THE WONDERS OF DIGITAL CATALYSIS

Bridging Chemistry and Machine Learning for Homogeneous Catalyst Design

ADARSH V. KALIKADIEN

The Wonders of Digital Catalysis

Bridging Chemistry and Machine Learning for Homogeneous Catalyst Design

The Wonders of Digital Catalysis

Bridging Chemistry and Machine Learning for Homogeneous Catalyst Design

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, prof. dr. ir. H. Bijl,
chair of the Board for Doctorates
to be defended publicly on
Monday 9 March 2026 at 12:30 o'clock

by

Adarsh Varun KALIKADIEN

Master of Science in Chemical Engineering, Delft University of Technology, The Netherlands
born in 's-Gravenhage, The Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. E.A. Pidko,	Delft University of Technology, *promotor*
Em. prof. dr. B. Dam,	Delft University of Technology, *promotor*
Dr. L. Lefort,	Johnson & Johnson USA, *external advisor*

Independent members:

Prof. dr. R. Eelkema	Delft University of Technology
Prof. dr. F.C. Grozema	Delft University of Technology
Prof. dr. ir. B. Ensing	University of Amsterdam
Dr. M. Podewitz	Technische Universität Wien
Prof. dr. ir. C.A. Ramirez-Ramirez	Delft University of Technology, reserve member

The research described in this dissertation has been carried out in the Inorganic Systems Engineering group at Delft University of Technology, the Netherlands. Financial support was provided by Janssen Pharmaceutica NV, a Johnson & Johnson company. High-performance computing resources were provided by the NWO Domein Exacte en Natuurwetenschappen at the national supercomputer Snellius.

Printed by: Proefschriftspecialist

Cover by: Fatih Paksoy

ISBN 978-94-93483-97-2

An electronic copy of this dissertation is available at
https://repository.tudelft.nl/.

Before enlightenment, chop wood, carry water.
After enlightenment, chop wood, carry water.

Contents

SUMMARY

Catalysis lies at the heart of modern society: from producing fuels and fertilizers to manufacturing pharmaceuticals and materials, it enables the chemical transformations that sustain our daily lives. Among the different forms of catalysis, homogeneous catalysis, where well-define molecular complexes drive the production of molecular products, plays a central role in both fundamental research and industrial applications. Yet, the discovery and optimization of catalysts remain resource-intensive, relying heavily on serendipity. The design of transition-metal based homogeneous catalysts remains a central challenge in modern chemistry. While recent advances in artificial intelligence have demonstrated transformative potential across domains such as natural language processing and image generation, their application to molecular design and catalysis has proven more limited. This dissertation explores the integration of high-throughput experimentation, computational chemistry, automation, and machine learning for *in silico* methodologies aimed at rational design of transition-metal based catalysts. Across eight Chapters, key challenges are addressed in the generation of descriptors, digital representations for machine learning, conformational and configurational flexiblity of ligands and practical examples of machine learning modeling in data-driven catalysis.

Chapter 1 lays the foundation by providing a perspective on the role of computational methods in homogeneous catalyst design. While computational tools and machine learning have become increasingly important, their application in catalysis is hindered by the scarcity of high-quality data, complex reaction mechanisms, and limited automation. We review existing tools for automated structure generation, highlighting the lack of a universally applicable workflow that integrates structure generation, descriptor computation, and analysis for transition-metal complexes. To address this, we introduced the Open Bidentate Ligand eXplorer (OBeLiX), a Python package designed to enable modular and automated high-throughput exploration of catalytic chemical space. This chapter concludes that automation, modularity, and the integration first principles of chemistry and catalysis into modern workflows is essential for rational catalyst design and democratizing data-driven approaches in catalysis.

Chapter 2 investigates the potential of machine learning to accelerate catalyst selection in Rh-catalyzed asymmetric hydrogenation of olefins. Using high-throughput experimentation, a large and reliable dataset of 192 chiral ligands and over 3500 reactions was generated, providing a robust foundation for modeling. Alongside, OBeLiX was developed to compute automated DFT-based descriptors, which were benchmarked against simpler representations in both out-of-domain and in-domain prediction tasks. The results revealed that predictive performance in out-of-domain tasks was largely driven by ligand differentiation, with limited benefits from computationally intensive descriptors, while in-domain applications showed modest success for conversion but persistent challenges for enantioselectivity. These findings revealed the limitations of general descriptors

and highlighted the critical role of dataset diversity, mechanistic insight and further improving the representation of catalyst structures in improving machine learning model performance.

Chapter 3 explored how substrate-specific interactions influence digital catalyst structures and derived descriptors. Using 11 Rh-bisphosphine complexes from Chapter 2, a comprehensive conformer ensemble analysis compared substrate-bound structures with generic precatalyst structures. The results revealed that while the substrate itself is relatively rigid, its inclusion induces substantial ligand flexibility, expanding the conformational landscape up to fivefold compared to the precatalyst. This flexibility significantly influences electronic descriptors such as the NBO charge on Rh and HOMO-LUMO gap, which vary with substrate coordination and cannot be captured by precatalyst-only models. The findings challenge the conventional reliance on simplified catalyst representations and emphasize the importance of ensemble-based and substrate-specific descriptors. Ultimately, the study underscores that neglecting catalyst-substrate interactions risks overlooking critical features for predictive modeling of reactivity and selectivity.

Building on this, Chapter 4 develops a generalizable workflow for automated selection of conformers from semi-empirical ensembles to enable accurate and computationally feasible representations of flexible transition-metal catalysts. Using 24 Rh-bisphosphine precatalysts from Chapter 2, CREST-generated conformers were systematically compared to their DFT-optimized counterparts, evaluating filtering strategies based on geometric descriptors, PCA of ensemble properties, and RMSD- and energy-based clustering via DBSCAN. The study revealed that CREST overestimates ligand flexibility and energy-based selection poorly correlates with DFT minima. Geometry- and RMSD-based approaches improved mapping to the DFT ensemble, but DBSCAN clustering provided the most effective balance between accuracy and computational cost, eliminating redundancies while preserving key configurations. The method remained robust when applied to more flexible substrate-bound complexes wich were introduced in Chapter 3, demonstrating its general applicability across different catalyst structures. Overall, this work establishes a computationally accessible and automated strategy for generating conformer ensembles suitable for high-throughput *in silico* screening of transition-metal catalysts.

Chapter 5 investigates the impact of stereoisomerism and configurational fluxionality on *in silico* high-throughput screening of octahedral TM complexes. Automated workflows generated ensembles of ligand configurations for 87 bisphosphine ligands coordinated to Ir(III), Ru(II), and Mn(I) centers, yielding 908 complexes. DFT calculations revealed distinct metal-dependent behavior: Ir(III) complexes favored a single configuration, while Mn(I)- and Ru(II)-complexes exhibited significant fluxionality with multiple configurations within 10 kJ/mol. Linear regression and descriptor analysis showed that global descriptors such as bite angle and HOMO-LUMO gap are transferable across configurations and metals, whereas local steric descriptors lack such transferability. Machine learning models successfully classified ligand configurations, but struggled to predict stability across metal centers, highlighting the need for improved descriptors of the first coordination sphere. Overall, the study demonstrates that ignoring stereoisomerism risks incomplete sampling of chemical space and underrepresentation of key catalyst features, emphasizing the importance of exhaustive configurational exploration in

screening workflows.

Chapter 6 builds on the findings from the previous Chapters to utilize our computational workflows to study the combination of high-throughput experimentation and computational chemistry combined with machine learning to guide ligand optimization for Ni-catalyzed arylketone formation. A chemically diverse dataset of monophosphine and bisphosphine ligands was generated, with descriptors explicitly incorporating conformational flexibility using our findings from Chapter 3 and 4. Principal component analysis identified key steric and electronic features governing variance within the ligand space, enabling rationalization of ligand selection. Machine learning models trained on experimentally validated ligands successfully classified catalytic activity and predicted performance of previously untested ligands. A transfer learning strategy, leveraging descriptors from the Rh-based model structures introduced in Chapter 2, provided a proof of concept for extending predictions across catalytic systems. Overall, this work demonstrates that integrating high-throughput experimentation and transfer learning offers a promising framework for accelerated rational catalyst design.

Chapter 7 addresses the prohibitive costs of exhaustive configurational and conformational exploration with quantum chemical methods in the context of *in silico* high-throughput catalyst screening. Here, we evaluate the Universal Machine-learning Potential for Atoms (UMA), a general-purpose interatomic potential developed by Meta, for transition-metal complexes with diverse bisphosphine ligands. Two datasets were analyzed: conformers of Ni-based catalysts in rigid and flexible model structures described in Chapter 6, and configurational isomers of Ru(II)- and Mn(I)-based complexes described in Chapter 5. UMA enables high accuracy for single-point energy calculations in seconds on consumer-grade GPUs, offering potential for high-throughput catalyst screening. While overall correlations with DFT energies were high, more stringent per-ligand ranking analyses revealed variability in highly fluxional systems where relative energy differences fall within only a few kJ/mol, a regime where even DFT shows limited resolution. These results highlight that machine learning interatomic potentials such as UMA are powerful tools for both rigid and flexible transition-metal complexes, provided their use is coupled with careful validation and chemical expertise. This chapter underscores the growing role of general-purpose MLIPs in accelerating computational catalysis workflows, highlighting both their transformative potential and the importance of domain-informed application.

Finally, Chapter 8 reflects on the broader context and societal relevance of this research, emphasizing the interdisciplinary nature of modern catalyst design. First, the growing importance of interdisciplinarity is illustrated. This is followed by a discussion on the parallels between Open Science and artifical intelligence-driven research, stressing the effort required to make data and workflows accessible and reusable while acknowledging the limitations and hype surrounding artificial intelligence models. Ultimately, a discussion of the role of Academia in providing incremental advances of human knowledge is provided.

SAMENVATTING

Katalyse vormt het hart van de moderne samenleving: van de productie van brandstoffen en meststoffen tot de vervaardiging van farmaceutica en materialen, zij maakt de chemische transformaties mogelijk die ons dagelijks leven ondersteunen. Binnen de verschillende vormen van katalyse neemt homogene katalyse, waarbij goed gedefinieerde moleculaire complexen de omzetting naar moleculaire producten aandrijven, een centrale positie in, zowel in fundamenteel onderzoek als in industriële toepassingen. Toch blijft de ontdekking en optimalisatie van katalysatoren arbeids- en kapitaalintensief en grotendeels afhankelijk van toeval. Het ontwerpen van overgangsmetaal-gebaseerde homogene katalysatoren vormt een blijvende uitdaging in de hedendaagse chemie. Hoewel recente ontwikkelingen in kunstmatige intelligentie een transformerende impact hebben laten zien in domeinen zoals natuurlijke taalverwerking en beeldgeneratie, is hun toepassing in moleculair ontwerp en katalyse vooralsnog beperkt. Dit proefschrift onderzoekt de integratie van high-throughput-experimenten, computationele chemie, automatisering en machine learning voor *in silico*-methoden gericht op het rationeel ontwerp van overgangsmetaal-gebaseerde katalysatoren. Verspreid over acht hoofdstukken worden kernuitdagingen besproken op het gebied van descriptorontwikkeling, digitale representaties voor machine learning, conformationele en configurationele flexibiliteit van liganden, en praktische toepassingen van machine learning in data-gedreven katalyse.

Hoofdstuk 1 legt de basis door een perspectief te bieden op de rol van computationele methoden in het ontwerp van homogene katalysatoren. Hoewel computationele hulpmiddelen en machine learning steeds belangrijker worden, wordt hun toepassing in katalyse beperkt door de schaarste aan hoogwaardige data, complexe reactiemechanismen en beperkte automatisering. Wij bespreken bestaande instrumenten voor geautomatiseerde structuur generatie en benadrukken het ontbreken van een universeel toepasbare workflow die structuur generatie, descriptorberekening en analyse integreert voor overgangsmetaalcomplexen. Om dit te adresseren, introduceerden wij de Open Bidentate Ligand eXplorer (OBeLiX), een Python-pakket dat modulair en geautomatiseerd high-throughputonderzoek naar katalytische chemische ruimte mogelijk maakt. Het hoofdstuk concludeert dat automatisering, modulariteit en de integratie van fundamentele chemische en katalytische principes essentieel zijn voor rationeel katalysatorontwerp en het democratiseren van data-gedreven benaderingen in de katalyse.

Hoofdstuk 2 onderzoekt het potentieel van machine learning om katalysatorselectie te versnellen in Rh-gekatalseerde asymmetrische hydrogenering van olefinen. Met behulp van high-throughput-experimenten werd een omvangrijke en betrouwbare dataset van 192 chirale liganden en meer dan 3500 reacties gegenereerd, die een solide basis vormt voor modellering. Parallel hieraan werd OBeLiX ontwikkeld om geautomatiseerde DFT-gebaseerde descriptoren te berekenen, die werden vergeleken met eenvoudige-

re representaties in zowel out-of-domain- als in-domain-voorspellingstaken. De resultaten toonden aan dat voorspellende prestaties in out-of-domain-taken voornamelijk werden gedreven door ligandenonderscheiding, met beperkte voordelen van computationeel dure descriptoren, terwijl in-domain-toepassingen bescheiden succes lieten zien voor conversie maar blijvende uitdagingen voor enantioselectiviteit. Deze bevindingen benadrukken de beperkingen van algemene descriptoren en onderstrepen de cruciale rol van datasetdiversiteit, mechanistisch inzicht en verbeterde representatie van katalysatorstructuren bij het verbeteren van machine learning-modellen.

Hoofdstuk 3 verkent hoe substraat-specifieke interacties digitale katalysatorstructuren en afgeleide descriptoren beïnvloeden. Met behulp van 11 Rh-bisfosfinecomplexen uit hoofdstuk 2 werd een uitgebreide conformationele ensemble-analyse uitgevoerd waarbij substraat-gebonden structuren werden vergeleken met generieke precatalysatorstructuren. De resultaten lieten zien dat, hoewel het substraat relatief rigide is, de aanwezigheid ervan aanzienlijke ligandenflexibiliteit induceert, waardoor het conformationele landschap tot vijf keer zo groot wordt vergeleken met de prekatalysator. Deze flexibiliteit beïnvloedt elektronische descriptoren zoals de NBO-lading op Rh en de HOMO-LUMO-kloof, die variëren met substraatcoördinatie en niet door modellen op basis van enkel prekatalysatoren kunnen worden vastgelegd. De bevindingen dagen het conventionele vertrouwen in vereenvoudigde katalysatorrepresentaties uit en benadrukken het belang van ensemble-gebaseerde en substraat-specifieke descriptoren. Uiteindelijk toont de studie aan dat het negeren van katalysator-substraatinteracties het risico met zich meebrengt cruciale kenmerken voor voorspellende modellering van reactiviteit en selectiviteit over het hoofd te zien.

Voortbouwend hierop ontwikkelt hoofdstuk 4 een generaliseerbare workflow voor de geautomatiseerde selectie van conformeren uit semi-empirische ensembles om nauwkeurige en computationeel haalbare representaties van flexibele overgangsmetaalkatalysatoren mogelijk te maken. Met behulp van 24 Rh-bisfosfineprekatalysatoren uit hoofdstuk 2 werden door CREST gegenereerde conformeren systematisch vergeleken met hun DFT-geoptimaliseerde tegenhangers. Daarbij werden filterstrategieën geëvalueerd op basis van geometrische descriptoren, PCA van ensemble-eigenschappen, en RMSD- en energiegebaseerde clustering via DBSCAN. De studie liet zien dat CREST de ligandenflexibiliteit overschat en dat energiegebaseerde selectie slecht correleert met DFT-minima. Geometrie- en RMSD-gebaseerde benaderingen verbeterden de mapping naar het DFT ensemble, maar DBSCAN clustering bood de meest effectieve balans tussen nauwkeurigheid en rekenkosten, door redundantie te verwijderen en toch belangrijke configuraties te behouden. De methode bleek robuust bij toepassing op flexibelere substraatgebonden complexen die geïntroduceerd waren in 3, wat de algemene toepasbaarheid over verschillende katalysatorstructuren bevestigt. Al met al biedt dit werk een computationeel toegankelijke en geautomatiseerde strategie voor het genereren van conformerenensembles die geschikt zijn voor high-throughput *in silico*-screening van overgangsmetaalkatalysatoren.

Hoofdstuk 5 onderzoekt de invloed van stereoisomerie en configurationele fluxionaliteit op *in silico* high-throughput-screening van octaëdrische overgangsmetaalcomplexen. Geautomatiseerde workflows genereerden ensembles van ligandenconfiguraties voor 87 bisfosfineliganden gecoördineerd aan Ir(III)-, Ru(II)- en Mn(I)-centra, resulterend in 908

complexen. DFT berekeningen onthulden duidelijk metaalafhankelijk gedrag: Ir(III)-complexen prefereerden èèn configuratie, terwijl Mn(I)- en Ru(II)-complexen aanzienlijke fluxionaliteit vertoonden met meerdere configuraties binnen 10 kJ/mol. Lineaire regressie en descriptoranalyse toonden aan dat globale descriptoren zoals bite angle en HOMO-LUMO-kloof overdraagbaar zijn over configuraties en metalen heen, terwijl lokale sterische descriptoren die overdraagbaarheid missen. Machine learning-modellen classificeerden ligandenconfiguraties succesvol, maar hadden moeite met het voorspellen van stabiliteit over metaalcentra heen, wat de noodzaak benadrukt van verbeterde descriptoren van de eerste coördinatiecirkel. De studie laat zien dat het negeren van stereoisomerie leidt tot onvolledige bemonstering van chemische ruimte en ondervertegenwoordiging van cruciale katalysatoreigenschappen, en benadrukt daarmee het belang van uitputtende configurationele exploratie in screeningsworkflows.

Hoofdstuk 6 bouwt voort op de bevindingen uit de voorgaande hoofdstukken en gebruikt onze computationele workflows om de combinatie van high-throughput-experimenten, computationele chemie en machine learning toe te passen bij de optimalisatie van liganden voor Ni-gekatalyseerde arylketoonsynthese. Een chemisch diverse dataset van mono- en bisfosfineliganden werd gegenereerd, met descriptoren die expliciet conformationele flexibiliteit integreerden, gebruikmakend van onze bevindingen uit Hoofdstuk 3 en 4. Principalecomponentenanalyse identificeerde belangrijke sterische en elektronische kenmerken die de variatie binnen de ligandruimte bepaalden, wat de rationele selectie van liganden mogelijk maakte. Machine learning-modellen, getraind op experimenteel gevalideerde liganden, classificeerden katalytische activiteit succesvol en voorspelden prestaties van voorheen niet-geteste liganden. Een transfer learning-strategie, gebruikmakend van descriptoren van de Rh-gebaseerde modelstructuren uit hoofdstuk 2, bood een proof-of-concept voor het uitbreiden van voorspellingen over verschillende katalytische systemen heen. Dit werk toont aan dat de integratie van high-throughput-experimenten en transfer learning een veelbelovend kader biedt voor versnelde, rationele katalysatorontwikkeling.

Hoofdstuk 7 behandelt de hoge kosten van uitputtende configurationele en conformationele verkenning met kwantumchemische methoden in de context van *in silico* high-throughput-screening. Hier evalueren wij het Universal Machine-learning Potential for Atoms (UMA), een algemeen interatomair potentiaal ontwikkeld door Meta, voor overgangsmetaalcomplexen met diverse bisfosfineliganden. Twee datasets werden geanalyseerd: conformeren van Ni-gebaseerde katalysatoren in rigide en flexibele modelstructuren uit hoofdstuk 6, en configurationele isomeren van Ru(II)- en Mn(I)-complexen uit hoofdstuk 5. UMA maakte het mogelijk om enkelpuntsenergieën met hoge nauwkeurigheid in seconden te berekenen op consumenten-GPUs, wat perspectief biedt voor high-throughput-screening. Hoewel de correlaties met DFT-energieën in het algemeen hoog waren, toonden strengere analyses per ligand variabiliteit aan in sterk fluxionele systemen waarin relatieve energiedifferentiaties slechts enkele kJ/mol bedragen, een regime waarin zelfs DFT beperkte resolutie biedt. Deze resultaten onderstrepen dat machine learning-interatomaire potentialen zoals UMA krachtige hulpmiddelen zijn voor zowel rigide als flexibele overgangsmetaalcomplexen, mits zij gepaard gaan met zorgvuldige validatie en chemische expertise. Dit hoofdstuk benadrukt de groeiende rol van algemene MLIPs in de versnelling van computationele katalyseworkflows, en plaatst hun trans-

formerende potentieel naast de noodzaak van domein-geïnformeerde toepassing.

Tot slot reflecteert hoofdstuk 8 op de bredere context en maatschappelijke relevantie van dit onderzoek, waarbij de interdisciplinaire aard van modern katalysatorontwerp centraal staat. Eerst wordt de toenemende betekenis van interdisciplinariteit geïllustreerd. Vervolgens worden parallellen besproken tussen Open Science en door kunstmatige intelligentie aangedreven onderzoek, met nadruk op de inspanning die vereist is om data en workflows toegankelijk en herbruikbaar te maken, terwijl tegelijkertijd de beperkingen en de hype rond AI-modellen worden erkend. Uiteindelijk wordt gereflecteerd op de rol van de academie in het bieden van incrementele bijdragen aan de menselijke kennisbasis.

1

PAVING THE ROAD TOWARDS AUTOMATED HOMOGENEOUS CATALYST DESIGN

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

This Chapter has been published as: Kalikadien, A. V.; Mirza, A.; Hossaini, A. N.; Sreenithya, A.; Pidko, E. A. *ChemPlusChem* **2024**, e202300702.[1]

1.1. INTRODUCTION

Numerous vital industrial processes rely on homogeneous catalysts. Their efficiency in steering a wide array of chemical transformations gives them a distinct status.[2] These catalytic systems find utility in the synthesis of pharmaceuticals, agrochemicals, bulk chemicals and fine chemicals.[2–7] Metal-ligand complexes are integral to modern chemistry, forming the cornerstone of homogeneous catalysis.[2,7] Despite their ubiquity and versatility, the field of homogeneous catalysis confronts an inherent challenge: the quest for the optimal catalyst.

The vast chemical and reaction space in catalysis poses a challenge to exploration.[8,9] It becomes evident that there are no singular candidates exhibiting unique catalytic performance for our applications. How to find the best performing homogeneous catalyst? The opportunity to perform brute-force exploration of potential candidates is always open. Fortunately, guidance by simple models such as the Bronsted-Evans-Polanyi (BEP) relationship, Hammett parameters and linear scaling relationships were established.[10–14] Together with chemical intuition and heuristics, these principles are often used to guide the screening process. They were originally developed to elucidate the correlation between the rate of a chemical reaction and the thermodynamic properties of the reaction constituents.[15,16] However, catalytic activity/selectivity is not straightforward and origins of high or low performance are often not easily explainable by conventional chemical principles.

In contrast to heterogeneous catalysts, homogeneous catalysts have a better defined structure that can be optimized for performance. For example, a wide range of ligands and modifiers that induce enantioselectivity have been developed for organometallic metal-ligand complexes, enabling high rates and selectivities.[17] Ligand engineering is the common strategy to optimize performance of the catalyst.[18] The modular architecture of transition-metal (TM) coordination complexes paves the way for larger-scale screening, achieved through methods such as fragment-based library construction and subsequent performance optimization.[19–21] Although often guided by mechanistic hypotheses and expert knowledge, ligand engineering has been a primary driver of reaction discovery and catalyst design. This identification of an optimal ligand and subsequent catalyst design is essential to achieve high performance for a desired reaction. However, beyond a specific application, the usability of these ligand engineering approaches becomes contentious. Can they be employed on out-of-sample datasets, e.g. on a new chemical reaction?

Despite the potential of the many automated tools for catalyst design[22], most use cases have been limited to retrospective analyses of the experimental results.[23] Recently, successful examples of computational design directly contributing to experimentally validated catalyst discoveries started emerging. Relevant examples include selective oligo-/polymerization, cross-coupling catalysis, and enantioselective Pauson-Khand reactions.[23–25] Generally, the aim is to optimize experimental targets condensed into a single metric, such as turnover frequency, turnover number, regioselectivity, product selectivity, yield, or enantioselectivity.[25] Rooted in classic principles, these computational catalyst design approaches usually involve featurizing the catalyst structure using chemical descriptors. In a reactivity model it is assumed that an experimental objective (e.g., yield or enantioselectivity) is a function of both

the experimental parameters and computational descriptors of the catalyst structure. This function can then be learned by a statistical model to enable predictions. The success of this approach relies strongly on accurate representations of the catalytic structure.[23,26]

Descriptors are chemically intuitive features of the catalyst structures that are known to be relevant for the catalytic activity. Two classical examples from organometallic chemistry are Tolman's electronic parameter (TEP) and the Tolman cone angle.[27] This cone angle was further adapted into White's solid angle which also takes the ligand's flexbility into account. Later, the derived concept of percent buried volume was introduced.[28,29] Many descriptors can be envisioned and the development of new descriptors, as well as derivatives of classical approaches and rapid calculation methods remains an ongoing endeavor.[30–32] These individual descriptor classes are usually categorized as being electronic, steric, geometric or thermodynamic. For a comprehensive overview of descriptors used in catalyst design, we refer the reader to a review by Durand et al.[30] These descriptors have played a pivotal role in homogeneous catalyst design since its inception. Remarkably, buried volumes, a fundamental descriptor, were already employed back in the '80s to gain insights and predict enantiomeric excesses and predict catalytic outcomes.[33] Another noteworthy example is the bite angle, used to describe the angle between two donor atoms and the metal (L-M-L, in the case of bidentate ligands). It was reported that this bite angle has a large impact on metal-centered reactivity in 1999.[31,34,35] More and more, individual descriptors of the chemical structures have progressed into more refined representations of chemical properties, which can be used to optimize particular objective(s).[26,36] For example, several libraries such as ReaLigands and the Ligand Knowledge Bases have been developed to elucidate ligand effects across a range of representative coordination environments[30,37]. The mapping of these descriptors provides an overview of the ligand space and a direction for more design within different ligand classes[38]. Additionally, less chemically intuitive descriptors such as graph-based representations[39] or derivations thereof[40] have also been applied in the field of TM-based catalysis.

Present day statistical methods used in catalyst design range in complexity from linear explainable models to advanced natural language processing (NLP) models for chemistry.[41,42] The former category is the traditional way automated catalyst design was tackled, while the latter emerged as a powerful tool only in the recent years. This was made possible by the introduction of the transformer architecture for neural networks which allowed processing of inputs of different sizes and interpretation of chemical languages (e.g. SMILES[43], DeepSMILES[44] or SELFIES[45]) in a similar way to human languages.

In the pursuit of understanding complex phenomena, human intuition has often led to the development of simplified and interpretable representations. In the domain of chemistry and cheminformatics, descriptors serve as static and compressed representations of specific chemical structures. Within the realm of catalysis however, every stage involved in constructing such a digital representation of a catalyst complex is susceptible to introducing significant deviations.[23,46] This process typically encompasses various steps, such as the extraction or creation of

the initial complex, density functional theory (DFT) optimization, and descriptor calculation. Together, these constitute the workflow utilized for the creation of a computational and condensed structure representation. Thus, in predictive approaches, the catalyst structure, computational representation, and modeling space are deeply intertwined.[20] It is important to acknowledge that these representations are often still influenced by expert bias, mainly due to the manual generation of the initial chemical structure. This inherent bias can limit the generalizability of published approaches. In addition, this enhances the streetlight effect. This phenomenon refers to the tendency to focus on areas that are well-illuminated, or well-understood, while neglecting less-explored regions, potentially hindering a comprehensive understanding of the chemical space.

To address and mitigate the biases and constraints inherent in the manual structure generation process, the integration of automated structure generation tools is critical in advancing the field of rational catalyst design. Numerous tools have emerged, facilitating the dependable generation of 3D structures. Notable examples include DENOPTIM, Aarontools, MolSimplify, MolAssembler, and the more recent addition of Architector.[47–51] However, the pursuit of an universally applicable computational approach that streamlines all aspects, ranging from structure generation to descriptor computation for organometallic complexes, remains a highly coveted goal within the research community. Such a tool would significantly enhance the efficiency and effectiveness of catalyst design endeavors. This is the philosophy behind our Python package called Open Bidentate Ligand eXplorer (OBeLiX).

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

1.2. THE FOUNDATION OF THE ROAD

The current advances in computational homogeneous catalyst design primarily stem from the integration of four scientific disciplines: experimental organometallic chemistry and catalysis, quantum chemistry (QC), artificial intelligence (AI), and cheminformatics. These are at the core of current state-of-the-art approaches. Their historical evolution has significantly influenced and shaped the modern landscape of this field. Figure 1.1 presents a timeline of selected seminal works and tools across this multidisciplinary field, highlighting the parallel development of key methodologies and tools alongside experimental discoveries in homogeneous

catalysis. Within this graphical representation, the progress in structure optimization methods is denoted by a red star, while experimental works are represented by a blue square. The integration of cheminformatics, which is crucial for data analysis and modeling, is symbolized by a yellow circle. Lastly, the emergence and growing influence of AI and ML techniques in catalyst design are depicted by a purple hexagon. In this section, we will delve into the progress and advancements made in each field, shedding light on their respective developmental journeys.

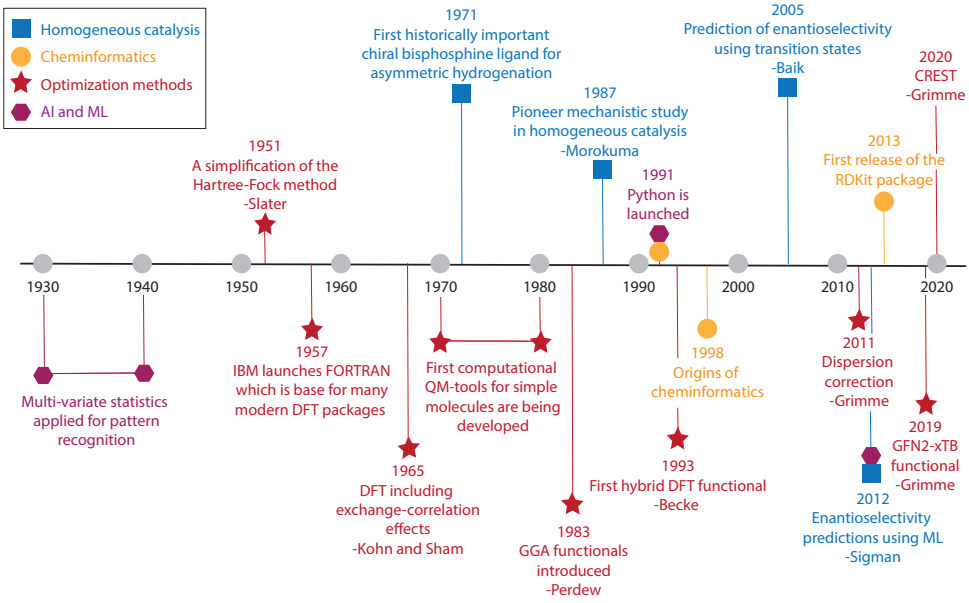


Figure 1.1.: A timeline showing the evolution of major fields contributing to the modern multidisciplinary research in homogeneous catalyst design.[52]

Electronic structure calculations play an important role in computational materials identification, characterization and optimization. For calculating properties of systems from first principles, DFT provides a powerful compromise between predictive power and computational cost.[53] Theoretical methods for studying catalysis have undergone significant development since then, with computational chemistry now regarded as an essential tool in the catalysis toolbox alongside laboratory techniques.[54–56] The origins of QC can be traced back to the pioneering work of Slater in 1951, marked by a red star on the left side of Figure 1.1. Slater's development of the Hartree-Fock method[57] marked the beginning of computational quantum mechanical (QM) methods by enabling feasible calculations for determining the energy minima of molecules.[52] The subsequent Kohn-Sham framework for approximating the electronic kinetic energy contribution proved especially useful.[58] A plethora of exchange-correlation potentials are currently advancing the frontier in accurate simulations.[59–65] For larger organometallic

complexes, these methods became particularly powerful after the introduction of Grimme's dispersion corrections.[66]

Theoretical frameworks and computational tools, must meet several criteria: (a) yielding reasonable outcomes, (b) operating efficiently within short timeframes, and (c) being applicable to a wide range of systems and physical-chemical properties.[67] The traditional DFT calculations are known to exhibit cubic scaling in computational time due to the diagonalization of the 3D Hamiltonian. This renders them inefficient for large molecular systems with a high number of electrons. Conventional force-field (FF) methods are frequently employed as a starting point, mainly for initial conformation searches.[68] These methods are not generally applicable since they lack parameterization for numerous elements, especially metals.[69,70] This has impeded the progression of the field.[67] In addressing this issue, low-level QC methods step in, offering an alternative to FFs, especially for systems of modest size, typically ranging from 500 to 1,000 atoms. For example, the GFNn-xTB methods are parameterized for applications to a wide range of chemical systems, including (organo)metallic systems[67,68,71] and polymers[72]. Grimme's Conformer-Rotamer Ensemble Sampling Tool (CREST), utilizes the GFNn-xTB methods for the creation and analysis of structure ensembles.[73] Conformational sampling via meta-dynamics simulations, regular MD simulations and genetic Z-matrix crossing have been implemented.[74] While exploring avenues to address the challenges in modeling organometallic systems, it is worth noting that the focus of this Chapter does not encompass machine learning interatomic potentials for electronic structure calculations, which is extensively discussed elsewhere.[75–79] Instead, a dedicated case study utilizing machine learning interatomic potentials is presented in Chapter 7.

After the modelling step and eventual conformer search, the discrete chemical structures need to be transformed into continuous representations for usage in statistical methods. This transformation is done by the calculation of chemical descriptors, as outlined in the introduction. These aim to capture the essential features of the catalyst for further analysis and design. Calculation of descriptors in a high-throughput manner was made possible by the invention of cheminformatics. Established in 1998[80], cheminformatics is an emerging domain of information technology. It focuses on the acquisition, organization, analysis, and management of chemical data. This discipline plays a crucial role in facilitating data-driven research and decision-making processes in chemistry. The advancement of cheminformatics is represented by a yellow circle in Figure 1.1. It has progressed in parallel with the field of machine learning in catalysis. Over the past two decades, continuous advancement of cheminformatics has significantly contributed to the progress achieved in the design and screening of homogeneous catalysts.[22,81] By leveraging the theoretical interpretation of chemical structures rather than relying solely on empirical measures, cheminformatics has enabled the derivation of meaningful relationships and the exploration of the vast chemical space.[82] This progress was mainly fueled by the invention of the Python programming language.[83] It allowed the creation of the OpenBabel[84] and RDKit[85] package, which are the backbone of many modern cheminformatics workflows. Additionally, the invention and widespread sharing of code via version control platforms such as Github, has

catalyzed the development of numerous innovative tools and workflows. These newly developed tools have empowered chemists with the capability to gather, analyze, and interpret chemical data in an efficient and systematic manner. Researchers have harnessed this powerful combination to build sophisticated algorithms for molecular descriptor calculation[18,86,87], virtual screening[88], reaction prediction[89–91] and many other aspects of catalyst design. The availability and integration of AI methods, represented by the purple hexagon in Figure 1.1, has further propelled the field by substantially enhancing the predictive power of these approaches. In its broadest definition, AI encompasses the theory and development of computer systems capable of performing tasks that traditionally require human intellect, such as speech recognition. As a prominent subset of computer science, AI has found significant applications in catalyst design, leveraging numerical methods and machine learning techniques to drive advancements in the field. While this Chapter will primarily focus on machine learning, it is important to acknowledge the vital role that numerical methods have played in enabling the development of DFT in the 1970s. These combined advancements have revolutionized catalyst design by augmenting the capabilities of computational models and enabling more sophisticated analyses and predictions.

The field of homogeneous catalysis has experienced a remarkable increase in the utilization and integration of AI techniques, driven by advancements in multi-variate statistics, quantitative structure-activity relationships (QSAR), and data science methodologies.[90,92,93] In recent years, there has been a notable transition within these modern machine learning approaches, as they have evolved from traditional *white-box* models to more sophisticated *black-box* models, where the emphasis is placed on the quality and size of the training data. White-box models are based on traditional statistics where causal effects are sought after and finding the most 'correct' model is the goal. On the other hand, black-box models prioritize predictive accuracy, aiming to find a highly performing model. Explainable models, such as QSARs, exemplify white-box models, where the model's performance is determined by the accuracy of physico-chemical parameters. Classic examples of explainable models include the Hammett equation, the Bronsted-Evans-Polanyi relationship, molecular volcano plots[94,95], and other linear free scaling relationships (LFERs). In contrast, black-box models often employ deep learning techniques, where descriptors derived from the molecular graph are utilized.[42,91,96] These black-box models focus on prediction quality and may lack explicit interpretability due to their complex internal representations. The exploration of both white-box and black-box models in automated catalyst design demonstrates the diverse strategies employed within the field, encompassing various approaches to predictive capability and performance. While white-box models provide interpretability and insights into causal relationships, black-box models offer greater predictive capabilities, leveraging vast amounts of data to make accurate predictions. The balance between white-box and black-box models in automated catalyst design represents a spectrum rather than a strict dichotomy. Figure 1.2 provides a visual representation of the data science continuum, showcasing various modeling techniques employed in catalyst design. These models encompass a range of methodologies, from explainable

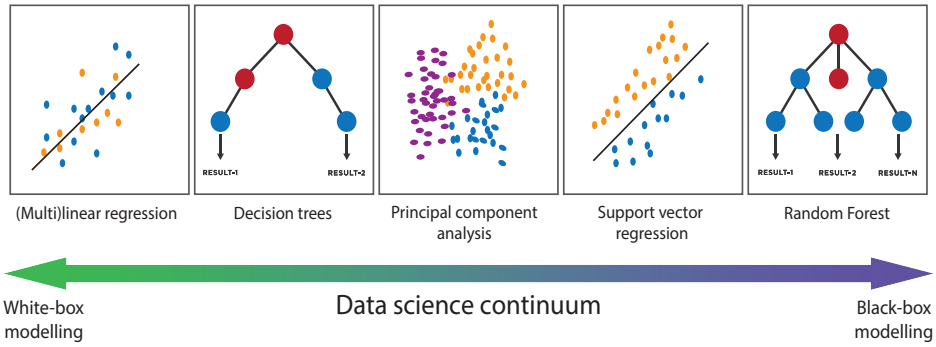


Figure 1.2.: The spectrum of data science techniques ranging from traditional white-box models that allow for explainability to black-box models that do not provide estimations on the importance of each feature or feature interactions.[97]

white-box models that prioritize interpretability, to more complex black-box models focused on predictability. Skilled scientists are capable of extracting valuable insights even from models traditionally classified as black box, such as when employing non-linear dimensionality reduction techniques. This continuous nature of modeling approaches highlights the interconnectedness and complementarity of different methodologies for catalyst design.

1.3. COMPUTATIONAL CATALYST DESIGN

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

Foscato et al. categorized catalyst design into two primary classifications: direct and inverse design.[22] In direct catalyst design, a direct causal relationship is established between a defined catalytic system and the observed experimental performance. Such performance is commonly quantified in terms of reactivity or selectivity, for example by measuring the conversion toward a desired product or the enantioselectivity achieved in forming a specific enantiomer. Since catalyst design is predominantly viewed as a nonlinear optimization problem, this endeavor often employs a diverse array of (non-)linear statistical methods.[22,98] On the other hand, inverse design starts from a known optimal performance and searches for systems with properties that match this performance.[98–100] Most inverse design strategies are still aimed at small organic molecules.[100,101] Only recently has the inverse strategy been applied to a subset of TM-based catalysts.[102] Since the focus of this Chapter lies on TM-based catalysts, our primary focus is on exploring direct catalyst design

strategies. These can generally be classified into two categories: mechanism-based and mechanism-agnostic approaches.

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

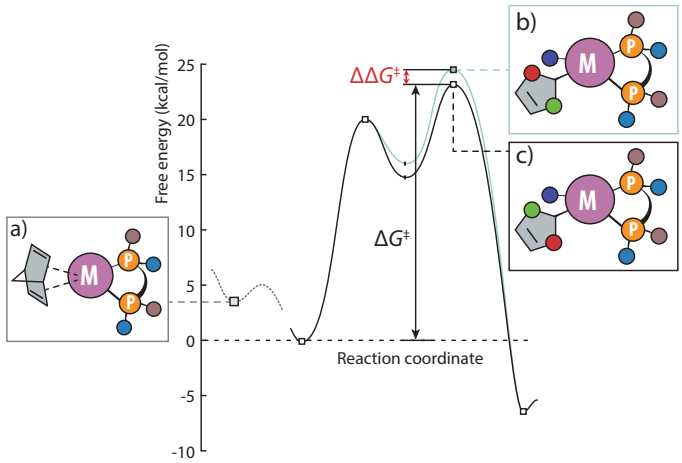


Figure 1.3.: A representative reaction profile diagram for an enantioselective reaction showing the structures used in mechanism-agnostic and mechanism-based computational catalyst design approaches. (a) Represents the precatalyst structure utilized in mechanism-agnostic approaches, while (b) and (c) depict competing prochiral transition state (TS) structures employed in the mechanism-based approach.

enantioselective reaction modeling as shown in Figure 1.3. In mechanism-agnostic approaches, a form of the precatalyst structure as shown in (a) which does not carry any mechanistic information can be utilized. The correlation of 3D descriptors calculated on this structure with selectivity has been utilized for the design and optimization of chiral ligands.[103] For mechanism-based approaches, TS structures of the selectivity determining step (b) or (c) are used. Small energy differences of 1-3 kcal/mol can significantly impact the preferred reaction pathway in enantioselective reactions, introducing additional complexities.[104] Achieving mechanistic insights thus entails the calculation of complex transition states from competing reaction pathways, followed by rigorous analysis. This makes the mechanism-based approach extremely problem-specific. In addition to these mechanistic insights, targeted DFT calculations are necessary for each new catalyst-substrate combination. On the contrary, the mechanism-agnostic approach is aimed to be more general. However, a deep understanding of the dataset and selected chemical descriptors for predictive modeling is necessary.[105] Despite these inherent disadvantages, both approaches have had successful applications in TM-based homogeneous catalyst design.[22,106]

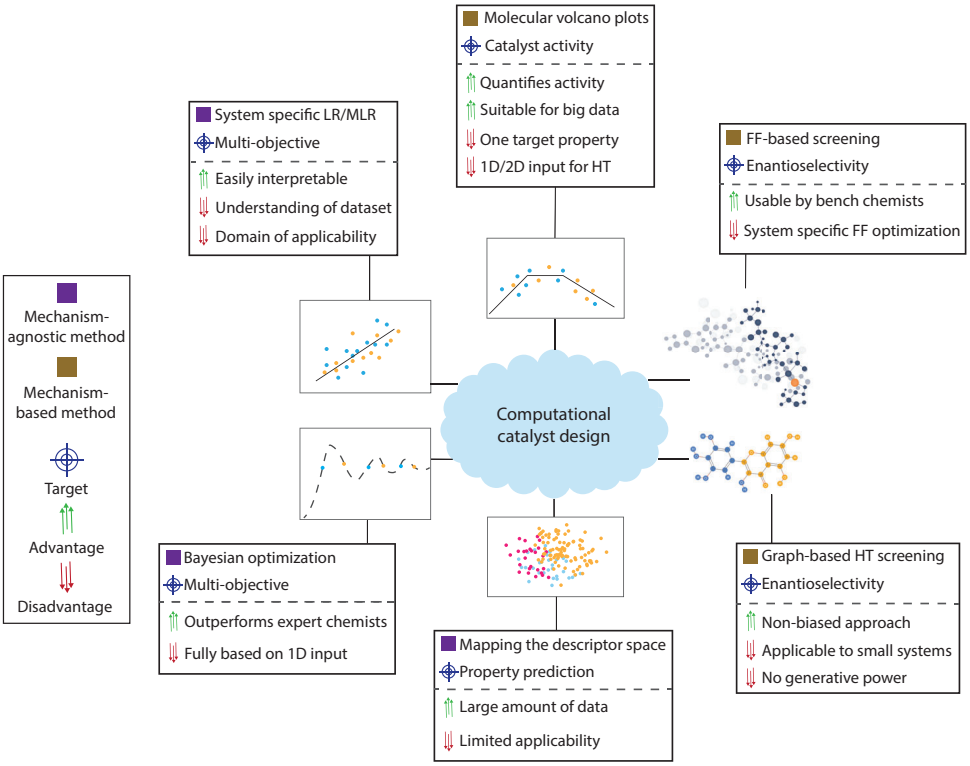


Figure 1.4.: A summary of the methodology, target applications, inherent advantages and disadvantages of several computational workflows for catalyst design. Mechanism-based approaches are indicated by a brown square. Selected examples are: FF-based[107] and graph-based high-throughput (HT) screening[108]. Mechanism-agnostic approaches are indicated by a purple square, selected examples are: Molecular volcano plots[109], system-specific linear regression/multi-linear regression (LR/MLR)[36], mapping the descriptor space[18] and bayesian optimization (BO)[110].

In the subsequent discussion, we provide a summary of a selection of state-of-the-art computational catalyst design approaches. These are presented in Figure 1.4.

1.3.1. MECHANISM-BASED APPROACHES

As mentioned, mechanism-based catalyst design approaches rely on mechanistic insights and are computationally intensive. However, they have been proven to predict the experimental enantioselectivity of TM-based catalysts well. As a first step in these approaches the transition state structures connecting two energy minima need to be found. Generally, most approaches first generate an approximate TS

structure and optimize towards a saddle point on the potential energy surface.[48] Automated and high-throughput localization of transition state structures has been enabled by tools such as AutoTST and AutoTS.[111–113] The need to sample for configurational and conformational freedom is particularly important in asymmetric catalysis. This sampling is, as far as we know, not implemented yet in these methods.[48] Separate tools enabling the sampling of transition state conformers are available, e.g. AARON[48,114], Mason[115] and MolAssembler[50]. These can be used to expedite and streamline the automated in-silico TS screening. These modules facilitate the conformational sampling, and transition state optimization for new catalyst-substrate variants. Combined with FF methods for transition states[116,117], these automated workflows contribute to faster and more efficient sampling of transition states. This is exemplified in several studies. A graph-based HT screening study was conducted by Laplaza et al.[108] An automated workflow was created to investigate multiple reaction pathways in a Rh-catalyzed asymmetric C-H functionalization and predict enantioselectivity. This was done on a set of 12 catalytic systems by sampling around 20 transition states per catalyst through unbiased conformational exploration with minimal human intervention.[108] The comparison of these computational predictions with experimental results shows that this workflow might be beneficial in the screening of new selective catalysts. Unfortunately, the computational cost of such an approach can run high due to the amount of transition states and DFT refinements needed. Alternatively, FF-based screening approaches aim to combine a QM treatment of a small core of atoms involved in the reaction, e.g. metal centers and donors of the ligands, with a force field treatment of the remaining molecule.[117–120] Virtual chemist, an approach by Patrascu et al., was specifically designed to empower experimental chemists with minimal computational chemistry knowledge.[107] This method combines Quantum-guided Molecular Mechanics (Q2MM) and molecular mechanics force field (MM3 FF) methods to model the transition states.[121,122] This approach enables bench chemists to virtually screen asymmetric reactions and make predictions about potential catalysts before conducting laboratory experiments. The study presents a significant advancement in the field of computational catalyst design through the development of a comprehensive virtual laboratory framework. This framework incorporates several modules, including Finders, React2D, Quemist, and Ace, each serving a specific function in the virtual design and evaluation of catalysts. However, Virtual Chemist does rely on a parametrized FF per reaction type. For example, the MM3 force field does not include parameterization for metals. To overcome this limitation, the force field parameters are automated using Hartree-Fock methods.[123,124] These Hartree-Fock methods are suboptimal for exploring TM-based catalysts.[107] Additionally, based on the flexibility of the catalyst, results can deteriorate since TSs are approximated as an energy minimum in Q2MM.[117] Shifts in the position of the transition state along the reaction coordinate, such as while facing significant steric hindrance, remain unaccounted in an energy minimum model.[117]

To study the activity of homogeneous catalysts, a well-established technique known as the volcano plot has been adopted from heterogeneous catalysis. This

approach originates from Sabatier's principle, which states that an ideal catalyst should exhibit an optimal level of bond strength with the substrate: neither too weak nor too strong.[94] The energies of reaction intermediates binding to the catalyst are interconnected through scaling relations, creating empirical mathematical relationships. These relationships allow the energies of all reaction intermediate and transition states to be expressed in terms of one or a few specific intermediates, forming linear free energy scaling relationships (LFSERs) based on a descriptor intermediate.[125] By analyzing the computationally calculated energy of the descriptor intermediate, the reaction rate can be estimated, resulting in the characteristic volcano shape. These plots have been used to define thermodynamic and kinetic profiles, aligning with experimental trends, both for smaller and large datasets.[95,126–129] The goal of quickly assessing the performance of prospective catalysts makes volcano plots well suited for big data analytics.[129] Meyer et al. employed a kernel ridge regression-based machine learning model to screen over 25,000 catalyst structures for the Suzuki-Miyaura C-C cross-coupling reaction.[109] They relied on a simplified thermodynamic profile, by using ML to learn the DFT-based reaction energy associated with oxidative addition which had been proven to be a descriptor variable in this catalytic cycle.[109] This approach allowed for rapid discrimination between catalysts with promising or inadequate energy profiles. Although this approach utilizing volcano plots yielded valuable insights into broad trends in catalyst behavior, it was only limited to screening of the catalyst activity. The application of volcano plots to a computational screening of enantioselectivity for TM-based homogeneous catalysts is still limited.[129]

1.3.2. MECHANISM-AGNOSTIC APPROACHES

Mechanism-agnostic approaches do not necessitate an understanding of the mechanism or the stereodetermining step for making reactivity or selectivity predictions.[130] For example, a promising approach in homogeneous catalyst design lies in various applications of quantitative structure-selectivity/activity relationships (QSSR/QSAR). A combination of quantum mechanical (QM) and statistical methods is the modern version of the QSAR approaches. In essence, QM-derived steric and geometric descriptors of the molecules are used to identify relationships between the catalyst structure and the observed experimental performance. Often, a general simplified catalyst structure with a 'dummy' substrate is used to derive these descriptors. Non-linear black-box models such as random forest, support vector machines, neural networks etc. have found to be successful in the prediction of target values such reaction yield or enantioselectivity.[131–133] More interpretable white-box univariate or multivariate linear regression have been successfully used in these applications as well.[18,103,130,134–138] A recently reported approach by Dotson et al. showcased an extensive workflow combining ML and HTE for multi-objective optimization.[36] The study focused on catalyst design and optimization, specifically targeting the yield and regioselectivity of chiral bisphosphine ligands. An extensive computational database consisting of 550 ligands was established, where diverse descriptors were computed for each ligand. The study was conducted on a Pd-catalyzed HayashiHeck reaction and a Rh-catalyzed alkene hydroformylation

reaction. Their methodology was shown to identify ligands with improved regioselectivity by ~ 1 kcal/mol compared to the previous best ligand. This novel methodology demonstrates the application of ML in addressing the simultaneous improvement of both yield and selectivity.[36] Unfortunately, although the results from the predictive model are readily interpretable, the construction of such a model requires a deep understanding of the calculated descriptors in relation to the reaction at hand. These descriptors depend on the computational catalyst structures, yet a detailed description of the process involved in their selection is often lacking. In addition, if the domain of applicability is limited and automation is minimal, the whole approach needs to be repeated for every addition of new catalysts.

It is well known that the chemical space is too large to be explored without automation. The development of Kraken, a comprehensive platform for mapping and predicting ligand properties, aimed at opening new avenues for understanding the catalytic chemical space.[18] Kraken encompasses a vast collection of 300,000 monodentate organophosphorus ligands, accompanied by 190 chemical descriptors that capture their conformational dependence. This mapping endeavor aims to encompass a broad range of conceivable structures relevant to organo(transition)metal reactions, providing valuable insights for catalyst design and optimization. The Kraken platform offers researchers access to computed data at different theoretical levels: semi-empirical QM, DFT, and ML. The database includes detailed information on 1,558 organophosphorus compounds, encompassing semi-empirical QM and DFT data, computed descriptors and properties, as well as coordinates information for associated conformers. Two versions of the compound were simulated, the free ligand and the ligand coordinated to the metal. To digitally represent structures, molecular descriptors are used. Additionally, the platform incorporates ML data, comprising 331,776 entries generated through combinatorial exploration of organophosphorus ligands with up to two distinct substituents. ML models are trained on the DFT dataset, enabling the on-the-fly prediction of properties for an extensive dataset of approximately 191 million distinct organophosphorus compounds. By utilizing the dataset and computational tools provided by Kraken, researchers can optimize reaction process parameters, inspire new ligand choices, and drive the synthesis of novel organophosphorus compounds.[88,135–137] The open-source nature of the Kraken platform and the accessibility of its extensive database facilitate collaboration and encourage contributions from the scientific community. Although this platform fosters ongoing advancements in the field of fully automated homogeneous catalyst design, it is currently limited to only monodentate organophosphorus ligands.

1.3.3. REACTION CONDITIONS & REAL-WORLD DATA

In the realm of automated homogeneous catalyst design, it is crucial to optimize towards reaction conditions which allow for maximum experimental productivity and efficiency of the catalyst.[139,140] This endeavor can be influenced by several factors such as the experimental error, the number of measured metrics, dataset size and data resolution.[141] Bayesian optimization (BO) in combination with a mechanism-agnostic approach introduced by Shields et al. enables optimization of reaction conditions.[110] The objective was to optimize the yield of the desired

product by exploring a combinatorial space of reaction conditions. The performance of their open-source BO framework could then be compared to a selected group of expert chemist. The framework employed different representations of reaction components, such as chemically descriptive fingerprint encodings based on quantum chemical properties computed via DFT, cheminformatics descriptors, and binary one-hot-encoded (OHE) representations generated using the Mordred package.[86] These reaction components were represented as a SMILES string and transformed into different representations using the Auto-Qchem Python package.[142] Remarkably, the BO framework incorporating DFT-derived features outperformed the chemists' expertise. Within the first 15 experiments, the framework consistently achieved higher average performance, yielding over 99% in all cases. The chemists, on the other hand, either prematurely terminated the optimization process or failed to identify the conditions that yielded the highest product yield.

While all the aforementioned approaches serve as exemplary methods, the utilization of non-structured real-world data, e.g. from electronic lab notebooks, for predictive endeavors raises concerns.[143] Determining whether the predictive value obtained is attributed to an inherent structure within the dataset presents a challenge. Additionally, biases can inadvertently manifest during the initial stages of data generation, such as when drawing catalyst structures for subsequent feature extraction or when making assumptions regarding reaction mechanisms. The discussed state-of-the-art approaches are automated to some extent. However, automating all steps from structure representation to prediction could make our work faster, more reproducible, and less prone to human error. The concept of modularity from the field of computer science can be useful here. The modular design of the workflows would mean that there is a logical partitioning of the steps that allows the separate parts to be integrated with easier implementation and maintenance. Though such an integrated workflow is more efficient, it is unfortunately not widely implemented yet in TM-based homogeneous catalyst design.

1.4. ROADBLOCKS

As introduced in the previous section, a direct catalyst design workflow usually consists of four components: structure generation, QC optimization, descriptor calculation, and a statistical method to relate these descriptors to properties of interest. Integrating these steps into a universally applicable computational workflow that streamlines all aspects, ranging from structure generation to descriptor computation seems trivial. However, the multidisciplinary nature of the involved steps and the modeling of variables that play a role in determining experimental catalytic performance form a hurdle.

This section will delve into some challenges that encompass key aspects of this automation task in TM-based homogeneous catalyst design: the representation of catalyst structures in computational workflows, the generation of reliable and diverse descriptors, and the inherent complexity of dynamics in catalysis. We attempt to address these challenges with our in-development tool focused on monodentate and bidentate ligand-containing structures, Open Bidentate Ligand eXplorer (OBeLiX).

With this Python package, we aim to automate and streamline the direct catalyst design workflow. Various of our in-house developed Python tools are currently integrated into OBeLiX, encompassing stand-alone modules for automated structure generation and subsequent descriptor calculation.

1.4.1. STRUCTURE REPRESENTATION

Regardless of the QSAR/QSSR approaches being implemented, there are three key parameters central to these workflows, as depicted in Figure 1.5. These are 1) the amount of data that is available, both computationally and experimentally, for the objective to be predicted 2) the interpretability of the prediction model and the associated computational cost and expertise required 3) the dimensionality of the computational representation of the catalyst structure.[144] These components are

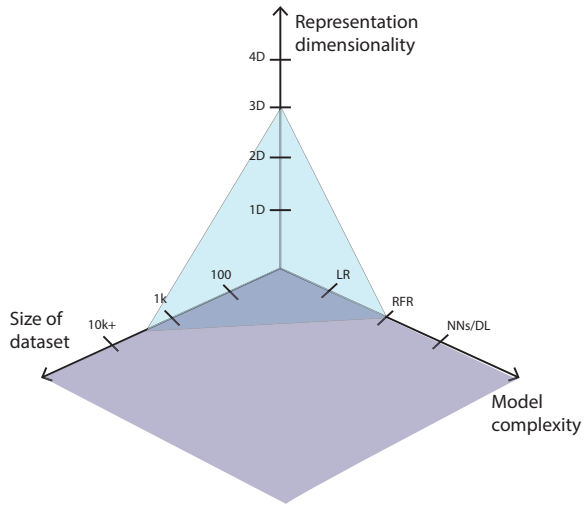


Figure 1.5.: An illustration of three critical parameters in computational catalyst design: the size of the dataset, quantified by the number of available data points; the model complexity, encompassing variations from simple linear regression (LR) to more complex non-linear models such as random forest regression (RFR), neural networks (NN), and other deep learning (DL) approaches; and lastly, the dimensionality of the computational structure representation, indicative of the level of detail captured by the representation. The shaded blue region signifies the size and complexity within which current computational catalyst design studies are mainly conducted.[144]

coupled and ever-changing, but more importantly, they can be a limiting factor. As an example, consider representation dimensionality. A 1D representation can be a SMILES string or one-hot encoded vector, a 2D representation is usually topology-based, in a 3D representation (QM-based) descriptors are derived from a 3D structure, while a 4D representation would also take the conformer ensembles

into account. In TM-based complexes, spin, oxidation state, coordinative bonds, and chirality can also be of importance to catalytic performance. Depending on the experimental objective to model, precise structural information from DFT-optimized 3D structures is required in a computational catalyst design workflow.[145] Since the catalyst structure, computational representation, and modeling space are deeply intertwined, it is critical to address and mitigate the biases that could be introduced in a manual structure generation approach.[20,146]

The automation of structure generation from string representations for TM complexes remains an active area of research, highlighting the ongoing efforts to overcome the challenges specific to these systems.[147] Computational packages have played a pivotal role in enabling various cheminformatics functions, including format conversion and other useful operations. Two widely recognized and extensively utilized tools in this domain are RDKit[85] and OpenBabel[84], which were introduced in 2013 and 2011 respectively. It is important to note that these tools primarily cater to organic molecules, reflecting the current focus of cheminformatics. However, as the field evolves, it is anticipated that future cheminformatics packages will expand their capabilities to handle more complex molecules and incorporate coordination bonds to cover the broader inorganic and organometallic chemistry landscape. Figure 1.6 visually presents the challenges of representing transition-metal (TM) complexes in three distinct formats: SMILES conversion, Morgan fingerprint, and graph representations. SMILES, or Simplified Molecular Input Line Entry System, is a concise notation for expressing chemical structures as text strings, offering a human-readable format. Morgan fingerprints, known as circular fingerprints, are a cheminformatics technique encoding molecular features based on substructures within a defined radius, producing a fixed-length binary vector. Graph representations in cheminformatics involve depicting molecules as graphs, portraying atoms as nodes and bonds as edges, capturing connectivity and topology. These representations have varying degrees of accuracy in capturing the intricate structure of TM complexes, as they have been successfully applied to organic molecules but often fail when applied to coordination complexes. When two coordinative bonds are formed with the metal center as shown in Figure 1.6, SMILES conversion and Morgan fingerprint representations fail to adequately represent the complex structure (indicated by crosses in both rows). The work of Sobez et al.[50] has demonstrated the effectiveness of graph representations in accurately encoding the 3D structure of TM complexes. While string representations have become commonplace in cheminformatics, they are not yet well-suited for predicting a delicate objective such as enantioselectivity in TM complexes, which is highly sensitive to structural variations.[108] Therefore, utilizing 3D representations for predictive models is desirable.

Building 3D representations of TM complexes is not always straightforward due to the possibility of multiple geometrical isomers, and coordination environment around the metal. Manual generation of these structures can introduce expert-bias by considering the chemical space partly. Utilizing the SMILES representation of components such as substrate, and ligand of the TM-based catalyst complex, a 3D structure can be built and configurationally explored. Two approaches are commonly employed for the automated generation of catalyst structures: 1) an exhaustive

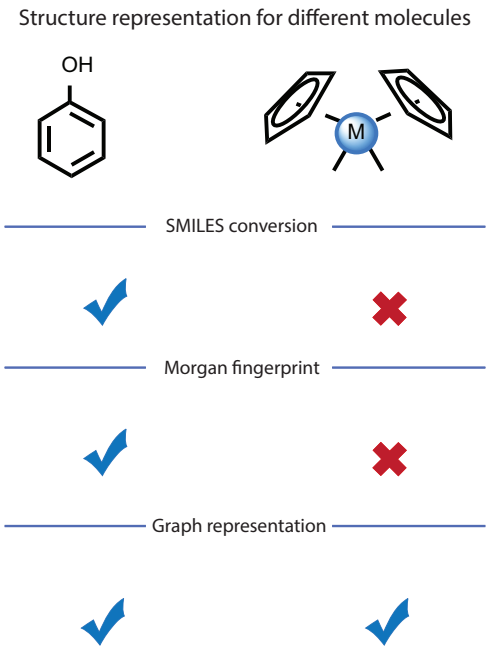


Figure 1.6.: Comparison of structure representations for TM Complexes. The figure illustrates the challenges of representing TM complexes using different file formats. While SMILES conversion and Morgan fingerprint representations are inadequate for capturing the geometric complexity, the graph representation accurately encodes the 3D structure.

search aided by heuristics and 2) searching algorithms aided by computational intelligence. However, neither method can generate perfect structures, and each has its own limitations. While ML algorithms may not achieve perfect or near-perfect accuracy, they are usually well-suited for designing small-scale systems, offering the advantage of speed. The molSimplify package is a notable example of a tool implementing an ML-based optimization tailored towards larger metal-ligand complexes. It employs a DFT-based pre-trained model to determine the skeleton structure, followed by selective force field optimization.[49] This approach allows for user-defined or program-determined ligand positioning on the metal center. DENOPTIM is an additional illustration of a computational intelligence-guided approach that combines fragment-building and genetic algorithms to construct hypothetical complexes with optimized fitness functions.[47] Examples of exhaustive algorithms are the Molassembler and Architector code.[50,51] Molassembler is a software tool that utilizes graph enumeration, stereopermuters, and the distance geometry algorithm to analyze and explore molecular structures, providing insights into connectivity, stereochemistry, and spatial arrangement.[50] It facilitates the generation of isomers and conformers, considers stereoisomeric configurations, and ensures physically realistic structures based on tabulated bond lengths. Architector

leverages metal-center symmetry analysis, distance geometry, fragment assembly, and ranking of conformers based on GFN2-xTB energies to capture the diversity of known experimental chemical space and design new complexes.[51]

Our approach to developing a platform for direct design of homogeneous catalysts is centered around enabling chemists to provide drawings of ligands and substrates in a chemically intuitive manner, from which the chemical space is automatically explored. The combined use of our in-house developed tools, MACE and ChemSpaX, facilitates this process as depicted in Figure 1.7. MACE is used

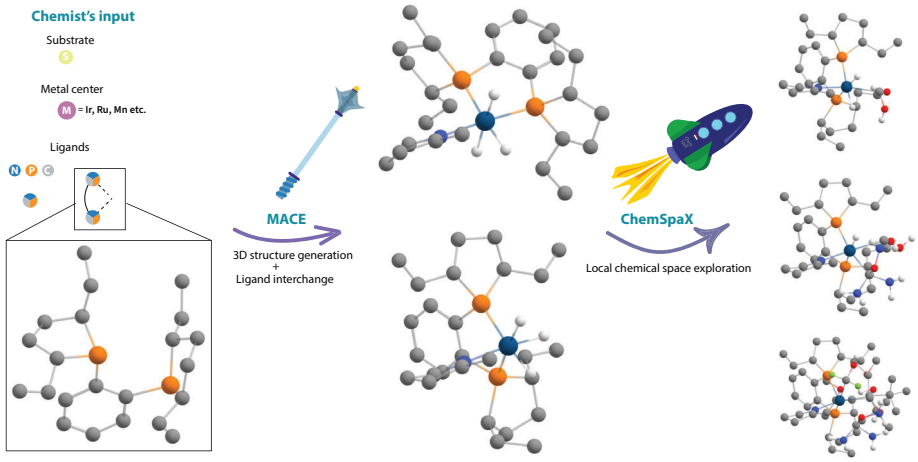


Figure 1.7.: An illustrative example of chemical space exploration of a TM-complex containing a bidentate ligand using MACE and ChemSpaX. The process begins with chemists providing 2D drawings of ligands and substrates. MACE generates 3D structures of TM-based metal-ligand scaffolds, conducting exhaustive searches of stereoisomers. This method, integrated into OBeLiX, enables expert-bias-free exploration of stereoisomers. ChemSpaX further explores the local chemical space by systematically placing substituent groups on the ligand. Together, these tools facilitate 3D structure generation and subsequent high-throughput screening for potential catalyst structures.

for generating 3D structures of TM-based metal-ligand scaffolds, starting from a 2D input. It is specifically aimed at conducting exhaustive searches of configurations and stereoisomers in square planar and octahedral complexes. MACE offers the advantage of generating a diverse range of stereoisomers, including those involving configurational isomers where ligands are swapped, while also providing computed energies through force-field calculations to rank these isomers. By incorporating the MACE protocol into computational workflows for organometallic complexes, we enable the expert-bias-free exploration of stereoisomers.[148–150] The introduction of structural variations at an early stage aligns with the complexity observed in real systems, allowing for the identification of likely stereoisomers in a high-throughput

manner. When 3D scaffold generation is done, the local chemical space of this structure can be explored by systematically placing substituent groups on the ligand. This approach creates close variations of the ligand structure. In OBeLiX, this is done by utilizing the ChemSpaX package.[151] Overall, these methodologies facilitate the exploration of an extensive chemical space for the screening of potential catalyst structures. Subsequent structural refinement can be achieved through geometry optimization, utilizing methods such as QC at any level of precision.

It is essential to note that current automated 3D structure generation methods do not inherently consider synthesizability or automatically adhere to chemical rules. Typically, manual error checking based on a random sampling is performed. Examples of automated error-checking workflows exist for small organic molecules cheminformatics, where SMILES and synthesizability scores can easily be generated. Such methodologies could involve 1) a blend of de novo design and synthesis planning or 2) a combination of biased generation with a synthesizability heuristic.[152] In the first approach, the structure generation method undergoes training on the existing data to incorporate knowledge of the synthetic steps involved in compound creation. This approach relies on reactivity rules encoded in a discrete action space of reaction templates, trained on artificial pathways generated from a pool of purchasable compounds and a list of expert-curated templates.[153] The second approach proposes the use of heuristics based on synthesizability to effectively bias generation towards synthetically tractable chemical space.[154] However, it has been observed that this may divert the generative model from its primary optimization objective.[154] In both cases, it is important to note that these automated error-checking methods are still in the development phase and have seen limited application even for small organic molecules. Looking forward, as our understanding of the inorganic chemical space advances, coupled with the availability of more experimental data and enhanced computational representations, these automated error-checking approaches may find application in the domain of TM-based homogeneous catalyst design.

1.4.2. GENERATION OF DESCRIPTORS

After geometry optimization, relevant electronic, steric, geometric, thermodynamic or combined descriptors can be extracted from the results of computations. Some classic examples such as the TEP, Tolman cone angle, and buried volume were mentioned in the introduction. It is possible that descriptors require manual input, e.g. when defining quadrants/octants of the buried volume for the prediction of enantioselectivity.[155] An example of a typical buried volume orientation is shown in Figure 1.8. The direction of axes here is defined with respect to the two donor phosphorus (P) atoms attached to the metal center. The relevance of this directional definition becomes apparent when there is a need to distinguish the quadrant occupied by 'P1' from the quadrant occupied by 'P2'. In the mechanism-agnostic multi-objective optimization approach discussed in the previous section, all atoms of the generated scaffold are manually mapped to define the orientation of this descriptor.[36] The indices of 'P1', 'P2' and the metal center, among others, are saved in an Excel file. If a new ligand is added to the dataset, this mapping has to be

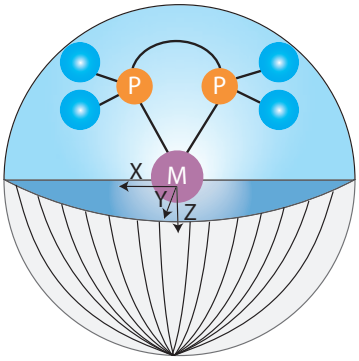


Figure 1.8.: Illustration depicting a representative orientation of a buried volume, defined with respect to the metal center and bidentate ligand donors. Typically, the buried volume is quantified at a designated radius and expressed as a percentage relative to the ligand occupying the encompassing sphere. Furthermore, the contributions of distinct quadrants and octants can be assessed by establishing a 3D axis and subdividing the sphere into discrete sections.

redone. To address this challenge, we employ a graph-based method to identify the ligand donor atoms using interatomic distances in the OBeLiX platform. These donor atoms are then numbered based on their charge and this definition is used in subsequent descriptor calculation. By leveraging this automated approach, we aim to reduce manual input in the calculation of descriptors, ensuring a more objective and efficient exploration of structures in catalyst design. This becomes especially relevant for smaller datasets, where the sensitivity of descriptors can impact the performance of predictive models.[26,40,156]

In addition to addressing biases within the descriptor generation part of catalyst design, it is important to acknowledge the challenges faced in the experimental domain which influence any predictive capabilities. While the use of AI has gained significant attention in various fields, its implementation in chemistry is still evolving.[157] Glorius et al. have highlighted the impact of systematic errors on dataset balance and completeness, which can severely limit the reliability of ML-based predictions.[158] In the context of direct design of catalysts, the objective is to create structures with desired properties. However, the accuracy of predictive models heavily relies on the quality of the training data.[159] If the experimentally measured target property is prone to significant errors or bias, it can introduce difficulties.[159] To address these limitations, it has been suggested that systematic reporting, including the documentation of underperforming reactions, can mitigate errors and improve dataset quality.[159,160] Additionally, due to the rapid advancements in ML, it may be more efficient to develop new algorithms that can overcome the challenges associated with unbalanced and missing data, facilitating accurate predictions of quantitative properties.[161]

1.4.3. THE COMPLEXITY OF CATALYSIS

A computer sees the catalytic system through the prism of the models and methods designed by humans, which are usually far from mimicking the complex chemical interactions in a real system. Based on the chosen settings and methods, DFT-based modeling might lack certain crucial aspects that are relevant at both micro and macro scales.[54] Real catalytic systems exhibit complexity arising from various factors, including solvent effects, conformational variations, and catalyst deactivation.[140] Modelling these complex phenomena might necessitate a combination of mechanism-agnostic and mechanism-based approaches. For example, in addition to the descriptors calculated on a general simplified catalyst structure, as is often done in the mechanism-agnostic approaches, those derived from specific TSs or reaction intermediates would make the model more realistic.[117,162] Mechanistic studies in catalysis are usually conducted using DFT methods[163], but the computational cost of exploring all possible reaction paths considering the aforementioned factors can be prohibitively high. Hence, a form of conformer searching or reaction network exploration based on semi-empirical methods might prove useful.[68,73,164–166] In such a conformer search, the presence of multiple conformations may itself be an important descriptor relating to the catalytic performance.[26,167] To enable such a high-throughput 4D-QSAR/QSSR approach, automated descriptor calculation should be facilitated on a conformer ensemble. In that context, our OBeLiX workflow uses the open-source cclib and Morfeus packages to calculate descriptors on DFT outputs, CREST conformer ensembles or XYZ files of TM complexes with monodentate and bidentate ligands.[18,168] This allows for an integrated and generalizable approach to a high-throughput 4D-QSAR/QSSR screening.

1.5. SCOPE OF THIS DISSERTATION

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

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

This dissertation presents our research on *in silico* methods for the rational design of TM-based catalysts. Since ML algorithms are inherently constrained by the quality of the data used for training, particular emphasis is placed on bias-free, data-driven approaches to investigate the dynamics of catalysts in the context of computational screening.

In Chapter 2, a comprehensive experimental and computational dataset for the Rh-catalyzed enantioselective hydrogenation of olefins is established. This dataset is used to evaluate the performance of ML models using descriptors of varying complexity to represent catalyst structures in a simplified and static form. Chapter 3 builds on this by examining the influence of conformational sampling and the inclusion of an asymmetric substrate in DFT calculations, thereby addressing the limitations of static catalyst representations.

Chapter 4 introduces a generalizable methodology for systematic conformer selection from ensembles obtained with semiempirical methods. This approach provides a computationally accessible framework for representing catalysts with significant conformational flexibility at the theory level of DFT, and its development is informed by the findings from Chapters 2 and 3.

Moving beyond square-planar Rh-based systems, Chapter 5 addresses the role of stereoisomerism in more fluxional TM complexes, focusing on Ir(III)-, Ru(II)-, and Mn(I)-based catalysts. This extension highlights the challenges posed by stereochemical diversity in predictive modeling. The distinct properties of different metal centers lead to varying stabilities of ligand configurations, making it essential to investigate to what extent exhaustive configurational exploration is required, as overlooking these effects risks incomplete sampling of chemical space and the omission of key catalyst features in modeling efforts.

In Chapter 6, insights from the preceding chapters are integrated into a transfer learning framework. Here, ML models informed by conformational dynamics are applied to predict the catalytic activity of previously unseen ligands, demonstrating a practical route toward ML-based catalyst discovery.

Looking ahead, machine learning interatomic potentials hold promise for significantly reducing both the computational and financial cost of *in silico* screening. Chapter 7 evaluates a recently released foundation model, trained on a state-of-the-art dataset developed by Meta (formerly Facebook), for its applicability in exploring the conformational and configurational landscapes of catalysts studied in Chapters 5 and 6.

Finally, Chapter 8 provides a broader perspective on the future of data-driven homogeneous catalysis. It reflects on the interdisciplinary challenges at the interface of chemistry and data science, the practical and cultural dimensions of Open Science, and the rapid developments in artificial intelligence.

DATA AVAILABILITY

The Python package for the generation and featurization of catalyst structures, OBeLiX, is available via the Github organization page of the ISE group at TU Delft: EPiCs-group (https://github.com/EPiCs-group/obelix). A full tutorial for the workflows can be found at https://github.com/EPiCs-group/obelix-tutorial/.

CONTRIBUTIONS

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

REFERENCES

(1) Kalikadien, A. V.; Mirza, A.; Hossaini, A. N.; Sreenithya, A.; Pidko, E. A. *ChemPlusChem* **2024**, e202300702.

(2) Crawley, M. L.; Trost, B. M., *Applications of Transition Metal Catalysis in Drug Discovery and Development: An Industrial Perspective*; John Wiley and Sons: 2012.

(3) Fernandes, R. A.; Jha, A. K.; Kumar, P. *Catal. Sci. Tech.* **2020**, *10*, 7448–7470.

(4) Green, A. I.; Tinworth, C. P.; Warriner, S.; Nelson, A.; Fey, N. *Chemistry* **2021**, *27*, 2402–2409.

(5) Tran, B. L.; Johnson, S. I.; Brooks, K. P.; Autrey, S. T. *ACS Sus. Chem. Eng.* **2021**, *9*, 7130–7138.

(6) Kuriyama, W.; Matsumoto, T.; Ogata, O.; Ino, Y.; Aoki, K.; Tanaka, S.; Ishida, K.; Kobayashi, T.; Sayo, N.; Saito, T. *Org. Process Res. Dev.* **2012**, *16*, 166–171.

(7) Keim, W., *Concepts for the Use of Transition Metals in Industrial Fine Chemical Synthesis*; Wiley-VCH Verlag GmbH: 2008, pp 15–25.

(8) Kirkpatrick, P.; Ellis, C. *Nature* **2004**, *432*, 823.

(9) Saldívar-González, F. I.; Pilón-Jiménez, B. A.; Medina-Franco, J. L. *Phys. Sci. Rev.* **2019**, *4*, 20180103.

(10) Hammett, L. P. *Chem. Rev.* **1935**, *17*, 125–136.

(11) Hammett, L. P. *J. Am. Chem. Soc.* **1937**, *59*, 96–103.

(12) Hammett, L. P. *Trans. Faraday Soc.* **1938**, *34*, 156–165.

(13) Bell, R. P.; Hinshelwood, C. N. *Proc. R. Soc. A* **1997**, *154*, 414–429.

(14) Evans, M. G.; Polanyi, M. *Trans. Faraday Soc.* **1938**, *34*, 11–24.

(15) Gerischer, H. *Bull. Soc. Chim. Belg.* **1958**, *67*, 506–527.

(16) Parsons, R. *Trans. Faraday Soc.* **1958**, *54*, 1053–1063.

(17) Blaser, H. U.; Pugin, B.; Spindler, F.; Saudan, L. A., *Hydrogenation*; Wiley Online Books, 2017, pp 621–690.

(18) Gensch, T.; Gomes, G. D. P.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S.; Aspuru-Guzik, A. *J. Am. Chem. Soc.* **2022**, *144*, 1205–1217.

(19) Maldonado, A. G.; Rothenberg, G. *Chem. Soc. Rev.* **2010**, *39*, 1891–1902.

(20) Hageman, J.; Westerhuis, J.; FrÃijhauf, H.-W.; Rothenberg, G. *Adv. Synth. Catal.* **2006**, *348*, 361–369.

(21) Foscato, M.; Occhipinti, G.; Venkatraman, V.; Alsberg, B. K.; Jensen, V. R. *J. Chem. Inf. Model.* **2014**, *54*, 767–780.

(22) Foscato, M.; Jensen, V. R. *ACS Catal.* **2020**, *10*, 2354–2377.

(23) Burrows, L. C.; Jesikiewicz, L. T.; Lu, G.; Geib, S. J.; Liu, P.; Brummond, K. M. *J. Am. Chem. Soc.* **2017**, *139*, 15022–15032.

(24) Kwon, D.-H.; Fuller, J. T.; Kilgore, U. J.; Sydora, O. L.; Bischof, S. M.; Ess, D. H. *ACS Catal.* **2018**, *8*, 1138–1142.

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

(26) Gallegos, L. C.; Luchini, G.; John, P. C. S.; Kim, S.; Paton, R. S. *Acc. Chem. Res.* **2021**, *54*, 827–836.

(27) Tolman, C. A. *Chem. Rev.* **1977**, *77*, 313–348.

(28) White, D.; Taverner, B. C.; Leach, P. G.; Coville, N. J. *J. Comput. Chem.* **1993**, *14*, 1042–1049.

(29) Clavier, H.; Nolan, S. P. *Chem. Commun.* **2010**, *46*, 841–861.

(30) Durand, D. J.; Fey, N. *Chem. Rev.* **2019**, *119*, 6561–6594.

(31) Lundgren, R. J.; Stradiotto, M., *Key Concepts in Ligand Design*; John Wiley & Sons, Ltd: 2016; Chapter 1, pp 1–14.

(32) SkoraczyÃśski, G.; DIttwald, P.; Miasojedow, B.; Szymkuc, S.; Gajewska, E. P.; Grzybowski, B. A.; Gambin, A. *Sci. Rep.* **2017**, *7*, 1–9.

(33) Koenig, K. E.; Sabacky, M. J.; Bachman, G. L.; Christopfel, W. C.; Bamstorff, H. D.; Friedman, R. B.; Knowles, W. S.; Stults, B. R.; Vineyard, B. D.; Weinkauff, D. J. *Ann. N.Y. Acad. Sci.* **1980**, *333*, 16–22.

(34) Leeuwen, P. W. V.; Kamer, P. C.; Reek, J. N.; Dierkes, P. *Chem. Rev.* **2000**, *100*, 2741–2769.

(35) Leeuwen, P. W. V.; Kamer, P. C.; Reek, J. N. *Pure Appl. Chem.* **1999**, *71*, 1443–1452.

(36) Dotson, J. J.; van Dijk, L.; Timmerman, J. C.; Grosslight, S.; Walroth, R. C.; Gosselin, F.; PÃijntener, K.; Mack, K. A.; Sigman, M. S. *J. Am. Chem. Soc.* **2022**, *145*, 110–121.

(37) Chen, S.-S.; Meyer, Z.; Jensen, B.; Kraus, A.; Lambert, A.; Ess, D. H. *J. Chem. Inf. Model.* **2023**, *63*, 7412–7422.

(38) Fey, N.; Koumi, A.; Malkov, A. V.; Moseley, J. D.; Nguyen, B. N.; Tyler, S. N.; Willans, C. E. *Dalton Trans.* **2020**, *49*, 8169–8178.

(39) Friederich, P.; dos Passos Gomes, G.; Bin, R. D.; Aspuru-Guzik, A.; Balcells, D. *Chem. Sci.* **2020**, *11*, 4584–4601.

(40) Janet, J. P.; Kulik, H. J. *J. Phys. Chem. A* **2017**, *121*, 8939–8954.

(41) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. *ACS Cent. Sci.* **2019**, *5*, 1572–1583.

(42) Singh, S.; Sunoj, R. B. *Digital Discovery* **2022**, *1*, 303–312.

(43) Weininger, D. *J. Chem. Inf. Comput.* **1988**, *28*, 31–36.

(44) O'Boyle, N.; Dalke, A. *ChemRxiv* **2018**.

(45) Krenn, M.; Ai, Q.; Barthel, S.; Carson, N.; Frei, A.; Frey, N. C.; Friederich, P.; Gaudin, T.; Gayle, A. A.; Jablonka, K. M.; Lameiro, R. F.; Lemm, D.; Lo, A.; Moosavi, S. M.; NÃąpoles-Duarte, J. M.; Nigam, A.; Pollice, R.; Rajan, K.; Schatzschneider, U.; Schwaller, P.; Skreta, M.; Smit, B.; Strieth-Kalthoff, F.; Sun, C.; Tom, G.; Falk von Rudorff, G.; Wang, A.; White, A. D.; Young, A.; Yu, R.; Aspuru-Guzik, A. *Patterns* **2022**, *3*, 100588.

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

(47) Foscato, M.; Venkatraman, V.; Jensen, V. R. *J. Chem. Inf. Model.* **2019**, *59*, 32.

(48) Guan, Y.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E. *J. Chem. Theory Comput.* **2018**, *14*, 5249–5261.

(49) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. *J. Comput. Chem.* **2016**, *37*, 2106–2117.

(50) Sobez, J. G.; Reiher, M. *J. Chem. Inf. Model.* **2020**, *60*, 3884–3900.

(51) Taylor, M. G.; Burrill, D. J.; Janssen, J.; Batista, E. R.; Perez, D.; Yang, P. *Nat. Commun.* **2023**, *14*, 1–11.

(52) Haunschild, R.; Barth, A.; French, B. *J. Cheminf.* **2019**, *11*, 72.

(53) Huang, B.; von Rudorff, G. F.; von Lilienfeld, O. A. *Science* **2023**, *381*, 170–175.

(54) Pidko, E. A. *ACS Catal.* **2017**, *7*, 4230–4234.

(55) Besora, M.; Maseras, F. *WIREs Comput Mol Sci* **2018**, *8*, e1372.

(56) Sameera, W. M.; Maseras, F. *WIREs Comput Mol Sci* **2012**, *2*, 375–385.

(57) Slater, J. *Phys. Rev.* **1951**, *81*, 385–390.

(58) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133–A1138.

(59) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.

(60) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(61) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.

(62) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(63) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(64) Perdew, J. P.; Kurth, S.; Zupan, A.; Blaha, P. *Phys. Rev. Lett.* **1999**, *82*, 2544.

(65) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

(66) Grimme, S. *WIREs Comput Mol Sci* **2011**, *1*, 211–228.

(67) Bursch, M.; Neugebauer, H.; Grimme, S. *Angew. Chem.* **2019**, *131*, 11195–11204.

(68) Bursch, M.; Hansen, A.; Pracht, P.; Kohn, J. T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2021**, *23*, 287–299.

(69) Spicher, S.; Grimme, S. *Angew. Chem. Int. Ed.* **2020**, *59*, 15665–15673.

(70) Hu, L.; Ryde, U. *J. Chem. Theory Comput.* **2011**, *7*, 2452–2463.

(71) Iribarren, I.; Trujillo, C. *J. Chem. Inf. Model.* **2022**, *62*, 5568–5580.

(72) Wilbraham, L.; Berardo, E.; Turcani, L.; Jelfs, K. E.; Zwijnenburg, M. A. *J. Chem. Inf. Model.* **2018**, *58*, 2450–2459.

(73) Pracht, P.; Bohle, F.; Grimme, S. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(74) Grimme, S. *J. Chem. Theory Comput.* **2019**, *15*, 2847–2862.

(75) Behler, J. *J. Chem. Phys.* **2016**, *145*, 170901.

(76) Behler, J.; Csányi, G. *Eur. Phys. J. B* **2021**, *94*, 1–11.

(77) Mishin, Y. *Acta Mater.* **2021**, *214*, 116980.

(78) Nagai, R.; Akashi, R.; Sugino, O. *npj Comput. Mater.* **2020**, *6*, 1–8.

(79) Eckhoff, M.; Reiher, M. *J. Chem. Theory Comput.* **2023**, *19*, 3509–3525.

(80) Brown, F. K. In Bristol, J. A., Ed.; Annu. Rep. Med. Chem. Vol. 33; Academic Press: 1998, pp 375–384.

(81) Dos Passos Gomes, G.; Pollice, R.; Aspuru-Guzik, A. *Trends in Chemistry* **2021**, *3*, 96–110.

(82) Polishchuk, P. *J. Chem. Inf. Model.* **2017**, *57*, 2618–2639.

(83) Van Rossum, G.; Drake Jr, F. L., *Python reference manual*; Centrum voor Wiskunde en Informatica Amsterdam: 1995.

(84) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminf.* **2011**, *3*, 33.

(85) Landrum, G. RDKit: Open-source cheminformatics, 2020.

(86) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. *J. Cheminf.* **2018**, *10*, 4.

(87) Luchini, G.; Patterson, T.; Paton, R. *Zenodo* **2023**, DOI: 10.5281/ZENODO.7702614.

(88) Gensch, T.; Smith, S. R.; Colacot, T. J.; Timsina, Y. N.; Xu, G.; Glasspoole, B. W.; Sigman, M. S. *ACS Catal.* **2022**, *12*, 7773–7780.

(89) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. *Science* **2018**, *360*, 186–190.

(90) Zuraski, A. M.; Alvarado, J. I. M.; Shields, B. J.; Doyle, A. G. *Acc. Chem. Res.* **2021**, *54*, 1856–1865.

(91) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J. L. *Mach. learn.: sci. technol.* **2021**, *2*, 015016.

(92) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D. A.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A. *Chem. Soc. Rev.* **2020**, *49*, 3525–3564.

(93) Hansch, C.; Fujita, T. *J. Biol. Chem* **1964**, *39*, 284.

(94) Sabatier, P., *La catalyse en chimie organique, Librairie Polytechnique*; Ch. Béranger Paris et Liège: 1913.

(95) Busch, M.; Wodrich, M. D.; Corminboeuf, C. *ACS Catal.* **2017**, *7*, 5643–5653.

(96) Hoque, A.; Sunoj, R. B. *Digital Discovery* **2022**, *1*, 926–940.

(97) Loyola-Gonzalez, O. *IEEE Access* **2019**, *7*, 154096–154113.

(98) Kuhn, C.; Beratan, D. N. *J. Phys. Chem.* **1996**, *100*, 10595–10599.

(99) Sanchez-Lengeling, B.; Aspuru-Guzik, A. *Science* **2018**, *361*, 360–365.

(100) Freeze, J. G.; Kelly, H. R.; Batista, V. S. *Chem. Rev.* **2019**, *119*, 6595–6612.

(101) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. *ACS Cent. Sci.* **2018**, *4*, 268–276.

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

(103) Sigman, M. S.; Harper, K. C.; Bess, E. N.; Milo, A. *Acc. Chem. Res.* **2016**, *49*, 1292–1301.

(104) Hamza, A.; Schubert, G.; SoÃşs, T.; PÃąpai, I. *J. Am. Chem. Soc.* **2006**, *128*, 13151–13160.

(105) Estrada, J. G.; Ahneman, D. T.; Sheridan, R. P.; Dreher, S. D.; Doyle, A. G. *Science* **2018**, *362*, eaat8763.

(106) Maloney, M. P.; Stenfors, B. A.; Helquist, P.; Norrby, P.-O.; Wiest, O. *ACS Catal.* **2023**, *13*, 14285–14299.

(107) Patrascu, M. B.; Pottel, J.; Pinus, S.; Bezanson, M.; Norrby, P. O.; Moitessier, N. *Nat. Catal.* **2020**, *3*, 574–584.

(108) Laplaza, R.; Sobez, J. G.; Wodrich, M. D.; Reiher, M.; Corminboeuf, C. *Chem. Sci.* **2022**, *13*, 6858–6864.

(109) Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. *Chem. Sci.* **2018**, *9*, 7069–7077.

(110) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, *590*, 89–96.

(111) Bhoorasingh, P. L.; West, R. H. *Phys. Chem. Chem. Phys.* **2015**, *17*, 32173–32182.

(112) Bhoorasingh, P. L.; Slakman, B. L.; Khanshan, F. S.; Cain, J. Y.; West, R. H. *J. Phys. Chem. A* **2017**, *121*, 6896–6904.

(113) Jacobson, L. D.; Bochevarov, A. D.; Watson, M. A.; Hughes, T. F.; Rinaldo, D.; Ehrlich, S.; Steinbrecher, T. B.; Vaitheeswaran, S.; Philipp, D. M.; Halls, M. D.; Friesner, R. A. *J. Chem. Theory Comput.* **2017**, *13*, 5780–5797.

(114) Ingman, V. M.; Schaefer, A. J.; Andreola, L. R.; Wheeler, S. E. *WIREs Comput. Mol. Sci.* **2021**, *11*, e1510.

(115) Chen, S.; Nielson, T.; Zalit, E.; Skjelstad, B. B.; Borough, B.; Hirschi, W. J.; Yu, S.; Balcells, D.; Ess, D. H. *Top. Catal.* **2022**, *65*, 312–324.

(116) Frank; Per-Ola *Theor. Chem. Acc.* **2003**, *109*, 1–7.

(117) Rosales, A. R.; Quinn, T. R.; Wahlers, J.; Tomberg, A.; Zhang, X.; Helquist, P.; Wiest, O.; Norrby, P.-O. *Chem. Commun.* **2018**, *54*, 8294–8311.

(118) Chung, L. W.; Sameera, W. M. C.; Ramozzi, R.; Page, A. J.; Hatanaka, M.; Petrova, G. P.; Harris, T. V.; Li, X.; Ke, Z.; Liu, F.; Li, H.-B.; Ding, L.; Morokuma, K. *Chem. Rev.* **2015**, *115*, 5678–5796.

(119) Reetz, M. T.; Meiswinkel, A.; Mehler, G.; Angermund, K.; Graf, M.; Thiel, W.; Mynott, R.; Blackmond, D. G. *J. Am. Chem. Soc.* **2005**, *127*, 10305–10313.

(120) Feldgus, S.; Landis, C. R. *J. Am. Chem. Soc.* **2000**, *122*, 12714–12727.

(121) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8566.

(122) Wahlers, J.; Rosales, A. R.; Berkel, N.; Forbes, A.; Helquist, P.; Norrby, P.-O.; Wiest, O. *The Journal of Organic Chemistry* **2022**, *87*, 12334–12341.

(123) Seminario, J. M. *Int. J. Quantum Chem.* **1996**, *60*, 1271–1277.

(124) Allen, A. E. A.; Payne, M. C.; Cole, D. J. *J. Chem. Theory Comput.* **2018**, *14*, 274–281.

(125) Anand, M.; NÃÿrskov, J. K. *ACS Catal.* **2020**, *10*, 336–345.

(126) Busch, M.; Wodrich, M. D.; Corminboeuf, C. *Chem. Sci.* **2015**, *6*, 6754–6761.

(127) Wodrich, M. D.; Busch, M.; Corminboeuf, C. *Chem. Sci.* **2016**, *7*, 5723–5735.

(128) Wodrich, M. D.; Sawatlon, B.; Solel, E.; Kozuch, S.; Corminboeuf, C. *ACS Catal.* **2019**, *9*, 5716–5725.

(129) Wodrich, M. D.; Sawatlon, B.; Busch, M.; Corminboeuf, C. *Acc. Chem. Res.* **2021**, *54*, 1107–1117.

(130) Reid, J. P.; Sigman, M. S. *Nature* **2019**, *571*, 343–348.

(131) Singh, S.; Pareek, M.; Changotra, A.; Banerjee, S.; Bhaskararao, B.; Balamurugan, P.; Sunoj, R. B. *Proc. Natl. Acad. Sci.* **2020**, *117*, 1339–1345.

(132) Xu, L. C.; Zhang, S. Q.; Li, X.; Tang, M. J.; Xie, P. P.; Hong, X. *Angew. Chem. Int. Ed.* **2021**, *60*, 22804–22811.

(133) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631.

(134) Newman-Stonebraker, S. H.; Smith, S. R.; Borowski, J. E.; Peters, E.; Gensch, T.; Johnson, H. C.; Sigman, M. S.; Doyle, A. G. *Science* **2021**, *374*, 301–308.

(135) Zell, D.; Kingston, C.; Jermaks, J.; Smith, S. R.; Seeger, N.; Wassmer, J.; Sirois, L. E.; Han, C.; Zhang, H.; Sigman, M. S.; Gosselin, F. *J. Am. Chem. Soc.* **2021**, *143*, 19078–19090.

(136) Christensen, M.; Yunker, L. P. E.; Adedeji, F.; HÃd'se, F.; Roch, L. M.; Gensch, T.; dos Passos Gomes, G.; Zepel, T.; Sigman, M. S.; Aspuru-Guzik, A.; Hein, J. E. *Commun. Chem.* **2021**, *4*, 112.

(137) Crawford, J. M.; Gensch, T.; Sigman, M. S.; Elward, J. M.; Steves, J. E. *Org. Process Res. Dev.* **2022**, *26*, 1115–1123.

(138) Matsuoka, W.; Harabuchi, Y.; Maeda, S. *ACS Catal.* **2022**, *12*, 3752–3766.

(139) Kuliaev, P. O.; Pidko, E. A. *ChemCatChem* **2020**, *12*, 795–802.

(140) Yang, W.; Filonenko, G. A.; Pidko, E. A. *Chem. Commun.* **2023**, *59*, 1757–1768.

(141) Jiscoot, N.; Uslamin, E. A.; Pidko, E. A. *Digital Discovery* **2023**, *2*, 994–1005.

(142) Zuraski, A. M.; Wang, J. Y.; Shields, B. J.; Doyle, A. G. *React. Chem. Eng.* **2022**, *7*, 1276–1284.

(143) Saebi, M.; Nan, B.; Herr, J. E.; Wahlers, J.; Guo, Z.; ZuraÅski, A. M.; Kogej, T.; Norrby, P.-O.; Doyle, A. G.; Chawla, N. V.; Wiest, O. *Chem. Sci.* **2023**, *14*, 4997–5005.

(144) Kulik, H. J. *WIREs Comput. Mol. Sci.* **2020**, *10*, e1439.

(145) Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. *WIREs Comput Mol Sci* **2022**, *12*, e1603.

(146) Janet, J. P.; Kulik, H. J. *Chem. Sci.* **2017**, *8*, 5137–5152.

(147) Jensen, J. GitHub - jensengroup/xyz2mol: Converts an xyz file to an RDKit mol object.

(148) Yang, W.; Chernyshov, I. Y.; van Schendel, R. K.; Weber, M.; MÃijller, C.; Filonenko, G. A.; Pidko, E. A. *Nat. Commun.* **2021**, *12*, 1–8.

(149) Yang, W.; Kalavalapalli, T. Y.; Krieger, A. M.; Khvorost, T. A.; Chernyshov, I. Y.; Weber, M.; Uslamin, E. A.; Pidko, E. A.; Filonenko, G. A. *J. Am. Chem. Soc.* **2022**, *144*, 8129–8137.

(150) Yang, W.; Chernyshov, I. Y.; Weber, M.; Pidko, E. A.; Filonenko, G. A. *ACS Catal.* **2022**, *12*, 10818–10825.

(151) Kalikadien, A. V.; Pidko, E. A.; Sinha, V. *Digital Discovery* **2022**, *1*, 8–25.

(152) Goldman, B.; Kearnes, S.; Kramer, T.; Riley, P.; Walters, W. P. *J. Med. Chem.* **2022**, *65*, 7073–7087.

(153) Gao, W.; Mercado, R.; Coley, C. W. *arXiv* **2022**.

(154) Gao, W.; Coley, C. W. *J. Chem. Inf. Model.* **2020**, *60*, 5714–5723.

(155) Antinucci, G.; Dereli, B.; Vittoria, A.; Budzelaar, P. H.; Cipullo, R.; Goryunov, G. P.; Kulyabin, P. S.; Uborsky, D. V.; Cavallo, L.; Ehm, C.; Voskoboynikov, A. Z.; Busico, V. *ACS Catal.* **2022**, *12*, 6934–6945.

(156) Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Langâat, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J.; Schrier, J. *Nature* **2019**, *573*, 251–255.

(157) Cole, J. M. *Nature* **2023**, *617*, 438.

(158) Strieth-Kalthoff, F.; Sandfort, F.; KÃijhnemund, M.; SchÃd'fer, F. R.; Kuchen, H.; Glorius, F. *Angew. Chem. Int. Ed.* **2022**, *61*, e202204647.

(159) Taniike, T.; Takahashi, K. *Nat. Catal.* **2023**, *6*, 108–111.

(160) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. *J. Am. Chem. Soc.* **2021**, *143*, 18820–18826.

(161) Cole, J. M. *Nat. Chem.* **2022**, *14*, 973–975.

(162) Oslob, J. D.; Ãkermark, B.; Helquist, P.; Norrby, P.-O. *Organometallics* **1997**, *16*, 3015–3021.

(163) Ryu, H.; Park, J.; Kim, H. K.; Park, J. Y.; Kim, S.-T.; Baik, M.-H. *Organometallics* **2018**, *37*, 3228–3239.

(164) Grimme, S. *J. Chem. Theory Comput.* **2019**, *15*, 2847–2862.

(165) Hashemi, A.; Bougueroua, S.; Gaigeot, M.-P.; Pidko, E. A. *J. Chem. Theory Comput.* **2022**, *18*, 7470–7482.

(166) Hashemi, A.; Bougueroua, S.; Gaigeot, M.-P.; Pidko, E. A. *J. Chem. Inf. Model.* **2023**, *63*, 6081–6094.

(167) Crawford, J.; Sigman, M. *Synthesis* **2019**, *51*, 1021–1036.

(168) O'boyle, N. M.; Tenderholt, A. L.; Langner, K. M. *J. Comput. Chem.* **2008**, *29*, 839–845.

2

PROBING ML MODELS BASED ON HTE DATA FOR THE DISCOVERY OF ASYMMETRIC CATALYSTS

ENANTIOSELECTIVE hydrogenation of olefins by Rh-based chiral catalysts has been extensively studied for more than 50 years. Naively, one would expect that everything about this transformation is known and that selecting a catalyst that induces the desired reactivity or selectivity is a trivial task. Nonetheless, ligand engineering or selection for any new prochiral olefin remains an empirical trial-error exercise. In this study, we investigated whether machine learning techniques could be used to accelerate the identification of the most efficient chiral ligand. For this purpose, we used high-throughput experimentation to build a large dataset of Rh-catalyzed asymmetric olefin hydrogenation results, specifically designed for machine learning applications, and validated it against existing literature while addressing observed discrepancies. Additionally, a computational framework for the automated and reproducible quantum-chemistry based featurization of catalyst structures was created. Together with less computationally demanding representations, these descriptors were fed into our machine learning pipeline for both out-of-domain and in-domain prediction tasks of selectivity and reactivity. For out-of-domain purposes, our models proved limited efficacy. It was found that even the most expensive descriptors do not impart significant meaning to the model predictions. The in-domain application, while partly successful for predictions of conversion, emphasizes the need for evaluating the cost-benefit ratio of computationally intensive descriptors and for tailored descriptor design. Challenges persist in predicting enantioselectivity, calling for caution in interpreting results from small datasets. Our insights underscore the importance of dataset diversity with broad substrate inclusion and suggests that mechanistic considerations could improve the accuracy of statistical models.

This Chapter has been published as: Kalikadien, A. V.; Valsecchi, C.; van Putten, R.; Maes, T.; Muuronen, M.; Dyubankova, N.; Lefort, L.; Pidko, E. A. *Chem. Sci.* **2024**, *15*, 13618–13630.[1]

2.1. INTRODUCTION

More than half a century ago, Knowles and Horner reported the first example of an enantioselective olefin hydrogenation catalyzed by Rh in combination with a chiral phosphine ligand.[2–4] Although the obtained enantiomeric excesses were modest, their seminal work started the field. Asymmetric hydrogenation immediately appeared as an attractive method to produce enantiopure compounds.[5–10] Compared to the competing classical resolution technology, it exhibits 100% theoretical yield, high atom economy, and good to excellent enantiomeric excesses. Over the last 50 years, the work from numerous industrial and academic groups resulted in the development of many efficient chiral ligands and in the implementation of this technology for large scale production.[11–18] In addition to ligand development, the mechanism of this reaction was extensively studied via experimental[19–28] and computational studies based on Density Functional Theory (DFT)[29–32] with the realization that the key elementary steps (i.e. the transition states governing selectivity and reactivity) vary with the ligands.

Despite the extensive knowledge built over the years, finding the right asymmetric hydrogenation catalyst for a new prochiral olefin remains a very empirical exercise and requires the screening of a large set of ligands and reaction conditions. High throughput experimentation (HTE) methodologies have successfully been implemented to rapidly explore the numerous parameters affecting the outcome of an asymmetric hydrogenation reaction.[33–40] Nevertheless, integrating *in-silico* assessments of catalyst candidates into HTE campaigns would be highly beneficial.[41] It could further accelerate the time-sensitive process development of active pharmaceutical ingredients and lower the consumption of substrates needed to perform the HTE screening, often available in low quantity at the start of a drug development program. Unfortunately, the *in-silico* design and development of homogeneous catalysts remains a challenging task.[42–44] Predictive strategies for catalyst design are generally categorized into two groups depending on whether or not they require knowledge of the underlying mechanism of the catalytic cycle.[45–50] The mechanism-based approaches rely on quantum chemical calculations of the key transition state intermediates. Consequently, they require knowledge of the reaction mechanism and are therefore very specific to the catalytic system under study. In addition, they are computationally expensive due to the complex energetic landscape of the transition metal-based catalysts. A few reports utilized this approach for the prediction of enantioselectivity of Rh based hydrogenation.[51–54] To make mechanism-based approaches practical at a larger scale, potential energy functions of the reactants and products such as force-fields are used to approximate the connecting transition state.[46] Recent implementations either mix the reactant and product potential energy surface with different weights/corrections to get an approximation of the stereo-determining transition state[51] or utilize transition-state force fields to approximately describe the transition state directly.[46,52,54]

The alternative approach that does not require any knowledge of the mechanism is the use of quantitative structure-property relationships (QSPR).[55–61] It consists in establishing a correlation between the structure of the catalyst and its performance e.g., with regards to its activity or selectivity. Originating from the traditional

LFERs, such as Hammett plots,[62–64] these methods have experienced a revival in the last decades with the advent of ML and its adoption by chemists.[47,56,65,66] Refined catalyst representations based on quantum chemical calculations combined with more sophisticated statistical approaches are challenging the *status quo* of homogeneous catalyst design.[56,59,60]

Recent studies utilized this approach for the design of selective Rh-based catalysts.[67,68] Xu et al. created a standardized database including over 12,000 data points on asymmetric hydrogenation of olefins from literature.[67] This database was utilized in a hierarchical learning approach to connect a large amount of related data from literature to the small amount of data from ongoing experimentation campaigns. It was shown that this hierarchical approach performs well for predicting the selectivity of reactions with closely related substrates. The tested catalyst and substrate representations were limited to 2D and 3D cheminformatics-based descriptors. Recently, Singh et al. showcased an approach rooted in quantum chemistry,[68] integrating quantum chemically derived molecular descriptors from five different asymmetric binaphthyl-derived catalyst families to predict the enantioselectivity of asymmetric olefin and imine hydrogenation. A random forest (RF) model trained on a set of 368 substrate-catalyst combinations demonstrated impressive predictive power compared to other linear and non-linear statistical methods, with a root-mean-square error in the predicted percent of enantiomeric excess (%ee) of about 8.4 ± 1.8 compared to experimental values.

Inspired by a recent publication of Sigman and coworkers together with Genentech,[69] we decided to evaluate whether we could build a predictive model to support our HTE workflow for the screening of asymmetric hydrogenation catalyst. Our standard library of 192 chiral Rh catalysts was used to generate high quality data (up 3552 data points) as sole input for our model since it has been recognized that literature data could induce biases.[70,71] In parallel, we developed a workflow to automatically generate a set of consistent DFT-based descriptors of our 192 ligands. Herein, we present our findings on the performance of our ML models in seven different cases including both out-of- and in-domain prediction tasks.

2.2. HTE DATA GENERATION AND DATA REPRODUCIBILITY

Experiments used in subsequent parts of this dissertation were carried out by Janssen's HTE group supporting chemical process development. Within the HTE group, several workflows were developed to expedite the screening of catalysts. For asymmetric hydrogenation, two plates are routinely screened, each containing 96 chiral ligands, in combination with Rh. These ligands were selected based on their documented activity in asymmetric hydrogenation and their commercial availability. To complete the plates, several ligands not typically used in such reactions were included. In the ligand library, bisphosphines (denoted as 'PP' ligands) were the most prevalent, comprising 142 entries (74%), followed by aminophosphines ('PN') with 25 entries (13%), phosphoramidites with 11 entries (6%), and monophosphines ('P') with 10 entries (5%). Each ligand contained at least one phosphorus donor atom.

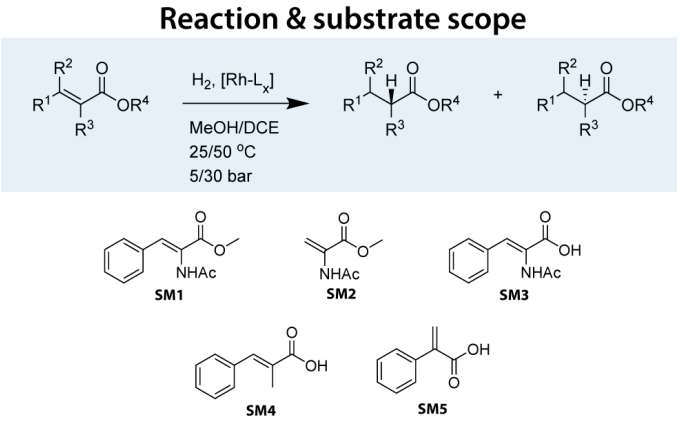


Figure 2.1.: Asymmetric hydrogenation reaction performed in this study. A set of varying substrates was selected to be tested with a wide range of Rh-based catalysts under varying conditions.

In this study, we tested our 192 Rh precatalysts against five model substrates: SM1-SM3, representing some of the most significant substrates in the development of asymmetric hydrogenation, and SM4-SM5, which are structurally related and pose slightly greater challenges (see Figure 2.1). Various substrates were tested under different reaction conditions. Although not a full factorial design, a total of 3552 data points were collected, representing to our knowledge the largest and most homogeneous dataset published for this crucial catalytic reaction. Table 2.1 summarizes the collected data points. To assess the stability of our Rh precatalysts, we evaluated the entire set with substrate SM1 immediately after their preparation, and again after six and twelve months of storage. Good reproducibility was observed for the enantiomeric excesses (ee), with coefficients of determination ranging between 0.87 and 0.94 across the experiments (see SI, Figure S3). Out of 576 data points, only 38 showed a discrepancy where the absolute change in ee ($|\Delta ee|$) measured in different runs exceeded 0.2.

As anticipated for a straightforward substrate like SM1, the conversion exhibited a strongly bimodal distribution, predominantly clustering around a value of 1 (indicating full conversion). To ensure a balanced classification, we categorized the data points into high conversion (conversion ≥ 0.8) and low conversion (conversion < 0.8). Out of 192 ligands, only 20 of them exhibited a variation in classification across different runs, resulting in an average Pearson correlation coefficient of 0.86 across the three runs. The accuracy levels for pairwise comparisons ranged from 0.92 to 0.96. For detailed information on the experimental procedure we refer the reader to Section S1 of the Supplementary Information.

Table 2.1.: Details of the 35 96-wells plates hydrogenation; data points selected for machine learning modeling in bold

Starting material	Solvent	T [°C]	H [bar]	Time [h]	#data points
SM1	DCE	25	5	1	192
				16	192
			30	16	192
	Methanol	25	5	1	**192**
				16	576
			30	16	192
SM2	DCE	25	5	1	192
				16	192
	Methanol	25	5	1	**192**
				16	192
SM3	DCE	25	5	1	192
				16	192
	Methanol	25	5	1	**192**
				16	192
SM4	Methanol	50	5	16	**192**
SM5	Methanol	25	5	16	96
		50	5	16	**192**
Total					**3552 (*960*)**

2.3. DATA ANALYSIS

2.3.1. INTERNAL DATA ANALYSIS OF EXPERIMENTAL RESULTS

A total of 3552 data points were generated during the data production exercise across 37 96-well plates (Table 2.1). SM1, SM2, and SM3 were tested in 2 solvents (methanol (MeOH) and 1,2-dichloroethane (DCE)) and 2 reaction times (1 and 16 h). For SM1, higher pressure (30 bar instead of 5 bar) was also explored. Additionally, the more challenging SM5 was tested at 2 temperatures (25°C and 50°C).

In addition to serving as high-quality input for machine learning models, this comprehensive dataset enables a systematic investigation of the effects of temperature, pressure and solvent across the entire set of ligands (see Figure 2.2). At first glance, it appears that a variation of the reaction conditions has less influence on the ee than on the conversion. As expected, increased temperatures lead to higher conversion while exerting minimal impact on enantioselectivity. Elevated hydrogen pressures were found to generally improve conversion but adversely affected enantioselectivity for a few ligands.[19,28,72] The solvent choice had the most significant effect on the performance; primarily on catalyst activity and, to a lesser extent, on enantioselectivity. These findings underscore the importance of solvent screening in HTE campaigns, advocating for its execution across a broad ligand set.

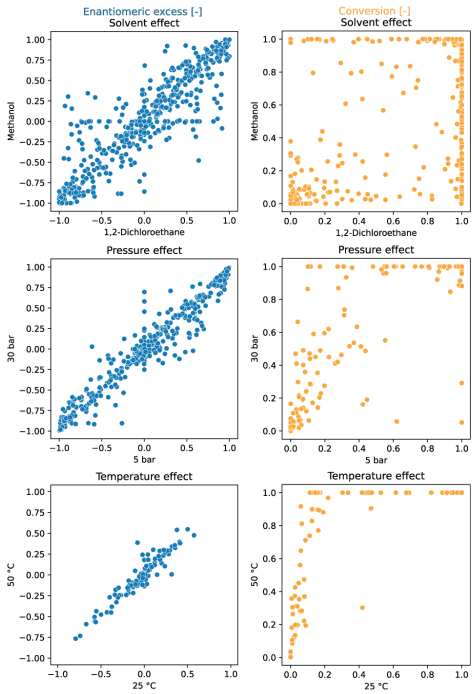


Figure 2.2.: The influence of various conditions on reactivity (conversion) and enantioselectivity (ee) in Rh-catalyzed asymmetric olefin hydrogenation. Solvent effect was evaluated on SM1-3 after 1 h reaction time. Pressure effect was evaluated on SM1 after 16 h. Temperature effect was evaluated on SM5 after 16 h and on one plate. An interactive version of this figure displaying the ligand structures corresponding to the data points can be found in the SI (see interactive figure 'Figure2.html' in the SI).

2.3.2. CONSISTENCY ANALYSIS OF EXPERIMENTAL RESULTS AND LITERATURE DATA

Since the substrates of our study have been extensively studied in Rh-catalyzed asymmetric hydrogenation, we conducted a comparison between our experimental results and those documented in the literature. This endeavor necessitated the aggregation and refinement of published data, a process that proved to be challenging. Utilizing the Reaxys database[73] (accessed in March 2023), we performed a reaction search for the conversion of SM1-SM5 into their corresponding hydrogenated products, omitting stereochemistry. This search yielded 2098 references, each requiring diligent formatting and cleaning to standardize reaction component labels and conditions. After discarding entries with missing information and focusing exclusively on Rh-catalyzed reactions, we obtained a dataset comprising 752 entries. Notably, 566 of these entries matched ligands from our library in conjunction with either MeOH or DCE as solvents. In line with observations made by other research

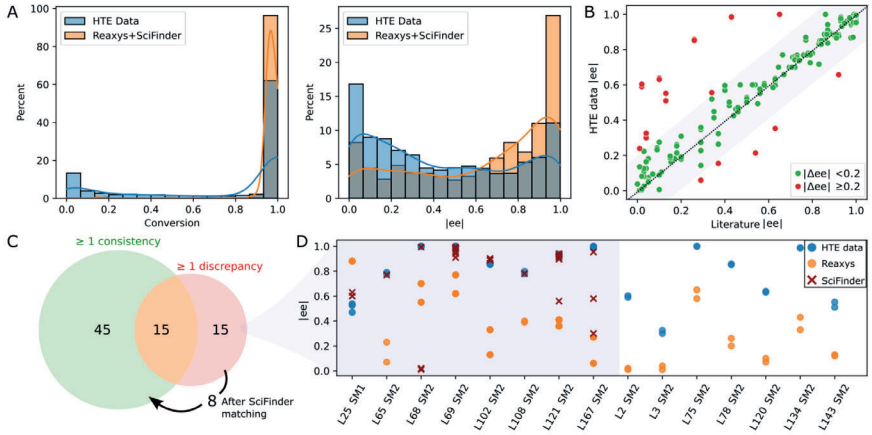


Figure 2.3.: Consistency Analysis: (A) Raw distributions of conversion and $|ee|$ in the current study (HTE data) - all data points in Table 2.1 - and in literature (Reaxys+Scifinder). (B) Scatter plot comparing the closest enantiomeric excess ($|ee|$) from literature with our experimental results under identical conditions (same catalyst, starting material, and solvent). (C) Venn diagram of ligand/substrate/solvent triplets divided into triplets with at least one consistent or discrepant Reaxys record (green and red set, respectively). The arrow shows that 8 triplets for which a consistency with Scifinder was found. (D) Comparative analysis of $|ee|$ discrepancies ($|\Delta ee| > 0.2$) across our data (blue), Reaxys (orange), and Scifinder (red) for the 15 triplets for which no consistency with Reaxys was found. An interactive version of this figure displaying the ligand structures corresponding to the data points can be found in the SI (see interactive figure 'Figure3.html' in the SI).

groups,[71,74] our HTE campaign contained more negative results (low conversion and/or enantioselectivity) as compared to those reported in literature (see Figure 2.3A), a disparity that augments the value of our dataset for machine learning model development.

The 566 literature entries comprised 75 unique ligand/substrate/solvent combinations, or "triplets", involving 39 distinct ligands. We compared the ee values reported in literature with our experimental values for each triplet (see Figure 2.3B). A "good match" was defined by an absolute ee difference of less than 0.2 (i.e., $|\Delta ee| < 0.2$). Among the 75 triplets, 45 exhibited complete concordance with no discrepancies observed between literature data and our experimental findings (see Figure 2.3C). Furthermore, 15 triplets showed partial agreement, with at least one literature result aligning with our experimental data. Overall, 80% of the literature data sourced from Reaxys aligned with our experimental results, reflected by a Pearson coefficient of 0.78. An additional search in Scifinder (accessed in March 2023) targeting the 15 triplets with discrepancies unearthed 152 new literature entries, enabling

the reconciliation for an additional 8 triplets (Figures 2.3C and 2.3D), thereby elevating the agreement with published studies to over 90%. For the remaining 7 discrepancies, no Scifinder records were found. As depicted in Figure 2.3D, our ee values were consistently higher than those documented in literature. A closer examination revealed that all discrepant literature data emanated from a singular source,[75] suggesting the possibility of a systematic experimental discrepancy within that study.

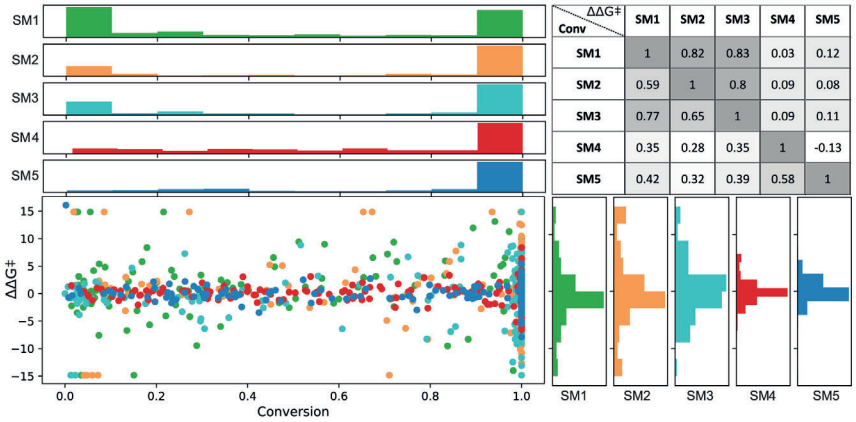
2.3.3. DATA SELECTION FOR PREDICTIVE MODELS

After the analysis of the consistency of our HTE dataset, 960 data points out of 3552 were selected as inputs for our models (Table 2.1). Despite the common advocacy for larger datasets in enhancing statistical model performance, this selection was driven by the need for data uniformity and the avoidance of redundancies that might compromise model effectiveness. Moreover, we opted to exclude data from the literature, such as those from Xu et al.,[67] to prevent the introduction of inconsistencies, adhering to findings discussed elsewhere regarding data bias.[70,76–79] In addition, this avoided the overrepresentation of a specific substrate. For simpler substrates SM1, SM2, and SM3, results from a 1-hour reaction time were chosen to ensure a balanced reactivity distribution. The empirically determined optimal temperature for each substrate was selected, and MeOH was consistently used as the solvent across all substrates to minimize variability from solvent effects, thereby allowing the model to more accurately focus on catalyst-specific features.

Considering the logarithmic nature of the enantiomeric excess, we chose to compute and model the $\Delta\Delta G^{\ddagger}$ values,[80] which followed a distribution that approximates normality.[49]

In the final set of 960 data points, enantioselectivity ($\Delta\Delta G^{\ddagger}$) demonstrated a distribution approximating a normal curve, with values ranging from -15 to 15 kJ/mol for SM1, SM2, and SM3, and -5 to 7 kJ/mol for SM4 and SM5.[81] The conversion results exhibited a bimodal distribution (see Figure 2.4), predominantly skewed towards higher values. Therefore, a classification model was built on a threshold of 0.8.

Spearmans rank correlation coefficient (see Figure 2.4) was employed to evaluate and compare the catalyst rankings across different substrates. SM1 and SM3 showed the highest correlation in their experimental values, with correlation coefficients of 0.77 for conversion and 0.83 for $\Delta\Delta G^{\ddagger}$. This was followed by SM1-SM2 (correlations of 0.59 for conversion and 0.82 for $\Delta\Delta G^{\ddagger}$) and SM2-SM3 (correlations of 0.65 and 0.80, respectively). In contrast, SM4 and SM5 did not exhibit significant correlations with the other substrates. For instance, among the top performers in enantioselectivity for substrates SM1-3 are ligands (R,R,S,S)-DuanPhos (L55) and (R,R)-Et-DuPhos (L68), whereas for substrates SM4 and SM5, the ligands providing highest enantioselectivity are SL-J505-1 (L18) and L186, respectively. Notably, L18 and L186 exhibit significantly lower selectivity with other substrates. As further detailed, conducting such an analysis provides critical insights for selecting training sets for out-of-domain tasks.



Conv \ ΔΔG‡	SM1	SM2	SM3	SM4	SM5
SM1	1	0.82	0.83	0.03	0.12
SM2	0.59	1	0.8	0.09	0.08
SM3	0.77	0.65	1	0.09	0.11
SM4	0.35	0.28	0.35	1	-0.13
SM5	0.42	0.32	0.39	0.58	1

Figure 2.4.: Distribution for conversion (%, on the top) and enantioselectivity ($\Delta\Delta G^{\ddagger}$ in kJ/mol, on the right) in red, blue, green, yellow, and magenta representing SM1, SM2, SM3, SM4, and SM5, respectively. The Figure includes a Spearman correlation matrix of experimental values for substrate pairs, with the upper triangle showing $\Delta\Delta G^{\ddagger}$ and the lower triangle indicating conversion.

2.4. CATALYSTS AND SUBSTRATE DESCRIPTORS

The next step towards a statistical model involved the featurization, i.e., mathematical representation, of the chemical entities. Descriptors were generated for the 192 catalysts and five substrates independently and then used as input features in our predictive models to encode the entire reaction space.[82]

2.4.1. LIGAND DESCRIPTORS

The numerical representation of chemical entities is key to the quality of machine learning models. We decided to test three different representations with increasing simplicity for our set of 192 catalysts: DFT-based descriptors derived from DFT optimized geometries, 2D cheminformatics based extended-connectivity fingerprints (ECFP4 with 512 bits) and one hot-encoding (OHE) of ligands and substrates.

For the DFT-based descriptors, we opted to compute well-established general descriptors for this reaction type without designing descriptors tailored to a specific reaction mechanism. We aimed at a state-of-the-art level of accuracy by performing DFT optimization with the PBE0-D3(BJ)/def2-SVP method (see SI section S3 for details and Appendix A for a benchmark study) on all 192 Rh- precatalyst, i.e., the cationic square planar $[Rh(L)(NBD)]^+$ (NBD=Norbornadiene) formed upon mixing $[Rh(NBD)_2]BF_4$ with a chiral ligand and therefore reflecting the precatalyst state in the experimental catalyst library. In addition to the alignment with the experimental workflow, the rigid and symmetrical nature of NBD was key to limit the conformational freedom and reduce the computational cost while featuring a Rh-olefin interaction.[83] Although this complex needs to lose NBD to enter the

catalytic cycle, we anticipated that the descriptors derived from such a metal-ligand complex would be close to the catalytically-relevant states where a square planar complex with a P-Rh-alkene and P-Rh-O bond is formed in the transition state. A Python package, Open Bidentate Ligand eXplorer (OBeLiX),[45] was developed to extract and calculate steric, geometric and electronic descriptors. A more detailed explanation about the featurization of the precatalyst structure is provided in section S4 of the SI. Among other features, OBeLix utilizes a graph-based method to identify the ligand in the complex. This ensured that steric descriptors, such as the buried volume, were only taking the ligand into account. For the non-symmetrical bidentate ligands, the two coordinating atoms were distinguished based on their charge with the label min/max denoting the least/most positively charged donor atom, respectively. In addition to descriptors derived from the $[Rh(L)(NBD)]^+$ complex, we also generated electronic descriptors for the ligand alone (labelled as free ligand). For this purpose, the ligand geometry was extracted from the optimized structure of the corresponding $[Rh(L)(NBD)]^+$ followed by a single-point (SP) DFT calculation. This entire workflow resulted in a total of 101 descriptors per catalyst. Highly correlated descriptors as well as descriptors judged redundant based on our computational chemistry intuition (e.g., Mulliken charges for atoms where NBO charges were already available) were removed leaving a final set of 34 descriptors per catalyst.[84] This set contained 15 steric, 8 geometric and 11 electronic descriptors. The steric descriptors include percent buried volumes calculated with either the donor atoms or metal as the center of the sphere, measuring the steric hindrance induced by a ligand.[85,86] The geometric descriptors include specific angles and distances, such as the bite angle,[87] the cone angle, a dihedral angle of NBD with respect to the donor atoms and distances between the donor atoms and the metal center in the complex. The Tolman cone angle[88] often serves as a method to assess the steric size of a ligand, however it was shown to be inaccurate for asymmetric ligands. The exact cone angle[89] as implemented in Morfeus was used instead. Finally, the electronic descriptors set consist of commonly used descriptors derived from electronic structure calculations such as the HOMO-LUMO gap, NBO charges of the donor and metal atoms, and lone pair occupancies of donor atoms. These descriptors represent either the metal-ligand bonding or the local electronic environment within the complex. Similarly, electronic parameters were extracted from the SP DFT calculation on the free ligand.

In addition to the 3D representation, 2D ECFP representations were generated from the metal-ligand complexes using RDKit.[90]

Recently, Sigman and coworkers published descriptors for 111 ligands present in our library.[69] Although these authors used $Pd(L)(Cl)_2$ in their DFT calculations, their global geometric and electronic descriptors correlate well with our descriptors. However, large discrepancies were observed for the local steric descriptors that are more sensitive to the procedure for the initial structure generation and to conformers search and selection (see SI section S5 for a detailed comparison).

2.4.2. PCA ANALYSIS OF 3D DFT-BASED DESCRIPTORS

We conducted a principal component analysis (PCA) on our dataset of DFT-based descriptors to investigate whether a dimensionality reduction would allow a visualization of the ligand space that aligns with human chemical intuition and understanding. Applying PCA directly to the 34 descriptors and selecting the first two principal components, which explain 37% of the variance, resulted in the formation of well-defined clusters (see Figure 2.5A) corresponding to chemically distinct classes of ligands (e.g., phosphoramidites, PN ligands, phosphine oxide). As expected, the loading plots for the first two principal components revealed a predominant influence of electronic descriptors, confirming that the clustering was primarily based on electronic differences (see SI Figure S11).

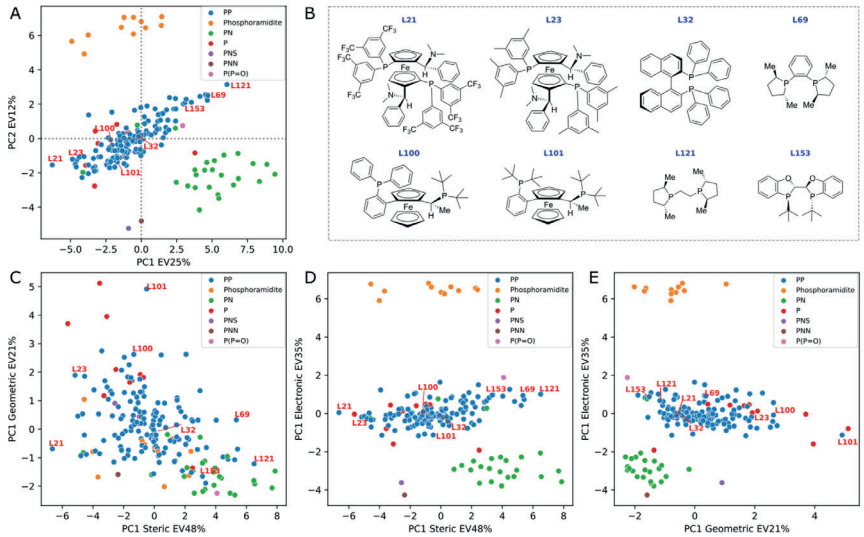


Figure 2.5.: PCA score plot (A) and cross-sections (C,D,E) based on binning descriptors into three categories: steric, geometric and electronic. Eight bisphosphine ligands are included as example (B). Percent of explained variance (EV) is reported in the axis label.

In an effort to discern smaller clusters, subsequent PCA analyses were performed on each category of descriptors (namely, the electronic, steric and geometric descriptors). The first principal component from each category was then used to construct cross sections and the results are summarized in Figure 2.5B-D (see PCA Interactive Figure in the SI for enhanced visualisation and analysis). The first component within each category accounts for a significant proportion of the variance explained by the descriptors, with values of 48%, 35%, and 21% for steric, electronic, and geometric descriptors, respectively. The cross section derived from geometric and steric descriptors clearly demonstrates that families of chemically distinct ligands occupy the entire geometric/steric space. The structures of a few ligands and their placements on the four PCA maps are depicted in Figure 2.5. Notably, similar

ligands, such as the MandyPhos (L21 and L23) and phospholane-based ligands (L69 and L121), are positioned in close proximity to one another. Along the steric axis, the bulky MandyPhos ligands (L21 and L23) are contrasted with the smaller DuPhos (L69) and BPE (L121), aligning with chemical intuition. Less intuitively, the WalPhos ligand (L101), characterized by two di-tert-butyl-phosphino groups and likely a large cone angle or Rh-P distance, is located at the extreme of the geometric axis. L100, another WalPhos ligand with only one di-tert-butyl-phosphino group, is appropriately positioned slightly below L101. At the opposite end of the geometric axis lies L153, a BIBOP ligand, presumably due to its compactness and small bite angle. These PCA maps could facilitate a data-driven approach for selecting a chemically diverse set of ligands for experimental screening. However, it is noteworthy that privileged structures,[91] such as for example the BINAP ligand (L32), are located in the center of all maps and thus are not distinguished by this methodology.

2.4.3. SUBSTRATE DESCRIPTORS

A static representation of all five substrates, SM1-SM5, was generated using four sets of descriptors: 3D DFT-based steric fingerprint, 3D SMILES-based steric fingerprint, ECFP and OHE. The steric fingerprints aimed to describe the local steric environment surrounding the olefinic bond. They were either created from a DFT optimized structure of the substrate alone or from a 3D structure generated by Openbabel[92] based on the SMILES of the substrate. The carbon atoms involved in the double bond (denoted as C_1 or C_2) as well as those directly connected to them were enumerated (denoted as R_1 to R_4). A buried volume for all these atoms and Sterimol parameters (B1, B5 and L) for each possible C and R pairing were calculated, resulting in a fingerprint consisting of six buried volumes and 12 Sterimol parameters (see our Github repository of the published ML pipeline for more details and code).

2.5. IN/OUT DOMAIN MODELING

As previously mentioned, our experimental data exhibits a bimodal distribution for conversion, biased towards higher conversion. Consequently, we opted for a classifier to model catalyst activity. The distribution of enantiomeric excesses ($\Delta\Delta G^{\ddagger}$) displayed a more normal pattern, rendering it suitable for regression analysis. Given the dataset encompassing five substrates (SM1-SM5) and 192 catalysts, our study presented a unique opportunity to explore both out-of-domain and in-domain modeling (see Figure 2.6).

In the out-of-domain approach, which would constitute the most impactful scenario, samples of the target starting material are excluded from the training dataset to evaluate the model's ability to predict reactivity and selectivity for starting materials that it has not encountered. This method involves numerically encoding both catalysts and substrates, followed by concatenating these encoded representations. The theoretically most informative feature set combines DFT-based descriptors for ligands with DFT-based steric fingerprints for substrates. Conversely, the simplest feature combination employs one-hot encoding for both ligands and substrates. Additionally, we explored the scenario when only a half of the target

starting material's samples is included in the training set, simulating the use of first HTE plate results before running a second plate. This approach is referred to as partially out-of-domain modeling. For conversion classification, we set a common threshold of 0.8 as the best balance between class distribution across substrate data.

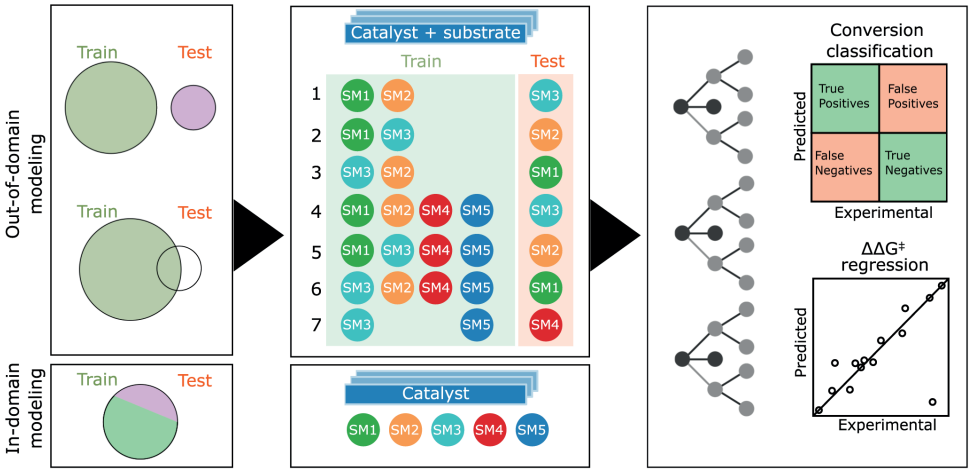


Figure 2.6.: Schematic Representation of the Machine Learning Workflow. In both fully and partially out-of-domain modeling scenarios, for each target starting material (SM), the model is trained on data from at least two additional SMs in accordance with seven specific cases. The feature matrix, is formed by concatenating descriptors of both catalyst and starting material. In partially out-of-domain modeling, half of the target SM samples are included in the training set. For in-domain tasks, each SM model undergoes training with an 80:20 training-test split, focusing solely on catalyst descriptors. We use of Random Forest for classification (reactivity) and regression (selectivity).

During our data analysis, we observed varying levels of correlation of the results obtained with the five starting materials (see Figure 2.4). This prompted us to investigate the impact of training set and target correlations on out-of-domain model accuracy. We detailed seven specific cases in Figure 2.6. Cases 1-3 focused solely on the related starting materials SM1, SM2, and SM3, using only the two most closely related substrates for training when predicting the target substrate's behavior. Cases 4-6 were still focused on predicting SM1, SM2, SM3 but with the inclusion of additional, less related substrates (SM4 and SM5) into the training set. Here, the goal was to assess the effect of substrate set diversity on model performance. Case 7 posed a more complex challenge, aiming to predict reactivity and selectivity for SM4 using the unrelated substrates SM3 and SM5 for training.

In the in-domain modeling, the objective is to predict catalyst performance for a specific substrate, employing a portion of the 192 results for model training and the remainder for evaluation. The conversion classifier's threshold is set based on the

median conversion specific to each substrate.

The efficacy of both out-of-domain and in-domain modeling approaches significantly depends on the choice of catalysts included in the training set. To mitigate this dependency and ensure the robustness of our findings, we tested three distinct random splits of the train/test set for each case. Hyperparameter tuning via a grid search within a predefined parameter space and k-fold cross-validation were performed for each split (see SI section S8 for more details on our ML pipeline).

Following preliminary screening with automated machine learning tools such as Auto-Sklearn[93] and TPOT,[94] we chose Sklearns Random Forest implementation as our study's algorithm.[95,96] Random Forest, an ensemble learning algorithm, harnesses multiple decision trees and randomness to construct a predictive model capable of handling diverse data types and excelling in classification and regression tasks.[68,97]

Furthermore, correlation analysis (see Excel file with experimental data and descriptors in the SI), revealed a limited univariate linear correlation (maximum absolute Spearman correlation coefficient of 0.58) between conversion and the DFT-based descriptors. The correlation of these descriptors with enantioselectivity was generally weaker (maximum absolute Spearman correlation coefficient of 0.15). A preliminary in-domain linear regression modeling (see SI Jupyter notebooks and pickle files with the final results in the SI) failed to accurately predict conversion and $\Delta\Delta G^{\ddagger}$, further justifying the selection of the Random Forest algorithm.

Considering four possible representations for the five starting materials, three ligand representations, seven cases and the prediction of both conversion and enantioselectivity, we trained in total over 700 models including 168 models for the fully out-of-domain task, 504 for the partially out-of-domain task (across three different training-test splits), and 90 for the in-domain task. To evaluate the predictive performance of our classifiers and regressors, we calculated the balanced accuracy (BA) and the coefficient of determination (R^2 score), respectively. BA is the average of recall obtained on each class and it ranges between 0 and 1 with 1 being the desired outcome (See the data availability statement for more information on all code and data).

2.6. RESULTS AND DISCUSSION

2.6.1. MACHINE LEARNING PREDICTIONS

OUT-OF-DOMAIN APPROACHES

The performance metrics of the models for the out-of-domain task are presented in Figure 2.7A and C (red dots). For reactivity modeling, we observed a BA ranging from 0.58 (case 7) to 0.85 (case 1, modeling SM3) with DFT-based descriptors for the catalysts, indicating that the models are able to estimate the experimental results with reasonable accuracy. Notably, SM3 modeling, when incorporating non-correlated substrates in case four, resulted in a marked decrease in performance (from 0.85 case 1 to 0.73 case 4), whereas this impact was minimal for other substrates (SM2, from 0.78 case 2 to 0.74 case 5; SM1, from 0.74 case 3 to 0.72 case 6). For the selectivity modeling using DFT-based descriptors for the catalysts, the highest R^2 score of 0.68, was observed in case 3.

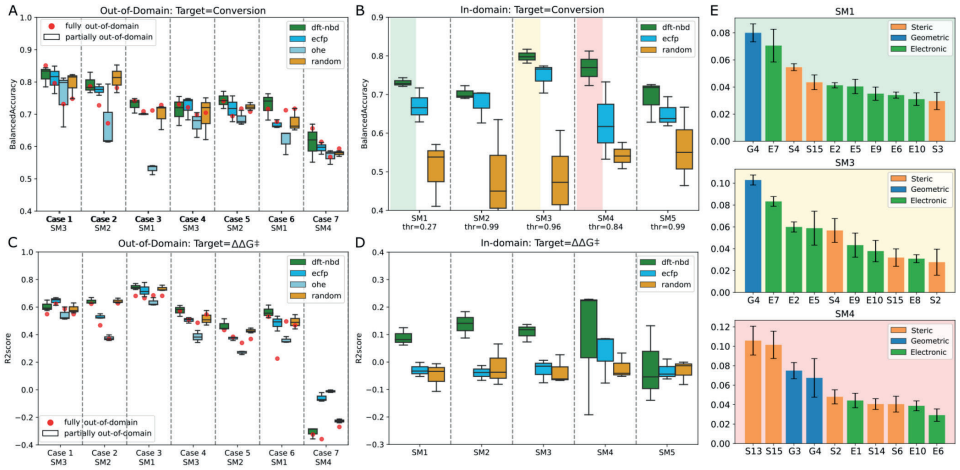


Figure 2.7.: Performance Metrics for out-of-domain and in-domain modeling. Panel A and C display the balanced accuracy and R2 score for out-of-domain modeling (A: Conversion; C: $\Delta\Delta G^{\ddagger}$), while Panel B and D illustrate the same for in-domain modeling (B: Conversion; D: $\Delta\Delta G^{\ddagger}$). In A and C the starting material's representation is one-hot encoded. Representations of fully out-of-domain results for DFT-based by red dots. E: Gini feature importance for RF in-domain classifiers trained on DFT-based descriptors to model conversion.

In this scenario, the target substrate SM1 exhibited a strong experimental correlation ($\Delta\Delta G^{\ddagger}$) with SM2 and SM3, as indicated by Spearman correlation scores of 0.82 and 0.83 respectively. Conversely, the performance of the model declined when introducing unrelated substrate in the training set (case 6: R^2 score of 0.53) and was very poor when trying to estimate the unrelated substrate SM4 (case 7: R^2 score of -0.3). Incorporating half of the catalyst set for the target substrate (partially out-of-domain task) generally did not significantly impact the balanced accuracy or the R^2 score (see Figure 2.7A and C, box plots). Indeed, the box plots obtained from random splits of the training set in the partially out-of-domain task almost always contain the red dots of the fully out-of-domain tasks.

Surprisingly, the more straightforwardly computed ECFP and trivial OHE exhibited BAs for conversion were largely consistent with those of the more costly DFT-based descriptors both for the fully and partially out-of-domain tasks. For case 2 and 3 only, the performance of OHE was notably inferior, suggesting that these models are influenced by the selection of training sets and hyperparameters. To mitigate overfitting, we restricted the maximum depth of the trees. This constraint, however, can be a bottleneck for OHE, adversely affecting its effectiveness, therefore for OHE only we let the model expand nodes until all leaves are pure. Models built on ECFP and OHE features, demonstrated performance that was for most cases comparable to DFT-based descriptors. To further examine the influence of

catalyst featurization, we introduced a new set of ligand descriptors consisting of 34 randomly generated values ranging from -100 to +100 for each ligand. A total of 192 vectors of random descriptors were generated and assigned to each ligand. In the out-of-domain modeling where multiple substrates are considered, a single ligand appears multiple times and was consistently represented by the same random vector. Interestingly, these random descriptors achieved the same performance as the DFT-based descriptors, with BAs ranging from 0.59 to 0.78 across the seven cases and R^2 score between 0.37 (case 5) and 0.68 (case 3) for the first six cases and a drop in performance for case seven. The maximum disparity noted in outcomes derived from DFT-based descriptors versus random descriptors within the fully out-of-domain approach is 0.06, as observed for case 7 for modeling conversion. These data suggests that our generated DFT-based descriptors were not able to capture the essential chemistry in this dataset as they do not impart significant meaning to the model estimates. The poorer performance demonstrated by OHE in certain cases suggests that binary descriptors may be less informative or less prone to chance correlations compared to other descriptor types. In general, the models effectiveness varied notably across different cases. The ability to estimate the experimental values correctly appears to be more related to the inherent correlation of catalyst performance across different substrates, rather than the intrinsic value of the descriptors. For instance, modeling SM4 using non-correlated substrates (SM3, SM5) in case 7 proved unsuccessful (Figure 2.7A,C), while outcomes based on correlated substrates (as in case 1) were more accurate (average BA for conversion of 0.8 and average R^2 score for selectivity prediction of 0.6). This outcome highlights that our machine learning models primarily estimates based on the principle of "what works now, will work in other cases" and vice versa (see SI Figure S13). Adding partial information about the target substrate to the training set (partially out-of-domain modeling) did not significantly enhance the accuracy nor alter the behavior of the model. In other words, a limited introduction of target substrate information does not substantially influence the performance of our models.

IN-DOMAIN APPROACH

Confronted with the unsuccessful out-of-domain modeling task, our focus shifted towards in-domain modeling, i.e., a more manageable, albeit less valuable, endeavor, in line with recent literature.[59,60,69] The objective was to test whether models constructed within a substrate-specific context, thus trained solely on catalyst descriptors, could effectively discern meaningful information. An additional goal was to assess whether DFT-based descriptors would introduce superior chemical information into the model compared to ECFPs and random descriptors. Results are summarized in Figure 2.7B and D. For reactivity, modeling using random descriptors achieved a BA around 0.5. Models utilizing DFT-based descriptors and ECFPs exhibited superior performance across all substrates. Specifically, for substrates SM1, SM3, and SM4, DFT-based descriptors surpassed ECFPs, achieving average BAs of 0.73, 0.80, and 0.77 on the test set, respectively, compared to ECFPs average BAs of 0.64, 0.71, and 0.63. For substrates SM2 and SM5, the performance of ECFPs was comparable to DFT-based descriptors, recording BAs of 0.65 versus 0.66 and 0.67

versus 0.69, respectively.

For SM1, SM3 and SM4, we investigated the feature importance of the trained models (see Figure 2.7E). Roughly, the same features are present for the related substrates SM1 and SM3 while different features are used by the model for unrelated SM4. For SM1 and SM3, most of the descriptors are electronic, for example, E7 (lone pair occupancy of the min donor atom calculated on free ligand), E5 (NBO charge of metal center) and E10 (NBO charge of the max donor atom calculated on the free ligand). The only geometric descriptor present in all four models is G4 (distance between Rh and the minimum donor). S15 (buried volume on minimum donor atom) seems to be the most important steric descriptors. Overall, it is difficult to derive any meaningful mechanistic considerations from these observations.

In-domain modeling for enantioselectivity was unsuccessful, as evidenced by R^2 scores not exceeding 0.2 on the test set. DFT-based descriptors for substrates SM1, SM2, and SM3 showed marginally better results than other descriptors. To investigate whether best-performing models for enantioselectivity could reside within smaller subsets of related catalysts, we implemented a Monte-Carlo data selection approach. This involved testing 1,000 random splits for each catalyst fraction, ranging from 90% to 10% of the entire catalyst set in decrements of 10%. Each subset was divided into an 80:20 training-test ratio, and RF models were trained exclusively using DFT-based descriptors. Our findings indicate the feasibility of deriving models with high R^2 scores (up to 0.98) on sets comprising merely 10% of the catalysts (only 25 and 4 data points in training and test set, respectively, see SI Figure S14). However, the lack of discernible pattern differentiating these catalysts from others, e.g. by ligand family or class, suggests that such high scores are solely due to chance correlation and test overfitting. This approach, deviating from standard machine learning practices, was employed to demonstrate the potential pitfalls when working with small datasets in machine learning pipelines, highlighting the risk of uncovering spurious, albeit appealing, correlations.

2.7. CONCLUSION

Using our high throughput experimentation workflow, we have generated and made available a large and reliable dataset of asymmetric hydrogenation results, encompassing most of the commercially available chiral ligands. Our experimental results align well with existing literature, affirming their validity but substantially exceed those in uniformity and comprehensiveness. Discrepancies observed were meticulously analysed and satisfactorily accounted for, ensuring the robustness of our dataset.

We proved that the application of machine learning modeling to estimate the reactivity of an unseen substrate relies solely on ligand differentiation with only a marginal improvement in performance observed for our set of DFT-based descriptors in the fully out-of-domain task for three out of seven cases. Several factors may contribute to this outcome. The dataset encompasses a limited range of substrates (five) and catalyst variability (192). This constraint hinders the models' ability to effectively interpret and differentiate features based on meaningful chemical

properties. Instead, the models tend to rely on mere object differentiation, which explains why random descriptors exhibit performance levels comparable to those of more expensive DFT-based descriptors. The outcomes of our models can be explained by the fact we derived our descriptors from the precatalyst with the goal to produce models with broad applicability. However, the combination of the reaction constituents, i.e., $[Rh(L)(NBD)]^+$ and substrate, may not be descriptive of reactivity-/stereo determining steps in the reaction mechanism. Given the current dataset, considering a transition state of catalysts with a specific substrate might be a more accurate alternative for modeling, albeit more computationally intensive and lacking the desired generality.

The "in-domain" strategy, while successful to some degree in modeling conversion, did not perform as well as anticipated. For certain substrates, we observed that ECFPs descriptors performed similarly to our DFT-based descriptors, indicating that our generated DFT-based descriptors might not significantly impact the model's effectiveness in certain scenarios. We acknowledge that our general, but still computationally intensive descriptors do not automatically outperform descriptors computed through more simplistic, naïve methods. This observation underscores the necessity for both a critical evaluation of the computational cost-benefit ratio and a tailored approach in the selection and engineering of descriptors for specific applications.

Enantioselectivity modeling remains a considerable challenge. We found no discernible correlations between our DFT-based descriptors and $\Delta\Delta G^{\ddagger}$, especially when working with small datasets (fewer than 20 data points). In such datasets, any strong correlations observed are likely due to chance, underscoring the necessity for cautious interpretation of results.

The insights garnered from our study highlight the importance of dataset diversity and mechanistic insights. Our findings suggest that expanding the dataset to include a broader range of substrates and ligands and applying an ad hoc DFT-based feature engineering process could potentially enhance model performance, particularly in out-of-domain scenarios. The library of 192 Rh catalysts is regularly being tested for new substrates of our development pipeline. The generated data are used to augment the dataset of the model with more diverse chemical entities. In addition, we are exploring alternative modeling approaches and new descriptors that might better capture the complexities of reactivity and selectivity modeling. Once our model will exhibit a high level of accuracy, we will include the screening of virtual ligands generated via *in-silico* modification of the existing ones[45] with the ambition to discover entirely new asymmetric hydrogenation catalysts.

DATA AVAILABILITY

The Supporting Information file for this Chapter is available at https://doi.org/10.1039/D4SC03647F.

The used machine learning pipeline is available at the Github organization page of the ISE group at TU Delft: EPiCs-group (https://github.com/EPiCs-group/obelix-ml-pipeline). Similarly, the used Python package for featurization of catalyst structures, OBeLiX, is available at the same page: EPiCs-group (https://github.com/EPiCs-group/obelix). In addition to this manuscript, supporting figures, code and all used datasets are available together with an extensive readme via 4TU.ResearchData at https://doi.org/10.4121/ecbd4b91-c434-4bdf-a0ed-4e9e0fb05e94:

- List and visualization of ligands ('ligand_list.pdf')
- Interactive figures ('Figure2.html',' Figure3.html' and 'PCA.html')
- DFT input and output files for metal-ligand complexes and extracted free ligand structures ('nbd_metal_ligand_dft_output.zip' and 'free_ligand_extracted_from_dft_output.zip')
- Excel file with experimental data and descriptors ('C=C_AH_dataset.xlsx')
- Excel file with ML results ('ml_results_tables.xlsx')
- Jupyter notebooks and pickle files with the final results ('data_analysis.ipynb', 'Literature_comparison_Reaxys_SciFinder.ipynb', 'dft_nbd_model_literature_comparison.zip', 'view.ipynb', 'dict_res_obj1.pkl', 'dict_res_obj2.pkl', 'dict_res_obj3.pkl', 'dict_res_obj4.pkl')

CONTRIBUTIONS

A.V. Kalikadien and **C. Valsecchi** contributed equally to this work. **A.V. Kalikadien:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration **C. Valsecchi:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration **R. van Putten:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization **T. Maes:** Conceptualization, Writing - Review & Editing **M. Muuronen:** Conceptualization, Methodology, Writing - Review & Editing **N. Dyubankova:** Conceptualization, Writing - Review & Editing, Supervision **L. Lefort:** Conceptualization, Methodology, Validation, Resources, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition **E.A. Pidko:** Conceptualization, Methodology, Validation, Resources, Data Curation, Writing Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

REFERENCES

(1) Kalikadien, A. V.; Valsecchi, C.; van Putten, R.; Maes, T.; Muuronen, M.; Dyubankova, N.; Lefort, L.; Pidko, E. A. *Chem. Sci.* **2024**, *15*, 13618–13630.

(2) Horner, L.; Siegel, H.; Büthe, H. *Angew. Chem. Int. Ed.* **1968**, *7*, 942.

(3) Knowles, W. S.; Sabacky, M. J. *Chem. Commun.* **1968**, 1445–1446.

(4) Knowles, W. S. *Angew. Chem. Int. Ed.* **2002**, *41*, 1998–2007.

(5) Yang, W.; Filonenko, G. A.; Pidko, E. A. *Chem. Commun.* **2023**, *59*, 1757–1768.

(6) Marianov, A. N.; Jiang, Y.; Baiker, A.; Huang, J. *Chem Catal.* **2023**, *3*, 100631.

(7) Seo, C. S. G.; Morris, R. H. *Organometallics* **2019**, *38*, 47–65.

(8) Imamoto, T., *Rhodium(I)-Catalyzed Asymmetric Hydrogenation*, 2019, pp 1–37.

(9) Brown, J. M. *Organometallics* **2014**, *33*, 5912–5923.

(10) Etayo, P.; Vidal-Ferran, A. *Chem. Soc. Rev.* **2013**, *42*, 728–754.

(11) Biosca, M.; Diéguez, M.; Zanotti-Gerosa, A., *Asymmetric hydrogenation in industry*; Diéguez, M., Pizzano, A., Eds.; Academic Press: 2021; Vol. 68, pp 341–383.

(12) Ager, D. J.; de Vries, A. H. M.; de Vries, J. G. *Chem. Soc. Rev.* **2012**, *41*, 3340–3380.

(13) Ager, D. J.; de Vries, J. G., *Industrial Applications of Asymmetric Reduction of C=C Bonds*; Carreira, E. M., Yamamoto, H., Eds.; Elsevier: 2012, pp 73–82.

(14) Busacca, C. A.; Fandrick, D. R.; Song, J. J.; Senanayake, C. H. *Adv. Synth. Catal.* **2011**, *353*, 1825–1864.

(15) Püntener, K.; Scalone, M., *Enantioselective Hydrogenation: Applications in Process R&D of Pharmaceuticals*, 2010, pp 13–25.

(16) Shimizu, H.; Nagasaki, I.; Matsumura, K.; Sayo, N.; Saito, T. *Acc. Chem. Res.* **2007**, *40*, 1385–1393.

(17) Johnson, N. B.; Lennon, I. C.; Moran, P. H.; Ramsden, J. A. *Acc. Chem. Res.* **2007**, *40*, 1291–1299.

(18) Shultz, C. S.; Krska, S. W. *Acc. Chem. Res.* **2007**, *40*, 1320–1326.

(19) Halpern, J. *Science* **1982**, *217*, 401–407.

(20) Landis, C. R.; Halpern, J. *J. Am. Chem. Soc.* **1987**, *109*, 1746–1754.

(21) Brown, J. M.; Chaloner, P. A. *Tetrahedron Lett.* **1978**, *19*, 1877–1880.

(22) Daubignard, J.; Lutz, M.; Detz, R. J.; de Bruin, B.; Reek, J. N. H. *ACS Catal.* **2019**, *9*, 7535–7547.

(23) Gridnev, I. D.; Yasutake, M.; Higashi, N.; Imamoto, T. *J. Am. Chem. Soc.* **2001**, *123*, 5268–5276.

(24) Gridnev, I. D.; Imamoto, T. *Acc. Chem. Res.* **2004**, *37*, 633–644.

(25) Gridnev, I. D.; Imamoto, T. *Chem. Commun.* **2009**, 7447–7464.

(26) Gridnev, I. D.; Imamoto, T. *Russ. Chem. Bull.* **2016**, *65*, 1514–1534.

(27) Imamoto, T.; Tamura, K.; Zhang, Z.; Horiuchi, Y.; Sugiya, M.; Yoshida, K.; Yanagisawa, A.; Gridnev, I. D. *J. Am. Chem. Soc.* **2012**, *134*, 1754–1769.

(28) Reetz, M. T.; Meiswinkel, A.; Mehler, G.; Angermund, K.; Graf, M.; Thiel, W.; Mynott, R.; Blackmond, D. G. *J. Am. Chem. Soc.* **2005**, *127*, 10305–10313.

(29) Gridnev, I. D.; Kohrt, C.; Liu, Y. *Dalton Trans.* **2014**, *43*, 1785–1790.

(30) Besora, M.; Maseras, F., *Computational insights into metal-catalyzed asymmetric hydrogenation*; Diéguez, M., Pizzano, A., Eds.; Academic Press: 2021; Vol. 68, pp 385–426.

(31) Feldgus, S.; Landis, C. R., *Catalytic Enantioselective Hydrogenation of Alkenes*; Maseras, F., Lledós, A., Eds.; Springer US: 2002, pp 107–135.

(32) Landis, C. R.; Hilfenhaus, P.; Feldgus, S. *J. Am. Chem. Soc.* **1999**, *121*, 8741–8754.

(33) Leitch, D. C.; Becica, J., *High-Throughput Experimentation in Organometallic Chemistry and Catalysis*; Parkin, G., Meyer, K., Ohare, D., Eds.; Elsevier: 2022, pp 502–555.

(34) Mennen, S. M.; Alhambra, C.; Allen, C. L.; Barberis, M.; Berritt, S.; Brandt, T. A.; Campbell, A. D.; Castañón, J.; Cherney, A. H.; Christensen, M.; Damon, D. B.; Diego, J. E. D.; García-Cerrada, S.; García-Losada, P.; Haro, R.; Janey, J.; Leitch, D. C.; Li, L.; Liu, F.; Lobben, P. C.; Macmillan, D. W.; Magano, J.; McInturff, E.; Monfette, S.; Post, R. J.; Schultz, D.; Sitter, B. J.; Stevens, J. M.; Strambeanu, I. I.; Twilton, J.; Wang, K.; Zajac, M. A. *Org. Process Res. Dev.* **2019**, *23*, 1213–1242.

(35) Renom-Carrasco, M.; Lefort, L. *Chem. Soc. Rev.* **2018**, *47*, 5038–5060.

(36) Krska, S. W.; DiRocco, D. A.; Dreher, S. D.; Shevlin, M. *Acc. Chem. Res.* **2017**, *50*, 2976–2985.

(37) Jäkel, C.; Paciello, R. *Chem. Rev.* **2006**, *106*, 2912–2942.

(38) Monfette, S.; Blacquiere, J. M.; Fogg, D. E. *Organometallics* **2011**, *30*, 36–42.

(39) Kallemeyn, J. M.; Hartung, J.; Connolly, T.; Ickes, A.; Kotecki, B.; Haandel, L. V.; Nazari, M.; Manjrekar, O.; Chen, S. *Org. Process Res. Dev.* **2022**, *26*, 2947–2956.

(40) Boogers, J. A. F.; Sartor, D.; Felfer, U.; Kotthaus, M.; Steinbauer, G.; Dielemans, B.; Lefort, L.; de Vries, A. H. M.; de Vries, J. G., *Asymmetric Hydrogenation of a 2-Isopropylcinnamic Acid Derivative en Route to the Blood Pressure-Lowering Agent Aliskiren*, 2010, pp 127–150.

(41) Eyke, N. S.; Koscher, B. A.; Jensen, K. F. *Trends Chem.* **2021**, *3*, 120–132.

(42) Poree, C.; Schoenebeck, F. *Acc. Chem. Res.* **2017**, *50*, 605–608.

(43) Hammes-Schiffer, S. *Acc. Chem. Res.* **2017**, *50*, 561–566.

(44) Houk, K. N.; Liu, F. *Acc. Chem. Res.* **2017**, *50*, 539–543.

(45) Kalikadien, A. V.; Mirza, A.; Hossaini, A. N.; Sreenithya, A.; Pidko, E. A. *ChemPlusChem* **2024**, e202300702.

(46) Maloney, M. P.; Stenfors, B. A.; Helquist, P.; Norrby, P.-O.; Wiest, O. *ACS Catal.* **2023**, *13*, 14285–14299.

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

(48) Foscato, M.; Jensen, V. R. *ACS Catal.* **2020**, *10*, 2354–2377.

(49) Ardkhean, R.; Fletcher, S. P.; Paton, R. S., *Ligand Design for Asymmetric Catalysis: Combining Mechanistic and Chemoinformatics Approaches*; Lledós, A., Ujaque, G., Eds.; Springer International Publishing: 2020, pp 153–189.

(50) Ahn, S.; Hong, M.; Sundararajan, M.; Ess, D. H.; Baik, M.-H. *Chem. Rev.* **2019**, *119*, 6509–6560.

(51) Patrascu, M. B.; Pottel, J.; Pinus, S.; Bezanson, M.; Norrby, P. O.; Moitessier, N. *Nat. Catal.* **2020**, *3*, 574–584.

(52) 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. *Nat. Catal.* **2019**, *2*, 41–45.

(53) Guan, Y.; Wheeler, S. E. *Angew. Chem. Int. Ed.* **2017**, *56*, 9101–9105.

(54) Donoghue, P. J.; Helquist, P.; Norrby, P.-O.; Wiest, O. *J. Am. Chem. Soc.* **2009**, *131*, 410–411.

(55) Williams, W. L.; Zeng, L.; Gensch, T.; Sigman, M. S.; Doyle, A. G.; Anslyn, E. V. *ACS Cent. Sci.* **2021**, *7*, 1622–1637.

(56) Dos Passos Gomes, G.; Pollice, R.; Aspuru-Guzik, A. *Trends Chem.* **2021**, *3*, 96–110.

(57) uraski, A. M.; Alvarado, J. I. M.; Shields, B. J.; Doyle, A. G. *Acc. Chem. Res.* **2021**, *54*, 1856–1865.

(58) Zahrt, A. F.; Athavale, S. V.; Denmark, S. E. *Chem. Rev.* **2020**, *120*, 1620–1689.

(59) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. *Science* **2018**, *360*, 186–190.

(60) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631.

(61) Crawford, J. M.; Kingston, C.; Toste, F. D.; Sigman, M. S. *Acc. Chem. Res.* **2021**, *54*, 3136–3148.

(62) Hammett, L. P. *Chem. Rev.* **1935**, *17*, 125–136.

(63) Hammett, L. P. *J. Am. Chem. Soc.* **1937**, *59*, 96–103.

(64) Hammett, L. P. *Trans. Faraday Soc.* **1938**, *34*, 156–165.

(65) Singh, S.; Sunoj, R. B. *Acc. Chem. Res.* **2023**, *56*, 402–412.

(66) Mace, S.; Xu, Y.; Nguyen, B. N. *ChemCatChem* **2024**, e202301475.

(67) Xu, L.-C.; Zhang, S.-Q.; Li, X.; Tang, M.-J.; Xie, P.-P.; Hong, X. *Angew. Chem. Int. Ed.* **2021**, *60*, 22804–22811.

(68) Singh, S.; Pareek, M.; Changotra, A.; Banerjee, S.; Bhaskararao, B.; Balamurugan, P.; Sunoj, R. B. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 1339–1345.

(69) Dotson, J. J.; van Dijk, L.; Timmerman, J. C.; Grosslight, S.; Walroth, R. C.; Gosselin, F.; Püntener, K.; Mack, K. A.; Sigman, M. S. *J. Am. Chem. Soc.* **2022**, *145*, 110–121.

(70) Saebi, M.; Nan, B.; Herr, J. E.; Wahlers, J.; Guo, Z.; Zuraski, A. M.; Kogej, T.; Norrby, P.-O.; Doyle, A. G.; Chawla, N. V. *et al. Chem. Sci.* **2023**, *14*, 4997–5005.

(71) Fitzner, M.; Wuitschik, G.; Koller, R.; Adam, J.-M.; Schindler, T. *ACS Omega* **2023**, *8*, 3017–3025.

(72) Shevlin, M., *High-Throughput Experimentation-Enabled Asymmetric Hydrogenation*; American Chemical Society: 2022; Vol. 1419, pp 107–130.

(73) Lawson, A. J.; Swienty-Busch, J.; Géoui, T.; Evans, D., *The Making of Reaxys-Towards Unobstructed Access to Relevant Chemistry Information*; American Chemical Society: 2014; Vol. 1164, pp 127–148.

(74) Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. *Angew. Chem. Int. Ed.* **2022**, *61*, e202204647.

(75) Alame, M.; Pestre, N.; de Bellefon, C. *Adv. Synth. Catal.* **2008**, *350*, 898–908.

(76) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J. L. *Mach. learn.: sci. technol.* **2021**, *2*, 015016.

(77) Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Langâat, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J.; Schrier, J. *Nature* **2019**, *573*, 251–255.

(78) Beker, W.; Roszak, R.; Woos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. *J. Am. Chem. Soc.* **2022**, *144*, 4819–4827.

(79) Fitzner, M.; Wuitschik, G.; Koller, R. J.; Adam, J.-M.; Schindler, T.; Reymond, J.-L. *Chem. Sci.* **2020**, *11*, 13085–13093.

(80) $\Delta\Delta G^{\ddagger} = -RTln\frac{(100+\%ee_R)}{(100-\%ee_R)}$ as in ref. 49.

(81) Our analytical method allows for the accurate determination of ee up to 99.5%, corresponding to $\Delta\Delta G^{\ddagger}$ of 15 kJ/mol.

(82) Mandal, D. K., *Transition metal-catalysed reactions: Diastereoselectivity and asymmetric synthesis*; Mandal, D. K., Ed.; Academic Press: 2021, pp 457–493.

(83) Baidun, M. S.; Kalikadien, A. V.; Lefort, L.; Pidko, E. A. *J. Phys. Chem. C* **2024**, *128*, 7987–7998.

(84) A comprehensive overview of all descriptors can be found in the Excel file with experimental data and descriptors in the SI. Additionally, a detailed explanation including code and data for the creation of this DFT NBD descriptors set is given in the SI section S4.

(85) Poater, A.; Cosenza, B.; Correa, A.; Giudice, S.; Ragone, F.; Scarano, V.; Cavallo, L. *Eur. J. Inorg. Chem.* **2009**, *2009*, 1759–1766.

(86) Falivene, L.; Credendino, R.; Poater, A.; Petta, A.; Serra, L.; Oliva, R.; Scarano, V.; Cavallo, L. *Organometallics* **2016**, *35*, 2286–2293.

(87) Durand, D. J.; Fey, N. *Chem. Rev.* **2019**, *119*, 6561–6594.

(88) Tolman, C. A. *Chem. Rev.* **1977**, *77*, 313–348.

(89) Bilbrey, J. A.; Kazez, A. H.; Locklin, J.; Allen, W. D. *J. Comput. Chem.* **2013**, *34*, 1189–1197.

(90) Landrum, G. RDKit: Open-source cheminformatics, 2020.

(91) Yoon, T. P.; Jacobsen, E. N. *Science* **2003**, *299*, 1691–1693.

(92) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminform.* **2011**, *3*, 33.

(93) Feurer, M.; Eggensperger, K.; Falkner, S.; Lindauer, M.; Hutter, F. *arXiv* **2022**.

(94) Le, T. T.; Fu, W.; Moore, J. H. *Bioinformatics* **2020**, *36*, 250–256.

(95) Breiman, L. *Mach. Learn.* **2001**, *45*, 5–32.

(96) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(97) Chuang, K. V.; Keiser, M. J. *Science* **2018**, *362*, eaat8603.

3

Impact of Model Selection and Conformational Effects on the Descriptors for In Silico Screening Campaigns

Data-driven catalyst design is a promising approach for addressing challenges in identifying suitable catalysts for synthetic transformations. Models with descriptor calculations relying solely on the precatalyst structure are potentially generalizable but may overlook catalyst-substrate interactions. This study explores substrate-specific interactions in the context of Rh-catalyzed asymmetric hydrogenation to elucidate the impact of substrate inclusion on catalyst structure and on the descriptors derived from it. We compare a catalyst-substrate complex with methyl 2-acetamido-acrylate as a model substrate with the generic precatalyst structure involving a placeholder substrate, norbornadiene, across 11 Rh-based catalysts with bidentate bisphosphine ligands. For these systems, a full conformer ensemble analysis reveals an intriguing finding: the rigid substrate induces conformational freedom in the ligand. This flexibility gives rise to a more diverse conformer landscape, showing a previously overlooked aspect of catalyst-substrate dynamics. Electronic descriptor variations particularly highlight differences between substrate-specific and precatalyst structures. This study suggests that generic precatalyst-like models may lack crucial insights into the conformational freedom of the catalyst. We speculate that such conformational freedom may be a more general phenomenon that can influence the development of generalizable predictive models of computational TM-based catalysis.

This Chapter has been published as: Baidun, M. S.; Kalikadien, A. V.; Lefort, L.; Pidko, E. A. *J. Phys. Chem. C* **2024**, *128*, 7987–7998.[1]

3.1. INTRODUCTION

Different approaches to data-driven models for catalysis exist, categorized as being based on the reaction-specific mechanism,[2–5] or mechanism-agnostic model structures.[6–9] The mechanism-agnostic approach utilized in Chapter 2 involves computing 3D descriptors to characterize catalyst structures.[7,10–12] These descriptors aim to identify ligands that optimize certain attributes such as reactivity or selectivity, often measured in terms of conversion or *ee*. Singh et al.[13] base their descriptor calculations on the ligands of different binaphthyl catalyst families to study the hydrogenation of different substrates bearing C=C and C=N bonds. Zahrt et al.[14] focus on predicting selective catalysis outside the range of selectivity values observed in the training data, with descriptor calculations based on a set of chiral phosphoric acids as model catalysts. Dotson et al.[7] calculate the descriptors based on a [ligand]$PdCl_2$ model system to predict both reactivity and selectivity. In these approaches, the substrate is included in the form of separate molecular parameters, thus considering the substrate separately from the precatalyst structure. The main hypothesis is that descriptors based on such a precatalyst structure, not including the substrate, capture essential catalyst characteristics and can adequately represent the performance.

Relying solely on the precatalyst structure may not be sufficient to predict and understand catalytic behavior as discussed in Chapter 2. Already in the 1980s it was suggested that valuable insights into enantioselectivity could be derived from the reversible substrate coordination, which is suggested to dictate stereoselection.[15–19] Moreover, the lowest energy conformer of the precatalyst structure is often taken as particularly important, which may not be an adequate assumption. Recent studies challenge the focus on the lowest energy conformer, aligning with the 'lock and key hypothesis'.[20] These studies suggest that catalyst flexibility, reflected in the existence of a conformer ensemble, enables adaptable chiral pockets, enhancing selectivity.[21–23] Acknowledging the significance of the substrate and catalyst flexibility for selectivity, it becomes plausible that structural variations induced by the substrate itself play a crucial role. This catalyst flexibility may introduce variations in descriptor values that are important to consider when predicting reactivity and selectivity in descriptor-based catalyst design approaches.

Hence, the primary objective of the present work is to answer the following research question: can we quantify the effects of the specific substrate on catalyst structure and descriptor values compared to the precatalyst with a model substrate? To address this question, the asymmetric hydrogenation reaction of methyl 2-acetamidoacrylate (referred to as S) with 11 Rh-based model catalysts is computationally investigated. These catalysts are referred to as L-Rh-S, with L being 11 different bidentate bisphosphine ligands (Figure 3.1A). Four coordination possibilities of S to the metal center are explored following Knowles quadrants[15] (Figure 3.1C). The L-Rh-S complexes are compared to the generic precatalyst structures with a model diene ligand, norbornadiene (NBD), referred to as L-Rh-NBD. The model ligand serves as a placeholder, fixing ligand geometry in such a way that it can be correlated to the preferred binding of the substrate. A comprehensive conformational search is conducted on all L-Rh-S and L-Rh-NBD structures to establish the effect of

the substrate on catalyst flexibility, followed by Boltzmann-averaged descriptor calculation to assess the effect of the conformer ensembles on substrate-specific descriptors.

The subsequent sections of the paper are organized as follows: the methods section outlines the applied workflow for structure generation, conformer search, structure comparison, and subsequent analyses. The results section starts with the conformer search outcomes of L-Rh-S compared to L-Rh-NBD. This investigation aims to study the influence of the specific substrate on the existence and characteristics of a conformer ensemble. Next, a detailed analysis of the catalyst flexibility is presented, including a separate evaluation of the ligand and substrate contributions to the structural flexibility in a conformer ensemble. Finally, the influence of a conformer ensemble on structural and electronic descriptors is shown. A direct comparison between L-Rh-S and L-Rh-NBD structures provides insights into how the specific substrate alters descriptor values. To close off, we derive some conclusions from these results regarding the effect of the specific substrate on structure and descriptor values, challenging conventional precatalyst-based approaches.

3.2. COMPUTATIONAL METHODS

For this study, we generated substrate-specific and precatalyst structures, performed conformer search and geometry optimization, and compared the conformers based on their energy, structure and descriptor values. The workflow for these steps is illustrated in Figure 3.2. An example is shown for one ligand, but the same workflow was applied on all other ligands. The figure connects the methodological steps with corresponding figures in the results section. Below, the subsequent steps are discussed, followed by the specific computational chemistry details.

3.2.1. WORKFLOW

LIGAND GENERATION

A total of 11 Rh-based catalysts with bidentate bisphosphine ligands were studied in this work, with NBD and S coordinated to the metal center. These ligands were taken from the dataset generated in Chapter 2. All studied ligands have a common backbone structure (Figure 3.1A). With this backbone structure, a total of five catalyst-substrate structures were built manually: two with S in the major coordination, two with S in the minor coordination, and one involving NBD (Figure 3.1D), referred to as L-Rh-S major, L-Rh-S minor, and L-Rh-NBD, respectively. Specific substituents, shown in Figure 3.1A, were introduced to these five structures to generate the unique ligands. This functionalization process was done using ChemSpaX,[24] an open-source software developed by our group, designed to automate the placement of substituents on ligands within a 3D space. This process is shown in step 1 of Figure 3.2, resulting in a total of 44 unique structures with S coordinated to the metal center and 11 structures involving NBD instead.

A

Ligand	R_1	R_2	R_3	R_4	R_5
L1	OCH_3	H	OCH_3	CF_3	H
L2	H	Cl	OCH_3	H	H
L3	H	H	OCH_3	tBu	OCH_3
L4	H	H	OCH_3	CH_3	H
L5	H	H	OCH_3	tBu	H
L6	H	H	OCH_3	tBu	H
L7	OCH_3	H	OCH_3	H	H
L8	OCH_3	H	OCH_3	CH_3	H
L9	OCH_3	H	OCH_3	tBu	OCH_3
L10	H	H	OCH_3	H	H
L11	H	H	OCH_3	F	CF_3

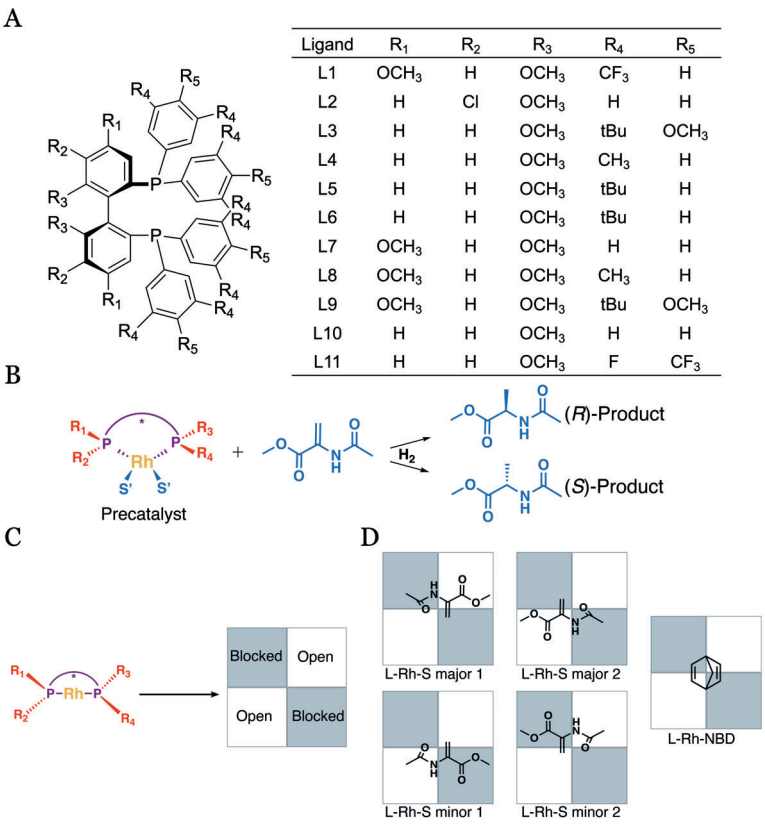


Figure 3.1.: Backbone of the 11 studied ligands with their respective substituents (A), catalyzing the hydrogenation of methyl 2-acetamidoacrylate (denoted as S) to the (*S*)- and (*R*)-products (B). The ligand groups lead to two open and two blocked quadrants (C), yielding four different coordination modes for S and one for NBD (D). According to the quadrant diagram, two S coordinations are less hindered and two are more hindered coordinations, referred to as the major and minor coordinations, respectively.

Geometry optimization & conformer search

The generated structures underwent DFT geometry optimization (see Section 3.2.2 for details), yielding four structures per ligand with different S coordinations, alongside one structure featuring NBD coordination (step 2). Conformational exploration was conducted on the 55 DFT-optimized structures, as shown in step 3, with subsequent DFT-geometry optimization on all generated conformers (step 4). Specific details regarding conformer search are described in Section 3.2.2. Within each conformer set, the DFT-based energies of the conformers (data in step 4) were compared to the DFT-based energies of the CREST[25,26] input structures (data

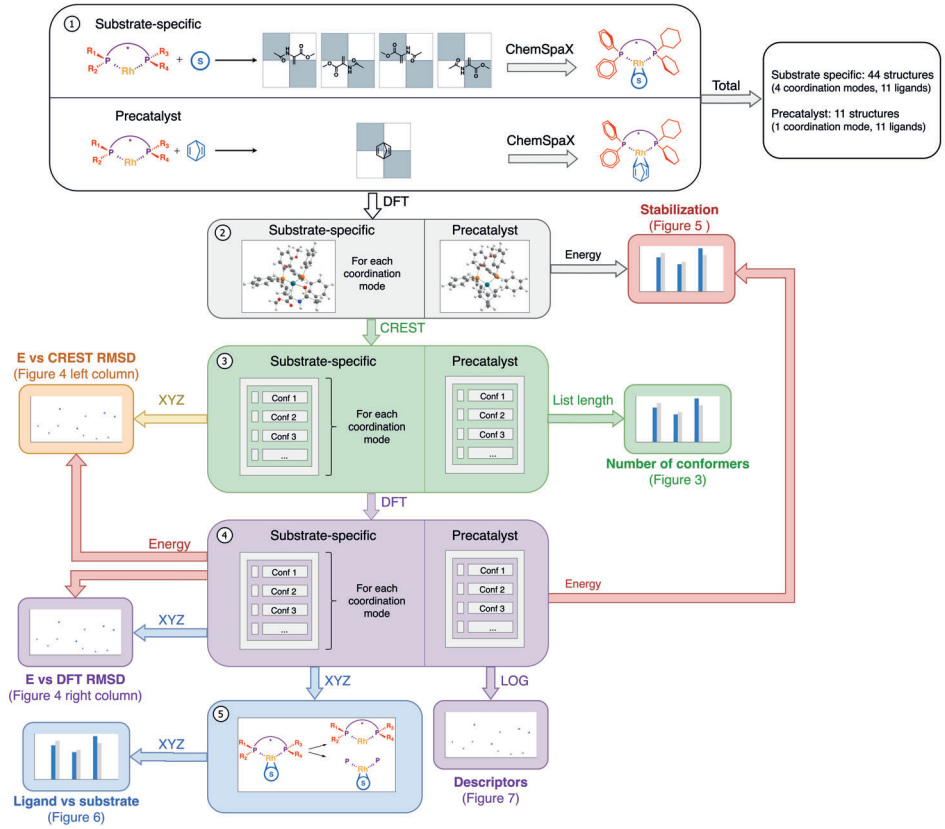


Figure 3.2.: Summary of the applied workflow. The entire workflow, shown as an example for one ligand, is employed on all 11 ligands. In step 1, the unique ligands featuring different substrate coordination modes are generated. In step 2, these structures are DFT-optimized, followed by conformer search in step 3. All generated conformers are DFT-optimized in step 4. In step 5, the DFT-optimized structures are separated into a substrate and ligand part for individual analysis. The workflow shows how the methodological steps are connected to figures in the remainder of the paper.

in step 2) to assess the degree of stabilization achieved through conformer search. Furthermore, the geometry-optimized conformers were compared to each other to assess variability within the generated conformer ensembles. This was done by analyzing the DFT-optimized energies together with structural differences. The structural differences were calculated on the structures both before (data in step 3) and after DFT optimization (data in step 4), with details on the structural difference calculations presented below.

Structural differences

Structural variations were assessed by calculating the minimal root-mean-square deviation (RMSD)[27] between conformers. The structural variations were computed relative to the conformer with the lowest DFT-based energy within the respective conformer set. All hydrogen atoms were ignored in the calculation of RMSD values with the option *–no-hydrogen*. These structural difference calculations were performed at three stages: on the conformer structures generated by CREST (data in step 3), on the DFT-optimized conformer structures (data in step 4), and on isolated ligand and substrate components (*vide infra*, data in step 5).

Ligand vs substrate

Following DFT optimization, step 5 illustrates how each conformer structure is separated into a substrate part and a ligand part. The ligand part comprises the atoms of the bidentate ligand and the metal center, while the substrate part consists of the substrate molecule (either S or NBD), the metal center, and the two P donor atoms. The inclusion of donor atoms aimed to establish an orientation reference for the substrate. Subsequently, structural difference calculations, as previously described, were performed on the isolated ligand and substrate parts.

Descriptor calculation

Descriptor calculations were performed on all DFT-optimized conformers (data in step 4) using the in-house developed method OBeLiX (Open Bidentate Ligand eXplorer).[28] OBeLiX is an automated and reproducible workflow for computing transition metal-based catalyst descriptors. The workflow accommodates descriptor calculations from xyz files, Gaussian[29] log files, or CREST output folders, with this work specifically relying on Gaussian log files. Using Morfeus[30] and cclib,[31] the workflow computes 75 descriptors categorized as steric, geometric, electronic, or thermodynamic.

This work focuses on a subset of five descriptors: NBO charge on Rh, NBO charge on the donor atoms of the ligand, buried volume on Rh, buried volume on the donor atoms of the ligand, and the HOMO-LUMO gap. These calculated descriptors were Boltzmann-weighted and averaged. With the goal of quantifying populations within the given substrate coordination ensembles, we have calculated the Boltzmann factors separately for L-Rh-S major, L-Rh-S minor, and L-Rh-NBD structures. The Boltzmann weights were determined using the formula $w = e^{-(E - E_{min})/k_b T}$. In this formula, w represents the weight, E is the energy obtained from DFT calculations, k_b is the Boltzmann constant, T is the temperature (289 K) and E_{min} is the corresponding DFT-based energy of the lowest energy conformer of L-Rh-S major, L-Rh-S minor, or L-Rh-NBD. The resulting weights were used to calculate the Boltzmann-averaged descriptor values, along with standard deviation values.

3.2.2. Computational methods details

All geometry optimization calculations were performed using the Gaussian 16 C.01 suite[29]. The calculations were executed at the PBE0[32]-D3(BJ)[33]/def2-SVPP[34] level

of theory. The chosen combination of functional and basis set have shown reliable results for similar transition metal-based complexes accompanied with low computational costs.[24,35] The nature of each stationary point was confirmed via frequency analysis. In case imaginary frequencies were present, these were removed with the pyQRC python package,[36,37] followed by an additional geometry optimization. All calculations were carried out in gas phase. A Natural Population Analysis (NPA) was performed using the NBO program version 3.1[38] as implemented in Gaussian.

Conformer search was done using the CREST software (version 2.12)[25,26] with xtb (version 6.6.1)[39] optimization. To efficiently screen the configurational space and find low-lying conformers, CREST makes use of MD simulations with a bias potential.[26] Generated conformer ensembles were selected within 6 kcal/mol of the lowest energy conformer, and calculations were carried out at the GFN2-xTB//GFN-FF level of theory. We evaluated conformer search for a set of representative ligand structures with the various GFNn-xTB methods, with the results in the ESI (Section S1.1) showing that the GFN2-xTB//GFN-FF method was the most suitable option in terms of computational cost and avoiding structural artifacts. Throughout this work, conformers generated with this method are referred to as "CREST" conformers. To preserve the chirality of the ligand, the aromatic rings on the chiral axis were fixed during xtb optimization and conformer search. Additional details about chirality preservation are provided in the ESI (Section S1.2). The *–noreftopo* option was employed to disable topology checks before conformer search, ensuring proper treatment of transition-metal complexes. Following conformer search, all generated conformers were checked on chirality, and L-Rh-S structures were checked on maintaining the initial substrate coordination mode. Two conformers from the L11-Rh-S major 1 conformer set were removed due to rotation of the substrate to another coordination mode.

3.3. RESULTS AND DISCUSSION

The results section is divided into two parts. The first part addresses the conformer search outcomes of L-Rh-S compared to L-Rh-NBD. The second part delves into the individual evaluation of ligand and substrate contributions to the configurational freedom, followed by a discussion of a set of descriptors calculated on the conformer ensembles. In the following, structural differences within a given metal-ligand complex are referred to as the flexibility of the system.

3.3.1. CONFORMER SEARCH

A comparative study of conformer ensembles involving S and NBD was performed on a total of 44 L-Rh-S and 11 L-Rh-NBD complexes. S can coordinate in four ways to the metal center, with two major and two minor coordinations (Figure 3.1D), yielding four input structures for conformer search and four separate conformer ensembles. NBD, having a single coordination mode, results in one input structure and conformer ensemble instead of four. Figure 3.3 shows the total number of conformers generated for L-Rh-S and L-Rh-NBD complexes. The two major and two

minor conformer ensembles are stacked on top of each other, and the cumulative major and minor counts are shown next to each other.

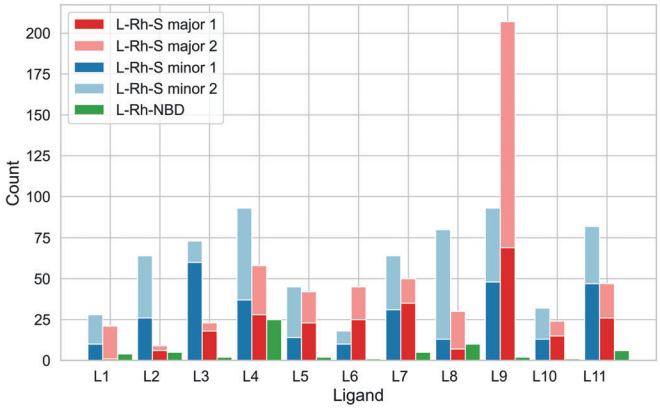


Figure 3.3.: Number of conformers from CREST output for L-Rh-S major, L-Rh-S minor, and L-Rh-NBD. For L-Rh-S major and minor, two major and two minor coordinations are considered, resulting in two conformer sets each, indicated by "1" and "2".

Across the ligands, we can first consider the two L-Rh-S major and two L-Rh-S minor conformer ensembles separately. For instance, L-Rh-S major for L3 exhibits a significant difference in conformer count between S major 1 and S major 2, with 18 and 5 conformers, respectively. This discrepancy is evident in ligands L1, L8, and L9 as well, despite having the same substrate coordination mode. Similar observations can be made when comparing the two minor conformer ensembles within each ligand, best visualized in the conformer ensembles of L3 and L8.

Secondly, we can compare L-Rh-S conformer ensembles to L-Rh-NBD ensembles for each ligand. A consistent observation is that at least one major and one minor set of conformers involving S contains significantly more conformers compared to the ensemble of the complex with NBD. This implies that the L-Rh-NBD precatalyst structure is rather rigid, while substrate coordination induces flexibility, resulting in a larger conformer ensemble. However, merely the count of generated conformers itself does not provide insight into ensemble diversity or the origin of catalyst flexibility.

To assess the diversity of the generated conformers, the structural differences are compared before and after DFT optimization, as illustrated in Figure 3.4. The y-axis in each subplot reflects the DFT-based energy differences, while the x-axis represents the structural differences, both relative to the lowest energy conformer within each conformer set. The structural differences are represented by the RMSD value calculated either on CREST-based (left column) or DFT-based (right-column) conformer structures. The upper and lower rows show the subplots regarding L4-Rh-S and L5-Rh-S, respectively. These ligands are chosen to analyze in detail as

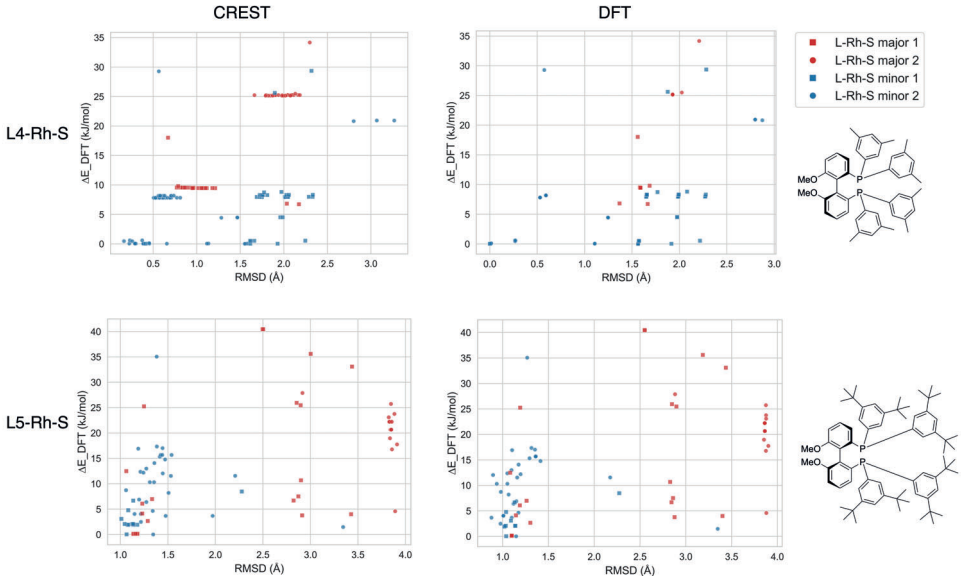


Figure 3.4.: Relative energies plotted against structural variations (RMSD) for the conformer ensembles of L4-Rh-S (upper row), and L5-Rh-S (lower row). RMSD values are obtained from CREST (left column) or DFT structures (right column), and the relative energy is DFT-based in all four subplots. Within each conformer ensemble, the conformer with the lowest DFT-based energy is taken as a reference point for the ΔE and RMSD values.

illustrative cases, with similar analyses for ligands L1-L3, L6-L11 presented in the ESI (Section S2).

When comparing RMSD values of the DFT-optimized structures to the values of the CREST-based structures, one can see whether the conformers identified by CREST remain true, distinct minima after DFT optimization. For L5-Rh-S (subplots in lower row), minimal differences between the left and right subplots indicate that the structural differences identified by CREST are preserved after DFT optimization. For this ligand, a total of 87 conformers identified by CREST converge to 62 distinct conformers after DFT optimization. L4-Rh-S (subplots in upper row) shows a contrasting picture. While CREST reveals distinct conformers close in RMSD values, DFT optimization converges many of these conformers to the same minima. Here, 151 conformers identified by CREST converge to 35 distinct conformers. These findings underscore the value of DFT optimization for a comprehensive understanding of the system's flexibility.

Analyzing the structural differences after DFT-optimization (right column) reveals many structures with varying RMSD values but minimal energetic differences. For instance, the DFT-optimized structures of L4-Rh-S (upper right subplot) contain several structures around ΔE of 0 kJ/mol, but with vastly different RMSD. These instances pose a challenge in selecting conformers for descriptor-based catalyst

design. Energetically identical structures with significant structural variations may yield diverging descriptor values, influencing predictions. Additionally, conformer ensembles with varying energy values but similar RMSD values can be found as well. An example are the conformers in L5-Rh-S with RMSD values around 4 Å and ΔE values spreading from 5 to 26 kJ/mol. These findings emphasize the importance of careful conformer consideration, especially given that a DFT uncertainty as low as 5 kJ/mol can invert predicted enantioselectivity trends.[40] A similar analysis for ligands L1-L11 with NBD coordination is presented in the ESI, highlighting that the majority of generated conformers exhibit ΔE values below 5 kJ/mol with a few outliers. An intriguing insight emerges from these observations - the inability of the precatalyst structure to capture the system's flexibility, translating into fewer conformers with less variability.

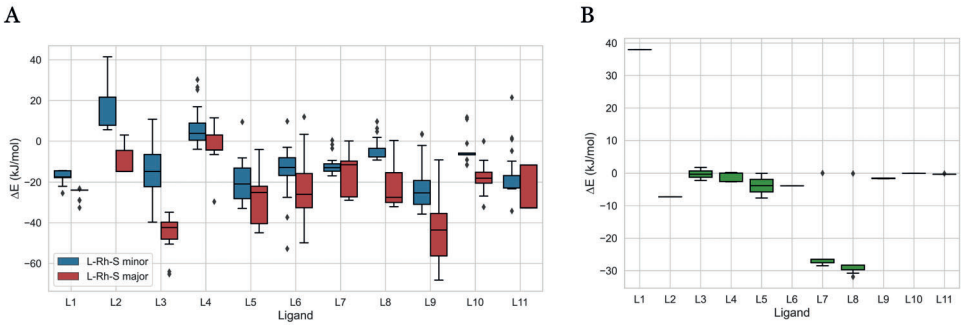


Figure 3.5.: Energy range of the conformer ensembles after DFT optimization per ligand for L-Rh-S major and L-Rh-S-minor (A) and L-Rh-NBD (B). Per conformer ensemble, the DFT-based energy of the structure used as CREST input is chosen as the baseline.

Conformer search is not only useful for generating a conformer ensemble, but it is also essential in locating the global minimum.[25] Figure 3.5 shows the energy distribution within each ligand's conformer ensemble relative to the energy of the CREST input structure. Except for L2-Rh-S minor in Figure 3.5A, conformer search with CREST successfully identifies conformers significantly lower in energy than the original structure on which conformer search was performed, surpassing 60 kJ/mol in some cases. With the exception of L1 this trend persists in Figure 3.5B for L-Rh-NBD, with CREST identifying conformers that are significantly lower in energy. In addition, the L-Rh-NBD conformers generally exhibit lower energies compared to the original structure, whereas the L-Rh-S conformer sets contain structures with both higher and lower stability than the original structure.

Comparing the number of generated conformers with the degree of energy minimization in the conformer ensembles offers insights into whether a high number of generated conformers corresponds to more significant stabilization. First, we can compare L-Rh-S major and L-Rh-S minor within each ligand. Taking L3 as an example from Figure 3.5A, L-Rh-S major exhibits a greater stabilization than L-Rh-S minor, despite the latter having a higher total number of conformers (Figure

3.3). This trend holds for 9 out of 11 ligands, where the coordination with a lower number of conformers exhibits greater stabilization than the coordination with a higher number of conformers, with the exception of L9 and L11. Secondly, we can compare all conformer ensembles across the ligands. For instance, L4-Rh-S minor is stabilized by approximately 5 kJ/mol, despite the discovery of nearly 100 conformers. Conversily, L9-Rh-S major is the most stabilized structure, correlating with the highest number of conformers. This indicates that the trend of high stabilization with a low number of conformers does not hold across all ligands.

A similar comparison can be drawn between the number of generated conformers and the energy range within the generated conformer ensembles. This analysis aims to discern whether the number of conformers offers insights into energetic variations in a conformer set. However, no consistent trend emerges across the ligands. While L9-Rh-S major displays a large number of conformers spanning from -68 to -9 kJ/mol, L4-Rh-S minor demonstrates the opposite. The large number of conformers results in a narrow energy range with a few outliers. This is corroborated by the top two subplots of Figure 3.4, where numerous conformers identified by CREST converge after DFT optimization to structures within a narrow range of 10 kJ/mol with a few outliers. These findings underscore the intricate relationship between energy minimization, energetic variability, and the number of conformers obtained with CREST.

3.3.2. CATALYST FLEXIBILITY

The previous section delved into the generation of extensive conformer ensembles with the inclusion of the specific substrate, revealing significant differences both in energy and structure. This section explores catalyst flexibility in detail. First, we disentangle the separate contributions of the ligand and substrate to the structural variations within the conformer ensembles. Next, the impact of these structural differences on a set of descriptors is examined.

LIGAND VS SUBSTRATE CONTRIBUTIONS

To discern the separate contributions of the ligand and substrate to structural variations within one metal-ligand system, the DFT-optimized structure of each conformer is separated into a ligand part and a substrate part, as outlined in the methods section. For each conformer, the structural difference of the substrate and ligand is calculated relative to the respective parts of the lowest energy conformer within the corresponding conformer set. These structural differences are represented by the RMSD values on the y-axis in Figures 3.6A (L-Rh-S major), B (L-Rh-S minor), and C (L-Rh-NBD) for each ligand.

The substantial increase in conformers introduced by the specific substrate compared to NBD (Figure 3.3) implies that the substrate induces conformational variation. The origin of this flexibility, whether from the ligand or the substrate, is clarified by examining the individual RMSD contributions in Figure 3.6. The substrate RMSD value rarely exceeds 1 Å across different ligands and major/minor coordinations, indicating a high degree of rigidity. A similar trend is observed when

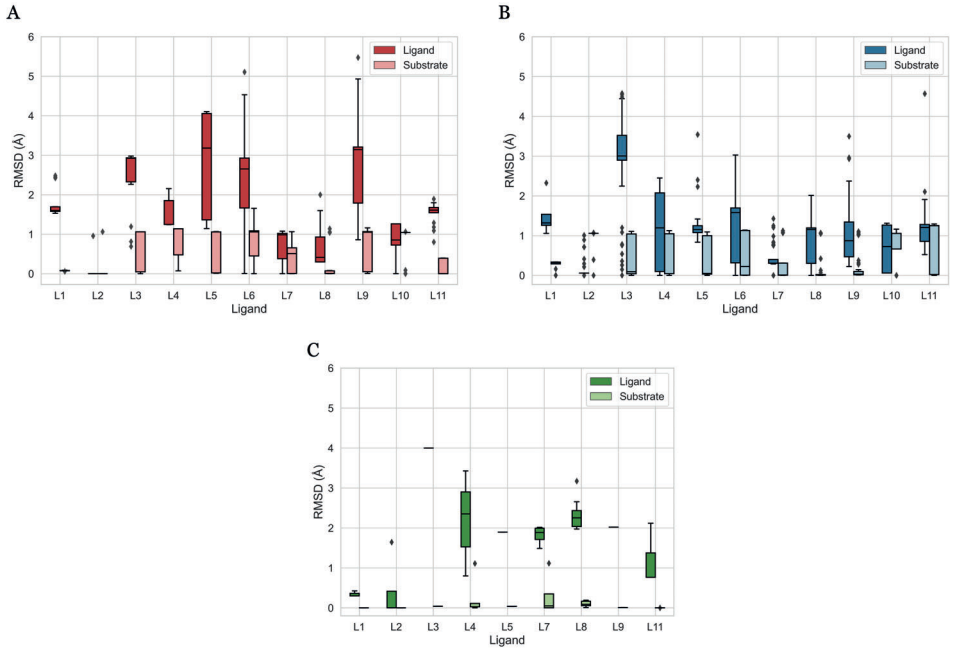


Figure 3.6.: RMSD range of the conformer ensembles after DFT optimization per ligand for L-Rh-S major coordination (A), L-Rh-S minor (B), and L-Rh-NBD (C). Within each conformer set, the ligand and substrate are isolated and the RMSD values of the ligand and substrate part are calculated relative to the conformer with the lowest DFT-based energy.

NBD is coordinated (Figure 3.6C), where the substrate displays high rigidity with consistently low RMSD values.

Contrasting to the consistently low substrate RMSD, the ligand exhibits broad RMSD ranges, reaching values up to 5 Å. These findings suggest that the ligand dynamically adapts to the rigid substrate. This ligand flexibility is also seen in the L-Rh-NBD structures, although to a lesser extent, given the generation of fewer conformers when NBD is involved. The comparison to NBD suggests that the specific substrate induces flexibility in the ligand. As an asymmetric structure, S may induce preferential orientations of the ligand to optimize non-covalent interactions. In contrast, the highly symmetric NBD structure does not lead to specific orientation preferences for the ligand. The substrate-induced flexibility yields a higher number of conformers in L-Rh-S complexes compared to L-Rh-NBD, even though the substrate exhibits rigidity in both cases. These findings on ligand flexibility support the observations of Crawford and Sigman, where the ligand's adaptability is suggested to stabilize intermediates and transition states throughout the catalytic cycle.[21]

The alignment between ligand flexibility and conformer diversity gains support

when comparing data in Figure 3.6 to Figure 3.4. Taking L5 as an example, Figures 3.6A and B reveal consistent substrate rigidity in both L-Rh-S major and L-Rh-S minor structures. However, there is a distinctive difference in ligand flexibility. L-Rh-S major exhibits a broad range of RMSD values (1.14 to 4.10 Å), while L-Rh-S minor demonstrates a narrower RMSD range (around 1.2 Å) with some outliers. This observation alligns with the structural differences after DFT optimization for L5-Rh-S in Figure 3.4 (lower right subplot). Here, L-Rh-S minor points are clustered within a narrow RMSD range with a few outliers, while L-Rh-S major points are spread across a broader range of RMSD values. Together, these findings strengthen the hypothesis that structural variability within a conformer set is primarily driven by ligand flexibility. The detailed analysis of the origin of energy differences within the conformer ensembles is presented in Section S3 of the ESI.

THE EFFECT OF FLEXIBILITY ON DESCRIPTORS

Given the significant structural and energetic variations within the conformer ensembles, the focus now shifts to understanding their impact on descriptors. Out of a total of 75 descriptors generated by OBeLiX,[28] we sought to test the impact of catalyst flexibility on representative descriptors for the transition metal complexes. We have selected a subset of five descriptors for detailed analysis, including the buried volume on Rh, buried volume on donor atoms, NBO charge on Rh, NBO charge on donor atoms, and the HOMO-LUMO gap. These five descriptors serve as an illustrative subset: the buried volume serves as a general steric descriptor,[41] the NBO charge has been previously used in catalytic investigations,[9,42–44] and the HOMO-LUMO gap represents kinetic stability.[9] Figure 3.7 summarizes the Boltzmann-averaged descriptor values, showing values for L-Rh-S major, L-Rh-S minor, and L-Rh-S-NBD, with error bars indicating standard deviations. Note that descriptor values involving donor atoms represent the average for two P atoms.

A first look at the figure reveals varying impacts of the conformer ensembles on different descriptors. The Rh buried volume (Figure 3.7A), donor buried volume (Figure 3.7B), and charge on the donor atoms (Figure 3.7D) are minimally affected by the conformer ensemble, with maximum standard deviation values of 0.008, 0.003, and 0.001, respectively. Comparing the results for L-Rh-S to those for L-Rh-NBD, generally smaller standard deviations are observed with L-Rh-NBD. L4, the ligand with the highest number of conformers involving NBD, shows the largest overall variety. The plots regarding the charge on Rh (Figure 3.7C) and the HOMO-LUMO gap (Figure 3.7E) warrant independent analysis due to significant influences from 1) the conformer ensemble, 2) initial coordination of S, and 3) the choice of the coordinating substrate.

Within a conformer set, interactions between the metal center and the substrate can influence the charge on the metal center significantly. The expectation is that the rigid substrate should manifest as a small standard deviation in the charge on Rh for the conformer ensembles of L-Rh-S. Indeed, this is seen in the maximum standard deviation of 0.03 a.u. for the L-Rh-S minor conformer ensemble of L1. However, when examining different substrate coordinations, significant differences emerge. With four different coordinations of S leading to four conformer sets (two

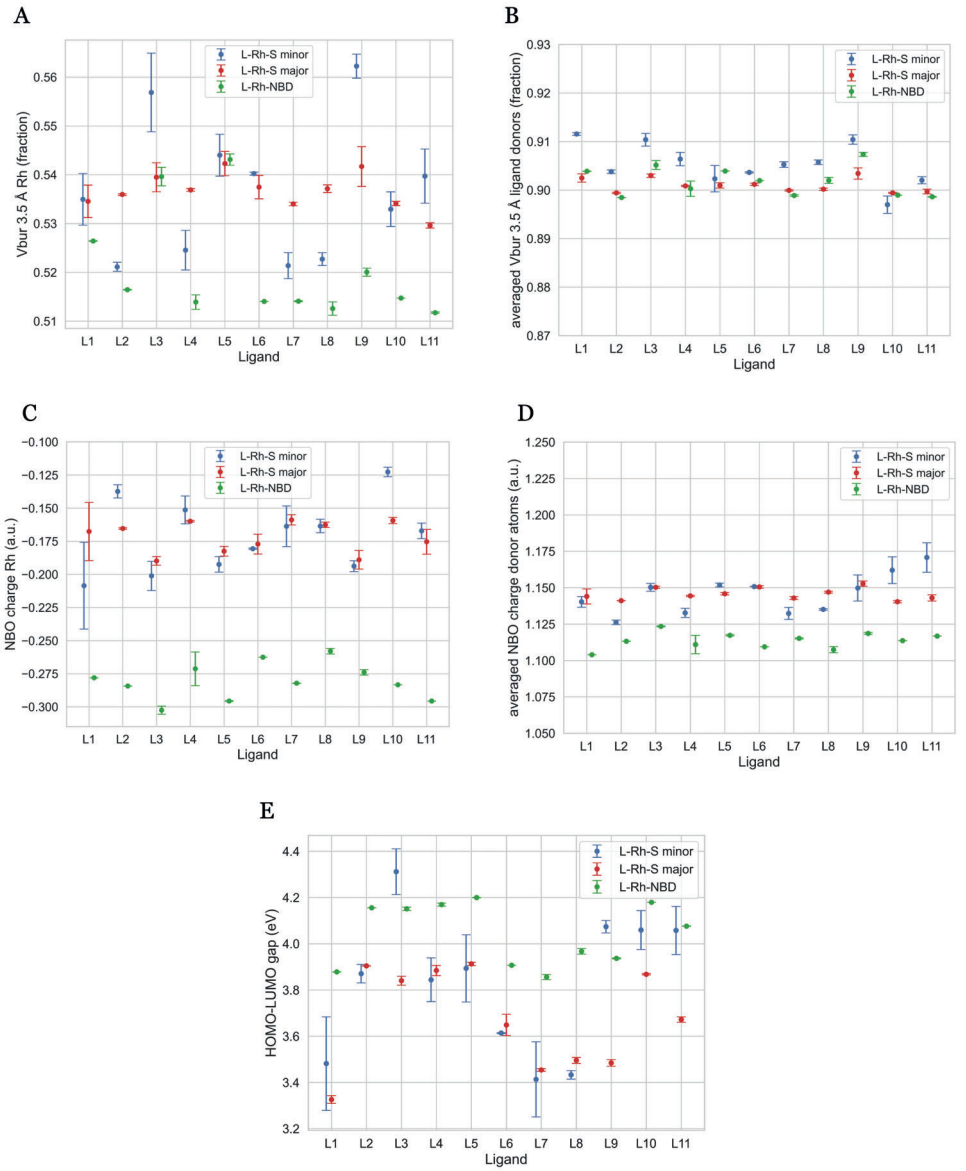


Figure 3.7.: Calculated values for the Rh NBO charge (A), donor NBO charge (B), Rh buried volume (C), donor buried volume (D), and HOMO-LUMO gap (E) Boltzmann averaged over the conformer sets of L-Rh-S minor, L-Rh-S major, and L-Rh-NBD. The descriptors concerning the donor atoms are an average of the two P-atoms.

L-Rh-S major and two L-Rh-S minor), a comparison of the lowest energy conformers reveals notable variations. For instance, the two lowest energy conformers of L-Rh-S major of L1 exhibit Rh NBO charge values of -0.185 and -0.180 a.u., while the lowest energy conformers of L-Rh-S minor of L1 exhibit values of -0.188 and -0.267 a.u. The difference of nearly 0.09 a.u. between major and minor coordination conformers within the same ligand is statistically significant given the maximum standard deviation of 0.03 a.u. across ligands. These observations emphasize that conformer search, due to the substrate's rigidity, has minimal impact on the metal center charge, while the initial coordination of the substrate significantly affects this descriptor value.

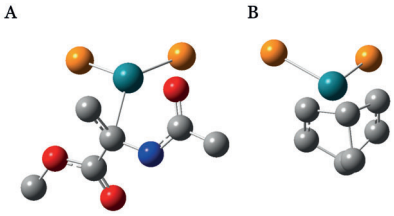


Figure 3.8.: S coordination in the overall lowest energy conformer of L1-Rh-S (A) and NBD coordination in the lowest energy conformer of L1-Rh-NBD (B), with Rh NBO charge values of -0.180 and -0.278 a.u., respectively. For clarity, hydrogen atoms and all atoms of the bidentate ligand except P omitted. Color coding of the atoms shows Rh (turquoise), P (orange), O (red), N (blue) and C (grey) atoms.

The metal center charge values of L-Rh-S can be compared to the L-Rh-NBD. The highly rigid NBD molecule leads to a maximum standard deviation of only 0.01 a.u. for ligand L4, revealing consistently lower L-Rh-NBD charge values for all ligands. The diverging descriptor values may be attributed to different substrate coordination interactions, as visualized in Figure 3.8. The coordination of both π-systems in the symmetric NBD structure (Figure 3.8B) is associated with a lower NBO charge of -0.278 a.u., whereas the more distorted coordination of S with the metal center (Figure 3.8A) leads to a higher NBO charge of -0.180 a.u.

Variations in the global HOMO-LUMO gap descriptor likely correlate with those of the charge on Rh. As discussed for L1, the lowest energy conformers of the four L-Rh-S coordinations show metal center charges ranging from -0.180 to -0.267 a.u. This variation is also reflected in diverging HOMO-LUMO gap values, and visualized with the HOMO and LUMO orbitals in Figure 3.9. The figure illustrates different spatial distributions of HOMO orbitals (left column) and LUMO orbitals (right column) for the lowest energy conformer of L1-Rh-S major (upper row), and L1-Rh-S minor (lower row), with charges on Rh of -0.180 and -0.267 a.u., respectively. In L1-Rh-S major, the orbitals have more spatial overlap, reflected in a lower HOMO-LUMO gap of 3.31 eV. Conversely, L1-Rh-S minor reveals less spatial overlap and a higher HOMO-LUMO gap of 3.86 eV. With a maximum standard deviation for the HOMO-LUMO gap of 0.2 eV within a conformer ensemble,

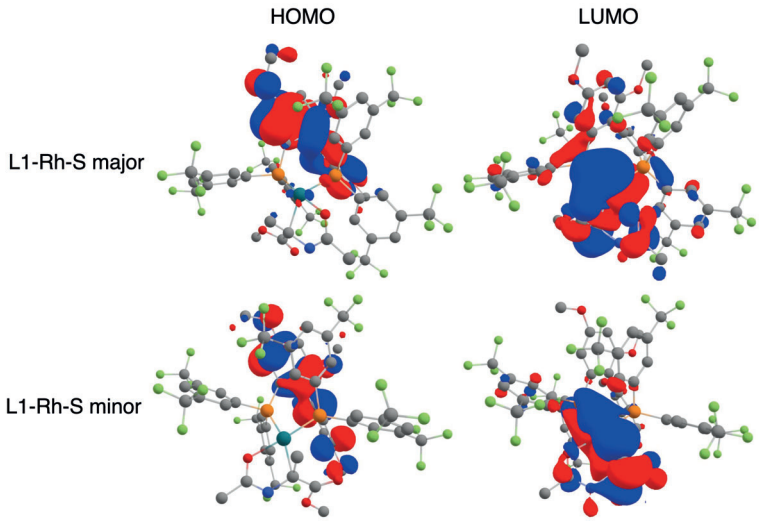


Figure 3.9.: Visualization of the HOMO orbitals (left column) and LUMO orbitals (right column) for the lowest energy conformer of L1-Rh-S major (upper row) and L1-Rh-S minor (lower row). For clarity, hydrogen atoms are not shown. Color coding of the atoms shows Rh (turquoise), P (orange), O (red), N (blue), C (grey), and F (green) atoms.

a difference exceeding 0.5 eV across different substrate coordination modes is significant, supporting previous claims that the electron distribution is sensitive to subtle conformational changes.[45] The pronounced influence of the substrate on this global descriptor is further evidenced by the comparison of HOMO-LUMO gaps for L-Rh-S and L-Rh-NBD (Figure 3.7E).

3.4. Conclusion

Herein we presented a comprehensive computational analysis of the impact of catalyst-substrate interactions on catalyst structure and the descriptors derived from it. Focusing on a family of Rh bisphosphine catalysts (11 members), we explored a representative model catalyst system of asymmetric acetamide hydrogenation using Me-2-acetamidoacrylate (denoted as S) as a model substrate, comparing it to the precatalyst structure with norbornadiene (NBD). The primary objective was to study the capability of mechanistically agnostic models to capture effects that are purely substrate-specific.

Conformer search was conducted to assess the influence of substrate inclusion on the size of the generated conformer ensembles. We found that the asymmetric nature of S induces catalyst flexibility, (i.e., conformational freedom within the metal-ligand system), reflected in the number of generated conformers. The maximum ensemble size involving S surpassed five times that of the maximum

ensemble size involving NBD. These findings indicate that the precatalyst structure may lack critical information about the system's flexibility, underscoring the impact of the substrate on the overall conformational landscape of the studied systems.

Delving deeper into the substrate-induced catalyst flexibility, the ligand and substrate contributions were investigated separately. We unveiled that the specific substrate is rather rigid, similar to NBD. Both structures show consistently low RMSD values below or around 1 Å compared to the lowest energy conformer per conformer ensemble. The ligand, on the other hand, exhibits remarkable flexibility, with RMSD values reaching up to 5 Å. This ligand flexibility challenges the traditional 'lock-and-key' model, supporting recent studies that highlight the importance of flexibility for achieving high selectivity and reactivity.

Finally, the influence of the found ligand flexibility was evaluated on a set of descriptors, underscoring the significance of considering the entire conformer ensemble, rather than focusing on solely the most stable conformer. While structural properties showed minimal sensitivity to various conformers, electronic properties, such as the charge on Rh and HOMO-LUMO gap, exhibited more substantial variations. More importantly, these descriptors were significantly influenced by the initially chosen substrate coordination mode. The charge on Rh can differ by almost 0.1 a.u. depending on the substrate coordination mode, leading to a HOMO-LUMO gap difference that exceeds 0.5 eV. These discrepancies show the sensitivity of these electronic properties to the coordination environment, influenced not only by the chosen coordination but also by the specific substrate compared to NBD. Notably, observed differences in electronic properties between the substrate-specific and precatalyst structures may not only impact enantioselectivity but also conversion, suggesting that substrate inclusion may influence descriptor-based catalyst design.

The detailed analysis of substrate-specific conformer ensembles, compared to the precatalyst structure, has provided valuable insights into the catalyst-substrate interactions of a family of Rh bisphosphine catalysts. With catalyst-substrate interactions often being omitted in conventional descriptor-based catalyst design strategies, this study may offer a starting point to understand the origin of ligand flexibility and its effects on descriptors. Future investigations could build upon our findings by exploring a broader spectrum of ligands and substrates. For instance, examining another symmetric, non-cyclic substrate could confirm whether substrate rigidity is an inherent property or arises from the asymmetric character of the substrate. Furthermore, integrating the substrate-specific and ensemble-averaged descriptors into data-driven catalyst design may deepen our understanding of the substrate's significance. Such investigations may help to advance the understanding of catalyst-substrate interactions in asymmetric hydrogenation, possibly contributing to more informed and effective catalyst design strategies.

DATA AVAILABILITY

The Supporting Information file for this Chapter is available at: https://doi.org/10.1021/acs.jpcc.4c01631. The following information is contained in the SI file: selection of GFNn-xTB for conformer search (S1.1); chirality and substrate coordination issues during conformer search (S1.2); supporting analyses of energy vs structural differences (S2); origin of energy differences (S3).

All inputs and outputs for DFT and CREST calculations, datasets and code are available together with an extensive readme via 4TU.ResearchData (https://doi.org/10.4121/ce7fb6ee-a10c-44c8-91c0-1d55d55882e3). The OBeLiX code for descriptor calculation is available via our Github page (https://github.com/epics-group).

CONTRIBUTIONS

M.S. Baidun and **A.V. Kalikadien** contributed equally to this work. **M.S. Baidun:** Investigation, Methodology, Formal analysis, Validation, Data Curation, Software, Writing - Original Draft, Writing - Review & Editing, Visualization **A.V. Kalikadien:** Investigation, Methodology, Conceptualization, Software, Writing - Original Draft, Writing - Review & Editing, Project administration **L. Lefort:** Supervision, Conceptualization, Resources, Funding acquisition, Writing - Review & Editing **E.A. Pidko:** Supervision, Conceptualization, Resources, Funding acquisition, Writing - Review & Editing, Project administration

REFERENCES

(1) Baidun, M. S.; Kalikadien, A. V.; Lefort, L.; Pidko, E. A. *J. Phys. Chem. C* **2024**, *128*, 7987–7998.

(2) Hansen, E.; Rosales, A. R.; Tutkowski, B.; Norrby, P.-O.; Wiest, O. *Acc. Chem. Res.* **2016**, *49*, 996–1005.

(3) Houk, K. N.; Cheong, P. H.-Y. *Nature* **2008**, *455*, 309–313.

(4) Cheong, P. H.-Y.; Legault, C. Y.; Um, J. M.; Çelebi-Ölçüm, N.; Houk, K. N. *Chem. Rev.* **2011**, *111*, 5042–5137.

(5) Ahn, S.; Hong, M.; Sundararajan, M.; Ess, D. H.; Baik, M.-H. *Chem. Rev.* **2019**, *119*, 6509–6560.

(6) Kozlowski, M. C.; Dixon, S. L.; Panda, M.; Lauri, G. *J. Am. Chem. Soc.* **2003**, *125*, 6614–6615.

(7) Dotson, J. J.; Van Dijk, L.; Timmerman, J. C.; Grosslight, S.; Walroth, R. C.; Gosselin, F.; Püntener, K.; Mack, K. A.; Sigman, M. S. *J. Am. Chem. Soc.* **2023**, *145*, 110–121.

(8) Freeze, J. G.; Kelly, H. R.; Batista, V. S. *Chem. Rev.* **2019**, *119*, 6595–6612.

(9) Durand, D. J.; Fey, N. *Chem. Rev.* **2019**, *119*, 6561–6594.

(10) Soyemi, A.; Szilvási, T. *Dalton Trans.* **2021**, *50*, 10325–10339.

(11) Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C. *Chem. Sci.* **2021**, *12*, 6879–6889.

(12) Niemeyer, Z. L.; Milo, A.; Hickey, D. P.; Sigman, M. S. *Nat. Chem.* **2016**, *8*, 610–617.

(13) Singh, S.; Pareek, M.; Changotra, A.; Banerjee, S.; Bhaskararao, B.; Balamurugan, P.; Sunoj, R. B. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 1339–1345.

(14) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631.

(15) Knowles, W. S. *Acc. Chem. Res.* **1983**, *16*, 106–112.

(16) Koenig, K. E.; Sabacky, M. J.; Bachman, G. L.; Christopfel, W. C.; Bamstorff, H. D.; Friedman, R. B.; Knowles, W. S.; Stults, B. R.; Vineyard, B. D.; Weinkauff, D. J. *Ann. N.Y. Acad. Sci.* **1980**, *333*, 16–22.

(17) Landis, C. R.; Halpern, J. *J. Am. Chem. Soc.* **1987**, *109*, 1746–1754.

(18) Gridnev, I. D.; Imamoto, T. *Chem. Commun.* **2009**, 7447.

(19) Halpern, J. *Science* **1982**, *217*, 401–407.

(20) Fischer, E. *Ber. Dtsch. Chem. Ges.* **1894**, *27*, 3189–3232.

(21) Crawford, J.; Sigman, M. *Synthesis* **2019**, *51*, 1021–1036.

(22) Gallegos, L. C.; Luchini, G.; St. John, P. C.; Kim, S.; Paton, R. S. *Acc. Chem. Res.* **2021**, *54*, 827–836.

(23) Shao, H.; Chakrabarty, S.; Qi, X.; Takacs, J. M.; Liu, P. *J. Am. Chem. Soc.* **2021**, *143*, 4801–4808.

(24) Kalikadien, A. V.; Pidko, E. A.; Sinha, V. *Digital Discov.* **2022**, *1*, 8–25.

(25) Grimme, S. *J. Chem. Theory Comput.* **2019**, *15*, 2847–2862.

(26) Pracht, P.; Bohle, F.; Grimme, S. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(27) Kromann, J. C. Calculate Root-mean-square deviation (RMSD) of Two Molecules Using Rotation, GitHub, v1.3.2 (accessed December 15, 2023), http://github.com/charnley/rmsd.

(28) Kalikadien, A. V.; Mirza, A.; Hossaini, A. N.; Sreenithya, A.; Pidko, E. A. *ChemPlusChem* **2024**, e202300702.

(29) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16 Revision C.01/C.02, 2016.

(30) Jorner, K.; Turcani, L. kjelljorner/morfeus: v0.7.2, version v0.7.2, (accessed December 15, 2023), 2022.

(31) O'Boyle, N. M.; Tenderholt, A. L.; Langner, K. M. *J. Comput. Chem.* **2008**, *29*, 839–845.

(32) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(33) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

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

(35) Sinha, V.; Laan, J. J.; Pidko, E. A. *Phys. Chem. Chem. Phys.* **2021**, *23*, 2557–2567.

(36) Goodman, J. M.; Silva, M. A. *Tetrahedron Lett.* **2003**, *44*, 8233–8236.

(37) Silva, M. A.; Goodman, J. M. *Tetrahedron Lett.* **2005**, *46*, 2067–2069.

(38) Glendening, E. D.; Reed, A. E.; Carpenter, J. E.; Weinhold, F. NBO Version 3.1.

(39) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. *WIREs Comput. Mol. Sci.* **2021**, *11*, e1493.

(40) Krieger, A. M.; Pidko, E. A. *ChemCatChem* **2021**, *13*, 3517–3524.

(41) Clavier, H.; Nolan, S. P. *Chem. Commun.* **2010**, *46*, 841.

(42) Guo, J.-Y.; Minko, Y.; Santiago, C. B.; Sigman, M. S. *ACS Catal.* **2017**, *7*, 4144–4151.

(43) Keylor, M. H.; Niemeyer, Z. L.; Sigman, M. S.; Tan, K. L. *J. Am. Chem. Soc* **2017**, *139*, 10613–10616.

(44) Du Toit, J. I.; Van Sittert, C. G. C. E.; Vosloo, H. C. M. *Monatsh. Chem.* **2015**, *146*, 1115–1129.

(45) Suresh, C. H. *Inorg. Chem.* **2006**, *45*, 4982–4986.

4

Data-Driven Virtual Screening of Conformational Ensembles of TM Complexes

Data-driven prediction models can accelerate high-throughput catalyst design but require computer-readable representations that account for conformational flexibility. This is typically achieved through high-level conformer searches followed by DFT optimization of the transition-metal complexes. However, conformer selection remains reliant on human assumptions, with no cost-efficient and generalizable workflow available. To address this, we introduce an automated approach to correlate CREST(GFN2-xTB//GFN-FF)-generated conformer ensembles with their DFT-optimized counterparts for systematic conformer selection. We analyzed 24 precatalyst structures, performing CREST conformer searches followed by full DFT optimization. Three filtering methods were evaluated: (i) geometric ligand descriptors, (ii) PCA-based selection, and (iii) DBSCAN clustering using RMSD and energy. The proposed methods were validated on Rh-based catalysts featuring bisphosphine ligands, which are widely employed in hydrogenation reactions. To assess general applicability, both the precatalyst and its corresponding acrylate-bound complex were analyzed. Our results confirm that CREST overestimates ligand flexibility, and energy-based filtering is ineffective. PCA-based selection failed to distinguish conformers by DFT energy, while RMSD-based filtering improved selection but lacked tunability. DBSCAN clustering provided the most effective approach, eliminating redundancies while preserving key configurations. This method remained robust across datasets and is computationally efficient without requiring molecular descriptor calculations. These findings highlight that DBSCAN-based clustering offers a computationally accessible strategy for rapid catalyst representations involving conformational flexibility.

This Chapter has been published as: Finta, S.; Kalikadien, A. V.; Pidko, E. A. *J. Chem. Theory Comput.* **2025**, *21*, 5334–5345.[1]

4.1. INTRODUCTION

One of the major challenges in applying data-driven algorithms to catalysis lies in creating accurate, computer-readable representations of molecules.[2–4] The universality of the resulting models and their predictive capabilities heavily depend on how well the specific features included in the molecular representation capture the fundamental characteristics and behavior of the catalytic species that ultimately determine the reactivity.[3–8] Although most models focus on features associated with static molecular representations, incorporating properties calculated on a conformer ensemble into the featurization step has gained traction as a means to better capture the fluxionality of molecular systems under reaction conditions and improve predictive accuracy.[9,10] Both experimental and computational studies underscore the importance of catalyst conformations for catalytic activity and enantioselectivity[11–13], as different conformers may exhibit unique steric effects and energy profiles.[9,14] As discussed in Chapter 3 given the sensitivity of physical-chemical properties to structural variations, the inclusion of conformational effects is essential for accurate feature acquisition.[15–17]

However, identifying suitable conformer searching algorithms for TM complexes remains challenging due to the complexity of these systems.[18,19] TM complexes are large and feature a wide variety of bond types, and there is a lack of fast, efficient methods that can effectively handle such systems, particularly in the context of high-throughput exploration of highly fluxional chemical environments. Broadly exploring possible conformations for large complexes comes with significant computational costs.[20,21] On the other hand, relying on chemical intuition to select conformers can introduce human bias, often leading to inaccurate representations and neglect of critical conformational effects.[20,22]

To achieve descriptors of conformers that accurately capture their physical-chemical properties, DFT-level calculations are typically utilized.[23] Quantum chemistry-based conformer searching methods, such as AARON,[24] use DFT calculations to produce precise conformer results, though they come at a high computational cost. As a result, many workflows begin with a less costly conformer exploration using force field or semi-empirical methods.[10,25] Common force field-based algorithms include RDKit,[26] OpenBabel,[27] and MOLASSEMBLER,[28] while CREST (ConformerRotamer Sampling Tool) is a widely used tool applying GFNn-xTB tight-binding semi-empirical methods[29]. Examples of methods for non-biased exploration of stereochemistry that utilize RDKit or Openbabel in the back-end are Architector[30] and MACE.[31] In most workflows, the ensembles generated in these initial steps are then refined with DFT to enhance accuracy.[32,33]

Selecting which conformers to refine is not straightforward. Ideally, the goal is to identify conformers that correspond to local minima on the DFT potential energy surface. A logical approach might involve selecting conformers with low relative energy within the ensemble based on energies calculated by a semi-empirical method. However, a significant challenge with current conformer searching methods is the unreliable energy ranking within the ensembles. Previous studies have highlighted the limitations of classical force fields (FF) and semi-empirical methods in accurately predicting energy ordering and global minima compared to DFT-level

calculations.[18,19,34,35] Consequently, relying solely on energy values for filtering could risk excluding important low-energy conformers that would otherwise be identified on the DFT potential energy surface. In CREST, an alternative option is based on principal component analysis (PCA) clustering, which performs PCA and then clusters conformers based on dihedral angles. However, as reported in the CREST documentation, the algorithm cannot accommodate non-covalent bonds, which often occur in transition metal complexes. Furthermore, the algorithm applies k-means clustering, where the number of clusters is a predetermined variable. Another commonly used filtering approach is the CENSO workflow.[23,36] This screening approach uses the obtained CREST ensemble as input and performs pre-filtering based on the energies obtained from DFT single-point calculations. The remaining conformers undergo DFT geometry optimization, during which several filtering steps are included based on energy thresholds. The final ensemble is obtained through a pruning step based on the Gibbs free energy. Although effective, this approach is based on constant re-evaluation of the energy of each conformer, increasing the computational cost with increasing flexibility of the molecule.

In our work, we aimed to investigate a practical and generic approach for streamlining the generation of a DFT-based conformer ensemble from lower-level ensembles in the context of high-throughput computational catalyst screening. Hence, we sought to answer the following question: what filtering method or combination of methods allows for a conformer selection workflow in which computational cost is kept low while a high accuracy is maintained? Several high-throughput and automated filtering approaches for conformer ensembles were investigated. Conformer ensembles were generated for 24 Rh-based catalysts originating from Chapter 2, utilizing bisphosphine ligands.[37] Parameters from the CREST-based ensemble were used to filter and the DFT refined ensemble was used as a ground truth, which enabled the quantification of the effectiveness of a filtering method. More specifically, by establishing a set of molecular descriptors we aimed to enable a data-driven filtering approach. Two data-driven filtering approaches were tested, the first one was a principal component analysis based on a set of steric, geometric and electronic descriptors calculated on the conformer ensemble. The second data-driven filtering approach was a heuristic approach based on the relative values of selected geometric and steric descriptors. Finally, a density-based clustering of relative energy and root-mean square deviation (RMSD) values was performed, constituting the simplest filtering approach investigated in this work.

4.2. COMPUTATIONAL METHODS

4.2.1. CONFORMER GENERATION AND FILTERING WORKFLOW

This study is based on a dataset from Chapter 2 on Rh-based catalyst employing primarily bidentate ligands.[37] From that study, 24 catalyst structures were randomly selected as the starting point for the current research. Each structure featured a Rh metal center with a bisphosphine ligand attached to it. A norbornadiene (NBD) moiety was coordinated trans to the bisphosphine ligand to reflect the precatalyst state.[37] These structures are referred to as L1 to L24, where the number corresponds

to the ligand identity. Visualizations of the ligands are available in the Data Availability section. In this study, the digital representation of the catalyst structures, i.e. in XYZ and MDL Molfile format, was utilized for further investigation via conformer searching and filtering methods.
An overview of the workflow for this study is presented in Figure 4.1, in which two stages can be identified: the first stage involves the generation of CREST conformer sets and the subsequent optimization based on DFT. This dataset served as a platform to test our conformer ensemble filtering approaches. The second stage explores various methods aimed at accurately modeling the contents of the refined DFT-based ensembles using features and parameters derived from the CREST ensembles.

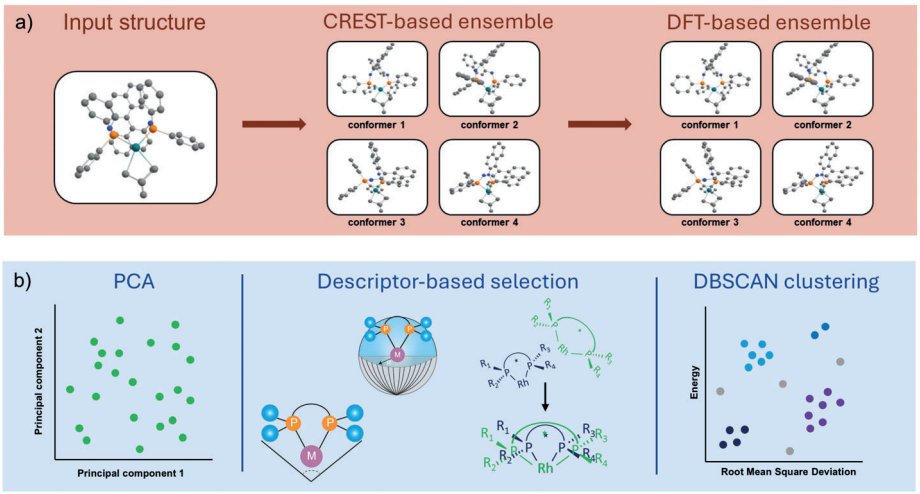


Figure 4.1.: An overview of the applied workflow with a representative illustration of the Rh-based catalyst structures. Part a) involves the creation of conformer ensembles via CREST and subsequent DFT refinement. In part b) various methods were tested to relate a representation of the CREST-based conformer ensemble to the DFT-based refined ensemble.

4.2.2. QUANTUM CHEMICAL METHODS

For stage one of the workflow, conformer generation and exploration were conducted using the Conformer-Rotamer Ensemble Sampling Tool (CREST) version 2.12[29,38] and xTB version 6.4.0.[39] CREST calculations were performed on all 24 Rh-based structures using Cartesian coordinates (*.xyz file) as input geometries for conformer ensemble creation. The GFN2-xTB//GFN-FF hybrid potential was chosen for its accurate performance at reasonable computational costs and universal applicability.[35] For readability purposes, the CREST(GFN2-xTB/GFN-FF)-generated conformer ensembles

are referred to as CREST-based conformer ensembles. Conformers generated by CREST were subsequently preprocessed using the MORFEUS Python package (version 0.7.1). The python package readily takes obtained CREST output folders as an input, which accommodates further filtering and analysis. To enable this, an explicitly added connectivity matrix was extracted from an MDL Mol file. Afterwards, structures that exhibited changes in chirality relative to the original input structure were removed from the ensemble.

The resulting CREST-based conformer ensembles were refined via DFT geometry optimization, performed using Gaussian 16 C.02.[40] The PBE0-D3(BJ)/def2-SVPP level of theory[41–43] was applied, known for its reliable accuracy and efficiency for the description of TM complexes[35,44,45]. The nature of each stationary point was confirmed via frequency analysis. Thermochemical parameters (e.g. ZPE, finite temperature corrections and entropy contributions to Gibbs free energies) were computed from analytical frequencies (Hessian) at 298.15K and 1 atm. For conformers displaying imaginary frequencies, the pyQRC Python script (version 1.0.3)[46,47] was employed to generate revised input geometries, which were then reoptimized with the same DFT settings. Conformers that retained imaginary frequencies after two attempts at reoptimization were excluded from further evaluation.

4.2.3. DATA ANALYSIS

The core objective of this study is to identify a subset of conformers from the CREST ensemble that best represent the DFT ensemble, using DFT-derived energy values from stage I (Figure 4.1a) as the reference. The main part of the workflow (stage II, Figure 4.1b) involves the evaluation of various algorithms selection methods to determine their effectiveness in capturing the most relevant conformers. In this context, assuming chemical accuracy of ca. 5 kJ/mol, conformers within this energy range were considered indistinguishable in the DFT ensemble.[48] An automated script was developed to perform this task, followed by additional manual adjustments. The finalized DFT ensembles are available in the Data Availability section.

Molecular descriptors of the CREST-based conformers were calculated using the OBeLiX (Open Bidentate Ligand eXplorer) open-source computational package.[15] With the MORFEUS conformer ensemble object as input, a total of 37 descriptor values for each individual conformer including steric, geometric and electronic properties were calculated. A comprehensive list of these descriptors is provided in the Data Availability section. Additionally, structural differences between conformers were incorporated into the analysis using the heavy-atom root-mean-square deviation (RMSD) relative to the first (lowest CREST energy) conformer in the ensemble. The RMSD calculations were performed with the MORFEUS package using its default settings.

As shown in Figure 4.1b, three approaches were used to identify a subset of conformers from the CREST ensembles that accurately represent the DFT ensemble: a principal component analysis (PCA), a molecular descriptor-based selection and a DBSCAN clustering of relative energy and RMSD values. For the PCA, the dataset of selected molecular descriptors supplemented by the RMSD values of the conformers was utilized. To standardize the dataset, a standard scaling procedure was applied to

the descriptors, ensuring uniform data ranges with a mean of zero and a standard deviation of one. This analysis focused on the first two principal components only. In the second approach, molecular descriptor-based selection methods, certain steric and geometric properties, such as the cone angle and buried volume, were used for conformer selection. This approach ensures that the selected conformer set includes conformers with varied steric and geometric profiles, including the extremes that define distinct accessible value ranges for these properties.[49] Based on this, it was chosen to select CREST-based conformers with the minimum and maximum values for both buried volume (calculated at the metal center with radius 4Å) and cone angle. The third approach applied DBSCAN clustering on the relative energy and RMSD values of conformers within the ensemble, with the minimum cluster size parameter set to 2, while the distance-to-centroid parameter was further optimized based on model performance.

The investigated methods were primarily evaluated by a confusion matrix. The following approach was used to determine the parameters of the confusion matrix:

- **True negative (TN)**: The number of conformers that are correctly eliminated by the algorithm: their DFT minima are already represented by other conformers in the predicted subset, making them redundant to cover the DFT ensemble.
- **False negative (FN)**: The number of conformers that are incorrectly eliminated by the algorithm: their DFT minima are not represented by other conformers in the predicted subset, making them necessary to cover the DFT ensemble.
- **False positive (FP)**: The number of conformers that are incorrectly included in the predicted subset by the algorithm: their DFT minima are already represented by other conformers, making them redundant to cover the DFT ensemble.
- **True positive (TP)**: The number of conformers that are correctly included in the predicted subset by the algorithm: their DFT minima are not represented by other conformers, making them necessary to cover the DFT ensemble.

The chosen evaluation parameters were True negative (TN) and false negative (FN) values. In a well-performing model, TN is maximized to ensure that all redundant conformers are removed, while FN is minimized to ensure that no DFT minimum is overlooked.

4.2.4. VALIDATION

The dataset from Chapter 3, in which both CREST-based conformer ensembles and their DFT-optimized structures were available, was used for further validation purposes. This dataset also consisted of Rh-based catalysts with bisphosphine ligands, but instead of using the precatalyst form with NBD coordinated to the metal center, a methyl 2-acetamidoacrylate substrate was coordinated to Rh. Based on the ligand-substrate configurations, four different coordination modes are possible of which two are more sterically restricted, and two are less sterically restricted.[35] Our workflow was tested on the 44 CREST ensembles from 11 different ligands.

Generally, the substrate coordination gives the structures more flexibility compared the precatalyst form with NBD. This makes the conformer ensembles extend beyond the possibilities of manual analysis within the restrictions of reasonable labor costs and thus serves as a representative case study where high-throughput conformer analysis would be useful.

4.3. RESULTS AND DISCUSSION

4.3.1. CONFORMER SEARCH AND DFT GEOMETRY OPTIMIZATION

The refinement of all conformers at a high level of theory after low-level conformational searches can significantly increase computational cost without justified gains. This can be demonstrated by comparing the conformer ensembles generated by CREST and refined at the DFT level of theory. CREST, employing xTB, generally predicts much greater conformational freedom, characterized by a broader range and higher number of individual conformers than those retained after DFT optimization (Figure 4.2). Specifically, while CREST generated a total of 678 conformers across the 24 input structures, the DFT ensembles retained a considerably smaller subset of these conformers. Among the 24 ensembles analyzed, the average

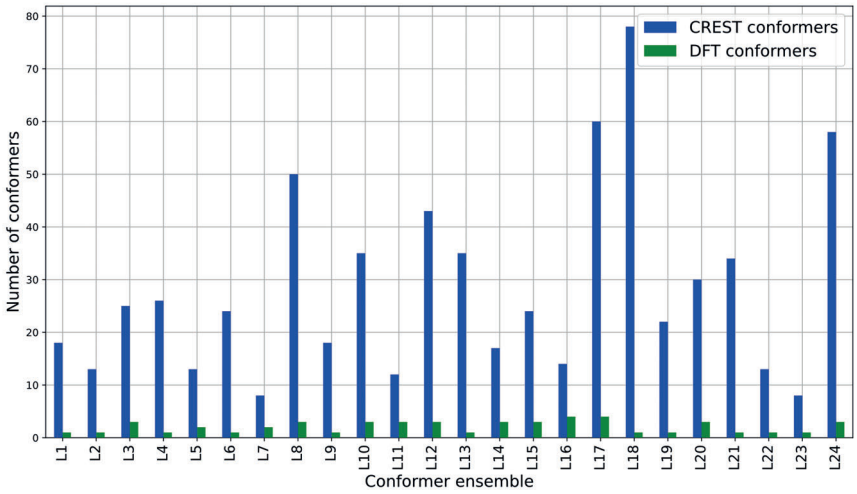


Figure 4.2.: Comparison of the number of conformers obtained from both CREST and DFT calculations. The number of conformers in each ensemble is indicated in blue for the CREST ensembles and in green for the DFT ensembles. The ensembles were named according to ligand numbering, which can be found in the list of ligands in the Data Availability section.

number of conformers per ensemble at the xTB level was 23, which was reduced to an average of only 2 conformers per ensemble after DFT refinement. The CREST

ensembles exhibited considerable variation in the number of conformers obtained; for example, the ensembles for L7 and L23 comprised only eight conformers, while the largest ensemble, L18, contained 78 conformers. Following DFT refinement, both L23 and L18 yielded a single conformer in the DFT ensemble, whereas the ensemble for L7 contained two conformers. The large reduction observed in ensemble size after DFT refinement is in line with our previous observations[35], which indicate that the size of conformer ensembles decreases greatly after the DFT refinement.

To investigate this in more detail, four representative ensembles are examined: L3, L8, L17, and L24. Figure 4.3 compares the relative stabilities of the conformers from CREST at the xTB level (ΔE_{xTB}) and after DFT refinement (ΔE_{DFT}). The broad

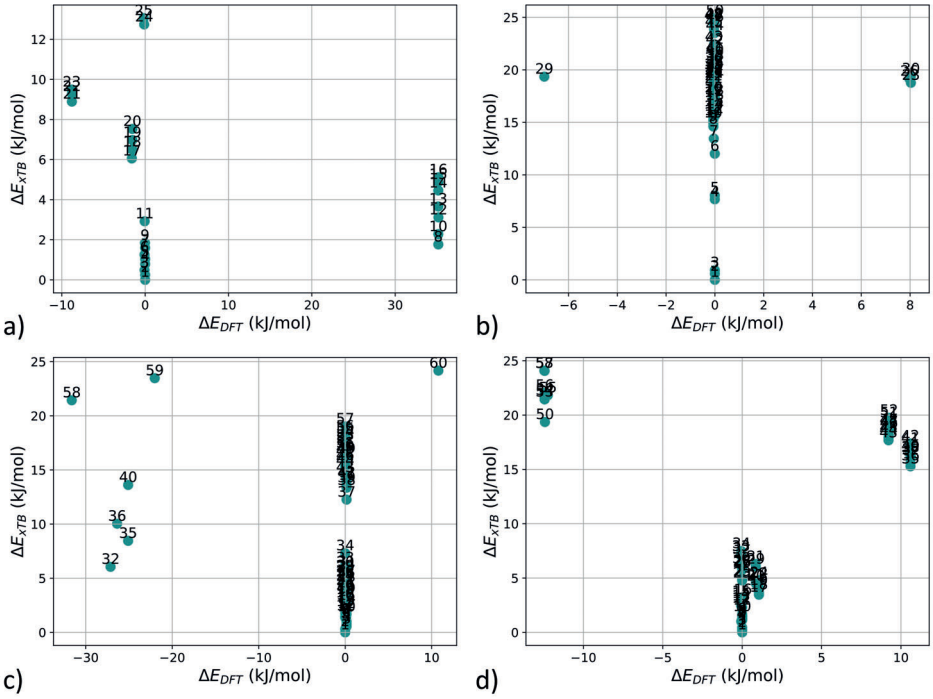


Figure 4.3.: DFT and xTB energies relative to the conformer with the lowest xTB energy of ensemble L3 a), ensemble L8 b), ensemble L17 c) and ensemble L24 d).

conformer space predicted by CREST collapses to only a few distinct conformers after DFT optimization (Figure 4.3). Furthermore, the relative stabilities predicted at the xTB level do not correlate with those computed at the DFT level. For example, in ensemble L17, the conformer ranked as lowest-energy by CREST is 21 kJ/mol higher than the lowest DFT energy conformer. Similarly, in ensemble L8, the CREST lowest-energy conformer has a higher energy by 19 kJ/mol compared to the lowest DFT conformer.

These examples highlight a key point: the apparent differences in flexibility predicted by the two methods stem from the fact that many of the CREST conformers, even those with large energy differences, converge to the same DFT conformer after optimization. This comparison reveals that the flexibility of the complexes obtained by xTB is overestimated, resulting in a much smaller conformer space at the higher level of theory.

The discrepancy between the conformer spaces predicted by xTB and DFT highlights a significant challenge when lower-level methods are utilized for conformer selection prior to further refinement: if one selects only the global minimum or a limited number of low-lying CREST conformers for subsequent refinement and physical-chemical descriptor calculation, there is a high probability of misrepresenting the actual higher-level ensemble. In the absence of more sophisticated conformer selection strategies, this approach risks overlooking relevant structural diversity and introducing bias into the results as highlighted by Laplaza et al.[22] Consequently, computational resources may be wasted, and a comprehensive understanding of the systems true conformational space may not be achieved.

4.3.2. METHODS BASED ON DESCRIPTORS

We introduce a systematic analysis framework to establish a more robust connection between the xTB and DFT conformer ensembles, with the objective of automating conformer selection while ensuring the retention of all unique configurations. To evaluate the correlation between the CREST-based ensemble and its DFT-optimized counterpart, we calculated a set of descriptors on the CREST conformer ensemble. These descriptors, including relative energy, RMSD, cone angle, and buried volume, were used to assess the effectiveness of filtering methods in generating a subset of conformers that closely mirror the DFT ensemble.

The RMSD and ΔE_{xTB} values of the conformers were employed to eliminate redundant conformers through geometry and energy pruning methods as implemented in the MORFEUS Python package. Similar approaches are implemented in the AQME package.[50] The RMSD pruning method targets structural redundancy, based on the hypothesis that conformers with similar geometries, indicated by an RMSD within 0.35Å of the lowest-energy conformer, are likely to converge to the same DFT minimum upon refinement. In contrast, the energy pruning method eliminates conformers with relative xTB-based energies exceeding a threshold of 12.55 kJ/mol (3.0 kcal/mol), suggesting that conformers with close relative energies may exhibit similar stabilities and thus contribute similarly to the conformational space. To further refine the conformer selection, we also considered geometric descriptors such as cone angle and buried volume, which are widely used to characterize the steric and geometric properties of catalysts. These parameters were selected based on the hypothesis that they would capture conformational variability in steric profiles that is not necessarily reflected in electronic properties.[49,51] The cone angle and buried volume are particularly sensitive to steric variations, which are crucial for understanding structural differences in catalytic environments. Therefore, we hypothesized that CREST-based conformers with extreme cone angles and buried volumes are more likely to converge to distinct DFT minima, reflecting significant

conformational differences.

To validate the use of cone angle and buried volume as key descriptors for distinguishing unique DFT minima, an initial analysis was conducted across the 24 conformer ensembles. Out of the 24 ensembles analyzed, 13 showed more than one DFT minimum. In 11 of these cases, the conformers with the highest and lowest buried volumes converged to distinct DFT minima, while in 2 cases (ensembles L3 and L17), they converged to the same minimum. For the cone angle descriptor, the conformers with the highest and lowest values converged to the same DFT minimum in 3 instances (ensembles L3, L8, and L12). This suggests that the combination of these two descriptors successfully differentiated at least two DFT minima in 12 of the 13 cases, providing a basis to utilize them in descriptor-based filtering methods. Three different pruning methods were used prior to the selection process, as shown in Figure 4.4. These methods vary by pruning approach: RMSD pruning (method 1), energy pruning (method 2), and a combined approach using both RMSD and energy pruning (method 3). In each method, the selected conformers retained were those with the highest and lowest cone angle and buried volume within the CREST ensemble.

In an ideal case, as many redundant conformers as possible are eliminated (true negatives) while minimizing the number of unique DFT minima missed (false negatives). Since the same descriptors were applied in for all selection methods, but the pruning methods differed, evaluating these parameters highlights the relative effectiveness of each pruning approach in balancing computational efficiency with accuracy.

Across the 24 CREST ensembles analyzed, a total of 644 redundant and 50 significant conformers were identified. The RMSD pruning method removed 364 (56%) of the redundant conformers, while the energy pruning eliminated only 240 (37%). A notable distinction between the two approaches is that RMSD pruning missed only one DFT minimum (in ensemble L16), while energy pruning failed to capture two DFT minima (one each in ensembles L8 and L15). Figure 4.4 shows that for ensembles L15 and L16, the missed DFT minima are not the lowest energy conformers, whereas in ensemble L8, the global DFT minimum is missed. Therefore, applying RMSD pruning is more effective in both reducing redundant conformers and capturing all minima of the DFT ensemble. This indicates that conformers that show strong structural similarities in the CREST space are more likely to converge into the minimum upon further geometry refinement than conformers that show similar energy values. In method 3, which combines both RMSD and energy pruning, all three of the previously mentioned DFT minima were missed (one each from ensembles L8, L15, and L16). However, this combined approach successfully removed 448 redundant conformers, representing 70% of the total redundancies. This indicates that the combined pruning method offers an effective option for applications where maximizing redundancy reduction takes precedence over capturing every DFT minimum.

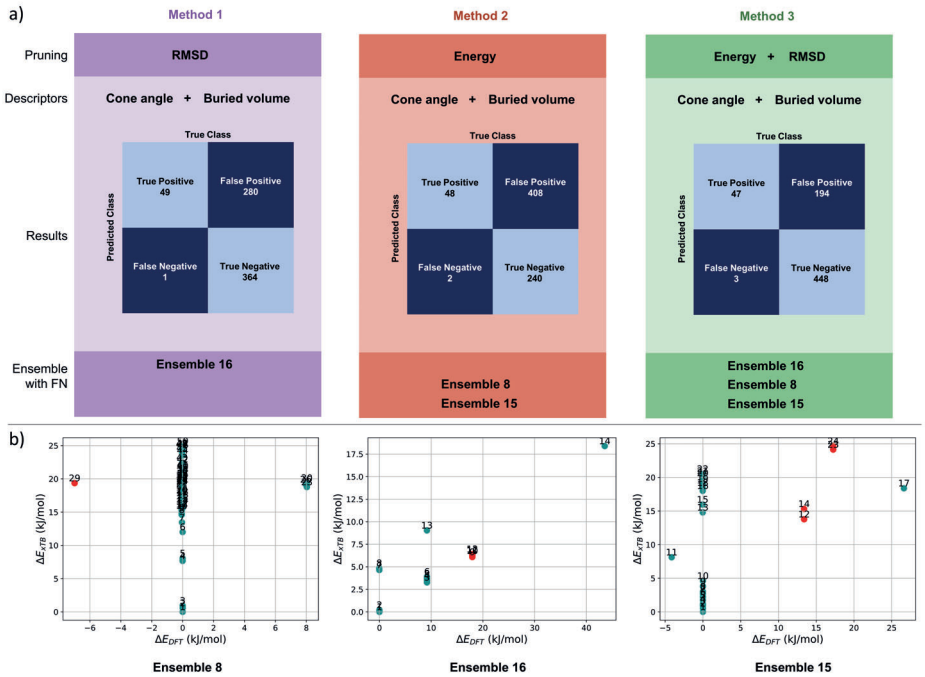


Figure 4.4.: Scheme and results of three descriptor-based filtering approaches are presented in a). In method 1 (left), RMSD pruning is applied; in method 2 (center), energy pruning is applied; and in method 3 (right), both RMSD and energy pruning are combined. Confusion matrices for each method are shown, highlighting the primary assessment parameters: false negatives (FN) and true negatives (TN). Additionally, each ensemble where a DFT minimum is missed (FN) is indicated. In b), CREST-DFT relative energy plots are provided for three ensembles where DFT minima are potentially missed, with conformers associated with a missed DFT minimum marked in red.

4.3.3. PRINCIPAL COMPONENT ANALYSIS

Although the descriptor-based filtering approach showed promising results in distinguishing unique DFT minima, its main limitation lies in the lack of flexibility to customize the balance between accuracy and computational cost, i.e. various pruning methods were utilized, but further downstream selection is based on two descriptors selected by chemical intuition. To address this limitation, a new method was developed, leveraging all descriptors calculated during the low-level CREST exploration. Since conformers often converge to the same DFT minimum after optimization, it can be hypothesized that such conformers share underlying similarities detectable from the CREST-derived descriptors. Energy alone did not

prove sufficient as a distinguishing feature; therefore, we employed a more advanced data-driven method to identify potential similarities among conformers.

This data-driven approach combined dimensionality reduction techniques with clustering methods to identify patterns among the CREST-derived conformers. Dimensionality reduction techniques, such as PCA, are commonly employed on molecular descriptors to facilitate the exploration of chemical space.[49,52,53] It was hypothesized that the variation in physicochemical properties captured by the descriptors contains information about the behavior of the refined DFT ensemble. As a result, the PCA space was expected to provide a more intuitive way to cluster conformers that are refined to similar DFT geometries. PCA was performed on the complete set of descriptors derived from the CREST-based structures, which included the full set of descriptors (see Data Availability section for the descriptor dataset) and the RMSD values of the conformers. Figure 4.5 presents the chemical

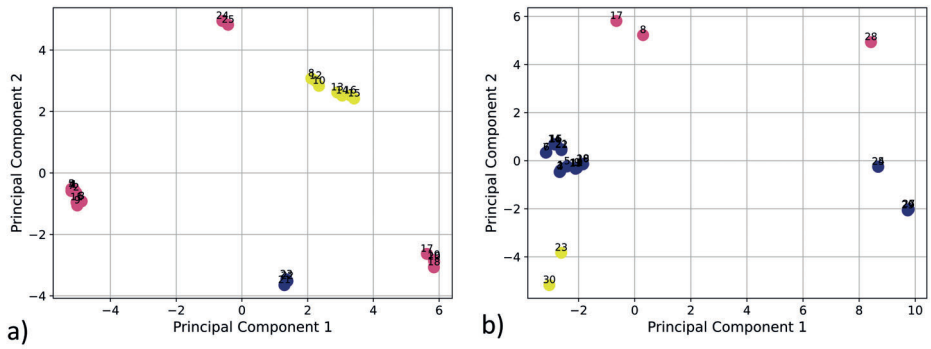


Figure 4.5.: PCA plots of 2 ensembles: ensemble L3 a) and L20 b), conformers that converge into the same DFT minimum are marked with the same color.

space derived from xTB-calculated features following PCA dimensionality reduction, with the coloring indicating the corresponding DFT minima of the conformers. In an ideal scenario, the PCA-reduced space would effectively capture the underlying DFT-defined energy minima, resulting in conformers with identical colors forming distinct clusters. However, the results reveal that this is not the case: the red-colored conformers fail to cluster cohesively, and similarly, the blue-colored conformers in Figure 4.5b are dispersed across two separate regions. These findings indicate that clustering within the PCA space does not yield an optimal selection of conformers. Furthermore, this observation underscores that the variability in the xTB-derived descriptors does not align well with the stability of conformers as determined by DFT calculations.

4.3.4. CLUSTERING

The PCA analysis did not provide a feasible alternative to the previously discussed descriptor-based methods, indicating that incorporating chemical heuristics, such

as filtering based on chemically intuitive descriptors, remains preferable. While the descriptor-based methods demonstrated efficiency, they suffer from a lack of flexibility in tuning the size of the ensemble for specific requirements. Additionally, these methods rely on molecular descriptors derived from CREST ensembles, which consequently adds an additional step to the workflow.

Building on the limitations of descriptor-based methods and PCA-based analysis, we explored an alternative filtering approach using unsupervised clustering techniques. Unlike previous methods that relied on a set of descriptors, this new approach focuses solely on the relative energy and RMSD values of the CREST-based conformers. By doing so, it captures both geometric and energetic features without the need for additional descriptor calculations, based on the assumption that conformers with similar geometries and energy values are likely to converge to the same DFT local minimum. An initial comparison of three clustering algorithms, K-means, K-medoids, and DBSCAN, revealed that DBSCAN is best suited for our dataset and objectives. Unlike K-means and K-medoids, which allocate all conformers to a cluster and thereby risk excluding key conformers, DBSCAN is designed to manage data with higher noise levels. Conformers are grouped only if they are sufficiently close in RMSD and energy, minimizing the likelihood of overlooking essential conformers in the ensemble. In particular, the cluster size parameter (ϵ) in DBSCAN provides a powerful mechanism to control the definition of "closeness," enabling the method to be fine-tuned for various objectives. This flexibility allows DBSCAN to strike a balance between precision and computational efficiency in conformer selection.

The results of the DBSCAN clustering (Figure 4.6a) show that the choice of the ϵ parameter and therefore the size of the clusters significantly influences the performance of the clustering model. The clustering results can be categorized into three parts based on the value of false negatives. In the initial range of ϵ, all DFT minima are successfully captured. As the cluster size increases, the number of redundant conformers eliminated also increases proportionally. At ϵ=0.19, 369 redundant conformers are filtered out, slightly surpassing the previously reported RMSD pruning method (364) and significantly exceeding the energy pruning method (240). However, as the cluster size is further increased, at ϵ=0.20, one DFT minimum remains uncaptured, specifically the global DFT minimum of ensemble L8 (see Figure 4.6b). Increasing the parameter further to ϵ=0.23 results in an additional missed DFT minimum, which corresponds to the highest energy minimum of ensemble 14. Despite its simplicity, this method outperformed all previously tested approaches while allowing for more precise performance tuning through the parameter ϵ. When capturing all DFT minima is critical, a lower ϵ value can be selected, with the filtering objective gradually shifting from accuracy toward cost-efficiency as ϵ increases.

4.3.5. VALIDATION

Although the clustering approach demonstrated promising results on the dataset, its applicability to a set of systems with higher conformational flexibility remains uncertain. To assess its generalizability, we validated the method using a dataset featuring methyl 2-acetamidoacrylate as the substrate. Switching from the precatalyst

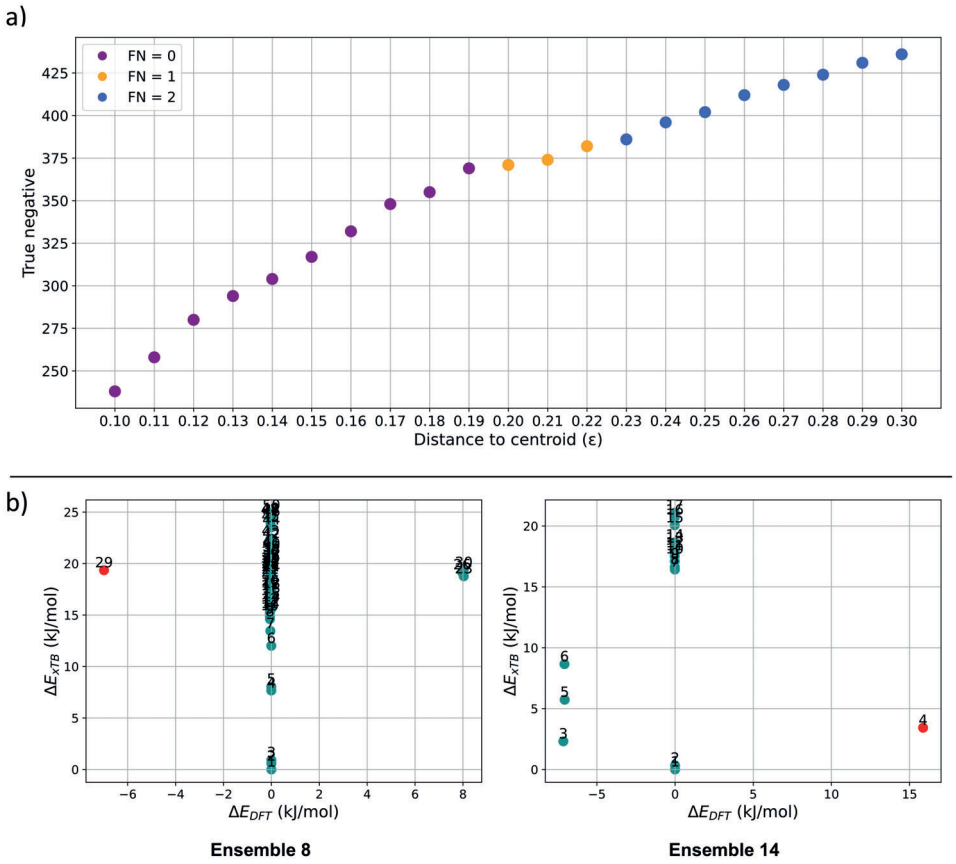


Figure 4.6.: Results on DBSCAN clustering. In a) the results of DBSCAN clustering on the dataset of 24 ensembles are presented. The x-axis represents the distance to centroid (ϵ) parameter, while the y-axis displays the true negative values. Data points are colored according to their false negative values: purple points indicate FN = 0, orange points represent FN = 1, and blue points correspond to FN = 2. In b), CREST-DFT relative energy plots are provided for three ensembles where DFT minima are potentially missed, with conformers associated with a missed DFT minimum marked in red.

to the actual substrate increases the ligands flexibility, resulting in a more complex and diverse conformational space.[35] This increased complexity provides a robust test for evaluating the transferability of our filtering approach and examining the sensitivity of the ϵ parameter across different structural types.

The 11 input structures, reflecting various ligand backbones, yielded 44 CREST ensembles, resulting in a total of 1271 conformers. Following DFT geometry

optimization, the refined ensembles contained 154 conformers, indicating that 1117 of the CREST conformers were redundant. Given that DBSCAN clustering within the range of ϵ = 0.10 to 0.19 successfully captured all DFT conformers from the original dataset, this algorithm was applied again with the same parameters. The outcome of this clustering approach is illustrated in Figure 4.7, which plots the ϵ parameter against the number of successfully eliminated redundant conformers. The color of the data points denotes the number of missed DFT minima. These

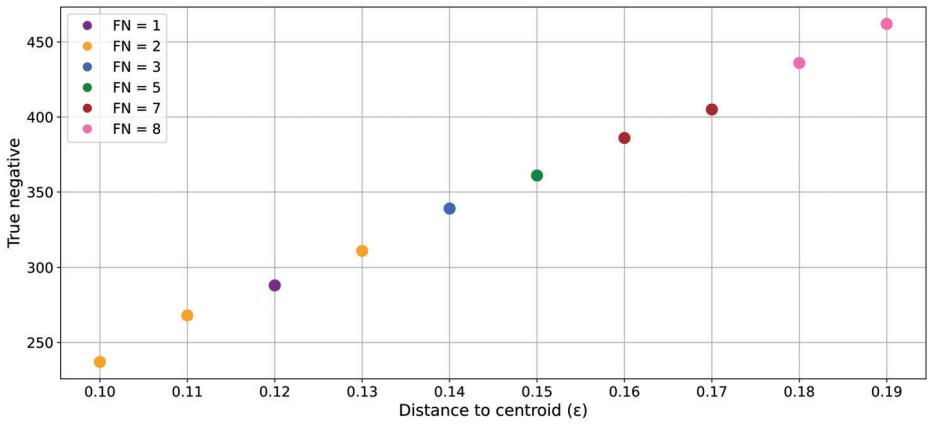


Figure 4.7.: The results of DBSCAN clustering on the validation dataset of 44 ensembles are presented. The x-axis represents the distance to centroid (ϵ) parameter, while the y-axis displays the true negative values. Data points are colored according to their false negative values: purple indicates FN = 1, orange represents FN = 2, blue corresponds to FN = 3, green denotes FN = 5, brown indicates FN = 7, and pink represents FN = 8.

results indicate that even in the best-case scenario with an epsilon value of 0.12, at least one DFT minimum remains uncaptured. However, given the larger number of DFT minima, this shortfall is proportionally less significant. When comparing the number of redundant conformers eliminated across both datasets using the same DBSCAN filtering approach (ϵ = 0.19), it becomes evident that although more redundant conformers are eliminated in absolute terms from the acrylate substrate dataset than from the original NBD substrate dataset, the relative reduction is lower. Specifically, 462 out of 1117 redundant conformers (41%) were removed from the acrylate substrate dataset, compared to 369 out of redundant 644 conformers (57%) in the NBD dataset. These findings suggest that although the acrylate substrate dataset exhibits more variations in the space of RMSD versus energy at the xTB level of theory, resulting in less straightforward clusters, our approach remains effective. A majority of DFT minima are captured via this simple clustering approach solely based on RMSD and relative energy as metrics.

4.4. CONCLUSION

Computer-readable representations of catalysts enable ML-based screening of widely utilized TM catalysts. The inclusion of conformational flexibility within these representations remains largely dependent on human decisions and assumptions for the filtering of 'relevant' conformers. Additionally, less accurate semi-empirical or force-field based approaches are preferred over DFT-based methods for the generation of these conformer ensembles due to lower computational cost. This study explored data-driven approaches to correlate conformer ensembles of a lower level of theory to their DFT optimized counterparts, enabling automated filtering of conformers. A dataset of 24 precatalyst structures based on our previous research was established for which conformer searching via CREST and subsequent DFT optimization of every resulting conformer was performed. The investigation was performed in three parts. Firstly, a combination of pruning and conformer selection based on geometric ligand descriptors was tested. Afterwards, a fully data-driven approach via PCA was tested for the mapping of the CREST-based conformers to their DFT optimized equivalents. Finally, RMSD- and energy-based clustering using DBSCAN was tested and then evaluated on a second dataset containing the same ligands, but the precatalyst structure was changed for one containing an acrylate substrate, inducing higher ligand flexibility.

Our research showed that the CREST-generated conformers, when compared to the DFT ensemble, do not reflect the flexibility of the structure. It proved difficult to identify the lowest energy conformer within a DFT optimized conformer ensemble directly based on the energy as calculated in CREST with the GFN2-xTB method. Additionally, CREST produced significantly more conformers compared to the DFT-based ensemble, thus overestimating the flexibility of ligands. Pruning methods demonstrated that pruning based on geometry, rather than energy, resulted in a more accurate mapping to the DFT-based ensemble. This highlighted issues with CRESTs energy calculations and the limitations of energy-based filtering. A fully geometry-based filtering method, using RMSD pruning and selection based on geometric descriptors, outperformed energy-based approaches. However, limitations remained such as limited tunability of this method and one of the DFT minima remaining uncaptured. Unfortunately, a second filtering approach using PCA on descriptors from the CREST ensembles failed to differentiate conformers based on their DFT energy. Remarkably, the simplest algorithm, clustering based on RMSD and energy values, performed exceptionally well. DBSCAN clustering with these features showed the best filtering, with the lowest false negative rate and the highest elimination of redundant conformers. This method can be fine-tuned using the cluster centroid distance parameter, balancing accuracy and computational cost for different applications. It also does not require the calculation of molecular descriptors for the CREST ensemble. When tested on a validation dataset containing an acrylate substrate with increased ligand flexibility compared to that of a precatalyst structure, the method remained effective, suggesting its general applicability across various catalyst structures employing bisphosphine ligands.

Overall, our findings bear significance for the dynamic representations involving conformational flexibility of catalyst structures in high-throughput virtual screening

workflows. A shortcoming of this approach is that the relationship between the distance to centroid parameter and the resulting accuracy-cost trade-off is highly dependent on the chemical structures themselves, making it challenging to tune. Additionally, when a very high accuracy is required, e.g. for the approximation of enantioselectivity, filtering based on constant energy refinement and reweighting conformers would be more advisable. Developments in conformer filtering approaches as researched in this study go hand-in-hand with developments in the field of conformer searching methods[10,54–56], ML-based energy calculations[57], and more efficient exchange-correlation functionals[58,59] where constant improvements are being made in the chemical space of transition-metal complexes. Nevertheless, a DBSCAN-based RMSD clustering approach remains the most simple and computationally feasible option for now. This approach is being utilized in our current and future research on dynamic representations of homogeneous catalysts for ML-based virtual screening.

DATA AVAILABILITY

The Python package for the featurization of catalyst structures, OBeLiX, is available through the GitHub organization page of the ISE group at TU Delft: EPiCs-group OBeLiX (https://github.com/EPiCs-group/obelix), with the specific version to calculate descriptors for individual conformers from a CREST ensemble contained on a separate branch (https://github.com/EPiCs-group/obelix/tree/confomer_searching_dev_final).

All datasets used in this study are provided with an extensive README via 4TU.ResearchData at https://doi.org/10.4121/45bb4e4b-272b-41ce-a090-2b6e4b1708fd. The following resources are included:

- A list and visualization of ligands ('ligand_description.docx')
- An Excel file categorizing and describing all descriptors ('descriptors_description.xlsx')
- Script for conformer filtering and creating figures used for analysis ('data_analysis.py')
- Pickled ConformerEnsemble objects created with the Morfeus package containing conformers, xTB energies and RMSD values ('conformer_ensemble_files.zip')
- Input and output of conformer searching with CREST ('CREST_structures.zip')
- CSV files with descriptors for each conformer calculated at the xTB level of theory ('descriptors.zip')
- Input and output of DFT optimized files with Gaussian 16 ('DFT_structures.zip')
- MDL Molfiles to extract the connectivity matrix per metal-ligand complex for conformer searching and analysis ('mol_files.zip')
- Files with energy values and tracking which conformers are pruned in a conformer ensemble ('pruning_files.zip')
- Data from the case study on a validation set from our previous research ('validation.zip')
- Figures for PCA-, clustering- and energy-based conformer selection approaches ('visualization.zip')

CONTRIBUTIONS

S. Finta and **A.V. Kalikadien** contributed equally to this work. **S. Finta:** Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization **A.V. Kalikadien:** Conceptualization, Methodology, Software, Validation, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration **E.A. Pidko:** Supervision, Conceptualization, Resources, Funding acquisition, Writing - Review & Editing, Project administration

REFERENCES

(1) Finta, S.; Kalikadien, A. V.; Pidko, E. A. *J. Chem. Theory Comput.* **2025**, *21*, 5334–5345.

(2) Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. *WIREs Comput Mol Sci* **2022**, *12*, e1603.

(3) Burrows, L. C.; Jesikiewicz, L. T.; Lu, G.; Geib, S. J.; Liu, P.; Brummond, K. M. *J. Am. Chem. Soc.* **2017**, *139*, 15022–15032.

(4) Gallegos, L. C.; Luchini, G.; St. John, P. C.; Kim, S.; Paton, R. S. *Acc. Chem. Res.* **2021**, *54*, 827–836.

(5) Durand, D. J.; Fey, N. *Chem. Rev.* **2019**, *119*, 6561–6594.

(6) Maloney, M. P.; Stenfors, B. A.; Helquist, P.; Norrby, P.-O.; Wiest, O. *ACS Catal.* **2023**, *13*, 14285–14299.

(7) Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C. *Chem. Sci.* **2021**, *12*, 6879–6889.

(8) Soyemi, A.; Szilvási, T. *Dalton Trans.* **2021**, *50*, 10325–10339.

(9) Brethome, A. V.; Fletcher, S. P.; Paton, R. S. *ACS Catal.* **2019**, *9*, 2313–2323.

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

(11) Shan, C.; Liu, X.; Luo, X.; Lan, Y. *Sci. Rep.* **2024**, *14*, 24031.

(12) Gallagher, J. M.; Roberts, B. M.; Borsley, S.; Leigh, D. A. *Chem* **2024**, *10*, 855–866.

(13) Peng, Q.; Wang, Z.; Zaric, S. D.; Brothers, E. N.; Hall, M. B. *J. Am. Chem. Soc.* **2018**, *140*, 3929–3939.

(14) Bursch, M.; Hansen, A.; Pracht, P.; Kohn, J. T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2021**, *23*, 287–299.

(15) Kalikadien, A. V.; Mirza, A.; Hossaini, A. N.; Sreenithya, A.; Pidko, E. A. *ChemPlusChem* **2024**, e202300702.

(16) Hoque, A.; Sunoj, R. B. *Digit. Discov.* **2022**, *1*, 926–940.

(17) Antinucci, G.; Dereli, B.; Vittoria, A.; Budzelaar, P. H. M.; Cipullo, R.; Goryunov, G. P.; Kulyabin, P. S.; Uborsky, D. V.; Cavallo, L.; Ehm, C.; Voskoboynikov, A. Z.; Busico, V. *ACS Catal.* **2022**, *12*, 6934–6945.

(18) Minenkov, Y.; Sharapa, D. I.; Cavallo, L. *J. Chem. Theory Comput.* **2018**, *14*, 3428–3439.

(19) Das, S.; Merz Jr, K. M. *J. Chem. Inf. Model.* **2023**.

(20) Laplaza, R.; Sobez, J.-G.; Wodrich, M. D.; Reiher, M.; Corminboeuf, C. *Chem. Sci.* **2022**, *13*, 6858–6864.

(21) Kammeraad, J. A.; Das, S.; Arguelles, A. J.; Sayyed, F. B.; Zimmerman, P. M. *Org. Lett.* **2024**, *26*, 2867–2871.

(22) Laplaza, R.; Wodrich, M. D.; Corminboeuf, C. *J. Phys. Chem. Lett.* **2024**, *15*, 7363–7370.

(23) Axelrod, S.; Gomez-Bombarelli, R. *Sci. Data* **2022**, *9*, 185.

(24) Guan, Y.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E. *J. Chem. Theory Comput.* **2018**, *14*, 5249–5261.

(25) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. *J. Comput. Chem.* **2016**, *37*, 2106–2117.

(26) RDKit https://www.rdkit.org/.

(27) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminf.* **2011**, *3*, 1–14.

(28) Sobez, J.-G.; Reiher, M. *J. Chem. Inf. Model.* **2020**, *60*, 3884–3900.

(29) Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.; Spicher, S.; Steinbach, P.; Wesoowski, P. A.; Zeller, F. *J. Chem. Phys.* **2024**, *160*, 114110.

(30) Taylor, M. G.; Burrill, D. J.; Janssen, J.; Batista, E. R.; Perez, D.; Yang, P. *Nat. Commun.* **2023**, *14*, 1–11.

(31) Chernyshov, I. Y.; Pidko, E. A. *J. Chem. Theory Comput.* **2024**, *20*, 2313–2320.

(32) Gillespie, A. M.; Morello, G. R.; White, D. P. *Organometallics* **2002**, *21*, 3913–3921.

(33) Jorner, K.; Brinck, T.; Norrby, P.; Buttar, D. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. Chem Sci 12: 1163–1175, 2021.

(34) Kanal, I. Y.; Keith, J. A.; Hutchison, G. R. *Int. J. Quantum Chem.* **2018**, *118*, e25512.

(35) Baidun, M. S.; Kalikadien, A. V.; Lefort, L.; Pidko, E. A. *J. Phys. Chem. C* **2024**, *128*, 7987–7998.

(36) Grimme, S.; Bohle, F.; Hansen, A.; Pracht, P.; Spicher, S.; Stahn, M. *J. Phys. Chem. A* **2021**, *125*, 4039–4054.

(37) Kalikadien, A. V.; Valsecchi, C.; van Putten, R.; Maes, T.; Muuronen, M.; Dyubankova, N.; Lefort, L.; Pidko, E. A. *Chem. Sci.* **2024**, *15*, 13618–13630.

(38) Pracht, P.; Bohle, F.; Grimme, S. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(39) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. *WIREs Comput Mol Sci* **2021**, *11*, e1493.

(40) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16 Revision C.01/C.02, 2016.

(41) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(42) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(43) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(44) Sinha, V.; Laan, J. J.; Pidko, E. A. *Phys. Chem. Chem. Phys.* **2021**, *23*, 2557–2567.

(45) Kalikadien, A. V.; Pidko, E. A.; Sinha, V. *Digit. Discov.* **2022**, *1*, 8–25.

(46) Goodman, J. M.; Silva, M. A. *Tetrahedron Lett.* **2003**, *44*, 8233–8236.

(47) Silva, M. A.; Goodman, J. M. *Tetrahedron Lett.* **2005**, *46*, 2067–2069.

(48) Krieger, A. M.; Pidko, E. A. *ChemCatChem* **2021**, *13*, 3517–3524.

(49) Gensch, T.; dos Passos Gomes, G.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-DAddario, M.; Sigman, M. S. *et al. J. Am. Chem. Soc.* **2022**, *144*, 1205–1217.

(50) Alegre-Requena, J. V.; V., S. S. S.; Pérez-Soto, R.; Alturaifi, T. M.; Paton, R. S. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2023**, *13*, e1663.

(51) Wilbraham, L.; Berardo, E.; Turcani, L.; Jelfs, K. E.; Zwijnenburg, M. A. *J. Chem. Inf. Model.* **2018**, *58*, 2450–2459.

(52) Van Dijk, L.; Haas, B. C.; Lim, N.-K.; Clagg, K.; Dotson, J. J.; Treacy, S. M.; Piechowicz, K. A.; Roytman, V. A.; Zhang, H.; Toste, F. D.; Miller, S. J.; Gosselin, F.; Sigman, M. S. *J. Am. Chem. Soc.* **2023**, *145*, 20959–20967.

(53) Dotson, J. J.; van Dijk, L.; Timmerman, J. C.; Grosslight, S.; Walroth, R. C.; Gosselin, F.; Püntener, K.; Mack, K. A.; Sigman, M. S. *J. Am. Chem. Soc.* **2022**, *145*, 110–121.

(54) De Souza, B. *Angew Chem. Int. Ed.* **2025**, *n/a*, e202500393.

(55) Otlyotov, A. A.; Rozov, T. P.; Moshchenkov, A. D.; Minenkov, Y. *Organometallics* **2024**, *43*, 2232–2242.

(56) Talmazan, R. A.; Podewitz, M. *J. Chem. Inf. Model.* **2023**, *63*, 5400–5407.

(57) Hölzer, C.; Oerder, R.; Grimme, S.; Hamaekers, J. *J. Chem. Inf. Model.* **2024**, *64*, 8909–8925.

(58) Gasevic, T.; Stückrath, J. B.; Grimme, S.; Bursch, M. *J. Phys. Chem. A* **2022**, *126*, 3826–3838.

(59) Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. *J. Chem. Phys.* **2021**, *154*, 064103.

5

Unveiling the Impact of Ligand Configurations and Structural Fluxionality on Virtual Screening of TM Complexes

While catalytic properties depend on ligand properties and spatial arrangement, the role of stereoisomerism in defining catalyst selectivity and reactivity has only been elucidated sporadically, leaving gaps in virtual screening workflows. This study investigates the necessity of exploring ligand configurations for virtual high-throughput (HT) screening of octahedral transition metal complexes. Using automated workflows, ligand configuration ensembles were generated for bisphosphine ligands with Ir(III), Ru(II), and Mn(I) metal centers. DFT calculations revealed that Mn(I)- and Ru(II)-complexes displayed significant fluxionality, with multiple configurations within a 10 kJ/mol energy range. Linear regression analysis showed that global descriptors, such as bite angle and HOMO-LUMO gap, are transferable across configurations and metal centers, while local steric descriptors lacked such transferability. ML models successfully classified ligand configurations but struggled to predict stability across metal centers, especially for Mn(I) and Ru(II). Overall, this study underscores the limitations of ignoring stereoisomerism in virtual HT screening, which may lead to incomplete exploration of chemical space and underrepresentation of key catalyst features. Until dynamic digital representations are developed, exhaustive stereoisomerism exploration should be implemented for screening workflows.

This Chapter has been published as: Kalikadien, A. V.; van der Lem, N. J.; Valsecchi, C.; Lefort, L.; Pidko, E. A. *Digit. Discov.* **2025**, *4*, 2033–2044.[1]

5.1. INTRODUCTION

Generally, catalyst screening workflows start with assumptions based on a specific reaction mechanism.[2] Considering the flexibility of these ligands, one thus assumes that a preferred ligand arrangement is retained for all members of a given ligand family and/or metal centers. Conformational search aiming at identifying low-energy rotamers and isomeric structures is commonly carried out for this selected coordination polyhedron with the pre-defined ligand configuration, which is preserved at this stage.[3,4] Although the kinetic trans-/cis-effect is well known,[5] the role of stereoisomerism of the catalyst in defining selectivity and reactivity has only been elucidated sporadically.[6–8]

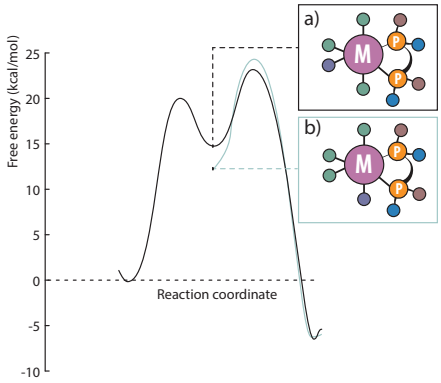


Figure 5.1.: Selectivity control by minor configurations: schematic representation of the impact of a minor (a) and major (b) isomer of a TM complex on the reactivity and selectivity in chemical transformations.

The relationship between the stability and observed catalytic properties of a complex is challenging to comprehend and is usually not known a priori. Consider a scenario where a meta-stable configuration of a TM complex, existing at a low concentration in the reactive system, establishes a favorable reaction channel. This minor catalytic component would provide a major impact on the reaction rate and would therefore determine the nature and characteristics of the primary reaction product (Figure 5.1).[9–15] As an example, two possible ligand arrangements are depicted in Figure 5.1a and b. Although complexes with a different ligand configuration are in equilibrium, one meta-stable configuration may provide a reaction path with an energy barrier that is significantly higher than that of the other configuration. The overall ensemble of ligand configurations ultimately contributes to the observed catalytic properties. In the context of virtual HT screening, the conformational isomerism of the organic ligand backbone has been recognized by the community[16–19] and various structure generation tools such as AARON, Architector, Molsimplify, Molassembler and AQME, contain on-the-fly conformer generation solutions.[20–24] However, due to the initial selection of a configuration, the influence and contribution of the metastable configurations featuring varied coordination environments and ligand arrangements might be overlooked. Furthermore, this

choice assumes that the preferred configuration does not change with relatively minor variations in the ligand structure and, often, even the nature of the metal center. Consequently, the question arises: can catalytic systems be fully accounted for when part of the chemical space is neglected due to initial human choices and intuition in structure generation?

To investigate this, we focused on TM-complexes that are relevant to homogeneous catalysis for hydrogenation reactions, where bidentate ligands are commonly employed to achieve high reactivity and enantioselectivity.[25–28] To ensure that the generated data is as bias-free and comprehensive as possible, we employed an automated workflow for construction, sorting and descriptor calculation of ensembles of ligand configurations for TM-complexes. We constructed ensembles containing different ligand configurations for 87 bidentate ligands, selected from the dataset generated in Chapter 2, connected to 3 different metal centers, namely Ir(III), Ru(II) and Mn(I), yielding a total of 908 octahedral TM-complexes. With these data, we set out to model the relations between stereoisomerism, stability, and descriptors.

The primary research question of this work is whether an exhaustive exploration of stereoisomerism is necessitated for Virtual HT screening of octahedral TM-based catalyst complexes, given that the degree of configurational fluxionality of a complex is not fully known a priori. To answer this, we investigated whether specific ligand configurations of the TM-complexes proved to be more energetically favorable and whether this could be modeled using physical-chemical descriptors and machine learning (ML). The Chapter is organized as follows. Initially, stability trends of different ligand configurations were analyzed on the basis of results from Density Functional Theory (DFT) calculations. The results were analyzed by means of linear regressions to identify relevant descriptors across different ligand configurations and metal centers. The descriptors were then utilized to construct ML models capable of distinguishing different types of ligand configurations and predicting energetic preferences for specific metal-ligand combinations. Our results highlight the challenge of configurational fluxionality for the virtual screening of TM complexes and provide practical directions to address them.

5.2. COMPUTATIONAL METHODS

5.2.1. LIGANDS AND TRANSITION METAL COMPLEXES

The investigated TM-complexes employed various bidentate ligands with isoelectronic Ir(III), Ru(II) and Mn(I) metal centers. The selected auxiliary ligands, next to the bidentate ligands, were hydrides and CO such that neutral TM-complexes were generated. Importantly, all complexes in this study were treated in their closed-shell singlet (i.e., diamagnetic) configurations. In particular, Mn(I) complexes with strong-field ligands such as CO are known to favor low-spin electronic configurations due to the large ligand field splitting they induce, consistent with their position on the spectrochemical series.[29,30] Acetonitrile served as a model substrate to ensure minor impact on the overall conformational freedom of the complexes. More specifically, we have explored the configurational freedom and physical-chemical properties of an extended catalyst dataset featuring 87 chiral

bisphosphine (PP) ligands coordinated to neutral transition-metal complexes. To maintain charge neutrality, Ir(III), Ru(II), and Mn(I) centers were stabilized with different auxiliary ligands, resulting in $PPIrH_3(CH_3CN)$, $PPRuH_2(CO)(CH_3CN)$, and $PPMnH(CO)_2(CH_3CN)$ complexes, respectively. The dataset was constructed without any a priori assumption of the preferred ligand arrangement or TM stereochemistry using a fully automated workflow for the generation of TM complexes.[31] Figure 5.2a illustrates the studied ligand configurations for the Ir, Ru and Mn complexes. The 87 selected bidentate PP ligands belong to different ligand families, a subset of which is shown in Figure 5.2b. The complete set of bidentate ligands is available in the SI (see Data Availability statement).

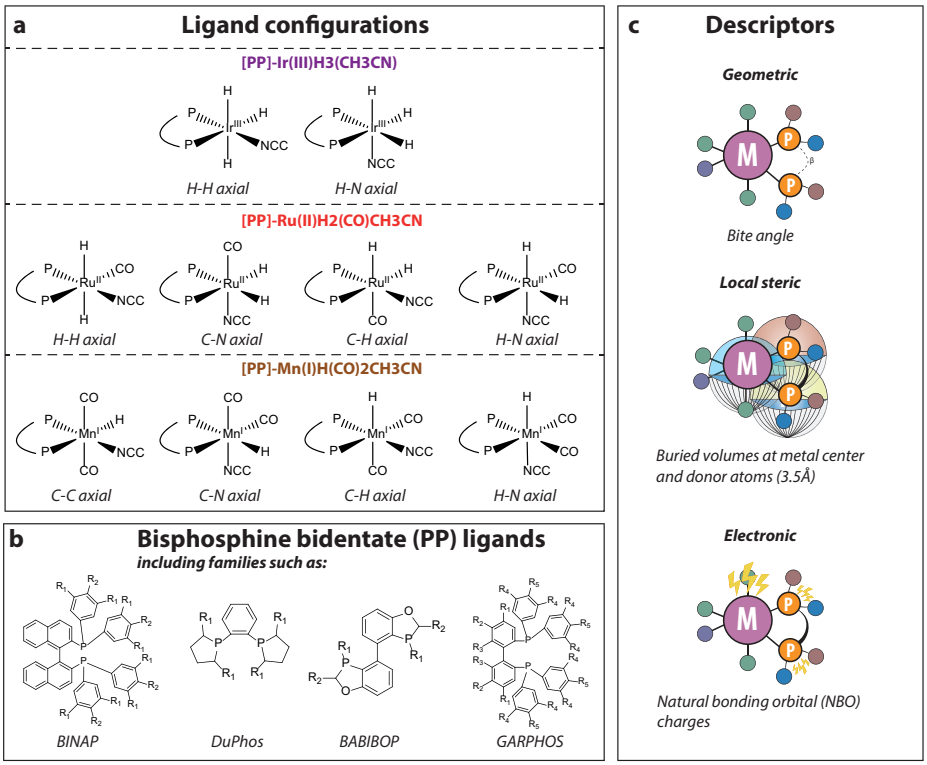


Figure 5.2.: (a) List of possible ligand configurations for each metal center (b) a selection of representative studied bisphosphine bidentate ligand families and (c) a selection of geometric, steric and electronic descriptors used in this study.

5.2.2. COMPLEX GENERATION AND SORTING WORKFLOW

The general workflow for generating and sorting TM-complexes as employed in this study is visualized in Figure 5.3. Structures for TM-complexes were generated

using the in-house Open Bidentate Ligand Explorer (OBeLiX) workflow (see Data Availability statement).[16] This workflow aims to aid computational exploration of the organometallic chemistry space through automated structure generation and descriptor calculation. OBeLiX utilizes the MACE python package for the automated generation of 3D structures and stereochemistry assessment of TM-complexes.[31,32] MACE is an open source python package, which allows bias-free generation of 3D TM-complexes starting from molecular SMILES strings[33] of ligands and metal centers. Furthermore, MACE generates all possible stereoisomers, explores conformations and filters out identical and "impossible" sterically hindered configurations for the given metal-ligand combination. In this study, 50 conformers were generated using the Universal Force Field as implemented in the RDKit package through MACE.

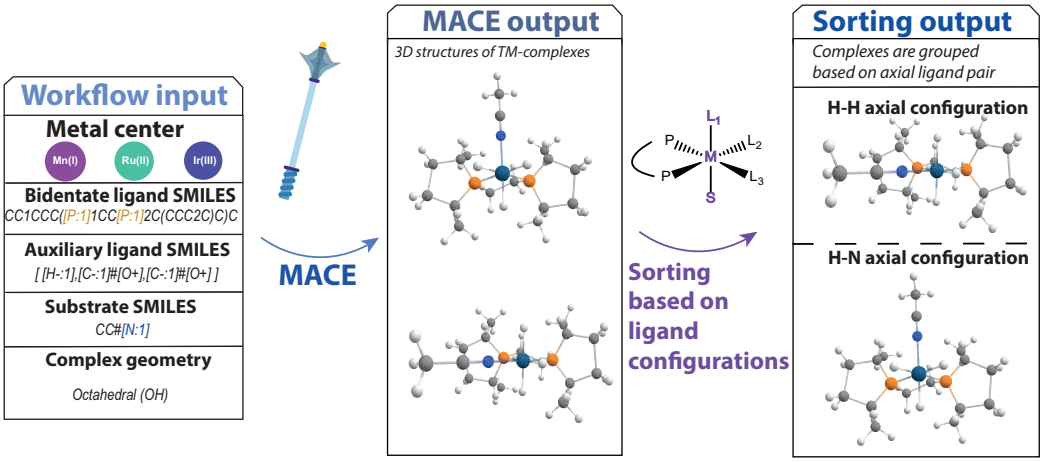


Figure 5.3.: Workflow for generating & sorting TM-complex geometries from specified user input.

After generating the structures for the TM-complexes via MACE, the complexes were named according to the first donor atoms axially bonded to the metal center as illustrated in Figure 5.2. These axial ligand configurations were identified using bite angles. The coordinate system of TM-complexes are defined with respect to the bidentate ligand, and hence the bidentate ligand is always present in the equatorial position. Therefore, the ligands in the axial position are the only non-bidentate ligand containing pair forming a bite angle of 180°. For ligands lacking C_2 symmetry, variations arising from asymmetric functionalization on the phosphine donor atoms were treated as distinct conformers within the same axial ligand configuration. After generating and sorting the TM-complexes, geometries were optimized at the DFT level of theory.

5.2.3. DENSITY FUNCTIONAL THEORY CALCULATIONS

The generated geometries were further refined by DFT calculations using Gaussian 16 C.02[34] software. For geometry optimizations, the PBE0[35] exchange-correlation

functional was used with Grimme's DFT-D3(BJ) dispersion corrections[36] and the def2-SVPP basis set.[37] The selected combination of basis set and exchange correlation functional have previously been established to generate reasonable energies and structures for similar TM-based complexes.[25,38–41] Normal mode analysis was carried out to confirm that the optimized geometries correspond to local minima on the potential energy surface. For structures with imaginary frequencies, the PyQRC python package[42,43] was used to remove these imaginary frequencies and restart geometry optimizations. After geometry optimization, energies were refined with single point (SP) calculations at the PBE0-D3 level using the def2-TZVPP basis set.[37,44]

Thermodynamic stabilities of ligand configurations were calculated by the difference in the DFT-based electronic energies with respect to a reference configuration. The H-N axial ligand pair structure is used as the reference, being the only common configuration present among studied metal centers. The difference in stability between the reference and alternative configurations is denoted as ΔE_{ref}.

In addition to screening the stability of the TM-complexes, a screening based on substrate binding energy was also conducted. Binding energies of the model substrate, acetonitrile, were computed as follows:

$$E_{bind} = E_{DFT,complex} - (E_{DFT,complex-nosub} + E_{DFT,sub}) \tag{5.1}$$

In this equation, substrate binding energy is described by the DFT optimized energy differences between the complex, ($E_{DFT,opt,complex}$) minus the sum of substrate-removed complex ($E_{DFT,opt,complex-nosub}$) and the energy of non-bonded substrate ($E_{DFT,opt,sub}$). More information about this screening approach can be found in SI section S1 and S2.

5.2.4. DESCRIPTOR CALCULATION

The OBeLiX descriptor calculator[16] was employed to automate the extraction of chemical-physical properties and descriptors of DFT-optimized complexes. This tool determines electronic, steric and geometric descriptors using Morfeus[45] and cclib.[46] A graph-based method is employed to locate and label the bidentate donor atoms based on charges calculated by a xTB single-point calculation. Based on these charges, the donor atoms in the bidentate ligand are labeled as either 'min' or 'max'. The structural and electronic descriptors were calculated on DFT-optimized structures. In total, 27 commonly used DFT-based descriptors were selected for the analysis (See Descriptors overview in the SI).

5.2.5. LINEAR REGRESSION

These descriptors were utilized for linear regression to model relationships between descriptors across different ligand configurations and metal centers. The Scikit-learn Python package with default settings was used, hence the coefficient of determination was used as a scoring function for performance.

5.2.6. MACHINE LEARNING

The calculated descriptors were also used in two ML modeling tasks: distinguishing different types of ligand configurations and predicting energetic preferences for specific metal-ligand combinations. The approaches leveraged a modified ML pipeline, adapted from our earlier work.[25] The first task was multi-class classification of configurations in which the TM-complexes were represented as a vector of descriptors and the target value was the axial pair of ligands, e.g. H-H and H-N for Ir-based complexes. This task was performed in two ways: 1) over the whole dataset containing all metal centers and ligand configurations or 2) divided per metal center. This enables highlighting of performance differences between TM-complexes and their respective metal centers. Additionally, the train/test split was done in two ways: 1) in-domain, in which the dataset was randomly divided into train- and test-set or 2) out-of-domain, in which a fixed set of 16 ligands and their configurations were kept out of the training set. This enables insights into the modeling performance on completely new ligands. In the second task, ML was employed for binary classification to model energetic preferences in ligand configurations. In this case the most stable configurations within a specific metal-ligand combination, would get a label 1, while the rest of the configurations for that combination would get a label 0. Again, this task was performed over either the whole dataset of all metal centers and ligand configurations or divided per metal center. The train/test split was also performed either in-domain or out-of-domain (*vide supra*).

The Random Forest (RF) and logistic regression (LR) algorithms were used. RF is an ensemble learning algorithm harnessing multiple decision trees and randomness to construct a predictive model, while logistic regression is a statistical method that models the probability of a binary outcome using a logistic function. In our study, all modeling tasks were attempted with both RF and logistic regression, with logistic regression serving as a simpler alternative to RF. The modeling tasks were evaluated with a balanced accuracy (BA) score which is a metric for evaluating classification models on imbalanced datasets. The score was calculated as follows:

$$\mathrm{BA} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{5.2}$$

Details about cross-validation, hyperparameter optimization and model intialization can be found in the SI section S3.

5.3. RESULTS AND DISCUSSION

To explore the role of structural fluxionality in the *in silico* screening of catalyst complexes and to assess whether trends in the energetic landscape could be visually discerned or systematically predicted using machine learning (ML), the research was structured into four key steps: 1) A detailed analysis of the DFT-calculated energetic landscape of ligand configurations to identify trends and patterns across different metal centers, 2) Statistical and linear regression analyses to examine the sensitivity of descriptors and their influence for combinations of specific configurations and metal centers, 3) ML-based classification to predict ligand configurations using these

descriptors, and 4) ML-based classification to identify the most stable configuration for various ligand-metal center combinations.

5.3.1. ENERGETIC PREFERENCES IN LIGAND CONFIGURATION

To investigate whether a specific configuration is generally more favorable compared to others, i.e. a global minimum on the Potential Energy Surface (PES), the relative stability of possible complex configurations is analyzed across our selection of bidentate bisphosphine ligands. The relative stability of alternative configurations

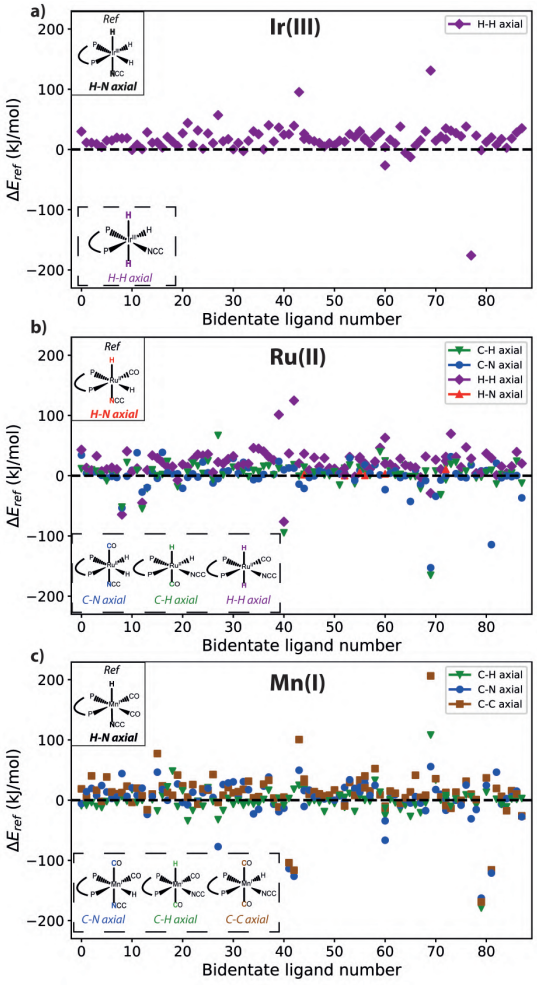


Figure 5.4.: Relative stability of ligand configurations, shown at the bottom of a graph, and a reference structure, shown at the top of a graph, for set of bidentate ligands for (a) Ir(III), (b) Ru(II) and (c) Mn(I) complexes

with respect to the selected reference structure per metal center is presented in Figure 5.4. At the top of each figure, the reference structure is depicted, with the alternative configurations shown at the bottom.

The results for Ir(III) complexes (Figure 5.4a) reveal that for most ligands the reference N-H axial ligand pair is more stable than the alternative H-H arrangement. In the case of Ru(II) (Figure 5.4b), the presence of additional auxillary ligands expands the configurational space, which now includes C-N, C-H and H-H as alternative axial ligand pairs next to the reference H-N. Upon assumption that hydrides are indistinguishable, Ir(III) complexes can only form two distinctive configurations, whereas four different configurations of the Ru complex can be formed. A significant variety in the data set is observed, as the complex configuration with the lowest stability often varies among the different bidentate ligands.

Similar to Ir(III), the H-H axial ligand configuration exhibits a positive ΔE_{ref} for the majority of bidentate ligands, signifying a lower stability compared to the reference H-N configuration. However, a notable difference from Ir(III) complexes is that alternative configurations exhibit higher stability for many systems. For instance, the C-N axial configuration commonly exhibits a negative ΔE_{ref}, indicating their higher stability than the H-N reference. Furthermore, our workflow identifies multiple Ru complexes with alternative configurations varying more than 50 kJ/mol compared to the reference case. These outliers are the result of unfavorable conformations imposed by the specific ligand arrangement on the metal center. In particular, 6 ligands were identified as outlier for multiple metal centers (L86, L87, L119, L134 and L171), but no noteworthy trends were observed. Data on these outliers are contained in the data_analysis and descriptor_analysis directory in the SI.

A similar analysis for Mn(I) complexes (Figure 5.4) reveals that the most stable preferred configuration varies between different bidentate ligands. Distinctive to Mn(I) complexes is the C-C configuration which shows a lower overall stability for most ligands. Nevertheless, as opposed to Ir(III) complexes, the reference H-N axial ligand configuration is not shown to be the most stable configuration in all cases. Instead, the C-H and C-N configurations are energetically more favorable.

Figure 5.5 summarizes the axial ligand configurations along with the percentage of bidentate ligands for which those specific configurations are found to be the global minimum on the PES. For the Ir(III) complexes 92% of bidentate ligands show a clear global minimum in energy for the H-N configuration. The remaining 8% favor the single alternative H-H configuration. For Ru(II) complexes, the H-N axial configuration is also frequently identified as the global minimum, but this now only accounts for 50% of the bidentate ligands. Both the C-N and C-H axial ligand configurations emerge as the global minimum for a notable number of bidentate ligands, 31% and 18% respectively. The H-H axial ligand configuration is the global minimum for a single bidentate ligand. Unlike Ru(II) and Ir(III), related Mn(I) complexes do not display a pronounced majority of minima containing the H-N configuration. This geometry is preferred for only 26% of Mn(I) complexes, while the alternative C-H configuration is the global minimum for 45% of the bidentate ligands in this case. The C-N and C-C axial arrangement are preferred by 24% and 5% of the Mn(I) complexes respectively. These findings underscore that even though

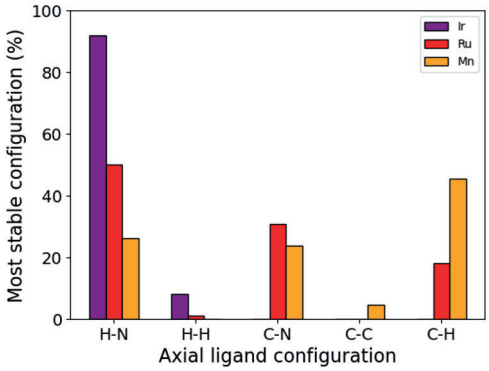


Figure 5.5.: Distribution of most stable ligand configuration over all possible ligand configurations for Mn(I), Ru(II) and Ir(III) complexes, alongside 87 bisphosphine (PP) bidentate ligands.

bidentate bisphosphine ligands are studied exclusively, no clear trend in the stability of a ligand configuration can be observed across the studied metal centers.

5.3.2. TRANSFERABILITY OF DESCRIPTORS TO DIFFERENT LIGAND CONFIGURATIONS AND METAL CENTERS

Next, a statistical analysis of different physical-chemical descriptors was performed to identify relevant descriptors that are affected by changes in the configurations of various metal-ligand combinations. In total, a selection of 8 electronic, 4 geometric and 15 steric descriptors are considered in this study. Examples that will be discussed in more detail in this work are: the buried volume which comprises a measure of the steric occupation of a ligand, the NBO charges of the bidentate ligand's donor atoms and metal center in the TM-complex, the bite angle between metal center and bidentate ligand's donor atoms and finally the HOMO-LUMO gap. These descriptors are commonly utilized in studies of the reactivity and selectivity of homogeneous catalysts. In previous research, we have elucidated the relation between conformational flexibility and physical-chemical descriptors.[47]. We now focus on the transferability of descriptors between different configurations, metal centers and combinations thereof. Transferability in this context, thus, means that a descriptor can be reliably predicted from a selected configuration of a metal-ligand combination, from which it can be inferred which descriptors are sensitive to variations in stereoisomerism.

Linear regression models were constructed to predict specific descriptors of the complexes across different combinations of metals and ligand configurations. The models are scored using a Coefficient of Determination (R^2) ranging from 0 to 1. Since there are 10 possible metal and ligand configuration combinations, the performance of (10x10) - 10 = 90 distinct linear models per descriptor is reported. Figure 5.6 shows a heatmap for four selected descriptors and the resulting R^2 for

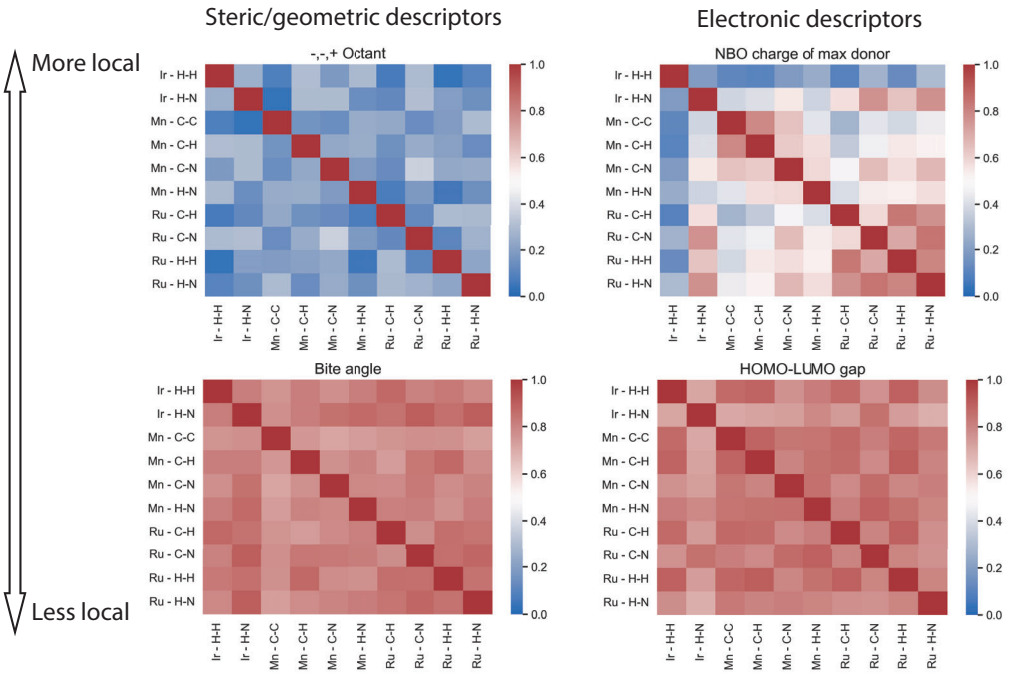


Figure 5.6.: Matrices for R^2 scores of linear models between specific descriptor from one set of bidentate ligands with a specific metal and ligand configuration to another set with a different combination of metal and ligand configuration. An example is shown for four selected descriptors that range in locality. An image for matrices of all calculated descriptors can be found in the SI section S4.

the 100 models, the steric descriptors are depicted on the left and the electronic descriptors on the right. The top two descriptors in the figure represent local properties of the bidentate ligand, while the bottom two are more global. Each heatmap shows all possible 'metal, configuration' combinations on the x- and y-axes. Similar heatmaps for all calculated descriptors are provided in the SI (Section S4).

Figure 5.6 reveals a clear distinction in the transferability of the steric descriptors. The calculated physical-chemical descriptors differ in the level of locality that is captured, where a buried volume can be separated into quadrant- and octant-based contributions which offer a local view on the steric occupancy of a ligand, the bite angle remains a more global geometric measure. Although a local octant of the buried volume shows a low R^2 ($R^2 < 0.5$) for models across all metal and ligand configuration combinations, the bite angle shows a relatively high R^2 ($R^2 > 0.7$) across all metal and ligand configuration combinations. This is in line with the inherent low variance (33.5°) of the bite angle across the whole dataset. Similar observations are made and reported in the SI section S4 for the percentage buried

volume with a radius of 3.5 Å at the metal center or ligand donor atoms.

For the electronic descriptors shown on the right in Figure 5.6, the distinction in transferability is less pronounced. This is evidenced by the presence of red regions in the heatmap of the NBO charge, which deviates from the uniform blue observed in the local steric descriptors. The NBO charge at an atom describes the local electronic environment of the specified atom, while the HOMO-LUMO gap remains a global descriptor depicting the difference in energy of the frontier orbitals of the whole complex. The NBO charge at the ligand donor atom labeled 'max' shows varying modeling performance. Starting at the top left of the heatmap, the 'Ir, H-H' metal and configuration combination shows no transferability across any other combination. The 'Ir H-N' metal and configuration combination shows moderate R^2 ($R^2 \approx 0.6$) for Ru(II)-based C-N and H-N configurations while a low R^2 is observed for all other combinations. On the bottom right, the Ru(II)-based configurations show a relatively moderate to high R^2 across other Ru(II)-based configurations. For Mn(I)-based configurations, the trend differs, as moderate to high R^2 values are observed only between the C-C, C-H, and C-N configurations. This highlights the sensitivity of certain descriptors to stereoisomerism and the nature of the metal center in the TM complex. In contrast, the HOMO-LUMO gap consistently exhibits high R^2 values ($R^2 > 0.7$) across all metal and ligand configuration combinations. Although the HOMO-LUMO gap itself seems transferable across metal and ligand configuration combinations, visualization of the frontier orbitals showed that the nature of the respective frontier orbitals may substantially differ with varied configurations. This analysis can be found in the SI section S5.

5.3.3. ML MODELING OF LIGAND CONFIGURATIONS

Given that certain descriptors are sensitive to changes in stereoisomerism and the nature of the metal center, the focus now shifts to the use of machine learning models to classify and predict the stability of ligand configurations based on these descriptors. Before applying this modeling approach, it is necessary first to assess the ability of machine learning to leverage the selected descriptors to distinguish between different ligand configurations. A comprehensive classifier was trained for either the whole dataset comprising all ligands and metals or metal-specific by dividing the dataset metal-wise. This leads to a five-class classification for the axial ligand pairs using either Random Forest or Logistic Regression algorithms.

The performance evaluation is shown in Figure 5.7, where the x-axis depicts whether modeling was performed on the dataset comprising all metals or by metal-specific division. The BA on the test set for RF and LR are shown by a red and blue bar respectively. The modeling on the dataset containing all ligands, metals and ligand configurations reveals a gap in performance between the non-linear RF and the linear LR models. Where the RF models yielded a remarkable BA of 0.87 - 0.89, the LR models yielded a good BA of 0.73-0.79. Inspecting the performance of metal-specific models going towards the right in the figure, it can be observed that although all models perform good to excellent, a drop in performance is observed for Mn(I)-specific modeling. Nevertheless, these results suggest that the descriptors employed allow ML to effectively distinguish between different axial configurations.

This holds true even for out-of-domain modeling cases where 16 ligands were kept out of the training set, simulating a case of applying the trained models to fully new ligands. The out-of-domain modeling results are contained in SI section S7.

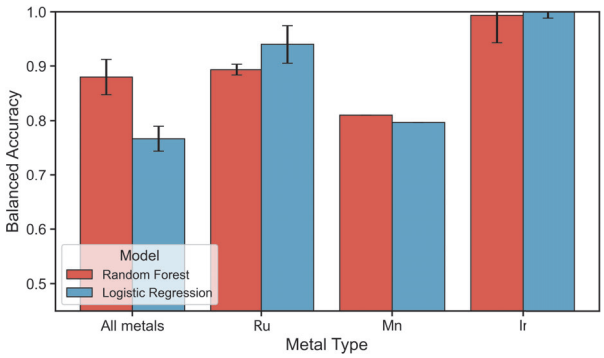


Figure 5.7.: Performance metrics for the in-domain modeling of ligand configurations. The performance of RF and LR are displayed in a red and blue bar respectively. The y-axis denotes the Balanced Accuracy score and the x-axis specifies whether modeling is done on the dataset containing all metal centers and ligand configurations or on a metal-specific subset.

An examination of the feature importances (see SI section S6) revealed that for modeling over all metal centers, descriptors such as the dipole moment, NBO charges on the metal or bidentate ligand donor atoms, and distances between the donor atoms and the metal center are of the highest importance. Thus, these importances reveal that the polarity of the complex (dipole moment) and the local electronic environment surrounding the metal and ligand donor atoms (all other mentioned descriptors) are informative enough to distinguish different ligand configurations for ML. This is in line with our findings on the transferability of descriptors, where those same descriptors are observed to exhibit a high sensitivity to changes in ligand configuration and metal center. However, it should be noted that an important difference is observed in the feature importances of Mn(I)-specific models. Where high importance is given in Ru(II)- and Ir(III)-specific models to the dipole moment and afterwards mainly descriptors of the local electronic environment surrounding the metal center, these seem of relatively lower importance in Mn(I)-specific models. Here, a higher importance is observed for more global descriptors such as the bite angle, cone angle and HOMO-LUMO gap. This observation points at a difference in which ML is able to distinguish ligand configurations of 3d TM-complexes compared to their 4d counterparts and is indicative of the gap in performance of Mn(I)-specific models compared to Ru(II)- and Ir(III)-specific models.

5.3.4. THERMODYNAMIC ACCESSIBILITY OF METASTABLE CONFIGURATIONS AND ML MODELING OF ENERGETIC STABILITY

Knowing that ML has the ability to distinguish different ligand configurations based on the given set of descriptors, we set out to model the stability of ligand configurations. However, the results in Figure 5.4, reveal that multiple isomers of the same metal-ligand pair can exhibit similar stability. This finding suggests that, under the reaction conditions, multiple ligand configurations may contribute to the population of the coordination complex, thereby impacting the overall observed catalytic behavior. To quantitatively assess this factor, we have analyzed the proportion of systems for which multiple ligand configurations were obtained within an energy threshold of 10 kJ/mol from the global minimum state. The choice of the 10 kJ/mol threshold is based on the assumption that a catalyst population follows a Boltzmann average, resulting in at least 5%, and up to 50% of the total population to be in a metastable state under the reaction conditions commonly employed in homogeneous catalysis. For each ensemble of configurations, the number of configurations within a 10 kJ/mol range of the lowest-energy isomer is obtained.

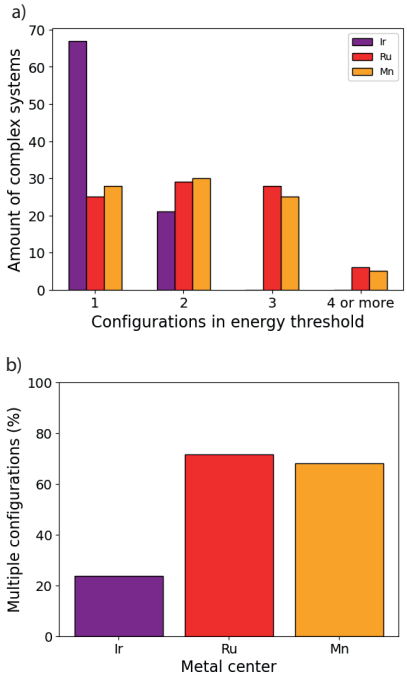


Figure 5.8.: (a) Number of ligand configurations within 10 kJ/mol of the most stable ligand configuration for the researched bidentate ligands, and (b) the percentage of bidentate ligands for which multiple ligand configurations are found within the specified 10 kJ/mol energy range.

Figure 5.8a reports the number of ligand configurations within the 10 kJ/mol

energy range from the most stable configuration for the Ir(III), Ru(II) and Mn(I) complexes, while the fraction of the respective complexes featuring multiple ligand configurations within this energy range is given in Figure 5.8b. For the majority of Ir(III) complexes, only a single configuration is observed within the specified energy range. This finding is in line with the significant stability differences and small number of available ligand configurations. However, even in this case, 24% of Ir(III) complexes are expected to exhibit substantial structural isomerism, i.e. present multiple ligand configurations with stability difference <10 kJ/mol, under the reaction conditions. The fraction of such systems is much higher for Ru(II) and Mn(I) complexes, where multiple ligand configurations within 10 kJ/mol stability range were found for 72% and 68% of the cases, respectively.

To enable machine learning models to classify ligand configurations based on their relative stability, we treated all configurations within 10 kJ/mol of the most stable structure as a single class. This threshold reflects a design choice based on the assumption that such configurations are thermally accessible and thus potentially relevant under catalytic conditions. Similar to the previous modeling approach, a binary classifier was trained either on the whole dataset comprising all ligands and metals or metal-specific by dividing the dataset metal-wise. This leads to a binary classification where the model has to predict whether a ligand configuration is within the stability range of 10 kJ/mol. Again, both the Random Forest and Logistic Regression algorithms were utilized.

Performance evaluation is shown in Figure 5.9, where the x-axis depicts whether the modeling was performed on the dataset that includes all metals or by metal-specific division. The BA on the test set for RF and LR are again shown by a red and blue bar

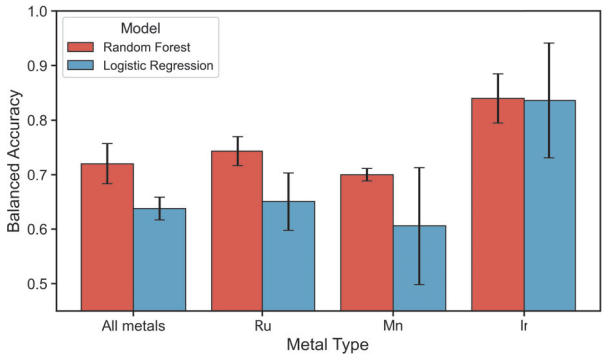


Figure 5.9.: Performance metrics for the in-domain modeling of the stability of ligand configurations. The performance of RF and LR are displayed in a red and blue bar respectively. The y-axis denotes the Balanced Accuracy score and the x-axis specifies whether modeling is done on the dataset containing all metal centers and ligand configurations or on a metal-specific subset.

respectively. All results, except for Ir(III)-specific models, reveal a gap in performance between RF and LR models. Where the RF models on the dataset of all metal centers

and ligand configurations yield a moderate BA of 0.69 - 0.74, the LR models yielded a worse BA of 0.60 - 0.68. Inspecting the performance of metal-specific models going towards the right in the figure, it can be observed that although all models perform moderately, again a drop in performance is observed for Mn(I)-specific modeling. Additionally, the performance of Ir(III)-specific models has a large range in BA of 0.21 for both RF and LR. Since only 24% of Ir(III) complexes are expected to exhibit substantial structural isomerism, the variation in performance depends on whether and how many, of these exceptions are present in the test set. These results suggest that utilizing these descriptors for modeling the stability of ligand configurations is only moderately possible with RF models. The modeling performance is exacerbated in the out-of-domain modeling approach, where a performance drop in the BA is observed for all RF models (see SI section S9). The performance or LR models remained similar in the out-of-domain modeling. Nevertheless, in both RF and LR modelling approaches for all cases except Ir(III)-specific models, this performance points at a modeling ability that is only marginally better than random selection for predicting the stability of a fully unseen ligand.

Given that modeling of the stability was only performing sufficiently for Ir(III)-specific models, the feature importances (see SI section S8) of these models give an insight into which descriptors are strongly linked to stability. The high standard deviations in the feature importance of LR models for Ir(III) make interpretations non-trivial. Nevertheless, the feature importance of RF models reveal that the same descriptors that enabled the modeling of different ligand configurations, which capture the polarity of the complex and the local electronic environment surrounding the metal and ligand donor atoms, are now also of high importance. However, since these descriptors only moderately allow for in-domain modeling and do not allow the reliable out-of-domain modeling of the stability of configurations for Ru(II) and Mn(I), the universality of these descriptors can be questioned.

5.4. CONCLUSION

In this study, we investigated whether an exhaustive exploration of stereoisomerism is necessitated for Virtual HT screening of octahedral TM-based catalyst complexes, since the degree of configurational fluxionality of a complex is not fully known a priori. Hence, ligand configurations of the TM-complexes were investigated for energetic preferences and the ability to model this using chemically intuitive physical-chemical descriptors and ML with an emphasis on explainability. This investigation was performed in four parts. Firstly, the preferences for certain ligand configurations in terms of stability was investigated. Secondly, simple linear regression models were employed to investigate sensitivity to changes in the metal center, ligand configuration or a combination thereof. Thirdly, it was investigated whether ML models could utilize these descriptors to distinguish different ligand configurations. Finally, the ability of ML to model global minima of DFT-based energy in ligand configurations was tested.

Using our automated workflows, ensembles of possible ligand configurations were generated for a library of bisphosphine bidentate ligands with Ir(III), Ru(II) and Mn(I)

metal centers. For the study of stability-based preferences in ligand configuration, our findings based on DFT calculations revealed that Ir-complexes displayed a clear preference in ligand configuration, whereas Mn(I)- and Ru(II)-complexes lacked this preference. Thus, it can be concluded that it is incorrect to assume a particularly fixed ligand configuration as the most stable one across these metal centers.

Investigating the transferability of physical-chemical descriptors across ligand configurations and metal centers revealed that local steric descriptors such as the octant contribution of the buried volume are hardly transferable across metal centers or even ligand configurations with the same metal center. However, local electronic descriptors such as the NBO charge on donor atoms of the ligand exhibited transferability between varying ligand configurations and the same metal center. More global steric, electronic and geometric descriptors, such as the bite angle, HOMO-LUMO gap, indicated a high degree of transferability between all metal centers and lignad configurations. These findings emphasized that the exploration of stereoisomerism in virtual HT screening is of importance if local descriptors are of interest to the screening task at hand.

Since the descriptor set was sensitive to variations in ligand configurations, they could prove useful in modeling the energetic preference of ligand configurations. Hence, it was first established whether the descriptor set allowed ML to distinguish between ligand configurations. Based on our results, where a BA of > 0.8 for RF models on the dataset containing all metal centers was achieved, it can be concluded that the employed descriptors and non-linear models allow for effective out-of-domain modeling where 16 ligands were kept out of the training set. In a case where descriptors of completely new bidentate ligands are given to the trained ML model, it is thus able to effectively predict its axial ligand configuration pair. However, metal-specific models underline challenges in the potential applications to 3d TM-complexes since a performance drop was observed for Mn(I)-specific models.

For a majority of Mn(I)- and Ru(II)-complexes, multiple ligand configurations were found within a 10 kJ/mol energy range from the most favorable one, indicating that multiple ligand configurations may coexist under reaction conditions typically employed in homogeneous catalysis, all influencing catalyst properties. Although a single configuration is predominantly observed within this energy range for most Ir(III)-complexes, a significant portion (24%) is shown to still exhibit substantial structural isomerism. To model the stability of ligand configurations, all ligand configurations with a stability difference of lower than 10 kJ/mol within an ensemble were thus treated as equal and indistinguishable. The modeling attempts proved to be only marginally better than random selection for predicting whether the configuration of a fully unseen ligand would fall within the 10 kJ/mol stability range. Since these descriptors only moderately allow for in-domain modeling and do not allow the out-of-domain modeling of the stability of configurations for specific models of configurations with a Ru(II) and Mn(I) metal center, it is concluded that these descriptors are not universally applicable across metal centers to model the stability of ligand configurations. Since the feature importances of Ir(III)-specific models, where modeling was successful, indicate that the local environment of the metal center and ligand donor atoms hold the highest importance, there is

a large potential for representations containing improved descriptors of the first coordination sphere surrounding the metal center.

Overall, our findings are significant for the virtual high-throughput screening of homogeneous catalysts, which remains heavily reliant on human decision making. Our results demonstrate that focusing on a single ligand configuration during this process may lead to insufficient coverage of the chemical space and an inadequate representation of key catalyst features, thereby limiting the predictive power of in silico catalyst screening campaigns. Furthermore, understanding the flexibility and fluxionality of novel metal-ligand combinations a priori is important for accurate statistical modeling, yet this information is often unavailable beforehand. The modeling approaches described in this study rely on descriptors of individual ligand configurations, creating a 'chicken-and-egg' problem: the flexibility and fluxionality are unknown a priori, yet without accounting for them, it remains unclear how comprehensively they should be explored in the digital representation of catalysts. This underscores the current absence of dynamic digital representations in screening workflows. Hence, screening campaigns should prioritize an exhaustive exploration of stereoisomerism when assessing properties sensitive to structural flexibility and fluxionality.

DATA AVAILABILITY

The Supporting Information file for this Chapter is available at: https://doi.org/10.1039/D5DD00093A.

The core machine learning pipeline used in this study is publicly accessible via the GitHub organization page of the ISE group at TU Delft: EPiCs-group ML Pipeline (https://github.com/EPiCs-group/obelix-ml-pipeline). Additionally, the Python package for the featurization of catalyst structures, OBeLiX, is also available through the same GitHub organization: EPiCs-group OBeLiX (https://github.com/EPiCs-group/obelix).

All supporting information and datasets used in this study are provided along with an extensive README via 4TU.ResearchData at https://doi.org/10.4121/216555e8-5f8b-48a0-b92d-9c08505ceacd.

- A list and visualization of ligands ('ligand_list.pdf')
- An Excel file categorizing and describing all descriptors ('descriptors_overview.xlsx')
- A directory containing the version of OBeLiX used, alongside Python scripts for structure generation and manipulation ('code.zip')
- A directory with DFT data, including xyz, log, and where applicable Gaussian .chk files ('dft_data.zip')
- A directory with Excel files of DFT results for each ligand configuration and a Jupyter notebook for stability analysis ('data_analysis.zip')
- A directory with Excel files containing all descriptors for all generated complexes ('descriptor_data.zip')
- A directory with descriptor, energy, and angle data for all studied complexes, alongside scripts and data for ML analysis ('descriptor_analysis.zip')

CONTRIBUTIONS

A.V. Kalikadien: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration **N.J. van der Lem:** Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization **C. Valsecchi:** Conceptualization, Software, Validation, Formal analysis, Investigation, Writing - Review & Editing **L. Lefort:** Supervision, Conceptualization, Resources, Funding acquisition, Writing - Review & Editing **E.A. Pidko:** Supervision, Conceptualization, Resources, Funding acquisition, Writing - Review & Editing, Project administration

REFERENCES

(1) Kalikadien, A. V.; van der Lem, N. J.; Valsecchi, C.; Lefort, L.; Pidko, E. A. *Digit. Discov.* **2025**, *4*, 2033–2044.

(2) Ahn, S.; Hong, M.; Sundararajan, M.; Ess, D. H.; Baik, M.-H. *Chem. Rev.* **2019**, *119*, 6509–6560.

(3) Rosales, A. R.; Quinn, T. R.; Wahlers, J.; Tomberg, A.; Zhang, X.; Helquist, P.; Wiest, O.; Norrby, P.-O. *Chem. Commun.* **2018**, *54*, 8294–8311.

(4) Pracht, P.; Bauer, C. A.; Grimme, S. *J. Comput. Chem.* **2017**, *38*, 2618–2631.

(5) Kauffman, G. B. *J. Chem. Educ.* **1977**, *54*, 86.

(6) Coe, B. J.; Glenwright, S. J. *Coord. Chem. Rev.* **2000**, *203*, 5–80.

(7) Torker, S.; Khan, R. K. M.; Hoveyda, A. H. *J. Am. Chem. Soc.* **2014**, *136*, 3439–3455.

(8) Bhaskararao, B.; Sunoj, R. B. *J. Am. Chem. Soc.* **2015**, *137*, 15712–15722.

(9) Kumari, S.; Alexandrova, A. N.; Sautet, P. *J. Am. Chem. Soc.* **2023**, *145*, 26350–26362.

(10) Zhang, Z.; Zandkarimi, B.; Alexandrova, A. N. *Acc. Chem. Res.* **2020**, *53*, 447–458.

(11) Brown, J. M.; Chaloner, P. A. *J. Chem. Soc., Chem. Commun.* **1979**, 613–615.

(12) Brown, J. M.; Chaloner, P. A. *J. Chem. Soc., Chem. Commun.* **1980**, 344.

(13) brown, J. M.; chaloner, P. A.; glaser, R.; geresh, S. *Tetrahedron* **1980**, *36*, 815–825.

(14) Chan, A. S.; Halpern, J. *J. Am. Chem. Soc.* **1980**, *102*, 838–840.

(15) Gridnev, I. D.; Imamoto, T. *Chem. Commun.* **2009**, 7447–7464.

(16) Kalikadien, A. V.; Mirza, A.; Hossaini, A. N.; Sreenithya, A.; Pidko, E. A. *ChemPlusChem* **2024**, e202300702.

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

(18) Burai Patrascu, M.; Pottel, J.; Pinus, S.; Bezanson, M.; Norrby, P.-O.; Moitessier, N. *Nat. Catal* **2020**, *3*, 574–584.

(19) Harden, I.; Neese, F.; Bistoni, G. *Chem. Sci.* **2022**, *13*, 8848–8859.

(20) Ingman, V. M.; Schaefer, A. J.; Andreola, L. R.; Wheeler, S. E. *WIREs Comput. Mol. Sci.* **2021**, *11*, e1510.

(21) Taylor, M. G.; Burrill, D. J.; Janssen, J.; Batista, E. R.; Perez, D.; Yang, P. *Nat. Commun.* **2023**, *14*, 1–11.

(22) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. *Journal of Computational Chemistry* **2016**, *37*, 2106–2117.

(23) Sobez, J. G.; Reiher, M. *J. Chem. Inf. Model.* **2020**, *60*, 3884–3900.

(24) Alegre-Requena, J. V.; V., S. S. S.; Pérez-Soto, R.; Alturaifi, T. M.; Paton, R. S. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2023**, *13*, e1663.

(25) Kalikadien, A. V.; Valsecchi, C.; van Putten, R.; Maes, T.; Muuronen, M.; Dyubankova, N.; Lefort, L.; Pidko, E. A. *Chem. Sci.* **2024**, *15*, 13618–13630.

(26) Wen, J.; Wang, F.; Zhang, X. *Chem. Soc. Rev.* **2021**, *50*, 3211–3237.

(27) Coetzee, J.; Dodds, D. L.; Klankermayer, J.; Brosinski, S.; Leitner, W.; Slawin, A. M.; Cole-Hamilton, D. J. *Chem. - Eur. J.* **2013**, *19*, 11039–11050.

(28) Lawrence, S. A., *Amines: Synthesis, properties and applications*; Cambridge University Press: 2006.

(29) Narro, A. L.; Arman, H. D.; Tonzetich, Z. J. *Organometallics* **2019**, *38*, 1741–1749.

(30) Kinzel, N. W.; Demirbas, D.; Bill, E.; Weyhermüller, T.; Werlé, C.; Kaeffer, N.; Leitner, W. *Inorg. Chem.* **2021**, *60*, 19062–19078.

(31) Chernyshov, I. Y.; Pidko, E. A. *J. Chem. Theory Comput.* **2024**, *20*, 2313–2320.

(32) Chernyshov, I. MACE: MetAl Complexes Embedding, https://github.com/EPiCs-group/epic-mace, 2020.

(33) Weininger, D. *J. Chem. Inf. Comput.* **1988**, *28*, 31–36.

(34) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16 Revision C.01/C.02, 2016.

(35) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(36) Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. *Chem. Phys.* **2019**, *150*, 154122.

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

(38) Wilson, P. J.; Amos, R. D.; Handy, N. C. *Phys. Chem. Chem. Phys.* **2000**, *2*, 187–194.

(39) Minenkov, Y.; Sharapa, D. I.; Cavallo, L. *J. Chem. Theory Comput.* **2018**, *14*, 3428–3439.

(40) Sinha, V.; Laan, J. J.; Pidko, E. A. *Phys. Chem. Chem. Phys.* **2021**, *23*, 2557–2567.

(41) Kalikadien, A. V.; Pidko, E. A.; Sinha, V. *Digital Discovery* **2022**, *1*, 8–25.

(42) Silva, M. A.; Goodman, J. M. *Tetrahedron Lett.* **2005**, *46*, 2067–2069.

(43) Goodman, J. M.; Silva, M. A. *Tetrahedron Lett.* **2003**, *44*, 8233–8236.

(44) Weigend, F. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057.

(45) Jorner, K. Morfeus: molecular features for machine learning, https://github.com/digital-chemistry-laboratory/morfeus?tab=readme-ov-file, 2022.

(46) O'Boyle, N. M.; Tenderholt, A. L.; Langner, K. M. *J. Comput. Chem.* **2008**, *29*, 839–845.

(47) Baidun, M. S.; Kalikadien, A. V.; Lefort, L.; Pidko, E. A. *J. Phys. Chem. C* **2024**, *128*, 7987–7998.

6

ML-Guided Optimization of Phosphine-based Ligands for Nickel-Catalyzed Addition of Arylboronic Acids to Nitriles

Arylketones are valuable intermediates in synthesis and drug development, yet conventional methods for their preparation often rely on stoichiometric, highly reactive reagents. Here, we report a combined high-throughput experimental and computational study of Ni-catalyzed couplings of nitriles with arylboronic acids under mild conditions. A chemically diverse dataset of monophosphine and bisphosphine ligands was digitally generated and translated into molecular descriptors that explicitly account for conformational flexibility. Principal Component Analysis revealed that specific steric and electronic features (e.g. buried volume around the metal center and dipole moment) dominate the variance within the dataset, providing chemically interpretable dimensions for ligand space. Machine learning models trained on experimentally validated ligands achieved predictive performance sufficient to classify catalytic activity and were further applied to assess untested ligands. A transfer learning strategy, leveraging descriptors from Rh-based model structures, was demonstrated as a proof of concept for extending predictive scope across catalytic systems. While challenges such as false positives and negatives remain, this study establishes an integrated HTE-ML framework for ligand discovery and highlights the potential of transferable descriptor sets to accelerate catalyst optimization in homogeneous catalysis.

This Chapter will be published as: Pedrazzi, F.; Kalikadien, A. V.; Valsecchi, C.; Lefort, L.; Pidko, E. A. *in preparation* **2025**.[1]

6.1. Introduction

Arylketones are important organic compounds that serve as versatile building blocks in both chemical synthesis and drug development. An established synthetic pathway for arylketones is founded upon the nucleophilic addition of organometallic reagents, such as organolithium species, to various electrophilic substrates, typically followed by a hydrolytic work-up to afford the desired ketone products.[2–4] While these classical methodologies are undeniably robust and well-validated, their inherent reliance on stoichiometric quantities of highly reactive organometallic reagents presents operational and synthetic challenges. Significant drawbacks encompass issues related to chemoselectivity, the requirement for rigorously anhydrous conditions, and a restricted tolerance towards sensitive functional groups, thereby limiting their broad utility in the synthesis of complex molecules. Furthermore, growing concerns regarding the sustainability of reagents and the scalability of these processes have considerably stimulated research into developing alternative approaches that aim to mitigate the environmental and operational burdens associated with these conventional synthetic paradigms. Some of these alternative approaches involve the use of TM-based homogeneous catalysis.[5,6]

In particular, nickel-based catalysts emerge as a highly compelling alternative to the more established systems based on noble metals such as rhodium and palladium.[7–9] Furthermore, Ni's distinctive reactivity enables a spectrum of chemical transformations that prove notably challenging for Pd or Rh complexes.[10] In this work, we present the development of Ni-based catalytic systems for the coupling of nitriles and boronic acids to arylketones under relatively mild conditions.

The design and selection of ligands are important for precisely dictating the catalytic activity and selectivity of TM catalysts, with phosphine-based ligands historically maintaining a dominant position in the field. Conventional mono- and bisphosphine ligands, exemplified by triphenylphosphine, dppe, dppp, dppf, and BINAP, have been extensively used in both Ni- and Pd-catalyzed cross-coupling reactions. In these systems, their electronic and steric properties are critical, influencing both catalyst stability and reactivity profiles. The subsequent emergence of more sterically demanding ligand scaffolds, such as Xantphos, has further expanded catalyst tunability, proving access to challenging bond-forming transformations.[11]

Another important class of ligands is N-heterocyclic carbenes (NHCs).[12] Although their application in nitrile coupling with boronic acid derivatives is less widespeard than phosphines, NHC-based ligands have demonstrated the capacity to broaden the reactivity profile of Ni catalysts, offering novel opportunities to achieve transformations that remain inaccessible to traditional phosphine-based catalytic systems.[13]

Unfortunately, identification of the most efficient ligand is still largely driven by trial-and-error strategies in practical catalysis research. Data-science-driven approaches to experiment design are beginning to reshape this landscape,[14–16] particularly through the integration of High throughput experimentation (HTE) with computational chemistry, automation and machine learning (ML) for the construction and analysis of transferable ligand databases.[17–19] Notably, Sigman

and co-workers pioneered surrogate strategies for data-driven modeling, employing [ligand]Pd-Cl_2 complexes as model structures to calculate generalizable descriptors for Rh- and Pd-based catalysis.[19] More recent efforts have begun incorporating dynamic features such as conformational flexibility into these databases, enhancing their chemical fidelity and predictive utility.[20,21] Yet, the application of such approaches to systematically generated experimental datasets for Ni-based catalysis remains limited to small-scale examples with monophosphine ligands.[22–24]

Herein, we report the construction of a systematic experimental dataset using high-throughput experimentation (HTE) to screen Ni-based catalysts for the coupling of nitriles with boronic acids to form aryl ketones under relatively mild conditions. Building on this, we developed a computational dataset of molecular descriptors that explicitly incorporates conformational flexibility. Linear dimensionality reduction using Principal Component Analysis (PCA) was employed to identify key ligand features governing variance within the dataset, thereby enabling rationalization of the search space. ML models were subsequently trained on a curated set of ligands with experimentally determined performance metrics, yielding predictive models for product conversion.

Extending their applicability, these models were then used to classify the activity of previously untested chiral ligands, with descriptors derived from an updated version of our Rh-based asymmetric hydrogenation dataset presented in Chapter 2.[25] While the fully out-of-domain strategy described in Chapter 2 proved unsuccessful for unseen substrates, the present study demonstrates that a transfer learning framework, leveraging descriptors from Rh-based model structures to guide predictions for Ni-catalyzed reactions, can serve as a proof of concept for an ML-driven recommendation system to identify ligand candidates predicted to promote aryl methyl ketone formation.

6.2. DATA GENERATION

6.2.1. EXPERIMENTAL DATASET

As described in Chapter 2, experiments were carried out by Janssen's HTE group supporting chemical process development where several workflows for the screening of catalysts were developed.[25] For this arylation reaction, four plates were screened, with three plates each containing 24 ligands in combination with a Ni salt. The fourth plate, containing 96 chiral ligands in combination with a Ni salt, was screened in order to test the predictivity power of ML models (*vide infra*). All ligands were selected from a pre-assembled 96-ligands library, based on existing literature and commercial availability. In the ligand library, bisphosphines were the most prevalent comprising 47 (49%) entries, followed by monophosphines with 14 (15%) entries.

In this study, the selected ligands were pre-mixed with a Ni(II) source, and the obtained complexes tested at mild conditions (60 °C) (Figure 6.1), following prior published procedures.[8,9] All monodentate ligands (including phosphines and NHCs) were tested in both 1:1 and 2:1 stoichiometric ratios to the metal.

The outcome of the reactions was evaluated by UPLC-MS analysis under batch conditions. Initially, conversion was assessed on the basis of consumption of the

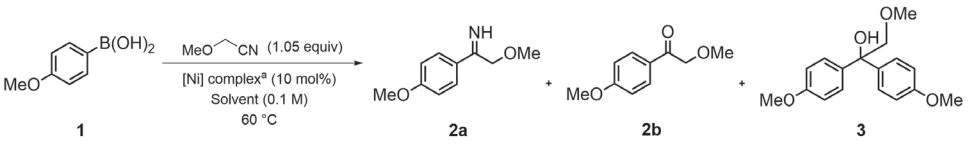


Figure 6.1.: Selected conditions for HTE screening. Reactions were carried out at 60 °C in selected solvent (100 µL) with **1** (10 µmol) and methoxyacetonitrile (1.05 equiv, 10.5 µmol), in presence of pre-formed Ni-complex (0.1 equiv, 1 µmol), for 6 h. [(a)] Pre-complexation performed by mixing 40 µL of a solution of Ni perchlorate hexahydrate in MeTHF (0.025 M, 1 µmol) with 40 µL of ligand's solution in appropriate solvent, either DCM or THF (0.025-0.05 M, 1-2 µmol). The mixture was then sirred at room temperature overnight and dried in parallel evaporator.

starting material (**1**). However, multiple products were observed in addition to the desired ketone, including the imine intermediate (**2a**), which proved relatively stable under the nearly anhydrous reaction conditions, the alcohol derivative (**3**) arising from sequential addition of a second equivalent of substrate **1**, along with traces of unidentified byproducts resulting in part from decomposition of **1** (Figure 6.2).

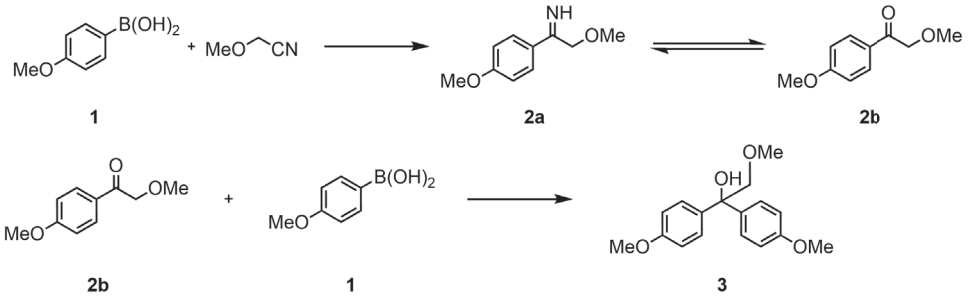


Figure 6.2.: Hypothetical formation pathways for major products observed.

Consequently, product distribution was assessed by integrating the UPLC peaks corresponding to substrate **1**, the combined imine/ketone fraction (**2a**+**2b**), and compound **3** relative to an internal standard. The conversion to product **2** was calculated as the ratio of its peak area to the sum of all three areas (**1**+**2**+**3**), with an analogous procedure applied to product **3**. Although this approach does not provide an absolute quantification, as it relies solely on net UV absorbance, it offers a consistent and practical metric for evaluating reaction output under batch conditions.

From these experimental results (Figure 6.3) an overall of 69 data points were collected, including repetitions and varying ratios of monodentate ligands. The three screenings show robustness in conversion into product **2** with some variability for a repeated screening of L1 (dCyhpe). The results indicated that six bisphosphine

ligands showed more than 80% combined product conversion to product **2a**/**2b**, namely: L1 (dCyhpe), L2 (dCyhpp), L3 (d^{i}Prpp), L4 (dCyppp), L37 (QuinoxP) and L39 ((R,R)-Me-DuPhos), which were labeled as 'active'. All remaining ligands exhibited either 0% or negligible conversion under the tested conditions and were therefore classified as 'inactive'. Of the 53 ligands, 25 bisphosphine ligands and 11 monophosphine ligands were selected for model development, which included L1, L2, L3, L4 and L39 as active samples. The selected ligands contained simple structural motifs that were well captured by the SMILES notation, making them readily usable in our developed workflows for fully automated computational screening.

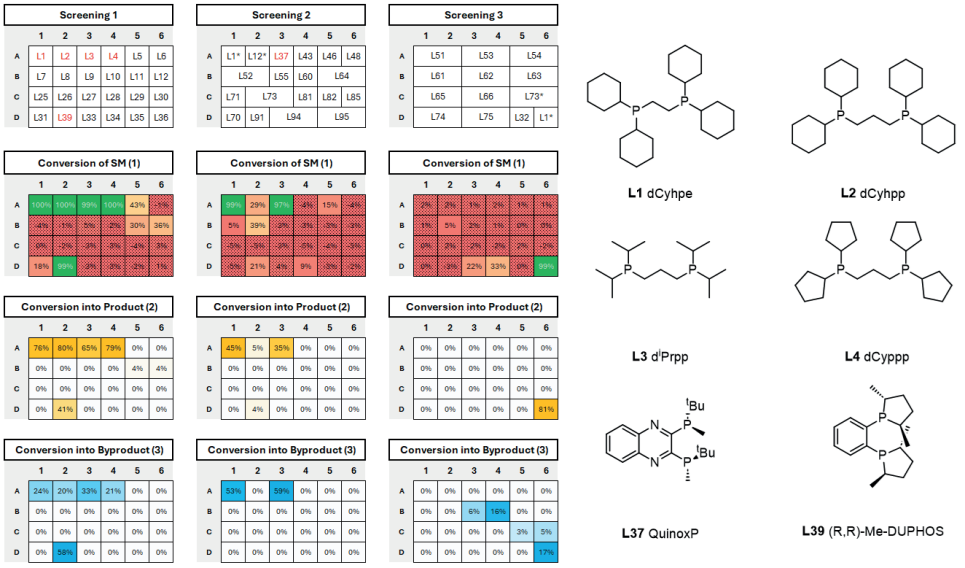


Figure 6.3.: Experimental results of ligands screenings. (*) Ligand repeated from previous plates, L1 showed some variability on Screening 2.

6.2.2. DESCRIPTOR GENERATION

Two surrogate 'model' structures were used to computationally represent the ligands and a substrate, providing a steric constraint on the conformational space of the ligand bound to the metal. Each model structure exhibits different degrees of conformational flexibility: a rigid [ligand]-Ni(II)-Cl_2 complex, in which Cl_2 creates a symmetric coordination environment, and a more flexible [ligand]-Ni(II)-(CH_3CN)(−pOMe(C_6H_4)) complex, representing a plausible reaction intermediate featuring an asymmetric coordination environment with increased conformational freedom.[26]

The atomic structures of the coordination complexes including all possible stereoisomers were generated with our MACE Python package,[27] followed by exhaustive conformer search utilizing CREST using the GFN2-xTB/GFN-FF hybrid

potential.[28–31] While bisphosphine complexes generally adopted a square-planar geometry, CREST occasionally generated tetrahedral distortions for these systems, and produced both cis- and trans-coordination modes for the monodentate ligands. The resulting CREST-based conformer ensembles were refined via DFT geometry optimization, performed using Gaussian 16 C.02.[32] The PBE0-D3(BJ)/def2-SVPP level of theory[33–35] was applied, known for its reliable accuracy and efficiency for the description of TM complexes[26,36,37]. The nature of each stationary point was confirmed via frequency analysis. Thermochemical parameters (e.g. ZPE, finite temperature corrections and entropy contributions to Gibbs free energies) were computed from analytical frequencies (Hessian) at 298.15K and 1 atm. For conformers displaying imaginary frequencies, the pyQRC Python script (version 1.0.3)[38,39] was employed to generate revised input geometries, which were then reoptimized with the same DFT settings. Conformers that retained imaginary frequencies after two attempts at reoptimization were excluded from further evaluation.

All conformers were subjected to descriptor calculation using the Open Bidentate Ligand eXplorer (OBeLiX) Python package.[25,40] The two coordinating atoms of the ligand were distinguished based on their charge, with the label min/max denoting the least/most positively charged donor atom, respectively. In addition to descriptors derived from the Ni-based model structures, we also generated electronic descriptors for the ligand alone (labelled as 'free_ligand'). For this purpose, the ligand geometry was extracted from the optimized structure of the corresponding Ni-based model structure followed by a single-point (SP) DFT calculation. With the DFT-optimized conformers as input, this workflow resulted in a total of 24 descriptors for each individual conformer including steric, geometric and electronic properties (see SI 'Descriptor Overview' for a full list of descriptors). Simultaneously, the DFT-based energy for each individual conformer was extracted. This resulted in four datasets of descriptors: individual descriptors per conformer and Boltzmann averaged descriptors for both the rigid and flexible model structure. Afterwards, conformers were filtered such that only those within 5 kJ/mol of the global minimum energy were retained for further analysis. The ensemble was thus filtered to include only those conformers that are thermodynamically accessible and significantly populated under ambient conditions. RMSD calculations were performed using the SpyRMSD Python package.[41] Together with the energy criterion, an RMSD threshold of 1 Å was applied to select 'unique' conformers representing the ensemble (see SI S2).

6.3. Data Analysis

A PCA was conducted on our dataset of DFT-based descriptors to visualize the ligand space and characterize the search space. Since the results were closely related for both the Boltzmann averaged and lowest-energy conformer datasets (see SI S2), the dataset with descriptors for the lowest-energy conformer was selected for further analysis. Applying PCA directly to the 24 descriptors and selecting the first two principal components, explaining in total 62% of the variance for both the rigid and flexible model structure, resulted in identical well-defined clusters in both cases (see SI S2 for the PCA on descriptors based on the flexible model

structure). Hence, the results for only the rigid model structure are discussed in the remainder of this section. The resulting PCA (Figure 6.4a) shows a distinct cluster of active catalysts (in red) on the left side, characterized by negative values for the first principal component. A catalyst was labelled as active if at least 80% combined product conversion was observed. Interestingly, the inactive catalysts (in grey) within this cluster (L7 (dppm), L8 (dppe), L10 (dppeO), L11 (dppp), L12 (dppb)) are distinguished by aromatic substituents connected to the phosphorus donor atoms. The loading plot (Figure 6.4b) indicates that Principal Component (PC)

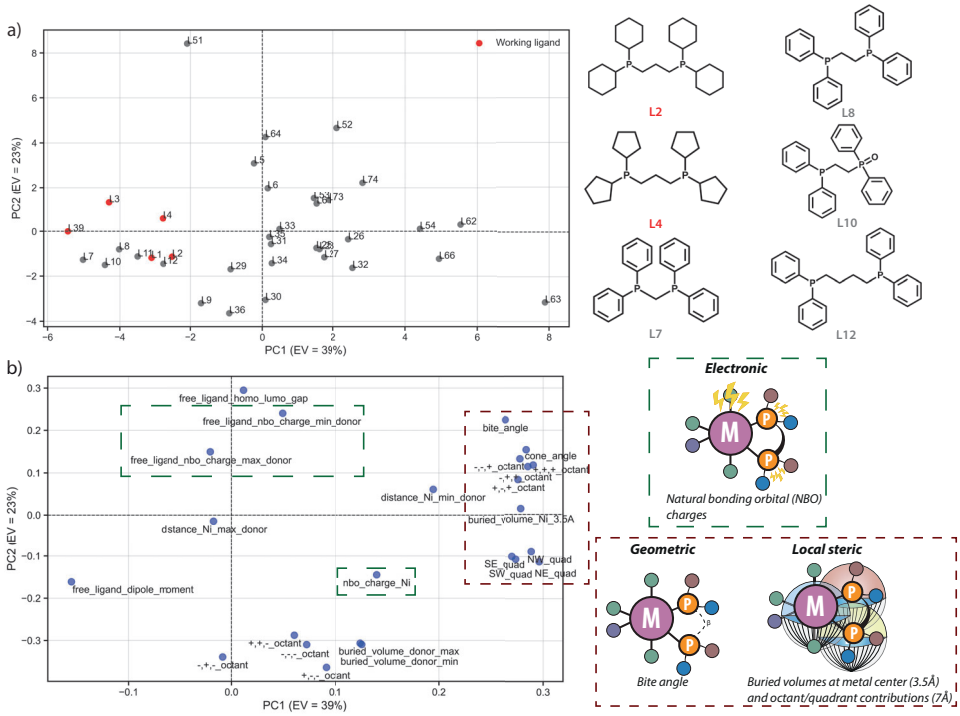


Figure 6.4.: PCA of ligand descriptors. (a) Scores plot showing clustering of active (red) and inactive (grey) ligands. Representative ligand structures are highlighted, with active ligands (L2, L4) located within the cluster on the left side. (b) Loading plot illustrating the contributions of steric (red dashed box) and electronic (green dashed box) descriptors to the first two principal components. PC1 (39% variance) is dominated by steric features such as bite angle, cone angle, and buried volumes, whereas PC2 (23% variance) reflects local steric and electronic contributions including HOMOLUMO gap and NBO charges.

1, accounting for 39% of the total variance, is primarily dominated by geometric and steric descriptors. The strongest contributions arise from bite and cone angles (geometric), the buried volume centered around the metal center (global steric),

and the local buried volumes describing steric occupation of octants around the substrate coordination site (i.e., Cl_2 in the rigid model structures), each with loading values around 0.3. In addition, smaller but non-negligible contributions are observed for electronic descriptors, including the NBO charge of Ni and the dipole moment. This axis therefore mainly separates compact ligands such as L39 ((R,R)-Me-DuPhos) and L7 (dppm), which show low steric occupation, from bulky ligands such as L63 ($(o\text{-}MeO_2Ph)_3P$), which is found at the opposite extreme.

In contrast, PC2, which explains 23% of the variance, is strongly influenced by electronic descriptors. Positive contributions arise from the HOMO-LUMO gap and the NBO charge on the donor atoms, while negative contributions stem from buried volumes centered at the phosphorus donor atoms as well as steric occupation at the minus Z direction of the octants. The separation in local electronic properties by PC2 is illustrated by L51 ($(EtO)_3P$) and L36 (Pseudo-XantPhos). The triethyl phosphite has a large HOMO-LUMO gap and acts as a nucleophile and σ donor with very low π acceptor ability compared to aromatic phoshporus ligands such as Pseudo-Xantphos where the aromatic backbone reduces localization of the electron density on the phosphorus atoms.

Taken together, the PCA provides a chemically intuitive view of the descriptor space - suggesting that active ligands are compact ligands (negative PC1) with intermediate electronic behavior (PC2 in [-2,2] range) - but does not directly establish whether these features can be exploited to predict catalyst activity. To address this, we applied supervised ML classification models to test whether the descriptor space together with information of conformational flexibility can indeed distinguish active from inactive ligands.

6.4. ML Model Development

The prediction of catalyst activity was formulated as a binary classification task using a reduced descriptor set consisting of the first two principal components (PC1 and PC2) together with the conformer count per ligand obtained after energy- and RMSD-based filtering. In this way, both the dominant steric and electronic axes identified by PCA and the extent of conformational flexibility were incorporated as features. Again, a catalyst was labelled as active if at least 80% combined product conversion was observed (see SI 'Ligands Overview' for an overview of ligands and experimental results). Logistic regression (LR) and Decision Tree (DT) classifiers were selected as simple yet interpretable representatives of linear and non-linear models for small datasets. In LR, the logarithmic-odds of the class probabilities are modeled as a linear function of the input descriptors, allowing for direct interpretation of the sign and magnitude of coefficients. In contrast, DT models partition the features space recursively, ranking descriptors by their ability to separate the training data, which provides a non-linear perspective on feature importance.

The dataset was divided multiple times into training and test sets using an 80/20 stratified split to preserve the class balance. Each test set therefore contained eight catalysts, of which only one was active. In training, the weight of active catalysts was higher than inactive ones, accounting for the class imbalance in the dataset. In

particular, weights were adjusted to be inversely proportional to class frequencies. The classification performance was quantified using the Balanced Accuracy (BA), which accounts for the imbalance between active and inactive catalysts. A BA value above 0.5 thus indicates that the model correctly identified the active catalyst in the test set. To ensure robustness and assess variability in the dataset, 500 independent random training/sest splits were tested to assess model performance.

To further verify that the models captured meaningful structure-activity relationships rather than spurious correlations, *y*-scrambling tests were performed. In this procedure, the activity labels were randomly shuffled, and the models retrained, providing a baseline distribution of performance expected from chance correlations.

Both LR and DT models resulted in a BA equal to 0.90 and 0.86 on average respectively, outperforming their *y*-scrambled counterparts showing an average BA equal to 0.52 and 0.54 respectively. This highlights that the models learned meaningful patterns in the data (Figure 6.5a). LR coefficients (Figure 6.5b) consistently assigned a strong negative weight to PC1 (around −1.5), indicating that ligands with low steric crowding near the metal center were more likely to be active. The conformer count also contributed negatively (−0.75), suggesting that high conformational flexibility is detrimental to activity. By contrast, PC2 coefficients were close to zero, suggesting that the modeling of catalytic activity is minimally influenced by electronic descriptors. The DT feature importances echoed this trend (Figure 6.5c): PC1 dominated the classification with an importance of 0.8, while conformer count contributed moderately (0.18) and PC2 was negligible (0.02).

6.5. TRANSFER LEARNING

6.5.1. DATASET FOR TRANSFER LEARNING

Having established that our ML models can identify meaningful structure-activity relationships within the Ni dataset, we next sought to evaluate their ability to generalize to an entirely new chemical space. This transfer learning step provides a stringent test of model generalizability, as it involves predicting the performance of ligands not included in training.

For this purpose, we selected an existing descriptor set based on a library of 96 previously investigated chiral ligands in the context of Rh-catalyzed asymmetric hydrogenation presented in Chapter 2.[25] These ligands had not been tested at Janssen's HTE lab for the Ni-catalyzed aryl ketone formation, making them an ideal benchmark for prospective predictions. To ensure comparability, the computational protocol mirrored that of the Ni dataset, with the difference being that descriptors were generated on a Rh-NBD model structure, as used in Chapter 2.[25] These Rh-based model structures were subjected to conformer sampling, DFT optimization, and descriptor calculation using identical settings (*vide supra*). The only modification was the application of our workflow presented in Chapter 4 which utilized DBSCAN clustering to filter conformers prior to DFT refinement. This reduced computational cost and improved automation in handling conformational ensembles.[42]

The resulting dataset contained 96 chiral ligands (labeled as 'Unseen Ligand',

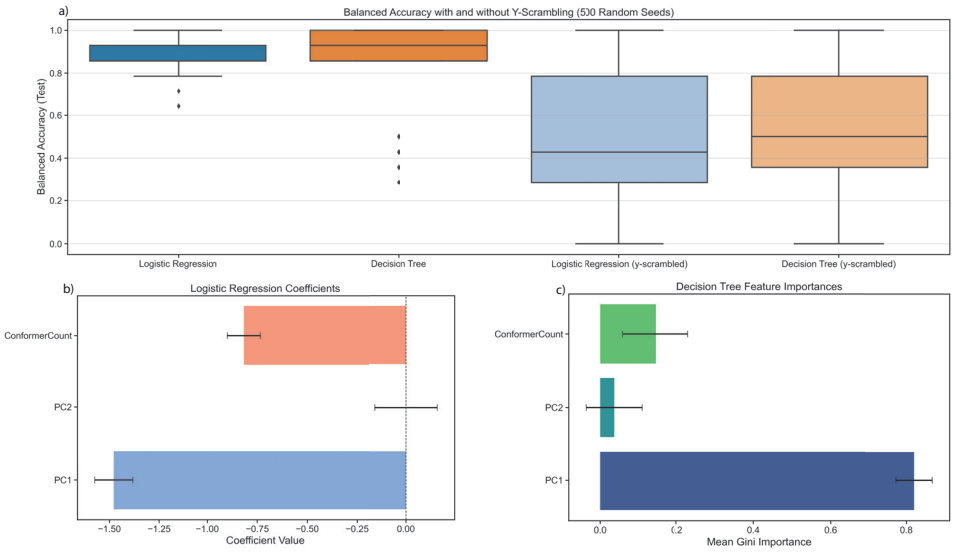


Figure 6.5.: Classification performance and feature importance analysis of logistic regression and decision tree models. (a) Distribution of balanced accuracies (BA) across 500 stratified train/test splits compared with *y*-scrambled controls. (b) Coefficient magnitudes for logistic regression and (c) feature importances for decision trees.

UL1UL96), each described by the same descriptor set as the Ni complexes ensuring consistency between datasets. For model evaluation, ligands were again classified into active and inactive categories based on a threshold of 80% combined product conversion.

6.5.2. Cross-projection of PCA and Ensemble Predictions

To place the Rh-based chiral ligand descriptor library in the context of the Ni-derived chemical space, the descriptors of the new dataset were projected onto the PCA model built from the Ni complexes. Since conformer count values were not available for the Rh dataset, the analysis was restricted to PC1 and PC2. The resulting projection is shown in Figure 6.6, where the coloring also reflects ensemble classification outcomes (*vide infra*).

As an internal reference, ligand L39 ((R,R)-Me-DuPhos) from the Ni dataset coincides with UL69 ((R,R)-Me-DuPhos) from the new ligand library. Their positions in the PCA space were highly similar, with only minor deviations attributable to changes in electronic (PC2) and geometric/steric (PC1) descriptors caused by the different square planar Ni and Rh model structures. This consistency demonstrates that the cross-projection preserves the essential steric and electronic features of this ligand across metal centers. The projected distribution reproduced chemically

meaningful trends. Along PC1, steric variation is captured: for instance, the small ligand UL67 ((R)-QuinoxP) appears at the left-hand extreme, while the bulkier MandyPhos-type ligand UL21 (SL-M003-1) lies toward the right, though not as far as the very bulky L63 ((R,R,R)-Xyl-SKP) from the Ni dataset. PC2 reflects electronic effects: UL21 (SL-M003-1), bearing strongly electron-withdrawing substituents, is located at the lower end of PC2, whereas L51 ($(EtO)_3P$), a trialkyl phosphite with electron-donating ethoxy groups, appears at the opposite extreme.

Before applying the Ni-trained models, they were retrained without the conformer count feature and their performance was verified. Although a performance drop was observed, both LR and DT classifiers retained average BA values of approximately 0.8 on the test set (see SI S3), supporting their suitability for prediction. An ensemble scheme was subsequently constructed from models exceeding 0.8 BA, classifying ligands as active if at least 75% of the selected LR and DT models agreed. This confidence score represents the average fraction of LR and DT models that classify a ligand as active. This threshold resulted in the inclusion of 400 LR and 404 DT models in the ensemble. Analysis of the corresponding test sets revealed that the five active ligands were sampled in a balanced manner across LR models, with each ligand appearing in approximately 1823% of the test sets. In contrast, the DT models displayed a more uneven distribution: ligand L2 (dCyhpp) appeared in only 5% of the test sets, whereas L4 (dCyppp) and L3 (D^iPrpp) were included more frequently, at 28% and 24%, respectively. While this imbalance suggests that DT models are more sensitive to the particular data partitions, the ensemble approach mitigates these biases by combining predictions across a large number of diverse models. In this way, the consistent coverage provided by the LR models compensates for the uneven sampling in the DT models, and the 75% consensus rule ensures that final activity predictions are not driven by the idiosyncrasies of individual models, but rather by reproducible agreement across the ensemble.

This ensemble prediction is overlaid in Figure 6.6: ligands from the first plate are shown in grey (observed inactive) and red (observed active), while predictions on the second plate with unseen ligands are shown in light blue (predicted inactive) and blue (predicted active). A full overview of all experimental results can be found in the SI (see SI 'A list and visualization of all ligands'). This figure is also provided as an interactive plot with ligands displayed upon hovering over the datapoints (see SI 'Interactive Figure 6.6'). Five chiral ligands were predicted to be active by 100% of the models: UL53 ((S,S)-iPr-BPE), UL66 ((R,R)-Et-BPE), UL67 ((R)-QuinoxP), UL69 ((R,R)-Me-DuPhos), and UL84 ((3S,3'S)-BABIBOP). Of these, two (UL67 and UL69) were experimentally confirmed as active. It should be noted, however, that UL69 is identical to L39 from the first plate and therefore not a truly unseen case. Instead, its correct classification demonstrates the internal consistency of the cross-projection. Among the remaining ligands, UL84 was misclassified but displayed moderate conversion (40%), just as UL66 (71%) while UL53 was experimentally inactive. Conversely, five experimentally active ligands, UL4 (SL-J004-1, a JosiPhos ligand), UL55 ((1R,1'R,2S,2'S)-DuanPhos), UL64 ((R,R)-Ph-BPE), UL68 ((R,R)-Et-DuPhos), UL95 ((S,S)-XylSKEWPHOS) were missed, despite close structural similarity between some misclassified and correctly classified ligands (e.g.,

UL68 and UL69 from the DuPhos family).

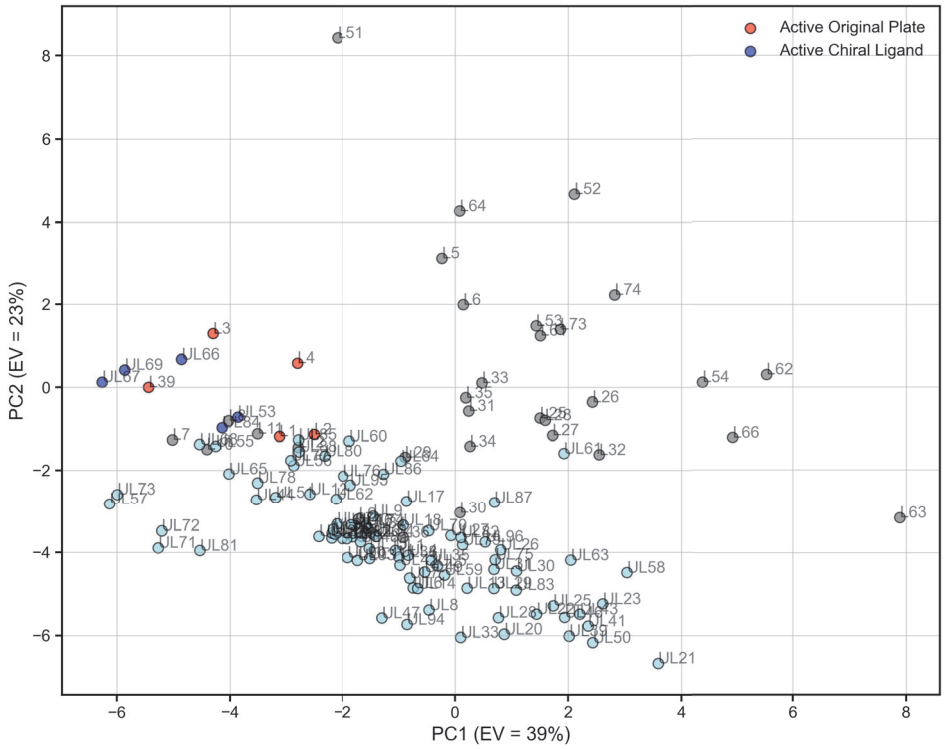


Figure 6.6.: Cross-projection of unseen chiral ligands (light blue/blue) into the PCA space derived from Ni complexes (grey/red). Coloring indicates ensemble predictions, with active ligands shown in red/blue and inactive ligands in grey/light blue. The projection preserves steric (PC1) and electronic (PC2) trends across both datasets, while ensemble predictions highlight candidate chiral ligands for experimental validation.

Predictions made based on ML models trained on the flexible model structure showed a similar pattern, where 12 chiral ligands were predicted to be active by 100% of the models. The detailed analysis for these models can be found in the SI (see SI S3). Five of these were experimentally confirmed to be active: UL55 ((1R,1'R,2S,2'S)-DuanPhos), UL67 ((R)-QuinoxP), UL68 ((R,R)-Et-DuPhos) and UL69 ((R,R)-Me-DuPhos). The remaining ligands were false positives of which UL84 was misclassified but displayed moderate conversion (40%). In this case, only three experimentally active ligands, UL4 (SL-J004-1, a JosiPhos ligand), UL64 ((R,R)-Ph-BPE) and UL95 ((S,S)-XylSKEWPHOS) were missed.

Overall, this cross-projection and ensemble prediction analysis demonstrates that the trained models can generalize to unseen ligands using a set of descriptors generated on a different Rh-based model structure, capturing steric and electronic

trends across chemical space. While not without false negatives and false positives, the workflow provides a systematic approach to prioritize ligand candidates and highlights the potential of transfer learning for homogeneous catalysis design.

6.6. CONCLUSION

Through the integration of high-throughput experimentation and computation, we established a chemically diverse dataset for the nickel-catalyzed addition of arylboronic acids to nitriles, yielding arylketones as target products. This dataset provided a robust foundation for the development and testing of machine learning models aimed at identifying active ligands for this reaction.

Our computational workflow enabled the systematic generation of ML-ready ligand representations through descriptor sets that explicitly incorporate conformational dynamics. By linking chemically intuitive descriptors with dimensionality reduction and ML modeling, we demonstrated that ML models capture the key steric and electronic features relevant to catalytic activity on a small dataset of 36 ligands.

The transfer learning approach presented here represents a proof of concept for extending descriptors derived from Rh-based model structures to predict ligand performance on Ni-based reactions. While successful in identifying chemically meaningful trends and highlighting some promising candidates, several limitations remain. The degree of predicted false negatives and positives was relatively high, especially when descriptors based on the rigid model structures were used for training of the ML models. Involving a mechanistically representative model structure with relevant conformational flexibility is thus of importance. The domain of applicability for these models could be extended by expanding the training of ML models used for prediction, e.g. through an active learning scheme. Additionally, the conformer count was excluded as a feature for predictions, which although not detrimental, impacts the performance of the models slightly. The concept of transfer learning can be extended beyond the used Rh-NBD model structure from our previous studies,[25,26,42] alternatives such as the Pd-π-allyl surrogate proposed by Stenfors et al.[20] should also be tested in the context of transferable representations.

Taken together, this study demonstrates that combining high-throughput experimentation, descriptor-based ML modeling, and ensemble learning can provide a powerful framework for ligand discovery. With further refinements in model structures and descriptor selection, such approaches have the potential to accelerate catalyst optimization and enable more predictive, generalizable models for homogeneous catalysis.

DATA AVAILABILITY

The Python package for the featurization of catalyst structures, OBeLiX, is available through the GitHub organization page of the ISE group at TU Delft: EPiCs-group OBeLiX (https://github.com/EPiCs-group/obelix).

All datasets used in this study are provided with an extensive README via 4TU.ResearchData at https://doi.org/10.4121/e77cddf1-7ffc-4cbb-a3c9-bf8adc352192. The following resources are included:

- A PDF with supporting information on experimental and computational procedures ('supporting_information.pdf')
- A list and visualization of all ligands with experimental results ('ligand_overview.xlsx')
- An Excel file categorizing and describing all descriptors ('descriptors_overview.xlsx')
- An interactive version of Figure 6.6 ('interactive_figure5.html')
- Python code, numerical data, (interactive) figures and CREST/DFT data together with instructions on how to reproduce this study ('code_and_data.zip')

CONTRIBUTIONS

F. Pedrazzi and **A.V. Kalikadien** contributed equally to this work. **F. Pedrazzi:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing Original Draft, Writing Review & Editing, Visualization **A.V. Kalikadien:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing Original Draft, Writing Review & Editing, Visualization, Project administration **C. Valsecchi:** Conceptualization, Software, Writing Review & Editing **L. Lefort:** Supervision, Conceptualization, Resources, Funding acquisition, Writing Review & Editing, Project administration **E.A. Pidko:** Supervision, Conceptualization, Resources, Funding acquisition, Writing Review & Editing, Project administration

REFERENCES

(1) Pedrazzi, F.; Kalikadien, A. V.; Valsecchi, C.; Lefort, L.; Pidko, E. A. *in preparation* **2025**.

(2) Nahm, S.; Weinreb, S. M. *Tetrahedron Lett.* **1981**, *22*, 3815–3818.

(3) Smith, A. B.; Boldi, A. M. *J. Am. Chem. Soc.* **1997**, *119*, 6925–6926.

(4) Lee, B.; Chirik, P. J. *J. Am. Chem. Soc.* **2020**, *142*, 2429–2437.

(5) Willis, M. C. *Chem. Rev.* **2010**, *110*, 725–748.

(6) Wu, X.-F.; Neumann, H.; Beller, M. *Chem. Soc. Rev.* **2011**, *40*, 4986–5009.

(7) Ananikov, V. P. *ACS Catal.* **2015**, *5*, 1964–1971.

(8) Wong, Y.-C.; Parthasarathy, K.; Cheng, C.-H. *Org. Lett.* **2010**, *12*, 1736–1739.

(9) Tu, D.-H.; Li, Y.; Zhao, B.; Gu, Y.-J.; Wang, B.; lu, J. *Synlett* **2018**, *29*, 593–596.

(10) Tasker, S. Z.; Standley, E. A.; Jamison, T. F. *Nature* **2014**, *509*, 299–309.

(11) Van Leeuwen, P. W.; Kamer, P. C.; Reek, J. N.; Dierkes, P. *Chem. Rev.* **2000**, *100*, 2741–2770.

(12) Fortman, G. C.; Nolan, S. P. *Chem. Soc. Rev.* **2011**, *40*, 5151–5169.

(13) Wang, H.; Lu, G.; Sormunen, G. J.; Malik, H. A.; Liu, P.; Montgomery, J. *J. Am. Chem. Soc.* **2017**, *139*, 9317–9324.

(14) Williams, W. L.; Zeng, L.; Gensch, T.; Sigman, M. S.; Doyle, A. G.; Anslyn, E. V. *ACS Cent. Sci.* **2021**, *7*, 1622–1637.

(15) Ickes, A. R.; Liles, J. P.; Borlinghaus, N.; Henle, J.; Swiatowiec, R.; Kaushik, N. P.; Braje, W. M.; Harper, K. C.; Shekhar, S.; Sigman, M. S. *J. Am. Chem. Soc.* **2025**, *147*, 28981–28992.

(16) Romer, N. P.; Min, D. S.; Wang, J. Y.; Walroth, R. C.; Mack, K. A.; Sirois, L. E.; Gosselin, F.; Zell, D.; Doyle, A. G.; Sigman, M. S. *ACS Catal.* **2024**, *14*, 4699–4708.

(17) Fey, N. *Chem. Cent. J.* **2015**, *9*, 38.

(18) Gensch, T.; dos Passos Gomes, G.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-DAddario, M.; Sigman, M. S. *et al. J. Am. Chem. Soc.* **2022**, *144*, 1205–1217.

(19) Dotson, J. J.; van Dijk, L.; Timmerman, J. C.; Grosslight, S.; Walroth, R. C.; Gosselin, F.; Püntener, K.; Mack, K. A.; Sigman, M. S. *J. Am. Chem. Soc.* **2022**, *145*, 110–121.

(20) Stenfors, B.; Cadge, J.; Aikonen, S.; Luchini, G.; Wahlers, J.; Koh, K.; Muuronen, M.; Menche, M.; Pfeifle, M.; Keto, A.; Paton, R.; Sigman, M.; Wiest, O. *ChemRxiv preprint* **2025**.

(21) Cadge, J. A.; Hart, S. D.; Walroth, R. C.; Mack, K. A.; Sigman, M. S. *ChemRxiv preprint* **2025**.

(22) Wu, K.; Doyle, A. G. *Nat. Chem.* **2017**, *9*, 779–784.

(23) Newman-Stonebraker, S. H.; Smith, S. R.; Borowski, J. E.; Peters, E.; Gensch, T.; Johnson, H. C.; Sigman, M. S.; Doyle, A. G. *Science* **2021**, *374*, 301–308.

(24) Newman-Stonebraker, S. H.; Wang, J. Y.; Jeffrey, P. D.; Doyle, A. G. *J. Am. Chem. Soc.* **2022**, *144*, 19635–19648.

(25) Kalikadien, A. V.; Valsecchi, C.; van Putten, R.; Maes, T.; Muuronen, M.; Dyubankova, N.; Lefort, L.; Pidko, E. A. *Chem. Sci.* **2024**, *15*, 13618–13630.

(26) Baidun, M. S.; Kalikadien, A. V.; Lefort, L.; Pidko, E. A. *J. Phys. Chem. C* **2024**, *128*, 7987–7998.

(27) Chernyshov, I. Y.; Pidko, E. A. *J. Chem. Theory Comput.* **2024**, *20*, 2313–2320.

(28) Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.; Spicher, S.; Steinbach, P.; Wesoowski, P. A.; Zeller, F. *J. Chem. Phys.* **2024**, *160*, 114110.

(29) Pracht, P.; Bohle, F.; Grimme, S. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(30) Bannwarth, C.; Ehlert, S.; Grimme, S. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(31) Grimme, S. *J. Chem. Theory Comput.* **2019**, *15*, 2847–2862.

(32) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16 Revision C.01/C.02, 2016.

(33) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(34) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(35) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(36) Sinha, V.; Laan, J. J.; Pidko, E. A. *Phys. Chem. Chem. Phys.* **2021**, *23*, 2557–2567.

(37) Kalikadien, A. V.; Pidko, E. A.; Sinha, V. *Digit. Discov.* **2022**, *1*, 8–25.

(38) Goodman, J. M.; Silva, M. A. *Tetrahedron Lett.* **2003**, *44*, 8233–8236.

(39) Silva, M. A.; Goodman, J. M. *Tetrahedron Lett.* **2005**, *46*, 2067–2069.

(40) Kalikadien, A. V.; Mirza, A.; Hossaini, A. N.; Sreenithya, A.; Pidko, E. A. *ChemPlusChem* **2024**, e202300702.

(41) Meli, R.; Biggin, P. C. *Journal of Cheminformatics* **2020**, *12*, 49.

(42) Finta, S.; Kalikadien, A. V.; Pidko, E. A. *J. Chem. Theory Comput.* **2025**, *21*, 5334–5345.

7

Performance of Meta's Universal Model for Atoms Across the Conformational and Configurational Space of Diverse TM Catalysts

Machine learning interatomic potentials (MLIPs) promise to transform computational catalysis by delivering near-DFT accuracy at a fraction of the computational cost. Here, we evaluate the Universal Machine-learning Potential for Atoms (UMA) on two datasets of transition-metal complexes presented in Chapter 5 and 6. UMA enables high-throughput evaluations in seconds per structure on consumer-grade GPUs. Analysis of per-ligand Spearman rank correlations ($\rho > 0.6$, $p < 0.05$) reveals variability in ranking reliability that is not captured by aggregate metrics such as R^2 or RMSE. However, these inaccuracies are shown to mainly occur in the near-DFT accuracy regime where these complexes are practically indistinguishable. For square-planar Ni complexes, reliable rankings are obtained for 79% of ligands in rigid Ni-Cl_2 complexes and drop to 58% for flexible asymmetric coordination environments, particularly only when conformers differ by < 2 kJ/mol. Dataset 2 shows a similar trend, with 63% and 52% reliability for Ru(II) and Mn(I) complexes, respectively, and, as expected, challenges for fluxional systems with small (< 5 kJ/mol) relative energy gaps. These findings highlight the promise of MLIPs for both rigid, well-defined systems and highly flexible or fluxional catalysts, while underscoring the need to combine the speed of ML with validation and domain expertise to ensure robust and meaningful chemical insights.

This Chapter will be published as: Kalikadien, A. V.; Pidko, E. A. *J. Phys. Chem. A.* **2026**, *accepted.*[1]

7.1. INTRODUCTION

DFT is indispensable for modeling catalytic reactions.[2–4] In homogeneous catalysis, energetic barriers and molecular properties are typically derived from optimized geometries and their electronic structure.[2,5] Although the molecular structures of these catalysts are relatively well-defined, their accurate calculations using even most modern quantum chemical methods come at a considerable computational cost.[6] In fact, exhaustive computational sampling of the conformational and configurational space of a metal-ligand complex with DFT calculations can incur costs similar to the operating costs of wet-lab experimentation.

To illustrate this, consider a simple thought experiment. In high-performance computing (HPC), Standard Billing Units (SBUs) are used to quantify computational usage and cost, defined as the product of the number of CPU cores and the number of wall-time hours. In our experience, a typical TM complex requires approximately 4 hours of computation using 32 CPU cores to optimize structure and carry out frequency analysis with hybrid DFT functional and double-zeta quality basis set. Given an optimistic estimate cost of €0.01 per SBU in the Netherlands, optimizing 10,000 such complexes would amount to roughly €12,800. In high-throughput screening efforts, such scales are common.[7,8] It is therefore essential to develop reductionist approaches that maintain chemical accuracy while significantly lowering computational demand.[7,9]

Machine Learning Interatomic Potentials (MLIPs) offer a promising solution, enabling the approximation of DFT-level energies within seconds.[10–13] In the early seminal papers, MLIPs were limited to generating interatomic potentials for highly specific systems.[13,14] Unfortunately, to this day, a central challenge in developing MLIPs still lies in achieving sufficient generalization across the diverse domains and tasks for which DFT is employed.[15–17] Recently, Meta released large, chemically diverse datasets designed to support general-purpose models.[18,19] Alongside these datasets, a family of Universal Models for Atoms (UMA) was presented.[19] These general-purpose models have demonstrated competitive or superior performance in terms of accuracy, inference speed, and memory efficiency when benchmarked against specialized models across a wide range of molecules.[19] An exciting feature of UMA is that automated workflows were used to include an extremely large number of TM complexes in the Open Molecules 2025 (OMOL25) dataset, allowing for extensive sampling of conformer space and configurational flexibility across different DFT datasets and tasks.[19] However, the ranking of conformers has only been evaluated for a subset of the dataset (GEOM) aimed at drug-like molecules.[19,20]

An ideal long-term goal would be to employ MLIPs directly in geometry optimization where both energies and forces are evaluated by the MLIP, bypassing the need for DFT calculations. However, while active research is being devoted to this area, the methodology is not yet stable enough for routine application to complex systems.[21,22] In the present work, therefore, we restrict our focus to assessing how well UMA can reproduce DFT-calculated energies on DFT-optimized geometries for catalytically-relevant organometallic complexes with a particular focus on the correct description of conformational and configurational ensembles.

Specifically, we investigate the ability of UMA to rank the relative stability of TM

complexes with varying conformational and configurational flexibility. Given the high inference speed of UMA, we assess whether the predicted relative stabilities of different configurations of diverse transition metal complexes are accurate enough for practical use cases. To address this question, we evaluate the performance of the smallest UMA model using our previously published datasets containing DFT-optimized geometries and energies for a broad array of TM complexes with varying bisphosphine ligands.

7.2. COMPUTATIONAL METHODS

In this study, the small UMA model (UMA-sm, v1.0) pretrained on the OMol25 dataset with 150M total parameters was used. This model was chosen because we were mainly interested in a single-point energy calculation of structures with less than 1k atoms. Single point energy calculations were performed via the Atomic Simulation Environment (ASE) in Python on a Dell XPS 15 9520 laptop with a 12th generation Intel Core i5 processor, 32GB of RAM and an Nvidia RTX 3050 GPU.

All generated DFT datasets focus on transition-metal based catalysts with bisphosphine ligands (Figure 7.1). Dataset 1 consists of conformers for square-planar Ni-based catalyst structures with 25 varying bisphosphine ligands presented in Chapter 6. Two surrogate "model" structures were used to represent the catalyst, each exhibiting different degrees of conformational flexibility: a rigid [ligand]-Ni(II)-Cl_2 complex (top left), in which Cl_2 creates a symmetric coordination environment, and a more flexible [ligand]-Ni(II)-(CH_3CN)(−pOMe(C_6H_4)) complex (top right), where asymmetry is introduced by the coordination environment which increases conformational freedom.[24] These complexes are viewed as representative models of the pre-catalysts and relevant intermediate in the Ni-catalyzed arylation of nitriles. Dataset 2 contains various octahedral complexes with Ir, Ru or Mn metal centers in combination with 88 bisphosphine ligands relevant for homogeneous hydrogenation catalysis presented in Chapter 5.[23] The studied ligand configurations (bottom) for the Ir, Ru and Mn complexes are named according to the donor atoms of ligands in the axial position.[23] For both datasets the atomic structures of the coordination complexes including all possible stereoisomers were generated with our MACE Python package,[25] followed by exhaustive conformer search. For Dataset 1 this conformer search was performed with CREST at the GFN2-xTB level of theory,[26–28] after which all resulting structures were subjected to density functional theory calculations. For Dataset 2 conformers were generated using RDkit[29] and the lowest-energy conformer per ligand was selected for further analysis. Density functional theory calculations in gas phase were then used to optimize all the resulting structures at the PBE0-D3(BJ)/def2-SVP level of theory.[30–33] Normal mode analysis was carried out to confirm that the optimized geometries correspond to local minima on the potential energy surface. For structures with imaginary frequencies, the PyQRC python package was used to remove these imaginary frequencies and restart geometry optimizations.[34,35]

Our case studies focus on accurately ranking the relative stabilities of either conformers (Dataset 1) or ligand configurations (Dataset 2). This ranking is

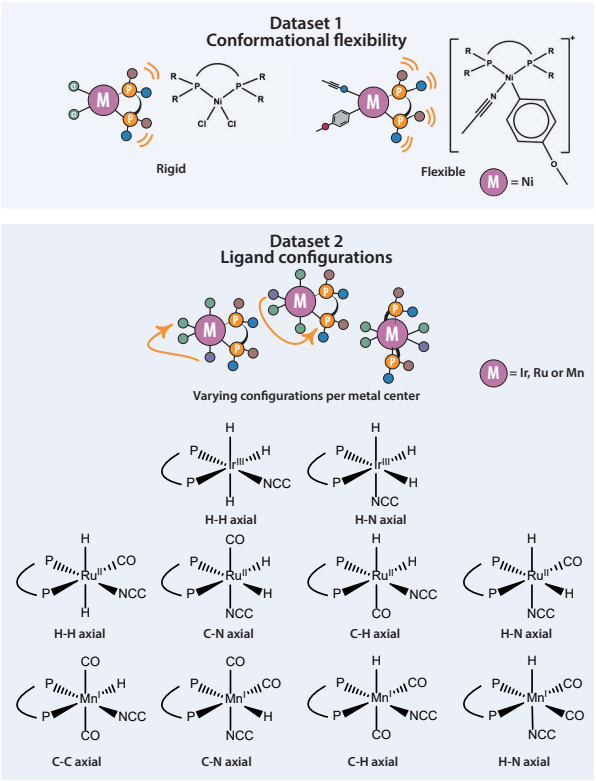


Figure 7.1.: Overview of the two datasets used in this study. Top: Dataset 1 contains conformers of square planar Ni-based rigid and flexible model structures relevant to nitrile arylation, each with one of 25 bisphosphine ligands. Bottom: Dataset 2 includes octahedral Ir, Ru, and Mn complexes relevant to hydrogenation chemistry, each with one of 88 bisphosphine ligands. Details on Dataset 2 are available in our previous work.[23]

performed by calculating the energy of the i-th conformer or configuration relative to a reference structure for the same ligand: $\Delta E_{DFT/UMA} = E_i - E_{\text{ref}}$. All relative energies are evaluated on DFT-optimized geometries to ensure consistency in structural input. To assess how well UMA reproduces DFT-based relative stabilities, conventional statistical metrics are employed, in particular, the pearson correlation coefficient (R^2) and the root-mean-square error (RMSE). The strength of the linear correlation between DFT and UMA energies was quantified using Pearson's correlation coefficient (r), and its square (R^2) is reported as a measure of how well the data follow a linear trend. with values closer to 1 indicating better agreement. The RMSE provides a direct measure of the average deviation between UMA and DFT energies, expressed in kJ/mol.

7.3. RESULTS AND DISCUSSION

Dataset 1 consists of 23 ligands with 2100 conformers of the rigid dichloride model structure and 21 ligands with 3505 conformers of the flexible model structure. For comparative analysis, only ligands for which both the rigid and flexible model structures had fully converged conformer geometries were considered. This filtering resulted in 19 ligands, comprising 746 conformers for the rigid model structure and 1260 conformers for the flexible model structure. A comparison of relative energies

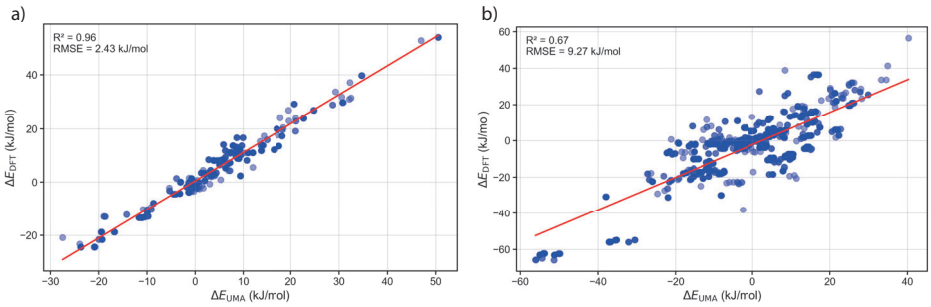


Figure 7.2.: Correlation between DFT and UMA relative single-point energies for Dataset 1 conformers: **(a)** rigid and **(b)** flexible model structures.

computed by DFT and UMA on these structures (Figure 7.2) reveals excellent agreement, particularly for the rigid model's conformers (R^2 = 0.96, RMSE = 2.4 kJ/mol), and reasonable correlation for the flexible model structures (R^2 = 0.67, RMSE = 9.2 kJ/mol). As expected, these results indicate that UMA has difficulties capturing energetic trends when the flexibility of complexes increase. While R^2 and RMSE capture the overall quality of the linear correlation and average energy deviation, they do not directly assess whether the relative ranking of conformers per ligand is preserved, which is an essential criterion for reliable conformational analysis. Therefore, to more directly evaluate the ability of UMA to rank conformers correctly, we also computed Spearmans rank-order correlation coefficient (ρ), a nonparametric metric that measures the monotonic relationship between two variables. In contrast to parametric measures such as Pearsons correlation coefficient, which assumes linearity and normally distributed variables, Spearmans rank-order correlation coefficient (ρ) operates solely on the ranked values. It therefore assesses whether two variables exhibit a monotonic relationship, regardless of the shape of their distributions. To ensure statistical robustness, we considered only ligands containing at least four conformer structures and required that the correlation be statistically significant (p-value < 0.05). In addition, we adopted a threshold of $\rho > 0.6$ to define a "trusted" correlation, reflecting a moderate to strong monotonic agreement between UMA and DFT rankings. This cutoff is chosen to distinguish meaningful performance from weak or inconsistent ordering, as lower ρ values may indicate substantial deviations in the predicted energy ranking. By combining statistical significance with a minimum ranking strength criterion, this approach highlights cases where UMA is reliably predictive of DFT-level ranking of relative

stabilities.

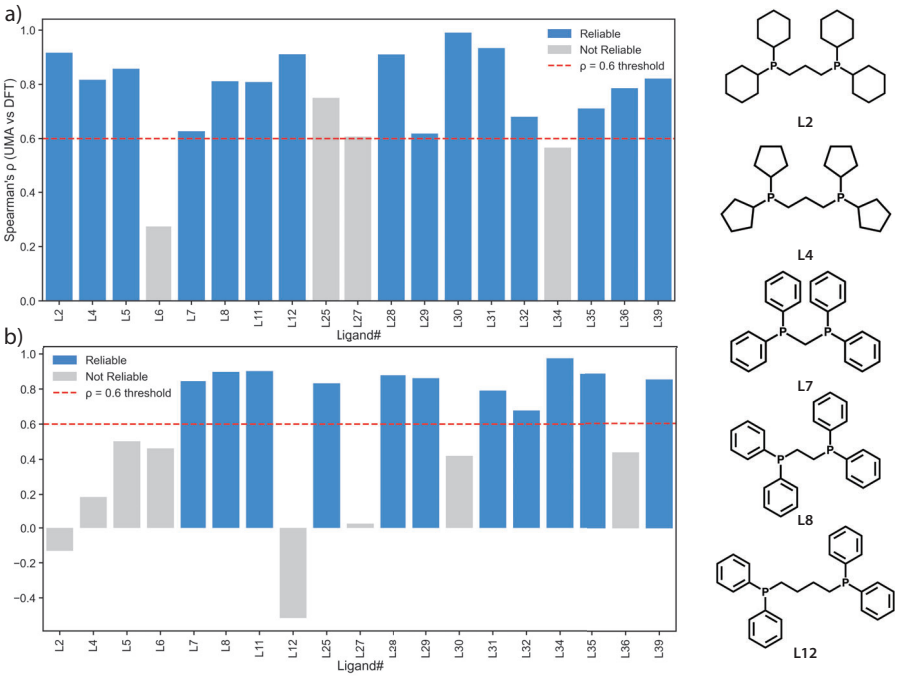


Figure 7.3.: Spearman rank correlation (ρ) between UMA and DFT conformer energy rankings for Dataset 1: **(a)** rigid and **(b)** flexible model structures. Bars are colored by reliability based on $\rho > 0.6$ (blue = reliable, $p < 0.05$ and $\rho > 0.6$; grey = not reliable). The red dashed line indicates the reliability threshold. The right-hand side shows representative bisphosphine ligands L2 (dCyhpp), L4 (Dcyppp), L7 (dppm), L8 (dppe), and L12 (dppb), illustrating differences in phosphorus substituents (phenyl, cyclohexyl, or pentyl) and linker length between donor atoms, which influence conformational flexibility and energy ranking performance.

Our results show that for the rigid model structures, UMA reliably predicts the ranking for 79% of the ligands (Figure 7.3). However, this performance declines when applied to the flexible model structures, where 58% of the ligands exhibit a reliable ranking correlation. Notably, in two cases, a negative Spearman ρ was observed, indicating that UMA predicts the reverse energy ranking compared to DFT.

To illustrate these trends, the right-hand side of Figure 7.3 presents five representative bisphosphine ligands (L2, L4, L7, L8, and L12), which differ primarily in the nature of the substituents on the phosphorus donors (phenyl, cyclohexyl, or pentyl) and in the length of the carbon linker between them. L2 (dCyhpp) and L4 (Dcyppp) are reliably predicted for the rigid model structure but not for the flexible model structure, with L2 even exhibiting a negative ρ. However, only eight conformers were identified for L2 in the flexible model structure, with relative DFT

energies differing by less than 2 kJ/mol. As expected, these are conditions under which UMA struggles to reproduce DFT-level precision. L7 (dppm) and L8 (dppe), which have short linkers of one and two carbons respectively, perform well for both rigid and flexible models, owing to their restricted conformational space. L12 (dppb) performs well in the rigid model structure but yields a negative ρ for the flexible model structure. In this case, only six conformers were found, again with relative DFT energies within 2 kJ/mol.

The tight energy ranges between conformers within Dataset 1 provided a relatively strict test of UMAs ranking capability, as energy differences between conformers are low and often approach chemical accuracy. To examine UMAs performance under conditions where structural and energetic variations are greater, we next turned to Dataset 2, which contains ligand configurations that span a much broader range of relative energies.

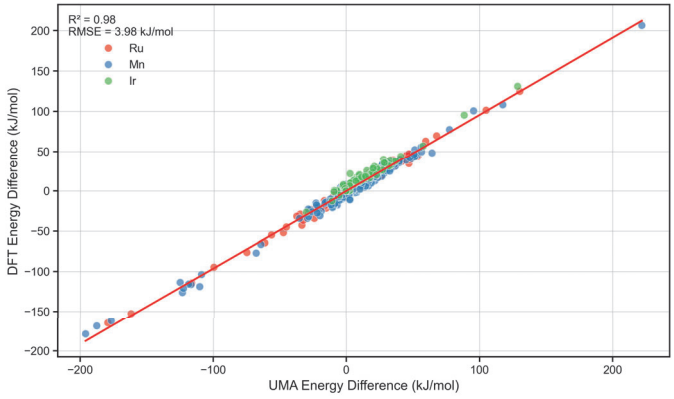


Figure 7.4.: Correlation between UMA and DFT relative energies for Dataset 2 (configurations). Points are colored by metal center: Ir (green), Ru (red), and Mn (blue).

Dataset 2 features 88 chiral bisphosphine ligands coordinated to various transition-metal complexes. After sampling stereoisomerism, a total of 909 geometries were obtained. The metal centers, Ir(III), Ru(II), and Mn(I), are stabilized with different auxiliary ligands, resulting in three distinct classes of complexes: [ligand]-$IrH_3(CH_3CN)$, [ligand]-$RuH_2(CO)(CH_3CN)$, and [ligand]-$MnH(CO)_2(CH_3CN)$. A comparison of relative energies computed by DFT and UMA across this dataset (Figure 7.4) again reveals excellent agreement (R^2 = 0.98, RMSE = 4.0 kJ/mol), highlighting UMAs general applicability across multiple transition metals and ligand environments.

To more closely examine the reliability of energy rankings for individual ligands, we excluded the Ir(III)-based complexes due to the presence of only two configurations per ligand, which is insufficient for rank correlation analysis. For the Mn(I)- and Ru(II)-based complexes, we again applied the significance and strength thresholds (p-value < 0.05, $\rho > 0.6$) to define a "trusted" ranking. Based on this analysis,

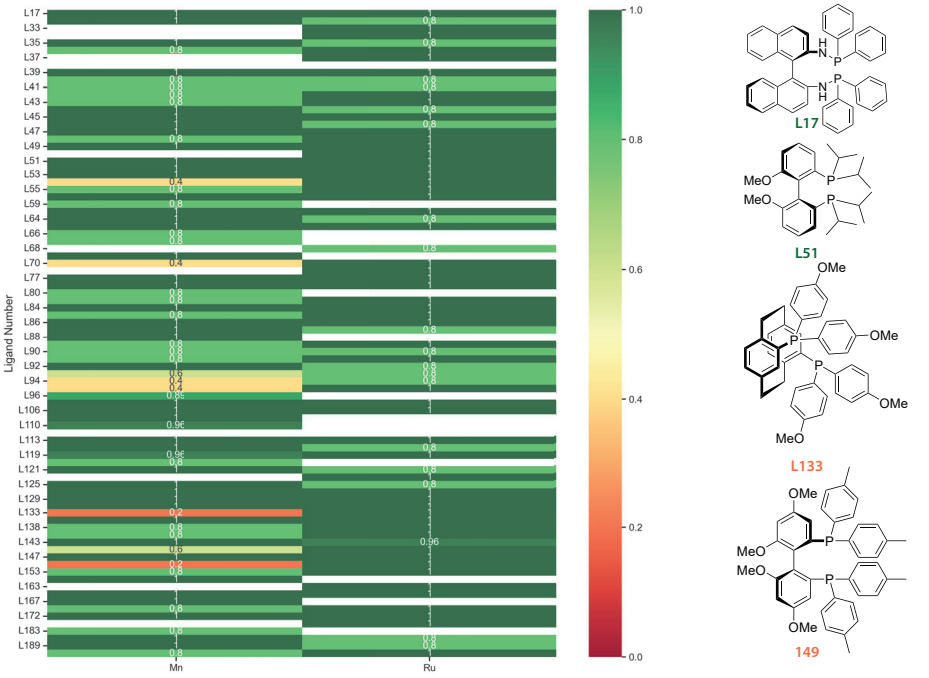


Figure 7.5.: Spearman rank correlation (ρ) between UMA and DFT energy rankings for Dataset 2: Mn(I)- and Ru(II)-based complexes. Cells are colored from red (low ρ) to green (high ρ). Ir(III) complexes are excluded due to insufficient configurations for ranking. The right-hand panels show representative chiral bisphosphine ligands: L17 ((R)-BINAM-P) and L51 ((S)-iPr-BIPHEP) (green) achieve perfect agreement for both metals, while L133 ((R)-An-PhanePhos) and L149 ((R)-Tol-GarPhos) (orange) are reliable for Ru(II) but not for Mn(I).

7

UMA is found to be reliably predictive for 57% of the ligand-metal combinations across the two metal centers (Figure 7.5). A closer examination per metal center reveals that 63% of the ligands are reliably predicted for Ru(II)-based complexes, whereas this drops to 52% for Mn(I)-based complexes. This difference is rooted in fundamental organometallic differences between the metal centers, where Mn(I) complexes, are known to exhibit more fluxional behavior. This can lead to a more complex and shallow potential energy surface, making the energy landscape harder to learn and predict reliably, especially with a general-purpose machine-learned potential such as UMA. In contrast, Ru(II) complexes tend to be more rigid and structurally well-defined under the same ligand field, contributing to the improved energy ranking accuracy observed for this class.

The right-hand side of Figure 7.5 highlights four representative bisphosphine ligands that exemplify these trends. L17 ((R)-BINAM-P) and L51 ((S)-iPr-BIPHEP) are shown in green because they achieve a perfect ρ of 1.0 for both Mn(I) and

Ru(II) complexes. In both cases, the relative energies between configurations are substantial, with L17 showing differences of at least 6 kJ/mol (up to 43 kJ/mol for the HH configuration and 34 kJ/mol for the C-N configuration in Ru complexes) and L51 showing differences of at least 4 kJ/mol (C-N configuration in Mn) and typically exceeding 15 kJ/mol for both metals. These large energy differences appear to be easily predicted by UMA.

In contrast, L133 ((R)-An-PhanePhos) and L149 ((R)-Tol-GarPhos) illustrate cases where UMA performs well for Ru(II) complexes but fails for Mn(I). For L133, Ru complexes display relative energy differences of at least 6 kJ/mol, whereas Mn complexes that feature C-C and C-H configurations differ by less than 2 kJ/mol from the reference. This aligns with our earlier observation made in Chapter 5 that these Mn complexes can exhibit substantial structural isomerism, presenting multiple ligand configurations within 10 kJ/mol.[23] For L149, the Ru C-N configuration is separated from the reference by only 2.3 kJ/mol yet is still ranked correctly, while the Mn C-C configuration differs by just 0.2 kJ/mol and the remaining Mn configurations lie within 5 kJ/mol, leading to inaccurate rankings. However, these examples reinforce the observation that only small energy differences that are near-DFT accuracy pose a challenge for UMA.

7.4. CONCLUSION

Our results demonstrate that UMA represent a significant step forward in the development of MLIPs. With near-DFT accuracy in energy prediction across a wide range of ligand conformers and configurations, UMA enables rapid and scalable analysis in computational chemistry. For example, single-point energy evaluations using UMA can be performed in seconds on consumer-grade GPUs, in stark contrast to the CPU time required for equivalent DFT calculations. This speed-up offers clear advantages in high-throughput screening and early-stage catalyst design workflows. However, the current generation of MLIPs, including UMA, is not without limitations. While they are highly effective at predicting relative energies and forces, they do not yet provide access to electronic structure information such as the electron density, which remains essential for understanding charge distribution, reactivity and spectroscopic properties. In addition, UMA is trained exclusively on gas-phase data and does not currently incorporate solvation effects. Because solvent can substantially influence conformational energetics and catalytic behavior, extending UMA with implicit- or data-driven solvent models represents an important direction for future development. Such advances would broaden the applicability of MLIPs to solution-phase catalysis and other realistic chemical environments.

In Dataset 1, reliable rankings are achieved for 84% of ligands in the rigid [ligand]-Ni(II)-Cl_2 model structures, but this drops to only 53% for the more flexible [ligand]-Ni(II)-(CH_3CN)(−pOMe(C_6H_4)) model structures, where asymmetric coordination and increased conformational freedom pose additional challenges. Ligands such as L7 (dppm) and L8 (dppe) are well-predicted in both rigid and flexible cases due to their short linkers and restricted conformational space, whereas L2 (dCyhpp) and L12 (dppb) are ranked less accurately for the flexible model structures,

often when relative DFT energies between conformers differ by less than 2 kJ/mol. This is a regime where errors in ranking relative stabilities may have limited impact on equilibrium ground-state populations and similar differences in ranking could appear from any DFT methodology by changing basis sets or functionals within the same rung on Jacob's ladder.

A similar pattern is observed in Dataset 2, where UMA reliably ranks configurations for 61% of Ru(II)-based and 44% of Mn(I)-based complexes. Ligands with large relative configuration energy gaps, such as L17 ((R)-BINAM-P) and L51 ((S)-iPr-BIPHEP), achieve near-perfect correlations for both metals, while ligands with small energy separations within the error range of DFT perform worse. These findings underscore that UMA as a general MLIP can be a powerful tools to achieve near-DFT accuracy for both rigid and structurally well-defined systems as well as fluxional and highly flexible complexes.

Finally, it is important to acknowledge the paradigm shift that MLIPs introduce. Although DFT itself is an approximation, the behavior and limitations of exchange-correlation functionals, basis sets, and dispersion corrections have been extensively studied and understood over decades. These methods are rooted in physics. In contrast, general-purpose ML models such as UMA are trained as black boxes on vast datasets, and their performance is not easily interpretable. This abstraction risks concealing underlying failures if used uncritically. As the field transitions into a post-UMA era, it will be crucial to combine the speed of ML with validation and domain expertise to ensure robust and meaningful chemical insights. Importantly, the instances where UMA fails to reproduce DFT rankings occur predominantly in near-degenerate energy regimes where DFT itself cannot provide a uniquely reliable ranking. This exemplifies that expert judgment remains essential when interpreting such cases, even as MLIPs like UMA enable rapid and accurate exploration of broader regions of chemical space.

DATA AVAILABILITY

The datasets, an overview of ligands and the code for reproducing the analysis presented in this study are available with an extensive readme via 4TU.ResearchData at https://doi.org/10.4121/14bcdfc0-dd25-4945-9cb9-5d862b47a784.

CONTRIBUTIONS

A.V. Kalikadien: Investigation, Methodology, Conceptualization, Software, Validation, Data Curation, Formal analysis, Visualization, Writing - Original Draft, Writing - Review & Editing, Project administration **E.A. Pidko:** Supervision, Conceptualization, Resources, Funding acquisition, Writing - Review & Editing, Project administration

REFERENCES

(1) Kalikadien, A. V.; Pidko, E. A. *J. Phys. Chem. A.* **2026**, *accepted.*

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

(3) Kalikadien, A. V.; Mirza, A.; Hossaini, A. N.; Sreenithya, A.; Pidko, E. A. *ChemPlusChem* **2024**, e202300702.

(4) Foscato, M.; Jensen, V. R. *ACS Catal.* **2020**, *10*, 2354–2377.

(5) Butera, V. *Phys. Chem. Chem. Phys.* **2024**, *26*, 7950–7970.

(6) Bellonzi, N.; Kunitsa, A.; Cantin, J. T.; Campos-Gonzalez-Angulo, J. A.; Radin, M. D.; Zhou, Y.; Johnson, P. D.; Martínez-Martínez, L. A.; Jangrouei, M. R.; Brahmachari, A. S.; Wang, L.; Patel, S.; Kodrycka, M.; Loaiza, I.; Lang, R. A.; Aspuru-Guzik, A.; Izmaylov, A. F.; Fontalvo, J. R.; Cao, Y. *arXiv preprint* **2024**.

(7) Matsuoka, W.; Harabuchi, Y.; Maeda, S. *ACS Catal.* **2022**, *12*, 3752–3766.

(8) Gensch, T.; Gomes, G. D. P.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S.; Aspuru-Guzik, A. *J. Am. Chem. Soc.* **2022**, *144*, 1205–1217.

(9) Pidko, E. A. *ACS Catal.* **2017**, *7*, 4230–4234.

(10) Behler, J. *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930–17955.

(11) Behler, J. *Int. J. Quantum Chem.* **2015**, *115*, 1032–1050.

(12) Kocer, E.; Ko, T. W.; Behler, J. *Annu. Rev. Phys. Chem.* **2022**, *73*, 163–186.

(13) Eyert, V.; Wormald, J.; Curtin, W. A.; Wimmer, E. *J. Mater. Res.* **2023**, *38*, 5079–5094.

(14) Handley, C. M.; Popelier, P. L. A. *J. Phys. Chem. A* **2010**, *114*, 3371–3383.

(15) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. *Chem. Rev.* **2021**, *121*, 10073–10141.

(16) Li, Y.; Zhang, X.; Liu, M.; Shen, L. *J. Mater. Inform.* **2025**, *5*.

(17) Kulichenko, M.; Nebgen, B.; Lubbers, N.; Smith, J. S.; Barros, K.; Allen, A. E. A.; Habib, A.; Shinkle, E.; Fedik, N.; Li, Y. W.; Messerly, R. A.; Tretiak, S. *Chem. Rev.* **2024**, *124*, 13681–13714.

(18) Levine, D. S.; Shuaibi, M.; Spotte-Smith, E. W. C.; Taylor, M. G.; Hasyim, M. R.; Michel, K.; Batatia, I.; Csányi, G.; Dzamba, M.; Eastman, P.; Frey, N. C.; Fu, X.; Gharakhanyan, V.; Krishnapriyan, A. S.; Rackers, J. A.; Raja, S.; Rizvi, A.; Rosen, A. S.; Ulissi, Z.; Vargas, S.; Zitnick, C. L.; Blau, S. M.; Wood, B. M. *arXiv preprint* **2025**.

(19) Wood, B. M.; Dzamba, M.; Fu, X.; Gao, M.; Shuaibi, M.; Barroso-Luque, L.; Abdelmaqsoud, K.; Gharakhanyan, V.; Kitchin, J. R.; Levine, D. S.; Michel, K.; Sriram, A.; Cohen, T.; Das, A.; Rizvi, A.; Sahoo, S. J.; Ulissi, Z. W.; Zitnick, C. L. *arXiv preprint* **2025**.

(20) Axelrod, S.; Gómez-Bombarelli, R. *Sci. Data.* **2022**, *9*, 185.

(21) Wu, Z.; Zhou, L.; Hou, P.; Liu, Y.; Guo, T.; Liu, J.-C. *ChemRxiv preprint* **2023**.

(22) Goldsmith, B. R.; Esterhuizen, J.; Liu, J.-X.; Bartel, C. J.; Sutton, C. *AIChE J.* **2018**, *64*, 2311–2323.

(23) Kalikadien, A. V.; van der Lem, N. J.; Valsecchi, C.; Lefort, L.; Pidko, E. A. *Digit. Discov.* **2025**, *4*, 2033–2044.

(24) Baidun, M. S.; Kalikadien, A. V.; Lefort, L.; Pidko, E. A. *J. Phys. Chem. C* **2024**, *128*, 7987–7998.

(25) Chernyshov, I. Y.; Pidko, E. A. *J. Chem. Theory Comput.* **2024**, *20*, 2313–2320.

(26) Pracht, P.; Bohle, F.; Grimme, S. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(27) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. *WIREs Comput. Mol. Sci.* **2021**, *11*, e1493.

(28) Grimme, S. *J. Chem. Theory Comput.* **2019**, *15*, 2847–2862.

(29) Landrum, G. RDKit: Open-source cheminformatics, 2020.

(30) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16 Revision C.01/C.02, 2016.

(31) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(32) Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. *Chem. Phys.* **2019**, *150*, 154122.

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

(34) Silva, M. A.; Goodman, J. M. *Tetrahedron Lett.* **2005**, *46*, 2067–2069.

(35) Goodman, J. M.; Silva, M. A. *Tetrahedron Lett.* **2003**, *44*, 8233–8236.

8

Outlook

Whenever I mention that I am a PhD candidate at TU Delft, the inevitable next question is: *"Oh, interesting! What is your research about?"* This seemingly simple inquiry often requires a careful balance in providing an answer that is both accurate and accessible to the inquirer, particularly since this work spans multiple disciplines. At its core, this research focuses on catalysis, a field fundamental to the production of chemical products, yet it is deeply intertwined with computational chemistry, data science, and automation. Explaining this intersection reveals a broader challenge: as chemistry becomes increasingly digital, effective communication and the development of knowledge that bridges disciplinary boundaries are more important than ever.

AI-driven materials design has received significant momentum in recent years, with major technology companies such as Google, Microsoft, and Meta using it as a testbed for new algorithms and hardware. The research presented in this dissertation contributes to a specific subset of this domain: *in silico* methods for the rational design of transition-metal-based catalysts. By leveraging bias-free, data-driven approaches, ML algorithms can support the optimization and design of new catalysts while remaining interpretable to chemists. However, as demonstrated throughout the preceding chapters, this field remains in constant development. In this Chapter, we look beyond the specific work described in this dissertation to reflect upon broader developments in the field and their societal relevance. A recurring theme that emerges from this perspective is interdisciplinarity: should chemists acquire data science skills, or should data scientists immerse themselves in chemistry?

As datasets grow in both size and complexity, this question becomes increasingly pressing. Large-scale data collection no longer permits manual inspection for quality control, and both experimental and computational workflows require greater levels of automation in data generation, engineering, and analysis. While this may sound straightforward, the challenges are illustrated by a case study explored during this project. Specifically, we required a system to automatically store molecular descriptors generated by the OBeLiX workflow in a manner that would allow data accumulation and facilitate downstream tasks, such as dashboard-based analysis. The most future-proof and low-maintenance option was to employ Amazon Web Services (AWS). Yet this seemingly simple decision quickly expanded into a much

larger design exercise, raising questions that had not been considered during data generation. Should descriptor names be standardized? Should each dataset generated by OBeLiX carry a unique identifier? Which metadata are not automatically included and therefore require separate processing? For those familiar with cloud architecture, the implemented solution ultimately connected two AWS S3 buckets (one for automatically generated data and one for manually curated metadata) with Lambda functions that wrote all inputs into a standardized DynamoDB table, while notifying the workflow owner via email upon submission. All this effort, just for the purpose of storing descriptors, highlights the complexity underlying what first appeared to be a simple task. The case underscores a broader point: achieving interoperable, reusable computational datasets, even when generated by the same workflow, demands foresight and expertise spanning multiple computer science disciplines.

Open Science, much like the design of interoperable data systems, is an endothermic exercise: it invariably requires an input of energy from the researcher. While the ultimate goal of most scientists is for their work to be used, reused, and built upon, realizing this vision demands effort beyond the immediate research itself. Writing clear documentation, developing reusable code, and depositing data into accessible repositories all require both time and technical expertise. Relying solely on intrinsic motivation is unlikely to guarantee widespread adoption of such practices. Yet if research outputs are not usable by others, the openness of science risks remaining nominal rather than functional, and the collective progress of a field is hindered. Despite these challenges, Open Science reflects how science was always meant to operate: as a cumulative and collaborative enterprise. By investing additional effort to make data, methods, and code accessible, researchers enable reproducibility, accelerate discovery, and extend the societal relevance of their work. The additional burden is undeniable, but so too is the long-term benefit to both the scientific community and the broader public. This balance between effort and reward is also mirrored in current debates around AI. Like Open Science, AI promises to accelerate discovery and reshape how research is conducted, but it also carries the risk of misplaced expectations and unsustainable hype if its limitations are not carefully acknowledged.

Following the release of OpenAI's GPT-3, developments in artificial intelligence have advanced at a pace that mirrors the trajectory of an economic bubble. Each week brings models of increasing complexity, yet the fundamental questions remain: what purpose do these models serve, and when does it make more sense to perform chemical experiments guided by serendipity rather than relying on opaque algorithms whose internal reasoning cannot be fully understood? Even within state-of-the-art high-throughput experimentation laboratories, chemical data are subject to error and noise, and thus should not be treated as infallible. As with most technological hypes, it is crucial to look beyond the horizon and ask which applications will remain meaningful once the bubble inevitably bursts. The grand challenges we face as a society will not be solved through short-lived bursts of hype, but through sustained, incremental progress that integrates technology with domain expertise.

Ultimately, human intelligence is not a product of scale or computational power, but of genuine curiosity. As Antoni van Leeuwenhoek once reflected (translated from Dutch): *"My work, which I have long done, was not pursued to gain the praise I now enjoy, but chiefly from a desire for knowledge, which I find remarkably more in me than in most other men."* It is this pursuit of curiosity that grants academia its societal relevance, not the accumulation of citations, grants, or institutional rankings. Academia's task is to advance knowledge, and in doing so it must continually reinvent itself to remain trusted and relevant to society. Maintaining this relevance requires perseverance as much as inspiration: success in science is always a combination of grit and luck. However, a scientist's true impact lies in their genuine engagement with a small niche of knowledge that, perhaps by sheer luck, contributes to the broader progress of humankind. That is what science is about, and that is how it should present itself to the public.

A

THE COMPROMISE BETWEEN SPEED AND ACCURACY FOR QM METHODS IN HIGH-THROUGHPUT SCREENING OF TM COMPLEXES

This Appendix presents a benchmarking study of DFT methods, evaluating both basis sets and exchange-correlation functionals for geometry optimization of the 192 Rh-NBD catalyst structures introduced in Chapter 2. The reference level of theory, PBE0/def2-SV(P), employed in Chapter 2, provides the foundation for comparison. Performance was assessed in terms of computational efficiency (time to convergence) and the statistical distributions of selected steric and electronic descriptors, including bite angle, buried volume, and HOMO-LUMO gap. For the basis set benchmark, def2-SV(P) was compared against two larger members of the Ahlrichs family, def2-TZVPP and def2-QZVPP. These basis sets were selected because they provide consistent accuracy across nearly the entire periodic table up to radon.[1] Among them, def2-SV(P) is the least extensive and def2-QZVPP the most extensive, making this comparison a direct test of the trade-off between computational cost and accuracy. For the benchmark of functionals, five functionals spanning different rungs of Jacobs ladder were considered. At the GGA level, PBE was chosen due to its close relation to PBE0.[2,3] TPSS was selected as a representative meta-GGA,[4] while B3LYP,[5] one of the most widely applied functionals in computational chemistry, was included as a canonical hybrid functional. PBE0 served as the reference hybrid functional, and MN15 was incorporated as a modern hybrid specifically parametrized for TM complexes.[6]

A.1. COMPUTATIONAL METHODS

Geometries were optimized via DFT, performed using Gaussian 16 C.02[7] on the Dutch national supercomputer Snellius. Calculations were carried out in the gas

phase with an ultrafine integration grid, and natural bond orbital (NBO) analyses were included for all structures.

To assess the consistency of the computed descriptors across different methods, two types of statistical analyses were performed. First, the Pearson correlation coefficient (R^2) was calculated to quantify the degree of linear correlation between descriptor values obtained with different methods. Second, an analysis of variance (ANOVA) statistical test was used to assess whether descriptor distributions were statistically distinguishable. The ANOVA test operates under the null hypothesis that two groups are drawn from the same distribution; a p-value below 0.05 was taken as evidence to reject this null hypothesis.

All statistical analyses were performed in Python using the scipy package.[8] For the ANOVA comparisons, descriptor values obtained with each functional or basis set were tested against those obtained with the reference method, PBE0/def2-SV(P).

A.2. BASIS SETS

To evaluate the effect of different basis sets on geometry optimization, the functional was kept constant. All calculations employed the PBE0 functional with Grimmes D3-BJ dispersion correction.[9] The following three basis sets from the Ahlrichs family were benchmarked:

- def2-SV(P): split valence, polarization on heavy atoms
- def2-TZVPP: triple zeta valence, two sets of polarization functions
- def2-QZVPP: quadruple zeta valence, two sets of polarization functions

Def2-SV(P) is the standard basis set commonly used within our research and was therefore included as the reference. The other two represent systematic expansions, with additional functions added per atomic orbital. All 192 RhNBD catalyst structures were optimized with each basis set. On Snellius, the maximum allowed wall time per job is five days, hence jobs exceeding this limit were restarted from checkpoint files containing the geometry and wavefunction.

The first comparison concerns the computational cost, measured in CPU hours required for geometry optimization. Table A.1 reports the wall time for the fastest converged structure with each basis set.

Functional/Basis set	CPU hours (fastest optimization)
PBE0-D3BJ/def2-SV(P)	5
PBE0-D3BJ/def2-TZVPP	175
PBE0-D3BJ/def2-QZVPP	9640

Table A.1.: CPU hours required for the fastest optimization with each basis set. Times are rounded up to the nearest whole hour.

As expected, computational time increased dramatically with basis set size: def2-SV(P) required the least time, while def2-QZVPP was prohibitively expensive.

After approximately 10,240 CPU hours, only two structures had converged with def2-QZVPP, leading to the decision to terminate further calculations with this basis set. For def2-TZVPP, a maximum of 4600 CPU hours yielded 88 converged structures. These 88 structures were used as the comparison set against def2-SV(P), for which all optimizations converged within 100 CPU hours. Based on time-to-convergence alone, def2-SV(P) already appears more practical for high-throughput applications.

In further comparison, descriptor values were compared between def2-SV(P) and def2-TZVPP. Violin plots in Figure A.1 summarize the distributions of the bite angle, buried volume (3.5 Å radius), and HOMO-LUMO gap across the 88 structures. The

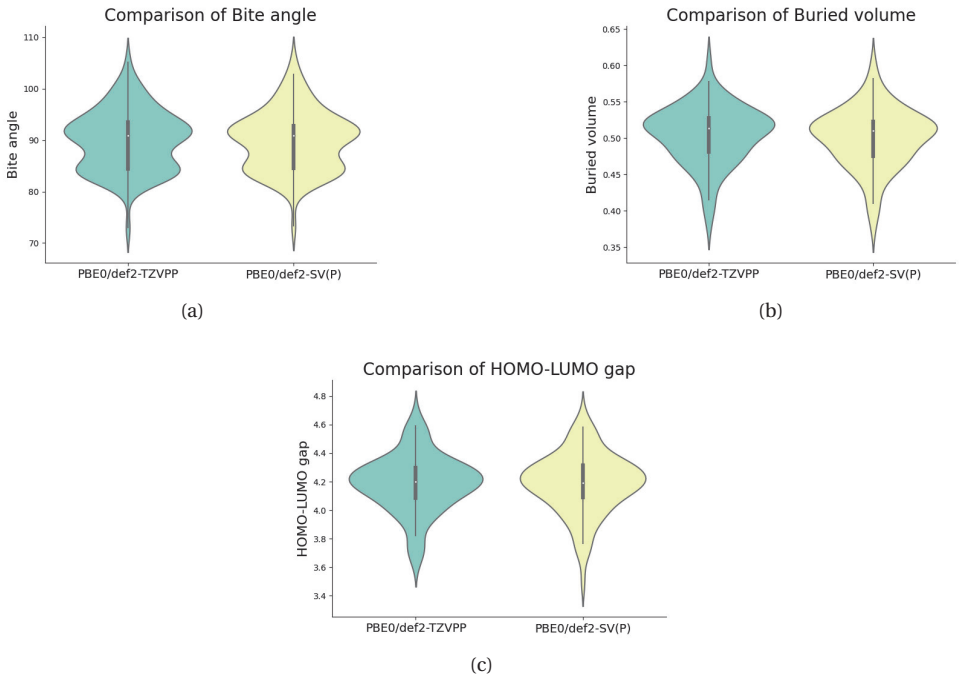


Figure A.1.: Violin plots comparing the 88 ligands optimized with def2-SV(P) and def2-TZVPP: (a) bite angle (°), (b) percentage buried volume with a radius of 3.5 Å(%), and (c) HOMO-LUMO gap (eV).

results indicate near-identical distributions for both basis sets. For bite angle, the correlation coefficient was $R^2 = 0.99$, with only a few unexplained outliers. The ANOVA test yielded a p-value of 0.97, confirming no significant statistical difference. For buried volume, $R^2 = 0.98$, and the ANOVA p-value was 0.52, suggesting minor variation but no statistically significant difference. For the HOMO-LUMO gap, $R^2 = 0.96$ and $p = 0.97$, again showing excellent agreement between basis sets.

In summary, the results demonstrate that def2-SV(P) provides nearly identical descriptor values compared to def2-TZVPP, while requiring substantially less computational effort. Therefore, def2-SV(P) is the superior choice for geometry

optimization in high-throughput screening of TM complexes.

A.3. FUNCTIONALS

Having established def2-SV(P) as the most efficient basis set for geometry optimizations, the next step was to assess the effect of different exchangecorrelation functionals. Geometry optimizations were performed for all 192 ligands with the def2-SV(P) basis set using the five functionals listed in Table A.2, followed by a comparison of computational cost and descriptors.

Functional	Dispersion	Level of theory	Degree of parameterization
PBE	D3-BJ	GGA	Low
TPSS	D3-BJ	mGGA	Low
B3LYP	D3-BJ	Hybrid	Medium
MN15	Intrinsic	Hybrid	Very high
PBE0	D3-BJ	Hybrid	Low

Table A.2.: Overview of the functionals benchmarked in this study, including dispersion correction, theoretical level, and degree of parameterization.

All functionals were used with Grimmes D3-BJ dispersion correction, except for MN15, which incorporates an intrinsic correction. As shown, the set covers generalized gradient approximation (GGA), meta-GGA (mGGA), and hybrid functionals, thereby spanning a representative range of commonly applied levels of theory in TM-based catalysis.

The CPU time requirements varied substantially across functionals (Figure A.2). MN15 was by far the slowest, consistent with its hybrid nature and high degree of parameterization. In contrast, PBE required the least time, as expected from its lower theoretical level.

Table A.3 provides a more quantitative assessment. PBE and TPSS were the fastest, followed by PBE0 and B3LYP, with MN15 the slowest. Ordering the functionals by efficiency yields:

1. PBE
2. TPSS
3. PBE0
4. B3LYP
5. MN15

This ranking aligns with expectations: hybrids are more expensive than GGA/mGGA, and within the hybrids, cost increases with parameterization (PBE0 < B3LYP < MN15).

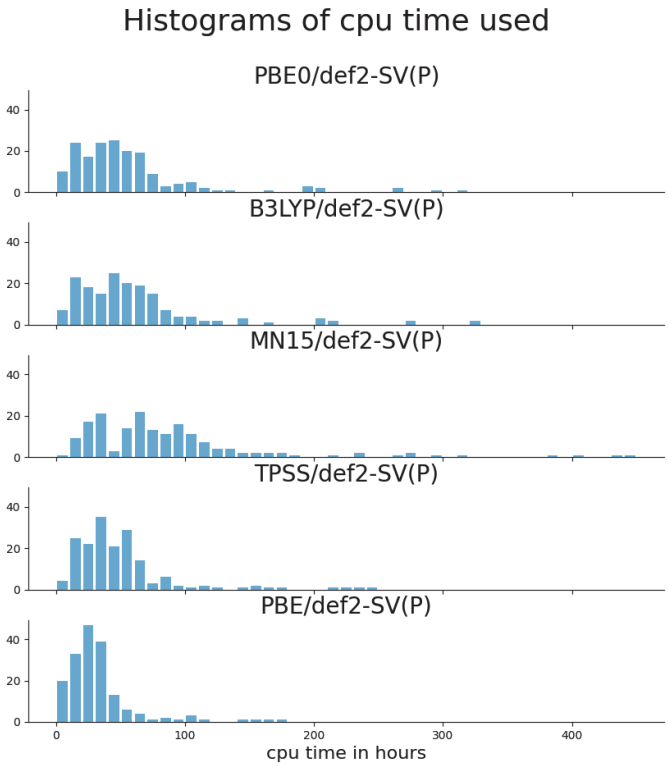


Figure A.2.: Histogram of CPU time required per functional to optimize all ligands (bin size = 10 CPU hours).

Functional	% completed after 100 CPU h	CPU h to reach 90% convergence
PBE0	89	105.5
B3LYP	87	113.1
MN15	72	152.7
TPSS	93	88.6
PBE	95	58.9

Table A.3.: Performance comparison of functionals based on optimization times: percentage of converged optimizations after 100 CPU hours, and total time required to reach 90% convergence.

To determine whether the choice of functional affects optimized structures, key descriptors were compared: bite angle, buried volume, and HOMO-LUMO gap. Violin plots in Figure A.3 illustrate the distributions across all functionals. For the bite angle, all functionals yielded nearly identical results. Correlation coefficients

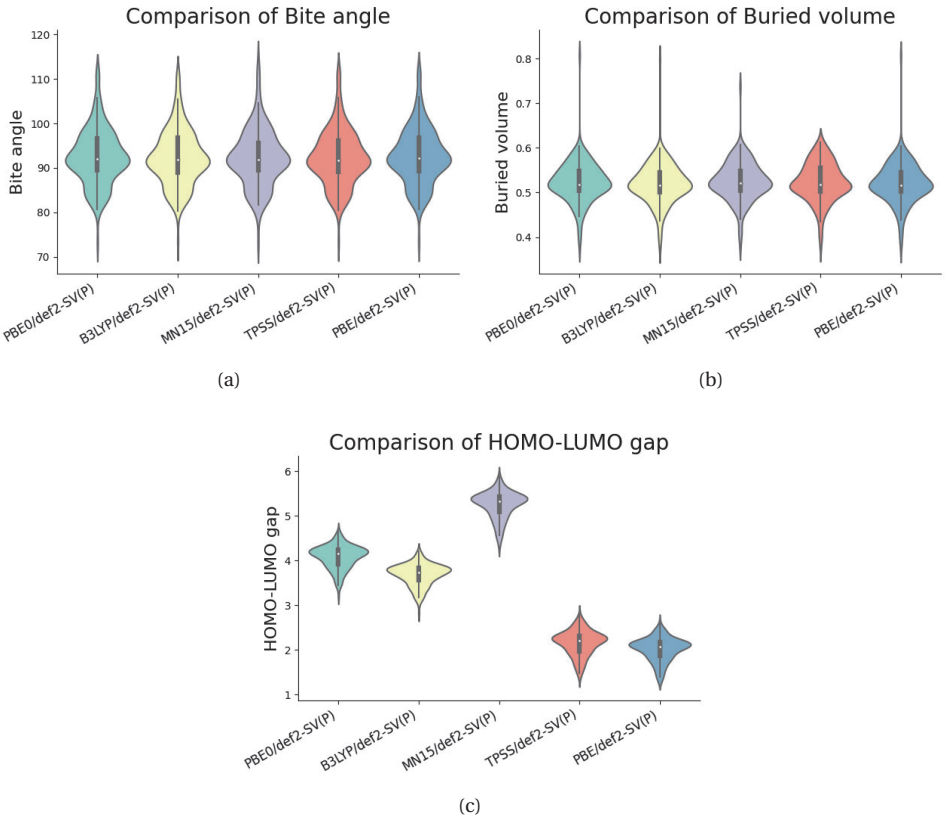


Figure A.3.: Violin plots comparing descriptor distributions across functionals: (a) bite angle (°), (b) percentage buried volume with a radius of 3.5 Å(%), and (c) HOMO-LUMO gap (eV).

were extremely high (R^2 = 0.97-0.99) with p-values between 0.72 and 0.88, confirming no significant differences.

Buried volume values were also highly consistent, with the exception of a single outlier structure (L103), in which norbornadiene dissociated in PBE0, B3LYP, and MN15 but not in TPSS (Figure A.4). This event caused a spike in buried volume but did not affect the overall agreement, as correlations remained high (R^2 = 0.83-0.98) and ANOVA tests confirmed no statistically significant differences. In contrast, the HOMO-LUMO gap showed significant dependence on the functional (Figure A.3c). This outcome is consistent with literature reports, e.g., PBE0 systematically producing larger gaps than B3LYP.[10] Strong correlations were observed among hybrid functionals (R^2 = 0.83-0.88), while correlations with GGA/mGGA functionals were weak ($R^2 \approx 0.5$). ANOVA confirmed that all functional pairs produced statistically distinct HOMO-LUMO gap distributions.

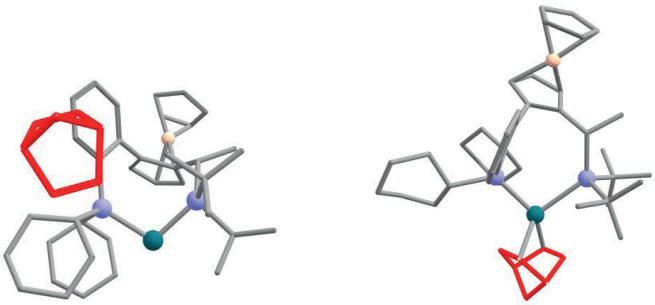


Figure A.4.: Outlier structure L103: (a) dissociation observed with PBE0, (b) intact binding with TPSS.

Overall, these results show that the tested geometric and steric descriptors (bite angle, buried volume) are largely insensitive to the functional, whereas the tested electronic descriptor (HOMO-LUMO gap) is highly sensitive. This implies that structural optimizations can reliably be performed with a computationally efficient functional, provided electronic properties are subsequently refined at a higher level of theory.

Among the hybrid functionals, PBE0 strikes the best balance between accuracy and efficiency. It performs faster than B3LYP and substantially faster than MN15, while providing more reliable electronic properties than GGA or mGGA functionals. For this reason, PBE0/def2-SV(P) was chosen for geometry optimization throughout this dissertation, followed by single-point energy calculations at the PBE0/def2-TZVPP level to ensure the highest feasible quality of electronic descriptors. This protocol ensures both computational efficiency and reliable accuracy for TM complexes.

REFERENCES

(1) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(2) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.

(3) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(4) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(5) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Chem. Phys.* **1994**, *98*, 11623–11627.

(6) Yu, H. S.; He, X.; Li, S. L.; Truhlar, D. G. *Chem. Sci.* **2016**, *7*, 5032–5051.

(7) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16 Revision C.01/C.02, 2016.

(8) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat,.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; SciPy 1.0 Contributors *Nat. Methods* **2020**, *17*, 261–272.

(9) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456–65.

(10) Novikov, N.; Maslov, M.; Katin, K.; Prudkovskiy, V. *Lett. Mater.* **2017**, *7*, 433–436.

ACKNOWLEDGMENTS

When I started this journey, I had no idea what lay ahead. My life has always revolved around walking new paths, whether shaped by my immigrant background or by the studies I pursued. I would not be where I am today if I had listened to every opinion offered about my career or my choices in life. Before you lies the result of hard work, dedication, and a passion for science. Of course, I am not only referring to myself: this journey would not have been possible without the support of the brilliant people I had the privilege to meet along the way. This dissertation was shaped by collaborations, scientific discussions, and countless social moments that made the past years so memorable. A heartfelt thank you to everyone who contributed in any way. Know that I truly value each and every one of you, even those not mentioned by name in this acknowledgment section.

Firstly, I would like to start by thanking my promotor **Evgeny Pidko**. I had never heard of the world of academic research, but during my master's thesis he enabled me to explore this world as much as I wanted to. He saw something in me during my master's thesis at his group and offered me a PhD position. Evgeny, I clearly remember how you offered me to join your research group for a fun and simple project where we will calculate some catalyst features and use them for machine learning. It was certainly not simple, but I am happy to have had you as my supervisor through it. Your support and trust allowed me to develop myself and the project in ways that I had not considered possible. I will never forget the random conversations that were started by you entering our office, which often led to a fiery scientific discussion. I am grateful for all of these discussions on scientific, philoshopical, political, professional, and personal topics. I could not have imagined a better promotor, supervisor, mentor and friend to guide me through the past four years, thank you for this wonderful experience.

To the team from Janssen: **Laurent Lefort**, **Robbert van Putten**, **Cecile Valsecchi**,**Mikko Muuronen**, **Tor Maes**, **Natalia Dyubankova** and (a bit later) **Francesco Pedrazzi**, thank you for the wonderful collaborations. A special thanks goes out to you **Laurent**. This project would not have existed in its current form without your experimental efforts, funding acquisition and supervisory role. I am endlessly grateful for your involvement. To this day I cannot believe that I got to work with the chemistry version of the Marvel's Avengers team. I clearly remember our project kickoff meeting at Janssen in Beerse. It was in the first week of my PhD project (5th of October 2021) and was supposed to last from 13:00 to 15:00. If memory serves me right, we finished somewhere around 17:30. This set the precedent for the many meetings that followed. The scientific discussions were always intense and stimulating to the point that everyone became enthusiastic about this research. I remember being nervous about what was to come when driving towards the kickoff meeting. Who are these people? Why would these renowned scientists care about what I have to say? Fortunately, now I know that all these doubts were completely unjustified. I could not have wished for a better industrial collaboration; I had

complete ownership over the project and everyone was very engaged and interested in my results. The individual regular discussions that I had with some of you really pulled me through this PhD. Thank you all so much, it has been an absolute privilege to work with you.

I would like to thank the external and reserve committee members: **Rienk Eelkema**, **Ferdinand Grozema**, **Bernd Ensing**, **Maren Podewitz** and **Andrea Ramirez-Ramirez**. Thank you so much for taking the time from your busy schedules to read this dissertation, give feedback and participate in the defense ceremony. I am delighted to be in the presence of such a diverse external committee and look forward to the scientific discussion.

Dear BSc. and MSc. thesis students, this dissertation would not have been possible without you. It is my name that is on the front of this dissertation, but in reality the main contributors to this work were: **Mark Heezen**, **Britt van Dongen**, **Niels van der Lem** (twice), **Adrian Mirza**, **Aydin Najl Hossaini**, **Margareth Baidun**, **Imme Schoot Uiterkamp**, **Joyce Sweere**, **Jan de Korte**, **Sára Finta** and **Tai Hong Chow**. Each of you was special in their own way and contributed in your own special way. Over the years, I asked each of you to take a picture with me during or after your project, which I promised to put into a collection in the acknowledgment section of my dissertation. I could honestly dedicate a book to our collaborations, but a picture says more than a thousand words (see below). I learned a lot from working with you and had a lot of fun. Good luck with all future endeavors, I know you will all do great.

Current and former group members/students of the ISE group, thank you for the support, collaborations and many moments of laughter we could share. **Nithya** and **Sasha**, you guys were amazing to work with. I thoroughly enjoyed our (scientific) discussions in the office; they made my time at TU Delft memorable. The international lunches and housewarming dinners were also a great hit, thank you for teaching me about your cultures. **Evgeny Jr.**, I feel like I missed out on the opportunity to do really cool experiments under your mentorship in the lab. Nevertheless, thank you for your critical thoughts on my computational results and the conversations we could enjoy over a coffee. **Liliana**, we did not get to interact much in a professional setting (since again, I barely entered our labs), but I am very grateful for the interactions we had during group activities or when we randomly bumped into each other in the hallway. **Els**, I do not know where I would be without your administrative support. Thank you for taking care of the many questions and requests that I had, also related to my PhD council stuff. Finally, I have to acknowledge the new generation of PhD candidates that are now taking a seat in the office where I worked for the past years. **Sumeia**, even though we only got to sit next to each other for a short time, I had a lot of fun. Thank you for having a sense of humor that is just as messed up as mine (Muhammad Sumbul) and thank you for all the discussions we had. I wish you good luck in your journey, but I know for sure that with your level of discipline, determination, and creativity you will do well. **Margareth**, we have discussed this before, but you have been with me through my PhD for quite a while. As a master student, writing a paper for your research internship (during my 2nd year), and afterwards as a PhD candidate yourself (during my 3rd and 4th year). I could not have asked for a better colleague and friend, it was wonderful to work alongside you and collaborate with you. I love seeing the developments that you have been through over the years, and I know for sure that you have all the skills to make your PhD project a huge success. I will greatly miss sitting across of you and being met with the occasional 'psst, I have a question for you' or your messed up posture leading to your feet somehow reaching all the way to my damn side of the office. Thank you for everything; I hope you can finally enjoy being the favorite PhD candidate of the group.

Speaking of favorites, thank you for being my favorite weirdo, **ass. prof. Georgy Filonenko**. I am not joking when I say that you are genuinely the weirdest person I have ever met in my life, but that makes you ever so fun and memorable. You helped me design the very first front cover I have ever made in Adobe Illustrator, which activated my interest in learning how to make creative illustrations myself. On many occasions at the start of my PhD I came to your lab to ask chemistry questions and you always answered them in a motivating and constructive way. Thank you for the many occasions where you pulled me from behind my computer to show me cool things in the lab and thank you for custom-building the Nerf foam blasters with me. Those friday evenings where we would nerd out with 3D printed pieces and metal barrels will forever remain in my memory. You are an amazing person and I hope you will find true happiness in life, you deserve it.

Support staff and members of the Digital Competence Center and data stewardship at TU Delft that I got to work with: **Esther Plomp**, **Maurits Kok**, **Niket Agrawal**, and **Selin Kubilay**. Esther and Maurits, thank you for being great mentors during the Open Life Science program. I learned a lot from you and you elevated my software engineering

skills to a new level. Niket and Selin, thank you for the great support in the revamping of OBeLiX and inventing a storage solution in AWS. I remember that we spent a large time of the project just to understand what OBeLiX did, what needed to be tested, and what questions we should ask from a user perspective. The hand-drawn block diagram to guide our discussions was super funny and I still have it stored in case I need to explain how OBeLiX works on a technical level. It was awesome to have you guys around and I greatly enjoyed working on the project with you and all the chats we had.

My fellow PhD council members, the most fun sidequests during my PhD found their origin here. Starting with the department PhD council: **Anand Raja**, **Jasmeen Nespoli**, **Christel Koopmans**, **Dominik Goldstein**, **Henri Pelzer**, **Sophie de Boer**, **Sven Weerdenburg** and **Kalani Ostermeijer**, thank you guys so much for the fun times. We did some great things for the department and had a lot of fun in the process. The running jokes such as the Pelzer formula for ordering snacks or the 100% representation of the ISE section due to me being the only PhD candidate will forever be etched in my brain. I thoroughly enjoyed our meetings on Tuesday afternoon and the events we managed to come up with. I wish you all good luck with the remainders of your PhD projects and hope to hear great things about you in the future.

The Applied Sciences faculty's PhD council: **Arent Kievits**, **Davide Costa**, **Sercan Deve**, **Pierfrancesco Ombrini**, **Siddharth Singh**, **Hector Maldonado**, **Chris Soukaras**, **Koushik Sreenivasa**, **Miriam Cammaert** and later **Kalani Ostermeijer**, **Tim Lugtenburg**, **Jianyao Jin** and **Monika Molnar**, thank you all for the great times we had. After the COVID-19 pandemic the PhD council was bleeding to death and I think we did a great job at reviving it. We have more members than ever and achieved great things for future PhD candidates. It was awesome to see the developments happen over the years. I greatly cherish all your input and it was wonderful to, next to the research work, have the possibility to work with such social, hardworking and productive people. The events always sounded insane on paper, but we managed to pull them off every time. It was an amazing opportunity to get to know people from all over the campus. Each of you are amazing people and I wish you good luck with the remainder of your PhD projects or other current endeavors. I hope you continue building communities and connections, the world needs more people like you.

A special thanks goes out to my mentor **Gabrie Meesters**. Most PhD candidates in our faculty are assigned a mentor and maybe speak to them once during their PhD journey. By sheer luck, I was assigned a mentor that truly cared about me and my career. Gabrie, I am glad that you offered meaningful guidance outside of the research work and helped me reflect on my own skills, personal life and career path. You helped me greatly throughout this journey and I am forever grateful for that.

While writing this acknowledgment section I suddenly realized that this is my dissertation, so I can do whatever I want. I would like to give a shoutout to my longtime friends **Werner Kastelein** and **Timothy Lugtenburg**, with whom I had the pleasure to enjoy various deep philosophical discussions and random events throughout my PhD journey. Thank you guys for all the adventures and I hope to never see the inside of De Kurk with you guys again. I would also like to give a shoutout to **Kalani Ostermeijer**, my homie. Always there to say yes to social plans and always down to discuss the deeper meaning of life. I wish we could have gone to more conferences together, because I am

very jealous of your adventurous way of approaching travels and holidays. Your vibe-driven approach to tackle challenges is deeply disturbing to someone as rational and data-driven as me, but somehow everything always worked out and we worked super well together. I will miss having you around, but I hope we can continue hanging out together. All the best with the rest of your PhD.

During my PhD I lived on the TU Delft campus, right next to the sports center X. I cannot forget all the fun times that I have had at this sports center playing dodgeball, beach volleyball, board games or cooking random dishes at cultural events. To everyone involved, thank you so much for the great time. To the **Delftsche Dodgers**, I hope you guys keep having fun playing dodgeball and win the national cup at least once. A special thanks to **Evi Marian Bulters** and **Marvin Dee**, who have grown to become friends to me through X. I greatly enjoyed your singing lessons Marvin. I definitely cannot unsee the video you showed me about the inner workings of vocal cords, I will never forget how they work.

My PhD has led me to various places around the world, in fact, it was my first time setting foot in some countries, for example the USA, the UK or Switzerland. At conferences, workshops and summer schools I had the honor of meeting a lot of renowned people in the field and I am deeply grateful for all the discussions, advice and fun I got to enjoy. **Robert Pollice**, thank you for your feedback on my ideas since the first conference where we met. It was fun to have a continuous discussion over the years. **Kjell Jorner**, thank you for the creation of Morfeus and the deeply technical discussions we had, I really appreciate it. **Arghya Bhowmik**, **Bardi Benediktsson** and **Francois Cornet**, thank you for the many discussions we had on diffusion models for organometallic complexes. Although it was a side project to all of us, I learned a lot from you. **Juan Vicente Alegre**, **David Dalmau Ginesta** and **Brenda Manzanilla**, thank you for letting me in on the project with your qdescp module of AQME. Even though it was in the middle of a very hectic period of my life, I greatly enjoyed it. Also, thank you guys for organizing the CAMLC24 workshop in Jaca, it is an amazing initiative which I think will continue to be very successful. **Fiona Kearnes**, **Haley Michel**, **Henning Remm**, **Soumik Das**, **Sandeep Dash**, **Cher-Tian Ser**, **Wojtek Trede** and **Filip Szczypiński**, thank you guys for the incredible experience at the GRC in Maine. It was super intense, but I am glad to have experienced it with you. I do not think I will ever play as much cornhole ever again in my life. Filip, thank you for the scientific discussions and the many attempts at trying to get me to do a postdoc in the UK afterwards. I am glad to have discussed our science in so much depth and I wish you all the luck with your new research group.

I would like to thank the close friends and family members that supported me throughout this journey, especially my brothers **Prajesh Kalikadien** and **Shravan Kalikadien**. Know that the hardships we have faced are always temporary and the main thing that matters in this world is your own happiness and well-being. Thank you for your endless support and always standing by my side. As your brother, I am incredibly proud of the people you have become and wish you all the success and happiness you deserve. Additionally, I would like to thank my parents **Bina Ramdiansingh** and **Dilip Kalikadien** for raising us and, in doing so, providing sufficient fuel for the development of an independent mindset and the ability to think critically and freely, qualities essential to both science and personal growth. Aunts, uncles and cousins who have greatly supported

me and prevented me from going absolutely insane during the past years, thank you for your unconditional support. **Dayant Ramkalup**, perhaps the future prime minister of the Netherlands, thank you for our meaningful dinners and strolls through various cities. It was wonderful to have a non-STEM friend by my side who still understood the challenges I was facing. I found great joy in your company and hope we will remain friends for a long time to come. As I mention in Proposition 9: "First-generation students and scientists face structural disadvantages in academia, as their environments often lack the cultural capital to navigate academic systems." Yet, I have learned that with the right people around you, those disadvantages can be softened. A PhD is not only an academic journey but also a deeply mental one, something not everyone realizes.

Last but not least, I would like to thank the main person who managed to pull me through the majority of challenges that I have faced over the past years, my amazing girlfriend and partner: **Julia Bustillo**. Dear Julita, never in my life would I have imagined that I could meet someone so beautiful, intelligent and social as you. You often had no clue what I was complaining about, but you still managed to always steer me towards solutions. We have gone through very rough but also very happy times during the past years and there is no one else who I would have preferred to be by my side through it all. Thank you for all your support and I am sorry for working crazy hours. You are my favorite person in the world and I look forward to (finally) spending some free time with you. I am excited for our future adventures, whether it is about our house renovations, travels, or random dates in the city, I know that with you it will always be full of meaning and fun. This dissertation and all of my achievements over the past four years would not have been possible without your unwavering support and trust in me. Thank you for being you.

To everyone I may have forgotten to mention by name: thank you. Your support, conversations, and presence have certainly contributed to this work. This dissertation is dedicated to you all.

I would like to end with a quote from *The Office (US)* that feels especially fitting as I write these final lines: "I wish there was a way to know you're in the good old days before you've actually left them." Looking back, I realize I have been in those good old days all along.

- Adarsh Varun Kalikadien

List of Publications

Publications Within the Scope of this Dissertation

Kalikadien, A. V.; Pidko, E. A. *J. Phys. Chem. A.* **2026**, *accepted*

Pedrazzi, F.; Kalikadien, A. V.; Valsecchi, C.; Lefort, L.; Pidko, E. A. *in preparation* **2025**

Kalikadien, A. V.; van der Lem, N. J.; Valsecchi, C.; Lefort, L.; Pidko, E. A. *Digit. Discov.* **2025**, *4*, 2033–2044

Finta, S.; Kalikadien, A. V.; Pidko, E. A. *J. Chem. Theory Comput.* **2025**, *21*, 5334–5345

Kalikadien, A. V.; Valsecchi, C.; van Putten, R.; Maes, T.; Muuronen, M.; Dyubankova, N.; Lefort, L.; Pidko, E. A. *Chem. Sci.* **2024**, *15*, 13618–13630

Baidun, M. S.; Kalikadien, A. V.; Lefort, L.; Pidko, E. A. *J. Phys. Chem. C* **2024**, *128*, 7987–7998

Kalikadien, A. V.; Mirza, A.; Hossaini, A. N.; Sreenithya, A.; Pidko, E. A. *ChemPlusChem* **2024**, e202300702

Publications Outside the Scope of this Dissertation

Manzanilla, B.; Dalmau, D.; Kalikadien, A. V.; Pidko, E. A.; Sigman, M. S.; Alegre-Requena, J. V. *in preparation* **2025**

Anker, A. S.; Aspuru-Guzik, A.; Bechtel, T.; Bigi, F.; Briling, K. R.; Das, B.; David, N.; Day, G. M.; Deringer, V. L.; Dyer, M.; Eardley-Brunt, A.; Evans, M. L.; Evans, R.; Franklin, B. A.; Ganose, A. M.; George, J.; Goulding, M.; Hickey, N.; James, G.; Kalikadien, A. V. *et al. Faraday Discuss.* **2025**, *256*, 639–663

Albornoz, R. V.; Antypov, D.; Blanke, G.; Borges, I.; Bran, A. M.; Cheung, J.; Collins, C. M.; David, N.; Day, G. M.; Deringer, V. L.; Draxl, C.; Eardley-Brunt, A.; Evans, M. L.; Fairlamb, I.; Fieseler, K.; Franklin, B. A.; George, J.; Grundy, J.; Johal, J.; Kalikadien, A. V. *et al. Faraday Discuss.* **2025**, *256*, 520–550

Anker, A. S.; Aspuru-Guzik, A.; Mahmoud, C. B.; Bennett, S.; Briling, K. R.; Changiarath, A.; Chong, S.; Collins, C. M.; Cooper, A. I.; Crusius, D.; Darmawan, K. K.; Das, B.; David, N.; Day, G. M.; Deringer, V. L.; Duarte, F.; Eardley-Brunt, A.; Evans, M. L.; Evans, R.; Fairlamb, I.; Franklin, B. A.; Frey, J.; Ganose, A. M.; Goulding, M.; Hafizi, R.; Hakkennes, M.; Hickey, N.; James, G.; Jelfs, K. E.; Kalikadien, A. V. *et al. Faraday Discuss.* **2025**, *256*, 373–412

Aspuru-Guzik, A.; Bechtel, T.; Bernales, V.; Biggin, P. C.; Bigi, F.; Borges, I.; Briling, K. R.; Cheung, J.; Collins, C. M.; Darmawan, K. K.; David, N.; Day, G. M.; Deringer, V. L.; Draxl, C.; Dyer, M.; Eardley-Brunt, A.; Evans, R.; Fairlamb, I.; Franklin, B. A.; George, J.; Goulding, M.; Grundy, J.; Hafizi, R.; Hakkennes, M.; Hickey, N.; James, G.; Juraskova, V.; Kalikadien, A. V. *et al. Faraday Discuss.* **2025**, *256*, 177–220

Kalikadien, A. V.; Pidko, E. A.; Sinha, V. *Digit. Discov.* **2022**, *1*, 8–25

Krieger, A. M.; Sinha, V.; Kalikadien, A. V.; Pidko, E. A. *Z. anorg. allg. Chem.* **2021**, *647*, 1486–1494

ORAL CONTRIBUTIONS

Homogeneous catalyst design in the digital age: insights from machine learning and data representation, The Netherlands' Catalysis and Chemistry Conference **2025**, Noordwijkerhout, the Netherlands

Mainstreaming open science - what's our next step? Fireside chat, Open Science Community Delft **2025**, Delft, the Netherlands

Workshop: ML-ready representations for homogeneous catalysis, Han-sur-Lesse Winterschool in Theoretical Chemistry **2024**, Han-sur-Lesse, Belgium

Automated homogeneous catalyst design: navigating catalytic chemical space, International Conference on Theoretical Aspects of Catalysis **2024**, Sevilla, Spain

Data-driven approaches for chemical space exploration in catalysis, Gordon Research Seminar on Computational Chemistry: Modelling and Simulations to Understand Chemical Systems **2024**, Portland, the United States of America

The effects of conformational and configurational flexibility on descriptors for in-silico screening campaings, Cheminformatics, Automation and Machine Learning in Chemistry Workshop **2024**, Jaca, Spain

Open Science in homogeneous catalyst design: the Open Bidentate Ligand eXplorer in Python, TU Delft Civil Engineering and Geosciences Faculty Colloquium **2023**, Delft, the Netherlands

Automated homogeneous catalyst design: navigating the catalytical chemical space (and getting lost in the forest), NWO CHAINS-IUPAC **2023**, The Hague, the Netherlands

POSTER CONTRIBUTIONS

ML for navigating the forest of TM-based homogeneous catalyst design, Data-driven Discovery in the Chemical Sciences Faraday Discussion **2024**, Oxford, the United Kingdom

ML for navigating the forest of TM-based homogeneous catalyst design, International Congress on Catalysis **2024**, Lyon, France

ML for navigating the forest of TM-based homogeneous catalyst design, Amsterdam AI for Sustainable Molecules and Materials kickoff **2024**, Amsterdam, the Netherlands

ML for navigating the forest of TM-based homogeneous catalyst design, NWO CHAINS **2024**, Veldhoven, the Netherlands

Navigating the forest of TM-based homogeneous catalyst design with ML, Conference: The Path of Quantum Chemistry Into the 21st Century **2024**, Zürich, Switzerland

Navigating the forest of TM-based homogeneous catalyst design with ML, ChemAI **2023**, Amsterdam, the Netherlands

Rapid data-driven chemical space exploration for transition-metal catalysts, The Netherlands' Catalysis and Chemistry Conference **2022**, Noordwijkerhout, the Netherlands

Paving the road towards automated homogeneous catalyst design, International Conference on Theoretical Aspects of Catalysis **2022**, Lyon, France

Student Theses

Tai Hong Chow: *Data, Representation, Models and Analysis: the four horsemen of machine learning for homogeneous catalysis*,
https://resolver.tudelft.nl/uuid:a23acf39-14ca-413b-9e8b-1351505312bd

Sára Finta: *Research in high-throughput conformer search methods for homogeneous catalysis*, https://resolver.tudelft.nl/uuid:60bac9b8-6a44-4cc2-8984-ab311c059f1f

Jan de Korte: *Application of Language Models to homogeneous catalysis*,
https://resolver.tudelft.nl/uuid:b0c07527-6e87-414a-a4be-fd222532f57b

Joyce Sweere: *Exploring the configurational space of homogeneous catalysts*,
https://resolver.tudelft.nl/uuid:d9cc72f9-e343-4b93-a572-c6db35a2a069

Niels van der Lem: *The influence of ligand configurations on TM-complex chemical space exploration*, https://resolver.tudelft.nl/uuid:a0e89deb-4811-4d1e-aa30-c6e5e9e3343d

Imme Schoot Uiterkamp: *Benchmarking QC optimisation methods for homogeneous TM-based catalysts*,
https://resolver.tudelft.nl/uuid:466f4ad2-cac1-410c-b580-f29832369d6a

Aydin Najl Hossaini: *Transferability of descriptors for in silico catalyst screening*,
https://resolver.tudelft.nl/uuid:70cb53d5-ba87-4e6c-b962-736a6442020c

Adrian Mirza: *Featurization of chemical reactions for in silico catalyts screening*,
https://resolver.tudelft.nl/uuid:b6aa9ae1-b602-4978-a3e7-0ea271def445

Niels van der Lem: *Describing the workflow of high-throughput kinetic experimentation data*, https://resolver.tudelft.nl/uuid:7a82dce3-d4e4-4e91-85fb-2cb9f8fef517

Britt van Dongen: *Study of the imine hydrogenation mechanism over an Ir-based catalyst using DFT*, https://resolver.tudelft.nl/uuid:e1a706e0-c014-4825-a640-d9b1eb103a31

Mark Heezen: *Optimisation of in silico techniques for homogeneous transition metal based catalysis research*, https://resolver.tudelft.nl/uuid:7a3259cb-b2cc-4538-9b3e-ee4282a6fddd

Curriculum Vitae

Adarsh Kalikadien was born in the Hague, the Netherlands. In 2016 he enrolled at Delft University of Technology, where he obtained a joint Bachelor of Science degree in Molecular Science and Technology from Delft University of Technology and Leiden University in 2019, with a major in Chemical Technology and a minor in Computer and Data Science.

In 2019, Adarsh began his Master of Science in Chemical Engineering at TU Delft, specializing in process engineering. His master thesis, conducted during the COVID-19 pandemic within the Inorganic Systems Engineering group under the supervision of prof. dr. Evgeny Pidko and dr. Vivek Sinha, focused on developing computational tools for data-driven exploration of chemical space to accelerate homogeneous catalyst discovery. His pitch on this research won the online Poster Pitch Prize awarded by the KNCV. The work was awarded the runner-up position for the KNCV Golden Master Award and led to multiple scientific publications. Adarsh's academic achievements have been complemented by practical experience through an honors program and internships in quantitative research at Van Lanschot Kempen and process engineering at Air Liquide. In 2021 he graduated with honors.

Following his master's degree, Adarsh joined prof. dr. Evgeny Pidko's group at TU Delft as a PhD candidate in 2021. His doctoral research focused on data-driven *in silico* methods for the rational design of TM-based catalysts. The work was carried out in collaboration with Janssen Chemical Process R&D. The main findings of this work are documented in multiple publications and described in this dissertation. During his research, he has given various (inter)national scientific lectures and workshops. He participated in the KNCV Spotlight Competition, where his pitch was awarded the runner-up position. Throughout his PhD program, he has been actively involved in teaching activities, scientific outreach, Open Science and organizational initiatives for the Delft University of Technology.

Beyond AI for chemistry, he experiments in the kitchen, where combining ingredients and lowering barriers between international cuisines often leads to delicious reaction products.


TU Delft
ADARSH V. KALIKADIEN