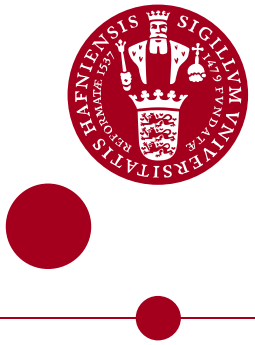
UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE



PhD thesis

# Computational Evolution of Efficient Catalysts

## Exploring chemical space beyond enumerated libraries

Julius Seumer

Advisor: Jan H. Jensen

Submitted: 31.05.2024

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen

PHD THESIS

# COMPUTATIONAL EVOLUTION OF EFFICIENT CATALYSTS

## EXPLORING CHEMICAL SPACE BEYOND ENUMERATE LIBRARIES

AUTHOR:
Julius Seumer
Department of Chemistry, University of Copenhagen

SUPERVISOR:
Professor Jan H. Jensen,
Department of Chemistry, University of Copenhagen

This thesis has been submitted to the PhD School
of the Faculty of Science, University of Copenhagen

Submission date: May 31, 2024

# ABSTRACT

The design of novel catalysts is an active field of chemical research, crucial for approximately 90% of industrial chemical processes. More efficient catalysts have the potential to decrease energy consumption, increase reaction yields, and enable currently unfeasible reactions, particularly those relevant to the green energy transition, such as power-to-x processes and carbon capture. Historically, the discovery of novel catalysts has been driven by trial and error and empirical observations.

Since the advent of reliable computational chemistry tools, high-throughput virtual screening and optimization algorithms, such as genetic algorithms, have been used to explore defined chemical spaces for promising catalysts. Relevant chemical constraints, such as stability and synthesizability, can be considered through careful selection of these chemical spaces. However, this approach does not facilitate the de novo discovery of catalysts.

In this context, machine learning-based tools offer a promising avenue for discovering novel chemical motifs and molecules. The real-world impact of these models on the design of efficient catalysts remains to be seen, as there is often no attempt at computational or experimental verification.

In the first part of this thesis, we present a method for the de novo discovery of efficient catalysts, moving beyond predefined chemical spaces using a graph-based genetic algorithm approach. We explicitly incorporate relevant chemical constraints and verify the success of the optimization computationally. Furthermore, we synthesize the catalyst and experimentally confirm its superior performance. This work represents a significant advancement towards more effective and efficient de novo catalyst design and its real-world application. Furthermore, we extend the approach to handle transition metal-based catalysts and show that we can efficiently find promising catalysts.

The second part of this thesis introduces an automated, fast, and user-friendly workflow designed to predict the regioselectivity of catalyzed C—H activations with directing groups. This workflow leverages semi-empirical quantum mechanical calculations to provide accurate predictions efficiently. By automating complex computational tasks, this approach streamlines the process of determining regioselectivity, making it accessible to researchers without extensive expertise in computational chemistry.

# DANSK RESUMÉ

Design af nye katalysatorer er et aktivt forskningsområde indenfor kemi og afgørende for cirka 90% af de kemiske processer i industrien. Ved at skabe mere effektive katalysatorer kan man reducere energiforbruget, øge reaktionsudbyttet og gennemføre reaktioner der ikke tidligere har været mulige. Dette er blandt andet særligt relevant for den grønne omstilling, hvor ny og bedre kemi er væsentlig til power-to-x og kulstoffangst. Historisk set har opdagelsen af nye katalysatorer været drevet af "trial and error" og empiriske observation.

Men med pålidelige og avancerede kvantekemiske beregningsværktøjer kan vi nu udføre virtuelle undersøgelser i stor skala og bruge optimeringsalgoritmer, såsom genetiske algoritmer, til at udforske veldefinerede kemiske rum for lovende katalysatorer. Desuden kan man tage højde for relevante kemiske problematikker såsom stabilitet og syntetiserbarhed ved omhyggelig at udvælge det kemiske rum. Denne tilgang letter dog ikke opdagelsen af helt nye katalysatorer også kaldet 'de novo design'.

I denne sammenhæng tilbyder maskinlæringsbaserede værktøjer en mere lovende vej til opdagelsen af nye kemiske strukturer og molekyler. Disse modellers indflydelse på den virkelige verden er dog stadig ikke klarlagt, da der ofte ikke fremvises nogen beregningsmæssig eller eksperimentel verification.

I første del af denne PhD afhandling præsenteres en metode til de novo design af effektive katalysatorer, som kan finde molekyler udenfor det foruddefinerede kemiske rum ved hjælp af en graf-baseret genetisk algoritme. Vi har inkorporeret et udvalg af metoder til at tackle de førnævnte kemiske problematikker og verificerer optimeringens succes ved brug af kvantekemiske beregninger. Desuden har vi syntetiseret en nyopdaget katalysator og eksperimentelt bekræftet dens overlegne ydeevne. Dette arbejde repræsenterer et betydeligt fremskridt hen imod en mere effektiv opdagelse nye katalysatorer med anvendelse i den virkelige verden. Desuden udvides tilgangen til at kunne håndtere katalysatorer baseret på overgangsmetaller og viser, at vi også her effektivt kan finde nye lovende katalysatorer.

Anden del af denne PhD afhandling introducerer en automatiseret, hurtig og brugervenlig metode til at forudsige regioselektiviteten af katalyseret C−H-aktiveringer med styregrupper. Denne metode benytter semi-empiriske kvantekemiske beregninger til effektivt at give præcise forudsigelser. Ved at automatisere komplekse beregningsopgaver, strømlines processen til at bestemme regioselektivitet, hvilket gør metoden tilgængeligt for forskere uden omfattende erfaring.

# LIST OF PUBLICATIONS

**Paper 1** **Julius Seumer**, Jonathan Kirschner Solberg Hansen, Mogens Brøndsted Nielsen, Jan H Jensen. "Computational evolution of new catalysts for the Morita–Baylis–Hillman reaction"
In: *Angewandte Chemie*, **2023**, 62 (18).
DOI: 10.1002/anie.202218565

**Paper 2** **Julius Seumer** and Jan H. Jensen. "Beyond Predefined Ligand Libraries: A Genetic Algorithm Approach for De Novo Discovery of Catalysts for the Suzuki Coupling Reactions"
*Preprint*, **2024**.
DOI: 10.26434/chemrxiv-2024-9xh38

**Paper 3** **Julius Seumer** and Jan H. Jensen. "Enhancing Chemical Synthesis Planning: Automated Quantum Mechanics Based Regioselectivity Prediction for C-H Activation with Directing Groups"
*Draft*, **2024**.

The following publications are not part of the main text of this PhD thesis but Paper 4 is included in the appendix:

**Paper 4** Maria H Rasmussen, **Julius Seumer**, Jan H Jensen. "Toward De Novo Catalyst Discovery: Fast Identification of New Catalyst Candidates for Alcohol-Mediated Morita-Baylis-Hillman Reactions"
In: *Angewandte Chemie*, **2023**, 135 (49).
DOI: 10.1002/anie.202310580

**Paper 5** Magnus Strandgaard, **Julius Seumer**, Bardi Benediktsson, Arghya Bhowmik, Tejs Vegge, Jan H Jensen "Genetic algorithm-based re-optimization of the Schrock catalyst for dinitrogen fixation"
In: *PeerJ physical chemistry*, **2023**, 5.e30.
DOI: 10.7717/peerj-pchem.30

**Paper 6** Magnus Strandgaard, **Julius Seumer**, Jan H Jensen "Discovery of molybdenum based nitrogen fixation catalysts with genetic algorithms"
*Preprint*, **2024**.
DOI: 10.26434/chemrxiv-2024-p835f-v2

# ACKNOWLEDGMENTS

I would like to begin by expressing my gratitude to my supervisor, Jan. Your support, guidance, and willingness to discuss ideas have been invaluable throughout my research journey. Thank you for providing insightful feedback and always making time for me.

I am also thankful to my colleagues, Maria, Mads, Nicolai, Magnus, Rasmus, Dominik, and Jacob. I really enjoyed working and not working with you and am looking forward to more days in good company.

To my family, friends, and Filippo, your unwavering support and encouragement have been essential in helping me complete this journey. Thank you for believing in me, for always being there when I needed you, and for making every moment brighter and more enjoyable.

# CONTENTS

# ACRONYMS

GA genetic algorithm

SA synthetic accessibility

TS transition state

RDS rate-determining step

LESR linear energy scaling relationship

QM quantum mechanical

SQM semiempirical quantum mechanical

DFT density functional theory

XTB extended tight binding

MBH Morita-Baylis-Hillman

TM transition metal

ML machine learning

SMILES simplified molecular-input line-entry system

SELFIES self-referencing embedded strings

HPC high-performance computing

DG directing group

CMD concerted metalation deprotonation

# 1

# GENERAL INTRODUCTION

Catalysts are essential in modern chemistry, playing a crucial role in accelerating chemical reactions. They are vital in numerous industries, such as pharmaceuticals, agriculture, and both fossil and renewable energy sectors. By enabling the efficient production of fuels, chemicals, and medicines, catalysts propel technological and industrial progress.

For catalysts to be effective, they must exhibit high activity and selectivity. Activity refers to a catalyst's ability to increase the rate of a chemical reaction. A highly active catalyst can significantly reduce the energy barrier for a reaction, allowing it to proceed more quickly and efficiently.

Selectivity, on the other hand, pertains to a catalyst's ability to direct a reaction to yield a specific product among multiple possible outcomes. High selectivity ensures that the desired product is obtained with minimal by-products, which is particularly important in the pharmaceutical industry, where high yields in multi-step reactions are crucial. Selective catalysts also reduce the need for extensive purification steps, thereby lowering production costs and minimizing waste.

This thesis is structured into two main parts, each focusing on a different critical aspect of catalyst design using computational techniques.

The objective of Part i is to develop and apply generative models to optimize catalysts with regard to their activity and also showcase the real-world impact of computational de novo generation of catalysts. We optimize the catalytic activity of an organic catalyst in Paper 1 and experimentally verify its superior performance. Paper 4 combines this approach with a reaction network exploration approach, which constitutes the end-to-end de novo discovery of efficient organic catalysts. In Paper 2, we turn our attention to transition metal (TM)-based catalysts and provide a workflow for the generation and optimization of TM-based catalysts, focusing on the Suzuki reaction.

In Part ii, we focus on selectivity in catalysed reactions, specifically in the context of C–H activation reactions, with the aim of developing a computational model for regioselectivity prediction. C–H activations and functionalizations have seen a substantial rise in popularity also due to their importance in late-stage functionalization, essential for modifying complex molecules in pharmaceutical development and other applications. We present a workflow for predicting

the regioselectivity of catalysed and directed C–H activation reactions. By integrating hierarchical quantum mechanical calculations, we offer tools to accurately predict the most probable sites for activation on a substrate. This facilitates the design of novel catalytic systems and the planning of selective reactions.

Part I

# GENETIC ALGORITHMS FOR CATALYST DESIGN

# 2

# INTRODUCTION

Catalysis is a cornerstone of modern chemistry, playing a crucial role in the production of approximately 90% of all chemical compounds.[1–3] It underpins a vast array of industrial applications, demonstrating its significant impact on sectors ranging from energy to healthcare. In petroleum chemistry, catalysts are vital in refining oil and transforming petroleum into useful materials, including polymers, which are indispensable in numerous everyday products. In fertilizer production, catalysts enable the synthesis of ammonia, a key component of nitrogen-based fertilizers that sustain agricultural productivity worldwide. Additionally, the development of catalysts plays a central role in emerging technologies, such as green hydrogen production, which promises to revolutionize energy systems with cleaner alternatives. Carbon capture and utilization technologies also rely on advanced catalysis to efficiently remove carbon dioxide from the atmosphere, converting it into valuable chemicals, thereby addressing environmental challenges. In the pharmaceutical industry, catalysts enable selective reactions that yield specific compounds, accelerating drug development processes to deliver innovative pharmaceutical solutions.

Historically, the discovery and development of catalytic systems have often relied on trial and error, where chemists experimented with different chemicals to identify effective catalysts.[4] Systematic experimental screening of catalyst libraries has also played a significant role in identifying promising candidates.[5] Rational design, guided by empirical observations and chemical intuition, has sometimes provided a more focused approach to developing new catalysts.[6] However, the advent of computational tools has profoundly transformed the field, enabling scientists to probe the mechanisms of catalytic reactions at the atomic level. This computational insight allows for the rational optimization of catalyst structures to enhance performance.[7] In addition, virtual screening of extensive libraries containing potential catalysts through computational models has accelerated the discovery process. More recently, the field has seen rapid advancements in the de novo design of entirely new catalysts using computational chemistry, machine learning (ML) and generative models.[8–15] This approach enables new possibilities in innovative catalyst design.

Despite these advances, many homogeneous catalytic systems currently in use still have major limitations. They often require high temperatures, specific (and sometimes toxic) solvents, and expensive or

harmful transition metals (TMs). These systems also frequently exhibit low turnover rates, require inert atmospheres to operate, and lack stability during storage.

In stark contrast, nature has evolved enzymes, which are highly selective and productive catalysts. These biological systems function efficiently under mild conditions of ambient temperature and pressure, using safe and environmentally friendly solvent. Enzymes have been perfected over countless iterations through the combinatorial process of mutations and natural selection, achieving catalytic efficiency and specificity unmatched by many synthetic catalysts. The design and engineering of synthetic catalysts that emulate these natural systems' efficiency and selectivity offer the potential to revolutionize catalysis across multiple industries.

# 3 CATALYST OPTIMIZATION

## 3.1 INTRODUCTION

In the work presented in this chapter, we take inspiration from nature and develop an automated workflow that emulates nature's process of evolution. The evolutionary process is guided using quantum mechanical (QM) methods to calculate catalytic activity and steer the optimization process towards molecules with high activity.

Optimizing catalysts is a multifaceted task since various properties can or must be optimized simultaneously to meet specific needs. One key property is activity, which refers to the catalyst's ability to increase the rate of a chemical reaction. Improving activity involves making the catalyst more efficient so that smaller amounts are needed to achieve the desired reaction rate, leading to cost savings and more sustainable processes.

Another critical property is selectivity, the catalyst's ability to favour the production of a specific product over others. This is particularly crucial in reactions where multiple pathways and products are possible. Enhancing selectivity can minimize by-product formation, reduce the need for costly and wasteful purification steps, and increase the overall yield of the desired product.

Generality is also important, as a general catalyst can facilitate a wide range of reactions, making it highly versatile and valuable across different chemical syntheses. Enhancing a catalyst's generality can broaden its applications, reducing the need to develop and use multiple specific catalysts. This often involves designing catalysts that are robust under various chemical conditions and substrates.

Stability refers to the catalyst's ability to remain effective over time, either while storing it, or while using it in potentially harsh chemical environments. Increasing stability is essential for processes that require long reaction times or continuous operations.

Furthermore, the cost is a critical factor, especially for large-scale industrial processes. Reducing the cost involves not only minimizing the amount of expensive materials, such as precious metals, used in the catalyst but also designing catalysts that can be easily regenerated and have long lifespans. Economic considerations also include the ease of catalyst preparation and the feasibility of recovering and recycling the catalyst.

Optimizing these properties often involves a trade-off. For example, enhancing stability might come at the expense of activity. Thus,

the challenge is to achieve an optimal balance that suits the specific application. In our works presented in this thesis, we have focused on optimizing the catalyst's activity while aiming to maintain synthesizability and, therefore, economically attractive catalysts. Other works in the field focused on optimizing enantioselectivity or generality using various computational approaches.[12, 16]

Recent advancements in ML have introduced a variety of generative models for molecular design, ranging from 1D models using simplified molecular-input line-entry system (SMILES) or self-referencing embedded strings (SELFIES) representations to 2D graph-based models and 3D models that act on point clouds. However, many of these ML-based works lack experimental or computational validation of the generated structures, which makes assessing their impact on real-world applications difficult.[15] This might be due to a neglect of relevant chemical constraints such as chemical stability or synthesizability.[17] Another potential limitation of ML-based models for de novo molecular generation is their tendency to produce molecules that resemble those in their training data, exploiting existing structural motifs but failing to explore new regions of chemical space. This issue becomes particularly pronounced when the available training data is limited.

An alternative approach to exploring chemical space is through the use of genetic algorithms (GAs), which have been widely used for catalyst design.[8–12, 18]

In these works, novel compounds are designed by recombination of fragments. The fragments are often extracted from experimental databases of known compounds and can be recombined by certain user-defined rules. These enumerated search spaces encompass between several thousands to billions of compounds.[8, 11] The stability and, to some degree, the synthesizability of the resulting molecules can be controlled by carefully selecting the fragments and (re-)combination rules. GAs have been shown to efficiently search vast spaces containing several billion compounds and rediscover the best-performing solutions by evaluating as little as 5% of the total space.[19] However, even in spaces containing over 100 billion compounds, it's possible that molecules with the desired target properties do not exist.[19] Additionally, selecting specific fragments from experimental databases inherently biases the exploratory process toward already considered regions of chemical space.

To overcome these limitations, an unrestricted and free search through chemical space using graph-based GAs can uncover previously unconsidered novel chemical motifs. A detailed description of graph-based GAs is presented in Section 3.2. This approach can provide potentially viable solutions in scenarios where enumerated spaces fail to yield results, opening new avenues for molecular

discovery.

The following chapter will showcase our work using unconstrained GAs for catalyst design. In Paper 1, we develop a workflow for the optimization of an organic catalyst for the Morita-Baylis-Hillman (MBH) reaction. We establish that novel, synthesizable and efficient catalysts can be discovered using an unrestricted graph-based GA. Then, in Paper 2, we turn our attention to the class of TM-based catalysts and extend the workflow to also handle such complexes.

In the following sections, the fundamentals of GAs for molecular optimization are outlined, as well as two models that were used to calculate the catalytic activity.

## 3.2 GENETIC ALGORITHMS

GAs are a class of optimization algorithms inspired by the principles of natural selection and genetics. They offer a robust search mechanism for finding optimal or near-optimal solutions to complex problems. GAs mimic the process of natural evolution, harnessing biological principles like inheritance, crossover, mutation and selection to evolve a population of potential solutions toward better solutions over successive generations. They belong to the family of evolutionary algorithms and are widely used in optimization problems where the search space is vast and not differentiable.

The core components of GAs mirror biological concepts, and their roles in the algorithm are as follows:

INDIVIDUAL / CHROMOSOME Represents a potential solution to the optimization problem.

POPULATION A set of individuals.

SCORING FUNCTION Measures the quality of individuals, directing the evolutionary process towards optimal solutions.

SELECTION A method to choose individuals for reproduction, based on their score.

CROSSOVER A method of merging parts of two parent chromosomes to produce offspring, thereby introducing genetic diversity.

MUTATION A random modification of genes within a chromosome to introduce new genetic structures, helping to avoid premature convergence.

Most commonly in the context of molecular optimization, a chromosome represents a list of molecular fragments combined with each other or with a scaffold, selected from a predefined set of 10 to 100 fragments.[8, 11, 19] Despite the limited number of fragments, the

combinatorial potential allows for the definition of chemical spaces containing millions to billions of possibilities. GAs have proven highly effective at navigating these extensive spaces and can identify near-optimal solutions while evaluating less than 5% of the total search space.[19]

Notably, a graph-based GA approach stands out as one of a few GAs that allows for the exploration of chemical space beyond enumerated libraries. In the work presented in Paper 1 and Paper 2 we build upon the graph-based GA approach developed by Brown et al. [20] and Jensen [21]. Here, the chromosomes are represented as molecular graphs, and the reproduction operators directly manipulate these graphs, as shown in Figure 3.1. The crossover operation in-

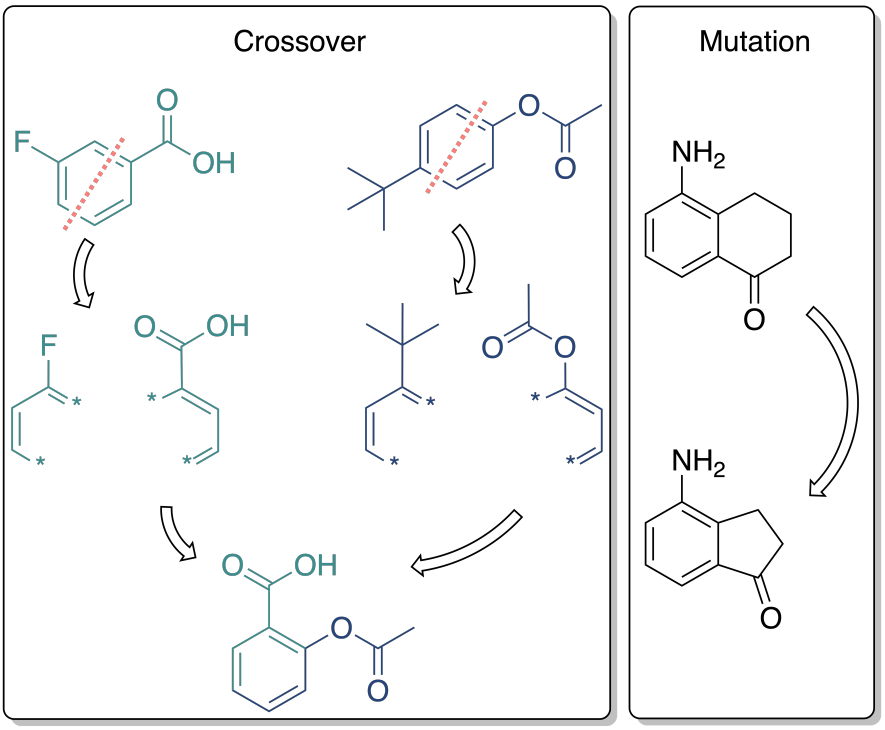



Figure 3.1: Reproduction rules for graph-based GAs on molecular graphs

volves cutting two molecular graphs at random points - either at ring or non-ring bonds - and recombining a fragment from each to form a new molecular graph. The mutation operation modifies a molecular graph by adding, removing, or altering a node (atom) or edge (bond). This allows for the complete exploration of chemical space since all molecules can be formed by sequentially applying these operators. In practice, bridged compounds or extensively fused systems are difficult to obtain due to the implemented recombination rules. Yet, this represents an unconstrained approach since it goes far beyond the screening of enumerated libraries.

A significant challenge with this approach is the proposed molecules' validity, stability and synthesizability. We address this

issue as each molecule's validity is confirmed using RDKit's sanitization protocol, while stability and synthesizability are assessed using empirical measures.[22] Herefore, we utilize the synthetic accessibility (SA)-score, developed by Ertl and Schuffenhauer [23], which quantifies the similarity of a molecule to known stable and synthesizable molecules based on the frequency of radial fingerprints in a precompiled database and incorporates a structural penalty for complex molecules like bridged or spiro compounds.

As we show in Paper 1, this approach is sufficient to steer the GA towards stable and synthesizable molecules.

## 3.3 CALCULATING CATALYTIC ACTIVITY

Catalytic activity is influenced by many factors, such as the reaction conditions, the substrate and the structure of the catalyst. Quantifying a catalyst's performance using a computational approach is not straightforward as it relies on several assumptions. These assumptions must be validated by comparison with experimental data, such as experimental rate constants. Often, computational models fail to calculate accurate experimental values, but they can still be used to qualitatively assess the performance of one catalyst against another catalyst since trends in energies and activities are often accurately captured.

In the work presented in this chapter, we have used two such models, which will be detailed in the following.

### 3.3.1 *Rate-determining step*

Computational chemistry methods are pivotal for calculating the reaction profiles of chemical reactions. In it, the energies of all stationary points along the reaction pathway are contained. Under kinetic control, an effective catalyst should minimize activation barriers throughout the reaction coordinate. A single activation barrier substantially higher than others typically identifies a reaction's rate-determining step (RDS). Given that the reaction rate is exponentially dependent on the energy difference between the reactants and the transition state (TS), even minor changes in this energy difference yield a significant change in reaction rate. Therefore, the calculated activation barrier can be used as a highly sensitive measure of a catalyst's activity in reaction.

This method proves effective in Paper 1 where we calculate the activation barrier of the RDS in the MBH reaction. We show that it is sufficient only to consider the difference in electronic energy calculated with an implicit solvent model to rank the catalysts accurately based on their activity.

Notably, this reaction's RDS step does not involve bond formation or breaking with the catalyst but instead a remote proton transfer between the substrate and a solvent molecule. This is important, since the activation barrier for the initial attack of the catalyst on to the substrate and the release of the product tend to naturally compete with each other in a catalytic cycle. This means that lowering the barrier for the reaction step where the bond between the catalyst and the product is broken tends to increase the barrier for the reaction step in which a bond is formed between the catalyst and the reactant.

In these cases, more sophisticated models might be necessary, which take into account the energetics of several reaction steps with potentially competing barriers.

### 3.3.2 *Vulcano Plots*

Volcano plots are such tools that relate the catalytic activity to the energetics of the whole catalytic cycle. Originally designed for heterogeneous catalysis, they were later adapted for homogeneous systems.[24–26] They are so named because their shape resembles a volcano, where the catalytic activity (*y-axis*) is shown in relation to the interaction strength between the catalyst and a specific reaction intermediate (*x-axis*).

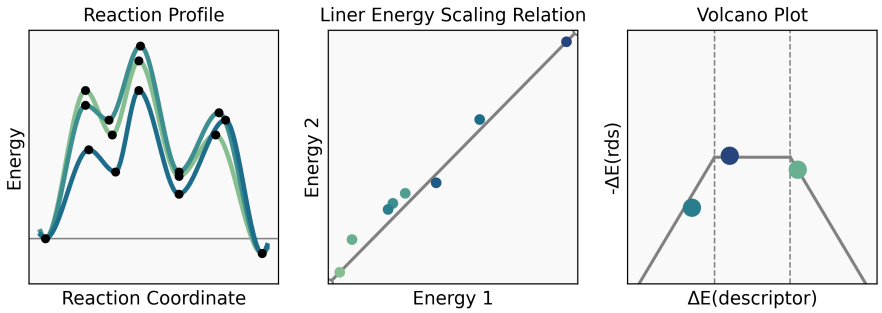



Figure 3.2: Construction of volcano plots: Starting from a series of reaction profiles for different catalysts (*left*), linear energy scaling relationships (LESRs) are established between the different relative intermediate energies (*middle*). Based on these, a volcano plot is constructed where the best-performing catalysts appear at the peak of the volcano (*right*).

Here, we focus on thermodynamic volcano plots, which only consider the relative energies of reaction intermediates (local minima on the potential energy surface).

To construct a volcano plot, one first must calculate all intermediates along the catalytic cycle for a series of different catalysts, as shown in Figure 3.2. Secondly, linear energy scaling relationships (LESRs) are established between the different reaction intermediates. One intermediate whose energy has the best correlation with the other energies is chosen as the descriptor intermediate. Lastly, the

LESRs are used to construct the volcano plot, where the relative energy of the descriptor intermediate is on the x-axis, and the negative of the reaction energy of the RDS is shown on the y-axis. In this model, the RDS is the reaction step with the highest reaction energy. The lower the reaction energy for any particular reaction step, the less favourable any reverse reaction will be. Therefore, a catalytic cycle in which all reaction steps are equally exergonic will be the most efficient. Those catalysts will be located at the top of the volcano plot.

Since a volcano plot is based on LESRs, the accuracy of these is crucial to the predictive performance of the volcano plot. Wodrich et al. [26] have shown for exemplary systems that scaling relations hold for catalysts upon a change in oxidation or spin state and ligation. Yet, it is possible that other reaction pathways become preferred upon drastic changes to the catalyst, which would invalidate the LESRs.

In Paper 2, we utilize the volcano plot established by Meyer et al. [27], which only considers the electronic energies of the intermediates in the catalytic cycle. This allows us to evaluate a catalysis activity by only calculating a single energy difference.

## 3.4 ORGANIC CATALYST

In this section, we present our work on optimizing the structure of an organic homogeneous catalyst for the MBH reaction. We use the graph-based GA, as introduced in Section 3.2, and calculate the catalytic activity using the activation barrier of the RDS as described in Section 3.3.1.

### 3.4.1 *Summary of key findings*

In this study, we employ a genetic algorithm to discover novel homogeneous catalysts for the MBH reaction. We calculate the catalytic activity using semiempirical and quantum mechanical (SQM/QM) methods and leverage these calculations for the de-novo generation and optimization of novel, efficient catalysts. Following computational evaluation, we experimentally validate one such catalyst, demonstrating that it outperforms existing benchmarks.

First, we establish a strong correlation between the experimentally observed reaction rates and Gibbs activation energies calculated via density functional theory (DFT). To accelerate the evaluation process, we leverage the semiempirical extended tight binding (xTB) method to compute approximate electronic activation energies, based on approximate TS structures. This enables us to assess catalyst performance rapidly and accurately, as shown by a strong linear correlation between the exact DFT and approximate semiempirical quantum mechanical (SQM) activation energies. We execute a series of GA runs, using this SQM-based approach as a scoring function. This allows us to explore a vast chemical space, ultimately generating 448 unique catalyst candidates.

These candidates are subsequently filtered based on their exact activation energies computed at the GFN2-xTB level.[28] To further refine the list, we employ retrosynthetic analysis using the Manifold program to identify catalysts that can be synthesized easily, narrowing the selection to 132 candidates that can be synthesized in a single step.[29]

From this refined pool, we select two catalysts for experimental validation, prioritizing those with readily available building blocks from commercial vendors. We then calculate the full reaction profiles for these catalysts using DFT, confirming their superior catalytic efficiency for the MBH reaction compared to a commonly used catalyst.

Our collaborators synthesized the selected molecules and were able to isolate one of them (M19) with good yield. The other one (M10) could not be purified. Subsequently, they measured the conversion rates for the MBH reaction using a frequently used catalyst as a reference and the new catalyst (M19). From these measurements, we calculate the rate constants and expected conversions. The results re-

veal that the newly discovered catalyst is 7.8 times more proficient in catalyzing this reaction than the previous benchmark catalyst.

GDCh Research Articles Angewandte International Edition Chemie www.angewandte.org

Check for updates

How to cite: *Angew. Chem. Int. Ed.* **2023,** *62,* e202218565
International Edition: doi.org/10.1002/anie.202218565
German Edition: doi.org/10.1002/ange.202218565

Homogeneous Catalysis

# Computational Evolution Of New Catalysts For The Morita–Baylis–Hillman Reaction**

*Julius Seumer, Jonathan Kirschner Solberg Hansen, Mogens Brøndsted Nielsen, and Jan H. Jensen**

**Abstract:** We present a de novo discovery of an efficient catalyst of the Morita–Baylis–Hillman (MBH) reaction by searching chemical space for molecules that lower the estimated barrier of the rate-determining step using a genetic algorithm (GA) starting from randomly selected tertiary amines. We identify 435 candidates, virtually all of which contain an azetidine N as the catalytically active site, which is discovered by the GA. Two molecules are selected for further study based on their predicted synthetic accessibility and have predicted rate-determining barriers that are lower than that of a known catalyst. Azetidines have not been used as catalysts for the MBH reaction. One suggested azetidine is successfully synthesized and showed an eightfold increase in activity over a commonly used catalyst. We believe this is the first experimentally verified de novo discovery of an efficient catalyst using a generative model.

## Introduction

Homogeneous catalysts have transformed synthetic organic chemistry and many of the most popular chemical reactions require a catalyst, and catalyst discovery is one of the "holy grails" of computational chemistry.[1–5] Quantum chemistry (QM) has become an important aid in elucidating catalytic mechanisms and experimental mechanistic studies frequently include a modelling component. QM calculations can also be used to test ideas for new catalysts that arise from these mechanistic insights.[6] While quantitative predictions of reaction rates and yields are difficult, the QM calculations often serve as a "sanity check" for new ideas based on chemical understanding.

[*] J. Seumer, J. Kirschner Solberg Hansen, M. Brøndsted Nielsen, J. H. Jensen
Department of Chemistry, University of Copenhagen, Denmark
E-mail: jhjensen@chem.ku.dk

[**]A previous version of this manuscript has been deposited on a preprint server (https://doi.org/10.26434/chemrxiv-2022-ngwvt).



Another approach to computational catalyst discovery has focused on quantitative structureactivity relationships (QSAR) of organometallic catalysts (which represent the majority of homogeneous catalysts). In the original approach, pioneered by Tolman in the 1970s, observed structural features of the catalysts and chemical features of the ligands are correlated with the observed catalytic activity.[7] These QSAR models were then used to guide the discovery of new catalysts. As QM method became more powerful the observed features were augmented or replaced with QM-predicted features.[8]

With the advent of machine learning (ML) the QM and QSAR approaches are starting to merge. While the traditional QSARs had to be constructed based on expert chemical knowledge, ML models can learn these QSARs given enough data. Sufficient experimental reactivity data can be obtained for some reactions using high throughput techniques but in many cases, the catalytic activity is estimated using QM.[9,10] A typical state-of-the-art computational organometallic catalyst discovery study involves libraries of $10^4$–$10^6$ catalyst candidates constructed using predefined metals and ligands.[11,12] The activity is then calculated using QM for a small subset [$O(10^3)$] exploiting a linear free energy relation between energies of key intermediates in the catalytic mechanism and the reaction rate. This data is then used to train an ML model that is then used to predict the reactivity of the entire library and the most interesting candidates are selected for further QM calculations. In some cases the process is repeated, i.e. the new QM data is used to update the ML model and the library is re-screened.[12] Another option is to use search algorithms, such as genetic algorithms (GAs), rather than evaluating the entire library.[13–16] The efficiency of these search algorithms also allows for the use of QM, rather than ML, for reactivity.[13,16] However, all studies so far have focused on screening user-defined libraries of catalysts.[11–17] While experimental verification of catalysts predicted using these computational approaches are rare, Das et al. have recently successfully identified a frustrated Lewis pair catalyst for direct hydrogenation of $CO_2$.[17]

As noted by Poree and Schoenebeck: "Computationally driven evolution of catalysts is an exciting prospect because it would allow us to break free of the limitations of our ideas and preconceptions."[4] However, it is not clear that this prospect can be achieved by screening userdefined libraries. In this paper, we present a de novo discovery of a new tertiary amine catalyst for the methanol-mediated Morita–Baylis–Hillman (MBH) reaction using GA searches of the

entire chemical space of tertiary amines, with synthetic accessibility as the only constraint. The GA searches discover that azetidines are likely to be potent catalysts despite the fact that this moiety is not in the starting population of most of the searches. One of the molecules was successfully synthesized and is indeed more active than DABCO with an eight-fold increase in the rate-constant corresponding to a roughly 1 kcal mol$^{-1}$ lower barrier, in good agreement with the predictions.

The Morita–Baylis–Hillman (MBH) reaction of methyl acrylate (MA) with *p*-nitrobenzaldehyde (*p*NBA) catalysed by tertiary amines in methanol is chosen as the model reaction. We focus on the mechanism as outlined in Figure 1 which is supported by experimental and computational studies.[18,19] The rate-limiting step of the reaction is the proton transfer between intermediate **2** and **3** which is aided by a methanol molecule, the corresponding transition state (**TS3**) is shown in Figure 1 highlighted in grey.

## Results and Discussion

We start by demonstrating that the level of theory used in this study can predict relative standard activation free energies ($\Delta G^{o,+}$) that are in good agreement with experiments. The relative standard activation free energy is calculated as the difference in Gibbs free energy of **TS3** and the reactants at the B3LYP-D3/6–31+G(d,p)(SMD) level.[20–23] Here, for each catalyst, 50 conformers of **TS3** are embedded using a template as described in the Supporting Information (section S1.2) and an RMSD pruning cutoff of 0.1 Å is used before each conformer undergoes a constrained GFN2-xTB optimisation where the atoms corresponding to the template are fixed.[24] Furthermore, 50 conformers of the catalyst on its own are created and the same RMSD threshold is used for pruning and the retained conformers undergo GFN2-xTB optimisation. The lowest energy conformer of **TS3** and the catalyst on its own are further optimised using Gaussian 16 at the B3LYP-D3/6–31+G(d,p)(SMD) level of theory to the respective transition state or minimum.[25] The transition states are characterised by one imaginary frequency along the reaction coordinate corresponding to the abstraction of a proton from the tertiary carbon by a methoxy fragment.

The resulting $\Delta G^{o,+}$ values show strong linear correlations (Pearson's correlation coefficient = −0.99) with the logarithm of the experimentally measured reaction rate constants for six MBH reactions catalysed by quinuclidine-based catalysts (Figure S3).[26] Therefore, this level of theory is used as a computationally accessible measurement of the catalytic potential of other tertiary amine-based catalysts for the MBH reaction.

Next, we benchmark the performance of three barrier scoring functions on a set of 100 molecules sampled from a GA run over a range of ts_scoring energies of −40 to 20 kcal mol$^{-1}$. The relative standard activation free energy is calculated for 52 of the molecules, for the other 48 molecules either the structure of **TS3**, the catalyst or both could not be successfully located (42, 9 and 3 times respectively).

Figure 2 shows a strong linear correlation between the ts_scoring energy and $\Delta G^{o,+}$ with a Pearson correlation coefficient of 0.96. In the GA, the scoring function is used to rank the molecules and their fitness is derived from their rank within the population. The Spearman rank correlation coefficient of 0.98 indicates that the molecules are success-

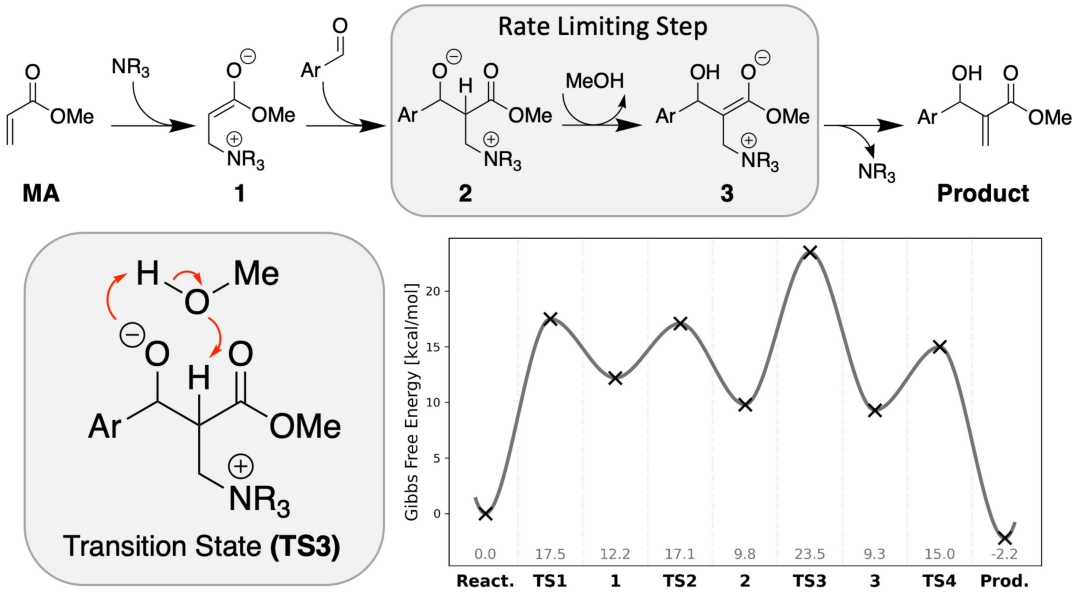



***Figure 1.*** Mechanism of the MBH reaction with the rate-limiting step and the corresponding transition state highlighted in grey. The Gibbs free energy of the reaction between MA and *p*NBA (as specific example of ArCHO) catalysed by DABCO in methanol as calculated with CCSD(T)//CBS1//B3LYP-D3(SMD) is shown in the bottom right.[18]

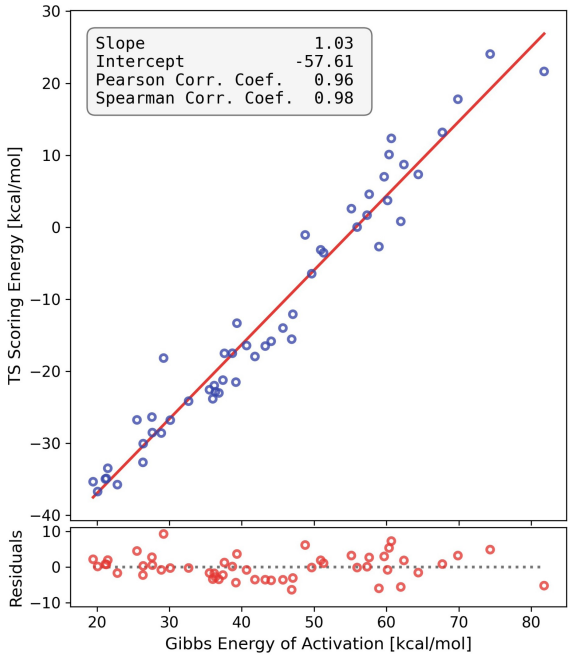



***Figure 2.*** Correlation between the relative standard activation free energy (B3LYP-D3/6–31 + G(d,p)(SMD)) and the activation energy calculated using the ts_scoring function for 52 molecules

fully ranked according to their $\Delta G^{o,\ddagger}$ by the ts_scoring function.

The performance and timings of the inter_scoring and path_scoring functions are discussed in the Supporting Information section S1.2. Due to the high rank correlation, low computational cost and robustness, the ts_scoring function is used in the GA going forward.

Five GA runs are performed with different starting populations of 100 molecules taken randomly from a subset of tertiary amines of the ZINC database.[27] Figure 3 shows the evolution of the ts-score (which includes a synthesizability penalty and is further described in Supporting Information section S1.1) of the best-performing molecule in the population over 100 generations for five separate GA runs.

In each run, the score drops drastically by up to 10 points in the first four generations. This jump can be largely attributed to the increase of the SA score component of the overall score to 1 [see Supporting Information Eq. (S2)]. In the subsequent generations, the overall score of the best-performing individuals decreases in smaller steps which are driven by a decrease in the calculated activation energy.

The best-performing molecule from the three runs with the lowest final score is shown in Figure 3 on the right. All three molecules have an azetidine moiety where they bind to the reactant. The azetidine moiety is present in one of the runs starting populations but does not survive after the first evolutionary step. Instead, the azetidine ring is rediscovered by the GA (via crossover and mutation operations) as a catalytic motif and virtually outcompetes all other motives as 498 out of the final 500 molecules from five runs contain the azetidine moiety. The first time a molecule containing an azetidine ring performs best in the population is marked with a black cross in Figure 3. Although the azetidine ring is discovered as a preferable binding motive within 25 generations, substantial improvement of more than 5 points are archived by the GA over the following 75 generations.

In order to create different structures and generate new ideas for potential catalysts, we perform an additional GA search where we remove any azetidine-containing molecules from the populations. This search yields molecules with ts-scores as low as −139 after 100 generations and the eighth bestperforming molecules in the final population are shown in the Supporting Information Figure S5. The molecules contain pyrrolidine or piperidine moieties or a tertiary amine with two methyl groups as preferred binding motifs.

The five GA searches result in a total of 448 unique molecules and the goal is now to select a handful of

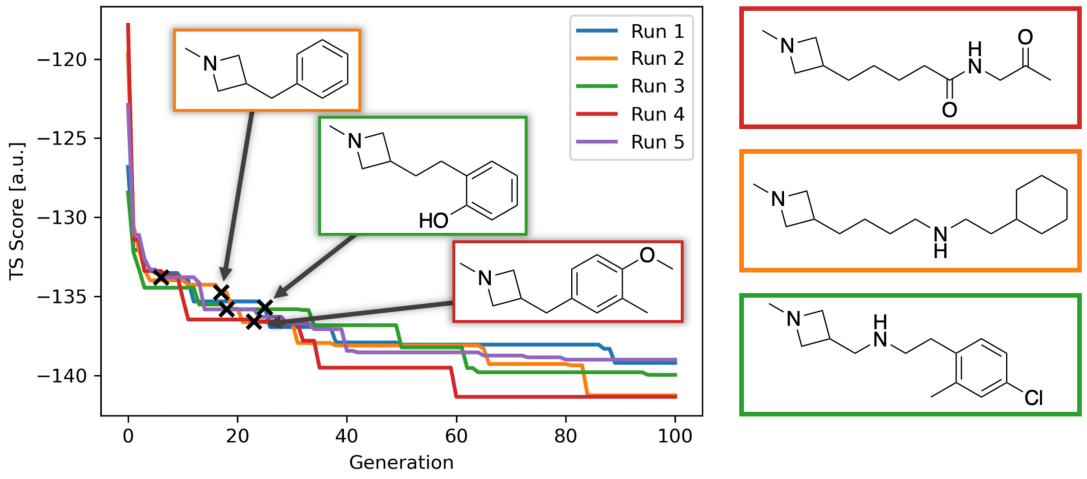



***Figure 3.*** Evolution of the ts-scoring score over 100 generations of the best performing molecule in five separate runs. The molecules on the right correspond to the best performing molecule in the final population of the three runs with the lowest final score.

molecules for which the entire reaction mechanism is computed at the DFT level and, if promising, for experimental validation. Figure 4A shows the distribution of ts_scoring energies for the 448 unique molecules. Next, for 435 out of 448 molecules, the structure of the true transition state **TS3** at the GFN2-xTB level is successfully located and the activation energy is calculated as the difference in electronic energy of **TS3** and the reactants. A histogram of the calculated activation energies is shown in Figure 4B. The activation energy for all 435 molecules is significantly lower than that of DABCO ($-37.4\,kcal\,mol^{-1}$), which is a known catalyst.

Next, the retrosynthesis tool Manifold is used to determine the minimum number of synthetic steps required to synthesise each molecule from commercially available building blocks.[28] The minimum number of steps is evaluated for synthetic routes involving molecular starting blocks from selected catalogues (generic, molport, emolecules, mcule) with a maximum lead time of 2 weeks. The histogram of the minimum number of synthetic steps is shown in Figure 4C. A number of 132 molecules can be synthesised in one step from commercially available building blocks and a subset is shown in the Supporting Information Figure S4.

Based on the availability of reactants from certain vendors two molecules, **M10** and **M19** from Figure S4, are selected for further potential synthesis. However, before commencing synthesis we computed the entire reaction path at the B3LYP-D3/6–31 + G(d,p)(SMD) level of theory for both molecules, as described next, and compare them to that for DABCO. To locate the relevant structure, a template structure of the intermediates or transition states from Liu et al. is used, the present catalyst is exchanged for the one that is to be evaluated and 50 conformers of the new catalyst are embedded.[18] The minimum energy conformer is found from GFN2-xTB constrained optimisations and the structure undergoes further optimisation at the B3LYP-D3/6–31 + G(d,p)(SMD) level. The transition states **TS1**, **TS2**, **TS3** and **TS4** as well as intermediates **1**, **2**, **3** are located using Gaussian16. From the transition states, IRC calculations are performed to confirm that the transition state connects the relevant reactants, intermediates or the product. The Gibbs free energy of each structure is calculated as described in the Supporting Information section S1.3 and the relative Gibbs free energy profile for the MBH reaction catalysed by DABCO (grey), **M10** (blue) and **M19** (orange) is shown in Figure 5. Both **M10** and **M19** show lower activation energies from the reactants to **TS3** of 19.5 and $18.8\,kcal\,mol^{-1}$ compared to $21.2\,kcal\,mol^{-1}$ for DABCO. This is in agreement with the previous calculations of the activation energy associated with **TS3** at the GFN2-xTB level. Furthermore, the reaction catalysed by **M19** shows a lower **TS3** barrier than the one with **M10** which is expected from the previous calculations. All transition states and intermediate structures from the reaction catalysed with **M10** have lower relative Gibbs free energies than the corresponding structures from the reaction catalysed with DABCO. From this thermodynamical point of view, one can expect the molecule **M10** to perform better as a catalyst than DABCO. The intermediates and transition states besides **TS3** for the reaction with **M19** are approximately 2–3 $kcal\,mol^{-1}$ higher in relative Gibbs free energy than the corresponding structures from the reaction catalysed by DABCO. Although the reaction energy from the reactants to **TS1** is with $15.2\,kcal\,mol^{-1}$ approximately $3\,kcal\,mol^{-1}$ higher than the one of the reaction with **M10** one can still expect **M19** to catalyse the MBH reaction effectively since the overall activation energy from the reactants to the highest energy transition state along the reaction path is lower than the one of the reaction catalysed by DABCO.

Visual inspection of the transition state structure of **TS3** with the catalyst **M19** reveals that the catalyst is forming a hydrogen-bond-like interaction between the secondary amine moiety of the catalyst and the hydroxyl moiety of the reactant as shown in Figure 6. When the secondary amine moiety of the catalyst is replaced with a $CH_2$ group the activation energy increases by $2.83\,kcal\,mol^{-1}$ since no stabilising interaction can be formed. Furthermore, the T-shaped interaction between the two aromatic rings stabilizes the transition state further.

Both **M10** and **M19** were synthesized, however, **M10** proved hard to purify. Catalyst **M19** was subjected to experimental validation. This compound was successfully synthesised and purified as described in the Supporting

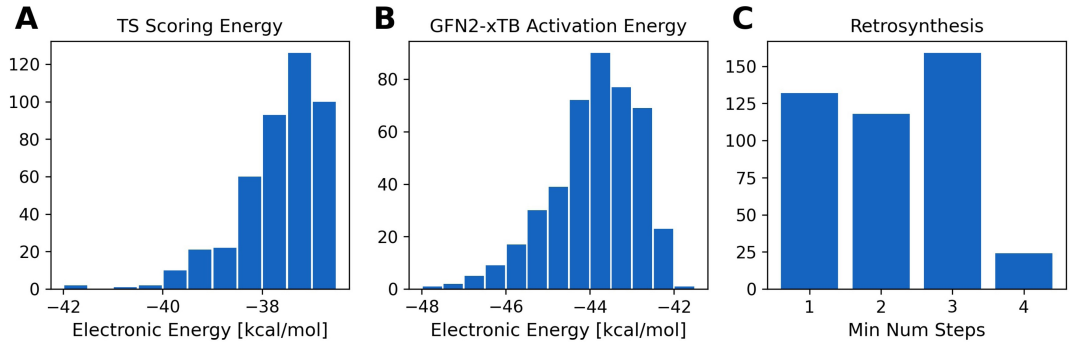



***Figure 4.*** A) Histogram of the electronic activation energy as calculated by the ts_scoring function for 448 molecules from 5 GA runs, B) electronic energy difference between TS3 and the reactants at the GFN2-xTB level for 435 molecules, C) Minimum number of synthetic steps from available building blocks as predicted by Manifold.

GDCh Research Articles Angewandte International Edition Chemie

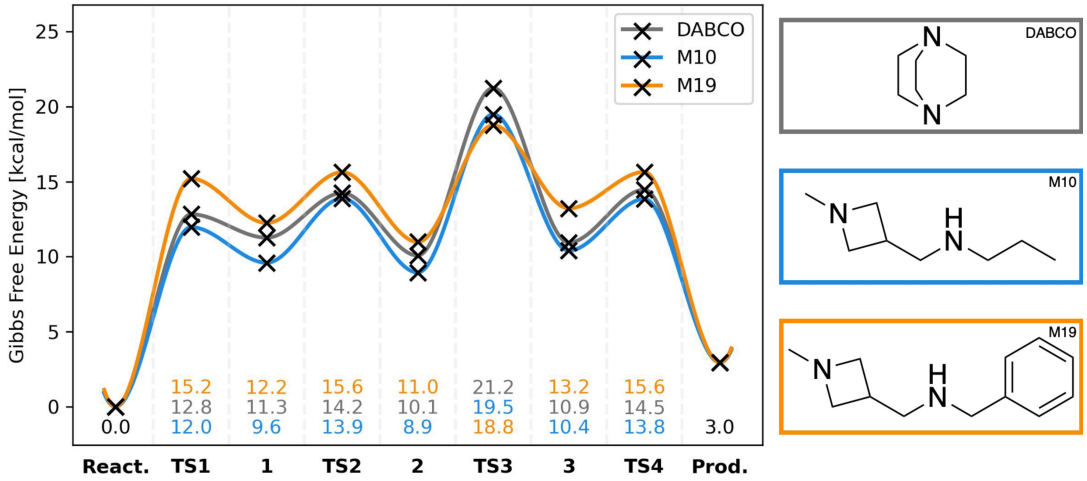



***Figure 5.*** Gibbs free energy profile of the MBH reaction with three different catalysts: DABCO in grey, **M10** in blue and **M19** in orange.

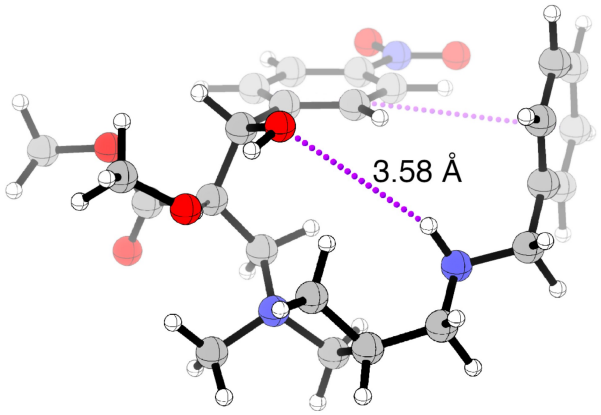



***Figure 6.*** Structure of **TS3** with non-covalent interaction (dashed purple line) between the catalyst **M19** and the hydroxyl moiety of the reactant and the two aromatic ring systems.

Information section S5. The MBH reaction between MA and *p*NBA in $d_4$-methanol is followed by $^1$H NMR spectroscopic measurements and the performance of **M19** as catalyst is evaluated by comparing conversion and reaction rate to reference experiments using DABCO as catalyst.

Figure 7 shows stacked $^1$H NMR spectra taken over the course of the reaction with the atoms in the reactant (MA) and product coloured like their corresponding signal. The concentration of MA is calculated based on the decrease of the signal at 5.87 ppm (purple). The concentration of the product at time $t$ is calculated as $[P]_t = [MA]_0 - [MA]_t$ with $[MA]_0$ as the starting concentration of MA and $[MA]_t$ as the concentration of MA at time $t$. The concentration of *p*NBA was calculated as $[pNBA]_t = [pNBA]_0 - [P]_t$.

The percent conversion to the product over reaction time is shown in Figure 8 with the reaction using DABCO as catalyst in grey and **M19** in orange. As described in the Supporting Information section S6.2, the rate constants $k$ are obtained by fitting to data points within the first 1.5 h of the reaction due to the incorporating of deuterium into the α-position of MA over the course of the reaction as described by Plata and Singleton which can be seen in the Supporting Information Figure S9.[19] The α-deuteriated MA is expected to yield the product significantly slower due to the kinetic isotope effect and the involvement of this atom in the rate-determining step of the reaction in **TS3**. The predicted conversion can be calculated using Equation (S5) and is shown as dashed lines in Figure 8. The corresponding third-order rate constant $k$ ($v = k$[MA] [*p*NBA] [Catalyst]) is $k = 0.00010\ M^{-2}s^{-1}$ for DABCO and $k = 0.00078\ M^{-2}s^{-1}$ for **M19**. The rate constant for the reaction with DABCO as catalyst is in reasonable agreement with the rate constant obtained using the same kinetic analysis of the $^1$H NMR spectroscopic measurements kindly provided by Prof. Singleton.

A comparison of the rate constants for the reaction using **M19** and DABCO shows that the rate constant for the reaction with **M19** is 7.8 times larger than the one for the reaction with DABCO. From the Eyring equation, we calculate that the Gibbs energy of activation at 22 C is 1.12 kcal mol$^{-1}$ lower for the reaction with **M19** than for the reaction with DABCO. This agrees well with the previously calculated value of 2.40 kcal mol$^{-1}$.

Furthermore, 54 % conversion could be reached after 7 h reaction time using **M19** as catalyst compared to 16 % using DABCO in deuterated solvent. As shown by Plata and Singleton, significantly higher conversion is reached in undeuterated solvent which would follow the predicted conversion in Figure 8 yielding 48 % and 87 % conversion after 24 h for the reaction with DABCO and **M19**, respectively.[19]

To further prove that **M19** is an excellent catalyst and the importance of the azetidine ring, we tested the diamine shown in Figure 9. This diamine was completely unable to function as a catalyst when using reaction conditions similar to those used for catalyst **M19**. The only observed reaction was a small amount of transesterification of methyl acrylate. This experiment further proved the importance of the

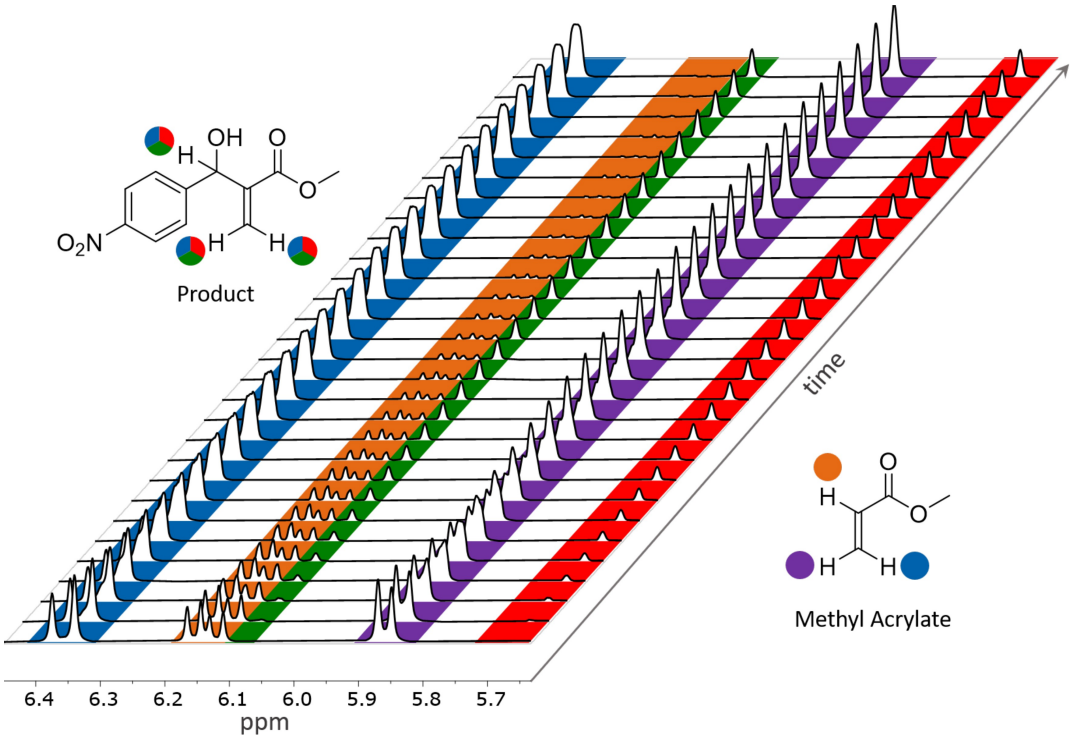



***Figure 7.*** Section of stacked $^1$H-NMR spectra (500 MHz, $CD_3OD$) from a mixture of 4-nitrobenzaldehyde, methyl acrylate, and DABCO in a 1:0.94:0.3 ratio with initial 4-nitrobenzaldehyde concentration of 0.578 M, measured approximately every 20 minutes for 11 hours and 40 minutes. The first spectrum starts from the bottom. In methyl acrylate, the α-proton (orange) disappears as product is formed and is exchanged for a deuterium in parallel. The blue and purple peaks change from a doublet to a singlet as product is formed, and the blue peak overlaps with 1 of 3 protons in the product. The three labelled protons in the product could not be distinguished and could hence correspond to peaks marked by any of the three colours (red, blue and green) in the spectrum.

structure of the catalyst being key to its functionality for the MBH reaction. The $^1$H NMR spectra recorded for this experiment can be found in the Supporting Information Figures S21 and S22.

## Conclusion

We present a de novo discovery of an efficient catalyst of the alcohol-mediated Morita–Baylis–Hillman (MBH) reaction by searching chemical space using a genetic algorithm (GA) starting from randomly selected tertiary amines. The GA searches for molecules that lower the estimated barrier of the rate-determining step, where the barrier is estimated by the semiempirical GFN2-xTB method using a model geometry of the transition state region. The barrier estimate is augmented by a function that rewards synthetic accessibility. We performed five independent GA searches, each for 100 generations with a population size of 100, which resulted in 448 unique molecules, for which we were able to locate 435 true transition states at the GFN2-xTB level of theory. The predicted activation energies of all 435 molecules were all lower than that of DABCO, which is a popular catalyst of the MBH reaction. Virtually all (498/500) of the molecules contain an azetidine N as the catalytically active site, which is discovered by the GA since it is not found in the initial population in four of the five runs (and in that run, it is discarded early only to be rediscovered as the search progresses). In addition, many of the GA searches also introduce an azetidine substituent with a hydrogen bond donor that helps to stabilize the transition state and thus lower the barrier. This demonstrates the power of free exploration of chemical space compared to more constrained fragment-based approaches.

Next, we predict retrosynthetic paths for the 435 molecules using the Manifold software package and select the 135 molecules which can be made in only one step from commercially available building blocks. From these 135 molecules, we select two (**M10** and **M19**) for further study, based on building block availability and cost. For these two molecules, we compute the entire free energy reaction profile at the DFT level and show that their rate-determining barriers are 1.7 and 2.4 kcal mol$^{-1}$ lower than that of DABCO. The molecule with the lowest barrier (**M19**) has higher barriers for the other steps compared to DABCO, but none of the barriers are competitive with the rate-determining barrier and is predicted to outperform DABCO.

Finally, the performance of **M19** as a catalyst for the MBH reaction was tested experimentally and it is shown

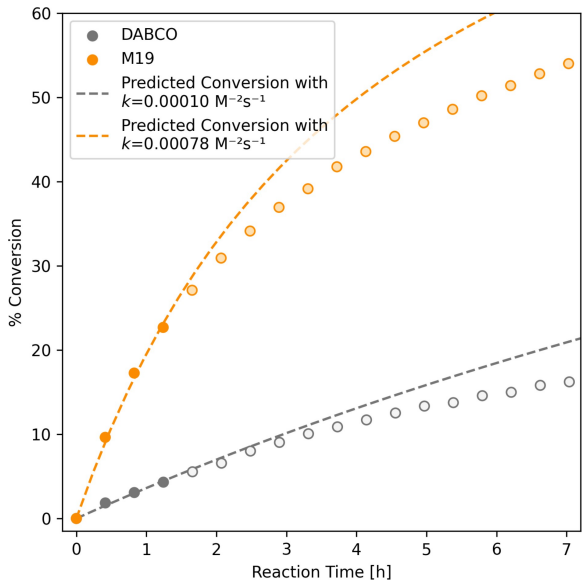



***Figure 8.*** Comparison between the kinetics of the MBH reaction with different catalysts: DABCO in grey and **M19** in orange. Conversion to the product is followed by $^1$H-NMR spectroscopic measurements in $d_4$-methanol. The dashed lines are theoretical curves based on the rate law $v = k$[MA] [*p*NBA] [Catalyst]. The rate constants $k$ were obtained by linear fits to the concentrations of the reactants within the first 1.5 h of the reaction.

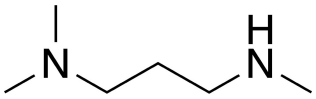



***Figure 9.*** Randomly choosen diamine but with a $C_3$ linker separating the two nitrogen atoms as in catalyst **M19.**

that it outperforms DABCO with significantly faster reaction times and a reaction rate constant that is 7.8 times larger. The observed difference in rate constants is consistent with the calculated difference in activation energy between DABCO and **M19** which highlights the accuracy of the here chosen computational approach.

We believe this is the first experimentally verified de novo discovery of an efficient catalyst using a generative model. Our study shows that generative models indeed can discover new chemistry with a minimum of empirical input, as long as the molecular property of interest (the fitness) can be computed accurately. It is important for experimental validation that synthetic accessibility is accounted for, both in the search process and in the final selection. Our fitness function presupposes that the rate-determining step is known. The de novo discovery of catalysts for new reactions would thus involve the de novo prediction of catalytic reaction mechanisms, which is an area we, and others, are currently working on.[29–31]

## Supporting Information

The data and code are available at https://sid.erda.dk/sharelink/hGBkdGdCy7 and https://github.com/jensengroup/mbh_catalyst_ga respectively.

## Acknowledgements

This work was supported by Novo Nordisk Fonden via grant number NNF20OC0064104. We thank Christian G. Tortzen (University of Copenhagen) and Daniel Singleton (Texas A&M University) for helpful discussions.

## Conflict of Interest

The authors declare no conflict of interest.

## Data Availability Statement

The data that support the findings of this study are available in the Supporting Information of this article.

**Keywords:** Chemical Space · De Novo Discovery · Genetic Algorithm · Organocatalysis

[1] D. G. Brown, J. Boström, *J. Med. Chem.* **2016**, *59*, 4443–4458.
[2] K. N. Houk, P. H.-Y. Cheong, *Nature* **2008**, *455*, 309–313.
[3] K. N. Houk, F. Liu, *Acc. Chem. Res.* **2017**, *50*, 539–543.
[4] C. Poree, F. Schoenebeck, *Acc. Chem. Res.* **2017**, *50*, 605–608.
[5] M. Foscato, V. R. Jensen, *ACS Catal.* **2020**, *10*, 2354–2377.
[6] S. Ahn, M. Hong, M. Sundararajan, D. H. Ess, M.-H. Baik, *Chem. Rev.* **2019**, *119*, 6509–6560.
[7] C. A. Tolman, *J. Am. Chem. Soc.* **1970**, *92*, 2953–2956.
[8] D. J. Durand, N. Fey, *Chem. Rev.* **2019**, *119*, 6561–6594.
[9] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **2018**, *360*, 186–190.
[10] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science* **2019**, *363*, eaau5631.
[11] B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld, C. Corminboeuf, *Chem. Sci.* **2018**, *9*, 7069–7077.
[12] A. Nandy, C. Duan, C. Goffinet, H. J. Kulik, *JACS Au* **2022**, *2*, 1200–1213.
[13] Y. Chu, W. Heyndrickx, G. Occhipinti, V. R. Jensen, B. K. Alsberg, *J. Am. Chem. Soc.* **2012**, *134*, 8885–8895.
[14] M. Foscato, V. Venkatraman, V. R. Jensen, *J. Chem. Inf. Model.* **2019**, *59*, 4077–4082.
[15] R. Laplaza, S. Gallarati, C. Corminboeuf, *Chem. Methods* **2022**, *2*, e202100107.
[16] K. J. Kron, A. Rodriguez-Katakura, P. Regu, M. N. Reed, R. Elhessen, S. Mallikarjun Sharada, *J. Chem. Phys.* **2022**, *156*, 184109.
[17] S. Das, R. C. Turnell-Ritson, P. J. Dyson, C. Corminboeuf, *Angew. Chem. Int. Ed.* **2022**, *61*, e202208987; *Angew. Chem.* **2022**, *134*, e202208987.
[18] Z. Liu, C. Patel, J. N. Harvey, R. B. Sunoj, *Phys. Chem. Chem. Phys.* **2017**, *19*, 30647–30657.
[19] R. E. Plata, D. A. Singleton, *J. Am. Chem. Soc.* **2015**, *137*, 3811–3826.
[20] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648–5652.

[21] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, *J. Chem. Phys.* **2010**, *132*, 154104.
[22] W. J. Hehre, R. Ditchfield, J. A. Pople, *J. Chem. Phys.* **1972**, *56*, 2257–2261.
[23] A. V. Marenich, C. J. Cramer, D. G. Truhlar, *J. Phys. Chem. B* **2009**, *113*, 6378–6396.
[24] C. Bannwarth, S. Ehlert, S. Grimme, *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.
[25] Gaussian 16, Revision C.01, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr, J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, **2016**.
[26] V. K. Aggarwal, I. Emme, S. Y. Fulford, *J. Org. Chem.* **2003**, *68*, 692–700.
[27] T. Sterling, J. J. Irwin, *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.
[28] PostEra, Medicinal Chemistry Powered by Machine Learning, en, https://postera.ai/ manifold/, Accessed: 2022-3-16.
[29] M. H. Rasmussen, J. H. Jensen, *PeerJ Phys. Chem.* **2022**, *4*, e22.
[30] M. Bensberg, M. Reiher, **2022**.
[31] S. Habershon, *J. Chem. Theory Comput.* **2016**, *12*, 1786–1798.

Manuscript received: December 15, 2022
Accepted manuscript online: February 14, 2023
Version of record online: March 22, 2023

# Supporting Information

## S1. Computational Methodology

### S1.1. Genetic Algorithm

A graph-based genetic algorithm (GA) is used to obtain tertiary amines that induce a low activation barrier (the energy difference between **TS3** and the reactants) in the MBH reaction and hence a low score (further details on the scoring function in the next section).[1] The GA is run for 100 generations with a population size of 100, the starting population is chosen randomly from a subset of tertiary amines from the ZINC database.[2] New individuals are created with a 50:50 chance either by mutation or crossover operations performed on molecules from the previous generation. The molecular sites for crossover and mutations are selected at random and thus offer a free exploration of chemical space. Individuals are selected for mutation/crossover operations with a probability proportional to their rank within the population (rank selection). Here, individuals are ranked in descending order based on their score (low score to high rank) and their probability for selection is calculated following Baker et al. with a selective pressure (SP) of 1.5.[3]

$$p_i = \frac{1}{N}\left(2 - \mathrm{SP} + 2 \cdot (\mathrm{SP} - 1) \cdot \frac{r_i - 1}{N - 1}\right) \tag{S1}$$

Here, $p_i$ is the probability for the selection of individual $i$, $r_i$ is the rank of individual $i$ and $N$ is the number of individuals in the population. Individuals with low scores will have a high probability of being selected for mutation/crossover operations. After the mutation/crossover operation, the new individual has to contain one tertiary amine and has to have more than 5 but less than 14±6 heavy atoms, otherwise, it is discarded. Duplicates are removed after 100 new individuals are generated and the 100 best-performing individuals from the current and the previous generation advance to the next generation.

Generative models such as GAs are known to generate molecules with unstable bonds and highly complex structures.[4,5] Here, the score from the scoring function is modified by a synthetic accessibility measurement to steer the GA towards molecules with high catalytic activity and which are also synthesisable. The SA score as defined by Ertl and Schuffenhauer is used in connection with a modified Gaussian function as proposed by Brown et al. with the parameters $\mu = 2.230044$ and $\sigma = 0.6526308$ as proposed by Gao and Coley.[4–6] This way molecules which are deemed easy to synthesise are assigned a value of up to 1 and molecules that are hard to synthesise a value of 0. The final score used in the GA is then obtained as the product of the electronic activation energy minus 100 and the modified SA score.

$$\mathrm{Score}_i = \left(\Delta E^{\ddagger}_{\mathrm{A}i} - 100\right) \cdot \exp\left(\frac{-\left(\max\left(\mathrm{SA}_i, \mu\right) - \mu\right)^2}{2\sigma^2}\right) \tag{S2}$$

Subtracting 100 from the activation energy ensures that all scores are negative as the calculated electronic activation energy typically ranges between -40 and 20 kcal $\cdot$ mol$^{-1}$. Without subtraction from the activation energy, most molecules in a randomly selected starting population will have a positive activation energy which would then be multiplied with the modified SA score ranging between 0 and 1 to give the final score which then would often be positive. This score could be easily minimised by generating molecules that are not synthesisable, therefore giving a modified SA score of 0 which again gives an overall score of 0. Hence, the GA could get trapped in an area of chemical space with low synthesisability and scores close to 0. With the subtraction, virtually all overall scores will be negative and the only way to minimise the score further is to increase the modified SA score while simultaneously decreasing the activation barrier.

### S1.2. Scoring Functions

Three scoring functions are developed to estimate the activation energy (called scoring energy in the following) of the MBH reaction with a tertiary amine as catalyst. Each function takes the molecular graph of the catalyst as an input and returns the scoring energy for the MBH reaction with that catalyst.

The `ts_scoring` function calculates the difference in electronic energy between the TS of the rate-limiting step (**TS3**) and the reactants (MA, *p*NBA, methanol, catalyst) using GFN2-xTB and

the GBSA solvent model for methanol (ts-scoring energy).[7,8] Instead of performing an exhaustive diastereomeric and conformational search for the lowest energy TS structure, a template structure of TS3 taken from Liu et al. is used.[9] The catalyst present in the template is replaced with the catalyst that is to be evaluated and 10 conformers of the catalyst are generated using ETKDG as implemented in RDKit.[10,11] Only conformers that differ by at least 0.5 Å in root-mean-square deviation (RMSD) on the heavy atoms from each other are retained. Each conformer is then optimised in methanol using GFN2-xTB and the GBSA solvent model while keeping the atoms of the template fixed at their respective position and the lowest energy conformer is retained. Moreover, 10 conformers of the catalyst on its own are generated, the same RMSD pruning is applied and each conformer is optimised in methanol. The activation energy is then calculated as the difference between the electronic energy of the approximated TS and the sum of electronic energies of the catalyst, *p*NBA, MA and methanol ($\Delta E_{\mathrm{A}}^{\ddagger} = E_{\mathrm{TS}} - (E_{\mathrm{catalyst}} + E_{\mathrm{MA}} + E_{p\mathrm{NBA}} + E_{\mathrm{MeOH}})$).

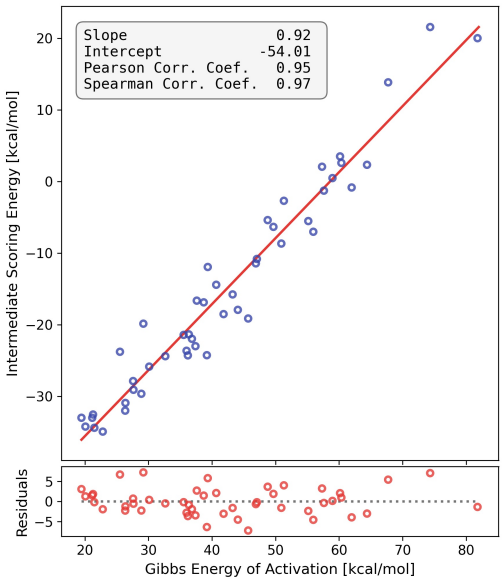



**Figure S1.** Correlation between the Gibbs energy of activation and the activation energy calculated using the `inter_scoring` function for 48 molecules

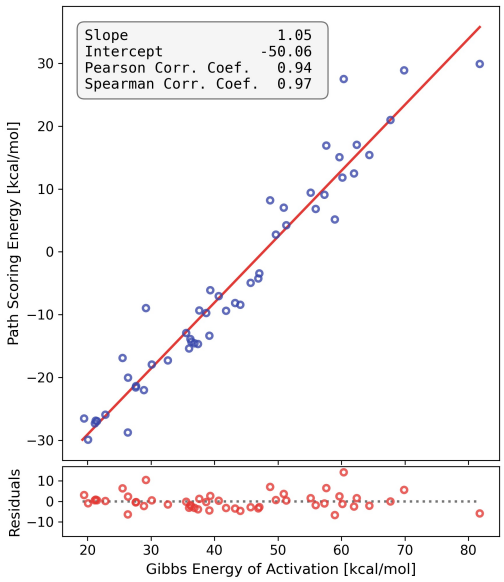



**Figure S2.** Correlation between the Gibbs energy of activation and the activation energy calculated using the `path_scoring` function for 50 molecules

The `inter_scoring` function follows a similar approach as the `ts_scoring` function, but here the difference in electronic energy between intermediate **2** and the reactants is calculated (intermediate-scoring energy). A template structure of intermediate **2** from Liu et al. is used and the present catalyst is replaced with the catalyst that is to be evaluated.[9] Ten conformers of the catalyst are generated while keeping the atoms of the template fixed to their respective position in the reference structure of intermediate **2**. After RMSD pruning, each conformer undergoes optimisation in methanol and the lowest energy conformer is retained. The lowest energy conformer of the catalyst on its own is obtained as previously described and the activation energy is calculated as the difference between the electronic energy of intermediate **2** and the sum of electronic energies of the catalyst, *p*NBA and MA ($\Delta E_{\mathrm{A}}^{\ddagger} = E_{\mathrm{Intermediate}} - (E_{\mathrm{catalyst}} + E_{\mathrm{MA}} + E_{p\mathrm{NBA}})$).

The correlation between $\Delta \mathrm{G}^{\circ,\ddagger}$ and the `inter_scoring` energy is comparably strong with a Pearson correlation coefficient of 0.95, see Figure S1. Again, the Spearman correlation coefficient of 0.97 is high and shows that the molecules can be successfully ranked according to their $\Delta \mathrm{G}^{\circ,\ddagger}$ by the `inter_scoring` function. For four out of the 52 molecules, the intermediate-scoring energy could not be calculated since the catalyst reacts with the oxygen-atom of the carbonyl-moiety of intermediate **2**. The reaction mechanism as studied by Liu et al. relies on this moiety in the rate-determining step and all molecules that react with the carbonyl-moiety are deemed ineffective catalysts and are assigned an overall score of 0.[9]

The `path_scoring` function calculates the difference between **TS3** and the reactants similar to the `ts_scoring`. Here, instead of using the template for **TS3**, the structure is obtained as the maximum energy structure along the reaction path as approximated by the RMSD-PP method

(path-scoring energy).[12] The reaction path is created for the abstraction of the proton by a methoxy moiety from a tertiary carbon atom which corresponds to **TS3**. To generate the reaction path, a template structure of the structure before proton abstraction is used, the present catalyst is replaced with the one that is to be evaluated and 10 conformers of the new catalyst are generated. After RMSD pruning all retained conformers undergo geometry optimisation in methanol and the lowest energy structure is retained. The structure of the catalyst is cut off the intermediate and attached to a template of an intermediate after TS3 (intermediate 10) to ensure that the RMSD between the structure before and after the TS is as small as possible with the catalyst being the same conformer and having the same orientation. The so-obtained structure of intermediate **3** is optimised in methanol. The reaction path method as implemented in xTB (version 6.4.1) with $k_{\mathrm{push}} = 0.003$, $k_{\mathrm{pull}} = -0.02$ and $\alpha = 1.6$ is used to obtain an approximate transition state structure as the structure with the highest energy along the reaction path.[12] 10 conformers of the catalyst on its own are generated, and after RMSD pruning the conformers undergo geometry optimisation in methanol and the lowest energy conformer is retained. The activation energy is then obtained as the difference between the electronic energy of the approximated transition state and the sum of electronic energies of the catalyst, *p*NBA, MA and methanol ($\Delta E_{\mathrm{A}}^{\ddagger} = E_{\mathrm{TS}} - (E_{\mathrm{catalyst}} + E_{\mathrm{MA}} + E_{p\mathrm{NBA}} + E_{\mathrm{MeOH}})$).

With the `path_scoring` function the activation energy could be calculated for 50 out of 52 molecules and the path-scoring energy is shown against $\Delta G^{\circ,\ddagger}$ in Figure S2. Here, the calculation of the path-scoring energy failed for two molecules because the pre-optimisation does not converge to the correct intermediate. Again, a strong linear correlation between the path-scoring energy and $\Delta G^{\circ,\ddagger}$ is shown in Figure S2 with a Pearson correlation coefficient of 0.94 and a Spearman rank correlation coefficient of 0.97.

On average, evaluating the score of a single molecule with the `ts_scoring`, `inter_scoring` or `path_scoring` function on a single core of an Intel® Xeon® E5-2643 v3 (3.4 GHz) takes 149±98 s, 178±158 s or 221±158 s respectively. Overall, the energies obtained from all three scoring functions show strong correlation with the Gibbs energy of activation for different molecules over a wide range of activation energies, but the `ts_scoring` function is the fastest and most robust and is thus used going forward.

### S1.3. DFT Calculations and Retrosynthetic Analysis

The best-performing molecules from separate GA runs are selected for further calculations of the electronic activation energy at the GFN2-xTB and B3LYP-D3/6-31+G(d,p)(SMD) level of theory.[13–16] First, the activation energy from the reactants to the true transition state (**TS3**) is calculated. The transition state optimiser in Gaussian16 is used with GFN2-xTB(methanol/GBSA) for the energy/gradient calculations.[17] The retrosynthesis tool Manifold is used to obtain the minimum number of synthetic steps to yield the desired product from commercially available building blocks.[18] The maximum lead time is set to 2 weeks and the catalogues 'generic', 'molport', 'emolecules' and 'mcule' are selected to search for building blocks.

For a few selected molecules, all relevant transition states and intermediates along the reaction path are located and harmonic frequencies calculated using Gaussian16, the B3LYP-D3 functional with the 6-31+G(d,p) basis set and the SMD solvent model for methanol. The Gibbs free energy is calculated with a standard state of 1 M for all solutes and 24.9 M for methanol. Furthermore, vibrational frequencies below 20 $\mathrm{cm}^{-1}$ are raised to 20 $\mathrm{cm}^{-1}$ before calculating vibrational contributions to the enthalpy and entropy following the approach of Ribeiro et al.[19]

## S2. Correlation between experimental rate constants and calculated activation energies

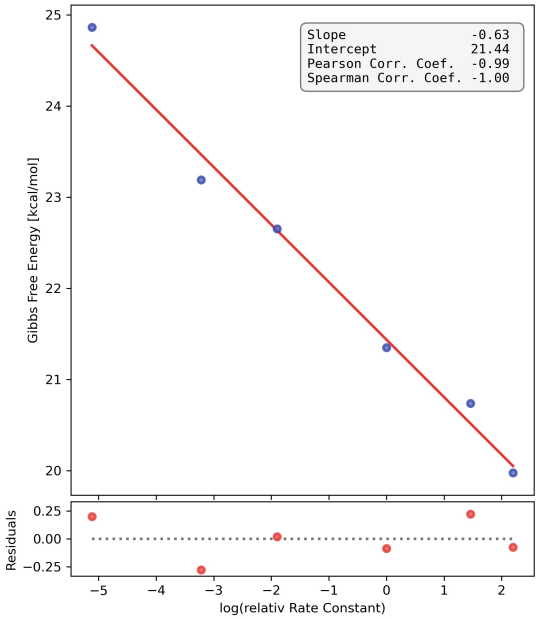



**Figure S3.** Correlation between the logarithm of the relative rate constant for the MBH reaction (values from Aggarwal et al.[20]) and the calculated Gibbs energy of activation (B3LYP-D3/6-31+G(d,p)(methanol/SMD)) for six quinuclidine-based catalysts

## S3. Molecules obtained from GA runs

## S4. GA run with azetidine filter

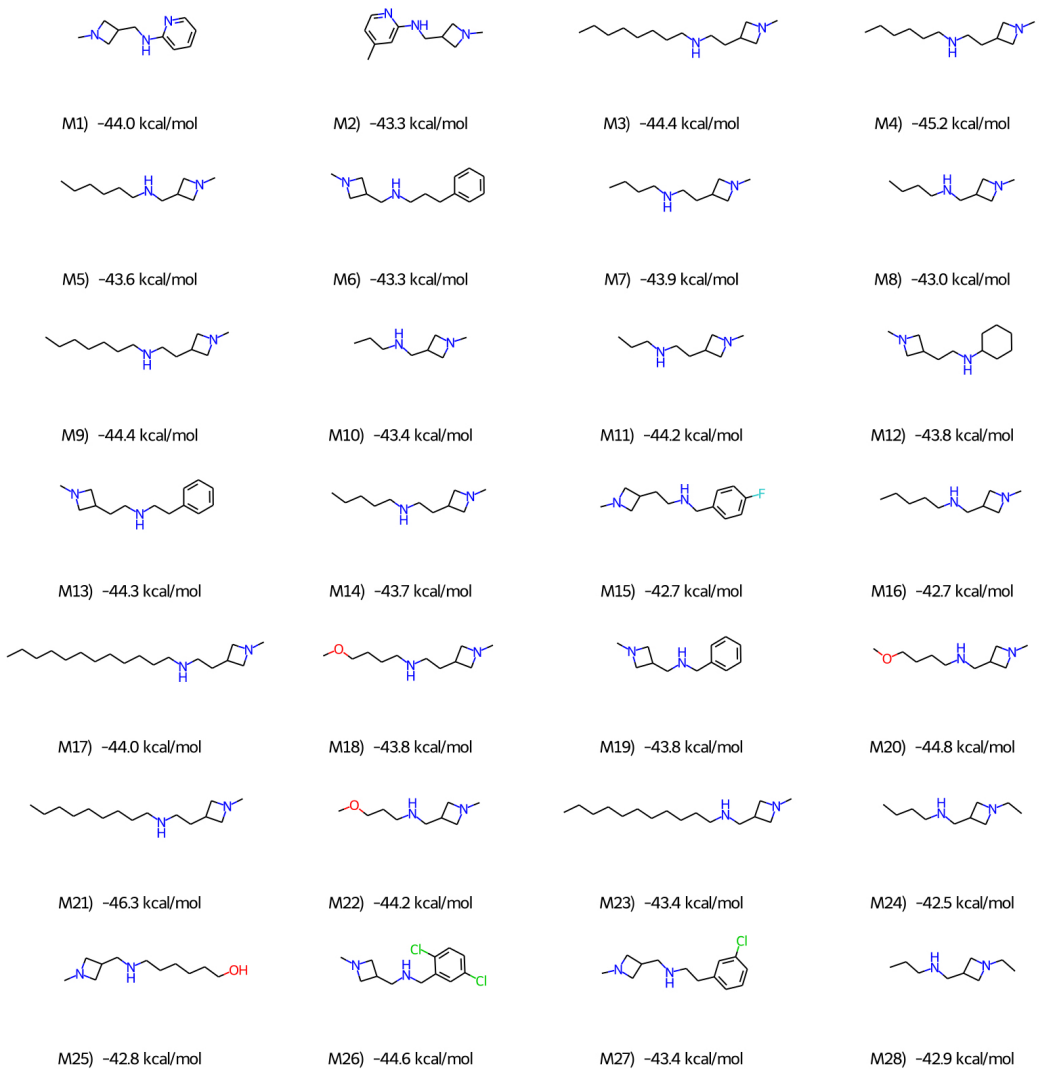



**Figure S4.** Molecules that can be synthesised in one step from building blocks and which induce low activation energies in the MBH reaction

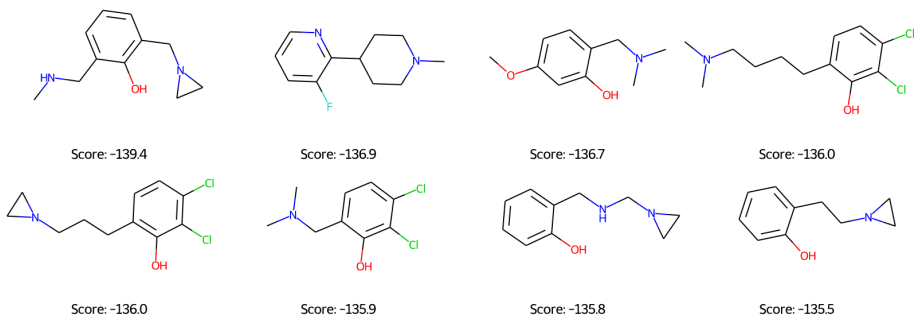
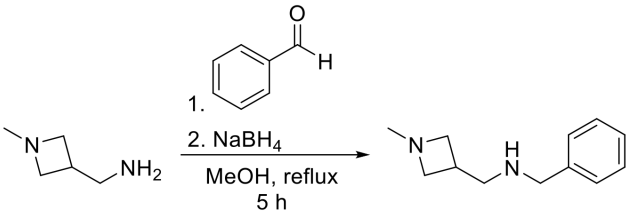



**Figure S5.** Top 8 molecules found by the GA when the azetidine filter is applied

## S5. Synthesis of M19

Starting materials are commercially available and were used without further purification. Methanol used as solvent was HPLC grade and was used without any purification or drying. Chromatographic purification was performed using silica gel (Flash column: 40–63 μm) and likewise, thin layer chromatography analysis (TLC) using aluminium sheets coated with silica gel containing a fluorescent indicator. NMR spectroscopic data were recorded at 500 MHz for $^{1}$H-NMR and 126 MHz for $^{13}$C NMR using a Bruker instrument equipped with a cryoprobe. Solvent references are: deuterated methanol ($CD_3OD$, $^{1}$H-NMR: $\delta$ = 3.31 ppm, $^{13}$C NMR: $\delta$ = 40.00 ppm). $^{13}$C-APT-NMR spectra were phased so negative signals correspond to an uneven number of protons on a carbon atom and positive signals correspond to zero or an even number of protons on a carbon atom. The instrument used for HRMS was a Bruker Solarix XR 7T ESI/MALDI-FT-ICR-MS instrument run in ESI-mode. Calibration was external using NaTFA cluster ions. Data processing was done using Bruker DataAnalysis version 5.0 SR1.




(1-Methylazetidin-3-yl)methanamine (100 μL, 85.6 mg, 0.85 mmol) was dissolved in methanol (20 mL) in a 50 mL round-bottomed flask. Benzaldehyde (91 μL, 0.90 mmol) was added, and the reaction mixture was heated to reflux. After 1 hour of refluxing, $NaBH_4$ (70.6 mg, 1.87 mmol) was added, and refluxing was continued for 1 hour. Another portion of $NaBH_4$ (79.8 mg, 2.10 mmol) was added and refluxing continued for another 30 minutes. The heating was stopped, and the reaction mixture was cooled to room temperature, and $CH_2Cl_2$ (30 mL) was added. The mixture was washed with water (50 mL), and the aqueous phase was extracted with $CH_2Cl_2$ (3 x 50mL) and the combined organic phases concentrated in *vacuo*. Purification by flash column chromatography ($SiO_2$, 49%/49%/2% of $CH_2Cl_2$/Heptane/$Et_3N$) furnished **M19** (114 mg, 70%) as a clear oil. $^{1}$H-NMR (500 MHz, $CD_3OD$) δ 7.36–7.28 (m, 4H), 7.28–7.21 (m, 1H), 3.73 (s, 2H), 3.51-3.43 (m, 2H), 2.96 (dd, J=8.1 Hz, 6.9 Hz, 2H), 2.74 (d, J=7.4 Hz, 2H), 2.62 (hept, J=7.1 Hz, 1H), 2.31 (s, 3H). $^{13}$C NMR (126 MHz, MeOD) δ 140.54, 129.53, 129.48, 128.22, 61.41, 54.48, 53.43, 45.68, 31.25. HRMS (ESI) $m/z$ = 191.15440 [M+H$^+$], calc. for [$C_{12}H_{19}N_2^+$]: 191.15428.

## S6. Kinetics

### S6.1. NMR Spectroscopic Measurements

The NMR spectroscopic kinetic measurements were performed by first making two stock solutions of DABCO (412.5 mg in 1 mL $CD_3OD$, concentration: 3.68 M) and 4-nitrobenzaldehyde (183.4 mg in 2 mL $CD_3OD$, concentration: 0.607 M). 4-Nitrobenzaldehyde was difficult to get into solution, and it was required to sonicate for a couple of minutes combined with waiting for 24 hours to allow equilibration of the reaction between the aldehyde and methanol to form the more soluble hemiacetal. To the 4-nitrobenzaldehyde solution, cyclohexane (22 μL) was added as an internal standard. To an NMR tube, the 4-nitrobenzaldehyde solution (0.572 mL), the DABCO solution (28.3 μL) and methyl acrylate (29.5 μL) were mixed, which gave a ratio of 1:0.94:0.3 (4-nitrobenzaldehyde/methyl acrylate/DABCO) and concentrations of 0.578 M, 0.543 M, and 0.173 M, respectively. The NMR tube was given a quick shake and measured at regular intervals. The internal standard (cyclohexane) could be used as a reference integral in each measurement, and the concentration of each species could be calculated.

During the experiment, a couple of side reactions took place and an overview of the products of the side reactions is shown below in Figure S6.

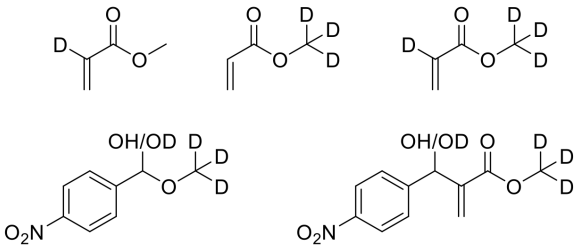


**Figure S6.** Deuterated reactants, intermediates and product

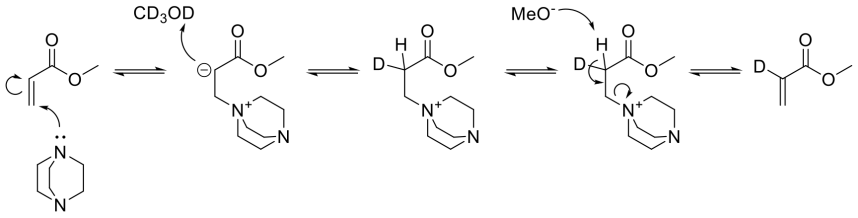


**Figure S7.** Proposed mechanism for deuterium exchange at α-C on methyl acrylate

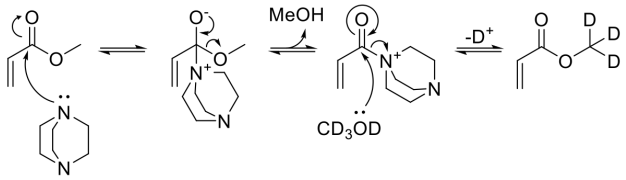


**Figure S8.** Proposed mechanism of $CD_3OD$ incorporation in methyl acrylate with DABCO as a catalyst. Using amine **M19** would result in an even faster incorporation of $CD_3OD$

All side reactions yield deuterated analogues of reactants, intermediates or the MBH product. As described in section 2, the signal corresponding to one of the β-protons of MA was used to determine the concentration of MA and to follow the course of the reaction. This β-proton is present in all side-products of MA and its chemical shift is barely affected by deuteration of MA, only a change from doublet to singlet is observed. Therefore, all loss in intensity of the signal corresponding to this β-proton can be attributed to conversion of MA to the MBH product.

The catalyst takes part in all those side reactions and the actual concentration of catalyst that is available for the MBH reaction is lower than assumed. In Figure S9, one can see that both side reactions, the incorporation of deuterium into the α-position of MA and the transesterification with deuterated methanol, are considerably faster using **M19** than using DABCO. However, this does not necessarily mean that the concentration of **M19** that is available for the MBH reaction is lower than that of DABCO, since that depends on the relative life-times of the intermediates.

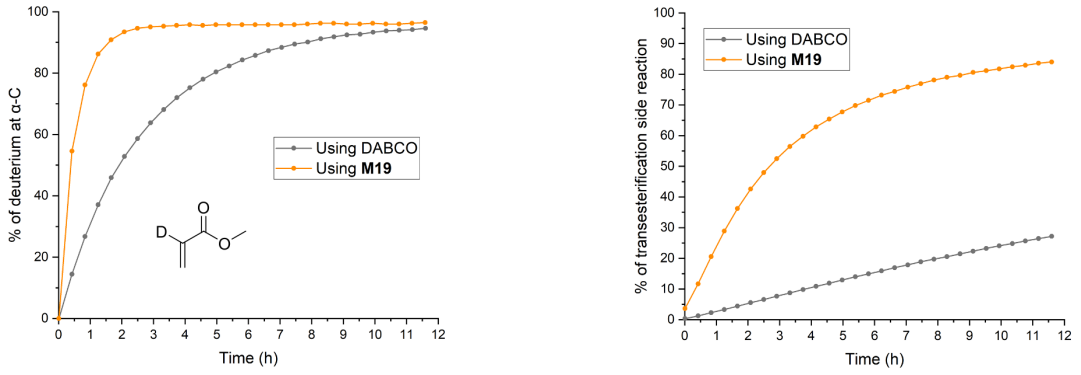



**Figure S9.** Left: percentage of deuterium incorporation of methyl acrylate in the α-position over time. Right: percentage of transesterification of methyl acrylate with $CD_3OD$ over time calculated from the free MeOH generated with this side reaction.

## S6.2. Calculation of Reaction Rates

The rate of the catalysed MBH reaction of MA with *p*NBA following the mechanism outlined in Figure 1 is described by the pseudo-second-order rate law in Eq. S3, following Plata and Singleton.[21]

$$\begin{aligned} v = \frac{d\,[\mathrm{P}]}{dt} &= k \cdot [\mathrm{MA}]\,[p\mathrm{NBA}]\,[\mathrm{Catalyst}] \\ &= k' \cdot [\mathrm{MA}]\,[p\mathrm{NBA}] \end{aligned} \tag{S3}$$

The integrated rate law is shown in Eq. S4 with the following variables and values:

$$\begin{aligned} [p\mathrm{NBA}]_0 &= \text{Starting Concentration of } p\mathrm{NBA} = 0.578\ \mathrm{M} \\ [\mathrm{MA}]_0 &= \text{Starting Concentration of MA} = 0.543\ \mathrm{M} \\ [\mathrm{Catalyst}] &= \text{Concentration of DABCO/M19} = 0.173\ \mathrm{M}/0.149\ \mathrm{M} \\ [\mathrm{P}] &= \text{Concentration of Product} \end{aligned}$$

$$k't = \frac{1}{[p\mathrm{NBA}]_0 - [\mathrm{MA}]_0} \cdot \ln\left(\frac{([p\mathrm{NBA}]_0 - [\mathrm{P}])\,[\mathrm{MA}]_0}{([\mathrm{MA}]_0 - [\mathrm{P}])\,[p\mathrm{NBA}]_0}\right) = y \tag{S4}$$

The constant $k'$ can be obtained from a linear fit to the right-hand side of Eq. S4 against the reaction time which is shown in Figure S10 ($y$ vs. $t$). Following Plata and Singleton, the expression $y = k' \cdot t$ is fitted only to the first four data points since the incorporation of deuterium into the α-position of MA over the course of the reaction decreases the reaction rate.[21]

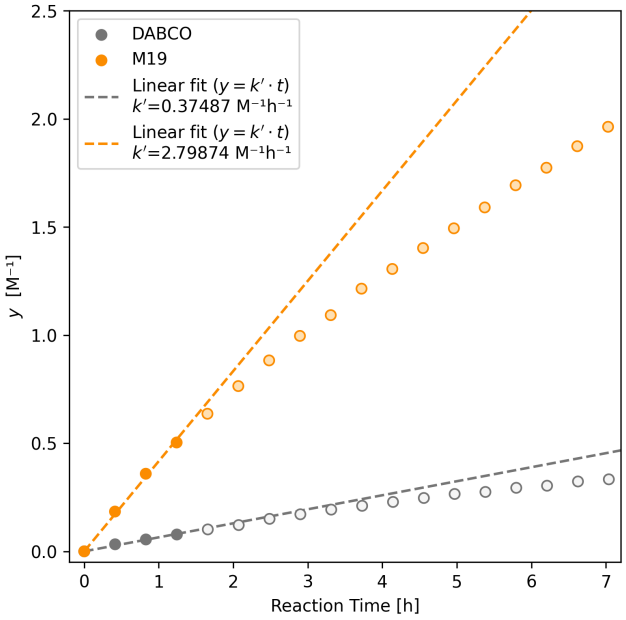



**Figure S10.** Fitting of constants $k'$ to $y$ (Eq. S4) with concentrations obtained from $^1$H-NMR spectroscopic measurements. Data for DABCO are shown in grey, data for **M19** in orange. The constants $k'$ are obtained as the slope of a linear fit to the first four data points.

The so-obtained constants $k'$ can be concerted into reaction rate constants by dividing it by [Catalyst] and converting from hours to seconds:

$$\begin{aligned} k'\,(\mathrm{DABCO}) &= 0.37487\ \mathrm{M^{-1}h^{-1}} \\ k\,(\mathrm{DABCO}) &= 0.00010\ \mathrm{M^{-2}s^{-1}} \\ k'\,(\mathrm{M19}) &= 2.79874\ \mathrm{M^{-1}h^{-1}} \\ k\,(\mathrm{M19}) &= 0.00078\ \mathrm{M^{-2}s^{-1}} \end{aligned}$$

With the constants $k'$ the concentration of the product ([P]) at any time $t$ can be obtained using Eq. S5 which was used to calculate the expected conversion to the product in Figure 8.

$$[\mathrm{P}] = \frac{[p\mathrm{NBA}]_0\,[\mathrm{MA}]_0\,(\exp([p\mathrm{NBA}]_0\,k't) - \exp([\mathrm{MA}]_0\,k't))}{[p\mathrm{NBA}]_0 \cdot \exp([p\mathrm{NBA}]_0\,k't) - [\mathrm{MA}]_0 \cdot \exp([\mathrm{MA}]_0\,k't)} \tag{S5}$$

## References


[1] J. H. Jensen, *Chemical science* **2019**, *10*, 3567–3572.

[2] T. Sterling, J. J. Irwin, *Journal of chemical information and modeling* **2015**, *55*, 2324–2337.

[3] J. E. Baker, *undefined* **1985**.

[4] W. Gao, C. W. Coley, *Journal of chemical information and modeling* **2020**, *60*, 5714–5723.

[5] N. Brown, M. Fiscato, M. H. S. Segler, A. C. Vaucher, *Journal of chemical information and modeling* **2019**, *59*, 1096–1108.

[6] P. Ertl, A. Schuffenhauer, *Journal of cheminformatics* **2009**, *1*, 8.

[7] C. Bannwarth, S. Ehlert, S. Grimme, *Journal of chemical theory and computation* **2019**, *15*, 1652–1671.

[8] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, S. Grimme, *Wiley interdisciplinary reviews. Computational molecular science* **2021**, *11*, DOI 10.1002/wcms.1493.

[9] Z. Liu, C. Patel, J. N. Harvey, R. B. Sunoj, *Physical chemistry chemical physics: PCCP* **2017**, *19*, 30647–30657.

[10] S. Riniker, G. A. Landrum, *Journal of chemical information and modeling* **2015**, *55*, 2562–2574.

[11] G. Landrum, P. Tosco, B. Kelley, Ric, sriniker, gedeck, R. Vianello, NadineSchneider, E. Kawashima, A. Dalke, N. Dan, B. Cole, D. Cosgrove, M. Swain, S. Turk, AlexanderSavelyev, G. Jones, A. Vaucher, M. Wójcikowski, D. Probst, V. F. Scalfani, G. Godin, A. Pahl, F. Berenger, JLVarjo, K. Ujihara, strets, JP, DoliathGavid, G. Sforna, rdkit/rdkit: 2021_09_4 (Q3 2021) Release, **2022**.

[12] S. Grimme, *Journal of chemical theory and computation* **2019**, *15*, 2847–2862.

[13] A. D. Becke, *The Journal of chemical physics* **1993**, *98*, 5648–5652.

[14] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, *The Journal of chemical physics* **2010**, *132*, 154104.

[15] W. J. Hehre, R. Ditchfield, J. A. Pople, *The Journal of chemical physics* **1972**, *56*, 2257–2261.

[16] A. V. Marenich, C. J. Cramer, D. G. Truhlar, *The journal of physical chemistry. B* **2009**, *113*, 6378–6396.

[17] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr, J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, Gaussian16 Revision C.01, **2016**.

[18] PostEra, Medicinal Chemistry Powered by Machine Learning, en, https://postera.ai/manifold/, Accessed: 2022-3-16.

[19] R. F. Ribeiro, A. V. Marenich, C. J. Cramer, D. G. Truhlar, *The journal of physical chemistry. B* **2011**, *115*, 14556–14562.

[20] V. K. Aggarwal, I. Emme, S. Y. Fulford, *The Journal of organic chemistry* **2003**, *68*, 692–700.

[21] R. E. Plata, D. A. Singleton, *Journal of the American Chemical Society* **2015**, *137*, 3811–3826.

### 3.4.3 *Summary and Outlook*

In this study, we demonstrated how GAs can effectively optimize the catalytic activity of molecules for specific reactions, using the well-documented MBH reaction as our model. This reaction was chosen due to its comprehensive exploration through both experimental and computational methods. However, a significant challenge arises when the mechanism of a reaction is not yet known, which is a more common scenario and requires not only optimizing but also identifying potential catalysts and their mechanisms.

In Paper 4, we address this challenge by employing an automated molecular meta-dynamics approach to explore the reaction network of the MBH reaction. Starting with a set of 11 diverse molecular templates, Rasmussen [30] explored the corresponding reaction networks, identified the catalytic cycle, selected a promising molecular template as a potential catalyst and located the RDS and corresponding TS. Utilizing this TS structure as a template in our GA's scoring function, we replicated the experiments from Paper 1 with identical parameters and initial populations. The results were consistent, revealing similar molecules and reidentifying azetidine moieties as advantageous structural motifs.

These findings underscore that the discovery of a novel catalyst does not rely on prior knowledge of the reaction mechanism or template structures from known catalysts, which marks a true de novo discovery. This breakthrough paves the way for the discovery of catalysts for reactions that are currently impractical or require harsh conditions, significantly broadening the scope of catalytic science and potential applications. This constitutes true de-novo discovery since no prior knowledge of the reaction mechanism is required.

Since the majority of homogeneous catalysts are TM-based, our next challenge is to extend the here presented workflow to those complexes, which we will show in Paper 2.

## 3.5 TM-BASED CATALYSTS

In the following section, we present our work on discovering novel ligands for catalysts based on the TMs Palladium, Copper and Silver. We use a fragment-based and graph-based GA approach and calculate the catalytic activity based on intermediate reaction energies for comparison to a volcano plot, as introduced in Section 3.2. We choose this metric for direct comparison with previous work done by Meyer et al. [27] and Schilter et al. [15]. We showcase how QM calculations in connection with GAs can be used efficiently to search through large enumerated spaces and generate novel ligands beyond them. This approach calculates the descriptor of the catalytic activity from first principle instead of relying on ML models to relate a featurized representation to an intermediate reaction energy. We expect this approach to be more robust and give reliable values for novel, out-of-domain catalysts. This feat is especially important for the de novo discovery presented here.

### 3.5.1 *Motivation for Paper 2*

Encouraged by our previous achievements in utilizing GAs to uncover novel organic catalysts, we aimed to expand this approach to include the discovery of TM-based catalysts. Given that the majority of homogeneous catalysts involve TMs, developing a workflow for the de novo generation of such catalysts represents a substantial advancement. To this end, we chose the Suzuki reaction as a test case to develop a workflow that optimizes the ligands of a TM-based catalyst. Traditionally, the design of TM-based catalysts has depended on predefined ligand libraries. Recombination of a set of less than 300 unique ligands can enumerate a vast space of up to billions of compounds.[8, 19] GAs have been successfully used to navigate through these virtual spaces. However, these predefined spaces limit the potential for discovering entirely new chemistry that would potentially allow for efficient catalysts based on cheaper or less toxic TMs.

### 3.5.2 *Summary of key findings and discussion*

We have demonstrated that GAs can be integrated with DFT for evaluations through efficient parallelization, offering a cost-effective alternative to high-throughput screening in predefined spaces by using a fragment-based GA. Next, we used a graph-based GA workflow that goes beyond traditional enumerated libraries. It includes automatic assembly, coordinate generation, and a hierarchical QM workflow that facilitates the calculation of a thermodynamic descriptor of the catalytic cycle. This methodology is expected to be more robust than ML-based approaches for exploring new areas of chemical space, as

it bases scoring on first principles rather than ML correlations. With this workflow, compute-intensive scoring processes can be applied to TM-based catalysts. Furthermore, we show how a graph-based GA traverses chemical space and generates ligands with new coordination modes that improve the catalyst's activity measure.

We utilize a SA measure developed for drug-like molecules to assess ligand's stability and synthesizability, which considerably constrains the exploration to a somewhat drug-like subset of chemical space. Additionally, this approach does not give insight into whether the resulting TM complex will be stable or synthesizable. Further development of this workflow requires the assessment of stability and synthesizability of TM-based complexes.

One way of addressing this challenge is to calculate how similar a complex and its fragments are to complexes that are known to be synthesizable. While the Cambridge Structural Database holds coordinates and connectivity information of several thousand TM-based complexes, no bond orders are assigned, which makes such an analysis difficult. Our group is currently working on assigning bond orders to those complexes. Upon completion, the resulting dataset containing molecular graphs, including bond orders of TM-based complexes, can be analysed, and a statistical model created that calculates the similarity of novel complexes to known complexes. We anticipate that such a development will facilitate the efficient exploration of chemical space relevant for TM-based catalysis.

Looking ahead, we envision using the here presented workflow to comprehensively explore the reaction profile of a Suzuki reaction, integrating all relevant activation barrier heights into our scoring. This approach would move us away from relying solely on LESRs, incorporating actual kinetics instead of thermodynamic descriptors.

# Beyond Predefined Ligand Libraries: A Genetic Algorithm Approach for De Novo Discovery of Catalysts for the Suzuki Coupling Reactions

Julius Seumer[a], Jan H. Jensen[a]*

[a] Julius Seumer, Jan H. Jensen
Department of Chemistry, University of Copenhagen, Denmark
* E-mail: jhjensen@chem.ku.dk, Twitter: @janhjensen

This study introduces a novel approach for the unrestricted de novo design of transition metal catalysts, leveraging the power of genetic algorithms (GAs) and density functional theory (DFT) calculations. By focusing on the Suzuki reaction, known for its significance in forming carbon-carbon bonds, we demonstrate the effectiveness of fragment-based and graph-based genetic algorithms in identifying novel ligands for palladium-based catalytic systems. Our research highlights the capability of these algorithms to generate ligands with desired thermodynamic properties, moving beyond the restriction of enumerated chemical libraries. Limitations in the applicability of machine learning models are overcome by calculating thermodynamic properties from first principle. The inclusion of synthetic accessibility scores further refines the search, steering it towards more practically feasible ligands. Through the examination of both palladium and alternative transition metal catalysts like copper and silver, our findings reveal the algorithms' ability to uncover unique catalyst structures within the target energy range, offering insights into the electronic and steric effects necessary for effective catalysis. This work not only proves the potential of genetic algorithms in the cost-effective and scalable discovery of new catalysts but also sets the stage for future exploration beyond predefined chemical spaces, enhancing the toolkit available for catalyst design.

## 1. Introduction

Catalysis plays a crucial role in synthetic chemistry and is fundamentally dependent on the formulation of catalysts that are both efficient and selective. A reaction of critical importance in this domain is the Suzuki coupling because of its ability to synthesize carbon-carbon bonds via the coupling of organohalides with boronic acids.[1, 2] The performance and specificity of this reaction are largely influenced by the ligand selection in palladium (Pd)-based catalytic systems.[3]

The traditional approaches to ligand discovery have been characterized by intensive experimental screening processes that are both time-intensive and demand significant resources.[4–6] Computational methodologies for identifying efficient catalysts have evolved, predominantly involving the screening of extensive enumerated libraries, typically ranging from $10^{3-5}$ catalysts.[7] These virtual screenings are executed through (semi-empirical) quantum mechanics (QM) calculations which limits the size of the screen library due to the computational cost.[8, 9] In response to this challenge, researchers have turned to machine learning (ML) models, trained on pre-existing data, to screen more extensive libraries, which may include up to $10^6$ catalysts or more.[10–12] While ML models have proven valuable within their trained domains, their efficacy often diminishes when applied to scenarios outside their original scope, particularly in the context of true de novo generation. Another approach to limit computational cost is to employ search algorithms aimed at efficient navigation of chemical space, such as genetic algorithms (GAs). However, a notable limitation of these methods is their tendency to restrict searches to predefined chemical spaces, often encompassing around $10^{5-7}$ catalysts.[13–18]

Recently, our research has introduced the de novo design of a highly efficient organic homogeneous catalyst, specifically devised for the Morita–Baylis–Hillman reaction.[19] Utilizing a genetic algorithm, we explored the unrestricted chemical space of tertiary amines, signifying a shift from traditional screening methods and restricted chemical spaces.

Following that, Strandgaard et al. [20] have shown the computational de novo design of fragments of ligands for the Schrock catalysts which extends the unrestricted genetic algorithm search to parts of inorganic homogeneous catalysts.

In this study we expands this concept to demonstrate the de novo design of whole ligands for transition metal catalysts at the example of the Suzuki reaction. While ML models trained on pre-existing data have proven effective in high-throughput screening, their use for unrestricted de novo design is challenging since their predictive performance on truly out of domain samples deteriorates. Another approach, relying on the correlation between geometric descriptors such as bond-lengths, cone angles or sterimol parameters and the thermodynamics of the catalytic sytem has been shown to be effective, yet it appears challening when moving beyond only one binding motif.[18, 21, 22] Therefore, we evaluate the performance of a catalysts by calculation of DFT level thermodynamic descriptors, instead of using ML models or (semi-empirical) QM calculated geometric descriptors. This distinctive approach expands the domain of our de novo design, allowing for the exploration of chemical space beyond the confines of ML model training or QSAR correlations.

## 2. Computational Methodology

We use a genetic algorithm (GA) to discover promising catalyst candidates for the Suzuki reaction. In the GA, the gene is represented as a list of two molecular fragments which are the ligands of the Suzuki catalyst. Two different GAs are used which differ only in their reproduction rules.

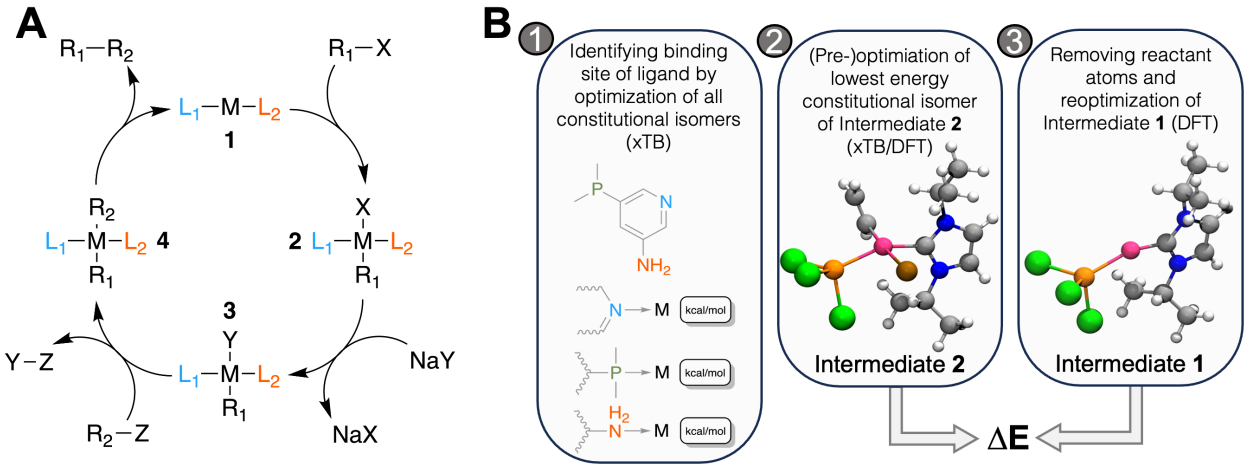



**Figure 1.** **A**: Catalytic cycle of the Suzuki reaction with key intermediates **1** and **2**, **B**: Workflow to calcualte ΔE for a catalyst: 1. for ligands with multiple potential binding site, all constitutional isomers are generated and their structures optimized with GFN2-xTB. 10 conformers of the lowest energy constitutional isomer are then optimized with GFN2-xTB and the lowest energy structure is further optimized at the DFT level in step 2. The atoms corresponding to the reactant (**$R_1$-X**) are removed and the structure is optimized in step 3. ΔE is then obtained as the difference in electronic energy at the DFT level.

In the fragment-based GA (FBGA), a crossover operation means that one ligand is exchanged for another ligand from another catalyst. The mutation operation exchanges one of the two ligands with another selected randomly from a pre-defined list of ligands.

The graph-based genetic algorithm (GBGA) utilizes the crossover and mutation operations as implemented by Jensen [23]. During a crossover operation, the graph of one of the ligands is cut at random points and recombined with a fragment of a molecular graph from another individual. The mutation operations act directly on the molecular graph, adding, changing or removing atom(-type) or connectivity. Newly generated ligands are considered valid if they contain at least one molecular pattern that is considered to be a potential coordination site for the ligand. These patterns are phosphines, amines, carbenes are carbonyls. When more than one potential coordination site is detected, all possible constitutional conformers of the ligand attached to a Pd-containing reference catalyst are generated and 25 conformers of each are embedded and optimized at the GFN2-xTB level of theory.[24] The coordination site with the highest binding energy is then chosen as the coordination site for that ligand.

All GAs are run with a population size of 25 for up to 50 generations. The mutation rate is set to 50 % which means that with a 50:50 chance either a crossover or mutation operation is chosen for the reproduction. Ligands with as many as 30 heavy atoms and/or 5 rotatable bonds are allowed. The starting population is created from combinations of ligands from a pre-defined list containing 91 different amines, phosphines, N-heterocyclic carbenes, pyridines and CO taken from Meyer et al. [7].

For both GAs, the fitness of each individual depends on the difference in electronic energy of intermediate **2** and **1**, see Figure 1. As Meyer et al. [7] established using linear-energy scaling relationships following an approach by Busch et al. [25] and Wodrich et al. [26], the optimal difference in electronic energy (ΔE) between the two intermediates is in the range of -32.1 and $-23.0\,\mathrm{kcal\,mol^{-1}}$ at the B3LYP-D3BJ/def2-TZVP//B3LYP-D3BJ/3-21 level.[27–34] We calculate ΔE at the B3LYP-D3BJ/def2-TZVP//B3LYP-D3BJ/3-21 level and convert it to a score between 0 and 1 using a Gaussian function centered around $-27.55\,\mathrm{kcal\,mol^{-1}}$ ($\mu_1$) and a standard deviation ($\sigma_1$) of $6.00\,\mathrm{kcal\,mol^{-1}}$, chosen empirically, see Equation 1. Therefore, the closer the calculated ΔE is to the target value of $-27.55\,\mathrm{kcal\,mol^{-1}}$ the closer the score is to 1.0.

To calculate ΔE for a molecule, 10 conformers of intermediate **2** are embedded using ETKDG as implemented in a slightly modified version of RDKit based on 2023.03.2, see subsection S1, and an RMSD pruning threshold of 0.25 Å is applied.[35, 36] The retained conformers are optimized at the GFN2-xTB level of theory.[24] The lowest energy conformer is further optimized at the B3LYP-D3BJ/3-21 level and its single point energy is calculated as the B3LYP-D3BJ/def2-TZVP level using ORCA 5.0.4.[37] The fragments **$R_1$-X** ($HC{=}CH_2$) and **X** (Br) are removed from the optimized structure to obtain a structure of intermediate **1** which undergoes optimization and single point calculation at the B3LYP-D3BJ/def2-TZVP//B3LYP-D3BJ/3-21 level of theory. ΔE is then obtained as the difference in electronic energy between intermediate **2** and intermediate **1** and **$R_1$-X**.

In the FBGA, the final score of individual $i$ is equal to the first factor in Equation 1 which only depends on $\Delta E_i$. In the GBGA, this score is multiplied with a modified synthetic accessibility (SA) score, which also ranges from 0.0 to 1.0 as shown in Equation 1. Here, the modified SA score is the mean of the two SA scores of the ligands which are calculated as described by Ertl and Schuffenhauer [38] and further modified using a modified Gaussian function as proposed by Brown et al. with the parameters $\mu_2 = 2.230044$ and $\sigma_2 = 0.6526308$ as used by Gao and Coley.[39, 40] Therefore, the final score in both GAs ranges from 0.0 to 1.0.

$$\mathrm{Score}_i = \exp\left(\frac{-\left(\Delta E_i - \mu_1\right)^2}{2\sigma_1^2}\right) \cdot \exp\left(\frac{-\left(\max\left(\bar{SA}_i, \mu_2\right) - \mu_2\right)^2}{2\sigma_2^2}\right) \tag{1}$$

We choose to calculate the overall score as the product of the two normalised objectives, the energy-dependent term and the synthesizability penalty so that an optimal solution can only be found when both objectives are satisfied since we are not interested in high synthesizable molecules that do not perform well as catalysts or molecules that are calculated to perform well as catalysts but are not synthesizable or show uncommon structural motifs. This is achieved by using the product as shown in Equation 1, compared to, for example, a sum of the two terms.

Based on the rank $r_i$ of each individual in each population ($N$ = population size), a normalized fitness value is calculated using Equation 2 from Baker [41] with a selection pressure (SP) of 1.5. Individuals are selected for reproduction with a frequency proportional to their normalized fitness value.

$$p_i = \frac{1}{N}\left(2 - \mathrm{SP} + 2 \cdot (\mathrm{SP} - 1) \cdot \frac{r_i - 1}{N - 1}\right) \qquad (2)$$

# 3. Results and Discussion

## 3.1. Fragment-based GA

Firstly, we assess the ability of a fragment-based GA to locate catalysts based on Pd within the defined ΔE range. The starting population is comprised of 25 molecules that

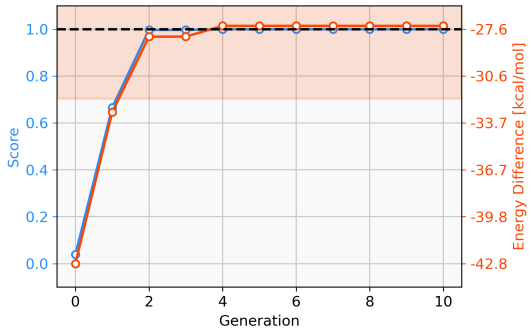



**Figure 2.** Evolution of the score (blue) and energy difference (red) of the best-performing individual of an FBGA run over 10 generations. The target range of the energy difference is shown as a red-shaded area.

all have a calculated ΔE below $-42.00\,\mathrm{kcal\,mol^{-1}}$ (ΔE far below target value). Subsequently, the GA explores if combinations of the present ligands can yield molecules with ΔE values closer to the target of $-27.55\,\mathrm{kcal\,mol^{-1}}$ and adds new ligands from the full list of available ligands via mutation operations. The evolution of the best-performing molecule over 10 generations is shown in Figure 2. The score of the best-performing individual quickly increases from 0.04 to 0.99 within two generations. Closer inspection of the GA run reveals that the molecule is created by two successive mutation operations as shown in Figure 3. After four generations the best-performing molecule has a calculated energy difference of $-27.33\,\mathrm{kcal\,mol^{-1}}$ which yields a score of 1.00. Over the next six generations, more molecules are found with a score of 1.00 and after ten generations all 25 molecules have a calculated energy difference in the target range as defined by Meyer et al. [7]. Here, we performed $2 \cdot 25 \cdot 10 = 500$ DFT optimization (two for each catalyst, 25 catalysts for 10 generations) to locate a total of 134 unique catalysts within the target range. Meyer et al.

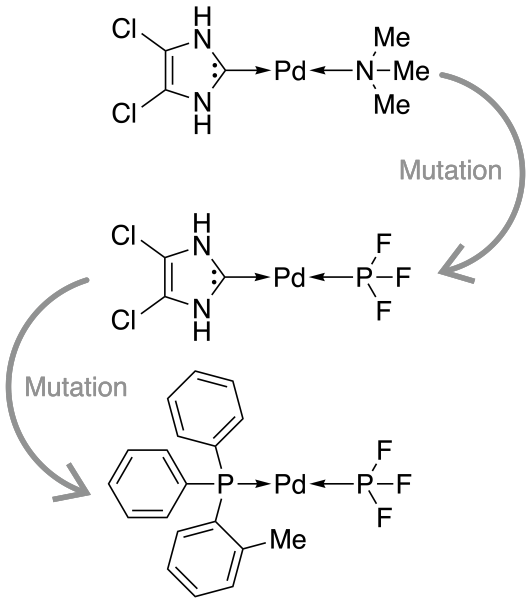



**Figure 3.** Evolution of the best-performing individual of an FBGA run after two generations

[7] were able to identify 265 unique Pd catalysts that are predicted to have an energy difference in the target range using an ML model to exhaustively screen the same library we search with this GA. Their ML model was trained on a total of 7054 molecules out of which 2595 contained Pd. If they restricted themselves to only Pd-containing catalysts and were able to obtain a model performing similarly on Pd-containing catalysts, they would have been able to find 265 catalysts while doing $2 \cdot 2595 = 5190$ DFT optimization (two for each molecule). This highlights how GAs can be used in connection with DFT for scoring to search a large pre-defined library of ligands with comparably little computational cost.

## 3.2. Graph-based GA

Next, we aim to discover novel ligands for Pd-containing catalysts for the Suzuki reaction and not just recombine pre-defined ligands. From the same starting population, a GBGA without synthetic accessibility constraint is run

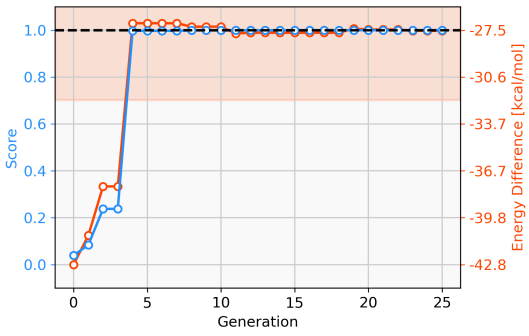



**Figure 4.** Evolution of the score (blue) and energy difference (red) of the best-performing individual of a GBGA run over 25 generations. The target range of the energy difference is shown as a red-shaded area.

for 20 generations. Here, the best-performing individual is created by subsequent crossover and mutation operations on the molecular graph of the ligands.

Since only parts of one ligand are changed in the GBGA instead of the whole ligand as in the FBGA, evolution of molecular structures happens in smaller steps through chemical space. This allows the discovery of novel structures. The score of the best-performing individual increases drastically over the first four generations and the calculated energy difference increases from $-42.8\,\mathrm{kcal\,mol^{-1}}$ to $-27.1\,\mathrm{kcal\,mol^{-1}}$, as shown in Figure 4. Over the following 16 generations, the energy difference of the best-performing molecule decreases slightly to $-27.59\,\mathrm{kcal\,mol^{-1}}$ which corresponds to a score of 1.00. After 20 generations, all molecules in the population have a calculated energy difference within the target range and the four best-performing individuals are shown in Figure 5. The ligands coordinate to the transition metal via phosphine or pyridine derivative sites and possess up to five hetero atoms. Although no

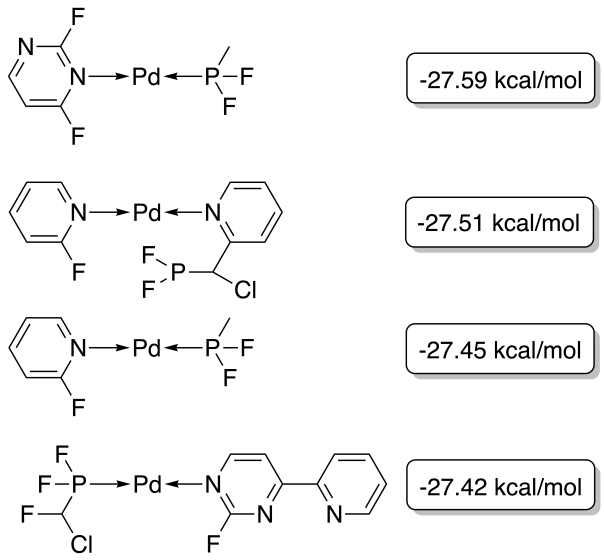



**Figure 5.** Best-performing individuals of the GBGA after 25 generations. The calculated energy difference is shown to the right of the structure.

synthetic accessibility constraint was applied, some purchasable ligands such as 2-fluoropyridine were discovered by the GA.[42] On the other hand, for many of the ligands containing highly fluorinated phosphines, no synthetic route can be found with the retrosynthesis software Manifold.[43]

We predict the binding site of each ligand as the site with the highest binding energy when coordinating as a monodentate ligand. This might not be a reasonable assumption for one of the ligands of the fourth molecule shown in Figure 5 which might coordinate as an N,N-bidentate ligand. One could perform more automated xTB calculations considering other coordination modes than simple monodentate coordination to verify this in the GA. This way the coordination site and mode could be identified by the highest binding energy across all sites and modes.

Here, we defer this additional consideration to verification and evaluation steps that are necessary after the molecular optimization, along with a more extensive conformational search, location of transition states and calculation of activation barriers, extensive retrosynthetic analysis and calculation of full catalytic cycles.

## 3.3. Graph-based GA with SA

To address this short-coming, a GBGA is started again from the same starting population with a synthetic accessibility constraint to the score as described in section 2 which stears the search into an area of chemical space that is deemed to be more accessible. Virtually all molecules of the initial population are deemed to be synthetically inaccessible by the modified SA score, as shown in Figure 6. This is not surprising since it was developed for drug-like molecules. Within ten generations, the GA discovers new ligands that

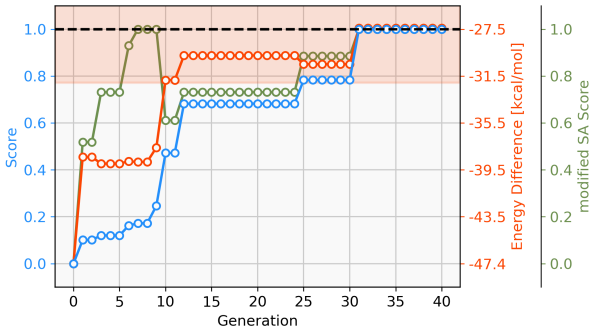



**Figure 6.** Evolution of the score (blue), energy difference (red) and modified SA score (green) of the best-performing individual of a GBGA run over 40 generations. The target range of the energy difference is shown as a red-shaded area.

are deemed to have moderate synthetic accessibility as well as an energy difference within the desired target range. Here, a trade-off has to be made between the two components of the score, the energy difference and the modified SA score. Although molecules with modified SA scores of 1.00 are found after seven generations, the modified SA score of the best-performing molecule decreases again to 0.6 after ten generations since the energy difference of the molecule is closer to the target value of $-27.55\,\mathrm{kcal\,mol^{-1}}$ which yields an increase to the overall score of 0.2. On the other hand, after 26 generations, the best-performing molecule has a less favorable energy difference than the best-performing molecule in the previous generation but this is compensated for by a higher modified SA score which yields an overall in-

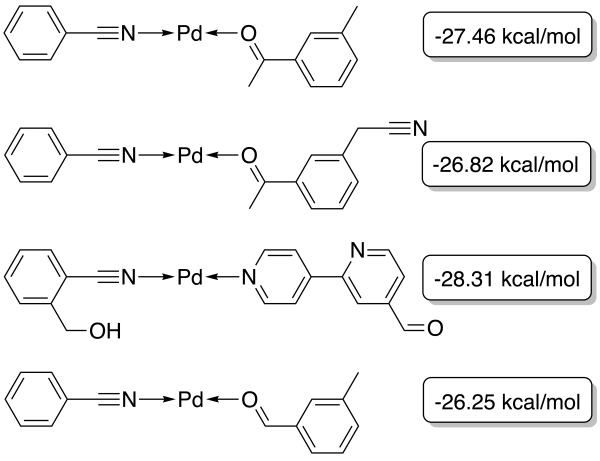



**Figure 7.** Best-performing individuals of the GBGA with synthetic accessibility constraint after 40 generations. The calculated energy difference is shown to the right of the structure.

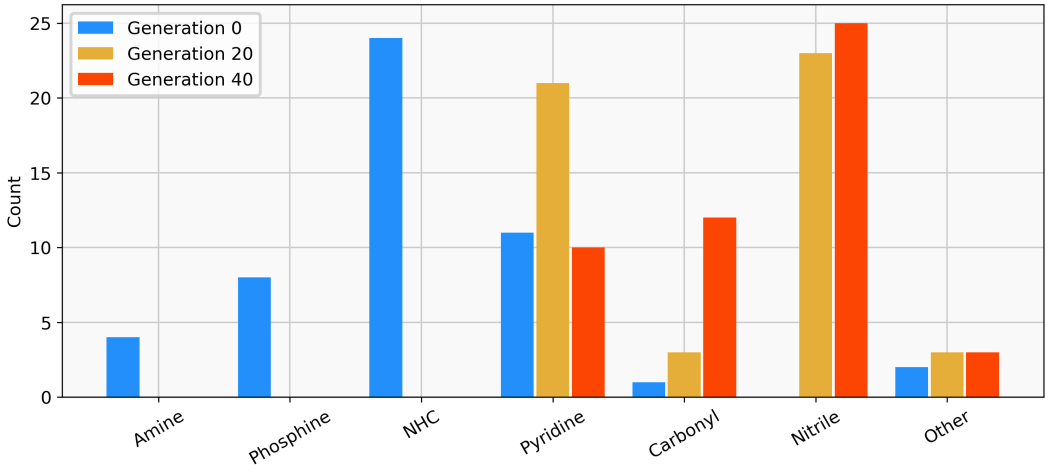



**Figure 8.** Distribution of the functional groups via which the ligands bind to the transition metal from a GBGA run with synthetic accessibility constraint for the initial population (blue), after 20 generations (yellow) and for the final population (red).

creased score of 0.1. After 30 generations, molecules with energy differences close to the target and near-perfect synthetic accessibility score are located.

The best-performing individuals from the final population are shown in Figure 7. All molecules possess one ligand coordinating via a nitrile group to the transition metal and another one coordinating via a carbonyl group or the nitrogen atom of a pyridine derivative. All six unique ligands are purchasable building blocks via Sigma-Aldrich and/or Mcule.[44] Figure 8 shows a bar chart of the different coordination sites of the discovered catalysts at different evolutionary steps. In the initial population of the GA, which was selected to have low $\Delta$E values, N-heterocyclic carbenes (NHCs) are the most common coordination sites followed by pyridine derivatives and phosphines. Neither NHCs nor phosphines are found as binding sites after 20 generations of the GA. This could be partially due to a low modified SA score for phosphines and NHCs. Instead more pyridine derivatives are found and nitriles are discovered as a favourable coordination site. After 40 generations, even more ligands coordinating via nitrile groups are found while the number of ligands containing pyridine derivatives as binding sites decreases. Instead, ligands that coordinate via a carbonyl group are preferred. This shows, that the GA actually traverses chemical space since the final population contains mainly coordination sites that are not present in the starting population and not just interpolated between chemical structures present in the starting population.

Schilter et al. [45] developed a variational-autoencoder trained on the dataset from Meyer et al. [7] and were able to discover novel catalysts with favourable energy differences by optimizing in a learned latent space. They show that the distribution of coordination sites for the generated molecules follows the distribution of the training data. This indicates, that they find novel ligands by interpolating in the latent space, but do not discover novel binding motifs in a different area of chemical space than what their training data contains.

## 3.4. Graph-based GA for Cu- and Ag-based Catalysts

With a GBGA, it is straightforward to discover novel catalysts utilizing other transition metals than Pd. Here, we show the generation of novel ligands for Cu- and Ag-based catalysts with a favourable thermodynamic profile for the Suzuki reaction. This appears to be a challenging task since Meyer et al. [7] were only able to find 20 and 0 catalysts in the desired energy range via screening of 18062 catalyst candidates, respectively. Furthermore, when calculating the actual energy difference at the B3LYP-D3BJ/def2-TZVP//B3LYP-D3BJ/3-21 level of theory, we could only confirm 6 out of 20 Cu-based catalysts with $\Delta$E values within -32.1 and $-23.0\,\mathrm{kcal\,mol^{-1}}$. The evolution of the score, calculated energy difference and the modified SA score of the best Cu-based catalysts over 40 generations are shown in Figure 9. The calculated $\Delta$Es of the best catalysts in the early generations are considerably higher ($>15\,\mathrm{kcal\,mol^{-1}}$) than the desired target range which yields low overall scores in the first six generations. After seven generations, a catalyst with a calculated $\Delta$E in the target range is identified. In the following 33 generations more catalysts within the target range and varying modified SA scores are discovered by tradeoff between the two objec-

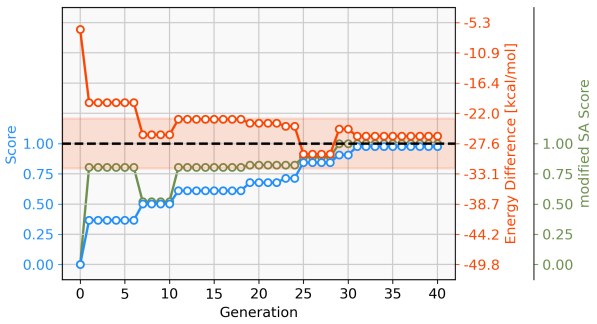



**Figure 9.** Evolution of the score (blue), energy difference (red) and modified SA score (green) of the best-performing individual of a GBGA run with Cu-containing catalysts over 40 generations. The target range of the energy difference is shown as a red-shaded area.

tives. Overall, 112 unique catalysts within the target range were discovered by evaluating 1000 catalysts with DFT, a subset is shown in Figure 10.

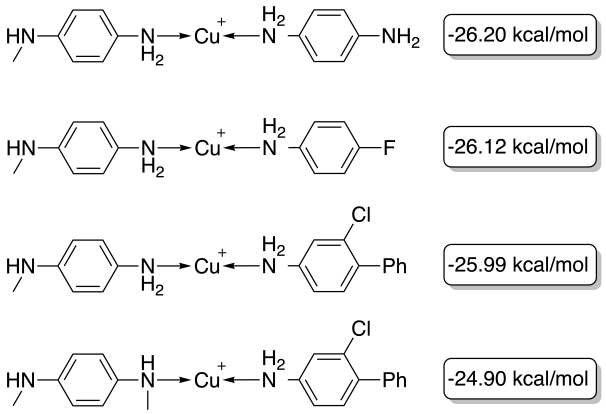



**Figure 10.** Best-performing individuals of the GBGA with synthetic accessibility constraint for Cu-based catalysts after 40 generations. The calculated energy difference is shown to the right of the structure.

For the generation of Ag-based catalysts, no synthetic accessibility constraint was applied since preliminary experiments proved the generation of catalysts with both high SA scores and ΔE values in the target range too challenging. We therefore show how the GBGA is used to generate structural motifs that yield catalysts within the desired ΔE range. Analysis of the generated structures yields insight into electronic and steric effets that would be necessary for Ag-based catalysts.

68 catalysts with ΔE values in the target range could be identified over 40 generations. Figure 11 shows the evolution of the score (blue) and the calculated ΔE (red) of the best catalyst over 40 generations. This task appears to

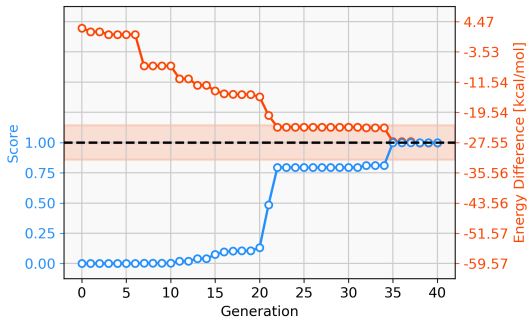



**Figure 11.** Evolution of the score (blue) and the energy difference (red) of the best-performing individual of a GBGA run with Ag-containing catalysts over 40 generations. The target range of the energy difference is shown as a red-shaded area.

be more challenging than previous ones, since it takes 22 generations without SA constraint until a catalyst with ΔE in the target range is discovered. The final catalyst candidates are all anionic which appears to stabilize the reactant in intermediate **2**. Figure 12 shows the structure of the best-performing catalyst candidate with the interaction between the ligand and reactant in blue. This can be seen as an example of what the structure of the ligand would need to look like in order for the catalyst to fall within the target energy range. The unusual structure with a deprotonated

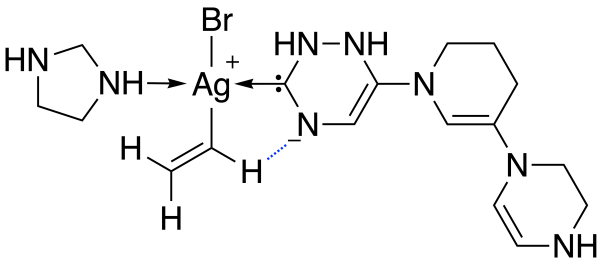


**Figure 12.** Lewis structure of the best-performing catalyst candidate from GA run containing Silver. The favourable interaction between one ligand and a reactant is shown as a blue dashed line.

NHC is not expected to be a stable complex and could most likely not be synthesised. Yet, the structure could be useful for further optimization while considering the need for non-covalent interactions between the ligand and the reactant.

## 4. Conclusion

In conclusion, the results of our study demonstrate the effectiveness of fragment-based genetic algorithms (FBGAs) and, especially, graph-based genetic algorithms (GBGAs) in the search for novel ligands for catalysts in the Suzuki reaction. The FBGA successfully evolved a population of molecules over 10 generations, yielding 134 unique catalysts with ΔE values within the target range. Our study demonstrates that GAs, requiring on the order of 500 evaluations or less, are effective in directly identifying competitive catalysts using DFT without the need for constructing a Machine Learning (ML) model.

The GBGA, without synthetic accessibility (SA) constraints, explored chemical space by iteratively applying crossover and mutation operations on ligand molecular graphs. The resulting ligands exhibited a diverse range of coordination sites, emphasizing the capability of GAs to discover novel structures. With SA constraints incorporated in the latter part of the study, the GA navigated towards ligands with improved synthetic accessibility while maintaining a ΔE within the target range. The final population of ligands, identified after 40 generations, showcased diverse binding motifs and confirmed the GA's ability to discover ligands with desirable properties for catalysis.

Furthermore, the application of GBGAs to explore ligands for Cu- and Ag-based catalysts in the Suzuki reaction revealed their potential to generate novel structures for different transition metals. Despite the challenges associated with the limited number of previously identified Cu-based catalysts, the GA successfully discovered 112 unique Cu-based catalysts within the desired ΔE range, demonstrating the versatility of the GA approach for exploring novel catalytic systems.

Generation of Ag-based catalysts with favourable thermondynamic profiles showed that the GBGA can discover structural motifs that yield catalysts within the target ΔE range without regard for stability or synthsizability. These structural motifs can yield insights for molecular discovery of further ligands.

Comparisons with existing machine learning (ML) models highlight the complementarity of GA methods, as GAs traverse chemical space, discovering ligands with coordination sites not present in the initial population. This is in contrast to ML models that interpolate within a learned la-

tent space but may struggle to explore entirely new binding motifs.

As with all generative models, the real-life performance of structures generated by this molecular optimization approach is limited by the applicability of the used scoring function or property prediction model. Here, we optimize the DFT-calculated energy of one specific reaction step in one well-defined catalytic cycle. To assess the real-world performance of a generated catalyst, an extensive computational strategy should be applied to assess the underlying assumptions.

The proposed catalyst structure should first be evaluated through an extensive conformational search covering all possible coordination modes and sites, potentially utilizing CREST for this purpose [46]. Subsequently, the catalyst's stability can be assessed using methods like the (meta-)dynamics approach outlined by Grimme et al. [47]. Also, the dominant oxidation state of the metal center needs to be considered. Next, the applicability of the linear energy scaling relation must be validated for the generated molecules by calculating all reaction intermediates. Additionally, it is crucial to locate the transition states throughout the catalytic cycle and compute their associated activation barriers, thereby avoiding reliance on linear correlations between intermediate energies and actual activation barriers. Finally, the entire reaction network of the reaction system should be investigated to identify any competing side reactions, similar to the work of Rasmussen et al. [48].

In summary, our study underscores the potential of genetic algorithms as powerful tools for ligand discovery in catalysis, showcasing their ability to efficiently navigate chemical space, discover novel structures, and generate ligands with desirable properties while minimizing computational costs.

Finally, the employed synthetic accessibility (SA) score in this study is observed to impose penalties on frequently utilized ligands, including phosphines and carbenes, redirecting the exploration towards drug-like chemical space. To achieve a more comprehensive exploration of the relevant chemical space, ongoing research in our group is dedicated to developing synthetic accessibility measures tailored for homogeneous inorganic catalysts.

## 5. Data availability

All code and data can be found at github.com/jensengroup/tmcat-design.

## References


(1) Miyaura, N.; Yamada, K.; Suzuki, A. *Tetrahedron letters* **1979**, *20*, 3437–3440.

(2) Miyaura, N.; Suzuki, A. *Chemical reviews* **1995**, *95*, 2457–2483.

(3) Suzuki, A. *Pure and applied chemistry* **1991**, *63*, 419–422.

(4) Porte, A. M.; Reibenspies, J.; Burgess, K. *Journal of the American Chemical Society* **1998**, *120*, 9180–9187.

(5) Sigman, M. S.; Vachal, P.; Jacobsen, E. N. *Angewandte Chemie (International ed. in English)* **2000**, *39*, 1279–1281.

(6) Renom-Carrasco, M.; Lefort, L. *Chemical Society reviews* **2018**, *47*, 5038–5060.

(7) Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. *Chemical science* **2018**, *9*, 7069–7077.

(8) Fu, R.; Nielsen, R. J.; Goddard III, W. A.; Fortman, G. C.; Gunnoe, T. B. *ACS catalysis* **2014**, *4*, 4455–4465.

(9) Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; Wiest, O.; Norrby, P.-O. *Nature catalysis* **2018**, *2*, 41–45.

(10) Nandy, A.; Duan, C.; Goffinet, C.; Kulik, H. J. *JACS Au* **2022**, *2*, 1200–1213.

(11) Nandy, A.; Duan, C.; Taylor, M. G.; Liu, F.; Steeves, A. H.; Kulik, H. J. *Chemical reviews* **2021**, *121*, 9927–10000.

(12) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, DOI: 10.1126/science.aau5631.

(13) Chu, Y.; Heyndrickx, W.; Occhipinti, G.; Jensen, V. R.; Alsberg, B. K. *Journal of the American Chemical Society* **2012**, *134*, 8885–8895.

(14) Foscato, M.; Venkatraman, V.; Jensen, V. R. *Journal of chemical information and modeling* **2019**, *59*, 4077–4082.

(15) Laplaza, R.; Gallarati, S.; Corminboeuf, C. *Chemistry-Methods* **2022**, *2*, DOI: 10.1002/cmtd.202100107.

(16) Kneiding, H.; Nova, A.; Balcells, D. *ChemRxiv* **2023**, DOI: 10.26434/chemrxiv-2023-k3tf2-v2.

(17) Gallarati, S.; van Gerwen, P.; Laplaza, R.; Brey, L.; Makaveev, A.; Corminboeuf, C. *Chemical science (Royal Society of Chemistry: 2010)* **2024**, DOI: 10.1039/d3sc06208b.

(18) Foscato, M.; Occhipinti, G.; Hopen Eliasson, S. H.; Jensen, V. R. *Journal of chemical information and modeling* **2024**, *64*, 412–424.

(19) Seumer, J.; Kirschner Solberg Hansen, J.; Brøndsted Nielsen, M.; Jensen, J. H. *Angewandte Chemie* **2023**, e202218565.

(20) Strandgaard, M.; Seumer, J.; Benediktsson, B.; Bhowmik, A.; Vegge, T.; Jensen, J. H. *PeerJ physical chemistry* **2023**, *5*, e30.

(21) Chu, Y.; Heyndrickx, W.; Occhipinti, G.; Jensen, V. R.; Alsberg, B. K. *Journal of the American Chemical Society* **2012**, *134*, 8885–8895.

(22) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. *ACS catalysis* **2019**, *9*, 2313–2323.

(23) Jensen, J. H. *Chemical science* **2019**, *10*, 3567–3572.

(24) Bannwarth, C.; Ehlert, S.; Grimme, S. *Journal of chemical theory and computation* **2019**, *15*, 1652–1671.

(25) Busch, M.; Wodrich, M. D.; Corminboeuf, C. *Chemical science* **2015**, *6*, 6754–6761.

(26) Wodrich, M. D.; Busch, M.; Corminboeuf, C. *Chemical science* **2016**, *7*, 5723–5735.

(27) Becke, A. D. *The Journal of chemical physics* **1993**, *98*, 5648–5652.

(28) Lee, C.; Yang, W.; Parr, R. G. *Physical review. B, Condensed matter* **1988**, *37*, 785–789.

(29) Vosko, S. H.; Wilk, L.; Nusair, M. *Canadian journal of physics* **1980**, *58*, 1200–1211.

(30) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *The journal of physical chemistry* **1994**, *98*, 11623–11627.

(31) Becke, A. D.; Johnson, E. R. *The Journal of chemical physics* **2005**, *123*, 154101.

(32) Grimme, S.; Ehrlich, S.; Goerigk, L. *Journal of computational chemistry* **2011**, *32*, 1456–1465.

(33) Binkley, J. S.; Pople, J. A.; Hehre, W. J. *J. Am. Chem. Soc.* **1980**, *102*, 939–947.

(34) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.

(35) Riniker, S.; Landrum, G. A. *Journal of chemical information and modeling* **2015**, *55*, 2562–2574.

(36) Landrum, G. et al. rdkit/rdkit: 2023_03_2 (Q1 2023) Release, 2023.

(37) Neese, F. *Wiley interdisciplinary reviews. Computational molecular science* **2022**, *12*, DOI: 10.1002/wcms.1606.

(38) Ertl, P.; Schuffenhauer, A. *Journal of cheminformatics* **2009**, *1*, 8.

(39) Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. *Journal of chemical information and modeling* **2019**, *59*, 1096–1108.

(40) Gao, W.; Coley, C. W. *Journal of chemical information and modeling* **2020**, *60*, 5714–5723.

(41) Baker, J. E. In *Proceedings of the 1st International Conference on Genetic Algorithms*, 1985.

(42) Availability of 2-Fluoripyridine, https://app.postera.ai/molecule-set/3c65a0f6-cbfe-4b0a-aaad-e88d148400c1.

(43) PostEra, Medicinal Chemistry Powered by Machine Learning, en, https://postera.ai/manifold/, Accessed: 2023-7-18.

(44) Availability of Generated Ligands (Figure 7), https://app.postera.ai/molecule-set/77612eb9-15e6-425d-9a48-6c22383287b7.

(45) Schilter, O.; Vaucher, A.; Schwaller, P.; Laino, T. *Digital discovery* **2023**, *2*, 728–735.

(46) Pracht, P.; Bohle, F.; Grimme, S. *Physical chemistry chemical physics: PCCP* **2020**, *22*, 7169–7192.

(47) Grimme, S. *Journal of chemical theory and computation* **2019**, *15*, 2847–2862.

(48) Rasmussen, M. H.; Seumer, J.; Jensen, J. H. *Preprint* **2023**, DOI: 10.26434/chemrxiv-2023-lchmv.

# Supplementary Material

## S1. Modifications to RDKit

All catalysts studied here are represented in RDKit as a molecule with a central transition metal atom to which two ligands bind with a dative bond each. The hybridization of the atoms from which the dative bonds start is determined incorrectly in RDKit $\leq$ 2023.03.2. The hybridization of an atom is determined by its number of bonds, lonepairs and radicals. When a dative bond starts from an atom, the sum of bonds, lonepairs and radicals is calculated wrong since the lonepair/radicals are the same electrons that are part of the dative bond. To account for this overcounting, we reduce the number of lonepairs by one or the number of radicals by two if a dative bond starts from an atom which has lonepairs or radicals, respectively.

The hybridization of an atom is used in the embedding process to determine ideal angles between atoms. The angle between atoms determines what lower and upper distance bounds are set in the bounds matrix Therefore, a wrong higher hybridization (SP3D instead of SP2) can lead to too small ideal angles (109.5 ° instead of 120.0 °) and to too large distance bounds. In the default implementation of RDKit, the atoms from which the dative bond starts are deemed to be higher hybridized than they are, for example, the carbon atom of a carbene which forms a dative bond to a transition metal is deemed to be SP3D hybridized, whereas one would consider the atom SP2 hybridized. This results in distorted geometries, especially around atoms such as the carbon atom of a carbene or a nitrogen atom of a pyridine which should be considered SP2 hybridized. The distortions are considerable such that an optimization using GFN2-xTB often does not yield an undistorted conformer and/or yields a conformer with a different bonding pattern than expected.

# 4

# DISCUSSION AND OUTLOOK

In this chapter, we showed how GAs can be used effectively to optimize homogeneous catalysts. We combined the optimization with a reaction network exploration approach in Paper 4, which constitutes an end-to-end workflow that can generate novel catalysts for a specific reaction without extensive prior knowledge of the reaction mechanism. As Corin Wagen notes in his blog post:

> It's difficult to underscore how groundbreaking this result is; as the authors dryly note, "We believe this is the first experimentally verified de novo discovery of an efficient catalyst using a generative model." On the spectrum discussed above, this is getting pretty close to "oracle."
>
> - Corin Wagen
> Blog post from 21.02.2023

An "oracle" in this context refers to a hypothetical computational system capable of answering key questions like:

- "What is a good catalyst for this reaction?"
- "What are the best reaction conditions for this reaction?"
- "How can I increase the yield for this reaction?"

Such an oracle would not only compute specific quantities but also generate innovative ideas and solutions. While we are far from the point of developing a general-purpose oracle with such capabilities, progressing towards this goal requires several critical elements.

Firstly, the accuracy of our computational models depends on high-quality experimental data to validate our approaches. This includes kinetic studies on various catalysts, substrates, and reaction conditions to ensure that our calculated catalytic activity measure aligns with experimental outcomes.

Secondly, we need accurate, robust and fast QM methods for calculating relative energies. While the SQM-based xTB methods have proven accurate for closed-shell organic molecules in the MBH reaction, their efficacy for TM-based complexes has been less reliable. For example, with square planar $Pd^{2+}$-complexes, we have observed discrepancies in ligand coordination preferences compared to those predicted at the DFT level.

DFT itself may also fall short in accurately modelling open-shell systems and those with multi-reference characters. These challenges

might be tackled using (Δ)-ML approaches, although their application is limited to specific domains.[31] Additionally, modelling solvent effects using implicit models may prove inadequate, particularly for reactions involving zwitterionic intermediates.

While our developed workflow has been successfully applied to a specific reaction, ongoing research aims to expand on this foundation. Other research groups are exploring applications such as optimizing the structure of molecular tweezers for sucrose detection or enhancing (ballistic) conduction through molecular junctions based on our provided codebase.[32, 33] Moreover, our own group is leveraging the developed code to propose catalysts for the conversion of nitrogen to ammonia and for carbon capture, both critical areas for a sustainable and green future.[34–36]

It is our hope that the work presented here will inspire further research using these developed workflows to innovate and solve various chemical challenges.

Part II

# AUTOMATED REGIOSELECTIVITY PREDICTION

# 5

# INTRODUCTION

As we have shown in Chapter 3, the computational optimization of a catalyst's activity is a challenging, yet achievable task. In this chapter, we investigate the selectivity of a catalytic system, which is another relevant property. A highly selective reaction yields only one relevant product which increases atom economy and decreases the need for extensive purification steps afterwards.

Some of the most selective reactions known to us can be found in nature. Enzymes are highly efficient and selective catalysts, achieving their remarkable selectivity through molecular recognition. In this process, a functional group of the substrate interacts with the enzyme, aligning the enzyme's active site with a specific location on the substrate where the reaction subsequently occurs. These interactions involve both attractive forces, such as hydrogen bonds, aromatic stacking, and ion pairing, and repulsive forces, including steric hindrance and specific shape matching.

Among synthetic catalysts, those based on transition metals (TMs) are often the most efficient and widely used. Their selectivity is typically adjusted by modifying their ligands; for instance, bulkier ligands tend to favour reactions at less sterically hindered sites on the substrate. However, this approach does not achieve the high level of selectivity seen in enzymatic reactions.

Particularly in C–H activation and functionalization reactions, the use of directing groups (DGs) has proven highly effective in controlling regioselectivity. Similar to molecular recognition, a functional group on the substrate coordinates to the catalyst, positioning it near the intended reaction site. Traditionally, the DG is located close to the reaction site, but several studies have explored the activation of remote sites using DGs, as demonstrated in works by Achar et al. [37] and others.[38, 39] This cutting-edge research holds the potential to enable chemists to design increasingly selective catalyst/substrate systems, enhancing atom economy and reducing chemical waste.

# 6

# PREDICTING REGIOSELECITIVITY

Predicting the regioselectivity of chemical reactions is a difficult but equally important task. Accurate predictions of where chemical transformations occur are essential for efficiently designing novel and complex molecules such as pharmaceuticals and agrochemicals. Several reactivity trends are known for various reactions; for example, the formation of a stable radical on an $SP^3$ carbon atom follows the reactivity trend tertiary > secondary > primary, which can be used to assess regioselectivity. While these rules are useful, their application to larger and more complex molecules is difficult. Here, computational models can give insight into the reactivity trends of specific substrates. A plethora of different computational approaches focusing on specific reaction mechanisms are available. Previously, Kromann et al. [40] developed RegioSQM, which predicts the regioselectivity of electrophilic aromatic substitution reactions. Ree, Göller, and Jensen [41] developed an automated workflow to predict the regioselectivity of the Heck reaction. The regioselectivity of nucleophilic aromatic substitutions can be predicted following a workflow by Liljenberg et al. [42].

In this context, we introduce an automated workflow to predict the regioselectivity of directed C−H activations via the concerted metalation deprotonation (CMD) reaction mechanism. This provides a structured approach to predicting where reactions will occur, which can enhance the synthesis planning process.

Previous works by Tomberg et al. [43] and Cao et al. [44] to predict the regioselectivity of this reaction either lack generalizability or require a high-performance computing (HPC) cluster to obtain predictions due to the high computational cost. Here, we aim to present a user-friendly, efficient and accurate semiempirical quantum mechanical (SQM) workflow that yields prediction within seconds to minutes on laptops. The use of DGs for the selective C−H activation and functionalization has seen a significant increase in popularity over the last decades, as shown in Figure 6.1. Increased understanding of the mechanism has helped to extend the reaction's scope to more substrates, DGs and remote activation sites. We hope that the here presented fast and user-friendly workflow for regioselectivity prediction will be useful to chemists with and without computational expertise.

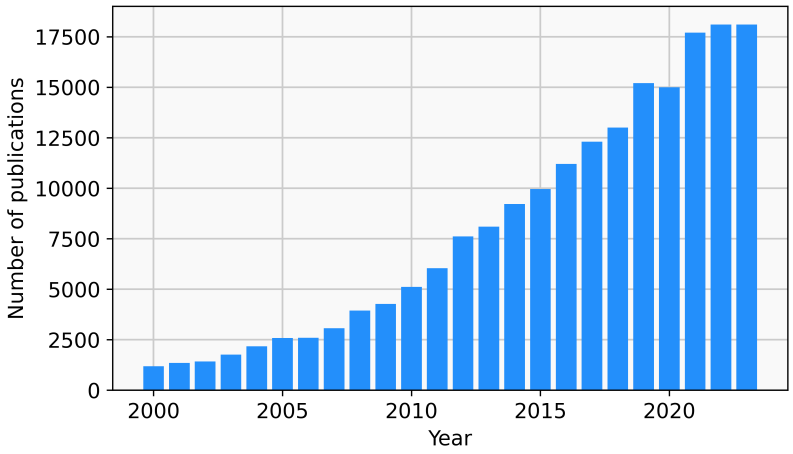



Figure 6.1: Number of scientific publications per year found on Google Scholar with the search term "directed ch functionalization activation"

## 6.1 C—H ACTIVATION VIA CONCERTED METALLATION DEPROTONATION

There are several possible and potentially competing reaction mechanisms for C—H activation, influenced by factors such as the nature of the metal, the steric and electronic effects of the ligands, and the acidity of the C—H bond. When DGs are present in the substrate and late TM catalysts with chelating bases are utilized, the CMD mechanism becomes viable, often yielding high regioselectivity. Lafrance et al. [45] and Gorelsky, Lapointe, and Fagnou [46] suggested the CMD mechanism as the most plausible one for the C—H activation using Pd-based catalysts with chelating bases. Here, several moieties work together to form the C—Pd bond in a concerted mechanism, as shown in Figure 6.2. The DG coordinates via a heteroatom to the TM centre of

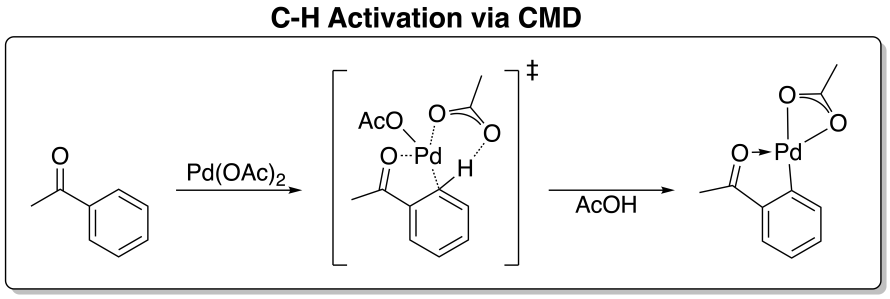



Figure 6.2: Reaction mechanism for the formation of a carbon-TM bond via the CMD mechanism

the catalyst, which brings the catalyst close to a specific C—H bond. The TM atom coordinates to the carbon atom of the C—H bond, which increases the acidity of the proton. The acetate moiety of the catalyst then abstracts the proton to form acetic acid, and the C—Pd bond is

formed. Upon C-H activation, a plethora of chemical transformations is possible to form new C-C and C-heteroatom bonds.

## 6.2 THIS WORK

Here, we present a fully automated workflow to predict the regioselectivity of $Pd(OAc)_2$ catalyzed C—H activations involving directing groups. A simplified molecular-input line-entry system (SMILES) representation of the substrate must be provided as input. The workflow identifies all possible reaction sites, creates molecular graph representations of the relevant catalyst-substrate complexes, generates several conformers and performs SQM calculations to identify the most likely reaction site. The workflow is accessible via an interactive command line interface or a web-based GUI in which the substrate can be drawn as a Lewis structure and predictions can run within seconds to minutes on laptops. Furthermore, the workflow can be deployed on a HPC cluster and the user can interact with it via an API to allow for even faster prediction for several substrates in parallel.

We evaluate the predictive workflow's performance on datasets curated from experimental data and achieve an accuracy of $\sim 80\%$. While we first only consider *ortho*-activation, we later extend the workflow to also consider activation on remote sites in the substrate.

# Enhancing Chemical Synthesis Planning: Automated Quantum Mechanics-Based Regioselectivity Prediction for C−H Activation with Directing Groups

Julius Seumer & Jan H. Jensen

May 20, 2024

The mild and selective functionalization of carbon-hydrogen (C−H) bonds remains a pivotal challenge in organic synthesis, crucial for developing complex molecular architectures in pharmaceuticals, polymers, and agrochemicals. Despite advancements in directing group (DG) methodologies and computational approaches, predicting accurate regioselectivity in C−H activation poses significant difficulties due to the diversity and complexity of organic compounds. This study introduces a novel quantum mechanics-based computational workflow tailored for the regioselective prediction of C−H activation in the presence of directing groups. Utilizing (semi-empirical) quantum calculations hierarchically, the workflow efficiently predicts outcomes by considering concerted metallation deprotonation mechanisms mediated by common catalysts like $Pd(OAc)_2$. Our methodology not only identifies potential activation sites but also addresses the limitations of existing models by including a broader range of directing groups and reaction conditions while maintaining moderate computational cost. Validation against a comprehensive dataset reveals that the workflow achieves high accuracy, significantly surpassing traditional models in both speed and predictive capability. This development promises substantial advancements in the design of new synthetic routes, offering rapid and reliable regioselectivity predictions that are essential for accelerating innovation in material science and medicinal chemistry.

## 1. Introduction

The activation and functionalization of carbon-hydrogen (C−H) bonds represent a fundamental challenge in modern organic chemistry, particularly due to the inherent stability and prevalence of these bonds in organic molecules. These bonds, which typically exhibit bond energies ranging from 90 to 110 kcal/mol, constitute the majority of bonds in organic chemicals. Therefore, their selective functionalization is central for advancing the synthesis of complex molecules like pharmaceuticals, polymers, or agrochemicals.[1–3]

Advancements in organometallic catalysis have facilitated significant progress in this area through C−H activation, transforming these inert bonds into reactive carbon-transition metal (C-M) bonds. Subsequent transformations of these complexes enable the formation of an array of new functional groups, such as carbon-carbon and carbon-heteroatom bonds, underpinning a plethora of synthetic applications.

Nevertheless, the high prevalence of C–H bonds in organic compounds presents a substantial challenge in achieving site-specific functionalization. A principal strategy to circumvent this challenge leverages directing groups (DGs) within the substrate, which coordinate to the metal centre of the catalyst, thereby dictating the site of C−H activation. Common DGs include unsaturated heteroatoms and alkenyl groups, which have proven effective in guiding the regioselectivity of these reactions.[4]

Mechanistic studies with $PdOAc_2$ as catalyst support the following mechanism of C−H activation, called Concerted metal deprotonation (CMD).[5–7] In a concerted mechanism, the Pd atom of the catalyst forms a sigma bond to an aromatic carbon, which increases the acidity of the adjacent proton. This allows for the simultaneous abstraction of this proton by a carboxylate ligand. A directing group facilitates this step as it stabilizes the complex through coordination to the Pd atom, thereby lowering the reaction barrier. A depiction

of the CMD step is shown in Figure 1. Upon C−H bond breaking, the Pd atom moves into the plane of the aromatic ring, forming a palladacycle intermediate and carboxylic acid. The palladacycle intermediate can undergo further (coupling) reactions and form a variety of products via reductive elimination.

In a computational study, the rate and regioselectivity controlling step was identified as the formation of the palladacycle. The regioselectivity could be correctly predicted by calculation and comparison of the activation barrier of this step by Davies et al. [8]. The reaction site for which the activation barrier is the lowest is predicted to be the most probable one. Tomberg et al. [9] established that the regioselectivity could equally be predicted by calculation and comparison of the relative energies of the proceeding palladacycle intermediate, as postulated in the Bell–Evans–Polanyi principle.[10, 11] Focussing on the intermediates allows for easier automation of the calculations since a minimum instead of a saddle point structure on the potential energy surface is located which can be easiyly done using standard optimization algorithms. Tomberg et al. [9] introduced a hierarchy of directing strength for 238 different ortho DGs, which can be used to rapidly predict the regioselectivity of C−H activation in complex molecules. The 238 directing groups are extracted from 150 molecules, taken from Chen et al. [4], for which reaction sites are known from experiments. For each directing group, the energy of the palladaycle intermediate with H-abstraction at a specific site is calculated using DFT and compiled into a hierarchical list for the determination of the reaction site with the lowest energy. Using the hierarchy, the regioselectivity of C−H activations could be rationalized for the 150 molecules with remarkable accuracy. While this approach performs well on this dataset, it doesn't generalize well to other molecules that were not used to extract DGs and precompute their relative directing strength. This is evidenced by our analysis using a dataset curated from Reaxys, where a prediction for only two out of ten molecules could be obtained. This is due to the specificity of the patterns, which only matched all potential C−H activation sites for two molecules. This underscores the necessity for more robust and versatile predictive models that can adapt to the broad spectrum of organic chemistry's structural variability.

Cao et al. [12] developed an automated workflow that predicts the regioselectivity of C−H activations using extensive DFT calculations on a HPC-cluster using up to 600 nodes each containing 16 Intel Xeon E5-2670 cores. They considered two possible reaction mechanisms, an electrophilic aromatic substitution and a proton abstraction mechanism via concerted metallation deprotonation (CMD), where they calculated the relative energies of the intermediates. Using their workflow, they were able to predict the regioselectivity for 18 tested substrates correctly. The main limitation of this work is the computational cost and usability since several DFT calculations need to be run on an HPC cluster in order to make a prediction.

In this study, we introduce a quantum mechanics-based computational workflow specifically developed to predict regioselectivity in C−H functionalization reactions involving directing groups following the CMD mechanism. This workflow employs (semi-empirical) quantum calculations in a hierarchical way to predict regioselective outcomes, delivering results within seconds to minutes. For substrates that are expected to follow the electrophilic aromatic substitution mechanism, we refer the reader to the work done by Kromann et al. [13]. Using RegioSQM the regioselectivity of reactions following the electrophilic aromatic substitution mechanism can be predicted within seconds to minuts using a web-interface or a python module.[14]

Similarly to previous works, we focus on the CMD step, the first and commonly the rate-determining step in C−H activation, and consider the prototypical $Pd(OAc)_2$ catalyst. Using a selective approach, we calculate the relative energies of all relevant palladacycle intermediates at the QM level. We determine the relevant reaction sites either by a set of SMART patterns or by screening all possible reaction sites using the Merck molecular force field calculated ring strain energy, for details see subsection 3.3. This restriction allows us to to rapidly predict the of regioselectivity for C−H-activations via the CMD mechanism within seconds to minutes on standard consumer hardware. The workflow accommodates various directing groups (DGs) and reaction conditions and can be extended to include not only ortho-activations, as detailed in section 3.

This development holds the potential to significantly accelerate the discovery and optimization of new synthetic routes, thereby impacting material science and medicinal chemistry by facilitating the synthesis of novel compounds with high precision and efficiency.

## 2. Computational Methodology

The predictive QM-based model calculates which potential reaction site it is most likely to react based on its corresponding activation energy. The site with the lowest activation energy is expected to correspond to the experimentally observed reaction site. Instead of locating the structure of the transition state, the preceding palladacycle intermediate structure is generated and optimized, as shown in Figure 1. Following the Bell–Evans–Polanyi principle, the relative energy of the intermediate should correlate linearly with the energy

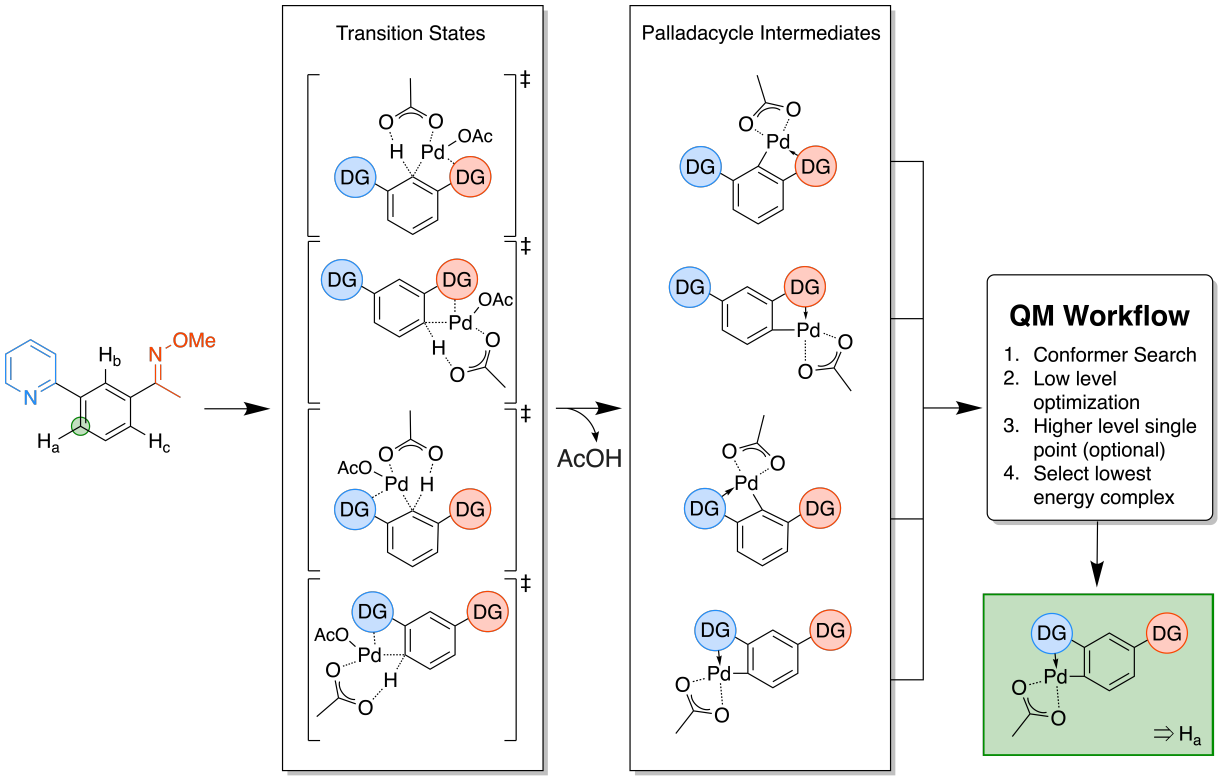



**Figure 1.** Overview of the predictive workflow: For the shown substrate on the left, three unique activation sites are possible (labeled $H_{a-c}$) with two directing groups, a pyridine (blue) and an oxime-ether (red). The latter has two potentially directing atoms, nitrogen and oxygen. The transition state structures of the rate-determining concerted metallation deprotonation (CMD) step are shown in the left column. In this work, we generate the structures of the proceeding palladacycle intermediate, shown in the right column. For each structure, we perform a conformer search followed by a low level optimization (GFN1-xTB) followed by an optional higher level single-point calculation (r2SCAN-3c). The lowest energy complex is selected and the corresponding reaction site is considered to be most likely to be activated.

of the transition state.[10, 11] Using this approximation the generation and optimization of structures simplifies greatly. In an automized workflow, all unique and possible combinations of C−H bonds and ortho-directing groups (heteroatom with lone pair) in the substrate are found following this procedure:

1. All combinations of C−H bonds from $sp^2$ hybridized C-atoms and directing groups (heteroatom with lone pair) in the substrate, which are between 2 and 5 bonds apart from each other, are detected with SMART patterns. These patterns are general enough to cover all directing groups that were encountered in the literature sample from Chen et al. [4]

2. Next, we identify all relevant palladacycle complexes for the C−H activation using ortho-directing groups. For each match, a complex with the substrate and Pd is formed, here the Pd atom is bonded to the carbon atom of the reaction site and the hetero atom of the directing group, as shown in Figure 2. A 2D embedding for the complex is generated with RDKit, here all atoms are within a plane, this embedding is usually used only for depictions. We measure the internal bond angles between bonds of the ring involving Pd, the heteroatom of the directing group, and the carbon atom of the reaction site in the 2D embedding. If any angle deviates more than 10% from the ideal 2D angle of a ring with $N_{\text{atoms}}$ atoms, the match is removed. The ideal angle is calculated as $\frac{(N_{\text{atoms}}-2)\cdot 180^\circ}{N_{\text{atoms}}}$. This allows us to filter out complexes with strained geometries, such as the one shown on the left in Figure 2, using a simple 2D approach.

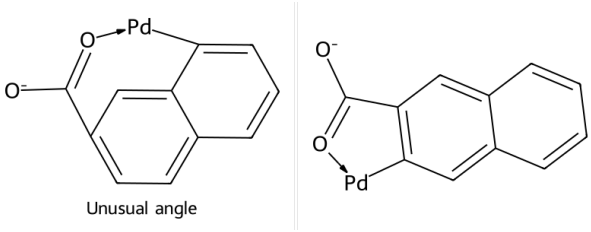



**Figure 2.** Example of a combination of C−H bond and DG that is discarded due to the angle constraint on the left and a combination that is considered valid on the right

3. Duplicate matches are removed when the reaction site is symmetric. Symmetry equivalent sites are determined by comparison of the canonical SMILES for the substrate with an explicit hydrogen atom added to the corresponding reaction site. When two SMILES with an added explicit hydrogen at different atom indices are identical, then the corresponding atoms are symmetry equivalent.

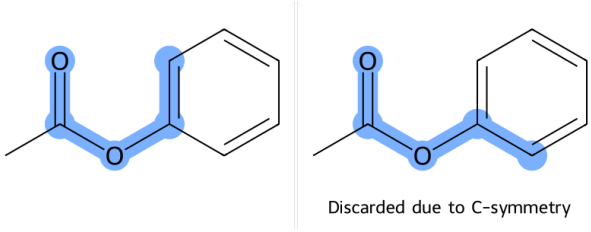



**Figure 3.** Example of combinations of C−H bonds and DGs that are considered identical due to symmetry of the C−H bond

4. Duplicates are removed when the directing group is symmetric. Again, symmetry equivalent atoms are determined by comparison of SMILES strings, here a bond to a dummy atom is added to the heteroatom of the directing group and the canonical SMILES representation is compared to all other SMILES with an added dummy atom.

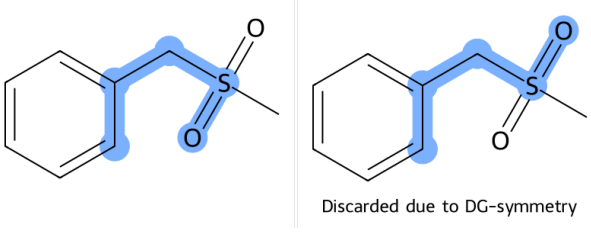



**Figure 4.** Example of combinations of C−H bonds and DGs that are considered identical due to symmetry of the DG

5. Duplicates are removed when the directing group has equivalent resonance forms, as shown in Figure 5. The equivalent heteroatoms are detected using SMARTS patterns for nitro- and carboxylate-groups.

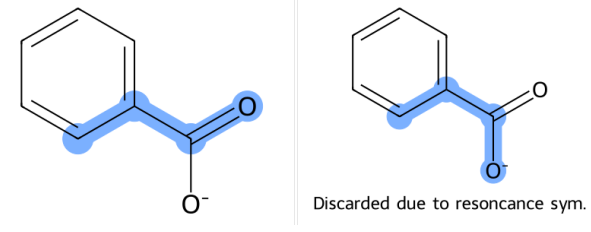



**Figure 5.** Example of combinations of C−H bonds and DGs that are considered identical due to resonance structures of the DG

For the remaining combinations of C−H bonds and directing groups, the corresponding intermediate substrate-Pd(OAc)-complex is generated. For each complex $3N_{\mathrm{rot}} + 3$ conformers are generated with ETKDG, here $N_{\mathrm{rot}}$ is the number of rotatable bonds in the substrate.[15, 16] The conformers are clustered based on their RMSD with a cutoff of 1.0 Å and the conformer corresponding to the centroid of each cluster is retained. The remaining conformers of each complex are optimized using GFN1-xTB in the implicit solvent model ALPB with parameters for $CH_2Cl_2$.[17, 18]

After each optimization, the geometry of the complex is analyzed to determine whether the connectivity has changed. The determination of connectivity to the transition metal of the complex is difficult to determine using either a radial distance cutoff or a cutoff on the overlap population of a Hueckel calculation, as shown in Figure 6.

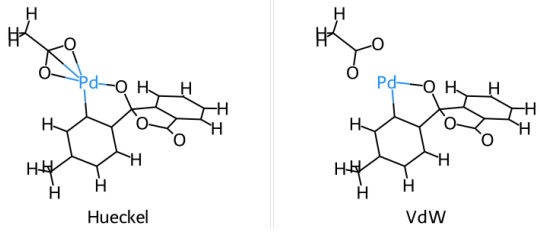



**Figure 6.** Example of challenges when determining the connectivity from a point cloud of atoms of transition metal complexes

Using the implementation of xyz2mol in RDKit with the flag useHueckel=True determines that both oxygen atoms of the acetate, as well as the adjacent carbon atom, are forming a bond towards the transition metal. In contrast, no bonds between the acetate moiety and the transition metal are found when using the radial distance cutoff, when one would expect the two oxygen atoms to be connected to the transition metal. Therefore, the connectivity of the complex before and after optimization is compared only for bonds not involving the transition metal. Instead, the geometry of the four atoms adjacent to the transition metal (the two oxygen atoms from the acetate moiety, the carbon atom from the reaction site, and the hetero atom from the directing group) is analyzed without regard for connectivity. All four atoms have to lie within a plane after the optimization for the optimization to be considered successful. This is determined by calculating the angle between the normal vectors of the plane spanned by Pd, the reaction site and the hetero atom of the directing group and the plane spanned by Pd and the two oxygen atoms of the acetate moiety. This angle has to be below 5° for the atoms to be considered to be within a plane.

Once all calculations for all conformers of all complexes are completed, the complex with the overall lowest energy conformer is selected and its corresponding reaction site is considered the most likely to react. All complexes that have conformers within a defined energy threshold of the overall lowest energy conformer are considered to correspond to potential reaction sites, for this study, we choose a threshold of 1 kcal/mol.

When several complexes with conformers within the energy threshold are found and the corresponding reaction site differs between the complexes, we allow the user to refine the prediction by running r2SCAN-3c single-point calculations on the lowest energy conformer of each complex within the energy threshold using ORCA.[19, 20] This allows us to refine the predicted binding sites at a higher level of theory when this is required.

## 3. Results

In the following, we tested our method on the dataset from Tomberg et al. [9] as well as on a new dataset that was curated from Reaxys.[21]

In the evaluation, we consider the three categories "correct", "semi-correct" and "incorrect". When the experimentally observed reaction site is the only reaction site that is predicted within the energy cutoff, the prediction is considered correct. When the experimentally observed reaction site is not the only reaction site that is predicted within the energy cutoff, the prediction is considered semi-correct, since we can't distinguish beyond what is considered the chemical accuracy. When the experimentally observed reaction site is not one of the predicted sites, the prediction is considered incorrect.

## 3.1. Dataset from Tomberg et al.

We consider 142 molecules with their experimentally determined reaction site from Tomberg et al. [9], which were originally curated by Chen et al. [4]. We are excluding cyclization reactions for which the regioselectivity is not only determined by the activation energy to form the palladacycle intermediate but also by which site is accessible for the intramolecular cyclization.

Using the previously described workflow, we were able to predict the experimentally observed reaction site with 78% accuracy over the whole dataset when using no energy threshold, meaning that only the reaction site corresponding to the lowest energy complex is predicted to be the reaction centre. In Figure 7A, the predictions, correct (green) or wrong (red), are shown as a stacked bar chart for molecules with different numbers of potential reaction sites. The expected number of correct predictions and the 95% confidence interval of a model that guesses one of the potential reaction sites is shown as a black cross.

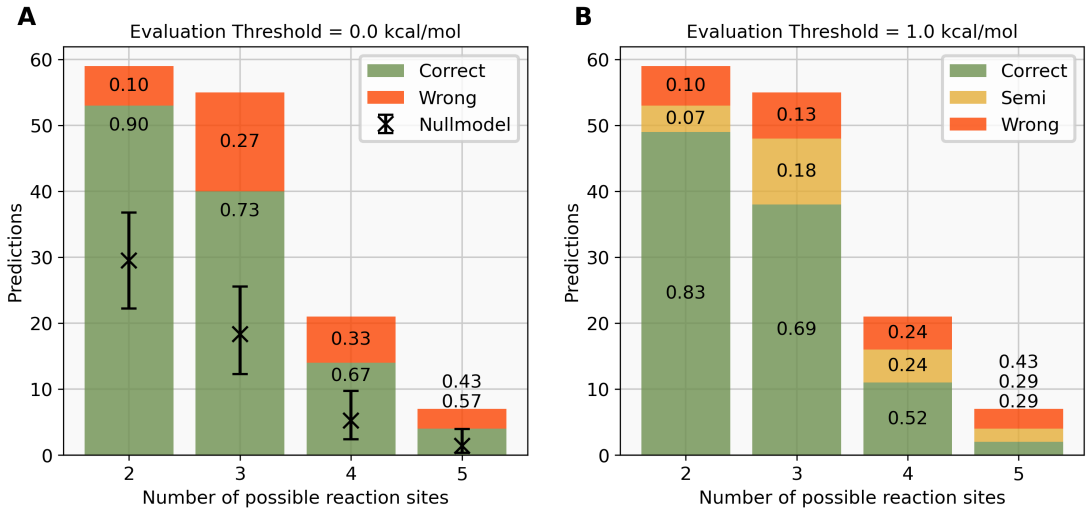



**Figure 7.** **A**: Distribution of correct (green) and wrong (red) predictions for molecules with two to five potential reaction sites, evaluated with an energy threshold of 0.0 kcal/mol. The numbers inside the bar plot correspond to the fraction of each label out of the total number of predictions. The expected performance of the null model with a 95% confidence interval is shown as a black cross. **B**: Distribution of correct (green), semi-correct (yellow), and wrong (red) predictions for the same molecules, evaluated with an energy threshold of 1.0 kcal/mol

For molecules with only two potential reaction sites, the null model is expected to correctly predict the reaction site for 30 out of 60 molecules. Our QM-based workflow can predict the correct reaction site for 54 out of 60 molecules with two potential reaction sites which corresponds to 90% correct predictions and lies outside of the confidence interval of the null model. Similarly, for molecules with three and four potential reaction sites, the QM workflow predicts between 73-67% of reaction sites correctly, when we would expect the null model to guess the correct reaction site with an accuracy of 33% and 25%. Notably, the QM workflow predicts the correct reaction site for only four out of seven molecules with five potential reaction sites, which corresponds to 57% accuracy. The three molecules with five potential reaction sites and wrong predictions are shown in Figure 8 with the experimentally observed reaction site in green and the predicted reaction site marked by a blue circle.

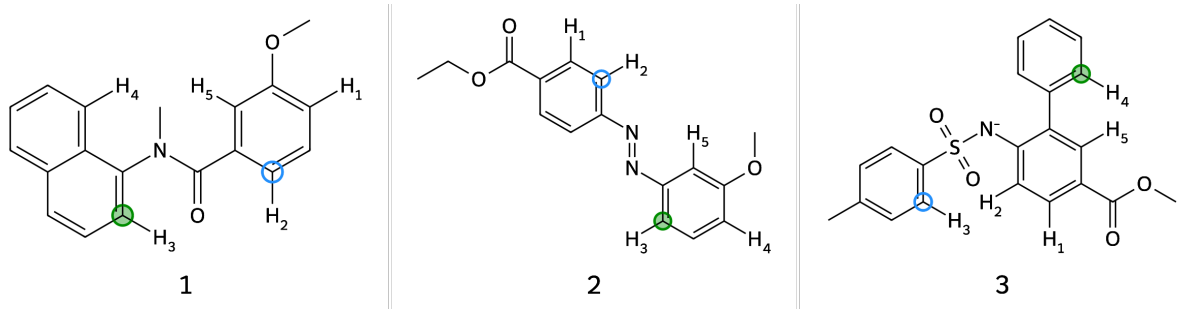



**Figure 8.** Molecules with five potential reaction sites that are predicted wrong by the QM workflow

For molecule **1**, we can see in the original paper from Yeung et al. [22] that the reaction proceeding the C−H activation is an intramolecular cyclization between the C-atom marked in green and the C-atom marked by a blue circle. This reaction was originally not marked as a cyclization reaction, which is why we did not remove it from the dataset. Nevertheless, upon inspection, our QM workflow correctly predicts the reaction site(s) of the intramolecular cyclization as it predicts one of the two reaction sites for the C−H activation.

The reaction site of molecule **2** from Dong et al. [23] can not be predicted correctly as the experimentally observed reaction site is 1.7 kcal/mol higher in energy than the predicted site at the r2SCAN-3c level. This would correspond to a ten times higher rate-constant of the reaction leading to the other regioisomer at the reaction temperature of 90°C. Experimentally, it is observed that the regioselective C−H activation happens on the more electron rich aromatic ring with the methoxy substituent as opposed to the one with the alkoxycarbonyl group. The wrong prediction here might indicate that the Bell–Evans–Polanyi relationship does not hold in this case and one would need to calculate the activation energy to the actual transition states.

Molecule **3** from Jiang et al. [24] is another intramolecular cyclization reaction which was not labelled as such. For such reactions, the regioselectivity is not only determined by the activation energy for the rate-determining step but also by the proximity of an intramolecular reaction partner, here the secondary amine.

From this in-depth analysis, we conclude that our QM workflow only predicted the wrong reaction site for one out of these three molecules investigated as the other "incorrect" predictions are due to a problem with the underlying dataset.

Since we don't assume that the energies obtained at the r2SCAN-3c(CPCM)//GFN1-xTB(ALPB) level are accurate enough to separate regioisomers which are close in energy, we consider all reaction sites that are within an empirically chosen threshold of 1 kcal/mol of the lowest energy reaction site as potential reaction sites. When more than one reaction site is within this threshold, we label the prediction as "semi-correct". Depending on the use case, the user might want to proceed with optimizing the structures of the relevant complexes at a higher level of theory or perform a transition state search to calculate the activation energy.

With this threshold, we obtain 70% correct, 14.5% semi-correct and 14.5% wrong predictions over the whole dataset. For 6 out of 17 molecules, all possible reaction sites are predicted as reaction sites within the threshold as shown in Figure S6. This means that these predictions do not yield any information, but for the other cases, the prediction rules out other potential reaction sites.

## 3.2. Dataset curated from Reaxys

From a query in Reaxys (see SI, subsection S1), we selected 10 C−H activation reactions with $Pd(OAc)_2$ as catalyst and multiple directing groups and/or symmetry unqeuivalent reaction sites. Using our QM workflow, we were able to predict the regioselectivity of 9 out of 10 molecules (semi-)correctly. Five reaction sites were predicted to be within 1 kcal/mol of another possible reaction site in the reactant and were therefore classified as semi-correct, meaning we can not predict with our model which of the two regioisomers will be the main product of the reaction. Refinement using r2SCAN-3c single-point calculations did not yield more accurate predictions.

The 10 molecules with their experimentally observed main reaction site in green and all predicted reaction sites within a 1 kcal/mol threshold are shown in Figure 9.

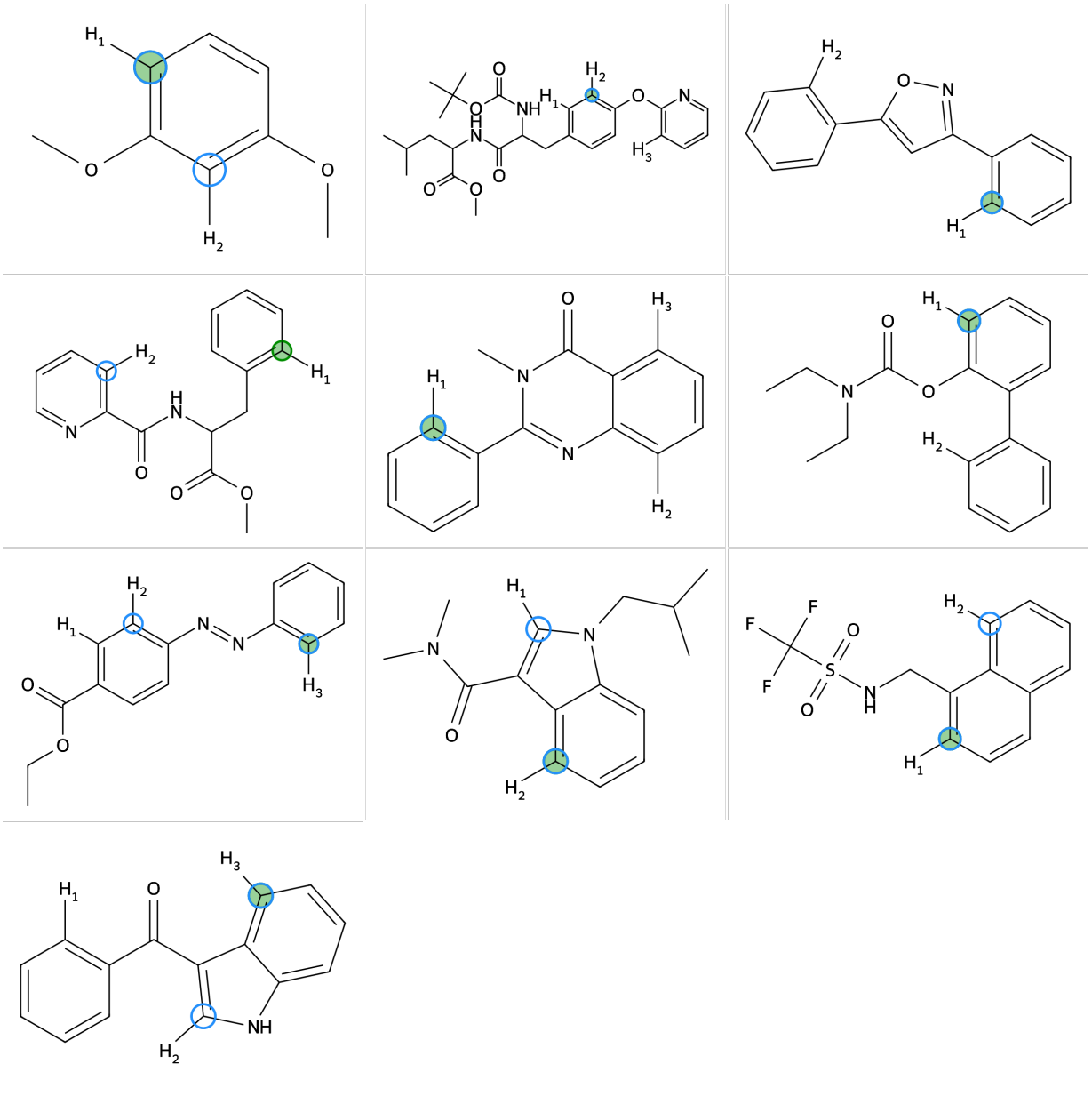


**Figure 9.** Predictions of reaction sites within a 1 kcal/mol threshold for 10 molecules are marked with a blue circle, and experimentally observed reaction sites are highlighted by a green circle.

## 3.3. Beyond ortho-directing groups

The presented workflow can be used to include the influence of all directing groups, not only ortho-directing groups. Here we extend the application of the workflow to a substrate with a meta-directing group. The substrate was investigated by Achar et al. [25] and the reaction site was determined by the authors to be the **$H_2$** with a meta:other regioselectivity of up to 25:1 and a yield up to 85%. In order to extend our approach to meta-/para- and remote-directing groups, we use a different approach to identify relevant palladacycle complexes as in points 1. and 2. outlined in section 2. Instead of using SMARTs patterns to detect pairs of ortho-directing groups and reaction sites, we detect all potential reaction sites by detecting all C−H bonds at $sp^2$ hybridized carbon atoms as well as all heteroatoms with lone pairs separately and remove symmetry equivalent sites. Then, we obtain all potential pairs of C−H bonds and heteroatoms as the Cartesian product of the two sets. Next, we filter out all pairs for which no reasonable 3D geometry can be generated. To determine whether or not a pair of C−H bond and heteroatom can form a reasonable 3D geometry, we generate a 3D geometry of a dummy

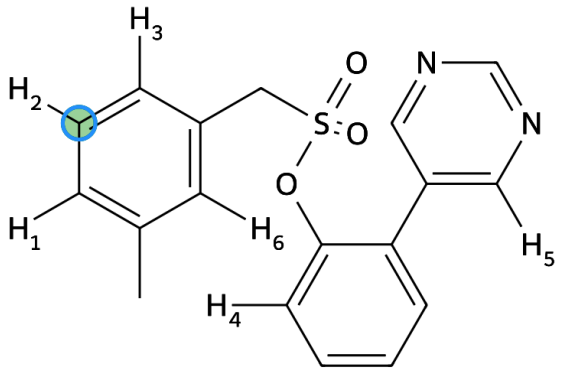



**Figure 10.** Substrate with six potential unique reaction sites for C−H functionalization. The experimentally determined reaction site is marked by a green circle.

"palladacycle"-intermediate between the substrate and a $CCl_2$ fragment using ETKDG. The $CCl_2$ fragment is used to mimic the $Pd(OAc)_2$ catalyst, which can't be used since the following step relies on the Merck molecular force field (MMFF, version MMFF94s), which is not parameterized for transition metals like Pd.[26, 27] If the embedding fails, the corresponding pair is removed. When a 3D geometry could be obtained, we optimized the structure using the MMFF94s. Next, we calculate the sum of (out-of-plane) angle terms and torsion terms of the MMFF94s forcefield for the optimized structure. The geometry is considered reasonable if the sum of the angle and torsion terms is below a threshold of 10 kcal/mol. From here on, we proceed with the workflow as described in section 2.

For the here considered substrate, this procedure reduces the number of complexes to optimize with GFN1-xTB from 30 to 9, the complexes are shown in Figure S1. This procedure involves several force-field optimizations, which increase the overall wall time by $\approx 10$ seconds for the here shown substrate compared to the previously reported approach. From here on, we follow the same procedure as for the ortho-directing groups and correctly predict the reaction site $\mathbf{H_2}$, which is the only one within the 1 kcal/mol energy threshold at the GFN1-xTB level.

## 4. Discussion

Our study demonstrates that the Quantum Mechanics (QM) workflow reliably predicts the reaction site as observed experimentally with 70% correct predictions and 14.5% semi-correct predictions on the dataset provided by Tomberg et al. [9]. Analysis of molecules where the reaction site was incorrectly predicted, particularly those with five potential sites, revealed that there might be issues with the underlying data in some cases. When only considering the lowest energy reaction site predicted by our workflow, we were able to achieve an accuracy of 78% on the same dataset. In contrast, a basic model making random guesses would achieve only 38% accuracy, within a 95% confidence interval from 36 to 40%, underscoring our workflow's superior performance.

Additionally, we applied the workflow to a new set of 10 molecules, achieving a 90% accuracy rate in predicting C−H activation sites. We also explored the tool's capability to predict regioselectivity in C−H activation with various directing groups, not limited to ortho-directing groups. By identifying potential reaction site-directing group pairs using an approach based on MMFF energies instead of simple SMARTS patterns, we illustrated the workflow's effectiveness with a case study from existing literature, accurately predicting the reaction site in a meta-directing C−H activation scenario.

In this study, we rely on several key assumptions that we will outline below. Firstly, we focus exclusively on the regioselective outcomes of reactions using the concerted metallation deprotonation (CMD) mechanism between the catalyst and the substrate. It is important to note that this approach does not allow us to predict the occurrence of the reaction, its yield, or confirm if the reaction might proceed via a different mechanism influenced by the substrate, catalyst, and ligands.

Secondly, we assume that the reaction is controlled kinetically, where the activation energy required to form the palladacycle intermediate determines the C−H activation regioselectivity. This assumption holds true primarily

when the reaction is irreversible, and the formation of the intermediate is the rate-limiting step. While previous studies support this assumption, it may not always apply universally across various substrates or catalysts.[8]

Thirdly, we consider the linear energy relationship between the intermediate and its preceding transition state as per the Bell–Evans–Polanyi principle. However, this relationship may not provide sufficient accuracy for making predictions when the energy difference between reaction sites is less than 1 kcal/mol. To enhance the reliability of our predictions, ideally, we would automate the process of locating transition state structures.

To enable rapid predictions of C−H activation sites regioselectivity, ranging from seconds to minutes on consumer hardware, we employ semi-empirical optimizations and, when necessary, DFT single-point calculations to reduce computational costs. In our analysis using the dataset from Tomberg et al. [9], we recorded median and mean prediction times of 2:02 minutes and 2:21 minutes, respectively, using four Intel Xeon E5-2643 v3 (3.4 GHz) CPUs. These times were significantly reduced to 22 and 34 seconds when exclusively using semi-empirical optimizations. The workflow benefits from parallelized QM programs and routines, demonstrating nearly linear reductions in wall time as the number of cores increases, tested up to 16 cores. However, for greater accuracy, particularly at reaction sites with energy differences less than 1 kcal/mol, DFT optimizations are recommended, as they may necessitate higher-level (re-)optimization for precise predictions.

The primary strength of the quantum mechanics (QM) workflow lies in its flexibility, which facilitates customization through various means, such as simulating different solvent effects or examining the impact of different catalysts and ligands, extending beyond $Pd(OAc)_2$. Additionally, variations reaction conditions, like conducting the reaction under acidic or basic environments, is possible by adjustments to the substrate-SMILES. Protonation states of substrates can be predicted using either machine learning models[28] or QM calculations[29].

This developed workflow is designed to be accessible not only to computational chemists but also to those without a computational background through multiple interfaces, including a command line interface, a web-based user interface, an API, and a stand-alone Python module for integration into more complex systems. For example, this workflow can be used in further molecular discovery and optimization to design specific directing groups that can facilitate the functionalization of remote C−H bonds, like meta- or para-functionalization. This can be done by using the workflow in the scoring function of a genetic algorithm, for example. Here, the absolute directing strength towards a specific site can be used to score different directing groups to each other and have the genetic algorithm design molecules that increase the directing strength of a directing group towards a specific site.

## References


(1) Arndtsen, B. A.; Bergman, R. G.; Mobley, T. A.; Peterson, T. H. *Accounts of chemical research* **1995**, *28*, 154–162.

(2) Halpern, J. *Discussions of the Faraday Society* **1968**, *46*, 7–19.

(3) Roudesly, F.; Oble, J.; Poli, G. *Journal of molecular catalysis. A, Chemical* **2017**, *426*, 275–296.

(4) Chen, Z.; Wang, B.; Zhang, J.; Yu, W.; Liu, Z.; Zhang, Y. *Organic chemistry frontiers: an international journal of organic chemistry* **2015**, *2*, 1107–1295.

(5) Gorelsky, S. I.; Lapointe, D.; Fagnou, K. *Journal of the American Chemical Society* **2008**, *130*, 10848–10849.

(6) Lapointe, D.; Fagnou, K. *Chemistry letters* **2010**, *39*, 1118–1126.

(7) Davies, D. L.; Donald, S. M. A.; Macgregor, S. A. *Journal of the American Chemical Society* **2005**, *127*, 13754–13755.

(8) Davies, D. L.; Macgregor, S. A.; McMullin, C. L. *Chemical reviews* **2017**, *117*, 8649–8709.

(9) Tomberg, A.; Muratore, M. É.; Johansson, M. J.; Terstiege, I.; Sköld, C.; Norrby, P.-O. *iScience* **2019**, *20*, 373–391.

(10) Bell, R. P. *Proceedings of the Royal Society of London* **1936**, *154*, 414–429.

(11) Evans, M. G.; Polanyi, M. *Transactions of the Faraday Society* **1936**, *32*, 1333.

(12) Cao, L.; Kabeshov, M.; Ley, S. V.; Lapkin, A. A. *Beilstein journal of organic chemistry* **2020**, *16*, 1465–1475.

(13) Kromann, J. C.; Jensen, J. H.; Kruszyk, M.; Jessing, M.; Jørgensen, M. *Chemical science (Royal Society of Chemistry: 2010)* **2018**, *9*, 660–665.

(14) RegioSQM, en, http://regiosqm.org/, Accessed: 2024-4-16.

(15) Riniker, S.; Landrum, G. A. *Journal of chemical information and modeling* **2015**, *55*, 2562–2574.

(16) Landrum, G. et al. rdkit/rdkit: 2023_03_2 (Q1 2023) Release, 2023.

(17) Grimme, S.; Bannwarth, C.; Shushkov, P. *Journal of chemical theory and computation* **2017**, *13*, 1989–2009.

(18) Ehlert, S.; Stahn, M.; Spicher, S.; Grimme, S. *Journal of chemical theory and computation* **2021**, *17*, 4250–4261.

(19) Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. *The Journal of chemical physics* **2021**, *154*, 064103.

(20) Neese, F. *Wiley interdisciplinary reviews. Computational molecular science* **2022**, *12*, DOI: 10.1002/wcms.1606.

(21) Reaxys, en, https://www.reaxys.com, Accessed: 2024-4-26.

(22) Yeung, C. S.; Zhao, X.; Borduas, N.; Dong, V. M. *Chem. Sci.* **2010**, *1*, 331–336.

(23) Dong, J.; Jin, B.; Sun, P. *Organic letters* **2014**, *16*, 4540–4542.

(24) Jiang, H.; Gao, H.; Liu, B.; Wu, W. *RSC advances* **2014**, *4*, 17222–17225.

(25) Achar, T. K.; Zhang, X.; Mondal, R.; Shanavas, M. S.; Maiti, S.; Maity, S.; Pal, N.; Paton, R. S.; Maiti, D. *Angewandte Chemie (International ed. in English)* **2019**, *58*, 10353–10360.

(26) Halgren, T. A. *Journal of Computational Chemistry* **1996**, *17*, 490–519.

(27) Halgren, T. A. *Journal of computational chemistry* **1999**, *20*, 720–729.

(28) Johnston, R. C.; Yao, K.; Kaplan, Z.; Chelliah, M.; Leswing, K.; Seekins, S.; Watts, S.; Calkins, D.; Chief Elk, J.; Jerome, S. V.; Repasky, M. P.; Shelley, J. C. *Journal of chemical theory and computation* **2023**, *19*, 2380–2388.

(29) Bochevarov, A. D.; Watson, M. A.; Greenwood, J. R.; Philipp, D. M. *Journal of chemical theory and computation* **2016**, *12*, 6001–6019.

# S1. Supporting Information

## S1. Reaxys Query for C−H activation

The query on Reaxys consists of the following steps:

1. Seach for "C−H activation" which yields 553472 reactions
2. Filter by catalyst "$Pd(OAc)_2$" which yields 705 reactions
3. Filter by reaction class "C-C bond formation" which yields 17 reactions
4. Selecting the one reaction that is actually a Pd-catalyzed C−H activation with a directing group and search for similar reactions with criterion "wide" which yields 37105 reactions
5. Filter by catalyst "$Pd(OAc)_2$" which yields 1955 reactions
6. Filter by reaction class "C-C bond formation" and the subcategories "Ar-H to Ar-CH2-", "ArH to Ar-C(=)-", "ArH + -C(=)-X to Ar-C(=)-", "ArH + -C(=)-O- to Ar-C(=)-" and "ArH + -CH2-O- to Ar-CH2-" which yields 330 reactions
7. Out of the 330 reactions we chose all examples that had several potential reaction sites and/or directing groups which yields 10 reactions

## S2. All potential complexes considered for C−H activation of a specific substrate

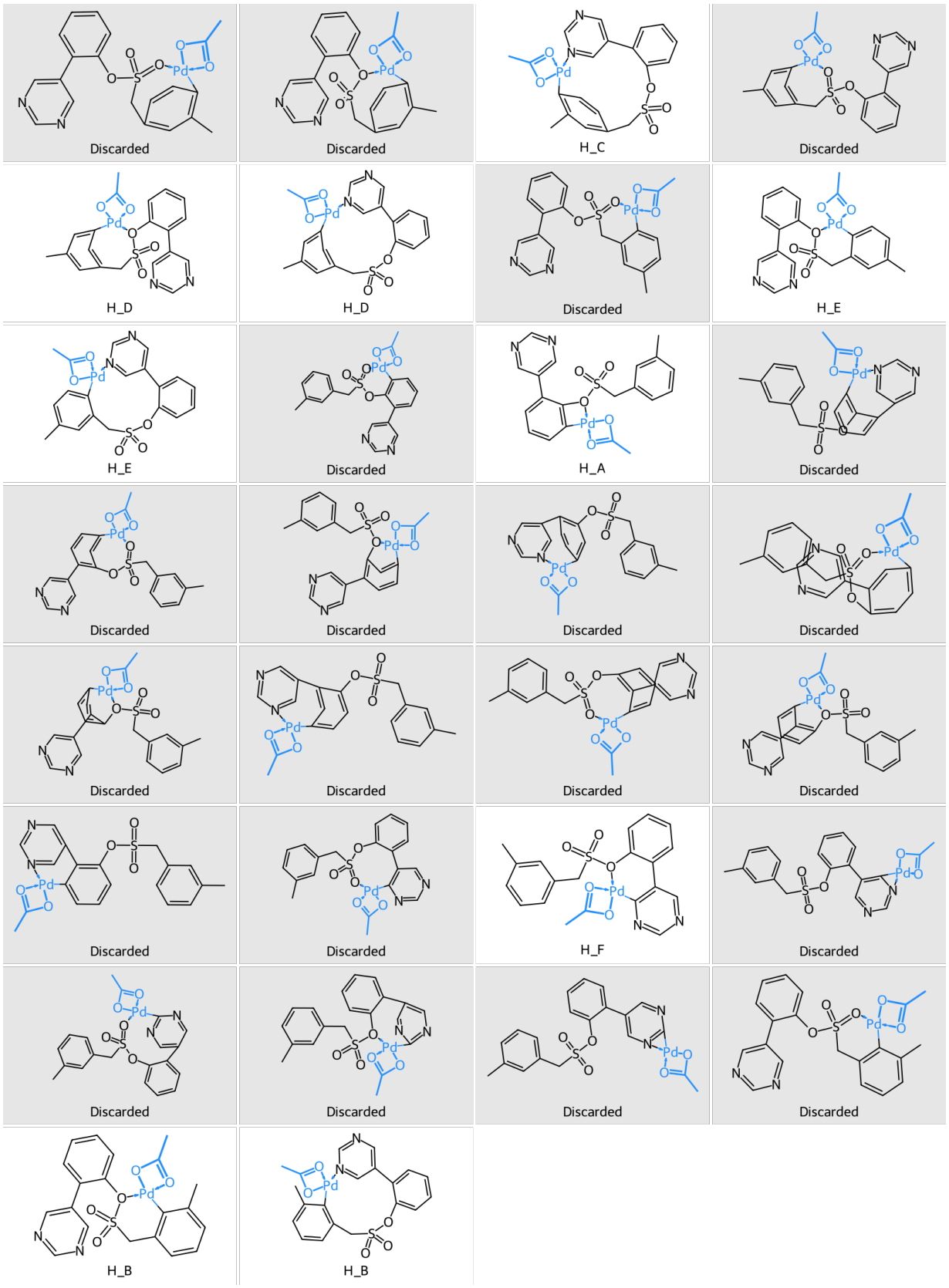



**Figure S1.** All potential palladacylce intermediates for a substrate from Achar et al. [1]. Via screening of the MMFFs energy contribution from the angle related terms, unreasonable combinations of C−H site and directing groups are discarded.

## S3. Correct regioselectivity prediction for the Dataset from Tomberg et al. [2]/Chen et al. [3]

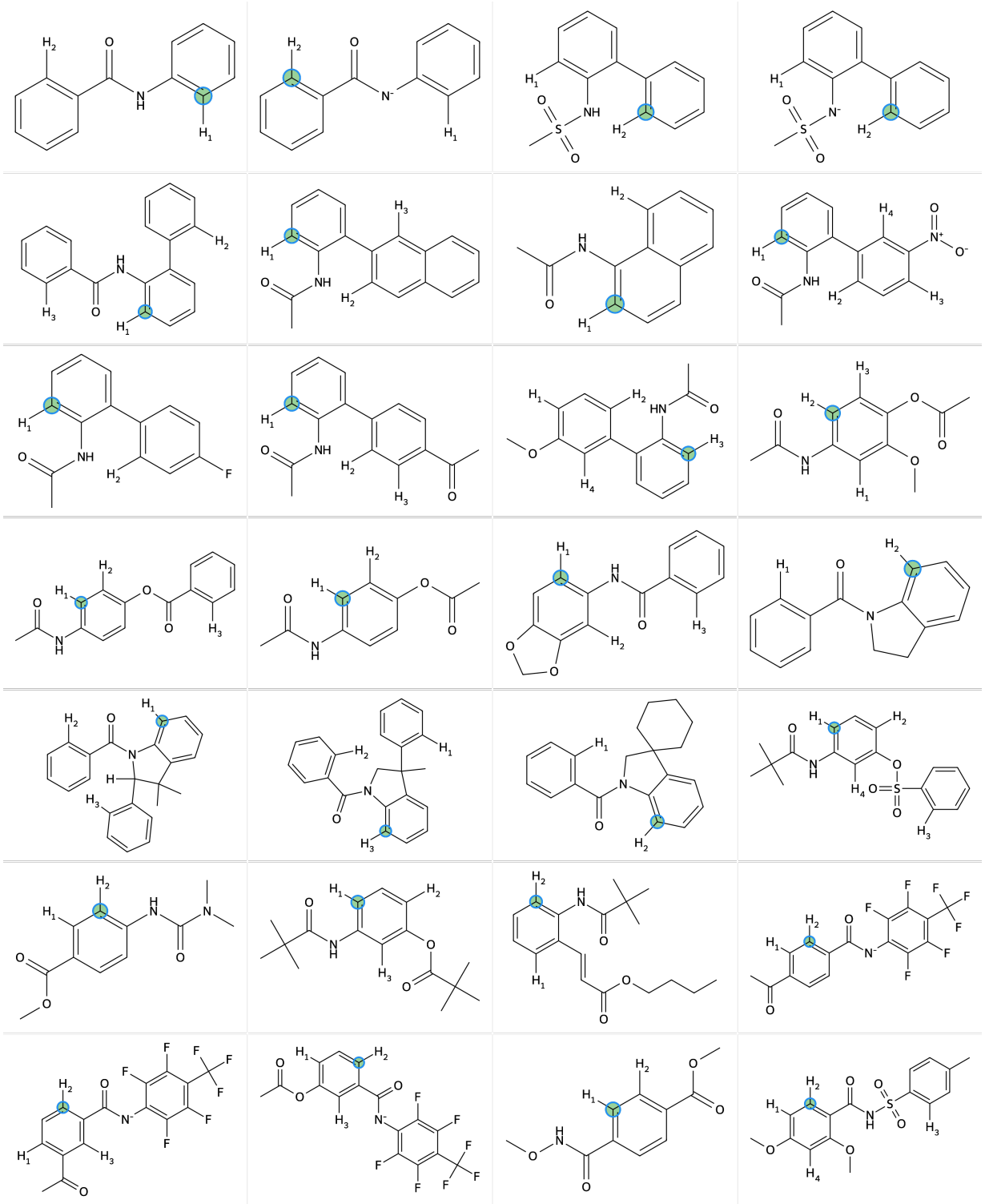


**Figure S2.** (*1/4*) Correct regioselectivity predictions for molecules from Tomberg et al. [2]/Chen et al. [3]. The predicted reaction site is marked by a blue circle, the experimentally observed reaction site is highlighted with a green circle.

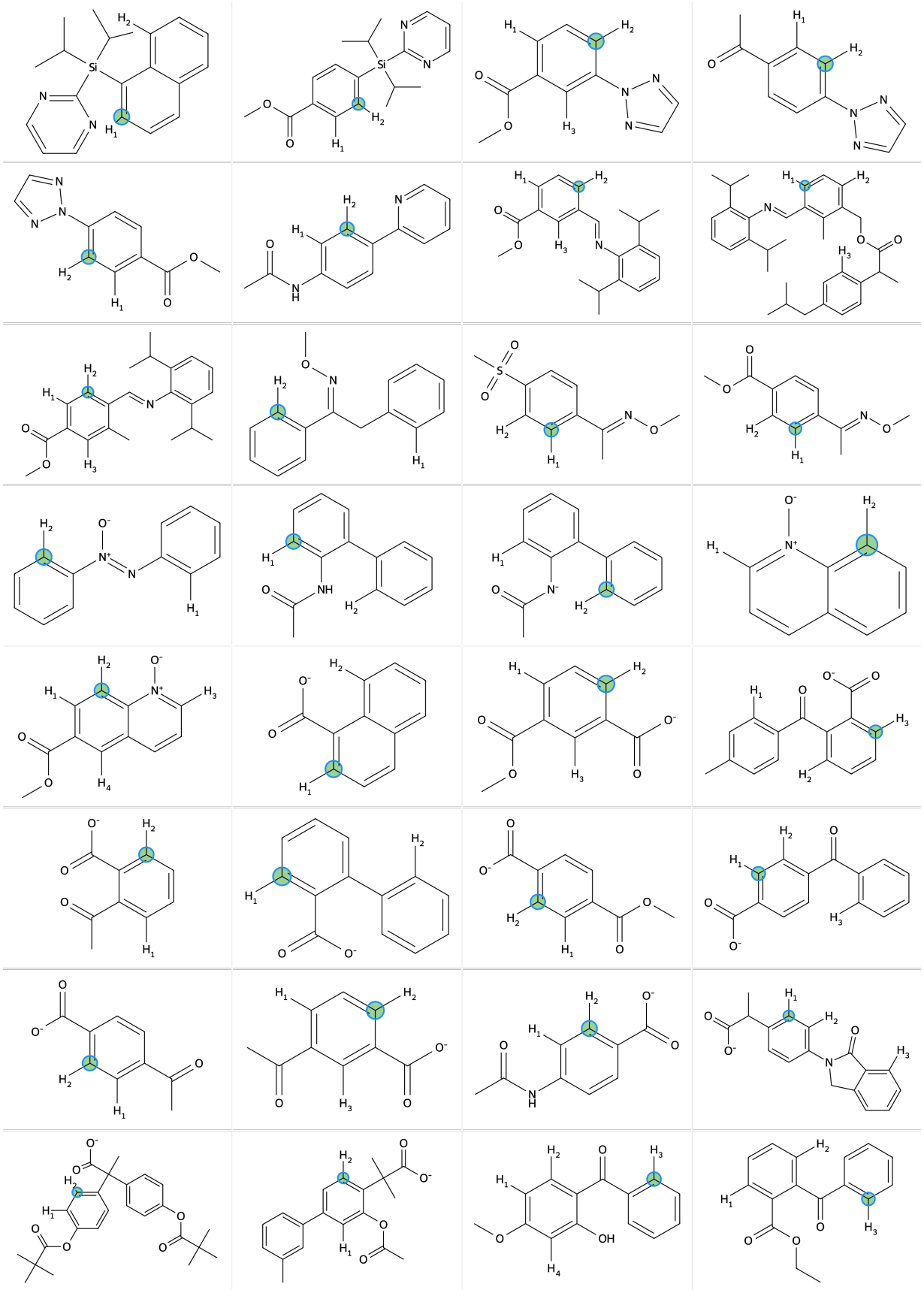


**Figure S3.** (*2/4*) Correct regioselectivity predictions for molecules from Tomberg et al. [2]/Chen et al. [3]. The predicted reaction site is marked by a blue circle, the experimentally observed reaction site is highlighted with a green circle.

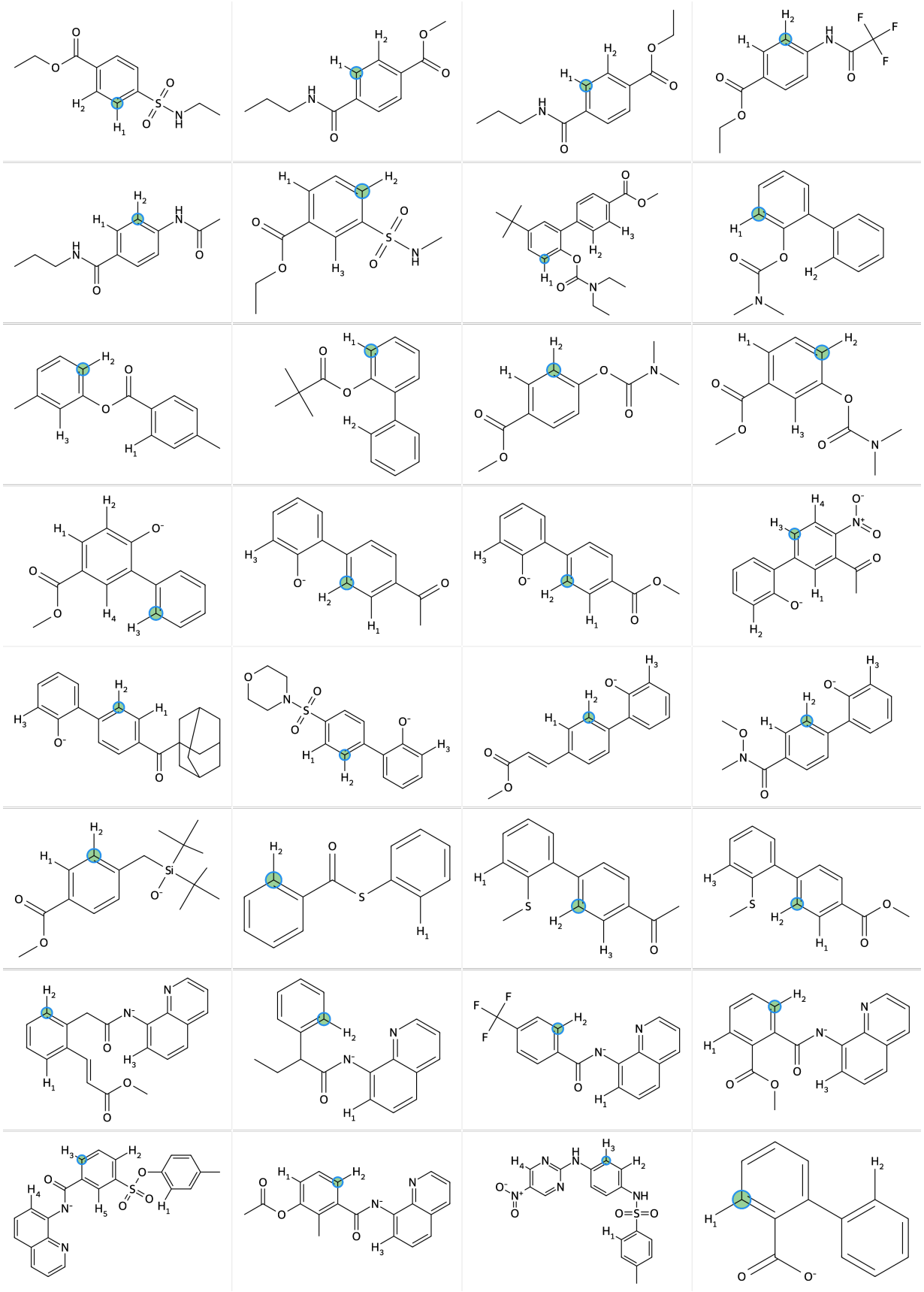


**Figure S4.** (*3/4*) Correct regioselectivity predictions for molecules from Tomberg et al. [2]/Chen et al. [3]. The predicted reaction site is marked by a blue circle, the experimentally observed reaction site is highlighted with a green circle.

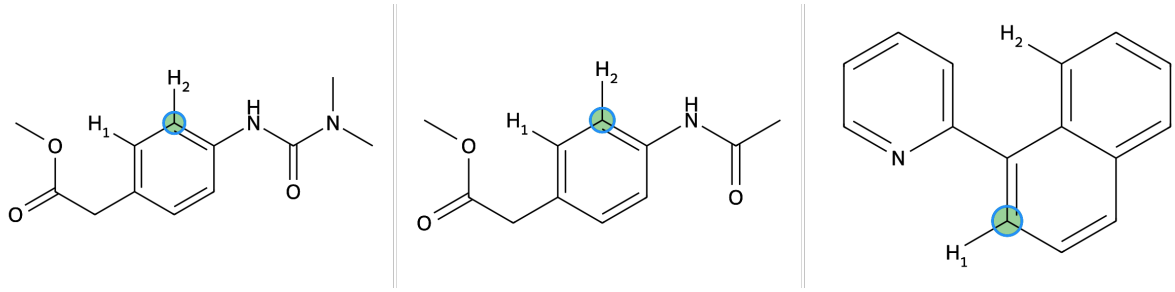



**Figure S5.** (*4/4*) Correct regioselectivity predictions for molecules from Tomberg et al. [2]/Chen et al. [3]. The predicted reaction site is marked by a blue circle, the experimentally observed reaction site is highlighted with a green circle.

## S4. Semi-correct regioselectivity prediction for the Dataset from Tomberg et al. [2]/Chen et al. [3]

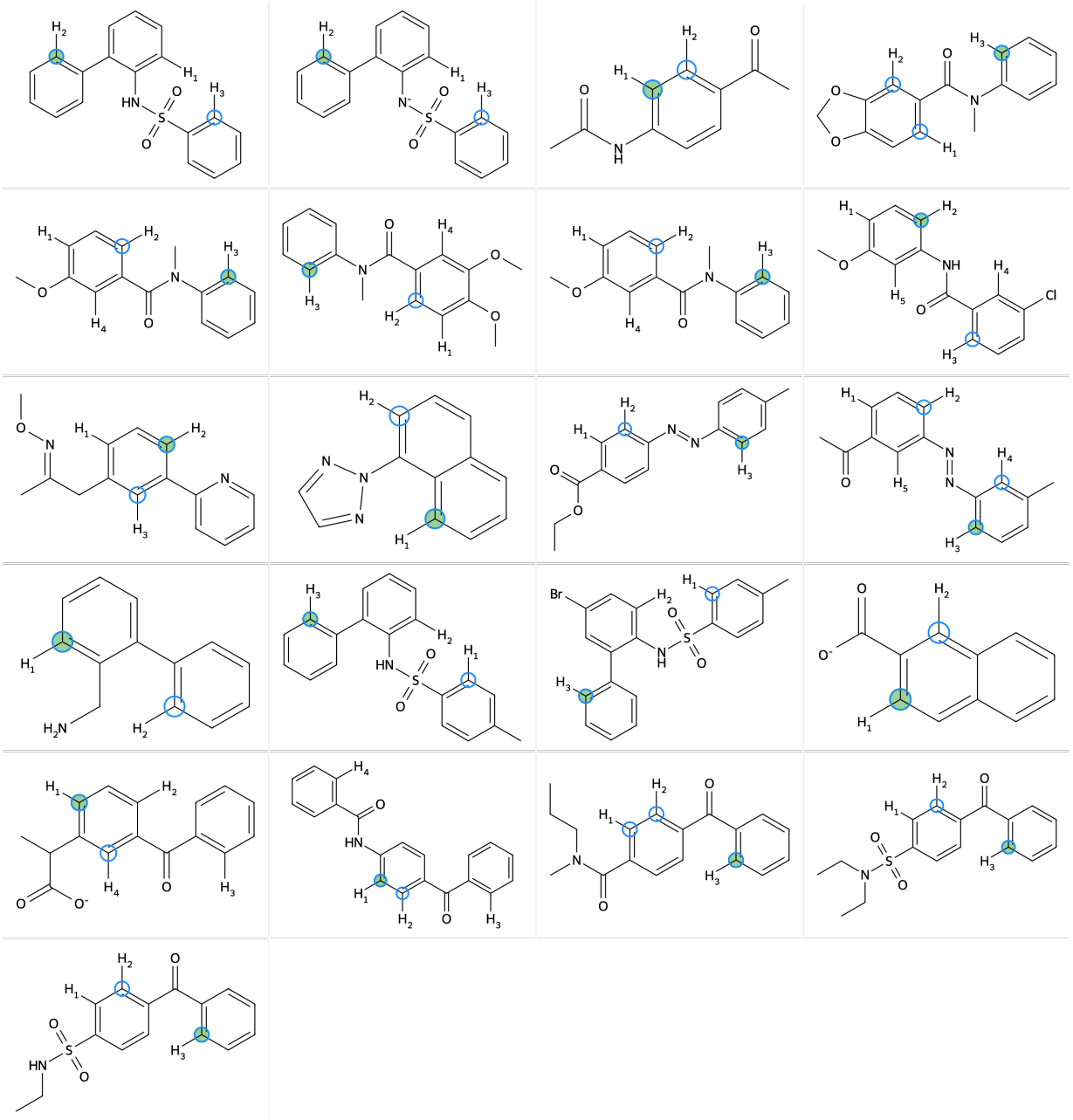


**Figure S6.** Semi-correct regioselectivity predictions for molecules from Tomberg et al. [2]/Chen et al. [3]. Predicted reaction sites are marked by a blue circle, the experimentally observed reaction site is highlighted with a green circle.

## S5. Wrong regioselectivity prediction for the Dataset from Tomberg et al. [2]/Chen et al. [3]

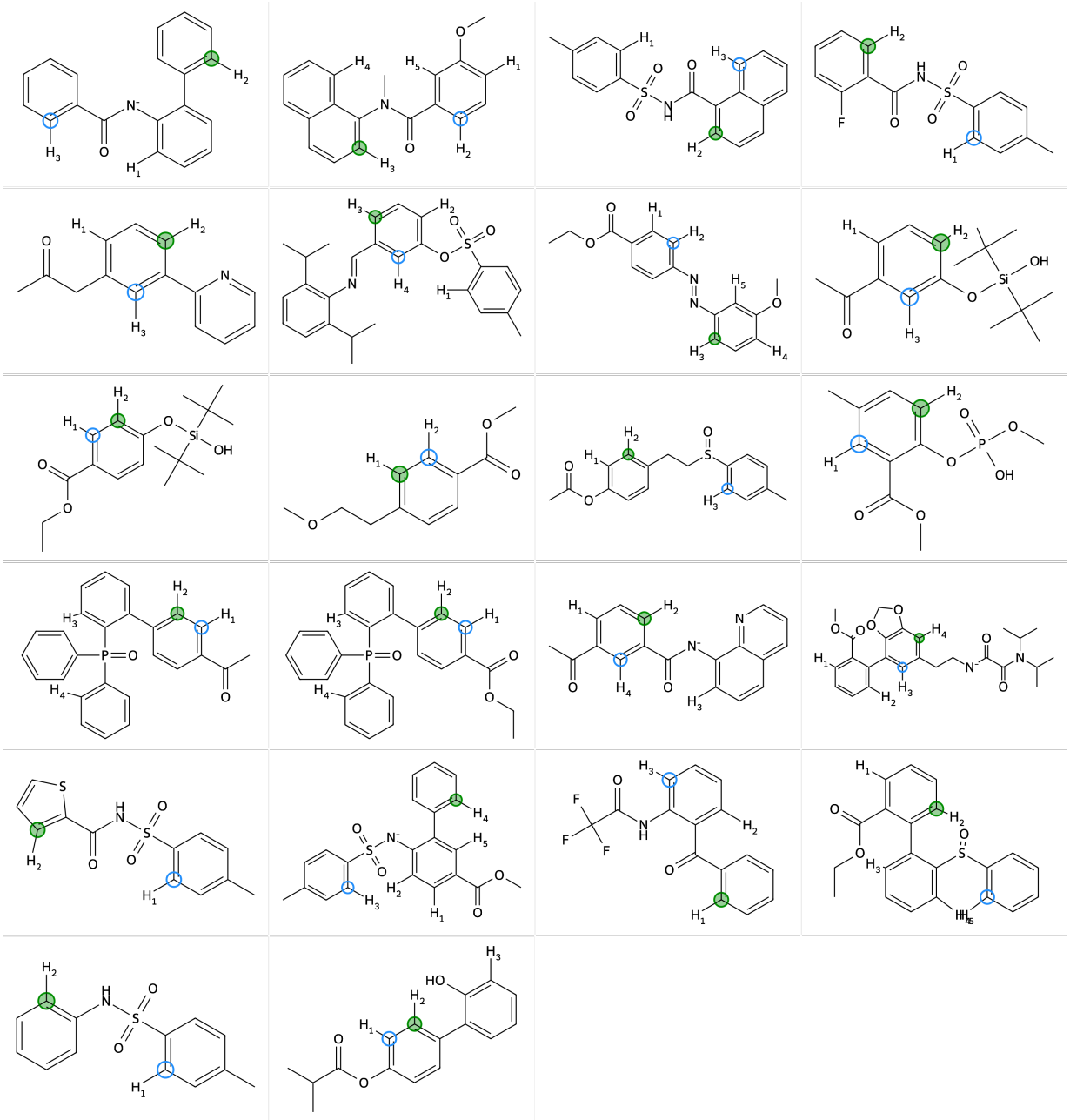


**Figure S7.** Wrong regioselectivity predictions for molecules from Tomberg et al. [2]/Chen et al. [3]. The predicted reaction site is marked by a blue circle, the experimentally observed reaction site is highlighted with a green circle.

## S6. Web-based interface

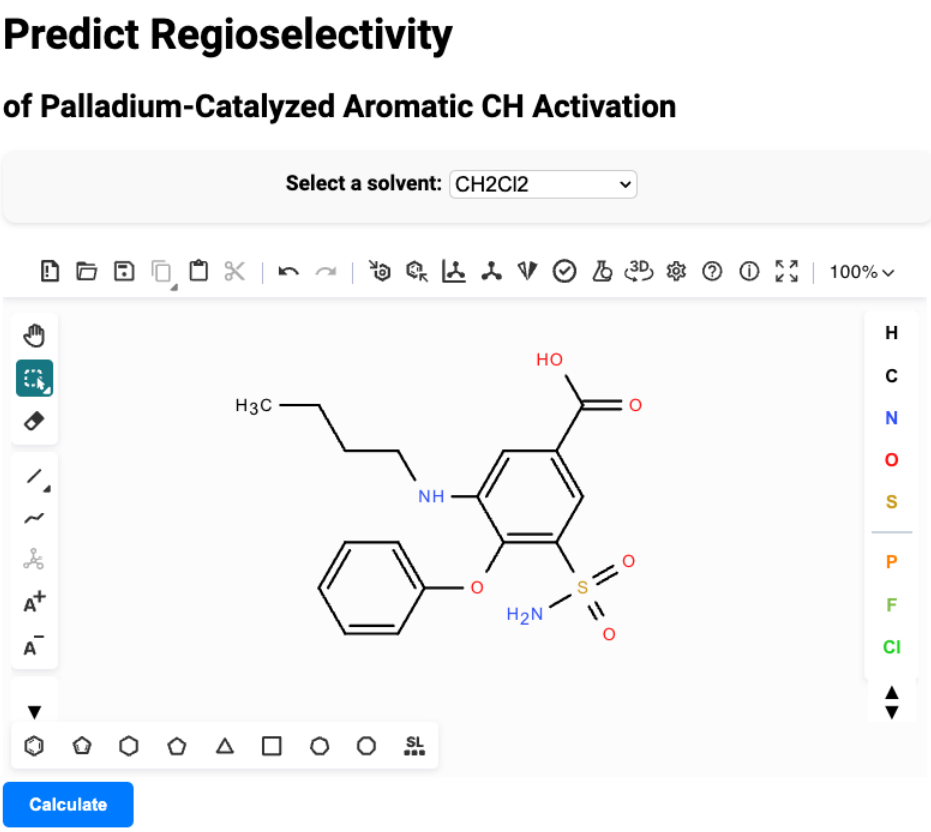



**Figure S8.** Overview of the regioselectivity prediction user interface.

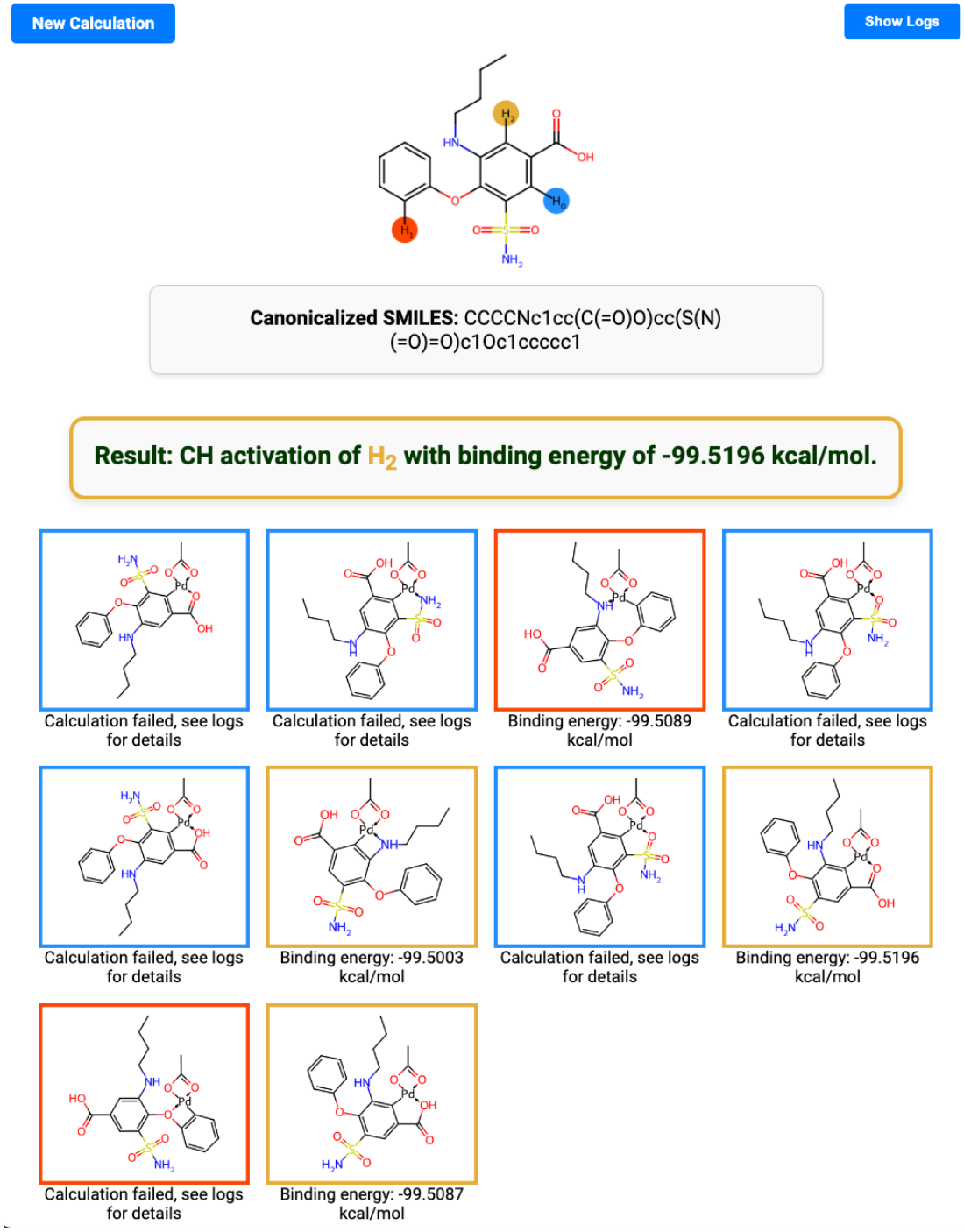



**Figure S9.** Overview of the regioselectivity prediction using the web-based user interface

# 7

# DISCUSSION AND OUTLOOK

We have developed a user-friendly, fast, and accurate workflow to predict the regioselectivity of directed C–H activation reactions. Like all machine learning (ML) or quantum mechanical (QM) prediction models, accurate results depend on posing the correct question. Specifically, our model determines the most likely site for C–H activation on a substrate according to a given mechanism. It does not, however, predict whether a reaction will occur or which reaction mechanism will be followed when a substrate and catalyst are combined experimentally. Instead, it identifies where a specific reaction is most likely to take place. Although this limits certain applications, the workflow holds significant potential for organic chemists, particularly when used in conjunction with previously developed tools like RegioSQM and other regioselectivity prediction models.

Looking ahead, we aim to create an integrated workflow that inputs a substrate and then labels and ranks potential reaction sites according to specific mechanisms. This feature will be especially beneficial for late-stage functionalization.

Take, for example, the pharmaceutical Clopidogrel, which reduces the risk of blood clots and presents multiple potential reaction sites on its $SP^2$ and $SP^3$ carbon atoms, as shown in Figure 7.1. Many reac-

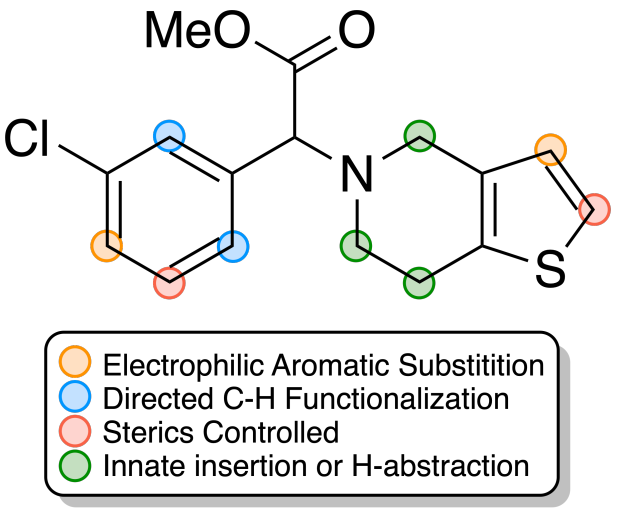



Figure 7.1: Structure of the drug Clopidogrel with potential reaction sites labelled by potential reaction mechanisms, from Cernak et al. [47]

tion mechanisms and principles are available, requiring careful selection based on the desired reaction site. Identifying the most effective method for modifying the existing molecular scaffold at a specific

site is essential to achieve efficient conversion. This integrated approach will help streamline the decision-making process in complex synthetic tasks, making it a valuable asset in organic synthesis.

# 8

# GENERAL CONCLUSIONS AND OUTLOOK

Throughout this thesis, I have explored the optimization of catalysts using QM calculations and the prediction of regioselectivity in catalysed reactions. The research presented has expanded our toolkit for catalyst optimization, introducing innovative techniques for advancing catalyst design and predicting regioselective reaction outcomes.

One of the most significant achievements of this research was the optimization of an organic catalyst, which led to an almost eightfold increase in reaction rates. Here, we explored chemical space beyond predefined fragment libraries, setting our approach apart from conventional evolutionary strategies in catalyst design.[8–13] Our experimentally verified improvement exemplifies the practical utility of computational models in real-world applications and highlights the potential of computational approaches to enhance catalytic processes substantially.

Despite these successes, several challenges remain in broadening the applicability and accuracy of these computational tools. While we have shown that computational methods can accurately reproduce experimental reactivity trends in a specific reaction, we are far from having a general method for quantifying a catalyst's activity. This would ideally include automating the calculation of the reaction profile of any arbitrary reaction and identifying relevant side reactions. We have shown in Paper 4 that a network exploration approach can identify the relevant reaction mechanism and locate the relevant transition states (TSs) for an organic catalytic system. Yet, this approach has not been tested for TM-based catalysts, which might present additional challenges with regard to the accuracy of the underlying SQM method.

Furthermore, the incorporation of relevant chemical constraints such as stability and synthesizability needs further consideration. This aspect is particularly critical for TM-based catalysts, where the practical synthesis of the designed catalysts remains a hurdle. We have successfully used heuristics, which were originally developed for drug-like molecules. More tailored solutions for catalysts, also TM-based ones, will improve the chemical exploration process in evolutionary algorithms.

Looking forward, the methodologies developed in this thesis lay the groundwork for more sophisticated, automated workflows that integrate both catalyst activity optimization and selectivity prediction. I am optimistic that the continued refinement of these computational tools will not only deepen our understanding of catalytic pro-

cesses but also lead to significant advancements in the efficiency and sustainability of chemical production. The real-world impact of these technologies has the potential to transform industries, streamline production methods, and lead to the development of more effective and environmentally friendly catalysts.

By embracing these challenges, continuing to innovate and taking inspiration from nature, we can further the role of computational chemistry in catalyst design and help push the boundaries of what is possible in chemical synthesis and beyond.

# APPENDIX

GDCh Research Articles Angewandte International Edition Chemie www.angewandte.org

Check for updates

How to cite: *Angew. Chem. Int. Ed.* **2023**, *62*, e202310580
doi.org/10.1002/anie.202310580

Catalysis

# Toward De Novo Catalyst Discovery: Fast Identification of New Catalyst Candidates for Alcohol-Mediated Morita–Baylis–Hillman Reactions**

*Maria H. Rasmussen,* Julius Seumer, and Jan H. Jensen**

**Abstract:** Recently we have demonstrated how a genetic algorithm (GA) starting from random tertiary amines can be used to discover a new and efficient catalyst for the alcohol-mediated Morita–Baylis–Hillman (MBH) reaction. In particular, the discovered catalyst was shown experimentally to be eight times more active than DABCO, commonly used to catalyze the MBH reaction. This represents a breakthrough in using generative models for catalyst optimization. However, the GA procedure, and hence discovery, relied on two important pieces of information; 1) the knowledge that tertiary amines catalyze the reaction and 2) the mechanism and reaction profile for the catalyzed reaction, in particular the transition state structure of the rate-determining step. Thus, truly de novo catalyst discovery must include these steps. Here we present such a method for discovering catalyst candidates for a specific reaction while simultaneously proposing a mechanism for the catalyzed reaction. We show that tertiary amines and phosphines are potential catalysts for the MBH reaction by screening 11 molecular templates representing common functional groups. The method relies on an automated reaction discovery workflow using meta-dynamics calculations. Combining this method for catalyst candidate discovery with our GA-based catalyst optimization method results in an algorithm for truly de novo catalyst discovery.

## Introduction

The search for new catalysts is instrumental in addressing some of humanities current challenges. The right catalyst in principle holds the key to selective and sustainable production of target molecules; an important task in all areas of chemical research and industry. However, several factors make catalyst discovery challenging, slow and expensive. Detailed knowledge about the reaction network including side reactions and solvent reactions is needed and these steps still rely heavily on chemical intuition and experimentation.

So far, the main contribution of computational chemistry methods to catalyst discovery has been in establishing a mechanism for catalytic cycles of known catalysts. Current state of the art within computationally driven catalyst discovery is generally focused on improving known catalysts based on the catalytic mechanism. One approach is to screen large libraries of molecules that are structurally similar to the known catalyst. For example Nandy et al. combined machine learning (ML) and density functional theory (DFT) methods to screen a library of 16 million candidates for the catalysis of the methane to methanol oxidation.[1] Another recent example of computationally driven catalyst optimization from Cramer et al. used mechanistic information of three intertwined catalytic cycles for $CO_2$ conversion to formic acid, formaldehyde 1 and methanol to generate theoretically founded rules for predicting catalytic activity and selectivity.[2] The theoretical framework was used to predict a catalyst that optimizes selectivity towards formaldehyde, which was verified experimentally with an 81 % yield. Das et al. mapped catalytic activity in $CO_2$ hydrogenation to acid/base properties of frustrated Lewis pairs to predict a specific combination of a Lewis acid and a Lewis base from ≈4000 candidate pairs which was experimentally verified to catalyze the reaction.[3]

An alternative to the screening approach is to use generative models to propose new catalyst candidates. The advantage of this approach is that one is not limited to discovering catalyst candidates already present in a predefined library. One such approach; a graph-based genetic algorithm (GA), was recently employed by our group to discover a new catalyst with a previously untested structural motif for the alcohol-mediated Morita–Baylis–Hillman (MBH) reaction.[4,5] The proposed catalyst candidate was experimentally verified to be eight times more active than 1,4-diazabicyclo[2.2.2]octane (DABCO), a commonly used catalyst for the reaction.[5]

Vital to these important findings is mechanistic insight into the relevant catalytic cycle. Typically, these catalytic cycles are generated by a combination of experimental

[*] M. H. Rasmussen, J. Seumer, J. H. Jensen
Department of Chemistry, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark
E-mail: mhr@chem.ku.dk
jhjensen@chem.ku.dk

[**]A previous version of this manuscript has been deposited on a preprint server (https://doi.org/10.26434/chemrxiv-2023-lchmv).

work, expert knowledge and quantum chemical calculations, making this step a laborious undertaking. Moreover, someone needs to actually have the idea of testing a specific molecular moiety as a catalyst for a reaction. Thus, a method for truly de novo catalyst discovery must also include a solution for this part of the discovery process.

Since catalytic activity can be extracted from a reaction network including the reactants and potential catalyst candidate, many groups (including us) have worked to develop methods for automated reaction discovery for the generation and exploration of reaction networks.[6–18] This in principle provides a way of discovering completely novel catalysts for reactions with no prior knowledge of catalytic activity.

Some proof-of-principle papers have been presented using these methods to discover the mechanism of a catalytic cycle. Specifically the cobalt-catalyzed alkene hydroformylation has been a popular example, since the system consists of only 18 atoms.[8,10,19–22] A more widespread application of these reaction discovery methods is mostly hindered by the vast computational cost of freely growing several reaction networks. In order to move from hypothesis-testing to discovery, the methods need to be efficient enough that screening of different potential catalyst candidates is possible.

Our method for exploring reaction networks is based on the meta-molecular dynamics (meta-MD) approach proposed by Grimme.[23] As we have demonstrated previously, an automated workflow that tracks the reactions occurring during the meta-MD simulations can be used as an efficient way of predicting which low-barrier (defined as $<30$ kcal/mol) reactions are possible.[11,12] The combination of using a method that focuses on low-barrier reactions while relying on a fast semi-empirical quantum chemistry method (GFN2-xTB)[24] means that we can grow reaction networks quite fast even for larger molecular systems, making screening applications possible.

In this work, we present a method for discovering that a certain catalyst-template such as a tertiary amine can be used to catalyze a certain reaction such as the MBH reaction. We build reaction networks for the reactants of the MBH reaction with 11 different catalyst-templates and use them to detect the presence/absence of catalytic activity. As part of the method we also map out a proposed mechanism for the catalyzed reaction. We demonstrate the method by re-discovering tertiary amines and phosphines as catalysts for the MBH reaction. From the found reaction profile we extract a transition state (TS) template for the ratedetermining step. With the TS-template available, the genetic algorithm (GA) applied by Seumer et al. can be used to optimize the catalytic activity. This represents an end-to-end approach for de novo catalyst discovery starting from no prior knowledge about how a reaction can be catalyzed and ending with a range of possible catalysts optimized for said reaction (Figure 1).

Within drug discovery, two of the crucial steps are lead identification and lead optimization. Lead identification is the task of finding a compound that is active against a specific drug target. Once a lead compound is identified, lead optimization represents the process of making structural changes that improves activity and selectivity while reducing toxicity and other unwanted side effects. The

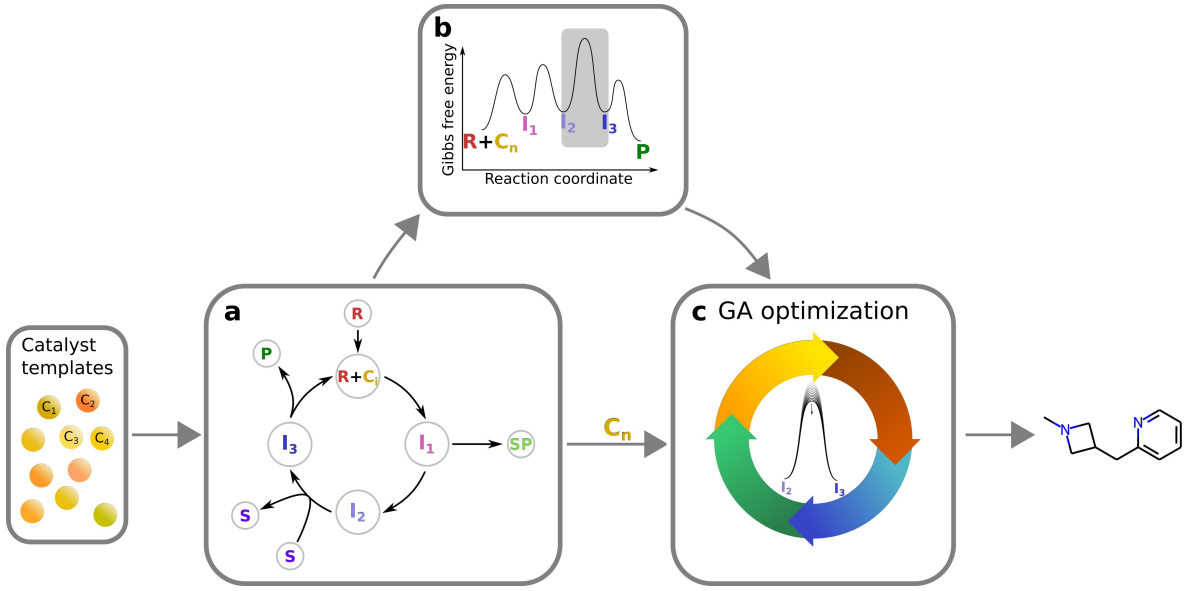



***Figure 1.*** Overview of the proposed method to discover new catalysts: (a) We grow a reaction network for each of the possible catalyst templates ($C_i$) and look for catalytic activity in each reaction network. The reactant system (R) reacts with the catalyst template ($C_i$) with the possibility to form several intermediates ($I_i$) before reaching a product (P) with the catalyst template regenerated. Potential side products (SP) are also identified. Solvent (S) mediated hydrogen transfer reactions are also included in the procedure. (b) For catalyst templates exhibiting catalytic activity, we extract the mechanism of the reaction and calculate the full reaction profile, considering possible side-reactions. (c) Finally the catalyst template and reaction mechanism are used as input for a genetic algorithm optimization resulting in new catalyst candidates.

analogy to catalyst discovery is clear and while current state-of-the-art computationally driven catalyst discovery has been focused on catalyst optimization (i.e. lead optimization), this work represents a step forward in filling in the gap for lead identification.

## Results and Discussion

To test the meta-MD based reaction discovery method described above we set out to rediscover tertiary amines and phosphines as catalysts for the MBH reaction. The MBH reaction is represented by the reaction of methyl acrylate (MA) with *p*-nitrobenzaldehyde (*p*NBA) following Seumer et al.[5] The task is then to find a catalyst template that can catalyze the reaction between MA and *p*NBA.

In the Supporting Information (section S1) we describe the iterative method based on meta-MD calculations used for growing the reaction networks. In short we find possible reactions by tracking bond changes during the meta-MD runs. From the intermediates of these reactions, intermediates formed from solvent mediated proton transfer reactions (tautomers) are added to the network. Note that in this work we choose not to screen the found reactions based on barriers but only reaction energies. These choices are possible since the number of intermediates generated by our reaction discovery procedure is low enough that we can grow the reaction network quite a bit with screening based only on reaction free energies (in this case three iterations of step 1–4 in the Supporting Information, section S1) before doing another iteration would become computationally unfeasible (months on our local cluster). Subsequently we can define the parts of the reaction networks found that seem most interesting. In this case the application is to look for potential catalysts, therefore we look for places in the reaction network where the catalyst is regenerated. Meanwhile the reaction energy for the catalyzed reaction should not be very endothermic, the threshold will depend on the accuracy of the DFT method used for calculating the reaction energies.

To demonstrate the rediscovery process, we initially provide a list of eight possible catalyst templates (Round 1), for which reaction networks with the reactants (MA and *p*NBA) are grown following step 1–4 in the Supporting Information (section S1). Based on an analysis of the most promising reaction networks grown in Round 1, three new possible catalyst templates are tested (Round 2).

We build reaction networks for eight possible templates (Figure 2, Round 1) with a variety of functionality. In particular, we include several different kinds of nucleophiles; ethene (representing a double bond), hydroxide (nucleophile and strong base), ammonia (nucleophile and weak base) and phosphine. We also include entries that act as acids; the hydronium ion (strong acid, conjugate base weak nucleophile), hydrogen sulfide (weak acid, conjugate base strong nucleophile) and formic acid (representing a carboxylic acid). Formaldehyde is included as an example of an electrophile.

In the first step of generating reaction networks for the eight template structures in Figure 2, we include MA and each of the catalyst templates in the meta-MD run. Before doing another step of the reaction discovery workflow, we add *p*NBA to the reaction mixture. Since we only consider elementary reactions between two molecules, in case more than two molecules are present in an intermediate, we only search for reactions between the two largest molecules. Note that in an application that is not rediscovery, one would also grow reaction networks starting from *p*NBA + catalyst template and later adding MA. This can be expected to roughly increase the computational cost with a factor of two.

The eight reaction networks vary a lot in size after three iterations of the procedure. For formic acid (**C8**), the network consists of only 14 intermediates (Figure S3a) while for phosphine (**C5**) the network consists of 298 intermediates (Figure S3b). For some of the catalyst templates (**C3**, **C5** and **C6**), the inclusion of tautomers adds a vast number of new intermediates; notice the number of yellow edges (proton transfer reactions) for phosphine (Figure S3b). Since we start the reaction discovery procedure for each intermediate, more intermediates found means a higher computational cost of growing the networks.

Of the eight catalyst templates tested, only the reaction network for ammonia (**C3**) shows clear signs of catalytic activity, defined by the presence of a node with the catalyst regenerated and reaction free energy $<10$ kcal/mol. For the remaining seven reaction networks, some show no intermediate/product with the template regenerated (**C2**, **C6** and

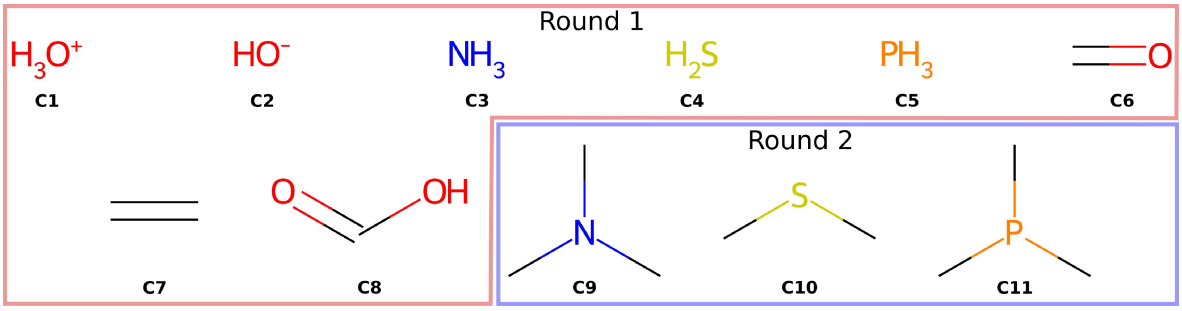



***Figure 2.*** The 11 templates tested as potential starting points for catalyst optimization of the MBH reaction.

**C7**) while some have 1–3 intermediates with the catalyst regenerated but all with reaction Gibbs free energies >10 kcal/mol (**C1**, **C4**, **C5** and **C8**). Phosphine (**C5**) is thus not recognized as a potential catalyst from the reaction network grown after three iterations. Below, we will analyze the behavior of phosphine in a bit more detail by comparing with the reaction network found for ammonia.

For ammonia, the primary reaction path found with meta-MD as the first step is nucleophilic attack at the $\beta$-carbon of MA to form intermediate **13** (Figure 3a). Another major reaction is transfer of a proton from ammonia to oxygen simultaneously with the nucleophilic attack forming intermediate **16**. For phosphine, we see another primary reaction path for the first step of meta-MD generated products which is the addition of phoshine to the double bond forming intermediate **12** (Figure 3b). While the second most observed reaction is nucleophilic attack of phosphine to the $\beta$-carbon of MA forming intermediate **25**, the Gibbs free energy at 27 kcal/mol is much higher and in fact very close to our cutoff of 30 kcal/mol. In ammonia's case we do find the next step of the MBH reaction when reacting intermediate **13** with *p*NBA forming intermediate **209**. However, when trying to optimize this intermediate at the DFT level, we observe a proton transfer between nitrogen

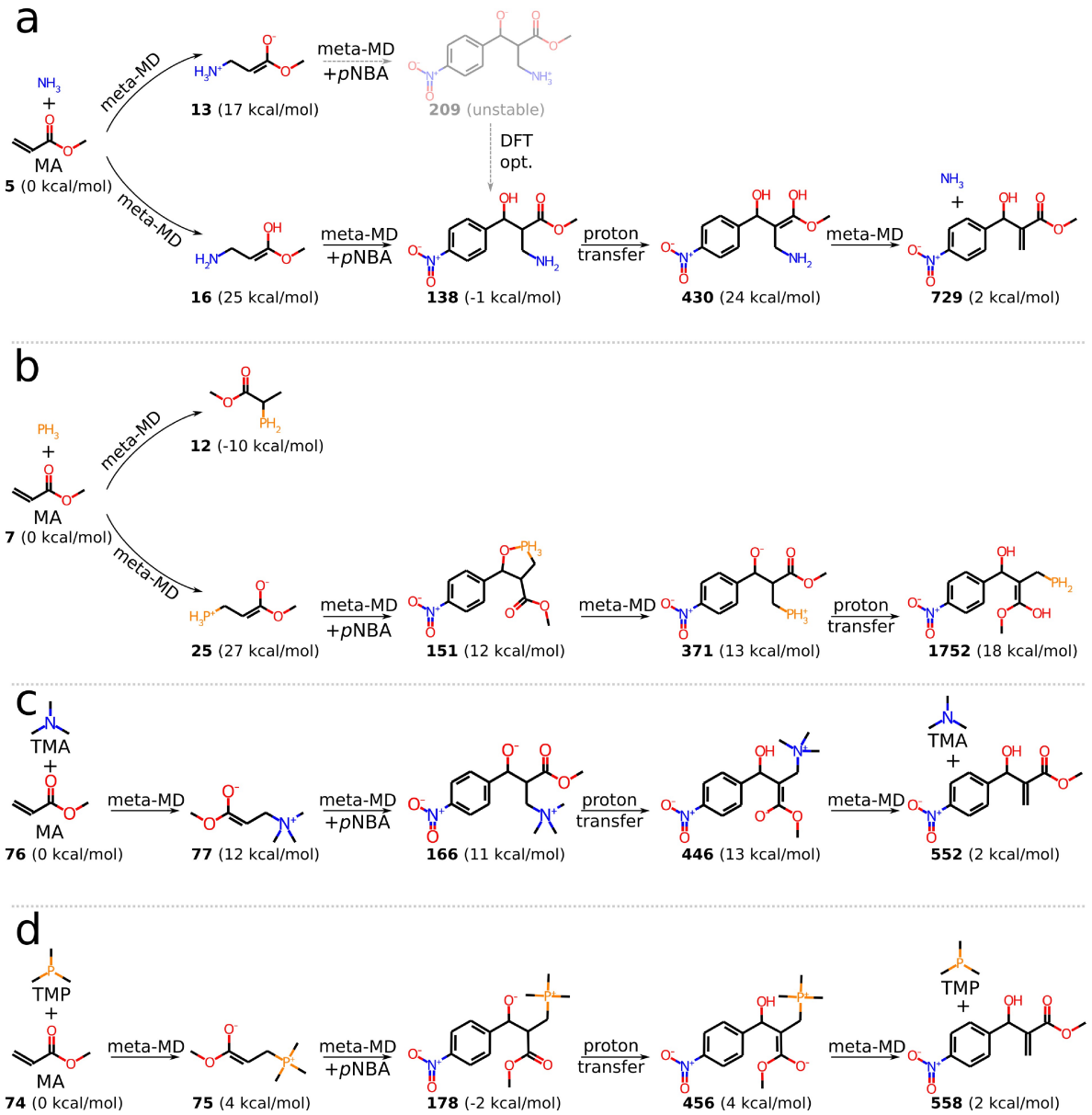



***Figure 3.*** Highlighted reaction paths towards the MBH product discussed in the manuscript for (a) **C3**: ammonia, (b) **C5**: phoshine, (c) **C9**: TMA and (d) **C11**: TMP. The intermediate energies are Gibbs free energies relative to the individual reactant molecules at the B3LYP-D3/6-31+G(d,p) level of theory following Seumer et al.[5] (details in the Supporting Information, section S1). Note that bold integers are merely intermediate labels.

and oxygen forming intermediate **138** (Figure 3a). Intermediate **430** is found within the same iteration as a tautomer to intermediate **138**. For phosphine on the other hand we observe another intermediate where a five-membered ring is formed from phosphor and oxygen binding (intermediate **151**) which is possible due to phosphors ability to form 5 bonds. We do find the expected MBH intermediate (intermediate **371**) in the next meta-MD iteration and with it intermediate **1752** from a proton transfer reaction which is the equivalent to intermediate **430** for ammonia. Thus, the reason we are not observing the MBH product for phosphine is that the extra intermediate observed on the path (intermediate **151**) means that it would take another, fourth, iteration of our reaction discovery method to get there. In the Supporting Information (section S2) we provide an analysis of the remaining reaction networks of Round 1; in particular their similarities and deviations to the path observed for ammonia.

Based on the initial analysis of reaction networks for the eight potential catalyst candidates (Figure 2, red box) ammonia is the most promising being the only candidate that shows catalytic activity after three iterations of the reaction discovery procedure. For both ammonia and phosphine we see that the possibility of a proton transfer from the nitrogen/phosphor atom creates a vast amount of reaction channels (Figures S4 and S3b). A way of hindering those reactions would be to exchange the hydrogens in ammonia and phosphine with methyl groups forming trimethylamine (TMA) and trimethylphosphine (TMP), respectively. For completeness, we also try dimethyl sulfide (DMS) formed by changing the hydrogens in hydrogen sulfide to methyl groups. Thus, in the second round of testing possible catalyst templates, we include TMA (**C9**), DMS (**C10**) and TMP (**C11**) (Figure 2, Round 2).

Figure 4 shows the reaction network grown for TMA (**C9**) after three iterations of the meta-MD procedure. Importantly, the MBH reaction is discovered as a four-step reaction (highlighted in green). From the initial reactant system (TMA + MA, intermediate **76** in the reaction network), the only elementary reaction found with meta-MD is the attack of TMA on the $\beta$-carbon (resulting in intermediate **77**), which is indeed the first step of the MBH reaction. From this point we observe four different reaction paths, all involving a nucleophilic attack of the enolate anion. Attack at the carbonyl carbon of *p*NBA results in the second intermediate of the MBH reaction (intermediate **166**). For

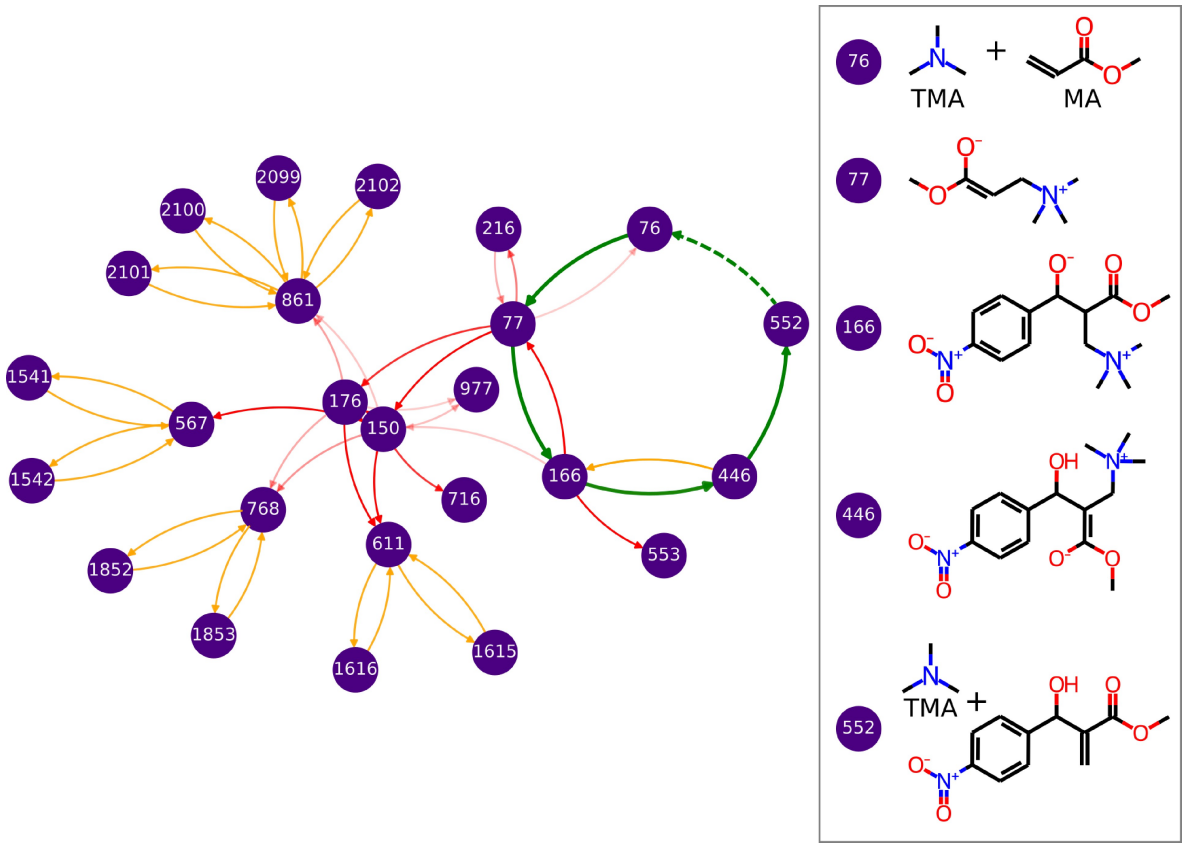



***Figure 4.*** The reaction network grown for TMA (**C9**). Red arrows represent reactions found with meta-MD, while orange arrows represent the proton-transfer reactions described in step 2 (Supporting Information, section S1). The transparency of the red arrows indicate how many times a reaction is found in the meta-MD simulations. A fully colored arrow is found at least 30 times. Only reactions found more than five times in the meta-MD simulations are included in the network. The MBH reaction path found as part of the network is highlighted in green.

the additional three elementary reactions originating at intermediate **77**, the enolate anion attacks at either a benzene carbon ortho to the nitro group (intermediate 216) or at an oxygen of the nitro group (intermediate **150** and intermediate **176**). The third step of the MBH reaction is found as a proton-transfer reaction from intermediate **166** (generating intermediate **446**) and from here on, the only reaction found by the meta-MD approach is the final step of the MBH reaction to form the product (intermediate **552**).

TMP (**C11**), the phosphor-equivalent to TMA, generally behaves similarly with key features of the reaction networks being identical (Figure S5). Importantly, the MBH reaction path is also identified; intermediate **74**→intermediate **75**→intermediate **178**→intermediate **456**→intermediate **558**. The reaction network for (**C10**) indicates no sign of catalytic activity.

For both TMA (Figure 3c) and TMP (Figure 3d) we find only a single elementary reaction for MA+TMA/TMP which is indeed the expected first step of the MBH mechanism (intermediates **77** and **75**, respectively). This is contrary to what we observed for ammonia and phosphine, where competing paths were found already from the first step. Also, the intermediate free energies are significantly lower for TMA/TMP. We generally find far fewer reaction paths for TMA and TMP compared to ammonia and phosphine resulting in the much simpler reaction networks (Figures 4 and S5). For both TMA and TMP we find the expected MBH mechanism (Figures 3c and 3d). Without considering any barriers but solely based on Gibbs free energies of the intermediates, we would expect TMA and TMP to be better starting points for catalyst optimization. Here, we focus on TMA as a starting point for catalyst optimization using a genetic algorithm as done in Ref. [5].

Having obtained a possible catalyst template (TMA) and reaction mechanism from the reaction networks grown, we need to (1) validate the mechanism by calculating transition states (TSs) for all steps in the proposed mechanism and (2) consider the side-reactions suggested by the reaction network. In particular, having obtained a barrier for the rate-determining step in (1), we can evaluate competing reactions on their barrier heights being higher/lower than the rate-determining step.

Finding TSs for testing possible mechanisms (typically suggested by experimental chemists) is an important part of computational chemistry. While many promising methods for automating this process have been proposed in the last couple of decades,[26–30] much work regarding finding TSs is still done manually. For a non-screening application like this one, where a mechanism is proposed and a handful of TSs need to be found, there would likely be some degree of manual adjustment/evaluation. Numerous computational studies have demonstrated this kind of work for the MBH mechanism.[31–33] The TS-structures for the TMA catalytic cycle (steps highlighted with a full green arrow in Figure 4) are based on the TS-structures found by Liu et al.[32] The full reaction profile, with the proton transfer from intermediate **166** to intermediate **446** conducted by methanol, is shown in Figure 5. The energies and barriers for equivalent steps using DABCO as the catalyst are shown for comparison. As expected, TMA is calculated to be a worse catalyst than DABCO. With a calculated activation energy 1.3 kcal/mol higher than DABCO, the reaction catalyzed by TMA is expected to be roughly 9 times slower. The catalyst templates chosen need to be rather generic in order to reduce the number of templates to be investigated by growing their reaction networks. Therefore, we generally do not expect to find good catalysts among the templates. Rather, we hope to find a starting point for further optimization.

The activation energy for TMA is found to be 23.4 kcal/mol. Side reactions having activation energies that are lower than or close to this value must thus be considered. This can be done by incorporating knowledge of side reactions in the scoring function of the genetic algorithm i.e. a given catalyst

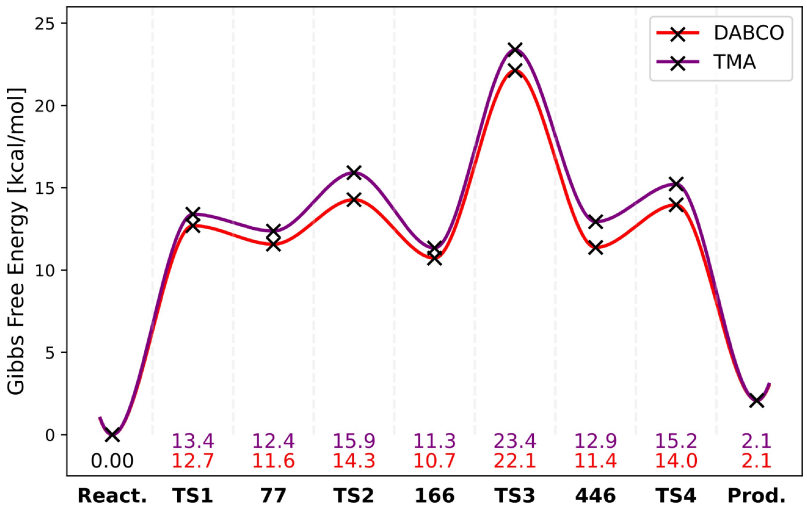



***Figure 5.*** Calculated reaction profile for the mechanism shown in Figure 3c for TMA and DABCO. The TS structures are optimized at the B3LYP/6-31+G(d,p)/SMD(methanol) level of theory using Gaussian 16.[25]

is optimized to both lower the activation energy for the target reaction while simultaneously raising it for side reactions.

From the reaction network of TMA, we see that any side reaction needs to go through intermediate **150**, **176**, **216** or **553** (Figure 4). The primary path to intermediates **150**, **176** and **216** originates from intermediate **77** while intermediate **553** is formed from intermediate **166** (Figure 6). We find neither new meta-MD nor proton-transfer reactions from intermediate **216**—only the back-reaction to intermediate **77** is observed. With a Gibbs free energy 14 kcal/mol higher than the reactants, the system will also not get stuck here meaning that this side-reaction is deemed unimportant. Both intermediates **150** and **176** have high Gibbs free energies close to the activation energy of the MBH reaction; 23 and 20 kcal/mol, respectively. However, if the barriers are low enough they could act as "gateways" to products with lower reaction energies than the MBH product (Figure 4). Thus we use a previously published method for finding TS guess structures based on the same kind of biasing potentials used in the meta-MD to find TSs for the reaction between intermediates **77** and **150** and between intermediates **77** and **176**.[11,12,23,29] For the **77**→**176** reaction we find a barrier of 29 kcal/mol. When searching for the **77**→**150** TS, we instead find the TS for the **176**→**150** reaction with a barrier of 25 kcal/mol. Looking at intermediates **150** and **177**, it makes sense that the **77**→**150** goes through intermediate **176**. Thus, we expect a 29 kcal/mol barrier for getting to either intermediates. Compared to a barrier of 23 kcal/mol for the MBH reaction we do not deem these side-reactions important and ignore them in the GA scoring function. Finally, intermediate **553** has a Gibbs free energy 24 kcal/mol higher than the reactant system. Since the intermediate Gibbs free energy is already higher than the MBH activation energy, this side-reaction is also deemed unimportant.

In this case, no side-reactions of the network are deemed important and we score the catalyst solely based on the reactant to **TS3** GFN2-xTB electronic energy barrier as done in the original study.[5] The five GA searches done by Seumer et al. are now repeated with the same starting populations (100 generations, a population size of 100 and a mutation rate of 50 %, details can be found in Ref [5]). The only difference is, that while the original study scored the catalysts based on a GFN2-xTB TS template from a known catalyst (DABCO) we score the catalysts based on a GFN2-xTB TS template from the simple and generic catalyst template, TMA. As the GA optimization has several stochastic elements, we do not expect to find the exact same final populations. However, molecules containing an azetidine ring are still dominating the final populations and in fact 20 of the catalysts are present in both this and the original study (Figure S6). In fact the similarity between the final populations (evaluated by the average Tanimoto similarity for Morgan extended-connectivity fingerprints with diameter 4 and length 1024)[34] between the five GA runs of Seumer et al.[5] ($0.34 \pm 0.04$) is statistically identical to the average similarity of the final populations from the DABCO and TMA template, respectively ($0.29 \pm 0.05$). We calculate the TS and barrier for the rate-determining step (**TS3**) for three of them (m5, m6 and M10) at the DFT level and find an activation energy $\approx$2 kcal/mol lower than DABCO ($\approx$20 kcal/mol vs. 22.1 kcal/mol) for all of them. The reductions are similar to what was found by Seumer et al. for the two molecules calculated at the DFT level (1.7 and 2.4 kcal/mol).[5] For M10 we compute barriers for the remaining steps in the reaction path and confirm that **TS3** still corresponds to the rate-determining step (Figure 7). This shows that the GA catalyst optimization procedure proposed by Seumer et al. is not dependent on the availability of an already known catalyst (e.g. DABCO).

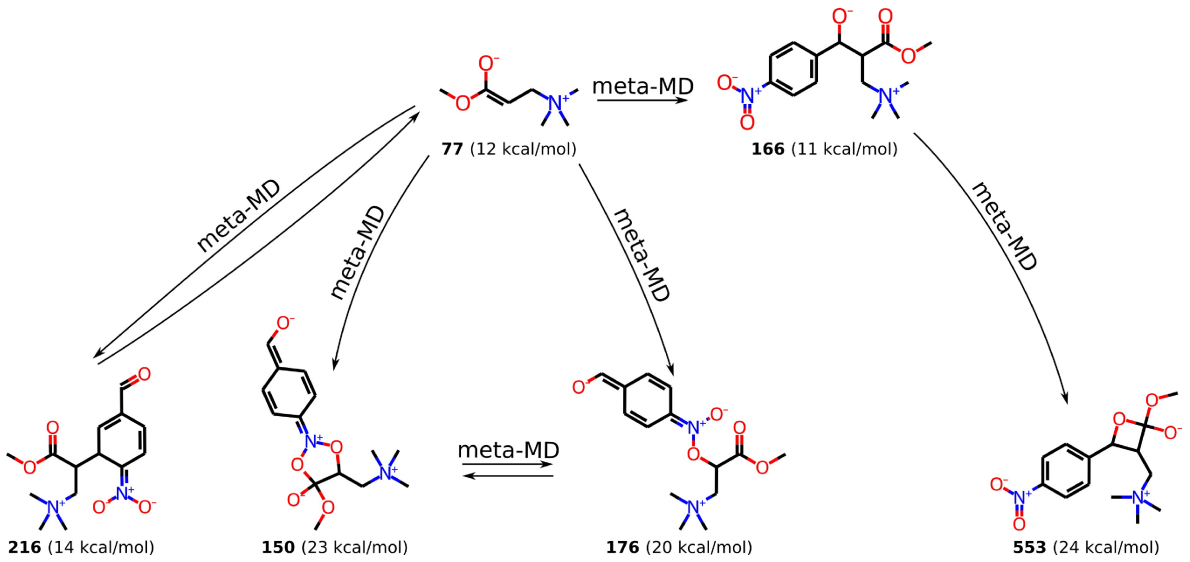



***Figure 6.*** Possibly relevant side reactions extracted from the TMA reaction network in Figure 4.

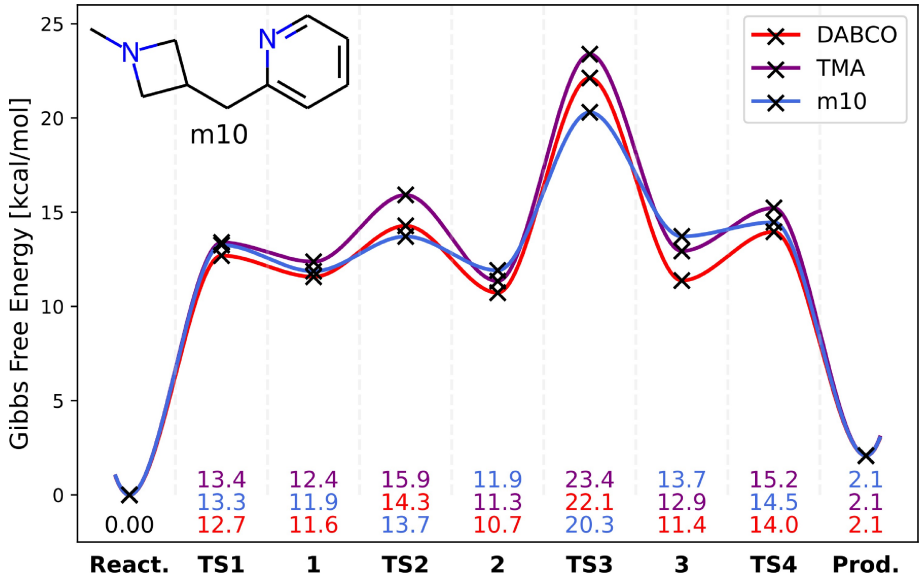



***Figure 7.*** Calculated reaction profile for the MBH mechanism for TMA, DABCO and m10. The TSs are optimized at the B3LYP/6-31+G(d,p)/SMD(methanol) level of theory using Gaussian 16.[25]

Rather, one can get to catalysts performing equally well on the DFT level starting from a bad catalyst (e.g. TMA).

## Conclusion

Full de novo catalyst discovery requires both automated lead identification and optimization. We have previously shown that the latter step can be done efficiently using a graph-based genetic algorithm, given the TS structure of the rate determining step in the mechanism.[5] Here, we present a method for the first step.

We use the meta-MD method developed by Grimme to automatically determine reaction networks for possible catalyst candidates.[11,23,29] From the reaction networks we extract the presence/absence of catalytic activity as well as a mechanism for the catalytic cycle. We find that the TS of the rate determining step for the found catalyst candidate can be used for catalyst optimization (lead optimization). None of these steps are specific to the MBH reaction and the method should be applicable to a wide range of reaction types.

We demonstrate the method by using it to rediscover tertiary amines and phosphines as catalysts for the MBH reaction. Building reaction networks for 11 possible catalysts candidates representing different functional groups (amines, phosphines, sulfides, alkenes, acids, bases and carbonyl groups), tertiary amines and phosphines clearly stand out as the most promising catalyst candidates. Furthermore, the reaction networks are used to extract information about possible side reactions; in this case we identified no relevant sidereactions that are competitive with the desired catalytic mechanism. The tertiary amine template, TMA, identified by the screening of possible catalyst candidates is used as a starting point for the GA proposed by Seumer et al. and we found that the optimized catalyst candidates performed similarly to the experimentally validated **M10** catalyst suggested by Seumer et al. at the B3LYP/6-31+G(d,p) level of theory with methanol modelled as a continuum solvent using the SMD model.[5]

A method for catalyst candidate identification requires that screening of possible candidates is practically possible, meaning that growing the reaction networks must be fast. We achieve this by focusing on screening intermediate Gibbs free energies rather than searching for TSs. Furthermore, the meta-MD simulations are performed at the fast semi-empirical level of theory GFN2-xTB.[24]

The combination of this method for catalyst candidate discovery with the GA-based catalyst optimization method by Seumer et al. results in an algorithm for truly de novo catalyst discovery.

We note that the MBH catalytic cycle is in many ways an ideal case for the proposed method. First of all, the family of molecules catalyzing the reaction (tertiary amines and phosphines) is quite simple in terms of functionality. One can easily imagine reactions requiring a much more complicated functionality of the catalyst. This affects the size and nature of the library of possible catalyst candidates, we need to test. It is very likely that some kind of iterative procedure in terms of updating the library will be necessary to implement in order to find something good enough that the genetic algorithm can take over. Second, this catalytic cycle is relatively simple (four steps in the mechanism) and no additives, acids, bases etc. is needed. Clearly, testing several conditions for each catalyst template will increase the cost significantly. However, the cost is unlikely to become prohibitive and even several of months of comput-

ing is an acceptable investment of effort for discovering a novel catalyst candidate.

Another thing to note is that organometallic catalysts are very important in the field of catalysis. How this method works for transition metal compounds is an important yet still unanswered question that we will continue working on providing an answer for.

Thus while this method by no means represents the final word for method development within de novo catalyst discovery, it is an important step in the right direction.

## Acknowledgements

This work was supported by Novo Nordisk Fonden via grant number NNF20OC0064104.

## Conflict of Interest

The authors declare no conflict of interest.

## Data Availability Statement

The code and data resulting from this study can be found here https://github.com/jensengroup/MBH_CatalystDiscovery, and https://sid.erda.dk/sharelink/C4RVLJdhC5, respectively

**Keywords:** catalysis · organocatalysis · de novo reaction discovery

[1] A. Nandy, C. Duan, C. Goffinet, H. J. Kulik, *JACS Au* **2022**, *2*, 1200–1213.
[2] H. H. Cramer, S. Das, M. D. Wodrich, C. Corminboeuf, C. Werlé, W. Leitner, *Chem. Sci.* **2023**, *14*, 2799–2807.
[3] S. Das, R. C. Turnell-Ritson, P. J. Dyson, C. Corminboeuf, *Angew. Chem. Int. Ed.* **2022**, *61*, e202208987.
[4] J. H. Jensen, *Chem. Sci.* **2019**, *10*, 3567–3572.
[5] J. Seumer, J. Kirschner Solberg Hansen, M. Brondsted Nielsen, J. H. Jensen, *Angew. Chem. Int. Ed.* **2023**, *62*, e202218565.
[6] P. M. Zimmerman, *J. Comput. Chem.* **2013**, *34*, 1385–1392.
[7] L.-P. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande, T. J. Martínez, *Nat. Chem.* **2014**, *6*, 1044–1048.
[8] S. Habershon, *J. Chem. Theory Comput.* **2016**, *12*, 1786–1798.
[9] I. Ismail, R. Chantreau Majerus, S. Habershon, *J. Phys. Chem. A* **2022**, *126*, 7051–7069.
[10] Y. Kim, J. W. Kim, Z. Kim, W. Y. Kim, *Chem. Sci.* **2018**, *9*, 825–835.
[11] M. Koerstz, M. H. Rasmussen, J. H. Jensen, *Sci. Post. Chem.* **2021**, *1*, 003.
[12] M. H. Rasmussen, J. H. Jensen, *PeerJ. Phys. Chem.* **2022**, *4*, e22.
[13] C. Lavigne, G. Gomes, R. Pollice, A. Aspuru-Guzik, *Chem. Sci.* **2022**, *13*, 13857–13871.
[14] E. Martínez-Núñez, G. L. Barnes, D. R. Glowacki, S. Kopec, D. Peláez, A. Rodríguez, R. Rodríguez-Fernández, R. J. Shannon, J. J. P. Stewart, P. G. Tahoces, S. A. Vazquez, *J. Comput. Chem.* **2021**, *42*, 2036–2048.
[15] R. J. Shannon, E. Martínez-Núñez, D. V. Shalashilin, D. R. Glowacki, *J. Chem. Theory Comput.* **2021**, *17*, 4901–4912.
[16] J. P. Unsleber, S. A. Grimmel, M. Reiher, *J. Chem. Theory Comput.* **2022**, *18*, 5393–5409.
[17] J. P. Unsleber, H. Liu, L. Talirz, T. Weymuth, M. Mörchen, A. Grofe, D. Wecker, C. J. Stein, A. Panyala, B. Peng, K. Kowalski, M. Troyer, M. Reiher, *J. Chem. Phys.* **2023**, *158*, 084803.
[18] S. Maeda, Y. Harabuchi, *WIREs Comput. Mol. Sci.* **2021**, *11*, e1538.
[19] R. F. Heck, D. S. Breslow, *J. Am. Chem. Soc.* **1961**, *83*, 4023–4027.
[20] S. Maeda, K. Morokuma, *J. Chem. Theory Comput.* **2012**, *8*, 380–385.
[21] S. Habershon, *J. Chem. Phys.* **2015**, *143*, 094106.
[22] J. A. Varela, S. A. Vázquez, E. Martínez-Núñez, *Chem. Sci.* **2017**, *8*, 3843–3851.
[23] S. Grimme, *J. Chem. Theory Comput.* **2019**, *15*, 2847–2862.
[24] C. Bannwarth, S. Ehlert, S. Grimme, *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.
[25] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, Gaussiañ16 Revision A.03, Gaussian Inc. Wallingford CT, **2016**.
[26] G. Henkelman, B. P. Uberuaga, H. Jónsson, *J. Chem. Phys.* **2000**, *113*, 9901–9904.
[27] P. Zimmerman, *J. Chem. Theory Comput.* **2013**, *9*, 3043–3050.
[28] S. Maeda, T. Taketsugu, K. Morokuma, *J. Comput. Chem.* **2014**, *35*, 166–173.
[29] M. H. Rasmussen, J. H. Jensen, *PeerJ. Phys. Chem.* **2020**, *2*, e15.
[30] T. A. Young, J. J. Silcock, A. J. Sterling, F. Duarte, *Angew. Chem. Int. Ed.* **2021**, *60*, 4266–4274.
[31] R. E. Plata, D. A. Singleton, *J. Am. Chem. Soc.* **2015**, *137*, 3811–3826.
[32] Z. Liu, C. Patel, J. N. Harvey, R. B. Sunoj, *Phys. Chem. Chem. Phys.* **2017**, *19*, 30647–30657.
[33] R. Robiette, V. K. Aggarwal, J. N. Harvey, *J. Am. Chem. Soc.* **2007**, *129*, 15513–15525.
[34] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

Manuscript received: July 25, 2023
Accepted manuscript online: October 13, 2023
Version of record online: October 31, 2023

# BIBLIOGRAPHY


[1] Meeri Kim. *The Ubiquity of Catalysis*. en. https://www.lindau-nobel.org/blog-the-ubiquity-of-catalysis/. Accessed: 2024-5-12. 2022.

[2] Carl D Millholland. *Industrial Uses Of Catalysts*. en. https://www.thermofisher.com/blog/materials/characterizing-the-effectiveness-of-industrial-catalysts/. Accessed: 2024-5-12. 2021.

[3] Leah Burrows. *Towards more efficient catalysts*. en. https://seas.harvard.edu/news/2024/02/towards-more-efficient-catalysts. Accessed: 2024-5-12. 2024.

[4] John N Armor. "A history of industrial catalysis." en. In: *Catalysis today* 163.1 (2011), pp. 3–9. ISSN: 0920-5861,1873-4308. DOI: 10.1016/j.cattod.2009.11.019.

[5] A H Hoveyda. "Catalyst discovery through combinatorial chemistry." en. In: *Chemistry & biology* 5.8 (Aug. 1998), R187–91. ISSN: 1074-5521,1879-1301. DOI: 10.1016/s1074-5521(98)90155-7.

[6] Varinder K Aggarwal, Ingo Emme, and Sarah Y Fulford. "Correlation between pK(a) and reactivity of quinuclidine-based catalysts in the Baylis-Hillman reaction: discovery of quinuclidine as optimum catalyst leading to substantial enhancement of scope." en. In: *The Journal of organic chemistry* 68.3 (2003), pp. 692–700. ISSN: 0022-3263. DOI: 10.1021/jo026671s.

[7] K N Houk and Fang Liu. "Holy Grails for Computational Organic Chemistry and Biochemistry." en. In: *Accounts of chemical research* 50.3 (2017), pp. 539–543. ISSN: 0001-4842,1520-4898. DOI: 10.1021/acs.accounts.6b00532.

[8] Hannes Kneiding, Ainara Nova, and David Balcells. "Directional multiobjective optimization of metal complexes at the billion-system scale." en. In: *Nature Computational Science* 4.4 (2024), pp. 263–273. ISSN: 2662-8457,2662-8457. DOI: 10.1038/s43588-024-00616-5.

[9] Marco Foscato, Vishwesh Venkatraman, and Vidar R Jensen. "DENOPTIM: Software for Computational de Novo Design of Organic and Inorganic Molecules." en. In: *Journal of chemical information and modeling* 59.10 (2019), pp. 4077–4082. ISSN: 1549-9596,1549-960X. DOI: 10.1021/acs.jcim.9b00516.

[10] Marco Foscato and Vidar R Jensen. "Automated in Silico Design of Homogeneous Catalysts." In: *ACS catalysis* 10.3 (2020), pp. 2354–2377. DOI: 10.1021/acscatal.9b04952.

[11] Marco Foscato, Giovanni Occhipinti, Sondre H Hopen Eliasson, and Vidar R Jensen. "Automated de Novo design of olefin metathesis catalysts: Computational and experimental analysis of a simple thermodynamic design criterion." en. In: *Journal of chemical information and modeling* 64.2 (2024), pp. 412–424. ISSN: 1549-9596,1549-960X. DOI: 10.1021/acs.jcim.3c01649.

[12] Simone Gallarati, Puck van Gerwen, Ruben Laplaza, Lucien Brey, Alexander Makaveev, and Clemence Corminboeuf. "A genetic optimization strategy with generality in asymmetric organocatalysis as a primary target." en. In: *Chemical science (Royal Society of Chemistry: 2010)* (2024). ISSN: 2041-6539,2041-6520. DOI: 10.1039/d3sc06208b.

[13] Ruben Laplaza, Simone Gallarati, and Clemence Corminboeuf. "Genetic optimization of homogeneous catalysts." en. In: *Chemistry-Methods* 2.6 (June 2022). ISSN: 2628-9725,2628-9725. DOI: 10.1002/cmtd.202100107.

[14] François Cornet, Bardi Benediktsson, Bjarke Hastrup, Mikkel N Schmidt, and Arghya Bhowmik. "Om-Diff: Inverse-design of organometallic catalysts with guided equivariant denoising diffusion." en. In: *ChemRxiv* (2024). DOI: 10.26434/chemrxiv-2024-882hh.

[15] Oliver Schilter, Alain Vaucher, Philippe Schwaller, and Teodoro Laino. "Designing catalysts with deep generative models and computational data. A case study for Suzuki cross coupling reactions." en. In: *Digital discovery* 2.3 (2023), pp. 728–735. ISSN: 2635-098X. DOI: 10.1039/d2dd00125j.

[16] Anthony R Rosales et al. "Rapid virtual screening of enantioselective catalysts using CatVS." en. In: *Nature catalysis* 2.1 (2018), pp. 41–45. ISSN: 2520-1158,2520-1158. DOI: 10.1038/s41929-018-0193-3.

[17] Wenhao Gao and Connor W Coley. "The Synthesizability of Molecules Proposed by Generative Models." en. In: *Journal of chemical information and modeling* 60.12 (2020), pp. 5714–5723. ISSN: 1549-9596,1549-960X. DOI: 10.1021/acs.jcim.0c00174.

[18] Jon Paul Janet, Lydia Chan, and Heather J Kulik. "Accelerating chemical discovery with machine learning: Simulated evolution of spin crossover complexes with an artificial neural network." en. In: *The journal of physical chemistry letters* 9.5 (2018), pp. 1064–1071. ISSN: 1948-7185. DOI: 10.1021/acs.jpclett.8b00170.

[19] Mads Koerstz, Anders S Christensen, Kurt V Mikkelsen, Mogens Brøndsted Nielsen, and Jan H Jensen. "High throughput virtual screening of 230 billion molecular solar heat battery candidates." en. In: *PeerJ physical chemistry* 3.e16 (2021), e16. ISSN: 2689-7733. DOI: 10.7717/peerj-pchem.16.

[20] Nathan Brown, Ben McKay, François Gilardoni, and Johann Gasteiger. "A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules." en. In: *Journal of chemical information and computer sciences* 44.3 (2004), pp. 1079–1087. ISSN: 0095-2338,1520-5142. DOI: 10.1021/ci034290p.

[21] Jan H Jensen. "A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space." en. In: *Chemical science* 10.12 (2019), pp. 3567–3572. ISSN: 2041-6520. DOI: 10.1039/c8sc05372c.

[22] Greg Landrum et al. *rdkit/rdkit: 2021_09_4 (Q3 2021) Release*. 2022. DOI: 10.5281/zenodo.5835217.

[23] Peter Ertl and Ansgar Schuffenhauer. "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions." en. In: *Journal of cheminformatics* 1.1 (2009), p. 8. ISSN: 1758-2946. DOI: 10.1186/1758-2946-1-8.

[24] Michael Busch, Matthew D Wodrich, and Clémence Corminboeuf. "Linear scaling relationships and volcano plots in homogeneous catalysis - revisiting the Suzuki reaction." en. In: *Chemical science* 6.12 (2015), pp. 6754–6761. ISSN: 2041-6520. DOI: 10.1039/c5sc02910d.

[25] Matthew D Wodrich, Michael Busch, and Clémence Corminboeuf. "Accessing and predicting the kinetic profiles of homogeneous catalysts from volcano plots." en. In: *Chemical science* 7.9 (2016), pp. 5723–5735. ISSN: 2041-6520. DOI: 10.1039/c6sc01660j.

[26] Matthew D Wodrich, Boodsarin Sawatlon, Michael Busch, and Clemence Corminboeuf. "The genesis of molecular volcano plots." en. In: *Accounts of chemical research* 54.5 (2021), pp. 1107–1117. ISSN: 0001-4842,1520-4898. DOI: 10.1021/acs.accounts.0c00857.

[27] Benjamin Meyer, Boodsarin Sawatlon, Stefan Heinen, O Anatole von Lilienfeld, and Clémence Corminboeuf. "Machine learning meets volcano plots: computational discovery of cross-coupling catalysts." en. In: *Chemical science* 9.35 (2018), pp. 7069–7077. ISSN: 2041-6520,2041-6539. DOI: 10.1039/c8sc01949e.

[28] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. "GFN2-xTB-An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions." en. In: *Journal of chemical theory and computation* 15.3 (2019), pp. 1652–1671. ISSN: 1549-9618,1549-9626. DOI: 10.1021/acs.jctc.8b01176.

[29] *PostEra, Medicinal Chemistry Powered by Machine Learning*. en. https://postera.ai/manifold/. Accessed: 2023-7-18.

[30] Maria H Rasmussen, Julius Seumer, and Jan H Jensen. "Toward De Novo catalyst discovery: Fast identification of new catalyst candidates for alcohol-mediated Morita-Baylis-Hillman reactions." en. In: *Angewandte Chemie, International Edition* 62.49 (2023), e202310580. DOI: 10.1002/anie.202310580.

[31] Chenru Duan, Daniel B K Chu, Aditya Nandy, and Heather J Kulik. "Two wrongs can make a right: A transfer learning approach for chemical discovery with chemical accuracy." In: *arXiv [physics.chem-ph]* (2022). arXiv: 2201.04243 [physics.chem-ph].

[32] *GA_DBA_Sucrose: Genetic algorithm for design of molecular tweezers based on phenylboronic acids for selective detection of Sucrose*. en. https://github.com/sugaralc/GA_DBA_Sucrose. Accessed: 2024-5-10.

[33] *chem-william/mbh_catalyst_ga: Graph-based genetic algorithm*. en. https://github.com/chem-william/mbh_catalyst_ga. Accessed: 2024-5-10.

[34] Magnus Strandgaard, Julius Seumer, Bardi Benediktsson, Arghya Bhowmik, Tejs Vegge, and Jan H Jensen. "Genetic algorithm-based re-optimization of the Schrock catalyst for dinitrogen fixation." en. In: *PeerJ physical chemistry* 5.e30 (2023), e30. ISSN: 2689-7733. DOI: 10.7717/peerj-pchem.30.

[35] Magnus Strandgaard, Julius Seumer, and Jan Halborg Jensen. "Discovery of molybdenum based nitrogen fixation catalysts with genetic algorithms." en. In: *ChemRxiv* (2024). DOI: 10.26434/chemrxiv-2024-p835f-v2.

[36] Dominik B Orlowski. *CarbonCaptureCatalystGA*. en. https://github.com/Tidodon/CarbonCaptureCatalystGA. Accessed: 2024-5-10.

[37] Tapas Kumar Achar, Xinglong Zhang, Rahul Mondal, M S Shanavas, Siddhartha Maiti, Sabyasachi Maity, Nityananda Pal, Robert S Paton, and Debabrata Maiti. "Palladium-catalyzed directed meta-selective C-H allylation of Arenes: Unactivated internal olefins as allyl surrogates." en. In: *Angewandte Chemie*

*(International ed. in English)* 58.30 (2019), pp. 10353–10360. ISSN: 1521-3773,1433-7851. DOI: 10.1002/anie.201904608.

[38] Jin-Quan Yu, Ramesh Giri, and Xiao Chen. "Sigma-chelation-directed C-H functionalizations using Pd(II) and Cu(II) catalysts: regioselectivity, stereoselectivity and catalytic turnover." en. In: *Organic & biomolecular chemistry* 4.22 (2006), pp. 4041–4047. ISSN: 1477-0539,1477-0520. DOI: 10.1039/b611094k.

[39] Fumitoshi Kakiuchi and Naoto Chatani. "Catalytic methods for C-H bond functionalization: Application in organic synthesis." en. In: *Advanced synthesis & catalysis* 345.9-10 (2003), pp. 1077–1101. ISSN: 1615-4169,1615-4150. DOI: 10.1002/adsc.200303094.

[40] Jimmy C Kromann, Jan H Jensen, Monika Kruszyk, Mikkel Jessing, and Morten Jørgensen. "Fast and accurate prediction of the regioselectivity of electrophilic aromatic substitution reactions." en. In: *Chemical science (Royal Society of Chemistry: 2010)* 9.3 (2018), pp. 660–665. ISSN: 2041-6539,2041-6520. DOI: 10.1039/c7sc04156j.

[41] Nicolai Ree, Andreas H Göller, and Jan H Jensen. "Automated quantum chemistry for estimating nucleophilicity and electrophilicity with applications to retrosynthesis and covalent inhibitors." en. In: *Digital discovery* 3.2 (2024), pp. 347–354. ISSN: 2635-098X. DOI: 10.1039/d3dd00224a.

[42] Magnus Liljenberg, Tore Brinck, Björn Herschend, Tobias Rein, Simone Tomasi, and Mats Svensson. "Predicting regioselectivity in nucleophilic aromatic substitution." en. In: *The Journal of organic chemistry* 77.7 (2012), pp. 3262–3269. ISSN: 0022-3263,1520-6904. DOI: 10.1021/jo202569n.

[43] Anna Tomberg, Michael Éric Muratore, Magnus Jan Johansson, Ina Terstiege, Christian Sköld, and Per-Ola Norrby. "Relative strength of common directing groups in palladium-catalyzed aromatic C-H activation." en. In: *iScience* 20 (2019), pp. 373–391. ISSN: 2589-0042. DOI: 10.1016/j.isci.2019.09.035.

[44] Liwei Cao, Mikhail Kabeshov, Steven V Ley, and Alexei A Lapkin. "In silico rationalisation of selectivity and reactivity in Pd-catalysed C-H activation reactions." en. In: *Beilstein journal of organic chemistry* 16.1 (2020), pp. 1465–1475. ISSN: 1860-5397,2195-951X. DOI: 10.3762/bjoc.16.122.

[45] Marc Lafrance, Christopher N Rowley, Tom K Woo, and Keith Fagnou. "Catalytic intermolecular direct arylation of perfluorobenzenes." en. In: *Journal of the American Chemical Society* 128.27 (2006), pp. 8754–8756. ISSN: 0002-7863,1520-5126. DOI: 10.1021/ja062509l.

[46] Serge I Gorelsky, David Lapointe, and Keith Fagnou. "Analysis of the concerted metalation-deprotonation mechanism in palladium-catalyzed direct arylation across a broad range of aromatic substrates." en. In: *Journal of the American Chemical Society* 130.33 (2008), pp. 10848–10849. ISSN: 0002-7863,1520-5126. DOI: 10.1021/ja802533u.

[47] Tim Cernak, Kevin D Dykstra, Sriram Tyagarajan, Petr Vachal, and Shane W Krska. "The medicinal chemist's toolbox for late stage functionalization of drug-like molecules." en. In: *Chemical Society reviews* 45.3 (2016), pp. 546–576. ISSN: 1460-4744,0306-0012. DOI: 10.1039/c5cs00628g.