, T.; Geerlings, P. *Modern Charge-Density Analysis*; Springer, 2011; pp 715–764.

[148] Corey, E. J.; Jorgensen, W. L. Computer-Assisted Synthetic Analysis. Synthetic Strategies Based on Appendages and the Use of Reconnective Transforms. *J. Am. Chem. Soc.* **1976**, *98*, 189–203.

[149] Pensak, D. A.; Corey, E. J. *Computer-Assisted Organic Synthesis*; ACS Symposium Series 61; American Chemical Society, 1977; Vol. 61; pp 1–32.

[150] Gasteiger, J.; Ihlenfeldt, W. D. In *Software Development in Chemistry 4*; Gasteiger, J., Ed.; Springer: Berlin, 1990; pp 57–65.

[151] Rücker, C.; Rücker, G.; Bertz, S. H. Organic Synthesis - Art or Science? *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 378–386.

[152] Fialkowski, M.; Bishop, K. J. M.; Chubukov, V. A.; Campbell, C. J.; Grzybowski, B. A. Architecture and Evolution of Organic Chemistry. *Angew. Chem. Int. Ed.* **2005**, *44*, 7263–7269.

[153] Todd, M. H. Computer-Aided Organic Synthesis. *Chem. Soc. Rev.* **2005**, *34*, 247–266.

[154] Chen, J. H.; Baldi, P. Synthesis Explorer: A Chemical Reaction Tutorial System for Organic Synthesis Design and Mechanism Prediction. *J. Chem. Educ.* **2008**, *85*, 1699.

[155] Chen, J. H.; Baldi, P. No Electron Left Behind: A Rule-Based Expert System To Predict Chemical Reactions and Reaction Mechanisms. *J. Chem. Inf. Model.* **2009**, *49*, 2034–2043.

[156] Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. Eur. J.* **2017**, *23*, 5966–5971.

[157] Segler, M. H. S.; Waller, M. P. Modelling Chemical Reasoning to Predict and Invent Reactions. *Chem. Eur. J.* **2017**, *23*, 6118–6128.

[158] Trygubenko, S. A.; Wales, D. J. A Doubly Nudged Elastic Band Method for Finding Transition States. *J. Chem. Phys.* **2004**, *120*, 2082–2094.

[159] Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. A Growing String Method for Determining Transition States: Comparison to the Nudged Elastic Band and String Methods. *J. Chem. Phys.* **2004**, *120*, 7877–7886.

[160] Kumeda, Y.; Wales, D. J.; Munro, L. J. Transition States and Rearrangement Mechanisms from Hybrid Eigenvector-Following and Density Functional Theory. *Chem. Phys. Lett.* **2001**, *341*, 185–194.

[161] Decker, P.; Schweer, H.; Pohlamnn, R. Bioids. *J. Chromatogr., A* **1982**, *244*, 281–291.

[162] Zweckmair, T.; Böhmdorfer, S.; Bogolitsyna, A.; Rosenau, T.; Potthast, A.; Novalin, S. Accurate Analysis of Formose Reaction Products by LC–UV: An Analytical Challenge. *J. Chromatogr. Sci.* **2014**, *52*, 169–175.

[163] Breslow, R. On the Mechanism of the Formose Reaction. *Tetrahedron Lett.* **1959**, *1*, 22–26.

[164] Bissette, A. J.; Fletcher, S. P. Mechanisms of Autocatalysis. *Angew. Chem. Int. Ed.* **2013**, *52*, 12800–12826.

[165] Kim, H.-J.; Ricardo, A.; Illangkoon, H. I.; Kim, M. J.; Carrigan, M. A.; Frye, F.; Benner, S. A. Synthesis of Carbohydrates in Mineral-Guided Prebiotic Cycles. *J. Am. Chem. Soc.* **2011**, *133*, 9457–9468.

[166] Socha, R. F.; Weiss, A. H.; Sakharov, M. M. Autocatalysis in the Formose Reaction. *React. Kinet. Catal. Lett.* **1980**, *14*, 119–128.

[167] Gregory Landrum, RDKit 2017.03.3. http://www.rdkit.org/, (Accessed: 8. July 2017).

[168] Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **2012**, *52*, 1146–1158.

[169] Simm, G. N.; Reiher, M. Context-Driven Exploration of Complex Chemical Reaction Networks. *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119.

[170] Kua, J.; Avila, J. E.; Lee, C. G.; Smith, W. D. Mapping the Kinetic and Thermodynamic Landscape of Formaldehyde Oligomerization under Neutral Conditions. *J. Phys. Chem. A* **2013**, *117*, 12658–12667.

[171] Eyring, H. The Activated Complex in Chemical Reactions. *J. Chem. Phys.* **1935**, *3*, 107–115.

[172] Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999–3094.

[173] Schrock, R. R. Catalytic Reduction of Dinitrogen to Ammonia by Molybdenum: Theory versus Experiment. *Angew. Chem. Int. Ed.* **2008**, *47*, 5512–5522.

[174] Hinrichsen, S.; Broda, H.; Gradert, C.; Söncksen, L.; Tuczek, F. Recent Developments in Synthetic Nitrogen Fixation. *Annu. Rep. Prog. Chem., Sect. A: Inorg. Chem.* **2012**, *108*, 17.

[175] Arashiba, K.; Miyake, Y.; Nishibayashi, Y. A Molybdenum Complex Bearing PNP-Type Pincer Ligands Leads to the Catalytic Reduction of Dinitrogen into Ammonia. *Nat. Chem.* **2011**, *3*, 120–125.

[176] Lee, Y.; Mankad, N. P.; Peters, J. C. Triggering N_2 Uptake via Redox-Induced Expulsion of Coordinated NH_3 and N_2 Silylation at Trigonal Bipyramidal Iron. *Nat. Chem.* **2010**, *2*, 558–565.

[177] Proft, F. D.; Ayers, P. W.; Geerlings, P. *The Chemical Bond*; Wiley-Blackwell, 2014; pp 233–270.

[178] Löwdin, P.-O. In *Advances in Quantum Chemistry*; Löwdin, P.-O., Ed.; Academic Press, 1970; Vol. 5; pp 185–199.

[179] Schenk, S.; Le Guennic, B.; Kirchner, B.; Reiher, M. First-Principles Investigation of the Schrock Mechanism of Dinitrogen Reduction Employing the Full $HIPTN_3N$ Ligand. *Inorg. Chem.* **2008**, *47*, 3634–3650.

[180] Reiher, M.; Le Guennic, B.; Kirchner, B. Theoretical Study of Catalytic Dinitrogen Reduction under Mild Conditions. *Inorg. Chem.* **2005**, *44*, 9640–9642.

[181] Le Guennic, B.; Kirchner, B.; Reiher, M. Nitrogen Fixation under Mild Ambient Conditions: Part I—The Initial Dissociation/Association Step at Molybdenum Triamidoamine Complexes. *Chem. Eur. J.* **2005**, *11*, 7448–7460.

[182] Schenk, S.; Kirchner, B.; Reiher, M. A Stable Six-Coordinate Intermediate in Ammonia-Dinitrogen Exchange at Schrock's Molybdenum Catalyst. *Chem. Eur. J.* **2009**, *15*, 5073–5082.

[183] Schenk, S.; Reiher, M. Ligands for Dinitrogen Fixation at Schrock-Type Catalysts. *Inorg. Chem.* **2009**, *48*, 1638–1648.

[184] Studt, F.; Tuczek, F. Energetics and Mechanism of a Room-Temperature Catalytic Process for Ammonia Synthesis (Schrock Cycle): Comparison with Biological Nitrogen Fixation. *Angew. Chem. Int. Ed.* **2005**, *44*, 5639–5642.

[185] Thimm, W.; Gradert, C.; Broda, H.; Wennmohs, F.; Neese, F.; Tuczek, F. Free Reaction Enthalpy Profile of the Schrock Cycle Derived from Density Functional Theory Calculations on the Full $[Mo^{HIPT}N_3N]$ Catalyst. *Inorg. Chem.* **2015**, *54*, 9248–9255.

[186] Peverati, R.; Truhlar, D. G. Quest for a Universal Density Functional: The Accuracy of Density Functionals across a Broad Spectrum of Databases in Chemistry and Physics. *Philos. Trans. R. Soc. London, Ser. A* **2014**, *372*, 20120476.

[187] Weymuth, T.; Couzijn, E. P. A.; Chen, P.; Reiher, M. New Benchmark Set of Transition-Metal Coordination Reactions for the Assessment of Density Functionals. *J. Chem. Theory Comput.* **2014**, *10*, 3092–3103.

[188] Zhang, W.; Truhlar, D. G.; Tang, M. Tests of Exchange-Correlation Functional Approximations Against Reliable Experimental Data for Average Bond Energies of 3d Transition Metal Compounds. *J. Chem. Theory Comput.* **2013**, *9*, 3965–3977.

[189] Furche, F.; Perdew, J. P. The Performance of Semilocal and Hybrid Density Functionals in 3d Transition-Metal Chemistry. *J. Chem. Phys.* **2006**, *124*, 044103.

[190] Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37*, 785–789.

[191] Becke, A. D. Density-functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

[192] Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38*, 3098–3100.

[193] Perdew, J. P.; Ernzerhof, M.; Burke, K. Rationale for Mixing Exact Exchange with Density Functional Approximations. *J. Chem. Phys.* **1996**, *105*, 9982–9985.

[194] Adamo, C.; Barone, V. Toward Reliable Density Functional Methods without Adjustable Parameters: The PBE0 Model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

[195] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

[196] Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. Comparative Assessment of a New Nonempirical Density Functional: Molecules and Hydrogen-Bonded Complexes. *J. Chem. Phys.* **2003**, *119*, 12129–12137.

[197] Schultz, N. E.; Zhao, Y.; Truhlar, D. G. Databases for Transition Element Bonding: Metal-Metal Bond Energies and Bond Lengths and Their Use To Test Hybrid, Hybrid Meta, and Meta Density Functionals and Generalized Gradient Approximations. *J. Phys. Chem. A* **2005**, *109*, 4388–4403.

[198] Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and Density Functional Theories for the Computation of Enthalpies of Formation. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

[199] Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-3 and Density Functional Theories for a Larger Experimental Test Set. *J. Chem. Phys.* **2000**, *112*, 7374–7383.

[200] Salomon, O.; Reiher, M.; Hess, B. A. Assertion and Validation of the Performance of the B3LYP* Functional for the First Transition Metal Row and the G2 Test Set. *J. Chem. Phys.* **2002**, *117*, 4729–4737.

[201] Zhao, Y.; Lynch, B. J.; Truhlar, D. G. Development and Assessment of a New Hybrid Density Functional Model for Thermochemical Kinetics. *J. Phys. Chem. A* **2004**, *108*, 2715–2719.

[202] Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Assessment of Gaussian-3 and Density-Functional Theories on the G3/05 Test Set of Experimental Energies. *J. Chem. Phys.* **2005**, *123*, 124107.

[203] Riley, K. E.; Merz, K. M. Assessment of Density Functional Theory Methods for the Computation of Heats of Formation and Ionization Potentials of Systems Containing Third Row Transition Metals. *J. Phys. Chem. A* **2007**, *111*, 6044–6053.

[204] Pernot, P.; Civalleri, B.; Presti, D.; Savin, A. Prediction Uncertainty of Density Functional Approximations for Properties of Crystals with Cubic Symmetry. *J. Phys. Chem. A* **2015**, *119*, 5288–5304.

[205] Klopper, W.; Manby, F. R.; Ten-No, S.; Valeev, E. F. R12 Methods in Explicitly Correlated Molecular Electronic Structure Theory. *Int. Rev. Phys. Chem.* **2006**, *25*, 427–468.

[206] Olivares-Amaya, R.; Amador-Bedolla, C.; Hachmann, J.; Atahan-Evrenk, S.; Sánchez-Carrera, R. S.; Vogt, L.; Aspuru-Guzik, A. Accelerated Computational Discovery of High-Performance Materials for Organic Photovoltaics by Means of Cheminformatics. *Energy Environ. Sci.* **2011**, *4*, 4849–4861.

[207] Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241–2251.

[208] Pernot, P.; Désenfant, M.; Hennebelle, F. Model's Output Variance Can Increase When Input Variance Decreases: A Sensitivity Analysis Paradox? 17th International Congress of Metrology. 2015; p 02004.

[209] Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972**, *56*, 2257–2261.

[210] Kruse, H.; Goerigk, L.; Grimme, S. Why the Standard B3LYP/6-31G* Model Chemistry Should Not Be Used in DFT Calculations of Molecular Thermochemistry: Understanding and Correcting the Problem. *J. Org. Chem.* **2012**, *77*, 10824–10834.

[211] Feller, D.; Peterson, K. A.; Crawford, T. D. Sources of Error in Electronic Structure Calculations on Small Chemical Systems. *J. Chem. Phys.* **2006**, *124*, 054107.

[212] Pernot, P. The Parameter Uncertainty Inflation Fallacy. *J. Chem. Phys.* **2017**, *147*, 104102.

[213] Bishop, C. M. *Pattern Recognition and Machine Learning*, 8th ed.; Information Science and Statistics; Springer: New York, 2009.

[214] Chernick, M. R. *Bootstrap Methods: A Practitioner's Guide*, 1st ed.; Wiley-Interscience: New York, 1999.

[215] Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, *112*, 289–320.

[216] Weymuth, T.; Reiher, M. Systematic Dependence of Transition-Metal Coordination Energies on Density-Functional Parametrizations. *Int. J. Quantum Chem.* **2015**, *115*, 90–98.

[217] Pernot, P.; Cailliez, F. A Critical Review of Statistical Calibration/Prediction Models Handling Data Inconsistency and Model Inadequacy. **2016**, arXiv:1611.04376.

[218] Perdew, J. P.; Schmidt, K.; Van Doren, V.; Van Alsenoy, C.; Geerlings, P. Jacob's Ladder of Density Functional Approximations for the Exchange-Correlation Energy. *AIP Conf. Proc.* **2001**, *577*, 1–20.

[219] Sun, J.; Ruzsinszky, A.; Perdew, J. P. Strongly Constrained and Appropriately Normed Semilocal Density Functional. *Phys. Rev. Lett.* **2015**, *115*, 036402.

[220] Medvedev, M. G.; Bushmarinov, I. S.; Sun, J.; Perdew, J. P.; Lyssenko, K. A. Density Functional Theory Is Straying from the Path toward the Exact Functional. *Science* **2017**, *355*, 49–52.

[221] Mardirossian, N.; Head-Gordon, M. ωB97X-V: A 10-Parameter, Range-Separated Hybrid, Generalized Gradient Approximation Density Functional with Nonlocal Correlation, Designed by a Survival-of-the-Fittest Strategy. *Phys. Chem. Chem. Phys.* **2014**, *16*, 9904–9924.

[222] Mardirossian, N.; Head-Gordon, M. Mapping the Genome of Meta-Generalized Gradient Approximation Density Functionals: The Search for B97M-V. *J. Chem. Phys.* **2015**, *142*, 074111.

[223] Pople, J. A.; Head-Gordon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. Gaussian-1 Theory: A General Procedure for Prediction of Molecular Energies. *J. Chem. Phys.* **1989**, *90*, 5622–5629.

[224] Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. Gaussian-2 Theory for Molecular Energies of First- and Second-row Compounds. *J. Chem. Phys.* **1991**, *94*, 7221–7230.

[225] Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. Gaussian-3 (G3) Theory for Molecules Containing First and Second-Row Atoms. *J. Chem. Phys.* **1998**, *109*, 7764–7776.

[226] Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 Theory Using Reduced Order Perturbation Theory. *J. Chem. Phys.* **2007**, *127*, 124105.

[227] Martin, J. M. L.; de Oliveira, G. Towards Standard Methods for Benchmark Quality Ab Initio Thermochemistry—W1 and W2 Theory. *J. Chem. Phys.* **1999**, *111*, 1843–1856.

[228] Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kállay, M.; Gauss, J. W3 Theory: Robust Computational Thermochemistry in the kJ/Mol Accuracy Range. *J. Chem. Phys.* **2004**, *120*, 4129–4141.

[229] Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. W4 Theory for Computational Thermochemistry: In Pursuit of Confident Sub-kJ/Mol Predictions. *J. Chem. Phys.* **2006**, *125*, 144108.

[230] Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vázquez, J.; Stanton, J. F. HEAT: High Accuracy Extrapolated Ab Initio Thermochemistry. *J. Chem. Phys.* **2004**, *121*, 11599–11613.

[231] Rauhut, G.; Pulay, P. Transferable Scaling Factors for Density Functional Derived Vibrational Force Fields. *J. Phys. Chem.* **1995**, *99*, 3093–3100.

[232] Neugebauer, J.; Hess, B. A. Fundamental Vibrational Frequencies of Small Polyatomic Molecules from Density-Functional Calculations and Vibrational Perturbation Theory. *J. Chem. Phys.* **2003**, *118*, 7215–7225.

[233] Irikura, K. K.; Johnson, R. D.; Kacker, R. N. Uncertainties in Scaling Factors for Ab Initio Vibrational Frequencies. *J. Phys. Chem. A* **2005**, *109*, 8430–8437.

[234] Neese, F. Prediction and Interpretation of the ^{57}Fe Isomer Shift in Mössbauer Spectra by Density Functional Theory. *Inorg. Chim. Acta* **2002**, *337*, 181–192.

[235] Han, W.-G.; Liu, T.; Lovell, T.; Noodleman, L. DFT Calculations of ^{57}Fe Mössbauer Isomer Shifts and Quadrupole Splittings for Iron Complexes in Polar Dielectric Media: Applications to Methane Monooxygenase and Ribonucleotide Reductase. *J. Comput. Chem.* **2006**, *27*, 1292–1306.

[236] Römelt, M.; Ye, S.; Neese, F. Calibration of Modern Density Functional Theory Methods for the Prediction of ^{57}Fe Mössbauer Isomer Shifts: Meta-GGA and Double-Hybrid Functionals. *Inorg. Chem.* **2009**, *48*, 784–785.

[237] Bochevarov, A. D.; Friesner, R. A.; Lippard, S. J. Prediction of ^{57}Fe Mössbauer Parameters by Density Functional Theory: A Benchmark Study. *J. Chem. Theory Comput.* **2010**, *6*, 3735–3749.

[238] Gubler, J.; Finkelmann, A. R.; Reiher, M. Theoretical ^{57}Fe Mössbauer Spectroscopy for Structure Elucidation of [Fe] Hydrogenase Active Site Intermediates. *Inorg. Chem.* **2013**, *52*, 14205–14215.

[239] Proppe, J.; Reiher, M. Reliable Estimation of Prediction Uncertainty for Physicochemical Property Models. *J. Chem. Theory Comput.* **2017**, *13*, 3297–3317.

[240] von Lilienfeld, O. A. First Principles View on Chemical Compound Space: Gaining Rigorous Atomistic Control of Molecular Properties. *Int. J. Quantum Chem.* **2013**, *113*, 1676–1689.

[241] Thiel, W. Semiempirical Quantum–Chemical Methods. *WIREs Comput. Mol. Sci.* **2014**, *4*, 145–157.

[242] Wu, X.; Thiel, W.; Pezeshki, S.; Lin, H. Specific Reaction Path Hamiltonian for Proton Transfer in Water: Reparameterized Semiempirical Models. *J. Chem. Theory Comput.* **2013**, *9*, 2672–2686.

[243] Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. Adiabatic Connection for Kinetics. *J. Phys. Chem. A* **2000**, *104*, 4811–4815.

[244] Mortensen, J. J.; Kaasbjerg, K.; Frederiksen, S. L.; Nørskov, J. K.; Sethna, J. P.; Jacobsen, K. W. Bayesian Error Estimation in Density-Functional Theory. *Phys. Rev. Lett.* **2005**, *95*, 216401.

[245] Brown, K. S.; Sethna, J. P. Statistical Mechanical Approaches to Models with Many Poorly Known Parameters. *Phys. Rev. E* **2003**, *68*, 021904.

[246] Frederiksen, S. L.; Jacobsen, K. W.; Brown, K. S.; Sethna, J. P. Bayesian Ensemble Approach to Error Estimation of Interatomic Potentials. *Phys. Rev. Lett.* **2004**, *93*, 165501.

[247] Petzold, V.; Bligaard, T.; Jacobsen, K. W. Construction of New Electronic Density Functionals with Error Estimation Through Fitting. *Top. Catal.* **2012**, *55*, 402–417.

[248] Wellendorff, J.; Lundgaard, K. T.; Møgelhøj, A.; Petzold, V.; Landis, D. D.; Nørskov, J. K.; Bligaard, T.; Jacobsen, K. W. Density Functionals for Surface Science: Exchange-Correlation Model Development with Bayesian Error Estimation. *Phys. Rev. B* **2012**, *85*, 235149.

[249] Wellendorff, J.; Lundgaard, K. T.; Jacobsen, K. W.; Bligaard, T. mBEEF: An Accurate Semi-Local Bayesian Error Estimation Density Functional. *J. Chem. Phys.* **2014**, *140*, 144107.

[250] Pandey, M.; Jacobsen, K. W. Heats of Formation of Solids with Error Estimation: The mBEEF Functional with and without Fitted Reference Energies. *Phys. Rev. B* **2015**, *91*, 235201.

[251] Lynch, B. J.; Truhlar, D. G. Robust and Affordable Multicoefficient Methods for Thermochemistry and Thermochemical Kinetics: The MCCM/3 Suite and SAC/3. *J. Phys. Chem. A* **2003**, *107*, 3898–3906.

[252] Lynch, B. J.; Truhlar, D. G. Small Representative Benchmarks for Thermochemical Calculations. *J. Phys. Chem. A* **2003**, *107*, 8996–8999.

[253] Lynch, B. J.; Zhao, Y.; Truhlar, D. G. Effectiveness of Diffuse Basis Functions for Calculating Relative Energies by Density Functional Theory. *J. Phys. Chem. A* **2003**, *107*, 1384–1388.

[254] Schultz, N. E.; Zhao, Y.; Truhlar, D. G. Density Functionals for Inorganometallic and Organometallic Chemistry. *J. Phys. Chem. A* **2005**, *109*, 11127–11143.

[255] Zhao, Y.; González-García, N.; Truhlar, D. G. Benchmark Database of Barrier Heights for Heavy Atom Transfer, Nucleophilic Substitution, Association, and Unimolecular Reactions and Its Use to Test Theoretical Methods. *J. Phys. Chem. A* **2005**, *109*, 2012–2018.

[256] Zhao, Y.; Truhlar, D. G. Benchmark Databases for Nonbonded Interactions and Their Use To Test Density Functional Theory. *J. Chem. Theory Comput.* **2005**, *1*, 415–432.

[257] Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

[258] Zhao, Y.; Truhlar, D. G. Density Functionals with Broad Applicability in Chemistry. *Acc. Chem. Res.* **2008**, *41*, 157–167.

[259] Zhao, Y.; Truhlar, D. G. Benchmark Energetic Data in a Model System for Grubbs II Metathesis Catalysis and Their Use for the Development, Assessment, and Validation of Electronic Structure Methods. *J. Chem. Theory Comput.* **2009**, *5*, 324–333.

[260] Korth, M.; Grimme, S. "Mindless" DFT Benchmarking. *J. Chem. Theory Comput.* **2009**, *5*, 993–1003.

[261] Goerigk, L.; Grimme, S. A General Database for Main Group Thermochemistry, Kinetics, and Noncovalent Interactions - Assessment of Common and Reparameterized (Meta-)GGA Density Functionals. *J. Chem. Theory Comput.* **2010**, *6*, 107–126.

[262] Goerigk, L.; Grimme, S. Efficient and Accurate Double-Hybrid-Meta-GGA Density Functionals—Evaluation with the Extended GMTKN30 Database for General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions. *J. Chem. Theory Comput.* **2011**, *7*, 291–309.

[263] Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*, 2nd ed.; Wiley: Chichester, West Sussex, England, 2004.

[264] Goerigk, L.; Grimme, S. A Thorough Benchmark of Density Functional Methods for General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6670–6688.

[265] Jaynes, E. T. In *Probability Theory: The Logic of Science*, 1st ed.; Bretthorst, G. L., Ed.; Cambridge University Press: Cambridge, UK, 2003.

[266] Gelman, A.; Carlin, J. B.; Stern, H. S.; Dunson, D. B.; Vehtari, A.; Rubin, D. B. *Bayesian Data Analysis*, 3rd ed.; Chapman and Hall, 2013.

[267] Aldegunde, M.; Kermode, J. R.; Zabaras, N. Development of an Exchange–Correlation Functional with Uncertainty Quantification Capabilities for Density Functional Theory. *J. Comput. Phys.* **2016**, *311*, 173–195.

[268] Sutton, J. E.; Guo, W.; Katsoulakis, M. A.; Vlachos, D. G. Effects of Correlated Parameters and Uncertainty in Electronic-Structure-Based Chemical Kinetic Modelling. *Nat. Chem.* **2016**, *8*, 331–337.

[269] Gautier, S.; Steinmann, S. N.; Michel, C.; Fleurat-Lessard, P.; Sautet, P. Molecular Adsorption at Pt(111). How Accurate Are DFT Functionals? *Phys. Chem. Chem. Phys.* **2015**, *17*, 28921–28930.

[270] Reiher, M.; Salomon, O.; Hess, B. A. Reparameterization of Hybrid Functionals Based on Energy Differences of States of Different Multiplicity. *Theor. Chem. Acc.* **2001**, *107*, 48–55.

[271] Reiher, M. Theoretical Study of the $Fe(Phen)_2(NCS)_2$ Spin-Crossover Complex with Reparametrized Density Functionals. *Inorg. Chem.* **2002**, *41*, 6928–6935.

[272] Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Insights into Current Limitations of Density Functional Theory. *Science* **2008**, *321*, 792–794.

[273] Baer, R.; Livshits, E.; Salzner, U. Tuned Range-Separated Hybrids in Density Functional Theory. *Annu. Rev. Phys. Chem.* **2010**, *61*, 85–109.

[274] Peach, M. J. G.; Helgaker, T.; Sałek, P.; Keal, T. W.; Lutnæs, O. B.; Tozer, D. J.; Handy, N. C. Assessment of a Coulomb-Attenuated Exchange–Correlation Energy Functional. *Phys. Chem. Chem. Phys.* **2006**, *8*, 558–562.

[275] Vydrov, O. A.; Heyd, J.; Krukau, A. V.; Scuseria, G. E. Importance of Short-Range versus Long-Range Hartree-Fock Exchange for the Performance of Hybrid Density Functionals. *J. Chem. Phys.* **2006**, *125*, 074106.

[276] Rohrdanz, M. A.; Herbert, J. M. Simultaneous Benchmarking of Ground- and Excited-State Properties with Long-Range-Corrected Density Functional Theory. *J. Chem. Phys.* **2008**, *129*, 034107.

[277] Stein, T.; Kronik, L.; Baer, R. Prediction of Charge-Transfer Excitations in Coumarin-Based Dyes Using a Range-Separated Functional Tuned from First Principles. *J. Chem. Phys.* **2009**, *131*, 244119.

[278] Srebro, M.; Autschbach, J. Does a Molecule-Specific Density Functional Give an Accurate Electron Density? The Challenging Case of the CuCl Electric Field Gradient. *J. Phys. Chem. Lett.* **2012**, *3*, 576–581.

[279] Srebro, M.; Autschbach, J. Tuned Range-Separated Time-Dependent Density Functional Theory Applied to Optical Rotation. *J. Chem. Theory Comput.* **2012**, *8*, 245–256.

[280] Autschbach, J.; Srebro, M. Delocalization Error and "Functional Tuning" in Kohn–Sham Calculations of Molecular Properties. *Acc. Chem. Res.* **2014**, *47*, 2592–2602.

[281] Leininger, T.; Stoll, H.; Werner, H.-J.; Savin, A. Combining Long-Range Configuration Interaction with Short-Range Density Functionals. *Chem. Phys. Lett.* **1997**, *275*, 151–160.

[282] Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. A Long-Range Correction Scheme for Generalized-Gradient-Approximation Exchange Functionals. *J. Chem. Phys.* **2001**, *115*, 3540–3544.

[283] Heyd, J.; Scuseria, G. E.; Ernzerhof, M. Hybrid Functionals Based on a Screened Coulomb Potential. *J. Chem. Phys.* **2003**, *118*, 8207–8215.

[284] Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. A Long-Range-Corrected Time-Dependent Density Functional Theory. *J. Chem. Phys.* **2004**, *120*, 8425–8433.

[285] Yanai, T.; Tew, D. P.; Handy, N. C. A New Hybrid Exchange–Correlation Functional Using the Coulomb-Attenuating Method (CAM-B3LYP). *Chem. Phys. Lett.* **2004**, *393*, 51–57.

[286] Arbuznikov, A. V.; Kaupp, M. Importance of the Correlation Contribution for Local Hybrid Functionals: Range Separation and Self-Interaction Corrections. *J. Chem. Phys.* **2012**, *136*, 014111.

[287] Lieb, E. H.; Oxford, S. Improved Lower Bound on the Indirect Coulomb Energy. *Int. J. Quantum Chem.* **1981**, *19*, 427–439.

[288] Zhang, Y.; Yang, W. Comment on "Generalized Gradient Approximation Made Simple". *Phys. Rev. Lett.* **1998**, *80*, 890–890.

[289] Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L. A.; Zhou, X.; Burke, K. Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces. *Phys. Rev. Lett.* **2008**, *100*, 136406.

[290] Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L. A.; Zhou, X.; Burke, K. Erratum: Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces [Phys. Rev. Lett. **100**, 136406 (2008)]. *Phys. Rev. Lett.* **2009**, *102*, 039902.

[291] Constantin, L. A.; Fabiano, E.; Laricchia, S.; Della Sala, F. Semiclassical Neutral Atom as a Reference System in Density Functional Theory. *Phys. Rev. Lett.* **2011**, *106*, 186406.

[292] Fabiano, E.; Constantin, L. A.; Della Sala, F. Two-Dimensional Scan of the Performance of Generalized Gradient Approximations with Perdew–Burke–Ernzerhof-Like Enhancement Factor. *J. Chem. Theory Comput.* **2011**, *7*, 3548–3559.

[293] Byrd, R.; Lu, P.; Nocedal, J.; Zhu, C. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.* **1995**, *16*, 1190–1208.

[294] Burke, K. Perspective on Density Functional Theory. *J. Chem. Phys.* **2012**, *136*, 150901.

[295] Chatt, J.; Dilworth, J. R.; Richards, R. L. Recent Advances in the Chemistry of Nitrogen Fixation. *Chem. Rev.* **1978**, *78*, 589–625.

[296] Magistrato, A.; Robertazzi, A.; Carloni, P. Nitrogen Fixation by a Molybdenum Catalyst Mimicking the Function of the Nitrogenase Enzyme: A Critical Evaluation of DFT and Solvent Effects. *J. Chem. Theory Comput.* **2007**, *3*, 1708–1720.

[297] Simm, G. N.; Reiher, M. Systematic Error Estimation for Chemical Reaction Energies. *J. Chem. Theory Comput.* **2016**, *12*, 2762–2773.

[298] Jiang, W.; DeYonker, N. J.; Determan, J. J.; Wilson, A. K. Toward Accurate Theoretical Thermochemistry of First Row Transition Metal Complexes. *J. Phys. Chem. A* **2012**, *116*, 870–885.

[299] Agrawal, P.; Tkatchenko, A.; Kronik, L. Pair-Wise and Many-Body Dispersive Interactions Coupled to an Optimally Tuned Range-Separated Hybrid Functional. *J. Chem. Theory Comput.* **2013**, *9*, 3473–3478.

[300] Steinmann, S. N.; Corminboeuf, C. A Generalized-Gradient Approximation Exchange Hole Model for Dispersion Coefficients. *J. Chem. Phys.* **2011**, *134*, 044117.

[301] Zhu, C.; Byrd, R. H.; Lu, P.; Nocedal, J. Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization. *ACM Trans. Math. Softw.* **1997**, *23*, 550–560.

[302] Valorani, M.; Goussis, D. A.; Creta, F.; Najm, H. N. Higher Order Corrections in the Approximation of Low-Dimensional Manifolds and the Construction of Simplified Problems with the CSP Method. *J. Comput. Phys.* **2005**, *209*, 754–786.

[303] Turányi, T.; Tomlin, A. S. *Analysis of Kinetic Reaction Mechanisms*; Springer Berlin Heidelberg, 2014; pp 183–312.

[304] Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J. Chem. Phys.* **2011**, *134*, 174105.

[305] Bowman, G. R. In *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*; Bowman, G. R., Pande, V. S., Noé, F., Eds.; Advances in Experimental Medicine and Biology 797; Springer Netherlands, 2014; pp 7–22.

[306] Lam, S. H.; Goussis, D. A. The CSP Method for Simplifying Kinetics. *Int. J. Chem. Kinet.* **1994**, *26*, 461–486.

[307] Kourdis, P. D.; Goussis, D. A. Glycolysis in Saccharomyces Cerevisiae: Algorithmic Exploration of Robustness and Origin of Oscillations. *Math. Biosci.* **2013**, *243*, 190–214.

[308] Schwartz, A. W.; de Graaf, R. M. The Prebiotic Synthesis of Carbohydrates: A Reassessment. *J. Mol. Evol.* **1993**, *36*, 101–106.

[309] Baly, E. C. C. Photosynthesis. *Ind. Eng. Chem.* **1924**, *16*, 1016–1018.

[310] Meinert, C.; Myrgorodska, I.; de Marcellus, P.; Buhse, T.; Nahon, L.; Hoffmann, S. V.; d'Hendecourt, L. L. S.; Meierhenrich, U. J. Ribose and Related Sugars from Ultraviolet Irradiation of Interstellar Ice Analogs. *Science* **2016**, *352*, 208–212.

[311] Kua, J.; Galloway, M. M.; Millage, K. D.; Avila, J. E.; De Haan, D. O. Glycolaldehyde Monomer and Oligomer Equilibria in Aqueous Solution: Comparing Computational Chemistry and NMR Data. *J. Phys. Chem. A* **2013**, *117*, 2997–3008.

[312] Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

[313] Balabin, R. M.; Lomakina, E. I. Neural Network Approach to Quantum-Chemistry Data: Accurate Prediction of Density Functional Theory Energies. *J. Chem. Phys.* **2009**, *131*, 074104.

[314] Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

[315] Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

[316] Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New J. Phys.* **2013**, *15*, 095003.

[317] Bartók, A. P.; Gillan, M. J.; Manby, F. R.; Csányi, G. Machine-Learning Approach for One- and Two-Body Corrections to Density Functional Theory: Applications to Molecular and Condensed Water. *Phys. Rev. B* **2013**, *88*, 054104.

[318] Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404–3419.

[319] Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to Represent Crystal Structures for Machine Learning: Towards Fast Prediction of Electronic Properties. *Phys. Rev. B* **2014**, *89*, 205118.

[320] Dral, P. O.; von Lilienfeld, O. A.; Thiel, W. Machine Learning of Parameters for Accurate Semiempirical Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2015**, *11*, 2120–2125.

[321] De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769.

[322] Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264.

[323] Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine Learning of Accurate Energy-Conserving Molecular Force Fields. *Sci. Adv.* **2017**, *3*, e1603015.

[324] Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*, 13890.

[325] Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3*, e1701816.

[326] Behler, J.; Martoňák, R.; Donadio, D.; Parrinello, M. Metadynamics Simulations of the High-Pressure Phases of Silicon Employing a High-Dimensional Neural Network Potential. *Phys. Rev. Lett.* **2008**, *100*, 185501.

[327] Handley, C. M.; Popelier, P. L. A. Potential Energy Surfaces Fitted by Artificial Neural Networks. *J. Phys. Chem. A* **2010**, *114*, 3371–3383.

[328] Botu, V.; Ramprasad, R. Adaptive Machine Learning Framework to Accelerate Ab Initio Molecular Dynamics. *Int. J. Quantum Chem.* **2015**, *115*, 1074–1083.

[329] Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331.

[330] Shen, L.; Wu, J.; Yang, W. Multiscale Quantum Mechanics/Molecular Mechanics Simulations with Neural Networks. *J. Chem. Theory Comput.* **2016**, *12*, 4934–4946.

[331] Li, Y.; Li, H.; Pickard, F. C.; Narayanan, B.; Sen, F. G.; Chan, M. K. Y.; Sankaranarayanan, S. K. R. S.; Brooks, B. R.; Roux, B. Machine Learning Force Field Parameters from Ab Initio Data. *J. Chem. Theory Comput.* **2017**, *13*, 4492–4503.

[332] Bleiziffer, P.; Schaller, K.; Riniker, S. Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *J. Chem. Inf. Model.* **2018**, *58*, 579–590.

[333] Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards Exact Molecular Dynamics Simulations with Machine-Learned Force Fields. **2018**, arXiv:1802.09238.

[334] Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.

[335] Zhou, Z.; Li, X.; Zare, R. N. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **2017**, *3*, 1337–1344.

[336] Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Cent. Sci.* **2017**, *3*, 283–293.

[337] Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3*, 1103–1113.

[338] Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120–131.

[339] Kayala, M. A.; Azencott, C.-A.; Chen, J. H.; Baldi, P. Learning to Predict Chemical Reactions. *J. Chem. Inf. Model.* **2011**, *51*, 2209–2222.

[340] Sadowski, P.; Fooshee, D.; Subrahmanya, N.; Baldi, P. Synergies Between Quantum Mechanics and Machine Learning in Reaction Prediction. *J. Chem. Inf. Model.* **2016**, *56*, 2125–2128.

[341] Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725–732.

[342] Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533*, 73–76.

[343] Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; pp 2607–2616.

[344] Fooshee, D.; Mood, A.; Gutman, E.; Tavakoli, M.; Urban, G.; Liu, F.; Huynh, N.; Van Vranken, D.; Baldi, P. Deep Learning for Chemical Reaction Prediction. *Mol. Syst. Des. Eng.* **2018**, *3*, 442–452.

[345] Bradshaw, J.; Kusner, M. J.; Paige, B.; Segler, M. H. S.; Hernández-Lobato, J. M. Predicting Electron Paths. **2018**, arXiv:1805.10970.

[346] Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning. *Science* **2018**, eaar5169.

[347] Rupp, M. Machine Learning for Quantum Mechanics in a Nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058–1073.

[348] Ramakrishnan, R.; von Lilienfeld, O. A. *Reviews in Computational Chemistry*; Wiley-Blackwell, 2017; Vol. 30; pp 225–256.

[349] Csányi, G.; Albaret, T.; Payne, M. C.; De Vita, A. "Learn on the Fly": A Hybrid Classical and Quantum-Mechanical Molecular Dynamics Simulation. *Phys. Rev. Lett.* **2004**, *93*, 175503.

[350] Li, Z.; Kermode, J. R.; De Vita, A. Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces. *Phys. Rev. Lett.* **2015**, *114*, 096405.

[351] Glielmo, A.; Sollich, P.; De Vita, A. Accurate Interatomic Force Fields via Machine Learning with Covariant Kernels. *Phys. Rev. B* **2017**, *95*, 214302.

[352] Glielmo, A.; Zeni, C.; De Vita, A. Efficient Non-Parametric n-Body Force Fields from Machine Learning. **2018**, arXiv:1801.04823.

[353] Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K. To Address Surface Reaction Network Complexity Using Scaling Relations Machine Learning and DFT Calculations. *Nat. Commun.* **2017**, *8*, 14621.

[354] Peterson, A. A.; Christensen, R.; Khorshidi, A. Addressing Uncertainty in Atomistic Machine Learning. *Phys. Chem. Chem. Phys.* **2017**, *19*, 10978–10985.

[355] Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure–Property Relationships. *J. Phys. Chem. A* **2017**, *121*, 8939–8954.

[356] Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.

[357] Mercer, J. XVI. Functions of Positive and Negative Type, and Their Connection the Theory of Integral Equations. *Philos. Trans. R. Soc. London, Ser. A* **1909**, *209*, 415–446.

[358] Ángyán, J. G.; Jansen, G.; Loss, M.; Hättig, C.; Heß, B. A. Distributed Polarizabilities Using the Topological Theory of Atoms in Molecules. *Chem. Phys. Lett.* **1994**, *219*, 267–273.

[359] Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B* **2013**, *87*, 184115.

[360] Ferré, G.; Haut, T.; Barros, K. Learning Molecular Energies Using Localized Graph Kernels. *J. Chem. Phys.* **2017**, *146*, 114107.

[361] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 140022.

[362] Simm, G. N.; Reiher, M. Error-Controlled Exploration of Chemical Reaction Networks with Gaussian Processes. *J. Chem. Theory Comput.* **2018**, accepted.

[363] Janet, J. P.; Kulik, H. J. Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks. *Chem. Sci.* **2017**, *8*, 5137–5152.

[364] Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *J. Phys. Chem. Lett.* **2018**, *9*, 1064–1071.

[365] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.

[366] Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J. Mol. Model.* **2013**, *19*, 1–32.

[367] Proppe, J.; Reiher, M. Mechanism Deduction from Noisy Chemical Reaction Networks. **2018**, arXiv:1803.09346.

[368] Riplinger, C.; Neese, F. An Efficient and near Linear Scaling Pair Natural Orbital Based Local Coupled Cluster Method. *J. Chem. Phys.* **2013**, *138*, 034106.

[369] Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F. Sparse Maps—A Systematic Infrastructure for Reduced-Scaling Electronic Structure Methods. II. Linear Scaling Domain Based Pair Natural Orbital Coupled Cluster Theory. *J. Chem. Phys.* **2016**, *144*, 024109.

[370] Keller, S.; Dolfi, M.; Troyer, M.; Reiher, M. An Efficient Matrix Product Operator Representation of the Quantum Chemical Hamiltonian. *J. Chem. Phys.* **2015**, *143*, 244118.

[371] Hedegård, E. D.; Knecht, S.; Kielberg, J. S.; Jensen, H. J. A.; Reiher, M. Density Matrix Renormalization Group with Efficient Dynamical Electron Correlation through Range Separation. *J. Chem. Phys.* **2015**, *142*, 224108.

[372] Stein, C. J.; Reiher, M. Automated Selection of Active Orbital Spaces. *J. Chem. Theory Comput.* **2016**, *12*, 1760–1771.

[373] Stein, C. J.; Reiher, M. Measuring Multi-Configurational Character by Orbital Entanglement. *Mol. Phys.* **2017**, *115*, 2110–2119.

[374] Stein, C. J.; von Burg, V.; Reiher, M. The Delicate Balance of Static and Dynamic Electron Correlation. *J. Chem. Theory Comput.* **2016**, *12*, 3764–3773.

[375] Stein, C. J.; Reiher, M. Automated Identification of Relevant Frontier Orbitals for Chemical Compounds and Processes. *Chimia* **2017**, *71*, 170–176.

[376] MongoDB Inc., MongoDB 3.2. www.mongodb.com, (Accessed: 30. September 2016).

[377] McKinney, W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. 2010; pp 51 – 56.

[378] Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95.

[379] Gansner, E. R.; North, S. C. An Open Graph Visualization System and Its Applications to Software Engineering. *Softw: Pract. Exper.* **2000**, *30*, 1203–1233.

[380] Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X.; Ghosh, D.; Goldey, M.; Horn, P. R.; Jacobson, L. D.; Kaliman, I.; Khaliullin, R. Z.; Kuś, T.; Landau, A.; Liu, J.; Proynov, E. I.; Rhee, Y. M.; Richard, R. M.; Rohrdanz, M. A.; Steele, R. P.; Sundstrom, E. J.; III, H. L. W.; Zimmerman, P. M.; Zuev, D.; Albrecht, B.; Alguire, E.; Austin, B.; Beran, G. J. O.; Bernard, Y. A.; Berquist, E.; Brandhorst, K.; Bravaya, K. B.; Brown, S. T.; Casanova, D.; Chang, C.-M.; Chen, Y.; Chien, S. H.; Closser, K. D.; Crittenden, D. L.; Diedenhofen, M.; Jr, R. A. D.; Do, H.; Dutoi, A. D.; Edgar, R. G.; Fatehi, S.; Fusti-Molnar, L.; Ghysels, A.; Golubeva-Zadorozhnaya, A.; Gomes, J.; Hanson-Heine, M. W. D.; Harbach, P. H. P.; Hauser, A. W.; Hohenstein, E. G.; Holden, Z. C.; Jagau, T.-C.; Ji, H.; Kaduk, B.; Khistyaev, K.; Kim, J.; Kim, J.; King, R. A.; Klunzinger, P.; Kosenkov, D.; Kowalczyk, T.; Krauter, C. M.; Lao, K. U.; Laurent, A. D.; Lawler, K. V.; Levchenko, S. V.; Lin, C. Y.; Liu, F.; Livshits, E.; Lochan, R. C.; Luenser, A.; Manohar, P.; Manzer, S. F.; Mao, S.-P.; Mardirossian, N.; Marenich, A. V.; Maurer, S. A.; Mayhall, N. J.; Neuscamman, E.; Oana, C. M.; Olivares-Amaya, R.; O'Neill, D. P.; Parkhill, J. A.; Perrine, T. M.; Peverati, R.; Prociuk, A.; Rehn, D. R.; Rosta, E.; Russ, N. J.; Sharada, S. M.; Sharma, S.; Small, D. W.; Sodt, A.; Stein, T.; Stück, D.; Su, Y.-C.; Thom, A. J. W.; Tsuchimochi, T.; Vanovschi, V.; Vogt, L.; Vydrov, O.; Wang, T.; Watson, M. A.; Wenzel, J.; White, A.; Williams, C. F.; Yang, J.; Yeganeh, S.; Yost, S. R.; You, Z.-Q.; Zhang, I. Y.; Zhang, X.; Zhao, Y.; Brooks, B. R.; Chan, G. K. L.; Chipman, D. M.; Cramer, C. J.; III, W. A. G.; Gordon, M. S.; Hehre, W. J.; Klamt, A.; III, H. F. S.; Schmidt, M. W.; Sherrill, C. D.; Truhlar, D. G.; Warshel, A.; Xu, X.; Aspuru-Guzik, A.; Baer, R.; Bell, A. T.; Besley, N. A.; Chai, J.-D.; Dreuw, A.; Dunietz, B. D.; Furlani, T. R.; Gwaltney, S. R.; Hsu, C.-P.; Jung, Y.; Kong, J.; Lambrecht, D. S.; Liang, W.; Ochsenfeld, C.; Rassolov, V. A.; Slipchenko, L. V.; Subotnik, J. E.; Voorhis, T. V.; Herbert, J. M.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *113*, 184–215.

[381] Dunning, T. H. Gaussian Basis Functions for Use in Molecular Calculations. I. Contraction of (9s5p) Atomic Basis Sets for the First-Row Atoms. *J. Chem. Phys.* **1970**, *53*, 2823–2833.

[382] Perdew, J. P. Density-Functional Approximation for the Correlation Energy of the Inhomogeneous Electron Gas. *Phys. Rev. B* **1986**, *33*, 8822–8824.

[383] Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

[384] Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. Electronic Structure Calculations on Workstation Computers: The Program System Turbomole. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

[385] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.;

Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian09 Revision D.01. Gaussian Inc. Wallingford CT 2009.

[386] Schaftenaar, G.; Noordik, J. H. Molden: A Pre- and Post-Processing Program for Molecular and Electronic Structures. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 123–134.

[387] Hanson, R. M. Jmol – a Paradigm Shift in Crystallographic Visualization. *J. Appl. Crystallogr.* **2010**, *43*, 1250–1260.

[388] Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M. Molpro: A General-purpose Quantum Chemistry Program Package. *WIREs Comput. Mol. Sci.* **2012**, *2*, 242–253.

[389] Woon, D. E.; Jr, T. H. D. Gaussian Basis Sets for Use in Correlated Molecular Calculations. IV. Calculation of Static Electrical Response Properties. *J. Chem. Phys.* **1994**, *100*, 2975–2988.

[390] Peterson, K. A.; Figgen, D.; Dolg, M.; Stoll, H. Energy-Consistent Relativistic Pseudopotentials and Correlation Consistent Basis Sets for the 4d Elements Y–Pd. *J. Chem. Phys.* **2007**, *126*, 124101.

[391] Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. A. NWChem: A Comprehensive and Scalable Open-Source Solution for Large Scale Molecular Simulations. *Comput. Phys. Commun.* **2010**, *181*, 1477–1489.

[392] Zhao, Y.; Truhlar, D. G. A New Local Density Functional for Main-Group Thermochemistry, Transition Metal Bonding, Thermochemical Kinetics, and Noncovalent Interactions. *J. Chem. Phys.* **2006**, *125*, 194101.

[393] Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta–Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91*, 146401.

[394] Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

[395] Andrae, D.; Häußermann, U.; Dolg, M.; Stoll, H.; Preuß, H. Energy-Adjusted Ab Initio Pseudopotentials for the Second and Third Row Transition Elements. *Theoret. Chim. Acta* **1990**, *77*, 123–141.

[396] Adler, T. B.; Werner, H.-J. An Explicitly Correlated Local Coupled Cluster Method for Calculations of Large Molecules Close to the Basis Set Limit. *J. Chem. Phys.* **2011**, *135*, 144117.

[397] Peterson, K. A.; Adler, T. B.; Werner, H.-J. Systematically Convergent Basis Sets for Explicitly Correlated Wavefunctions: The Atoms H, He, B–Ne, and Al–Ar. *J. Chem. Phys.* **2008**, *128*, 084102.

[398] Weigend, F.; Köhn, A.; Hättig, C. Efficient Use of the Correlation Consistent Basis Sets in Resolution of the Identity MP2 Calculations. *J. Chem. Phys.* **2002**, *116*, 3175–3183.

[399] Weigend, F. A Fully Direct RI-HF Algorithm: Implementation, Optimised Auxiliary Basis Sets, Demonstration of Accuracy and Efficiency. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285–4291.

[400] Zhao, Y.; Truhlar, D. G. Benchmark Data for Interactions in Zeolite Model Complexes and Their Use for Assessment and Validation of Electronic Structure Methods. *J. Phys. Chem. C* **2008**, *112*, 6860–6868.

[401] Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

[402] Stewart, J. MOPAC 2016. http://openmopac.net/, (Accessed: 20. April 2018).

[403] GPy, GPy: A Gaussian Process Framework in Python. http://github.com/SheffieldML/GPy, (Accessed: 18. November 2017).

Publications

The following publications (listed in chronological order) are included in parts or in an extended version in this dissertation:

- M. Bergeler, G. N. Simm, J. Proppe, M. Reiher "Heuristics-Guided Exploration of Reaction Mechanisms", *J. Chem. Theory Comput.* **2015**, *11*, 5712–5722.
- G. N. Simm, M. Reiher "Systematic Error Estimation for Chemical Reaction Energies", *J. Chem. Theory Comput.* **2016**, *12*, 2762–2773.
- J. Proppe, T. Husch, G. N. Simm, M. Reiher "Uncertainty Quantification for Quantum Chemical Models of Complex Reaction Networks", *Faraday Discuss.* **2016**, *195*, 497–520.
- G. N. Simm, J. Proppe, M. Reiher "Error Assessment of Computational Models in Chemistry", *CHIMIA* **2017**, 71, 202–208.
- G. N. Simm, M. Reiher "Context-Driven Exploration of Complex Chemical Reaction Networks", *J. Chem. Theory Comput.* **2017**, *13*, 6108–6119.
- G. N. Simm, M. Reiher "Error-Controlled Exploration of Chemical Reaction Networks with Gaussian Processes", arXiv:1805.09886.

The following publications (listed in chronological order) cover the research I did before pursuing my Ph.D. at ETH Zurich.

- E. O. Pyzer-Knapp, G. N. Simm, A. Aspuru-Guzik "A Bayesian Approach to Calibrating High-throughput Virtual Screening Results and Application to Organic Photovoltaic Materials", *Mater. Horiz.* **2016**, *3*, 226–233.
- S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Hachmann, A. Aspuru-Guzik "The Harvard Organic Photovoltaic Dataset", *Mater. Horiz.* **2016**, *3*, 160086.

Gregor Nils Christoph Simm

Personal details

Date of birth	December 20, 1991
Place of birth	Erlangen, Germany
Nationality	German

Education

2015–2018	Doctoral Studies in Chemistry, ETH Zurich, Switzerland
	Supervisor: Prof. Dr. Markus Reiher
2015	MSc in Interdisciplinary Sciences, ETH Zurich, Switzerland
2014–2015	Master's Thesis, Harvard University, USA
	Supervisor: Prof. Dr. Alán Aspuru-Guzik
2013	BSc in Interdisciplinary Sciences, ETH Zurich, Switzerland

Teaching

2015–2017	Teaching Assistant: *Introduction to Computer Science*
	ETH Zurich, Switzerland
2015–2017	Teaching Assistant: *Chemie für Rechnergestützte Wissenschaften*
	ETH Zurich, Switzerland
2014	Teaching Assistant: *Allgemeine Chemie Anorganische Chemie I*
	ETH Zurich, Switzerland
2014	Teaching Assistant: *Allgemeine Chemie Anorganische Chemie II*
	ETH Zurich, Switzerland
2013	Teaching Assistant: *Anorganische Chemie II*
	ETH Zurich, Switzerland

Awards and Distinctions

2016	Prize for best poster presentation at the Swiss Chemical Society Fall Meeting
2015–2017	Ph.D. Fellowship of the Fonds der Chemischen Industry
2015	Master's degree with distinction

Scientific Talks

2017	53rd Symposium on Theoretical Chemistry, Basel. Title: *Automated Exploration of Complex Chemical Reaction Networks.*
2017	Competence Center for Computational Chemistry Meeting, IBM Research, Rüschlikon. Title: *Heuristics-Guided Exploration of Reaction Mechanisms.*