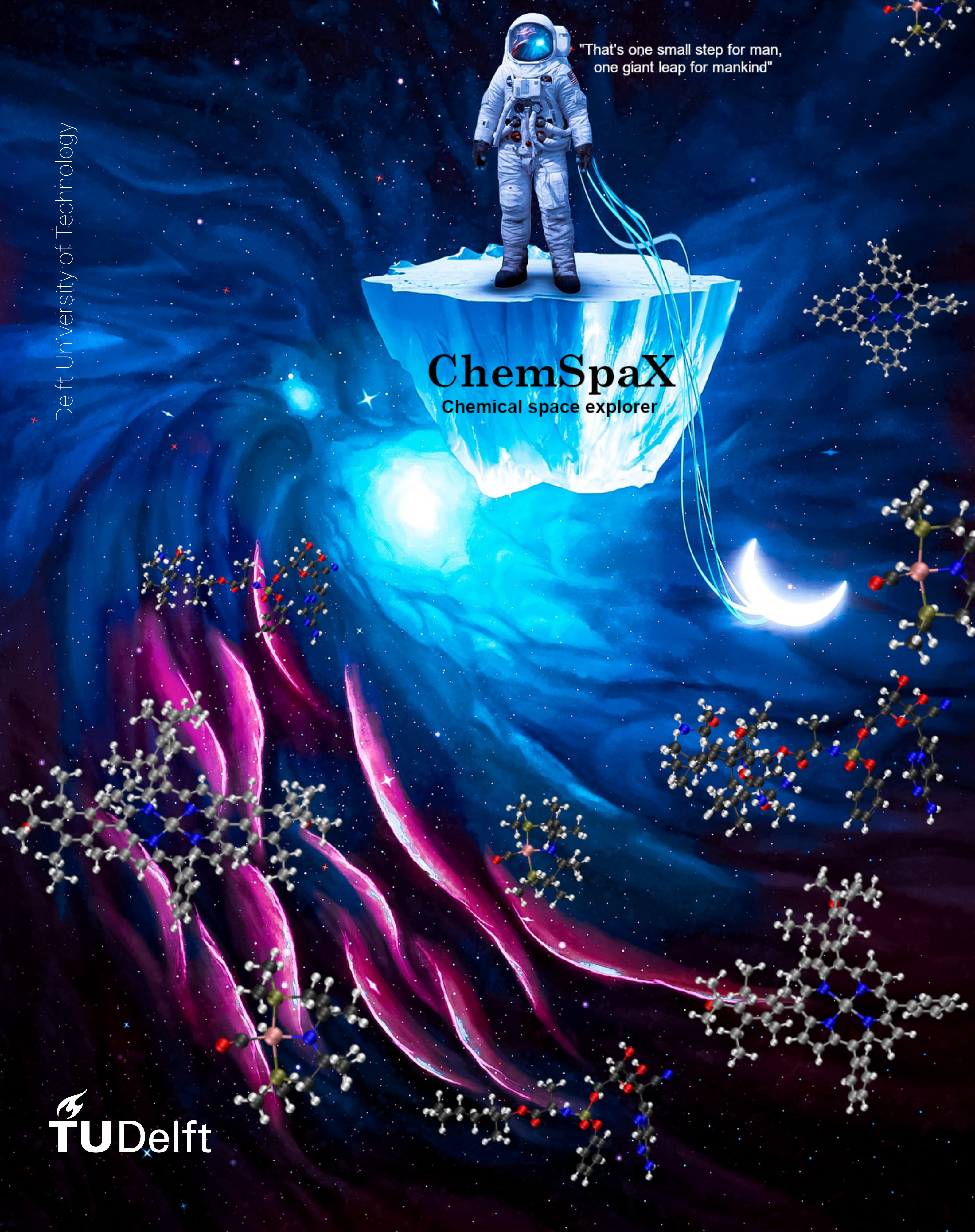
Automated data-driven exploration of chemical space for catalysts

Adarsh Kalikadien



Delft University of Technology

TU Delft

Automated data-driven exploration of chemical space for catalysts

by

Adarsh Kalikadien

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday April 22, 2021 at 9:30 AM.

Performed at:
Inorganic Systems Engineering
Faculty of Applied Sciences

Under supervision of:
Prof. Dr. E. A. Pidko
Dr. V. Sinha

Student number:	4545745	
Project duration:	August 10, 2020 – April 22, 2021	
Thesis committee:	Prof. Dr. E. A. Pidko,	TU Delft, supervisor
	Dr. A. M. Schweidtmann,	TU Delft
	Dr. A. Bansode,	TU Delft

Abstract

Catalysts play an essential role in the daily lives of humans. These catalysts are used in many industries to make processes more energetically favourable. Climate change is pushing humanity towards the usage of more green energy and catalysts play an important role in this transition. For example, in the hydrogenation reaction used for the storage of H_2, where the catalyst is involved in the storage and removal of H_2 on a storage medium like CO_2. The properties of the catalyst involved in this (de)hydrogenation reaction can affect the selectivity and yield of the reaction. Designing a catalyst that maximizes the property (yield for example) that we are interested in for a specific reaction, is an essential asset to tune catalyzed processes.

Computational screening of many catalysts has attracted the attention of academia and industry due to constant developments in the field of computational chemistry. In these computational methods, predictive models together with DFT and/or DFTB methods can be used to correlate a set of reaction descriptors with catalyst properties. The model has a higher probability to find novel molecules with a high activity when more (reliable) training data is used and when the search space of the model is confined to a local chemical space. This means that newly added molecules for screening should be structurally closely related to the molecule that was used to build the model. Unfortunately, large data sets are not readily available for transition-metal containing complexes although these complexes are widely applied in the field of homogeneous catalysis.

In this research a Python-based workflow, *ChemSpaX*, that is aimed at automating local chemical space exploration for any type of molecule is introduced. This workflow enables the user to place fragments on molecules based on 3D information, while staying close to the quality of the initial structure. This enables data-driven property calculations and prediction models, which could eventually be extended towards the automated design of new catalysts. Various representative applications of *ChemSpaX* are presented in which data-driven xTB and DFT property calculations are done. The found correlations between catalyst properties are shown and it is shown that *ChemSpaX* generates structures that have a reasonable quality for usage in data-driven prediction models for high-throughput screening.

Contents

Appendices

List of Figures

List of code listings

Abbreviations

TM	Transition metal
RMSD	Root mean square deviation of atomic positions
hRMSD	Cartesian heavy-atom root-mean-square deviation
HTS	High-throughput screening
VS	Virtual screening
DFT	Density functional theory
DFTB	Density functional based tight-binding
SCF	Self-consistent field
QC	Quantum chemistry
XC	Exchange-correlation
LDA	Local-density approximation
GGA	Generalized gradient approximation
PES	Potential energy surface
KS-DFT	Kohn-Sham density functional theory
GB	Generalized Born
SASA	Solvent accessible surface area
ML	Machine learning
MQN	Molecular quantum number
HOMO	Highest occupied molecular orbital
LUMO	Lowest unoccupied molecular orbital
FF	Force field
UFF	Universal force field
GAFF	Generalized amber force field
SVD	Singular value decomposition
CADD	Computer-aided drug discovery
QM	Quantum mechanics
SM	Statistical mechanics
CM	Coulomb matrix
SMILES	Molecular-input line-entry system
OLS	Ordinary least squares
GA	Genetic algorithms
LFESR	Linear free energy scaling relation

1

Introduction

Modern civilization would not exist without the invention of catalysts, as catalysis is involved in the processing of over 80% of all manufactured products [1]. Finding or creating a catalyst with the right catalytical properties is thus of great importance. A catalyst is defined as: "A substance which increases the rate at which a chemical reaction approaches equilibrium without becoming itself permanently involved" [2]. Catalysts can change the mechanism of a reaction which causes new barriers along the reaction coordinate to be lower [3]. Since the activation energy is lowered, the reaction can have a higher rate constant compared to the uncatalyzed reaction [3]. Although this may sound simple, the reaction mechanisms for catalyzed reactions can get really complex. Generally, two types of chemical catalysts are identified.

Heterogeneous catalysts are in a different phase than the reaction mixture, typically these catalysts are solids that are added in liquid or gas reactions mixtures. The reactant needs to bind onto the surface sites of the catalyst in order for the reaction to take place and the availability of these surface sites is also the limiting factor. Heterogeneous catalysts are mostly used in the processing of petrochemicals and fine chemicals [2]. In general, heterogeneous catalysts are preferred due to their easier recovery/separation from the products which reduces the operational costs [4].

Homogeneous catalysts are in the same phase as the the reaction mixture, these catalysts are molecular complexes. Organometallics are an example of homogeneous catalysts. The field of homogeneous catalysts has seen a lot of progress. In literature, a large number of chiral ligands and modifiers are recorded which achieve high enantioselectivity [5]. Additionally, the technology is now getting a better-defined scope and limitations for selectivity, activity and productivity [5, 6].

Catalysis research has been focused on noble-metal-based pincer complexes such as those of ruthenium, iridium or palladium [7–12]. Alternatives like manganese-based complexes are also being researched because manganese represents a cheap and earth-abundant alternative to precious-metal catalysts [13]. Society is shifting towards the usage of more green energy and these catalysts will play an important role in the success of this transition. This can be shown by an example from the field of energy storage where renewable energy is stored for later usage [14]. An example of such a storage process is the utilization of CO_2. This has been of particular interest in academia and industry, due to the renewable energy applications and the usage of CO_2 as a C1 carbon building block [15]. Efforts have been directed towards researching CO_2 neutral fuels like formic acid, where the same amount of CO_2 that was used to create the fuel is emitted. Apart from being used as a fuel in a fuel cell, formic acid can be used as a medium for hydrogen storage. The formic acid would be decomposed in CO_2 and H_2 at the desired location after which the CO_2 can be reused and the hydrogen can be used in an hydrogen oxygen fuel cell [16]. This cycle would make handling and transport of hydrogen easier [16, 17]. This cycle is shown graphically in Figure 1.1 [16]. This idea would be usable due to the simplicity of the cycle, since the homogeneous hydrogenation of carbon dioxide has long been studied which lead to the development of efficient procedures [17, 18].

Formic acid can be formed by the catalytic hydrogenation of CO_2 where the CO_2 is obtained from ambient air and the H_2 is obtained by electrolysis using green sources of electricity. For this process, a catalytic system consisting of a transition metal (TM) complex, solvent and base are needed. The catalyst is needed to harness the two electrons and two protons of the hydrogen and transfer it to CO_2. This is only one of the examples where the catalyst is indispensable to the process. As said before, many more catalytic processes are essential

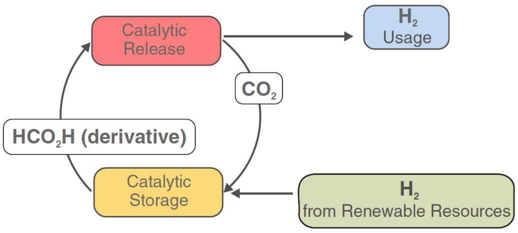


Figure 1.1: Graphical overview of a catalytical cycle for storing hydrogen in formic acid [16].

to humanity. Being able to fine-tune these catalysts, which directly affects the processes, is an important asset.

These catalysts can be broken into building blocks: the metal center(s) and its ligands. This is shown in Figure 1.2 for a pincer complex. Finding the optimal metal-ligand complex from nearly infinite possible combinations is experimentally impossible and time-consuming. Machine learning (ML), genetic algorithms (GA) and other computational methods are utilized in tackling the combinatorial problem of catalyst design [19].

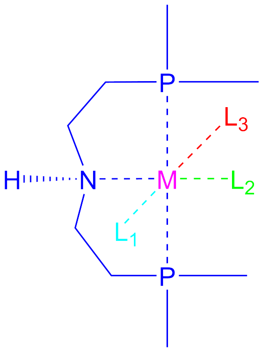


Figure 1.2: Example of a metal with 4 ligands coordinated to it. Each ligand has a distinct color. The pincer ligand backbone (dark blue) consists of an equatorial tridentate formed by posphorous, nitrogen and another phosphorous, hence the name PNP.

Using computational methods to design or select highly selective catalysts has been described as a holy grail of chemistry [20]. Thanks to molecular modeling tools that balance cost and accuracy, computational methods have taken a prominent role in the design of catalysts [21–30].

Analyzing a reaction's potential energy surface (PES) and an analysis using transition state theory (TST) are not sufficient to design the reactivity and selectivity of catalysts in most cases, dynamic effects need to be taken into consideration [20, 31]. Catalytic processes occur at a finite temperature, pressure and possibly with additives and solvents in the reaction mixture. The reaction conditions can affect the catalyst, for example, higher thermal energy induces vibrational motion which can be observed in the IR spectrum of the catalyst. Another example is the coordination to solvent molecules upon addition of a solvent. To take these dynamic effects into account a more extensive computational method is required, which increases the computational cost [32]. Thus, for screening thousands of catalysts a rational design strategy with low computational cost and high accuracy (by taking dynamic effects into account) is needed. Generally, 3 predictive strategies for catalyst design are described [25].

The first and the most primitive method is manual or interactive trial-and-error. Computational chemistry tools are becoming more accessible and are used by all sorts of chemists to test ideas and chemical intuitions [25]. There are a large number of potential interactions between reactants, catalysts and other intermediates. Additionally, the possible combinations of metals and ligands to create catalytic complexes are endless. Designing catalysts and determining the catalytic mechanism by trial-and-error, would require the chemist to set up and analyze at least the same amount of calculations as possible interactions [20]. This will be a time-consuming task if done manually.

Secondly, research is being done in the use of prediction models for catalyst design [33–36]. High-through put screening (HTS) techniques, allow hundreds of tests and can be expanded using predictive models. Predictive models seek to correlate a set of descriptors with catalyst properties. Quantitative structure–activity

/ -property relationships (QSAR/QSPR) are methods that predict chemical properties when only molecular structural information is available [37]. Macroscopic and microscopic properties of matter can be related through a set of mathematical equations. These properties can be physical, chemical, biological and technological [37]. Using these descriptors and correlations, new compounds can be designed. However, the chemical search space can be enriched if the molecular structures of newly researched compounds are similar to the structure of the compound that was used to build the model. QSAR/QSPR studies have shown that the chemical space of active compounds could be local, so the enriched local chemical space may correspond to a higher probability of discovering novel active molecules [38]. This means that this design strategy can be confined to a local chemical space, that strategy is known as local chemical space exploration [20, 39–43]. Finding the descriptors that represent structural properties is the major hurdle in making these models, some descriptors that are used in QSPR analysis can be found in literature [44–48].

Lastly, the most dedicated strategy is the automated design of catalysts, which can involve the use of generative models. To maximize their predictive power, the predictive models described above can be utilized in an automated manner. This also requires automated generation of candidate molecules, since predictive models get more reliable when more data is available. Therefore, the chemist would either have to automatically generate candidate molecules or rely on the existence of a large database of the specific catalyst the chemist is interested in. For complexes like TM based complexes, such a database does not exist yet and the computational chemist would need to generate one. The database would need to be non-biased and based on fundamental research, since difficult or unusual reaction pathways would require very customized catalysts [20]. In a systematic and objective automation workflow the predictive power can be maximized.

The goal of this research was to lay the foundations for a workflow which enables catalyst design using prediction models and could eventually be extended towards usage in automated catalyst design. The project was focused on homogeneous catalysts.

For past research, most of the exploration of the local chemical space of catalysts was done manually. This meant that the scientists had to open each structure's xyz coordinates, place the desired substituents and optimize the geometry afterwards. This process could be accelerated by automating the placement of substituents on an optimized structure, while staying close to the quality of a DFT optimized structure. This would mean that the manual labor and usage of computational resources are reduced, working towards a more efficient predictive model or an automated design workflow. For this purpose a workflow, *ChemSpaX*, which can place substituents on molecules in a high-throughput workflow was designed in this project. Although this project was mainly focused on the screening of catalysts, it is important to note that *ChemSpaX* can be applied to many geometries and thus could be utilized for other material discovery purposes like drug screening.

A simplified version of a data-driven catalyst design cycle is shown in Figure 1.3. This highlights how the designed computational workflow, *ChemSpaX*, could aid in the discovery of active catalysts. The topics highlighted in green show the parts of this cycle that were covered in this project. In this thesis the theoretical background necessary to understand and use the chemical space explorer *ChemSpaX* is given. Afterwards, representative use cases are presented in which *ChemSpaX* was applied to generate structures. In these use cases, data-driven methods are used to extract and correlate relevant descriptors of the generated structures. Finally, a summary of the project and an outlook for future research is given.

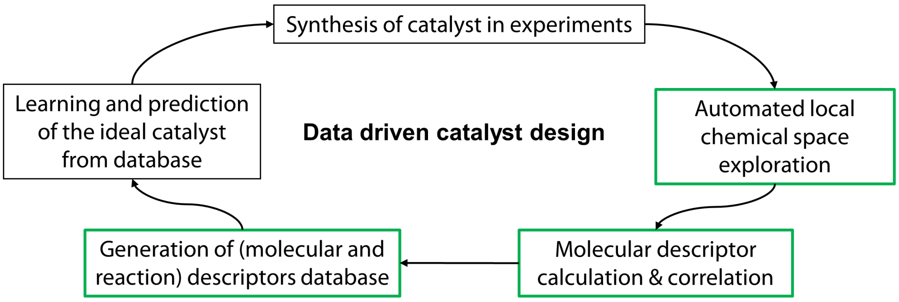

Synthesis of catalyst in experiments
Automated local chemical space exploration
Molecular descriptor calculation & correlation
Generation of (molecular and reaction) descriptors database
Learning and prediction of the ideal catalyst from database
Data driven catalyst design

Figure 1.3: Simplified version of a data-driven catalyst design cycle. The green boxes are topics that were covered in this project.

2

Theory

In this chapter an introduction into the used DFT and DFTB methods is given. Afterwards, the chemical properties that were studied in more detail are explained. Then, an explanation of the used force-field optimization methods and an explanation of the used methods to determine the quality of an optimized molecular geometry is given. Finally, the workflow of the chemical space explorer *ChemSpaX* is explained in more detail together with examples from the source code.

2.1. Density functional theory

The schrödinger equation is one of the fundamental equations in the field of quantum chemistry [49]. In its nonrelativistic, time-independent form the schrödinger equation is given by Equation 2.1.

$$\mathbf{H}\Psi_i(\vec{x}_1,\vec{x}_2,...,\vec{x}_N,\vec{R}_1,\vec{R}_2,...,\vec{R}_N) = E_i\Psi_i(\vec{x}_1,\vec{x}_2,...,\vec{x}_N,\vec{R}_1,\vec{R}_2,...,\vec{R}_N) \tag{2.1}$$

$\Psi_i(\vec{x}_1,\vec{x}_2,...,\vec{x}_N,\vec{R}_1,\vec{R}_2,...,\vec{R}_N)$ is called the wavefunction, in itself it is not a physical observable. However, the wavefunction can be squared to find the probability density:

$$|\Psi(\vec{x}_1,\vec{x}_2,...,\vec{x}_N)|^2 d\vec{x}_1 d\vec{x}_2...d\vec{x}_N \tag{2.2}$$

Which represents the probability that electrons 1,2,...,N are located in the volume $d\vec{x}_1 d\vec{x}_2...d\vec{x}_N$ with $\int ... \int |\Psi(\vec{x}_1,\vec{x}_2,...,\vec{x}_N)|^2 d\vec{x}_1 d\vec{x}_2...d\vec{x}_N = 1$.

H is the Hamiltonian operator for a system consisting of N electrons and M nuclei. The Hamiltonian operator for a system corresponds to its total energy (kinetic + potential), which is shown in Figure 2.1 [50–52]. In this equation, index i iterates over the electrons, while index A or A' iterate over the nuclei. Z_A is the charge of nucleus A and r_{ij} is the inter-electronic distance between the ith and jth electron. r_{iA} is the distance between electron i and nucleus A while $R_{AA'}$ is the inter-nuclear distance between nucleus A and A'.

$$\boldsymbol{H} = \overbrace{\boxed{-\frac{1}{2}\underbrace{\sum_{i=1}^{N}\nabla_i^2}_{\text{Electrons}} - \frac{1}{2}\underbrace{\sum_{i=1}^{N}\frac{\nabla_i^2}{M_A}}_{\text{Nuclei}}}}^{\text{Kinetic energy}} \overbrace{\boxed{-\underbrace{\sum_{i=1}^{N}\sum_{A\neq 1}^{M}\frac{Z_A}{r_{iA}}}_{\substack{\text{Coulomb attraction}\\ \text{(nucleus-electron)}}} + \underbrace{\sum_{i=1}^{N}\sum_{j\neq 1}^{M} r_{ij}^{-1}}_{\substack{\text{Coulomb repulsion}\\ \text{(electron-electron)}}} + \underbrace{\sum_{A=1}^{M}\sum_{A'>A}^{M}\frac{Z_{A'}Z_A}{R_{AA'}}}_{\substack{\text{Coulomb repulsion}\\ \text{(nucleus-nucleus)}}}}}^{\text{Potential energy}}$$

Figure 2.1: The Hamiltonian operator in the non-relativistic, time-independent Schrödinger equation [50–52]

The schrödinger equation can be solved analytically for small molecular systems. However, approximations are needed when expanding the equation to a many-body system. For example, Hartree-Fock methods provide an approximate solution to the schrödinger equation and make determination of the ground-state energy possible [53]. According to this method, the exact many-body wave function can be approximated by a single Slater determinant. A set of equations can be derived for the *N* spin orbitals and by solving these equations, the Hartree-Fock wave function and energy of the system is found. The Hartree-Fock method is often referred to as the self-consistent field (SCF) method, since the final field computed from the charge distribution needs to be consistent with the initial field. The non-linear Hartree-Fock equations are solved using iterative methods.

Density functional theory (DFT) is a popular Quantum Chemistry (QC) method to calculate properties of atoms and molecules using the electron density of the system as variable. DFT originates from the Hohenberg-Kohn theorems. These two theorems formally establish the electron density $\rho(\vec{\mathrm{r}})$ as the central quantity describing electron interactions. The first theorem states that the external potential, and hence the total energy is a unique functional of the electron density [54]. The second theorem states that the ground state energy can be obtained variationally: the density that minimises the total energy is the exact ground state density [54]. DFT uses various assumptions to approximate a solution to the schroödinger equation.

In the Born-Oppenheimer approximation it is assumed that nuclei have a kinetic energy of zero. This assumption can be done because even the lightest nucleus (a proton) is about 2000 times heavier than an electron. This approximation allows approximate separation of the wavefunction as a product of nuclear and electronic terms. The electronic wavefunction $\Psi_e(\mathbf{r}, \mathbf{R})$ is solved for a given set of nuclear coordinates. And the electronic energy obtained contributes a potential term to the motion of the nuclei described by the nuclear wavefunction $\Psi_N(\mathbf{R})$. The total energy is then obtained by adding the nuclear repulsion energy to the electronic energy [50].

Using the Born-Oppenheimer approximation and two Hohenberg-Kohn theorems, Kohn and Sham described a method to solve the Schrödinger equation using $\rho(\vec{\mathrm{r}})$ as the key variable. The Kohn-Sham energy term for the fully interacting Kohn-Sham system can be written as shown in Equation 2.3 [55]. This approach is rooted in the Hartree-Fock approach where it is assumed that the electrons move in an effective potential created by all other electrons and nuclei, thus creating a mean-field approximation for the electron-electron repulsion term J.

$$E_{KS}[\rho] = V_{AA} + V_{eA}[\rho] + J[\rho] + T_s[\rho] + E_{XC}[\rho] \tag{2.3}$$

Here, $\mathrm{V_{AA}}$ is the nuclei-nuclei repulsion term, $\mathrm{V_{eA}}$ the electrons-nuclei attraction term, J is the mean field electron-electron interaction, $\mathrm{T_s}$ is the kinetic energy functional and $\mathrm{E_{XC}}$ is the exchange-correlation functional [50]. A functional (given as '[ ]') is a function of a function. For example, $\mathrm{T_s}[\rho]$ is a functional of the electron density ρ which is a function of $\vec{\mathrm{r}}$.

$\mathrm{E_{XC}}$ describes exchange interaction and correlation effects of the electron-electron interaction that are not contained in the mean-field approximation by Hartree-Fock. The actual form of $\mathrm{E_{XC}}$ is not known, hence approximate functionals based upon the electron density are needed to describe this term. A selection of these approximate functionals will be explained in the next subsection.

2.1.1. Exchange-correlation functionals

Electrons interact via Coulomb repulsion, this repulsion stems from the Coulomb correlation between the spatial coordinates of electrons. Additionally, electrons follow the Fermi statistics. Fermi statistics state that the exchange of any two electrons reverses the sign of the total wavefunction. This is manifested in Pauli's exclusion principle, which states that two electrons of the same spin cannot be placed at the same point. This introduces a purely quantum mechanical interaction in the system, which is called the exchange interaction. Furthermore, a correlation is introduced by correlating the motions of electrons which have the same spin state. This is known as the exchange correlation or Fermi correlation.

The exchange-correlation (XC) energies within DFT can be approximated by making use of various approximation methods. These approximation methods are shown in the Jacob's ladder in Figure 2.2 where the chemical accuracy and computational cost of each approximation method are shown on a relative scale.

The simplest approximation method is the local-density approximation (LDA), where the value of $E_{XC}[\rho(\vec{r})]$ is approximated by the exchange-correlation energy of an electron in an homogeneous electron gas of the same density [56]. For systems where the density varies slowly, LDA performs well. In strongly correlated systems LDA is very inaccurate. LDA tends to find wrong ground states in many simpler cases and LDA does not account for van der Waals bonding. Additionaly, hydrogen bonding is poorly described, which is essential for most biochemistry applications [56]. These flaws were addressed with the introduction of the generalized gradient approximation (GGA). GGA also takes the gradient of the electron density into account as an additional term. Additionally, it is no longer assumed that the the electron gas is homogeneous. An example of a GGA functional is the BP86 functional, this functional was regularly used during the research reported in this thesis. BP86 is a combination of the Becke 1988 (B88) exchange functional and Perdew 86 correlation functional [57, 58]. An improvement to the GGA functionals was made by the meta-GGA functionals, which

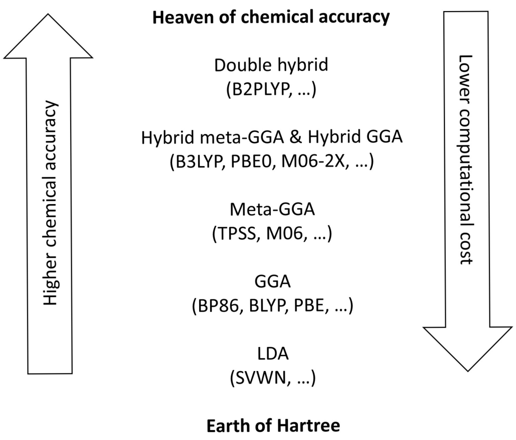


Figure 2.2: Jacob's ladder of density functional approximations which shows the increasing computational cost that go with increasing chemical accuracy for functional approximations.

also takes the Laplacian (second derivative) of the density or the kinetic energy density into account. In more recent advances the hybrid functionals were introduced. These functionals incorporate a part of the exact exchange energy from Hartree-Fock theory. These functionals are called implicit density functionals since the exact exchange energy functional is expressed as orbitals instead of a density. An example of a hybrid functional, that was used for research reported in this thesis, is the PBE1PBE (or PBE0) functional [59]. In the PBE1PBE functional 25% Hartree-Fock exchange energy and 75% Perdew–Burke-Ernzerhof (PBE) exchange energy is used along with the full PBE correlation energy [60]. More accurate hybrid functionals exist. For example, it is possible to use a combination of meta-GGA functionals with hybrid functionals or to use a double-hybrid functional. These functionals were outside the scope of this project.

2.1.2. Basis sets

A basis set is a set of linear algebra based functions that convert the HF differential equations into algebraic equations. These algebraic equations can be solved with matrix based methods, which is efficiently done by computers.

For the research reported in this thesis, the double zeta basis set def2-SVP and triple zeta basis set def2-TZVP were used [61]. The def2-SVP basis set consists of a split valence function with a polarization function on all atoms. Split valence indicates that each atomic orbital is represented by two algebraically solvable orbital-representations. Polarization functions describe the polarization of the electron density of the atom in molecules. The def2-TZVP basis set consist of a valence triple zeta function with a polarization function on all atoms. Triple zeta indicates that there are three algebraically solvable orbital-representations for each atomic orbital. In def2-TZVP these algebraically solvable orbital-representations represent larger spatial orbitals such that the spatial accuracy can increase with the tradeoff of more computational cost.

2.1.3. Geometry optimization

Geometry optimization is used to find the configuration of a molecule in which the total energy is minimized. This is done by exploring the potential energy surface (PES). The PES has local minima and maxima (peaks and valleys) which resemble optimized geometries and/or transition states. It is hard to distinguish a global minimum in the PES. A minimum of the PES is characterized by vanishing gradients of the potential energy with respect to position ($\frac{dV}{dx_i} = 0$) which means that the total force on each nucleus is vanishing. Additionally, positive second order partial derivatives are all positive ($\frac{d^2V}{dx_i^2} > 0$).

When optimizing a geometry, the method can choose whether the condition on the first derivative should apply only or if the condition on the second derivative should apply as well. Calculation of the second derivative makes the geometry optimization more computationally demanding. However, in order to distinguish between maxima, minima, transition states and higher order saddle points, it is necessary to examine these second derivatives. The matrix of second derivatives is called the hessian matrix.

2.1.4. Hessian

The hessian matrix can give additional information on a molecule and its geometry. This matrix can be diagonalized into eigenvectors and eigenvalues. The eigenvectors are normal modes of vibration and the square root of the eigenvalues corresponds to the frequency that vibrations would have in an infrared spectrum [62]. It is possible to find negative (imaginary) frequencies, this means that the geometry is at a maximum or saddle point (derivatives in all orthogonal directions are zero but geometry is not at a local extremum of the PES). (local) Minima on the PES have only positive (real) vibrational frequencies. A computational chemist that is using DFT can thus confirm that the geometry is in a (local) minimum by calculating the hessian matrix and checking if all vibrational frequencies are real. Hessian calculations can be used to calculate properties like the Gibbs free energy or enthalpy.

The vibrational free energy is one of the contributions to the Gibbs free energy and is defined as

$$F_{vib} = E_{zpe} - TS_{vib} \tag{2.4}$$

Where $\mathrm{E_{zpe}}$ is the zero-point-energy (defined as the lowest possible energy that a quantum mechanical system may have) which can be calculated by hessian calculations. And $\mathrm{TS_{vib}}$ is the vibrational contribution to the thermal energy which can be calculated by hessian calculations assuming that the ideal gas law applies to the molecular system.

2.1.5. Solvation

Different solvent models exist to simulate the effects of solvents on the system of interest. It is possible to use explicit or implicit solvent models to approximate thermodynamic properties of liquids. In explicit solvent models the system explicitly includes solvent molecules, this increases the number of interacting particles and the number of degrees of freedom of a system significantly [63]. Since the contribution to the computational effort required of these solvent molecules can get to over 90% of the total simulated system, implicit solvent models are more favourable to lower the computational cost [63]. In implicit solvent models the solvent is treated as a structureless continuum with certain dielectric and interfacial properties [63–66]. The size and shape of the continuum is subject to a tradeoff between computational cost and accuracy [65].

2.1.6. Dispersion corrections

Including the London dispersion interactions in the DFT approaches has shown to be important in order to reach high chemical accuracy of large systems like bio- or nanoarchitectures [67]. The dispersion energy is defined as a long ranged electron correlation effect and is not included in standard Kohn-Sham DFT (KS-DFT).

For the DFT calculations done for the research in this thesis, Grimme's D3 dispersion correction was used. The energy of this correction is given by [67]:

$$E_{D3} = E_{KS-DFT} - E_{disp} \tag{2.5}$$

Where $\mathrm{E_{KS-DFT}}$ is the self-consistent Kohn-Sham energy as obtained from the chosen density functional and $\mathrm{E_{disp}}$ is the dispersion correction as a sum of two- and three-body energies. Additionally, scaling factors and averaged nth-order dispersion coefficients are used together with two- and three-body inter-nuclear distances to construct $\mathrm{E_{disp}}$.

2.1.7. Density functional based tight-binding (DFTB)

DFTB uses an approximation to the KS-DFT scheme. Kohn-Sham equations are one-electron Schrödinger equations of a fictitious system of non-interacting electrons that generate the same density as any given system of interacting electrons [68, 69]. DFTB avoids any empirical parametrization by calculating the Hamiltonian and overlap matrices out of DFT-derived local orbitals (atomic orbitals, AO's) and corresponding atomic potentials [70, 71]. Although ab-intio DFT concepts are included in this method, efficiency and flexibility are improved by using concepts from the semiempirical tight-binding method [71].

Recently an extended tight-binding method, xTB, was introduced by Grimme et al. which was used throughout the research presented in this thesis [72]. Like closely related DFTB methods, the xTB methods use a semiempirical approximation to KS-DFT [72–75]. The GFNn-xTB methods (n = 0, 1, 2) focus on molecular properties that can be accurately described at a low level, namely geometries, (vibrational) frequencies and noncovalent interactions, hence the acronym GFN [72]. The first version of the GFNn-xTB methods, GFN1-xTB, uses the same approximations for the Hamiltonian and electrostatic energy as DFTB3, but does not rely on atom pair-wise parameterization [72, 74]. Instead, element specific empirical fitting is used to enable a parameterization that covers a large part of the periodic table [72]. The GFN2-xTB method was mostly used for the research presented in this thesis and brings several improvements (and an increase in computational cost) compared to GFN1-xTB. GFN2-xTB incorporates better physics, an extension of the latest D4 dispersion model and is completely pair-parameter-free [72, 76]. The total energy expression of the GFN2-xTB method is given by [77]:

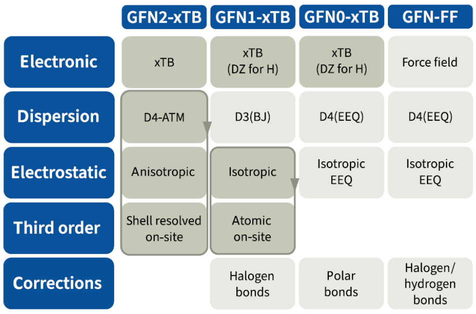
$$E_{GFN2-xTB} = E_{rep} + E_{disp} + E_{EHT} + E_{IES+IXC} + E_{AES} + E_{AXC} + G_{Fermi} \tag{2.6}$$

In the GFN2-xTB method, it is chosen to work with a spin-restricted wave function, which means that no spin density dependent terms are present. Hence it is possible that α (m_s = +1/2) and β (m_s = -1/2) orbitals can have a different occupation. It is chosen to use a finite temperature treatment, which means that electrons are treated at finite temperature. In this approach, the bands in energy are smeared such that the occupancies become continuous. Smearing means that the states of the system are occupied according to a smooth function, the Fermi distribution for example. There is an additional entropic contribution to the energy which must be calculated. The finite temperature treatment via fractional orbital occupation is chosen to be able to handle static correlations (nearly degenerate states). An example of systems with strong static correlation effects are: bond-breaking reactions, diradicals, conjugated polymers, magnetic materials, and transition metal compounds [78]. The last term G_{Fermi} is introduced by the choice of finite temperature treatment. This term describes the entropic contribution of an electronic free energy at finite electronic temperature T_{el} due to Fermi smearing [77]. The first term E_{rep} represents the classical repulsion energy which is an atom pairwise potential. The E_{disp} term describes the dispersion energy. The D3 dispersion correction calculates the inter- and intramolecular dispersion interactions only by employing the given system coordinates (and atomic numbers) as mentioned in subsection 2.1.6. The D4 dispersion correction was improved by using a less empiricial version of D3 dispersion correction together with the addition of atomic charge information. This introduced charge dependence of the dispersion coefficients improves thermochemical properties [76]. In GFN2-xTB the atomic partial charges are taken from a Mulliken population and are solved self-consistently, which allows for dropping a large three-body term from the dispersion energy equation. This noticeably decreases the computational cost of the method [72]. The E_{EHT} term is the extended Hückel contribution and is the crucial ingredient to describe covalent bonds in tight-binding methods [77]. $E_{IES+IXC}$ is the isotropic electrostatic and exchange-correlation energy, this term is treated with shell-wise partitioned Mulliken partial charge. Isotropic means that the electrostatic energy is independent of direction. The second-order charge density fluctuations are approximated by the orbital Mulliken charges [79]. This shell-wise treatment requires the definition of reference valence shell occupations. For the occupation of elements of group 1, 2, 12, 13, 16, 17, and 18, the aufbau principle is followed, whereas a modified method is used for treating transition metals and elements of group 14 and 15. E_{AES} describes anisotropic (direction dependent) electrostatic interactions. These terms are intended to improve the noncovalent interactions between the outer, less coordinated, atoms. This is used such that no extra hydrogen or halogen bond corrections nor any element-specific bond adaptations are required [77]. E_{AXC} describes the anisotropic exchange-correlation energy, this term is supposed to capture changes in the atomic exchange-correlation energy, which results from anisotropic density distributions (polarization).

In all GFNn-xTB methods, semiempirical parameters are not precomputed by first principle methods as in DFTB, but optimization on a large fit set is used to provide the best parameter combination for the desired GFN target properties [72]. An overview of the GFNn-xTB methods is given in Figure 2.3 [72].

Continuum solvation model (GBSA)

In xTB, a polar implicit solvation model based on the generalized Born (GB) model extended with the hydrophobic solvent accessible surface area (SASA) is implemented. In the GB model the solute is a continuous region surrounded by infinite solvent with a different dielectric constant than the continuous region. This GB model is introduced in the xTB Hamiltonian as a second-order fluctuation in the charge density. In addition to this polar contribution, a non-polar surface area contribution depending on the SASA of the molecule and the surface tension is added. Additionally, the SASA is also used in an empirical hydrogen-bond correction to the GB energy. Eventually, the total solvation free energy is fitted to reproduce COSMO-RS16 solvation free energies [72, 80].



	GFN2-xTB	GFN1-xTB	GFN0-xTB	GFN-FF
Electronic	xTB	xTB (DZ for H)	xTB (DZ for H)	Force field
Dispersion	D4-ATM	D3(BJ)	D4(EEQ)	D4(EEQ)
Electrostatic	Anisotropic	Isotropic	Isotropic EEQ	Isotropic EEQ
Third order	Shell resolved on-site	Atomic on-site		
Corrections		Halogen bonds	Polar bonds	Halogen/ hydrogen bonds

Figure 2.3: Summary of GFNn-xTB methods. The dark gray shaded areas denote a quantum mechanical description while light gray parts indicate a classical or semi-classical description. The parts surrounded by the arrows are treated in an iterative, self-consistent fashion. A more detailed explanation of individual components in this figure is given by Bannwarth et al. in their introduction of the xTB methods [72]

2.2. Descriptors

The calculation of molecular descriptors is needed to use QSAR and QSPR methods. A high-throughput screening (HTS) approach can be used to design the optimal catalyst [33–36]. In this method, large catalyst libraries are generated *in silico* and their catalytic performance is predicted by augmenting data from (a limited amount of) conducted experiments using statistical models. Additionally, molecular (structural or physico-chemical) descriptors, and QSAR/QSPR models are used in these predictions. Examples of used statistical methods include (linear) regression or machine learning (ML). The relations between reaction descriptors and molecular descriptors can be learned and be used to let the model predict which candidate will have optimal properties for the desired use case. For example, it can be investigated which set of structural properties (molecular descriptor) would drive the selectivity (reaction descriptor) of an investigated reaction into the desired direction. In the fully automated design of catalysts, this data can be used to generate new catalyst candidates that are likely to possess the defined optimal properties, which is called a generative model. An alternative to HTS is virtual screening, in which statistical methods are used to predict reaction descriptors (e.g. catalytic activity or selectivity) of conducted experiments based on molecular descriptors. This approach is limited due to a need for the same amount of experiments as investigated catalysts.

For material design, thousands of descriptors have been reported in the literature, which allows for limitless possbilities to construct and define a chemical space [81, 82]. Hence, an overview of molecular descriptors used in literature is given.

A group of important descriptors is fragment descriptors [83–85]. Fragment descriptors are extracted from a selected subgraph of a molecular graph and binary values (0, 1) can be used to indicate their presence or absence in the molecular graph [86]. For these fragments, the bond distance from a central atom of the fragment or SMILES strings can be used as descriptors. A major advantage of this method, breaking the catalyst into fragments, is the simplicity of descriptor calculation. Molecular quantum number (MQN) is another example of a simple molecular descriptor set consisting of atomic and bond counts and some other topological (2D information derived directly from molecular connectivity table) descriptors [87–89]. In literature, 42 of these MQN (counts for elementary constituents of molecules such as atoms, bonds, polar groups, and topological features) were used for a more focused approach to virtual screening in drug discovery [88]. Additionally, these descriptors were used to map the complex multi-dimensional descriptor space to a 2D space, which makes the chemical space and search for similarities more interpretable to humans [89].

Apart from these structural properties, physico-chemical properties like pK_a, redox potential, band gap and hydricity can be used as molecular descriptors. Some of these descriptors are dependent on the Gibbs free energy of reaction, which makes the Gibbs free energy in itself an useful descriptor to investigate [90–92]. The HOMO-LUMO gap (or band gap) is another useful descriptor which can be used to predict the strength and stability of TM complexes [93]. The HOMO-LUMO gap has been used in research to generate materials with desired electronic properties [94]. The Gibbs free energy and HOMO-LUMO gap for various structures were studied in more detail in this research next to the electronic energy. The electronic energy is already calculated when solving or approximating the schrödinger equation as explained in section 2.1. The Gibbs free energy and HOMO-LUMO gap are explained below. Both xTB and DFT were used for the calculation of these properties, with varying degrees of accuracy.

2.2.1. Gibbs free energy

When studying a (T,p) ensemble in molecular thermodynamics, the Gibbs free energy G(T, p) is defined as follows [95].

$$G(T,p) = E_{tot} + F_{vib} - TS_{conf} + pV \tag{2.7}$$

Where E_{tot} is the total electronic energy obtained from DFT as explained in section 2.1. F_{vib} is the vibrational free energy and accounts for the vibrational contributions as explained in subsection 2.1.4. Due to this factor, hessian calculations are required which make the calculation of G(T, p) computationally demanding. The third term TS_{conf} is the conformational free energy and includes configurational entropy. This term would require a power series expansion of the partition function, which is a function that describes the statistical properties of a system in thermodynamic equilibrium. The pV term accounts for expansion/compression and is negligible for solids.

2.2.2. HOMO-LUMO gap

The HOMO-LUMO gap (also called the band gap) is the difference between the energy of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecuular orbital (LUMO). This energy difference determines what type of light is absorbed and is thus crucial for photocatalysts and photovoltaic materials. In TM-complexes the HOMO-LUMO gap can be used to predict the stability, the HOMO is a potential place where electrophiles will attack so it is especially important in reaction chemistry and the LUMO is a potential place where nucleophiles may attack [96]. An illustration of the HOMO-LUMO gap is shown in Figure 2.4.

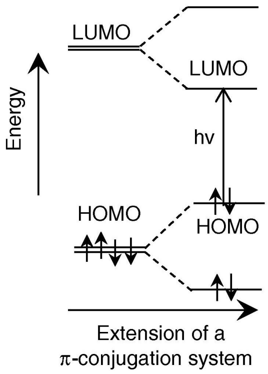


Figure 2.4: The HOMO-LUMO gap is shown for a π conjugated system [97].

2.3. Force-field (FF) optimization

Force-field methods are empirical methods that try to estimate the forces between atoms. Parameterization (either from DFT or experiments) and functional forms (relationship between a dependent variable and regressors) are essential to force-field methods. The energy landscape is described by the force field parameters. The acting forces on every atom are derived as a gradient of the potential energy with respect to the atom's coordinates [98]. Generally, the bonds between atoms are treated as springs. For the research presented in this thesis, two types of force-field methods are used.

In the Universal Force Field (UFF), the force field parameters are estimated using general rules based only on the element, its hybridization and its connectivity [99]. The original implementation of UFF determined the parameterization without an electrostatic model and is capable of reproducing most structural features across the whole periodic table with errors less than 0.1 Å in bond distances and 5° to 10° in angle bend [99].

The original Amber force fields were primarily developed for protein and nucleic acid systems and had limited parameters for organic molecules [100–104]. The generalized amber force field (GAFF) was designed to work for most pharmaceutical molecules and be as compatible as possible to the traditional Amber protein force fields. This was possible because the biomolecular parameters in the Amber force fields were developed using an extensible strategy and an extension had already been described [104]. GAFF uses 33 basic atom types and 22 special atom types to cover almost all the chemical space composed of H, C, N, O, S, P, F, Cl, Br, and I [100]. The parameterization is based on more than 3000 MP2/6-31G* optimizations and 1260 MP4/6-311G(d,p) single-point calculations [105].

2.4. Quality assessment of an optimized molecular geometry

The quality of a molecular geometry can be assessed in multiple ways. For the research presented in this thesis, two approaches were considered: 1) calculate the difference between the total electronic energy of a less accurate structure and a 'standard'. For example, $\mathrm{E}_{\mathrm{DFT}} - \mathrm{E}_{\mathrm{GFN2-xTB}}$ 2) calculate the root mean square deviation (RMSD) of atomic positions between a less accurate structure and a 'standard'. The first mentioned approach could be applied by simply using the optimization trajectory, for example: when optimizing a GFN2-xTB pre-optimized structure with DFT, the first and the last energy in the optimization trajectory correspond to the energy of the GFN2-xTB optimized and DFT optimized structure respectively. For the second approach, a more elaborate method was needed. A script developed by Dr. J. C. Kromann was used to calculate the RMSD between two structures [106]. For two molecules A and B which both have n atoms, the RMSD is defined as

$$RMSD(\mathbf{A},\mathbf{B}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((A_{ix}-B_{ix})^2+(A_{iy}-B_{iy})^2+(A_{iz}-B_{iz})^2)} \tag{2.8}$$

Note that it is also possible to calculate the RMSD between the cores of the structures, which is referred to as the Cartesian heavy-atom (all elements except H) root-mean-square deviation (hRMSD). To correctly calculate the true minimal RMSD between two structures, say molecule 1 and molecule 2, the following procedure was used:

1. The atoms of each molecule are recentered according to the centroid of the molecule (the centroid is the mean position of all the points in all of the coordinate directions (x, y and z), from a vectorset).
2. The Kabsch algorithm is used to calculate the rotation matrix that minimzes the RMSD between the two molecules [107].
3. Molecule 2 is recentered to the center of the 'view' (if no view is defined, the most outward atoms of molecule 1 determine the borders of the view).
4. The rotation matrix calculated in step 2 is used to rotate molecule 2 on top of molecule 1 such that the true minimal RMSD is calculated.

2.4.1. Calculating the optimal rotation matrix for RMSD

To calculate the rotation matrix that minimizes the RMSD between two molecules, the Kabsch algorithm is used. This algorithm requires that the centroid of molecule 1 and molecule 2 are at the origin, which is done in step 1. The molecules can then be represented in matrix notation where the x, y and z coordinates fit in a N × 3 (N = number of atoms) matrix and each row corresponds to an atom. Say for example that matrix A corresponds to the coordinates of molecule 1 and matrix B to the coordinates of molecule 2.

$$A = \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_N & y_N & z_N \end{pmatrix} \qquad B = \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_N & y_N & z_N \end{pmatrix} \tag{2.9}$$

Then the cross-variance matrix C is calculated. Which gives the covariance between all possible couples of random variables row-wise [108]. The covariance gives the directional relationship between two variables whenever one of them changes.

$$C = A^T B \tag{2.10}$$

$$C = \sum_{k=1}^{N} A_{ki}B_{kj} \tag{2.11}$$

The singular value decomposition (SVD) can then be used to get the optimal rotation matrix efficiently. First the cross-covariance matrix is decomposed.

$$C = USV^T \tag{2.12}$$

Where U is an orthogonal $m \times m$ matrix, S is an $m \times n$ diagonal matrix and V is an orthonormal (columns are an orthonormal set) $n \times n$ matrix [109]. Next, the determinant of V and U are used to check whether the rotation matrix needs to be corrected to ensure a right-handed coordinate system.

$$d = det(VU^T) \tag{2.13}$$

If d < 0 then the last row of S and V need to be multiplied by -1 to flip the z axis. So, the final optimal rotation matrix R can be calculated.

$$R = V \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{pmatrix} U^T \tag{2.14}$$

The script also supplies several options. Apart from the Kabsch algorithm, quaternions can be used [110]. Quaternions work by minimizing a single cost function associated with the sum of the orientation and position errors which was expected to improve both accuracy and speed [110]. The quaternion-based formula is equivalent to the formula derived by Kabsch, but quaternions might have a slight advantage in dealing with issues regarding chirality and degeneracy [111].

2.5. Local chemical space exploration

Local chemical space exploration is done by creating structural variations of an input structure. This approach creates structures that are closely related to each other and thus confines the search space 'locally'. In this research, this chemical space exploration was done by using various input complexes (referred to as 'skeletons') and placing substituents on indicated sites. To maximize predictive power by generating more data, it was decided that this process needed to be automated. A workflow needed to be designed that would take a skeleton complex as input together with functional groups, which could be substituted onto the skeleton. The workflow would need to output the functionalized version of the skeleton complex with a reasonable quality. After publication, the source code for the designed workflow, *ChemSpaX*, can be found on ISE's Github page (https://github.com/EPiCs-group/).

For *ChemSpaX*, it was necessary to rotate and align a substituent group that was to be placed on a molecule. The mathematical details for these operations are explained in this section together with examples from the *ChemSpaX* code. Two different approaches were taken in this code:

1. The first approach was to generate a tetrahedron on the indicated sites and place atoms on the vertices of this tetrahedron. With this approach only tetrahedral substituents could be created. The code for this approach is contained in generate_tetrahedron.py.

2. The second approach was to view the whole substituent group as a rigid block and attach it to the skeleton. The whole block is oriented correctly using the centroid vector of the substituent group. This approach works for substituents of all geometries, but requires a pre-made library of substituent groups. The code for this approach is contained in attach_substituent.py.

To orient newly placed substituents correctly in both approaches, the bond between the skeleton and the atom that will be replaced by the substituent is used. In *ChemSpaX*, these variables are called *bonded_atom* and *atom_to_be_functionalized* respectively. Note that indexing of atoms in *ChemSpaX* starts from 0. The indices of these atoms need to be given as input by the user for each functionalization to ensure correct placement of a substituent. The correct input format is explained in subsection 2.5.3 where a manual for *ChemSpaX* is given.

2.5.1. Approach 1: generate tetrahedron

The mathematics involved in approach 1 will be explained first. Although this approach was not actively used, the mathematical foundation of this approach was used to create approach 2.

In this approach a substituent is created by generating a correctly oriented tetrahedron on the skeleton and placing atoms on the vertices of this tetrahedron. In this explanation the placement of a CH3 substituent on a skeleton is taken as example. This would mean that in Figure 2.5, A would be the atom of the skeleton that is bonded to the substituent group (*bonded_atom*). The atom of the skeleton that will be replaced by the substituent group (*atom_to_be_functionalized*) is located at O. The central atom of the substituent group, carbon, would be placed at O when the functionalization occurs. The bond b between atom A and O together with the centroid of tetrahedron *ABCD* is used to rotate the new substituent group. The 3 hydrogens of CH3 would be placed on the vertices of equilateral triangle *BCD*.

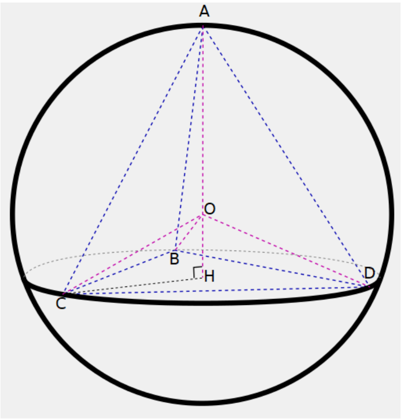


Figure 2.5: Example of a tetrahedron *ABCD*, which is used to illustrate the placement of a CH3 substituent group on a skeleton [112].

In the code, we start with a general equilateral triangle with its centroid at position [0, 0, 0]. The placement of a substituent group is done by scaling this triangle such that if the hydrogens of the substituent (3 hydrogens in our example of CH3) are placed on the vertices of the triangle *BCD* and the central atom of the substituent (carbon) replaces the *atom_to_be_functionalized* at *O*, the tetrahedron *ABCD* is formed as described in Figure 2.5. After this scaling is done, we want to find the centroid of the substituent group, which is used to rotate the substituent group correctly with respect to the skeleton.

To scale this equilateral triangle, the distance between the hydrogens of CH3, $a = CD = DB = BC$, needs to be found. The triangle law of vector addition is used to find this distance a.

$$a^2 = b^2 + b^2 - 2b^2 \cos(\theta) \tag{2.15}$$

Where b is the normalized bond length between *O* (*atom_to_be_functionalized*) and *A* (*bonded_atom*). Since we know that for a tetrahedral CH3, the H-C-H angle $\theta = 109.5$ °, this equation can be simplified.

$$a^2 = 2b^2(1 + 1/3) \tag{2.16}$$

$$a^2 = \frac{8b^2}{3} \tag{2.17}$$

$$a = 2\sqrt{\frac{2}{3}}b \tag{2.18}$$

By scaling the equilateral triangle by $2\sqrt{\frac{2}{3}}b$ we can thus create equilateral triangle *BCD* with the same distance between hydrogens as in CH3.

To calculate the centroid of the substituent group to rotate the group correctly, we need to take the central atom of the substituent into account since the central carbon atom is placed at *O*. Using that the distance from a vertex to the centroid of the tetrahedron in our case is given by $\frac{\text{length of each side}}{\sqrt{3}} = \frac{2\sqrt{\frac{2}{3}}b}{\sqrt{3}}$ together with the Pythagorean Theorem, one can find that the distance *OH* is equal to $\frac{\mathrm{b}}{3}$. This factor is used to find the actual centroid of the tetrahedral substituent. The logic for scaling the equilateral triangle and finding the new centroid is contained in *find_new_centroid* function inside the *Complex* class of *generate_tetrahedron.py*. This is shown in Code Listing 2.1.

After scaling of the equilateral triangle and calculation of the centroid, the rotation matrix for correctly rotating this tetrahedral substituent group onto the skeleton needs to be calculated. An explanation for this is given on the Mathematics Stack Exchange [113]. In our code the correct rotation is determined by two vectors, 1) the bond that will be functionalized (between *bonded_atom* and *atom_to_be_functionalized*) and 2) a normal vector pointing upwards in the z direction ([0, 0, 1]). The second vector is needed because we want the CH3 to be pointing outwards of the skeleton upon functionalization. In short, a rotation matrix R needs to be found which rotates unit vector $\vec{a}$ onto unit vector $\vec{b}$. Let $\vec{v} = \vec{a} \times \vec{b}$, $s = \|\vec{v}\|$ (sine of angle) and

```
self.bond_length = self.atom_to_be_functionalized_xyz - self.bonded_atom_xyz  # vector with origin on C and points to H in xyz plane
self.equilateral_triangle = np.array([[0, 1/np.sqrt(3.0), 0],
                                       [-0.5, -0.5/np.sqrt(3.0), 0],
                                       [0.5, -0.5/np.sqrt(3.0), 0]])  # equilateral triangle with centroid at origin
def find_new_centroid(self):
    # find new centroid and find where equilateral triangle needs to be translated to
    b = np.linalg.norm(self.bond_length)  # bond to be functionalized -H
    b = b * (2.0 * np.sqrt(2.0 / 3.0))
    self.equilateral_triangle = b*self.equilateral_triangle  # make side lengths equal to tetrahedral bond length
    centroid = self.atom_to_be_functionalized_xyz + (b/3.0) * self.normalized_bond_vector
    return centroid
```

Code Listing 2.1: In this code example the scaling of an equilateral triangle and finding the centroid of a tetrahedron is shown.

$c = \vec{a} \cdot \vec{b}$ (cosine of angle). If c is equal to -1 (which means $cos(\angle(\vec{a}, \vec{b})) = -1$, happens if $\vec{a}$ and $\vec{b}$ point in exactly the opposite directions), then the rotation matrix R is equal to the identity matrix. Else, R is given by

$$R = I + [v]_x + [v]_x^2 \frac{1-c}{s^2} \tag{2.19}$$

Where $[v]_x$ is the skew-symmetric cross-product matrix of v

$$[v]_x = \begin{pmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{pmatrix} \tag{2.20}$$

The last part of the formula can be simplified to

$$\frac{1-c}{s^2} = \frac{1-c}{1-c^2} = \frac{1}{1+c} \tag{2.21}$$

Which finally gives

$$R = I + [v]_x + [v]_x^2 \frac{1}{1+c} \tag{2.22}$$

The logic for calculating the rotation matrix is contained in *generate_substituent_vectors* function inside the *Complex* class of *generate_tetrahedron.py*. This is shown in Code Listing 2.2. Where we take a normal vector [0, 0, 1] as a and the bond length as b.

```
normal_vector = np.array([0, 0, 1])
normal_vector = normal_vector / np.linalg.norm(normal_vector)  # make unit vector
# construct rotation matrix
bond_length_norm = np.array(self.normalized_bond_vector.astype('float64'))
v = np.cross(normal_vector.T, bond_length_norm.T)  # v is perpendicular to normal vector and bond between C-H
v_x = np.array([[0, -v[2], v[1]], [v[2], 0, -v[0]], [-v[1], v[0], 0]])
v_xsq = np.dot(v_x, v_x)
c = np.dot(bond_length_norm.T, normal_vector.T)
if c != -1.0:
    rotation_matrix = np.eye(3) + v_x + v_xsq * (1 / (1 + c))
else:
    rotation_matrix = np.eye(3)
```

Code Listing 2.2: In this code example construction of a rotation matrix is shown.

After constructing this rotation matrix the rotation is applied to the tetrahedral group using the calculated centroid vector. This tetrahedral group is then translated to the correct distance from the skeleton, which completes the placement of a tetrahedral substituent. A disadvantage of the explained approach 1 is that only tetrahedral substituents can be created.

2.5.2. Approach 2: attach substituent block

In this approach, the substituent group is attached to the skeleton as a rigid block of atoms. This is done by aligning and translating the centroid vector of the substituent group. A visual example can be given using ice cream in a cone. By moving the cone, the ice cream contained in the cone is also moving in the same direction. The cone can be viewed as the centroid vector while the ice cream contained in the cone is the substituent group. Calculating this centroid vector for every substituent that can be attached is done before doing the functionalization, this is explained in the *ChemSpaX* manual in subsection 2.5.3. This centroid vector is generated and stored in a *CSV* file by methods in the *Substituent* class in *attach_substituent.py*. To generate the centroid vector, the *first_coordination* method of the *Substituent* class is used.

This method works by using the central atom of the substituent group to find all atoms bonded to this central atom. Carbon is the central atom of the substituent in the case of attaching CH3 and the 3 hydrogens are the atoms bonded to it. Based on this geometry of the central atom of the substituents and the atoms that are bonded to it, the centroid vector used to rotate and align the substituent group can be calculated. This calculation of the centroid vector is shown in Code Listing 2.3. A *CSV* file of substituents for which this calculation has been done is present in the *ChemSpaX* repository on Github. The name of the substituent, central atom and centroid vector are stored in *central_atom_centroid_database.csv* in the *substituents_xyz/manually_generated/* folder. Storing this data for substituents together with their xyz files and/or MDL molfiles allowed for less usage of computational resources when multiple functionalizations are done subsequently. The centroid vectors for substituents that are contained in the CSV database are used calculate the optimal rotation matrix. This approach was shown for *generate_tetrahedron.py* in Code Listing 2.2. For the current approach, the normal vector [0, 0, 1] is now replaced by the centroid vector of the substituent in Code Listing 2.2. Afterwards, the rotation is applied to the whole substituent group and the whole group is translated to the correct bonding distance from the skeleton. Finally, Universal Force Field (UFF) and Generalized Amber Force Field (GAFF), as implemented in the Openbabel package, are used to cheaply optimize the newly placed substituents [99, 100, 114, 115]. This FF optimization is done while the atoms of the skeleton are frozen, to preserve the input skeleton's quality.

```python
def scale_vector(starting_point, vector, length):
    """ Scales a vector with a given length
    :param starting_point:
    :param vector:
    :param length:
    :return: scaled vector
    """
    vector = vector/np.linalg.norm(vector)
    return starting_point + vector*length

def first_coordination(self):
# find atoms bonded to central atom of substituent, use mol file since graph representation is more accurate
    edges = get_bonded_atoms(self.path[:-4]+'.mol', self.central_atom_index)
    # scale bonds such that an hypothetical symmetrical molecule is created say C-X' C-Y' C-Z'
    for i in range(np.shape(edges)[0]):
        scale_vector(self.central_atom, (edges[i, :]-self.central_atom), self.bond_length)
    # calculate centroid of this hypothetical molecule, which will be similar to real molecule
    centroid = np.sum(edges, axis=0)/edges.shape[0]  # sum over rows and divide by amount of atoms found
    # get correct orientation of total group s.t. the centroid vector is pointing towards the central atom of the substituent group
    centroid = (centroid - self.central_atom)/np.linalg.norm(centroid - self.central_atom)
    return np.array(centroid)
```

Code Listing 2.3: In this code example the construction of a centroid vector pointing towards the central atom of the substituent group is shown.

This second approach was mainly used for the research reported in this thesis. The above mentioned steps are all contained in scripts which simplify the workflow.

2.5.3. ChemSpaX manual

To summarize, a step-by-step guide for placing substituents using *attach_substituents.py* is given. This explains the steps that were used to generate the complexes for the research presented in this thesis. Before explaining these steps in more detail, an important distinction should be made between xyz files and MDL molfiles. The xyz file format contains only the x, y and z coordinates per atom of a molecule. The MDL Molfile format contains these x, y and z coordinates together with a list of (single, double or triple) bonds formed between atoms. It is thus always desirable to use a MDL molfile for a molecule and convert it to xyz format instead of converting xyz to MDL molfile format. In the latter case the program used for conversion will make assumptions about bonds (and a computational graph representation of the molecule) based on inter-atomic distances. The code for *ChemSpaX* was initially designed to only work with xyz files and was extended to use MDL molfiles for the FF optimization. Hence, this explanation is currently focused on xyz files, but is subject to change.

1. It is recommended to start with a DFT optimized skeleton structure. After this optimization, the *atom_to _be_functionalized* and *bonded_atom* should be placed in a nested list on the comment line (the second line) of the optimized skeleton's xyz file. For example, if we want to replace the atom of index 3 with a substituent group that is bonded to the skeleton at index 2, and afterwards we want to replace index 10 that is bonded to the skeleton at index 9, then the *functionalizations_list* becomes: [[3, 2], [10, 9]]. After preparation of this skeleton, the skeleton's xyz file should be placed in the *skeletons* folder.

2. Substituents that can be placed on the skeleton are present in the *substituents_xyz/manually_generated* folder, if one of these pre-made substituents will be used then this step can be skipped.

   If a substituent needs to be added to this library, then the following procedure should be followed. If we would like to add methyl (which is already in the *substituents_xyz/manually_generated* folder): Take an xyz file for (optimized) CH4, then remove one of the hydrogens such that C (central atom of the substituent) has a lone pair of electrons. This lone pair of electrons will be used to form a bond with the skeleton. Let the C be the first atom in the xyz file of the substituent, since *data_preparation.py* assumes that the central atom of the substituent is always the first atom in the substituent's xyz file. Save the xyz file in *substituents_xyz/manually_generated* and run *data_preparation.py* which uses the *Substituent* class of *attach_substituent.py* to generate the centroid vector for the substituent group and adds this centroid to *central_atom_centroid_database.csv*.

3. If *data_preparation.py* has not been ran in the previous step, it should be ran now such that MDL molfiles from the skeleton and substituent xyz files are generated. These MDL molfiles are used to fetch correct bonding information to create the functionalized structure and FF optimize the newly placed substituent group.

4. If the same optimization (FF on newly placed substituents) method will be used as presented in this thesis, the user needs to modify the *functionalize_and_optimize_obabel.sh* script. For example, in step 1 we have defined that we want to place 2 substituents on the skeleton, thus we should make sure that we define the 2 substituents that need to be placed. This is done by modifying the *STARTING_C_SUBSTITUENT* and *RANDOM_C_SUBSTITUENTS* variables. If we want to place a CH3 on the first defined site and an OH on the second defined site, the variables become: *STARTING_C_SUBSTIT UENT="CH3"* and *RANDOM_C_SUBSTITUENTS="OH"*. If we had more than 2 substituents, the additional substituents need to be added to the *RANDOM_C_SUBSTITUENTS* variable as a string separated by a space.

5. After these preparation steps, the script can be called with the command:
   bash functionalize_and_optimize_obabel.sh C

   If the user wants to run xTB optimizations after every functionalization, the explanation in the previous step and this step should be applied to the *functionalize_and_optimize_xtb.sh* script.

3
Results

First, the introduction to *ChemSpaX* with various representative use cases is given in '*ChemSpax*: Exploration of chemical space by automated functionalization of molecular scaffold'. The SI for this section can be found in Appendix A. Secondly, an application of *ChemSpaX* to Mn based pincer complexes is shown in 'Metal-ligand cooperative activation of HX (X=H, Br, OR) bond on Mn based pincer complexes'. The SI for this section can be found in Appendix B. This application will be published in: A. Krieger, V. Sinha, A. Kalikadien, and E. A. Pidko, "Metal-ligand cooperative activation of HX (X=H, Br, OR) bond on Mn based pincer complexes," Zeitschrift für Anorg. und Allg. Chemie, 2021, doi: in press.

Apart from the presented research, several conferences, workshops and symposia were attended during this thesis project and the notes for these events can be found in Appendix C.

ChemSpaX: Exploration of chemical space by automated functionalization of molecular scaffold

Adarsh V. Kalikadien [1], Evgeny A. Pidko[1,*], and Vivek Sinha[1,*]

[1]Inorganic Systems Engineering, Department of Chemical Engineering, Faculty of Applied Sciences, Delft University of Technology, 2629 HZ, Delft, The Netherlands.

08-04-2021

Abstract

Generation of many molecular structures with reasonable quality, that resemble an existing (chemical) purposeful material, is needed for high-throughput screening purposes in material design. Large databases for complexes containing transition metals are not readily available, although these complexes are widely used in homogeneous catalysis. A Python-based workflow, *ChemSpaX*, that is aimed at automating local chemical space exploration for any type of molecule is introduced. The overall computational workflow of *ChemSpaX* is explained in more detail. *ChemSpaX* uses initial input of a molecular structure and 3D information, to place substituent groups on the input structure. The newly placed substituents are optimized using a cheap force-field optimization method. Representative applications of *ChemSpaX* are shown by the functionalization of transition-metal based pincer complexes, cobalt porphyrin complexes and a bipyridyl functionalized cobalt-porphyrin trapped in a M2L4 type cage complex. The relatively fast GFN2-xTB optimization method was used to compare structures generated by *ChemSpaX*. For selected use cases a comparison was also done against DFT optimized structures. Descriptors that can be used in data-driven material design were selected and studied in more detail for the selected use cases. It is shown that the structures generated by *ChemSpaX* have a reasonable quality for usage in high-throughput screening applications.

Keywords: catalysis; data-driven material design; density functional tight-binding theory; chemical space exploration; open source;
Article type: Software Focus

1 Introduction

The discovery of novel molecules is important for many industries. The usage of computational methods for the design or selection of highly selective catalysts has been described as a holy grail of chemistry [1]. Computer-aided drug discovery (CADD) led to the discovery of the HIV protease inhibitors ritonavir, indinavir and saquinavir. This discovery proved to be the key in reversing rapid growth in deaths due to AIDS in the US [2, 3]. This is one of the many examples that show that CADD has been playing a key role in the discovery of drugs and will continue to do so [4, 5]. Efforts are being made in the energy and chemicals sector to successfully apply computer-aided methods for discovery of new materials. For example, in the design of materials for lithium-ion batteries, hydrogen production and storage materials, superconductors, photovoltaics and thermoelectric materials [6, 7, 8, 9]. A roadmap for materials by computational design is given by Alberi et al., where it is discussed that a common need in material design is the need for high-throughput computational and experimental techniques as a foundation for the materials-by-design paradigm [10].

The chemical space is vast and global exploration of chemical space is difficult [11]. It is experimentally impossible and time-consuming to find the optimal molecule or material from nearly infinite possible combinations. Therefore, machine learning (ML) and other cost-effective computational methods are an attractive solution to the combinatorial problem of material and catalyst design [12]. Von Lilienfeld and coworkers propose that exploration and understanding of chemical space can be done by combining physical theories, data sets of quantum mechanics (QM) and statistical mechanics (SM) properties, and ML methods that incorporate physical and chemical knowledge [13]. These combinations of QM, SM and ML approaches are called QML models. In QML models, modern statistical learning theory is applied to predict electronic and atomistic properties and processess in molecules and materials [13]. However, there are challenges that need to be addressed before a complete workflow for *in silico* design of chemicals and materials can be achieved.

The first challenge is that computational methods should support scientists in adjusting their hypothesis after synthesis of a material has happened. This is part of the molecular design cycle [14]. To enable this cycle, a systemic approach for the local exploration of the chemical space of the synthesized material is needed to learn more about the chemistry involved. This approach can then be expanded to generate new candidate molecules and adjust the initial hypothesis. By building upon experimental knowledge in a systemic way together with automated computational high-throughput screening (HTS), larger subsets of chemical space can be covered.

The second challenge is that even when accurate simulations are available, the process of molecular design is still limited by the search strategy used to explore chemical space and the representation of a molecule in chemical space [15]. A differentiable continuous space is required to enable the use of gradient-based optimization and make larger jumps in chemical space [15]. An example of molecular representations is shown in literature, where autoencoders are used to map molecule structures onto a continuous latent space. The latent space preserves chemical similarity principle and thus can be used for the generation of analogue structures [16]. Another often used representation is the sorted Coulomb Matrix (CM), this representation was applied successfully in the ML screening of thousands of catalysts [17]. These representations require that the molecular structure that is represented, is of reasonable quality, while consumption of computational resources for the generation of the structure is kept to a minimum.

The third challenge is that advances in this field also highly depend on the availability of trustworthy QM data sets. For (small) organic molecules, reliable data sets like the GDB-13 or GDB-11 database exist which are being used to train generative models [18, 19, 20]. In these virtual screening workflows, 3D coordinates are generated from the simplified molecular-input line-entry system (SMILES) strings and these complexes are analyzed further. Large data sets are not readily available in the field of homogeneous catalysis and the alternative approach of using synthetic data generated *in silico* would become expensive due to the high computational cost of accurate QC methods like DFT [21, 22, 23]. This unavailability stems from the fact that transition metal (TM) complexes are regularly used in homogeneous catalysis and it is known that TM-complexes pose an issue for the SMILES. In current research, a toolkit is being developed to convert SMILES correctly to 3D *XYZ* coordinates for TM-complexes [24]. In the *molsimplify* code, a divide-and-conquer technique is used to get the correct 3D geometry of any complex. Force fields for organic components are being used together with a databases of quantum-mechanically derived rules for the metal–organic bonds [25, 26, 27, 28, 29]. For the exploration of local chemical space for TM complexes, using a SMILES string as starting point is thus something that needs additional research. Another approach that could be taken for exploration of these complexes, is using the *XYZ* coordinates directly. This is the approach that has been taken in this research, which allowed chemical space exploration for any type of geometry.

In this manuscript a tool that can be used for automated molecular design workflows is presented. *ChemSpaX*, a Python based workflow, aims to make the exploration of local chemical space of both organic and inorganic complexes as easy as possible. The exploration is done by automated placement of substituents on a given molecular scaffold while maintaining the quality of the initial scaffold. Several molecular scaffolds are already present in literature and can be used for an automated exploration of local chemical space. If a particular complex is known for its high catalytic activity, the 3D coordinates of this complex can be used as a starting point for exploration in the neighbourhood of its chemical space. With ChemSpaX it is possible to automate this exploration of the local chemical space. The user has full control of the placement of substituents groups and can thus guide the exploration of the local chemical space based on chemical intuition. A general overview of ChemSpaX is given in Figure 1.

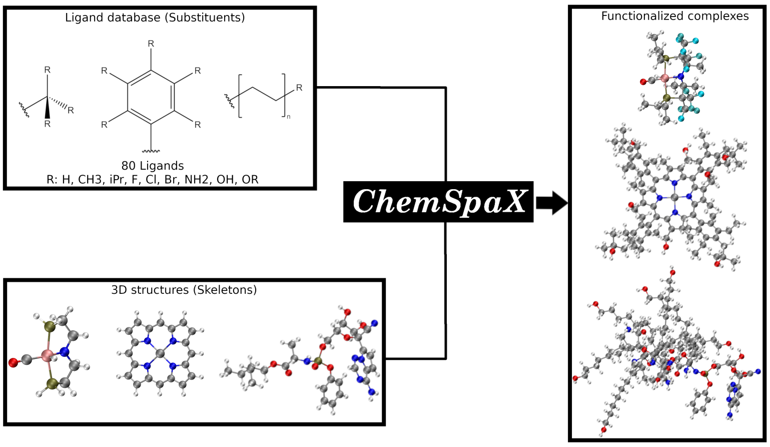

Ligand database (Substituents)
R
80 Ligands
R: H, CH3, iPr, F, Cl, Br, NH2, OH, OR
n
ChemSpaX
3D structures (Skeletons)
Functionalized complexes

Figure 1: A general overview of ChemSpaX. Using ligands and an user-defined complex, the local chemical space of this input geometry can be explored. Color code used for elements: gray = C, white = H, red = O, pink = Ru, dark-blue = N and turquoise = F.

The key features of *ChemSpaX* are presented in the next section. Subsequently, representative applications of *ChemSpaX* are presented. First, the functionalization of a RuPNP complex involved in a (de)hydrogenation reaction is studied [30, 31, 32, 33, 34, 35]. The (de)hydrogenation reactions are important in several industries. Catalytic hydrogenation has become a key technology for the manufacturing of pharmaceuticals and fine chemicals, this process replaces chemical reduction methods that generate large quanitities of waste [36]. Catalytic hydrogenation is currently the most widely applicable method for the reduction of organic compounds which causes it to belong to the most important transformations in chemical industry [37]. At healthcare company Roche, 10 to 20% of chemical reactions in fine chemical synthesis are catalytic (de)hydrogenations, catalytic hydrogenation is of importance for the economic production of carotenoids, sorbitol and vitamins A, E and C [38]. In the energy storage/conversion industry, (de)hydrogenation reactions play an important role in enabling a renewable energy-based hydrogen economy. For example, a formic acid based hydrogen battery allows efficient transportation of hydrogen [39, 40, 41]. Analyzing the properties of these RuPNP intermediates by exploration of the chemical space can thus be a valuable asset for multiple industries. For RuPNP, the quality of geometries generated by *ChemSpaX*, the energy of hydrogenation and a comparison of calculated HOMO-LUMO gaps are presented.

Secondly, the functionalization of Mn-pincer complexes as potential (de)hydrogenation catalysts is studied. This study is an extension of previous work by our research group [42]. With this application the chemical space of a more climate-friendly alternative to RuPNP is explored. Next to ruthenium pincer complexes, manganese-derived pincer complexes have also attracted the interest of the catalytic community. Manganese is known to be a cheap, abundant and biocompatible alternative to precious-metal catalysts [43]. The quality of geometry generation and energy of reaction with various adducts are investigated in more detail.

For both pincer complexes the activated catalyst (M –L) has a Lewis acid site on the metal and the ligand can act as a Bronsted base. This means that the metal can coordinate with an electron donating species while the ligand can accept a proton [42, 44]. For the (de)hydrogenation reactions, the outer-sphere Noyori-type mechanisms, involving proton transfer to the amido ligand and hydride transfer to the metal center are typically proposed. For the Mn-pincers, potential deactivation/inhibition through the metal-ligand cooperative addition of alcohol/water/base are studied [42]. The properties of the pincer complexes' intermediate with an electron donating species on the metal and a proton on the ligand are thus an important factor to ensure that the desired product is synthesized. The Noyori-type mechanism involved in these reactions together with an example 3D structure of the Mn-pincer is shown in Figure 2.

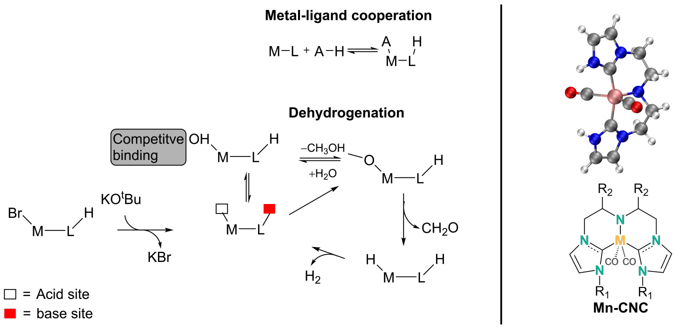


Figure 2: A representative proposed Noyori type cooperative catalytic cycle for dehydrogenation of methanol in aqueous phase is shown (left). Together with an example Mn-pincer complex (right). Color code used for elements: gray = C, white = H, red = O, pink = Mn, dark-blue = N.

Thirdly, the generation of a database of ∼1100 functionalized Cobalt Porphyrin (referred to as 'Co porphyrin' in the rest of this manuscript) complexes is shown. Co porphyrins are successful in the field of carbene and nitrene transfer reactions and its usage as a catalyst provides interesting possibilities [45, 46, 47, 48]. Automated generation of this database shows the possibility of systemic exploration of local chemical space. The generated database is used to investigate the correlation of root mean square deviation of atomic positions (RMSD) with other descriptors and to perform a regression analysis of HOMO-LUMO gaps of functionalized Co porphyrins.

Lastly, the functionalization of a bipyridyl functionalized Cobalt-porphyrin trapped in a M2L4 type cage complex (referred to as 'M2L4 cage' in the rest of this manuscript) is presented. This cage complex confines the Co porphyrin catalyst, which can lead to changed catalyst properties [49, 50, 48]. This case shows how ChemSpaX can be used to automatically functionalize structures that are more difficult to functionalize. The RMSD of various optimization methods is compared for the M2L4 cage.

2 Computational methods

2.1 Open Babel

Conversions between MDL Molfile and *XYZ* format were done using Open Babel [51, 52]. For structures generated by *ChemSpaX* the Generalized Amber Force Field (GAFF) followed by the Universal Force Field (UFF) optimization method as implemented in Openbabel was used [53, 54]. This order of optimization gave a reasonable geometry based on pre-defined molecular connections

2.2 Semiempirical tight-binding

Grimme lab's xTB package (version 6.3.3) was used for semiempirical tight-binding calculations [55]. The GFN2-xTB method and GFN-FF methods were used for geometry optimization [56, 57, 58, 59]. The RuPNP geometries were optimized using GFN2-xTB with *verytight* criteria, hessian calculations were also performed for these geometries to verify the absence of imaginary frequencies and that each geometry corresponds to a local minimum on its respective potential energy surface (PES). The Mn-pincer complexes and Co porphyrins were optimized using GFN2-xTB without hessian calculations. The M2L4 cage geometries were optimized using GFN2-xTB and GFN-FF.

The GBSA solvation method as implemented in xTB was used with THF as solvent for most optimizations, to implicitly account for solvent effects [60, 61]. These GFNn (n= 0, 1, 2, FF) methods are denoted as GFNn-xTB(THF) or GFNn-xTB(GAS) depending on whether GBSA solvation was used.

2.3 Density Functional Theory

2.3.1 Pincer complexes

Gaussian 16 C.01 was used to perform DFT calculations [62]. The BP86 exchange-correlation functional was used for geometry optimizations together with the def2SVP basis set [63, 64]. This combination of functional and basis set have shown reliable geometry predictions accompanied with low costs [65, 66]. Geometry optimizations were performed in the gas phase. Single point DFT calculations were performed using the SMD solvation (THF) model [67]. This was combined with either the BP86 or PBE1PBE (also known as PBE0) functional with the def2TZVP basis set to further refine the electronic energies [68]. All DFT calculations were performed with Grimme's D3 dispersion corrections [69]. These composite methods, BP86/def2-SVP//XC/def2-TZVP (THF), are denoted as XC(THF) or XC(GAS) depending on the exchange-correlation (XC) functional used and if a single-point calculation (SP) with solvation was done.

All geometries were pre-optimized with a combination of Openbabel's GAFF and UFF methods or GFN2-xTB before being subjected to full DFT based optimization. In this research no conformational search was conducted.

For the Ru-pincer complexes the PNP ligand was researched. Multiple ligands were selected for DFT calculations for the Mn-pincer complexes, namely: CNC, PNN and PCP ligands. The research was focused on the hydrogenation (addition of a H-H species) of the Ru based catalyst and the addition of H-X (X=Br, H, OH, iPrO) species to the Mn-based catalyst. The catalysts are represented as M-L where M represents the metal center and L the ligand. This lead to the formation of M(X)-L(H) species. The thermodynamic stability of the formed M(X)-L(H) species was estimated by computing the Gibbs free energy and total energy change under standard conditions upon addition of the H-X moiety.

$$\mathrm{H{-}X + M{-}L \longrightarrow M(X){-}L(H)} \tag{1}$$

$$\Delta G^{\circ}_{\mathrm{HX}} = G(M(X) - L(H)) - G(M - L) - G(\mathrm{H{-}X})... \tag{2}$$

$$\Delta E^{\circ}_{\mathrm{HX}} = E(M(X) - L(H)) - E(M - L) - E(\mathrm{H{-}X})... \tag{3}$$

2.3.2 Co porphyrins

TeraChem v1.94V-2019.08-beta was used to perform GPU-accelerated DFT SP calculations [70, 71, 72]. The PBE1PBE exchange-correlation functional was used with empirical dispersion corrections [68, 73]. The LANL2DZ basis set is a widely used effective core potential (ECP) type basis set and was used to model the Co metal centers [74]. All geometries were pre-optimized using the GFN2-xTB method before being subjected to DFT SP calculations.

2.4 Root-mean-square deviation of atomic positions (RMSD)

The RMSD is used to compare two molecular structures, the difference between the position of the same atom on both molecular structures is used. the RMSDs were calculated using a Python package made by Dr. Kromann [75]. First, the Kabsch or Quaternion algorithm can be used to align the cartesian coordinates [76, 77]. This ensures that real minimal average distance between atoms is calculated. Subsequently, the minimal average distance between atoms of two superimposed molecules can be calculated. If for example the two molecules **p** and **q** with n points are compared, the RMSD is defined as

$$RMSD(\mathbf{p}, \mathbf{q}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|p_i - q_i\|^2} \tag{4}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n} ((p_{ix} - q_{ix})^2 + (p_{iy} - q_{iy})^2 + (p_{iz} - q_{iz})^2)} \tag{5}$$

2.5 Linear regression

For selected Co porphyrin structures, the correlation between HOMO-LUMO gaps computed using DFT and GFN2-xTB was investigated. This was done with linear regression via ordinary least squares (OLS) fitting using the sklearn library in python [78].

3 Code implementation

ChemSpaX is a python tool that allows the automated functionalization of molecular structures, aimed at easing the creation of an automated workflow for quantum chemistry calculations. An overview of the overall workflow of ChemSpaX as described in this section is shown in Figure 3. The user has to supply: a molecule that needs to be functionalized (*skeleton*), which sites on the supplied molecule should be functionalized (*functionalization_list*) and what substituent should be placed on the supplied site (*substituent*). Substituents can be chosen from a premade database or users can supply new substituents in *XYZ* or MDL Molfile format. Based on the user input and molecular geometry calculations, the substituent is placed optimally on the skeleton. Information for the correct placement of a substituent is kept in a *CSV* file, here the central atom of the substituent group and its centroid vector is stored. After a geometrically correct placement, the GAFF and UFF optimization methods from Open Babel are used to optimize the newly placed substituent [51, 52]. This combination of GAFF and UFF was found by trial-and-error. This choice is explained in the SI.

It is recommended to use a DFT optimized geometry as input skeleton. When a new functional group is placed on the skeleton, the skeleton itself is fully constrained, hence the FF optimization only influences the newly placed functional group. This choice was made to keep the core of the geometry as close to its DFT optimized input structure as possible while preventing steric hindrance from newly placed substituents cheaply. The resulting geometries can be used for screening purposes or can be further optimized using semi-empirical methods or DFT.

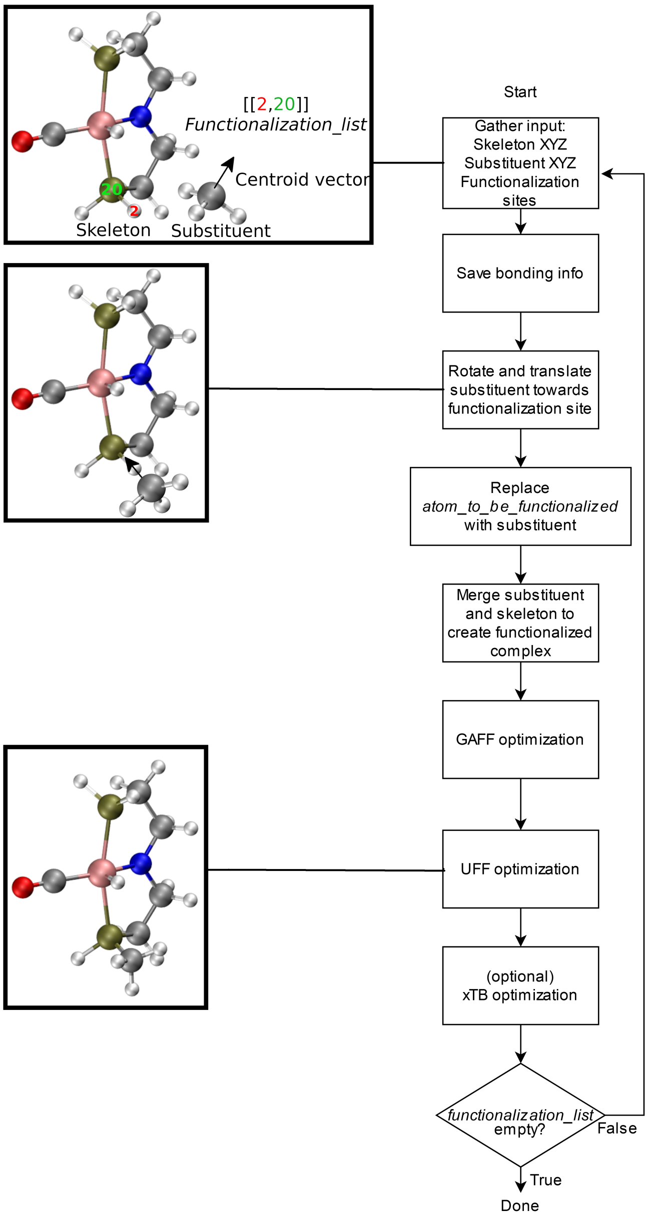


Figure 3: Overall workflow of ChemSpaX. 1) The user supplies a skeleton *XYZ*, *functionalization_list* and substituent. 2) The *XYZ* files are converted to MDL Molfiles to conserve correct bonding info. 3) The central atom of the substituent group and the centroid vector are used to rotate and translate the substituent group towards the functionalization site. 4) *atom_to_be_functionalized* is replaced by the substituent group. 5) The skeleton and substituent group are merged in one MDL Molfile with correct bonding information from input MDL Molfiles. 6) GAFF optimization is done to prevent steric hindrance. 7) Additionally, UFF optimization is done to prevent GAFF related issues. 8) Optionally, xTB optimization can be used for further optimization of the functionalized skeleton. 9) If there are no functionalizations left to do, the program is done and the functionalized skeleton is saved in MDL Molfile format. Else the functionalized skeleton will be used as input and the process starts again at step 1.

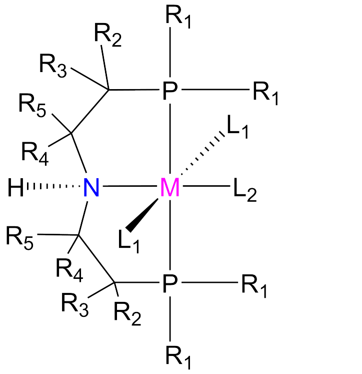
4 Results and discussion

4.1 Pincer complexes

In this research the functionalization of the ligand scaffold of TM pincer complexes was investigated. First, we looked at Ru based pincer complexes. Expanding on previous research done by our group, Mn-based pincer complexes functionalized with various adducts were investigated afterwards [42].

4.1.1 Functionalization strategy

The Ru-based complexes had a PNP-(bis(3-phosphaneylpropyl)amine)- backbone coordinated to a Ru(II) center stabilized by CO, PMe_3 and/or H ligands. For these RuPNP complexes, 288 geometries were generated by functionalization of the M-L complex and the M(H)-L(H) variant. 27 geometries were selected for BP86(GAS) optimization. BP86(THF) and PBE1PBE(THF) single-point (SP) calculations were done on the optimized geometries. Generally, the phosphorus sites were functionalized first and the carbon moieties on the PNP ligand backbone were functionalized second. This strategy is shown in Figure 4.



R_i = {Me, Et, ^{i}Pr, Ph, CF_3}; L_1 = H
L_2 = {CO, PMe_3}; M = Ru

Figure 4: Functionalization strategy for the RuPNP pincer complexes.

For the Mn-based complexes five representative ligand scaffolds were considered, PNP-(bis(3-phosphaneylpropyl)amine)-, SNS-(azanediylbis(ethane-1-thiol))-, CNC-(bis(2-(1H-3 4-imidazol-3-yl)ethyl)amine)-, PNN-(N1-(2-phosphaneylethyl)ethane-1,2-diamine)-, and PCP-(N1,N3-bis(phosphaneyl)benzene-1,3-diamine)- backbones coordinated to a Mn(I) center stabilized by CO ligands. 1225 geometries were generated using *ChemSpax* and optimized using GFN2-xTB(THF). Functionalizations were done on the M-L and M(X)-L(H) complex. From these 1225 geometries, 545 geometries were selected for DFT optimization using BP86(GAS). BP86(THF) (SP) calculations were done on the optimized geometries. Functionalizations were performed symmetrically, all four R1 sites were kept the same and both R2 sites (only 1 in case of PCP backbone) were functionalized with the same group. However, R1 and R2 were not constrained to be the same. The various backbones and the functionalization strategy is shown in Figure 5.

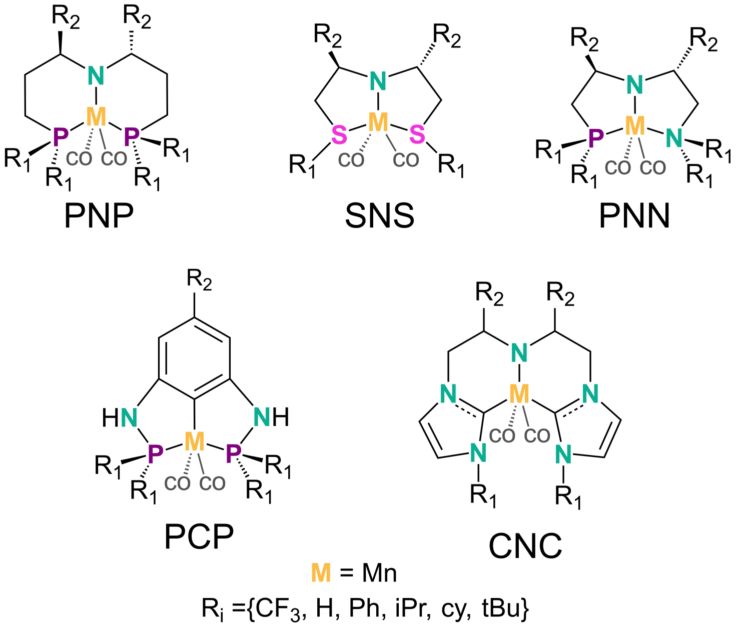


Figure 5: Functionalization strategy for Mn-pincers with various donor (R1) and backbone (R2) groups.

4.1.2 Quality assessment of generated geometries

For screening purposes and in generative models, an extensive database of accurate geometries is needed. These geometries need to be generated efficiently, so a computational efficient method with a performance close to full DFT optimization is desirable [17, 79, 80]. Methods like FF, xTB and genetic algorithms (GA) thus are attractive solutions for *in situ* structure generation. For this research all input skeletons were optimized on DFT level. *ChemSpaX* uses FF optimizations on newly placed substituent groups in each iteration, hence these geometries are referred to as 'FF geometries' in this manuscript. In this research the quality of the generated geometries was assessed by comparison with GFN2-xTB optimized structures. DFT optimized structures were used as a 'standard' for comparison.

First, the electronic energy was used for comparison. In this comparison, a FF optimized geometry is taken, its geometry is then optimized using DFT. By comparing the $\Delta \mathrm{E}$ before and after DFT by subtracting the energy from the FF optimized geometry from the energy of the DFT optimized geometry, the quality of the geometry can be assessed. By comparing $\Delta \mathrm{E}$ before and after DFT for a geometry optimized with GFN2-xTB by subtracting the energy from the GFN2-xTB optimized geometry from the energy of the DFT optimized geometry, a similar quality assessment can be done.

$$\Delta\Delta E_{GFN2-xTB} = \Delta E_{DFT} - \Delta E_{GFN2-xTB} \tag{6}$$

$$\Delta\Delta E_{FF} = \Delta E_{DFT} - \Delta E_{FF} \tag{7}$$

For RuPNP this $\Delta\Delta \mathrm{E}$ was calculated by comparing BP86(THF) and GFN2-xTB(THF) optimized structures. The mean of the $\Delta\Delta \mathrm{E}_{\mathrm{FF}}$ was found to be 7.20 kcal mol^{-1} with a relatively large standard deviation of 4.57 kcal mol^{-1}. The mean of the $\Delta\Delta \mathrm{E}_{\mathrm{GFN2-xTB}}$ was found to be 4.77 kcal mol^{-1} with a standard deviation of 2.57 kcal mol^{-1}. This indicates an overall good agreement between the GFN2-xTB optimized structures and the structures generated by *ChemSpaX*. For Mn-pincer complexes only the $\Delta\Delta\mathrm{E}_{\mathrm{FF}}$ was calculated. With an average of 27.2 kcal mol^{-1} and a very high standard deviation of 26.1 kcal mol^{-1}. The FF geometries were compared against BP86(GAS) optimized geometries because the database contained a substantial amount of structures and the DFT optimization was only done in gas phase to save computational resources. Due to this, there was no $\Delta\Delta\mathrm{E}_{\mathrm{GFN2-xTB}}$ available to compare the $\Delta\Delta\mathrm{E}_{\mathrm{FF}}$ against. This result shows that although the GFN2-xTB optimized structures and structures generated by *ChemSpaX* can be in good agreement as observed for the RuPNP structures, these structures can still be far off from the DFT optimized structure.

Apart from the electronic energy, the root-mean-square deviation of atomic positions (RMSD) can be used to assess the quality of a geometry. For the RuPNP complexes, the Cartesian heavy-atom (all elements except H) root-mean-square deviation (hRMSD) was calculated [55]. With this method the average distance between two geometries (in Å) is calculated. Again, DFT optimized structures were used as the standard for comparison. Both FF and GFN2-xTB structures had a similar average hRMSD when compared to DFT structures, 0.67 Å and 0.41 Å respectively. With low standard deviations of 0.30 Å and 0.34 Å respectively. As observed with the electronic energy, the structures generated by *ChemSpaX* are in good agreement with GFN2-xTB optimized structures. Generally, it was observed that the hRMSD's were mostly close to 1 Å and never exceeded 2.5 Å, both for FF and GFN2-xTB. Distributions of the hRMSD values are shown in the SI.

A selection of the geometries are visualized using structure overlay plots in Figure 6. The comparisons in these structure overlay plots are done in a similar way as calculating the RMSD, a) the FF optimized structure (silver), generated by *ChemSpaX*, is compared to a DFT optimized structure (green) and b) a GFN2-xTB optimized structure (silver) is compared to a DFT optimized structure (green). It was observed that FF optimization is worse at orienting the ligands on the phosphine sites correctly. However, the hRMSD mostly remains close to 1 Å. FF optimization on newly placed substituents only as employed by *ChemSpaX* can thus generate promising geometries for HTS applications since the geometries are in reasonable agreement with higher-level methods like GFN2-xTB while FF is relatively less resource consuming.

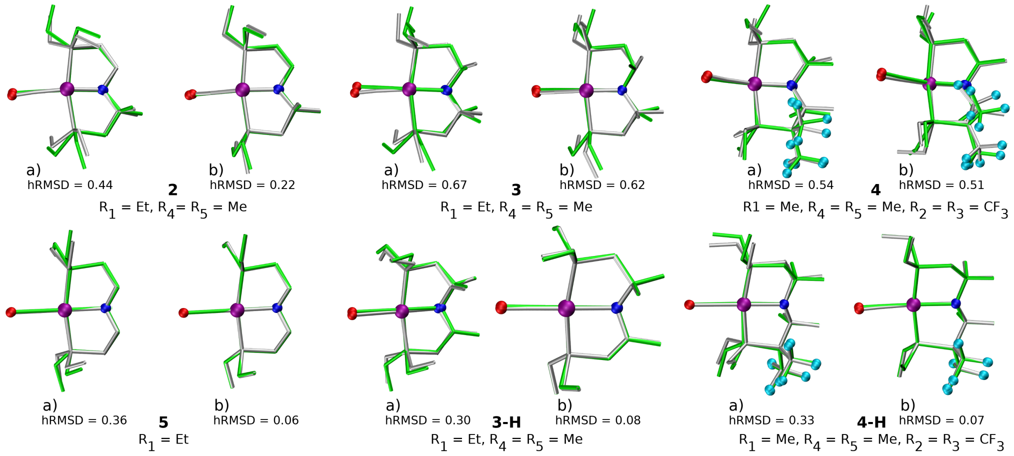

a)
hRMSD = 0.44
2
b)
hRMSD = 0.22
R_1 = Et, R_4= R_5 = Me
a)
hRMSD = 0.67
3
b)
hRMSD = 0.62
R_1 = Et, R_4 = R_5 = Me
a)
hRMSD = 0.54
4
b)
hRMSD = 0.51
R1 = Me, R_4 = R_5 = Me, R_2 = R_3 = CF_3
a)
hRMSD = 0.36
5
b)
hRMSD = 0.06
R_1 = Et
a)
hRMSD = 0.30
3-H
b)
hRMSD = 0.08
R_1 = Et, R_4 = R_5 = Me
a)
hRMSD = 0.33
4-H
b)
hRMSD = 0.07
R_1 = Me, R_4 = R_5 = Me, R_2 = R_3 = CF_3

Figure 6: Structure overlay plots of some selected TM complexes. a) shows FF optimized (silver) vs DFT optimized (green) structures and b) shows GFN2-xTB optimized (silver) vs DFT optimized (green) structures. The '-**H**' indicates that the complex is hydrogenated. Color code used for elements: red = O, purple = Ru, dark-blue = N and turquoise = F.

A comparison using the hRMSD was done in a similar manner for the Mn-pincer complexes. It was again observed that both FF and GFN2-xTB structures had a similar average hRMSD when compared to DFT structures, 0.70 Å and 0.78 Å respectively. With low standard deviations of 0.45 Å and 0.67 Å respectively. This confirms the previous observation seen from the RuPNP structures, that the structures generated by *ChemSpaX* are in good agreement with GFN2-xTB optimized structures.

The effect on the hRMSD of ligands bonded to the metal center was also compared for various backbones. This comparison is shown in Figure 7. As observed in previous research, most PNN based complexes resulted in hemilabile ligands, this hemilability could arise as an artifact of xTB based geometry optimization or it can be genuinely present in the system [42]. Due to this hemilability, the spread of the PNN backbone hRMSD data was larger compared to the other backbones. For the PNN and CNC backbones it was observed that functionalization with electron donating substituents on the R1 site resulted in a higher hRMSD. For the PCP backbone it was observed that functionalization with *t*Bu on the R1 site specifically gave a larger hRMSD. This observation is expected to have the following cause: although electron donating groups should increase basicity of the acceptor site and should hava a stabilizing effect, electron donating groups like *t*Bu are bulky and thus introduce a high amount of strain in the coordinated ligand geometry [42].

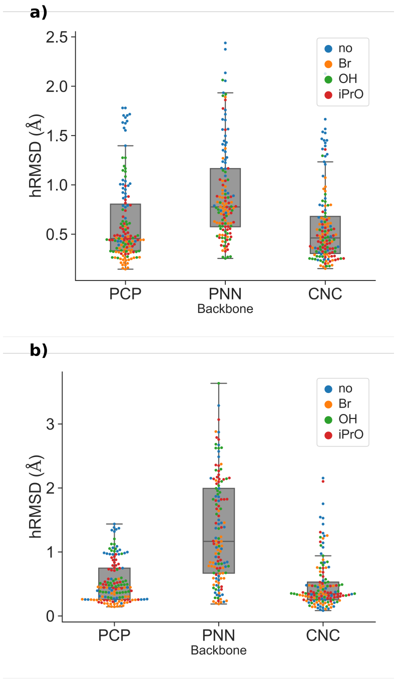


Figure 7: Comparison of hRMSD for the PCP, PNN and CNC ligand backbone of a) FF optimized structures compared to DFT optimized structures, b) GFN2-xTB optimized structures compared to DFT optimized structures. The various ligands bonded to the metal center are color coded.

4.1.3 Comparison of $\Delta\mathrm{G}$ and $\Delta\mathrm{E}$

A relation between the Gibbs free energy of reaction ($\Delta\mathrm{G}$) for species A and species B can indicate that $\Delta\mathrm{G}_\mathrm{A}$ can be used as descriptors to estimate the energetics of species B.

For the RuPNP complexes, it was observed that $\Delta\mathrm{G}$ correlates well with $\Delta\mathrm{E}$, this observation is shown in Figure 8. This is in line with a previous study done by our group where the same correlation was observed [81]. Calculation of $\Delta\mathrm{G}$ is relatively resource consuming since it involves a calculation of the full Hessian of the system. This correlation shows that $\Delta\mathrm{E}$ scales well with $\Delta\mathrm{G}$ and due to the observed minimal loss of accuracy, $\Delta\mathrm{E}$ can be used instead of $\Delta\mathrm{G}$ as descriptor in high-throughput screening (HTS) applications [81]. Thus, heavy resource-consuming hessian calculations can be skipped without losing a significant amount of accuracy.

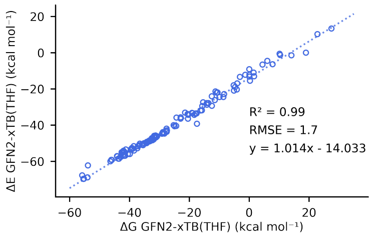


Figure 8: Comparison of the Gibbs free energy of reaction (ΔG) and the electronic energy of reaction (ΔE) in GFN2-xTB(THF) calculations for RuPNP complexes [81].

The same correlation between ΔG and ΔE was found for the Mn-pincers, regardless of the backbone and/or functionalization. Which confirms the finding that ΔE can be used in HTS applications without significantly losing accuracy. This correlation is shown in Figure 9.

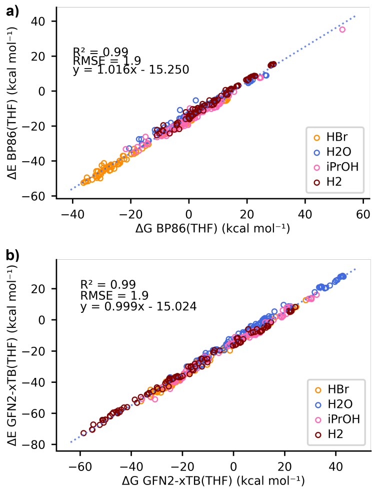


Figure 9: Comparison between the Gibbs free energy of reaction (ΔG) and the electronic energy of reaction (ΔE) in a) DFT calculations and b) GFN2-xTB calculations [81].

4.1.4 Comparison of HOMO-LUMO gap GFN2-xTB vs DFT

Functional inorganic materials that are used for photocatalysis or photovoltaics require knowledge of the HOMO-LUMO gap (also known as the band gap), since this gap determines which wavelength of light is absorbed [82, 83, 84]. Additionaly, the HOMO-LUMO gap can be used to predict the strength and stability of TM complexes [85]. HOMO-LUMO gap prediction based on only the molecular structure can thus be a great resource to screen and develop functional inorganic materials [86].

The HOMO-LUMO gap of the RuPNP pincers calculated for GFN2-xTB(THF) optimized geometries was compared to the HOMO-LUMO gap of BP86(THF) optimized structures. It was found that the HOMO-LUMO gap calculated by GFN2-xTB has a decent correlation with the HOMO-LUMO gap calculated by DFT. This result is shown in Figure 10 where a R^2 of

0.74 and a RMSE of 0.4 eV was found. This indicates a reasonable accuracy of the GFN2-xTB calculated HOMO-LUMO gap, which can be useful in HTS applications for replacing resource-consuming DFT calculations.

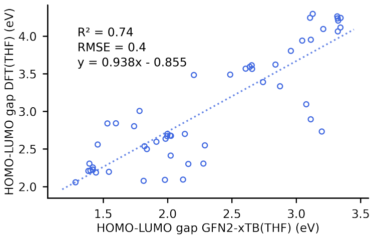


Figure 10: Comparison of the HOMO-LUMO gap calculated by GFN2-xTB and DFT.

For the Mn-pincers, the correlation between GFN2-xTB(THF) and BP86(THF) was worse $(\mathrm{R}^2 = 0.3)$. The correlation for various analyzed adducts on the metal site and for the various backbones are tabulated in the SI.

4.2 Co porphyrin

Serial functionalization of a Co porphyrins was done to create a relatively large database for analysis. Co porphyrins are reported to have a unique reactivity in carbene transfer reactions and are thus of importance for direct functionalization of C-H bonds compared to traditional hydro-carbon functionalization approaches [87, 88, 89, 90]. The potential of *ChemSpaX* is shown in the functionalization of a 2D Co porphyrin structure.

When functionalizing a structure as implemented in *ChemSpaX* (freezing the skeleton and performing FF on newly placed substituents), errors can be introduced. Stretching or compression of the skeleton structure is not taken into account since the skeleton is frozen. Relaxation of the skeleton upon placement of a new substituent group is thus barely taken into account. By investigating a structure that is close to 2D instead of 3D, the assessment of the introduced errors and their propagation by the workflow of *ChemSpaX* is simplified.

4.2.1 Functionalization strategy

Figure 11 shows the functionalization strategy for Co porphyrin, the Co porphyrin skeletons were functionalized with various phenyl groups on the R1 sites to generate 10 skeletons. These skeletons were then used to generate 1120 functionalized Co porphyrin complexes. The functionalization was done serially as described in Code implementation. The sites X1-X5 on the phenyl rings (R1) were functionalized first. Afterwards, functionalizations were done on R2 and R3 respectively. This functionalization strategy is shown for 3 different skeletons in Figure 12, where the skeleton, 5th functionalization and 15th functionalization are shown in a column. The 1120 geometries were optimized using GFN2-xTB(THF) and from these 1120 geometries 280 geometries were selected for PBE1PBE(GAS) calculations.

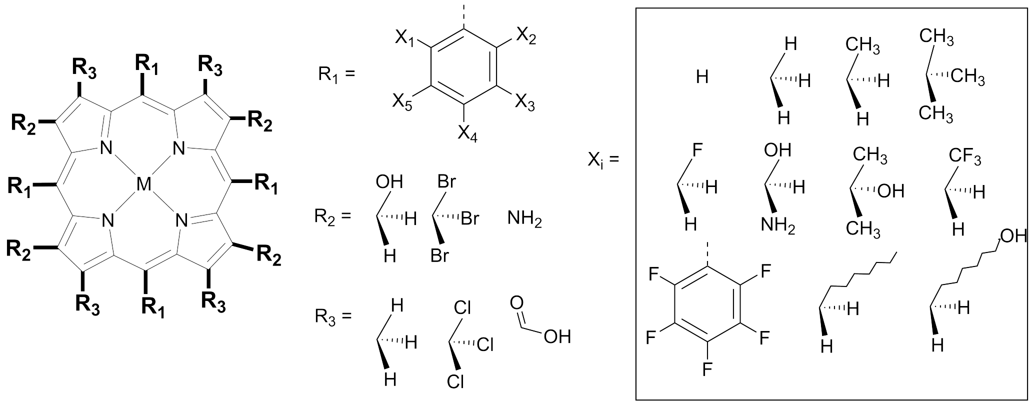


Figure 11: Functionalization strategy for Co porphyrin, phenyl groups were placed on the R1 sites and these newly placed phenyl groups were the first targets for functionalization. With this strategy a database of 1120 Co porphyrin structures was generated.

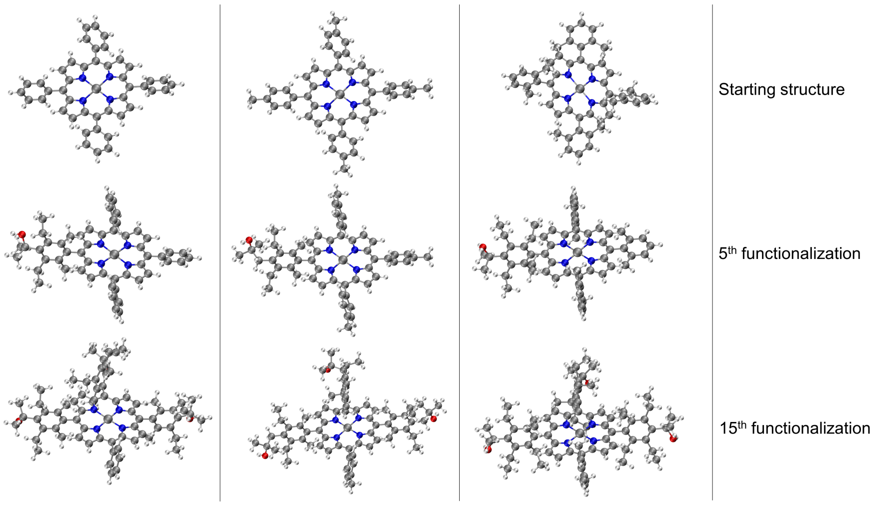


Figure 12: Functionalization strategy for Co porphyrin shown for 3 different skeletons. For each skeleton the 5th functionalization and 15th functionalization are shown in 1 column. The phenyl rings are functionalized symmetrically. In the 5th functionalization the most left phenyl ring of the skeleton is functionalized, in the 10th functionalization the same substituents are placed on the upper phenyl ring and in the 15th functionalization on the most right phenyl ring. Color code used for elements: gray = C (metal center = Co), white = H, red = O, dark-blue = N and turquoise = F.

4.2.2 Error propagation of serial functionalization

To compare the quality of the geometries generated by *ChemSpaX* against GFN2-xTB optimized geometries, the hRMSD was calculated. The average hRMSD was 1.43 Å with a relatively low standard deviation of 0.53 Å. It was observed that the hRMSDs were mostly close to 1.25 Å and the maximum was 2.69 Å.

Upon detailed analysis of the hRMSD it was observed that the hRMSD increases nearly linear for each subsequent functionalization on a skeleton. The error introduced by placing a new substituent group is thus propagated upon the next placement of a substituent. An example is shown in Figure 13 where the hRMSD for each skeleton is plotted. The first 10 blocks shows the increasing hRMSD for each functionalization on a given skeleton. The last block shows the hRMSDs for all 10 skeletons, showing how the error increases almost linearly upon each functionalization regardless of the used skeleton.

This finding could help researchers in getting the most accurate geometry when using *ChemSpaX*. One can determine when an extra geometry optimization with a higher-level method is needed in between functionalizations by setting a hRMSD treshold. When this treshold is reached, the higher-level optimization method can be used to reduce the hRMSD and the functionalization can be continued.

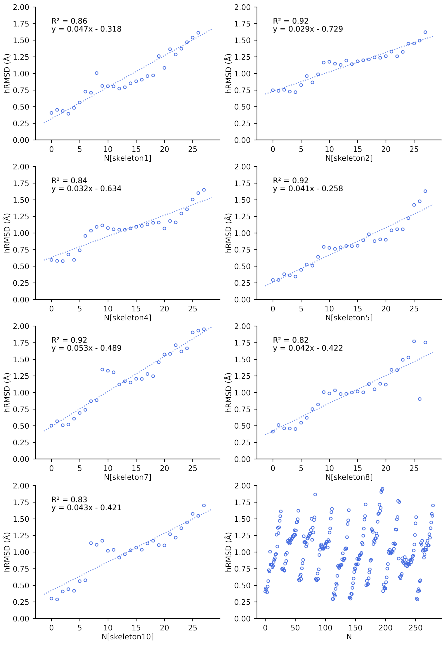


Figure 13: Increasing hRMSD for each functionalization on a given skeleton. Where N is the number of functionalizations, starting from 0. 10 skeletons were created and 28 functionalizations were done for each skeleton. The first 10 blocks each represent a skeleton, while the last block on the bottom shows the increasing hRMSD for each skeleton grouped in 1 figure. After every 28th functionalization ($0 \leq \mathrm{N} \leq 279$), a new skeleton is functionalized.

When analysing correlations of the hRMSD with other variables in more detail, clustering was observed. This observation was done when correlating hRMSD and the number of atoms in a structure. The data was divided into 3 groups, namely:

1. hRMSD $<$ 1.5 Å (low)
2. 1.5 Å $<$ hRMSD $<$ 2.0 Å (middle)
3. hRMSD $>$ 2.0 Å (high)

The clustering is shown in Figure 14. As expected from previous results, the highest hRMSDs are found in the heavy functionalized structures with a higher number of atoms. The error introduced by *ChemSpaX* upon functionalization is thus propagated, but the hRMSD remains below 2.69 Å, even after 28 functionalizations on the same skeleton.

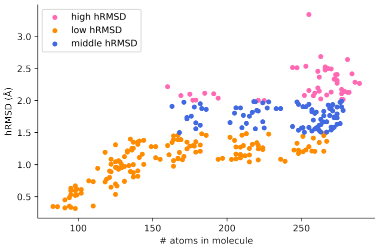


Figure 14: Observation of clustering when correlating the hRMSD with the number of atoms in a complex. 3 regimes were observed, which are color coded accordingly.

Structure overlay plots of the FF geometries (silver) and the GFN2-xTB optimized geometries (green) are shown in Figure 15. Within the 3 observed clusters there were 2 categories: structures with more than 200 atoms and structures with less than 200 atoms. For each cluster, 4 structure overlays were plotted. The upper half of the figure shows structures within these clusters with less than 200 atoms.

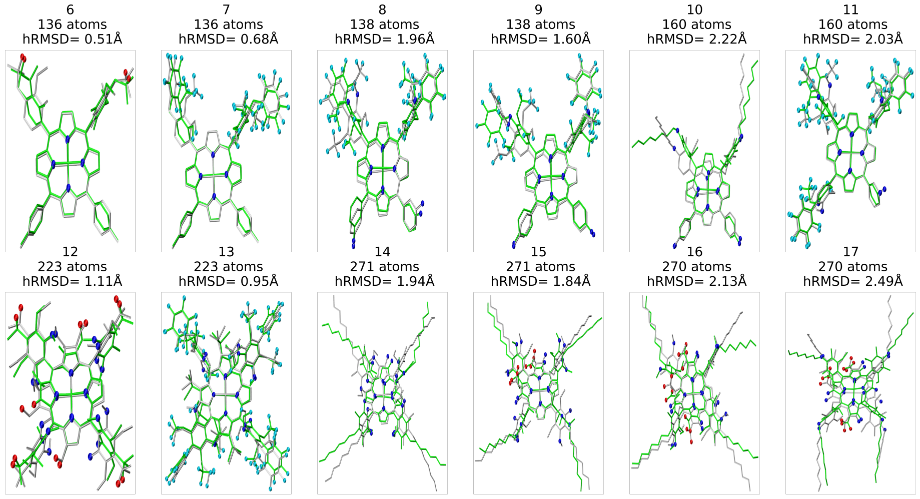


Figure 15: (Caption next page.)

Figure 15: (Previous page.) Structure overlay plots of selected Co porphyrin complexes. ChemSpaX generated (FF) structures (silver) are plotted against GFN2-xTB optimized (silver) structures. For each cluster 4 structures were plotted, the upper half of the figure consists of structures that have less than 200 atoms and the lower half of the figure shows structures that have more than 200 atoms. For example, structures 6, 7, 12 and 13 are in cluster 1 (hRMSD < 1.5 Å), here structure 6 and 7 have less than 200 atoms and structure 12 and 13 have more than 200 atoms. Color code used for elements: red = O, dark-blue = N and turquoise = F. .

4.2.3 HOMO-LUMO gap prediction

The correlation between HOMO-LUMO gaps for 280 selected structures was computed using DFT and GFN2-xTB. Three features (number of atoms in the structure, hRMSD and GFN2-xTB calculated HOMO-LUMO gap) were then used to apply linear regression via OLS fitting and predict the DFT calculated HOMO-LUMO gap. These features were chosen in an effort to select relevant and easily computable features from xTB calculations for HTS applications. 75% of the dataset was applied to learn the DFT calculated HOMO-LUMO gap, 25% of the dataset was used for testing the model. The results are presented in Figure 16. The model's $\mathrm{R}^2 = 0.71$ and the RMSE = 0.12 eV. This shows that the predictive power of xTB is reasonable for HTS applications as observed for the RuPNP complexes.

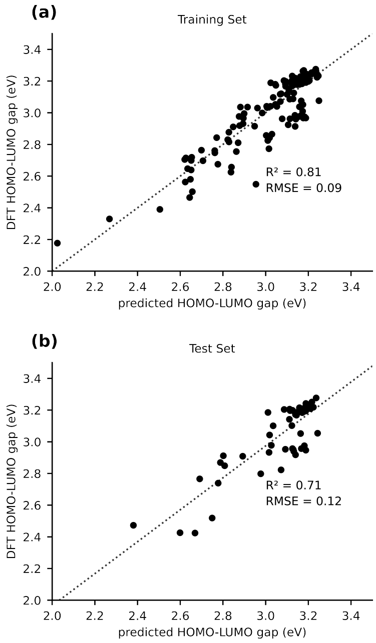


Figure 16: The comparison of the experimental and OLS-predicted HOMO-LUMO gap for (a) the training and (b) test sets

4.3 M2L4 cage

The real versatility of *ChemSpaX* is shown by the automated placement of substituents without introducing steric hindrance on a geometry that is more complex. An M2L4 cage was functionalized at 16 sites with various substituent groups. The M2L4 cage was functionalized on 16 sites using *ChemSpaX*, the results are shown in figure Figure 17. This serial functionalization yielded 16 structures and these were further optimized using GFN-FF and GFN2-xTB(GAS). The RMSDs between geometries generated by workflow of *ChemSpaX* (FF) were calculated against the GFN-FF and GFN2-xTB optimization methods. The results are shown in Table 1. This shows that the GFN-FF and FF geometries are in good agreement an reasonably close to the GFN2-xTB optimized geometries.

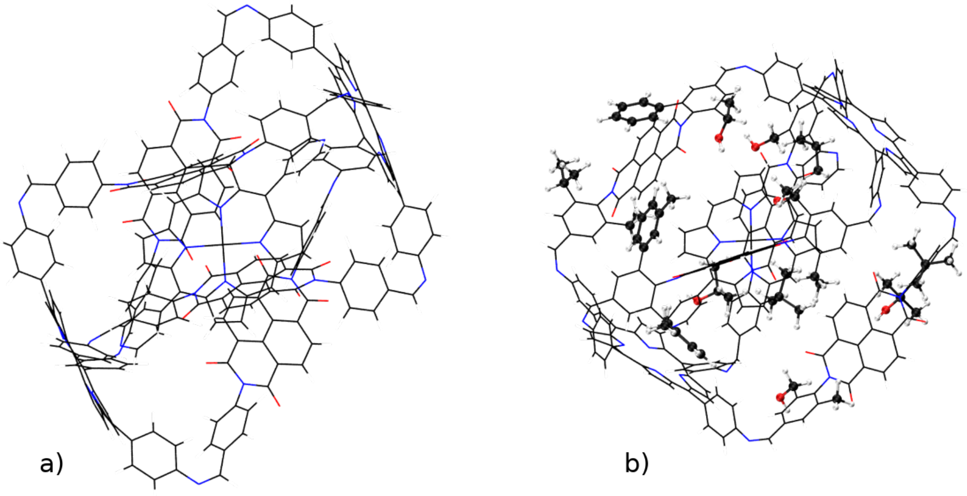


Figure 17: A visualization of the functionalized M2L4 cage which shows a) the input skeleton and b) the GFN2-xTB optimized geometry after placement of 16 substituents. The newly placed substituents are shown in a distinguished representation.

Table 1: Statistics for the RMSD between various methods. The two optimization methods that are compared to each other are shown in the first row.

	GFN-FF v GFN2-xTB	FF v GFN2-xTB	FF v GFN-FF
Average	2.54 Å	2.14 Å	0.83 Å
St. dev.	0.34 Å	0.25 Å	0.26 Å
Max.	3.18 Å	2.46 Å	1.37 Å

5 Summary and conclusions

In this research, an automated Python-based workflow for the exploration of local chemical space is presented. *ChemSpaX* can place substituents on a specific site of many structures based on initial user input and uses FF optimization to optimize newly placed substituents. Use cases were shown by using a data augmented approach which utilized fast GFN2-xTB optimizations to compare structures generated by *ChemSpaX*. For selected use cases a comparison was also done against DFT optimized structures. Descriptors that can be used for high-throughput screening were studied in more detail for some of the presented use cases.

For the pincer complexes a nearly linear scaling of $\Delta\mathrm{E}$ and $\Delta\mathrm{G}$ ($\mathrm{R}^2 = 0.99$ for RuPNP and the Mn-pincers) was found. This correlation was found for both DFT and GFN2-xTB calculations. Due to this correlation, $\Delta\mathrm{G}$ can be replaced by $\Delta\mathrm{E}$ in high-throughput screening, which would lower the consumption of computational resources significantly [81]. The HOMO-LUMO gap calculated by GFN2-xTB was compared against the DFT calculated HOMO-LUMO gap. The correlation ($\mathrm{R}^2 = 0.74$ and RMSE = 0.4 eV) indicated that GFN2-xTB has a reasonable accuracy for calculating the HOMO-LUMO gap and could potentially be used in high-throughput screening. This correlation was not observed for the Mn-pincers.

The investigated Co porphyrins showed a nearly linear increase in hRMSD for serial functionalizations done on the same skeleton. Using this linearity, the optimal moment to employ a higher-level optimization method on a geometry that is being functionalized can be determined. Clustering was observed when comparing the hRMSD against the number of atoms in a structure. Although a higher hRMSD was observed for structures with more atoms (= highly functionalized), the average hRMSD was 1.43 Å with a relatively low standard deviation of 0.53 Å. Additionally, three easily computable features (number of atoms in the structure, hRMSD and GFN2-xTB calculated HOMO-LUMO gap) were used to predict the DFT calculated HOMO-LUMO gap using linear regression via OLS. On the test set a reasonable correlation was found ($\mathrm{R}^2 = 0.71$ and RMSE = 0.12 eV), which again shows that the predictive power of GFN2-xTB is reasonable for high-throughput screening applications. Additionally, it has been presented that structures generated by *ChemSpaX* in this research are reasonably close to GFN2-xTB optimized structures and that all kinds of geometries can be functionalized using *ChemSpaX*.

As shown in this manuscript, *ChemSpaX* can be used in the generation of many structures in the local chemical space with a good quality for high-throughput calculations. The generated structures can be used to generate databases, which can play a role in enabling generative models for material design. There is room for improvement in *ChemSpaX*, which is in the early stages of development. Currently, the user is required to manually check their input geometry for correctness, this includes checking bond lengths/angles. There is work being done on a set of tools that check for correct bond lengths/angles against a quantum chemical database.

With the development of *ChemSpaX*, the door to data-driven catalyst design has been opened in our research group and this will stay an ongoing effort. Together with experimental chemists, work is being done on realizing an automated data-driven workflow where property calculations and/or predictions can be done with the simple click of a button.

Author contributions

The code for ChemSpaX was written by A. V. K. and V. S. DFT & xTB calculations were performed by A. V. K and V. S. Generation of functionalized structures, the compilation of datasets and analysis of DFT & xTB calculations was performed by A. V. K. under supervision of V. S. V. S. and E. A. P. conceived the project. E. A. P. played an advisory role and directed the project. All the authors discussed the results and wrote the manuscript.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgments

V. S. acknowledges the ARC-CBBC project 2016.008 for funding. E.A.P. acknowledges the financial support from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement no. 725686). The authors thank the Netherlands Organization for Scientific Research (NWO) for the access to SURFsara computational facilities and SURF Cooperative for support. The authors thank PRACE for DECI access in the DECI-16 call. The authors thank the PetaChem team for giving access to the TeraChem software package.

References

[1] Carl Poree and Franziska Schoenebeck. A Holy Grail in Chemistry: Computational Catalyst Design: Feasible or Fiction? *Accounts of Chemical Research*, 50(3):605–608, mar 2017. ISSN 0001-4842. doi: 10.1021/acs.accounts.6b00606. URL https://doi.org/10.1021/acs.accounts.6b00606.

[2] John H Van Drie. Computer-aided drug design: the next 20 years. *Journal of Computer-Aided Molecular Design*, 21(10):591–601, 2007. ISSN 1573-4951. doi: 10.1007/s10822-007-9142-y. URL https://doi.org/10.1007/s10822-007-9142-y.

[3] Gregory Sliwoski, Sandeepkumar Kothiwale, Jens Meiler, and Edward W Lowe Jr. Computational methods in drug discovery. *Pharmacological reviews*, 66(1):334–395, dec 2013. ISSN 1521-0081. doi: 10.1124/pr.112.007336. URL https://pubmed.ncbi.nlm.nih.gov/24381236https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3880464/.

[4] David E Clark. What has computer-aided molecular design ever done for drug discovery? *Expert Opinion on Drug Discovery*, 1(2):103–110, jul 2006. ISSN 1746-0441. doi: 10.1517/17460441.1.2.103. URL https://doi.org/10.1517/17460441.1.2.103.

[5] David E Clark. What has virtual screening ever done for drug discovery? *Expert Opinion on Drug Discovery*, 3(8):841–851, aug 2008. ISSN 1746-0441. doi: 10.1517/17460441.3.8.841. URL https://doi.org/10.1517/17460441.3.8.841.

[6] Anubhav Jain, Yongwoo Shin, and Kristin A Persson. Computational predictions of energy materials using density functional theory. *Nature Reviews Materials*, 1(1):15004, 2016. ISSN 2058-8437. doi: 10.1038/natrevmats.2015.4. URL https://doi.org/10.1038/natrevmats.2015.4.

[7] Lei Cheng, Rajeev S Assary, Xiaohui Qu, Anubhav Jain, Shyue Ping Ong, Nav Nidhi Rajput, Kristin Persson, and Larry A Curtiss. Accelerating Electrolyte Discovery for Energy Storage with High-Throughput Screening. *The Journal of Physical Chemistry Letters*, 6(2):283–291, jan 2015. doi: 10.1021/jz502319n. URL https://doi.org/10.1021/jz502319n.

[8] Hossein Beidaghy Dizaji and Hannaneh Hosseini. A review of material screening in pure and mixed-metal oxide thermochemical energy storage (TCES) systems for concentrated solar power (CSP) applications. *Renewable and Sustainable Energy Reviews*, 98:9–26, 2018. ISSN 1364-0321. doi: https://doi.org/10.1016/j.rser.2018.09.004. URL http://www.sciencedirect.com/science/article/pii/S136403211830652X.

[9] Johannes Hachmann, Roberto Olivares-Amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel S Sánchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna M Brockway, and Alán Aspuru-Guzik. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *The Journal of Physical Chemistry Letters*, 2(17):2241–2251, sep 2011. doi: 10.1021/jz200866s. URL https://doi.org/10.1021/jz200866s.

[10] Kirstin Alberi, Marco Buongiorno Nardelli, Andriy Zakutayev, Lubos Mitas, Stefano Curtarolo, Anubhav Jain, Marco Fornari, Nicola Marzari, Ichiro Takeuchi, and Martin L Green. The 2019 materials by design roadmap. *Journal of Physics D: Applied Physics*, 52(1): 13001, 2018. ISSN 0022-3727.

[11] Peter Kirkpatrick and Clare Ellis. Chemical space. *Nature*, 432(7019):823, 2004. ISSN 1476-4687. doi: 10.1038/432823a. URL https://doi.org/10.1038/432823a.

[12] Christopher Zhou, William Grumbles, and Thomas Cundari. Using Machine Learning to Predict the pKa of C–H Bonds. Relevance to Catalytic Methane Functionalization, jul 2020. URL https://chemrxiv.org/articles/preprint/Using_Machine_Learning_to_Predict_the_pKa_of_C_H_Bonds_Relevance_to_Catalytic_Methane_Functionalization/12646772https://chemrxiv.org/ndownloader/files/23820425.

[13] O Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. Exploring chemical compound space with quantum-based machine learning. *Nature Reviews Chemistry*, 4(7):347–358, 2020. ISSN 2397-3358. doi: 10.1038/s41570-020-0189-9. URL https://doi.org/10.1038/s41570-020-0189-9.

[14] Gisbert Schneider. Automating drug discovery. *Nature Reviews Drug Discovery*, 17(2): 97–113, 2018. ISSN 1474-1784. doi: 10.1038/nrd.2017.232. URL https://doi.org/10.1038/nrd.2017.232.

[15] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2):268–276, feb 2018. ISSN 2374-7943. doi: 10.1021/acscentsci.7b00572. URL https://doi.org/10.1021/acscentsci.7b00572.

[16] Thomas Blaschke, Marcus Olivecrona, Ola Engkvist, Jürgen Bajorath, and Hongming Chen. Application of Generative Autoencoder in De Novo Molecular Design. *Molecular Informatics*, 37(1-2):1700123, jan 2018. ISSN 1868-1743. doi: https://doi.org/10.1002/minf.201700123. URL https://doi.org/10.1002/minf.201700123.

[17] Benjamin Meyer, Boodsarin Sawatlon, Stefan Heinen, O Anatole von Lilienfeld, and Clémence Corminboeuf. Machine learning meets volcano plots: computational discovery of

cross-coupling catalysts. *Chemical Science*, 9(35):7069–7077, 2018. ISSN 2041-6520. doi: 10.1039/C8SC01949E. URL http://dx.doi.org/10.1039/C8SC01949E.

[18] Tobias Fink and Jean-Louis Reymond. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *Journal of Chemical Information and Modeling*, 47(2):342–353, mar 2007. ISSN 1549-9596. doi: 10.1021/ci600423u. URL https://doi.org/10.1021/ci600423u.

[19] Lorenz C Blum and Jean-Louis Reymond. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *Journal of the American Chemical Society*, 131(25):8732–8733, jul 2009. ISSN 0002-7863. doi: 10.1021/ja902302h. URL https://doi.org/10.1021/ja902302h.

[20] Josep Arús-Pous, Thomas Blaschke, Silas Ulander, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. Exploring the GDB-13 chemical space using deep generative models. *Journal of Cheminformatics*, 11(1):20, 2019. ISSN 1758-2946. doi: 10.1186/s13321-019-0341-z. URL https://doi.org/10.1186/s13321-019-0341-z.

[21] Pascal Friederich, Gabriel dos Passos Gomes, Riccardo De Bin, Alán Aspuru-Guzik, and David Balcells. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chemical Science*, 11(18):4584–4601, 2020. ISSN 2041-6520. doi: 10.1039/D0SC00445F. URL http://dx.doi.org/10.1039/D0SC00445F.

[22] Adrian Jinich, Benjamin Sanchez-Lengeling, Haniu Ren, Rebecca Harman, and Alán Aspuru-Guzik. A Mixed Quantum Chemistry/Machine Learning Approach for the Fast and Accurate Prediction of Biochemical Redox Potentials and Its Large-Scale Application to 315000 Redox Reactions. *ACS Central Science*, 5(7):1199–1210, jul 2019. ISSN 2374-7943. doi: 10.1021/acscentsci.9b00297. URL https://doi.org/10.1021/acscentsci.9b00297.

[23] Seoin Back, Kevin Tran, and Zachary W Ulissi. Toward a Design of Active Oxygen Evolution Catalysts: Insights from Automated Density Functional Theory Calculations and Machine Learning. *ACS Catalysis*, 9(9):7651–7659, sep 2019. doi: 10.1021/acscatal.9b02416. URL https://doi.org/10.1021/acscatal.9b02416.

[24] Jan Jensen. xyz2mol: Convert Cartesian coordinates to one or more molecular graphs, 2020. URL https://github.com/jensengroup/xyz2mol.

[25] Efthymios I Ioannidis, Terry Z H Gani, and Heather J Kulik. molSimplify: A toolkit for automating discovery in inorganic chemistry. *Journal of Computational Chemistry*, 37(22):2106–2117, 2016. ISSN 1096-987X. doi: 10.1002/jcc.24437. URL http://dx.doi.org/10.1002/jcc.24437.

[26] Jon Paul Janet, Fang Liu, Aditya Nandy, Chenru Duan, Tzuhsiung Yang, Sean Lin, and Heather J Kulik. Designing in the Face of Uncertainty: Exploiting Electronic Structure and Machine Learning Models for Discovery in Inorganic Chemistry. *Inorganic Chemistry*, 58(16):10592–10606, aug 2019. ISSN 0020-1669. doi: 10.1021/acs.inorgchem.9b00109. URL https://doi.org/10.1021/acs.inorgchem.9b00109.

[27] Aditya Nandy, Chenru Duan, Jon Paul Janet, Stefan Gugler, and Heather J Kulik. Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chem-

istry. *Industrial & Engineering Chemistry Research*, 57(42):13973–13986, oct 2018. ISSN 0888-5885. doi: 10.1021/acs.iecr.8b04015. URL https://doi.org/10.1021/acs.iecr.8b04015.

[28] Jon Paul Janet, Terry Z H Gani, Adam H Steeves, Efthymios I Ioannidis, and Heather J Kulik. Leveraging Cheminformatics Strategies for Inorganic Discovery: Application to Redox Potential Design. *Industrial & Engineering Chemistry Research*, 56(17):4898–4910, may 2017. ISSN 0888-5885. doi: 10.1021/acs.iecr.7b00808. URL https://doi.org/10.1021/acs.iecr.7b00808.

[29] Jon Paul Janet, Qing Zhao, Efthymios I Ioannidis, and Heather J Kulik. Density functional theory for modelling large molecular adsorbate–surface interactions: a mini-review and worked example. *Molecular Simulation*, 43(5-6):327–345, apr 2017. ISSN 0892-7022. doi: 10.1080/08927022.2016.1258465. URL https://doi.org/10.1080/08927022.2016.1258465.

[30] Chelsea A Huff and Melanie S Sanford. Catalytic CO2 Hydrogenation to Formate by a Ruthenium Pincer Complex. *ACS Catalysis*, 3(10):2412–2416, oct 2013. doi: 10.1021/cs400609u. URL https://doi.org/10.1021/cs400609u.

[31] Georgy A Filonenko, Robbert van Putten, Erik N Schulpen, Emiel J M Hensen, and Evgeny A Pidko. Highly Efficient Reversible Hydrogenation of Carbon Dioxide to Formates Using a Ruthenium PNP-Pincer Catalyst. *ChemCatChem*, 6(6):1485, jun 2014. ISSN 1867-3880. doi: https://doi.org/10.1002/cctc.201402265. URL https://doi.org/10.1002/cctc.201402265.

[32] Jacob Neumann, Christoph Bornschein, Haijun Jiao, Kathrin Junge, and Matthias Beller. Hydrogenation of Aliphatic and Aromatic Nitriles Using a Defined Ruthenium PNP Pincer Catalyst. *European Journal of Organic Chemistry*, 2015(27):5944–5948, sep 2015. ISSN 1434-193X. doi: https://doi.org/10.1002/ejoc.201501007. URL https://doi.org/10.1002/ejoc.201501007.

[33] Elisabetta Alberico, Alastair J J Lennox, Lydia K Vogt, Haijun Jiao, Wolfgang Baumann, Hans-Joachim Drexler, Martin Nielsen, Anke Spannenberg, Marek P Checinski, Henrik Junge, and Matthias Beller. Unravelling the Mechanism of Basic Aqueous Methanol Dehydrogenation Catalyzed by Ru–PNP Pincer Complexes. *Journal of the American Chemical Society*, 138(45):14890–14904, nov 2016. ISSN 0002-7863. doi: 10.1021/jacs.6b05692. URL https://doi.org/10.1021/jacs.6b05692.

[34] Zhihong Wei, Adiran de Aguirre, Kathrin Junge, Matthias Beller, and Haijun Jiao. Exploring the mechanisms of aqueous methanol dehydrogenation catalyzed by defined PNP Mn and Re pincer complexes under base-free as well as strong base conditions. *Catalysis Science & Technology*, 8(14):3649–3665, 2018. ISSN 2044-4753. doi: 10.1039/C8CY00746B. URL http://dx.doi.org/10.1039/C8CY00746B.

[35] Anastasiya Agapova, Elisabetta Alberico, Anja Kammer, Henrik Junge, and Matthias Beller. Catalytic Dehydrogenation of Formic Acid with Ruthenium-PNP-Pincer Complexes: Comparing N-Methylated and NH-Ligands. *ChemCatChem*, 11(7):1910–1914, apr 2019. ISSN 1867-3880. doi: https://doi.org/10.1002/cctc.201801897. URL https://doi.org/10.1002/cctc.201801897.

[36] Reinaldo Machado, Kevin Heier, and Robert Broekhuis. Developments in Hydrogenation

Technology for Fine-chemical and Pharmaceutical Applications. *Current opinion in drug discovery & development*, 4:745–755, dec 2001.

[37] Werner Bonrath, Jonathan Medlock, Jan Schütz, Bettina Wüstenberg, and Thomas Netscher. Hydrogenation in the Vitamins and Fine Chemicals Industry - An Overview. In *Hydrogenation*, pages 69–90. oct 2012. ISBN 978-953-51-0785-9. doi: 10.5772/48751.

[38] Felix Roessler. Catalysis in the Industrial Production of Pharmaceuticals and Fine Chemicals. *CHIMIA International Journal for Chemistry*, 50(3), 1996. URL https://www.ingentaconnect.com/content/scs/chimia/1996/00000050/00000003/art00012.

[39] Albert Boddien, Christopher Federsel, Peter Sponholz, Dörthe Mellmann, Ralf Jackstell, Henrik Junge, Gabor Laurenczy, and Matthias Beller. Towards the development of a hydrogen battery. *Energy & Environmental Science*, 5(10):8907–8911, 2012. ISSN 1754-5692. doi: 10.1039/C2EE22043A. URL http://dx.doi.org/10.1039/C2EE22043A.

[40] Björn Loges, Albert Boddien, Felix Gärtner, Henrik Junge, and Matthias Beller. Catalytic Generation of Hydrogen from Formic acid and its Derivatives: Useful Hydrogen Storage Materials. *Topics in Catalysis*, 53(13):902–914, 2010. ISSN 1572-9028. doi: 10.1007/s11244-010-9522-8. URL https://doi.org/10.1007/s11244-010-9522-8.

[41] Ferenc Joó. Breakthroughs in Hydrogen Storage—Formic Acid as a Sustainable Storage Material for Hydrogen. *ChemSusChem*, 1(10):805–808, oct 2008. ISSN 1864-5631. doi: 10.1002/cssc.200800133. URL https://doi.org/10.1002/cssc.200800133.

[42] Annika Krieger, Vivek Sinha, Adarsh Kalikadien, and Evgeny A Pidko. Metal-ligand cooperative activation of HX (X=H, Br, OR) bond on Mn based pincercomplexes. *Zeitschrift für anorganische und allgemeine Chemie*, 2021. doi: inpress.

[43] Marcel Garbe, Kathrin Junge, and Matthias Beller. Homogeneous Catalysis by Manganese-Based Pincer Complexes. *European Journal of Organic Chemistry*, 2017(30):4344–4362, aug 2017. ISSN 1434-193X. doi: https://doi.org/10.1002/ejoc.201700376. URL https://doi.org/10.1002/ejoc.201700376.

[44] Vivek Sinha, Nitish Govindarajan, Bas de Bruin, and Evert Jan Meijer. How Solvent Affects C–H Activation and Hydrogen Production Pathways in Homogeneous Ru-Catalyzed Methanol Dehydrogenation Reactions. *ACS Catalysis*, 8(8):6908–6913, aug 2018. doi: 10.1021/acscatal.8b01177. URL https://doi.org/10.1021/acscatal.8b01177.

[45] Monalisa Goswami, Christophe Rebreyend, and Bas de Bruin. Porphyrin Cobalt(III) "Nitrene Radical" Reactivity; Hydrogen Atom Transfer from Ortho-YH Substituents to the Nitrene Moiety of Cobalt-Bound Aryl Nitrene Intermediates (Y = O, NH). *Molecules*, 21(2):242, feb 2016. ISSN 1420-3049. doi: 10.3390/molecules21020242. URL http://dx.doi.org/10.3390/molecules21020242.

[46] Michael P Doyle and David C Forbes. Recent Advances in Asymmetric Catalytic Metal Carbene Transformations. *Chemical Reviews*, 98(2):911–936, apr 1998. ISSN 0009-2665. doi: 10.1021/cr940066a. URL https://doi.org/10.1021/cr940066a.

[47] Simone Fantauzzi, Alessandro Caselli, and Emma Gallo. Nitrene transfer reactions mediated by metallo-porphyrin complexes. *Dalton Transactions*, (28):5434–5443, 2009. ISSN 1477-9226. doi: 10.1039/B902929J. URL http://dx.doi.org/10.1039/B902929J.

[48] Matthias Otte, Petrus F Kuijpers, Oliver Troeppner, Ivana Ivanović-Burmazović, Joost N H Reek, and Bas de Bruin. Encapsulated Cobalt–Porphyrin as a Catalyst for Size-Selective Radical-type Cyclopropanation Reactions. *Chemistry – A European Journal*, 20(17):4880–4884, apr 2014. ISSN 0947-6539. doi: https://doi.org/10.1002/chem.201400055. URL https://doi.org/10.1002/chem.201400055.

[49] Valentinos Mouarrawis, Raoul Plessius, Jarl Ivar van der Vlugt, and Joost N H Reek. Confinement Effects in Catalysis Using Well-Defined Materials and Cages , 2018. URL https://www.frontiersin.org/article/10.3389/fchem.2018.00623.

[50] Matthias Otte, Petrus F Kuijpers, Oliver Troeppner, Ivana Ivanović-Burmazović, Joost N H Reek, and Bas de Bruin. Encapsulation of Metalloporphyrins in a Self-Assembled Cubic M8L6 Cage: A New Molecular Flask for Cobalt–Porphyrin-Catalysed Radical-Type Reactions. *Chemistry – A European Journal*, 19(31):10170–10178, jul 2013. ISSN 0947-6539. doi: https://doi.org/10.1002/chem.201301411. URL https://doi.org/10.1002/chem.201301411.

[51] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. The Open Babel Package, version 2.4.1, 2016. URL https://openbabel.org/.

[52] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, 2011. ISSN 1758-2946. doi: 10.1186/1758-2946-3-33. URL https://doi.org/10.1186/1758-2946-3-33.

[53] A K Rappe, C J Casewit, K S Colwell, W A Goddard, and W M Skiff. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*, 114(25):10024–10035, dec 1992. ISSN 0002-7863. doi: 10.1021/ja00051a040. URL https://doi.org/10.1021/ja00051a040.

[54] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, jul 2004. ISSN 0192-8651. doi: https://doi.org/10.1002/jcc.20035. URL https://doi.org/10.1002/jcc.20035.

[55] Christoph Bannwarth, Eike Caldeweyher, Sebastian Ehlert, Andreas Hansen, Philipp Pracht, Jakob Seibert, Sebastian Spicher, and Stefan Grimme. Extended tight-binding quantum chemistry methods. *WIREs Computational Molecular Science*, n/a(n/a):e01493, aug 2020. ISSN 1759-0876. doi: 10.1002/wcms.1493. URL https://doi.org/10.1002/wcms.1493.

[56] Philipp Pracht, Eike Caldeweyher, Sebastian Ehlert, and Stefan Grimme. A Robust Non-Self-Consistent Tight-Binding Quantum Chemistry Method for large Molecules. *ChemRxiv*, jun 2019. doi: 10.26434/chemrxiv.8326202.v1. URL https://chemrxiv.org/articles/preprint/A_Robust_Non-Self-Consistent_Tight-Binding_Quantum_Chemistry_Method_for_large_Molecules/8326202https://chemrxiv.org/ndownloader/files/15605534.

[57] Stefan Grimme, Christoph Bannwarth, and Philip Shushkov. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements

(Z = 1–86). *Journal of Chemical Theory and Computation*, 13(5):1989–2009, may 2017. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b00118. URL https://doi.org/10.1021/acs.jctc.7b00118.

[58] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671, mar 2019. ISSN 1549-9618. doi: 10.1021/acs.jctc.8b01176. URL https://doi.org/10.1021/acs.jctc.8b01176.

[59] Sebastian Spicher and Stefan Grimme. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. *Angewandte Chemie International Edition*, 59(36):15665–15673, sep 2020. ISSN 1433-7851. doi: https://doi.org/10.1002/anie.202004239. URL https://doi.org/10.1002/anie.202004239.

[60] W Clark Still, Anna Tempczyk, Ronald C Hawley, and Thomas Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society*, 112(16):6127–6129, aug 1990. ISSN 0002-7863. doi: 10.1021/ja00172a038. URL https://doi.org/10.1021/ja00172a038.

[61] T Ooi, M Oobatake, G Némethy, and H A Scheraga. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proceedings of the National Academy of Sciences*, 84(10):3086 LP – 3090, may 1987. doi: 10.1073/pnas.84.10.3086. URL http://www.pnas.org/content/84/10/3086.abstract.

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

[63] A D Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, 38(6):3098–3100, sep 1988. doi: 10.1103/PhysRevA.38.3098. URL https://link.aps.org/doi/10.1103/PhysRevA.38.3098.

[64] Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305, 2005. ISSN 1463-9076. doi: 10.1039/B508541A. URL http://dx.doi.org/10.1039/B508541A.

[65] Kasper P Jensen, Björn O Roos, and Ulf Ryde. Performance of density functionals for first row transition metal systems. *The Journal of Chemical Physics*, 126(1):14103, jan 2007. ISSN 0021-9606. doi: 10.1063/1.2406071. URL https://doi.org/10.1063/1.2406071.

[66] Michael Bühl and Hendrik Kabrede. Geometries of Transition-Metal Complexes from Density-Functional Theory. *Journal of Chemical Theory and Computation*, 2(5):1282–

1290, sep 2006. ISSN 1549-9618. doi: 10.1021/ct6001187. URL https://doi.org/10.1021/ct6001187.

[67] Aleksandr V Marenich, Christopher J Cramer, and Donald G Truhlar. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *The Journal of Physical Chemistry B*, 113(18):6378–6396, may 2009. ISSN 1520-6106. doi: 10.1021/jp810292n. URL https://doi.org/10.1021/jp810292n.

[68] Burkhard Miehlich, Andreas Savin, Hermann Stoll, and Heinzwerner Preuss. Results obtained with the correlation energy density functionals of becke and Lee, Yang and Parr. *Chemical Physics Letters*, 157(3):200–206, 1989. ISSN 0009-2614. doi: https://doi.org/10.1016/0009-2614(89)87234-3. URL http://www.sciencedirect.com/science/article/pii/0009261489872343.

[69] Eike Caldeweyher, Sebastian Ehlert, Andreas Hansen, Hagen Neugebauer, Sebastian Spicher, Christoph Bannwarth, and Stefan Grimme. A generally applicable atomic-charge dependent London dispersion correction. *The Journal of Chemical Physics*, 150(15):154122, apr 2019. ISSN 0021-9606. doi: 10.1063/1.5090222. URL https://doi.org/10.1063/1.5090222.

[70] Ivan S Ufimtsev and Todd J Martínez. Quantum Chemistry on Graphical Processing Units. 1. Strategies for Two-Electron Integral Evaluation. *Journal of Chemical Theory and Computation*, 4(2):222–231, feb 2008. ISSN 1549-9618. doi: 10.1021/ct700268q. URL https://doi.org/10.1021/ct700268q.

[71] Ivan S Ufimtsev and Todd J Martinez. Quantum Chemistry on Graphical Processing Units. 2. Direct Self-Consistent-Field Implementation. *Journal of Chemical Theory and Computation*, 5(4):1004–1015, apr 2009. ISSN 1549-9618. doi: 10.1021/ct800526s. URL https://doi.org/10.1021/ct800526s.

[72] Ivan S Ufimtsev and Todd J Martinez. Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics. *Journal of Chemical Theory and Computation*, 5(10):2619–2628, oct 2009. ISSN 1549-9618. doi: 10.1021/ct9003004. URL https://doi.org/10.1021/ct9003004.

[73] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics*, 132(15):154104, apr 2010. ISSN 0021-9606. doi: 10.1063/1.3382344. URL https://doi.org/10.1063/1.3382344.

[74] P Jeffrey Hay and Willard R Wadt. Ab initio effective core potentials for molecular calculations. Potentials for the transition metal atoms Sc to Hg. *The Journal of Chemical Physics*, 82(1):270–283, jan 1985. ISSN 0021-9606. doi: 10.1063/1.448799. URL https://doi.org/10.1063/1.448799.

[75] Jimmy Charnley Kromann. Calculate Root-mean-square deviation (RMSD) of Two Molecules Using Rotation, GitHub, v1.3.2, 2020. URL https://github.com/charnley/rmsd/releases/tag/rmsd-1.3.2.

[76] W Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, sep 1976. ISSN 0567-7394. URL https://doi.org/10.1107/S0567739476001873.

[77] Michael W Walker, Lejun Shao, and Richard A Volz. Estimating 3-D location parameters using dual number quaternions. *CVGIP: Image Understanding*, 54(3):358–367, 1991. ISSN 1049-9660. doi: https://doi.org/10.1016/1049-9660(91)90036-O. URL http://www.sciencedirect.com/science/article/pii/104996609190036O.

[78] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12(null):2825–2830, nov 2011. ISSN 1532-4435.

[79] Gabriel dos Passos Gomes, Robert Pollice, and Alán Aspuru-Guzik. Navigating through the Maze of Homogeneous Catalyst Design with Machine Learning. *Trends in Chemistry*, 3 (2):96–110, 2021. ISSN 2589-5974. doi: https://doi.org/10.1016/j.trechm.2020.12.006. URL https://www.sciencedirect.com/science/article/pii/S2589597420303166.

[80] Brian K Shoichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, 2004. ISSN 1476-4687. doi: 10.1038/nature03197. URL https://doi.org/10.1038/nature03197.

[81] Vivek Sinha, Jochem Jan Laan, and Evgeny Pidko. Accurate and Rapid Prediction of pKa of Transition Metal Complexes: Semiempirical Quantum Chemistry with a DataAugmented Approach, oct 2020. URL https://chemrxiv.org/articles/preprint/Accurate_and_Rapid_Prediction_of_pKa_of_Transition_Metal_Complexes_Semiempirical_Quantum_Chemistry_with_a_DataAugmented_Approach/13087463https://chemrxiv.org/ndownloader/files/25048262.

[82] Xijun Wang, Guozhen Zhang, Li Yang, Edward Sharman, and Jun Jiang. Material descriptors for photocatalyst/catalyst design. *WIREs Computational Molecular Science*, 8 (5):e1369, sep 2018. ISSN 1759-0876. doi: https://doi.org/10.1002/wcms.1369. URL https://doi.org/10.1002/wcms.1369.

[83] Albert Polman, Mark Knight, Erik C Garnett, Bruno Ehrler, and Wim C Sinke. Photovoltaic materials: Present efficiencies and future challenges. *Science*, 352(6283):aad4424, apr 2016. doi: 10.1126/science.aad4424. URL http://science.sciencemag.org/content/352/6283/aad4424.abstract.

[84] SeJin Ahn, Sunghun Jung, Jihye Gwak, Ara Cho, Keeshik Shin, Kyunghoon Yoon, Doyoung Park, Hyeonsik Cheong, and Jae Ho Yun. Determination of band gap energy (Eg) of Cu2ZnSnSe4 thin films: On the discrepancies of reported band gap values. *Applied Physics Letters*, 97(2):21905, jul 2010. ISSN 0003-6951. doi: 10.1063/1.3457172. URL https://doi.org/10.1063/1.3457172.

[85] J S Griffith and L E Orgel. Ligand-field theory. *Quarterly Reviews, Chemical Society*, 11(4): 381–393, 1957. ISSN 0009-2681. doi: 10.1039/QR9571100381. URL http://dx.doi.org/10.1039/QR9571100381.

[86] Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *The Journal of Physical Chemistry Letters*, 9(7):1668–1673, apr 2018. doi: 10.1021/acs.jpclett.8b00124. URL https://doi.org/10.1021/acs.jpclett.8b00124.

[87] Volodymyr Lyaskovskyy, Alma I Olivos Suarez, Hongjian Lu, Huiling Jiang, X Peter Zhang, and Bas de Bruin. Mechanism of Cobalt(II) Porphyrin-Catalyzed C–H Amination with Organic Azides: Radical Nature and H-Atom Abstraction Ability of the Key Cobalt(III)–Nitrene Intermediates. *Journal of the American Chemical Society*, 133(31):12264–12273, aug 2011. ISSN 0002-7863. doi: 10.1021/ja204800a. URL https://doi.org/10.1021/ja204800a.

[88] Shifa Zhu, Jason A. Perman, and X. Peter Zhang. Acceptor/Acceptor-Substituted Diazo Reagents for Carbene Transfers: Cobalt-Catalyzed Asymmetric Z-Cyclopropanation of Alkenes with α-Nitrodiazoacetates. *Angewandte Chemie International Edition*, 47(44): 8460–8463, oct 2008. ISSN 1433-7851. doi: https://doi.org/10.1002/anie.200803857. URL https://doi.org/10.1002/anie.200803857.

[89] Kamil Godula and Dalibor Sames. C-H Bond Functionalization in Complex Organic Synthesis. *Science*, 312(5770):67 LP – 72, apr 2006. doi: 10.1126/science.1114731. URL http://science.sciencemag.org/content/312/5770/67.abstract.

[90] Jay A Labinger and John E Bercaw. Understanding and exploiting C–H bond activation. *Nature*, 417(6888):507–514, 2002. ISSN 1476-4687. doi: 10.1038/417507a. URL https://doi.org/10.1038/417507a.

Metal-ligand cooperative activation of HX (X=H, Br, OR) bond on Mn based pincer complexes

Annika M. Krieger [a], Vivek Sinha *[a], Adarsh V. Kalikadien [a], Evgeny A. Pidko*[a]

Abstract: Reversible dissociation of H-X bond (M-L + H-X → M(X)-L(H); ΔG_{HX}) is an important step during pre-activation, catalysis and possible deactivation of acid-base cooperative pincer based transition metal catalysts (M-L). Herein we carried out a high-throughput computational investigation of the thermodynamic stability of different adducts in various functionalized Mn(I) based pincer complexes. We used a combination of density functional theory (DFT) and density functional tight binding (DFTB) calculations to analyze ΔG_{HX} of > 700 (M(X)-L(H)) intermediates based on functionalized variants of four pincer type ligand scaffolds derived from PCP, CNC, PNP and SNS ligands. We discovered linear scaling relations between ΔG_{HX} of various species. Strongest correlations were found between species of similar size and chemical nature e.g. ΔG_{tBuOH} correlated best with ΔG_{iPrOH} and worst with ΔG_{HBr}. Such scaling relations can be useful for property based screening of catalysts and selection of (co)solvent/substrate/base for optimized reaction conditions. We also investigated the influence of the ligand backbone and the functionalization of donor and backbone sites in the ligand. Our analysis reveals the crucial role of the second coordination sphere functionalization for the reactivity of the complexes with impact in some cases exceeding that of the variation of the functional groups directly attached to the donor atoms.

Introduction

Pincer complexes are important catalysts in organometallic chemistry for multiple applications such as transformation and synthesis of imines, amines, peptides, pyridines, pyrroles, acetals, and carboxylic acid derivatives, such as esters, ketones and amides.[1–3] Owing to their success with (de)hydrogenation of a wide scope of substrates, pincer complexes have been adopted favorably by the pharmaceutical, fine chemicals and the energy industry[4] (representative examples shown in Figure 1a,b). The well-defined geometry and tridentate coordination mode of this class of complexes offers a stable catalytic structure. Most highly active pincers such as Ru-MACHO complex[5,6] and Nozaki's Ir-PNP complex[7] are based on expensive Ru and Ir metals. Catalytic systems based on such metals are not desirable for large scale ubiquitous applications due to high cost and limited availability. Several successful examples of pincer catalysts based on earth abundant 3d transition metals (TMs) such as Fe and Mn have been realized in the last decade.[8,9] However, the activity and stability of such catalysts based on first row TMs remains a challenge. Therefore, the development and optimization of catalysts based on 3d transition metals is an active and highly sought after area of research.[10–12] Manganese is particularly attractive as the active metal in such catalysts in view of its high biocompatibility, which is of interest for industries in the food or pharmaceutical sector.

Functionalization of the ligand scaffold can be used to explore the chemical space of TM pincers in the pursuit of highly active and stable catalysts based on first row TMs. In such an approach one can start with a "skeleton" complex bearing a TM center coordinated to a pincer scaffold. Selected sites on the scaffold or metal center can be functionalized generating an ensemble of new TM complexes using various combinations of functional groups.[13,14] Experimentally only a handful of functionalized variants of pincer ligand scaffolds have been reported. Moreover, synthesis and subsequent testing of the catalytic activity of functionalized TM complexes quickly becomes intractable. Theoretical consideration of the functionalized variants that are synthetically not accessible can provide an insight into rational design principles. In this regard, computational methods are relevant and can be applied to screen through a large ensembles of functionalized TM complexes.[15,16] Recently such approaches have been applied to screen TM complex including pincer complexes for activity, regioselectivity and ligand effects.[17–19]

Herein we screened the effect of functionalization of the ligand backbone on the stability of potential catalytic intermediates on Mn(I)-pincer complexes. The focus was on the determination the effect of the type of functionalization and the functionalization site (backbone/donor site). We chose five representative pincer ligand scaffolds, namely, PNP- (bis(3-phosphaneylpropyl)amine)-, SNS- (azanediylbis(ethane-1-thiol))-, CNC- (bis(2-(1H-3λ^4-imidazol-3-yl)ethyl)amine)-, PNN- (*N*^{1}-(2-phosphaneylethyl)ethane-1,2-diamine)-, and PCP- (*N*^{1},*N*^{3}-bis(phosphaneyl)benzene-1,3-diamine)- backbones coordinated to a Mn(I) center stabilized by CO ligands as illustrated in Figure 1c. Our analysis included the pristine complexes as well as their catalytically relevant intermediates resulting in over ~1200 structures based on the five selected pincer ligand scaffolds (Figure 1c).

[a] Dr. V. Sinha, Mrs. A. Krieger, Mr. A. V. Kalikadien, Prof. Dr. E. A. Pidko
Inorganic Systems Engineering, Department of chemical engineering, Faculty of Applied Sciences, Delft University of Technology, van der Maasweg 9, 2629 HZ, Delft, The Netherlands.
V.Sinha@tudelft.nl; E.A.Pidko@tudelft.nl

Supporting information for this article is given via a link at the end of the document.((Please delete this text if not appropriate))

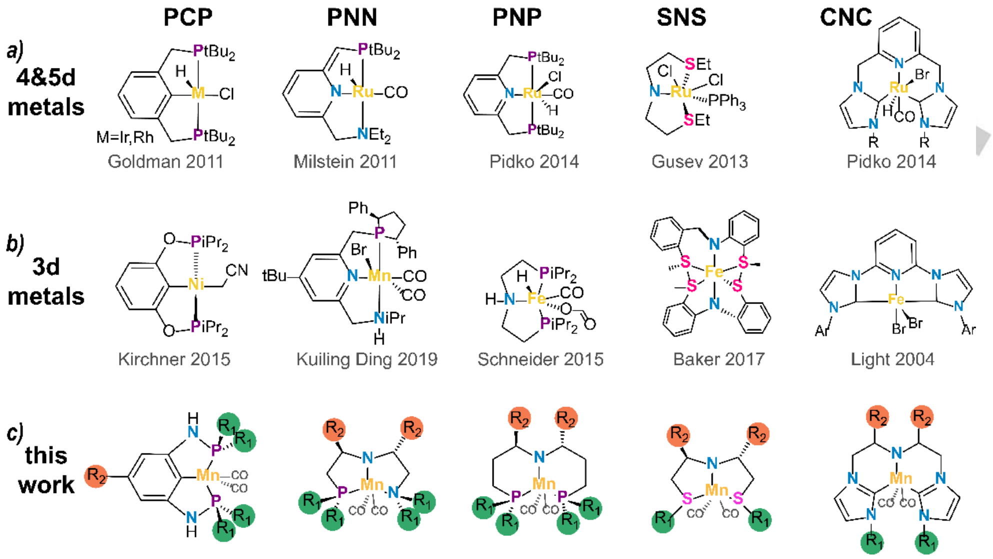


Figure 1. Representative transition metal pincer complexes of (a) 4d, 5d and (b) 3d metals, followed by (c) the catalyst scope of this work.[1]

Pincer complexes based on these ligands have been reported for various transition metals, including manganese (Figure 1b).[8,13,20–29] SNS-,[30–35] CNC-,[36–41] and PCP-ligands[42–46] are primarily known for their use in 4d and 5d transition metal catalysis (Figure 1a). Especially in 4d and 5d-transition metal catalysis high turnover frequencies and turnover numbers are reported for the complexes.[21,38,40,47]

Literature on 3d-transition metals generally reports lower catalytic efficiencies,[20,46] which indicates that there is opportunity to maximize their potential towards sustainable catalyst systems. Catalysts based on 3d metals are known to be more prone to deactivation and formation of resting states, limiting their reactivity.[4,48]

Herein, we investigated Mn(I) -pincers as potential (de)hydrogenation catalysts. Possible applications are the storage of H_2 in unsaturated moieties such as CO_2, the reduction of organic substrates with H_2 gas or with hydrogen donors such as ^iPrOH.[49–51] Figure 2 illustrates representative catalytic cycles of dehydrogenation of methanol to acetone and hydrogenation of acetone to isopropanol. The Br adduct (M(Br) – L(H)) is a common precursor to the active form of Mn-pincers.[11] The activation of the catalyst is commonly carried out by the reaction with a strong base (e.g. KOH or KO^{t}Bu). The activated catalyst (M – L) features a Lewis acid site on the metal and the ligand can act as a Bronsted base i.e. the metal can coordinate with an electron donating species while the ligand can accept a H^+. The activated catalyst is susceptible for potential deactivation/inhibition through the metal-ligand cooperative addition of alcohol/water/base resulting in the formation of -OR adducts.[52] Alkoxide adduct of hydrogen donating alcohols such ^iPrOH, MeOH and EtOH are often formed as intermediates in the course of catalytic hydrogenation reactions, and can even act as the resting states limiting the catalytic performance depending on their stability.[48,53–58] Competitive bonding of other species such as the solvent or the nucleophile base to the metal can slow down or even deactivate the catalyst.

For example, water can compete with methanol for the catalytically active site via the formation of a stable hydroxide adduct upon reaction with the catalyst or via ligand exchange with the methoxide adduct (Figure 2). To continue the catalytic cycle, the alkoxide adduct must convert to the hydride adduct (M(H)-L(H)), which in turn regenerates the catalyst by hydrogen evolution/transfer. Catalytic turnover is inhibited if the alkoxide adducts is very stable compared to the hydride adduct. Similarly, an excessively stable hydride adduct would render it inactive towards hydrogen liberation resulting in an adverse effect on the catalysis.

When employed as a hydrogenation catalysis, the hydride adduct is formed as the first step via heterolytic H_2 dissociation. A less stable hydride would be prone to H_2 recombination instead of catalytic turnover. On the other hand, excessive hydride stability would make it less reactive towards the hydride transfer steps of the catalytic cycle. The alkoxide adduct formed upon the hydride transfer to C=O must dissociate to regenerate the catalyst.

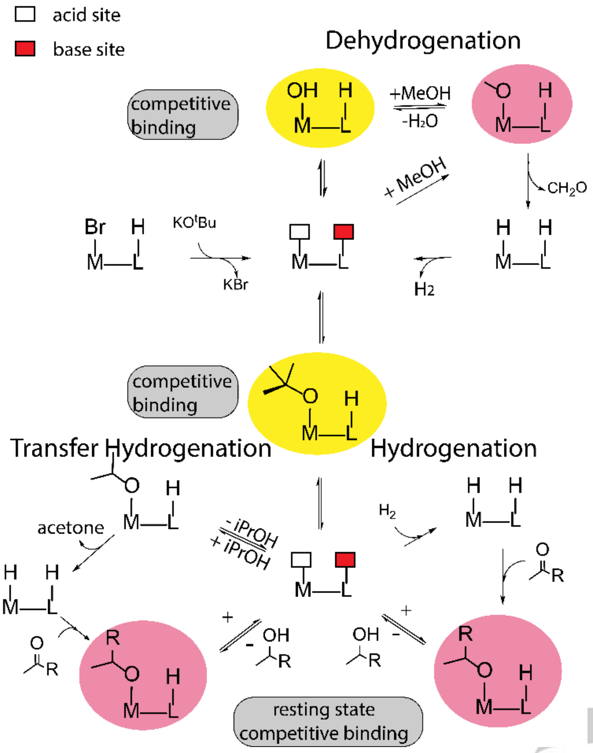


Figure 2. Representative catalytic cycles for dehydrogenation and hydrogenation reaction with possible competing and deactivation pathways. Dehydrogenation of methanol in aqueous phase and hydrogenation of acetone to iPrOH have been used as representative examples. Possible alkoxides and hydroxide based competing and resting states have been highlighted in yellow and red respectively

Transfer hydrogenation reactions with e.g. isopropanol often proceeds via the formation of an intermediate alkoxide species. Excessive stability of this alkoxide adduct would have a deactivating effect and adversely affect the catalytic turnover. From a mechanistic perspective the relative stability of hydride, hydroxide and alkoxides intermediates are important for the catalytic turnover. Their relative stability, in accordance with the Sabatier's principle should be balanced and any excessive stabilization/destabilization would have an inhibiting effect on the catalyst.

In the present work we investigate functional strategies to tune the stability of aforementioned intermediates. Our results allow comparison of different ligand scaffolds and functionalization strategies in a common framework. We have identified linear free energy scaling relations (LFESRs) between various intermediates. By analyzing relative stabilities of various adducts we analyze their competitive binding at the metal center. We describe the impact of functionalization of the complexes near the metal center and on the ligand backbone. Finally, we draw conclusions about catalyst behavior and formulate perspectives on catalyst activation strategies, choice of solvent environment and possible deactivation species. The paper is organized as follows: in the computational methods section we describe the functionalization approach, computational model to investigate the thermodynamic parameters, and details of the quantum chemical calculations applied. Next we describe and discuss the results from our calculations. Finally we summarize the results and present our conclusions.

Computational methods

Electronic structure calculations on the transition metal complexes were carried out either by using the extended density functional tight-binding (DFTB) or the density functional theory (DFT) methods. For a large number of calculations, a full DFT based approach is computationally expensive. Extended DFTB calculations via the xTB code from Grimme's group has recently emerged as a rapid tool with reasonable accuracy to predict geometry and thermochemistry of various chemical systems including TM complexes.[59] We therefore performed xTB calculations on all the complexes in our paper. On a selected number of TM complexes (432) based on CNC, PCP and PNN ligands, we also performed DFT calculations. xTB calculations on the same 432 complexes were compared with DFT based predictions to determine the accuracy of xTB.

Extended density functional tight-binding calculations

Extended DFTB calculations were performed using the xTB software suite (version 6.3.3).[60,61] The GFN2-xTB method was applied for geometry optimization, using the *verytight* criteria. Hessian matrix calculations were performed for all optimized geometries to verify the absence of imaginary frequencies and that each geometry corresponds to a local minimum on its respective potential energy surface (PES). The GBSA solvation model parametrized for THF as implemented in xTB was used to account for solvent effects.

Density functional theory calculations

Density functional theory (DFT) calculations were performed using the Gaussian 16 C.01 suite of software.[62] All geometries were optimized using the BP86 functional with a def2-SVP basis set in the gas phase.[63] The combination of functional and basis set have shown reliable geometry predictions accompanied with low computational costs.[64,65] Hessian calculations were performed to ensure that all optimized geometries were a minima on the PES (no imaginary mode). Zero-point energy and Gibbs free energy corrections to electronic energy were obtained from hessian calculations within the harmonic approximation under standard conditions (298.15K, 1bar). Single point (SP) energy calculations with the SMD[66] solvation (THF) model were carried out using the PBE0 hybrid functionals[67] with a triple zeta basis set (def2-TZVP) to further refine the electronic energies. SPs were also carried out using the BP86 functional.[68] We denote such composite methods BP86/def2-SVP//XC/def2-TZVP (THF), as bp86(thf) or pbe0(thf) depending on the exchange-correlation (XC) functional used for the SP single-point calculations. This allowed to investigate the impact of solvation (bp86(gas) vs bp86(THF)) and the functional (bp86(THF) vs pbe0(THF)) on the

WILEY-VCH

computed free energies. All DFT calculations were performed with dispersion correction (D3).[69]

DFT calculations were performed for a selected number of complexes bearing CNC, PNN and PCP ligands. We focused our investigation on the addition of H-X species (X=H, OH, MeO, EtO, ^{i}PrO, ^{t}BuO, Br) across the catalyst which is represented as M – L, where M represents the metal center and L represents the ligand. Addition of H – X across M – L leads to formation of M(X)- L(H) species, where metal forms an adduct with X and the ligand gets protonated. We estimated the thermodynamic stability of M(X) – L(H) by computing the Gibbs free energy change under standard conditions upon addition of H-X moiety across M-L bond (eq. (1)).

$$\boldsymbol{H - X + M - L \rightarrow M(X) - L(H)}$$

$$\boldsymbol{\Delta G^{\circ}_{HX} = G(M(X) - L(H) - G(M - L) - G(H - X)} \quad \dots \boldsymbol{(1)}$$

Results and Discussion

Functionalization strategy

All functionalized geometries were obtained via an *in-house* developed automated python based workflow.[70] We chose two different functionalization sites: four R_1 sites which are located near the metal center, and two R_2 sites (only one in case of PCP) which functionalize the ligand backbone as shown in Figure 1. We chose to perform symmetric functionalizations meaning all four R_1 sites were kept the same, and both R_2 sites were also functionalized with the same ligand. R_1 and R_2 were however not constrained to be the same. In addition to functionalization sites on the ligand, seven Mn-adducts were also considered which included vacant site (pristine complex), H, Br, MeO, iPrO, EtO and tBuO adducts. This functionalization scheme generated ~1200 geometries of metal complexes. Out of these 1225 geometries, we filtered geometries where the pristine complex had a hemi-labile ligand resulting in a total of 732 geometries which are discussed in this work. We found that most PNN based complexes resulted in hemi-labile ligand. Hemi-lability can arise as an artifact of xTB based geometry optimization or it can be genuinely present in the system. Since this would require further investigation we excluded all xTB based results for the PNN catalyst.

Comparison of xTB with DFT

Low computational cost and wide applicability of xTB calculations make them suitable for high throughput screening of TM complexes. The accuracy of xTB calculations has not been tested for Mn complexes. To investigate the accuracy of xTB calculations with respect to DFT based results, we computed CO stretching freqnecies and $\Delta G_{HX}$for selected complexes. Figure 3 compares xTB and DFT results for addition of HBr to Mn-PCP, Mn-PNN and Mn-CNC complexes. DFT and xTB computed ΔG_{HBr} agree well with R^2 = 0.82 and a RMSE (based on the linear fit; see SI) of 7.19 kcal mol^{-1}. The correlation coefficient between xTB and DFT computed ΔG_{iPrOH} is relatively poor (R^2 = 0.28; RMSE = 10.48 kcal mol^{-1}) and the two methods reach only qualitative agreement (see SI) for the addition of ^{i}PrOH.

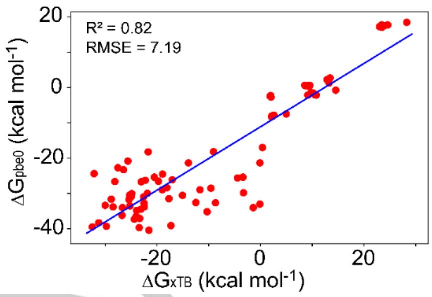


Figure 3. Correlation between Gibbs free energies for addition of HBr computed using xTB (x-axis) and DFT (y-axis).

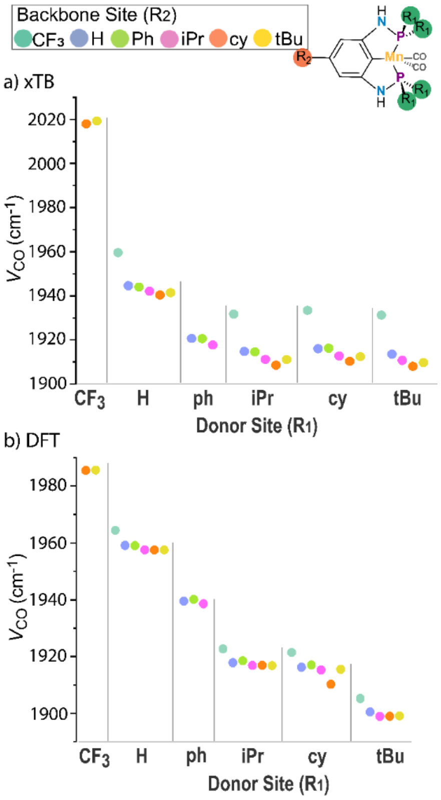


Figure 4. a) xTB and b) DFT computed *v*(CO) for Mn-PCP complexes sorted by electron donating properties of the functionalization group.

[70] The geometries of functionalized complexes were generated using an in-house developed automated python based workflow named "ChemSpaX". The related manuscript is currently in preparation.

To further compare the performance of xTB and DFT, we analyzed the computed CO stretching frequencies (*v*(CO)) for the carbonyl moieties in the Mn pincers. A comparison of the results obtained with the DFT and xTB methods for a representative case of Mn-PCP complex is presented in Figure 4. Both methodologies reveal a similar trend in computed *v*(CO). For a given R_1, electron donating R_2 groups give rise to lower *v*(CO). The plots also show that the electron donating effect of R_1 functionalziation in this case is more important that that of the R_2, because of the major role of the electronic effects at the metal center on the coordinated CO ligands. Furthermore, we observed a good agreement between t xTB and DFT results when all systems with CF_3 functionalization are excluded from the dataset (R^2=0.85). xTB calculations seem to overestimate the CO stretching for CF_3 functionalized ligands for all pincer complexes considered in this study (see SI). Nevertheless, the comparison of ΔG_{HX} and *v*(CO) parameters point to a qualitative agreemenet between the results obtained with the xTB and DFT methods.

Scaling Relations and competitive adduct formation

The activation of H – X bonds is assumed to proceed via a metal-ligand cooperative heterolytic cleavage over the M–L site in all of the pincer complexes discussed here. We therefore, expect similar trends in Gibbs free energy for addition of H – X species, with differences arising from the nature of M-X bonding. Such similarities practically manifests themselves in scaling or linear free energy relationships between different substrates. Such relationships imply that having computed the ΔG_{HX_1} for a substrate HX_1, one can estimate the ΔG_{HX_i} for all other species that follow a linear scaling relation with HX_1. Figure 5 shows a significant correlation between the DFT-computed ΔG_{HBr} and ΔG_{H2} using DFT (R^2=0.91), especially for PCP and CNC complexes. Therefore, the Gibbs free energy of the bromide adduct formation can be used to estimate the relative energy of formation of the hydride species.

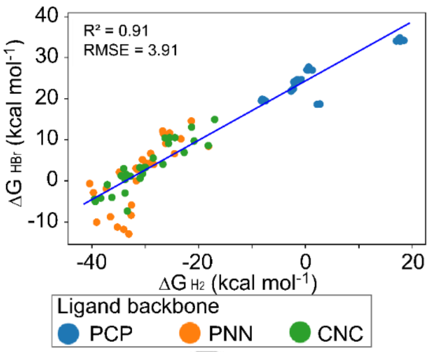


Figure 5. Comparison of xTB and DFT computed Gibbs free energy of formation of bromide and hydride adducts in Mn-PCP, Mn-PNN and Mn-CNC complexes.

In our experience metal hydrides complexes are not described well using xTB.[71] Often the M-H bond was found to be very elongated (> 2.8 Å) (see SI) in the structures obtained by the xTB-based geometry optimizations. However, Br complexes are described well and reasonable geometries were obtained with both xTB and DFT optimizations. Therefore, the scaling relationship that we observed between the bromide and hydride adduct becomes particularly practical because the ΔG_{HBr} computed using xTB can be directly used to estimate the stability of the active hydride intermediates in screening studies.

Figure 5 also shows that while both PCP and PNN complexes react with H_2 in an exergonic manner, the reaction is endergonic with Mn-CNC complexes. DFT calculations also revealed linear scaling relations among hydride, iso-propoxide, bromide and hydroxide adducts. The correlation of ΔG_{H_2} with OH and ^{i}PrO adduct was found to be rather weak (R^2 = 0.53 and 0.55) (see SI). We attribute this weaker correlation to differences in M – X bonds formed upon addition of H – X (M – OR vs. M – H) (*vide infra*). ΔG_{H_2O} and ΔG_{iPrOH} were found to have a moderate correlation with R^2 = 0.71. We also investigated the correlation coefficient between the xTB-computed ΔG_{HX}values. The resulting correlation matrix is shown in Figure 6. We observe that Gibbs free energies for addition of chemically similar species have higher correlation coefficient. For example, all alkoxides correlate well among each other but have relatively poor correlation with hydrides and bromides. Therefore, isopropoxide and ethoxide have the strongest correlation followed by the correlation between methoxide and ethoxide. The weakest correlation is between the Mn-alkoxides and the catalyst precursor (Mn-Br).

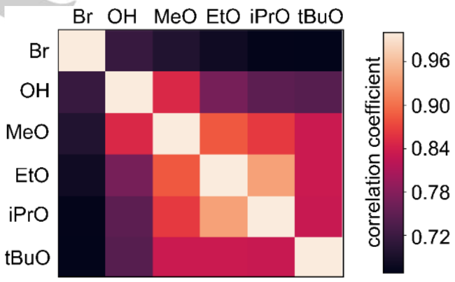


Figure 6. Correlation matrix of Gibbs free energy of formation in all ligand backbones investigated by xTB of the different metal adducts

Adduct stability modulations via ligand modifications

Next we investigated the impact of the functionalization of donor site vs backbone on ΔG_{HX}. For this purpose we examined ΔG_{HBr} in two scenarios: 1) R_2 = H with R_1 varied to investigate the impact of functionalization at the donor site 2) R_1 = H with R_2 varied to investigate the impact of functionalization on the ligand backbone (Figure 4). Functionalizations on the ligand either on donor or acceptor site can have a stabilizing or a destabilizing effect in two ways: 1) via (de)stabilization of the pristine catalyst 2) Electronic and/or steric (de)stabilization of the adduct moieties/. The pristine complex can be (de)stabilized via geometric distortions and the resulting strain in the ligand scaffold introduced by functionalization. The adduct moiety such as an alkoxides can be destabilized by increased electronic density at the metal center. On the other hand, the L-H bond in the adducts can be favorably stabilized by an increased electronic density at the ligand. Larger alkyl groups such as tBu/cy can also destabilize bulky alkoxides by steric repulsion. An accurate consideration of electronic effects is necessary to examine the impact of functionalization. Therefore we examined the impact of functionalization of ΔG_{HX} using DFT calculations.

Functionalization on donor site R_2 = H resulted in five PCP (R1 = H, Ph, ^{i}Pr, cy, ^{t}Bu), four CNC (R_1 = CF_3, Ph, ^{i}Pr and cy) and three PNN (R_1 = Ph, iPr and tBu) ligand complexes. The resulting ΔG_{HX} are presented in Table-S5. For both PCP and PNN complexes, electron donating groups at R_1 destabilize the adduct leading to higher (more positive) ΔG_{HX}. In contrast to PCP and PNN complexes, the CNC complexes exhibit a different trend. The electron withdrawing CF_3 groups also destabilize Br and OH adducts. A detailed analysis of this divergent trend is beyond the scope of this paper. We speculate that these differences are related to the nature of Mn-C coordination in the CNC complexes. Furthermore, the R_1 functionalizations in this case are not performed on the C moiety coordinating the Mn center unlike for the other complexes where the functionalized P/N/S are directly bound with the metal site.

Functionalization on ligand backbone

Primarily two factors control the impact of functionalization on the ligand backbone: 1) electronic effect which influences the basicity of the acceptor site and 2) strain in the geometry upon functionalization. Electron releasing groups should increase the basicity of the acceptor site and, therefore, are expected to have a stabilizing effect. At the same, since electron releasing groups such as a ^{t}Bu are bulkier they are also expected to introduce a high amount of strain in the coordinated ligand geometry. R1 = ^{i}Pr resulted in 6 CPC and PNN, and 5 PCP complexes with Br coordinated to the Mn. The resulting ΔG_{HBr} are presented in Table S7.

Variation of R_2 seems to have minimal impact on the PCP scaffold where ΔG_{HBr} shows little variation with the R_2 = H being the most stable complex, and R_2 = CF_3 being the least stable. This behavior is expected since the R_2 site is located further away from the proton acceptor site on the ligand. For both Mn-PNN and Mn-CNC complexes, R_2 = cy results in most favorable adduct formation. R_2 = CF_3 leads to most destabilized adduct for the Mn-CNC, whereas R_2 = ^{t}Bu is most destabilizing functionalization for Mn-PNN. The additional stabilization of *cy* substitution is in contrast with the destabilization introduced by other electron donating groups namely Ph, ^{i}Pr and ^{t}Bu. The Hamett constants of cy and ^{i}Pr are -0.05 and -0.04, respectively, indicating similar electron releasing behavior via inductive effect.[72] We attribute the observed destabilizing effect of Ph, ^{i}Pr and ^{t}Bu groups, despite their electron donating nature, to the geometric strain. Analysis of geometric strain showed that R_2 = cy scaffold had an effective stabilization of 4.15 kcal mol^{-1} relative to R_2 = H. In contrast while R_2 = ^{i}Pr suffered a destabilization of 1.37 kcal mol^{-1} (Table S5-S6). Therefore both steric and inductive effects play an important role in stabilization of the Br adducts.

Heatmap plots for the functionalization on donor and ligand backbone sites respectively represent the average impact of the choice of the functionalization groups (Figure S9). Both donor and backbone site functionalization leads to a great spread in average Gibbs free energy of formation, which can vary by up to 20 kcal mol^{-1} between different functionalizations. This reflects the degree of tunability of the catalytic properties that these ligand scaffolds can offer. Especially interesting are the qualitative changes in the average stability, where for example the substitution of CF_3 on the backbone site can turn the adduct formation from exergonic to endergonic. The relative range and standard deviation does not change significantly within the data sets.

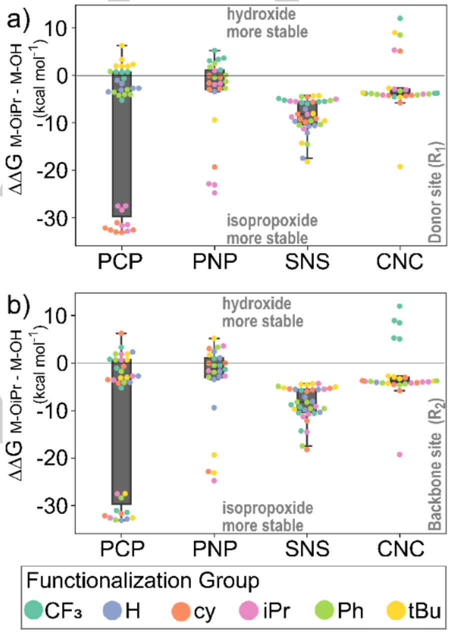


Figure 7. Comparison of isopropxide stabilities with hydroxide per catalytic ligand group. Swarmplots are added indicaton the functionalization on a) catalyst donor site R_1 and b) catalyst backbone site R_2.

Catalytic adduct species

With regards to the catalytic cycle, different adduct stabilities that play a significant role in activity and efficiency are examined. The heatmaps of the average formation Gibbs free energies shows the difference between the investigated adducts (Figure S9). The results indicate that the bromide adduct sustains substantial stability upon functionalization, especially for CNC complexes. The stabilities of the other adducts with similar structure such as MeO and EtO, or ^{t}BuO and ^{i}PrO are quite similar.

Boxplots comparing the difference formation Gibbs free energies for the catalysts with different ligands are shown in Figure 7 and 8 allow to analyze the trends in more detail. One interesting aspect in (de)hydrogenation chemistry is the choice of the base activator / promotor for the catalytic reaction. An enhanced stability of a complex formed with the base can result in inhibition and the catalyst can remain in a resting state. Here, we considered a model (de)hydrogenation of iPrOH reaction as an example.

In Figure 7 and 8, the stability of isopropxide to the adduct of two potential bases (KOH and KOtBu, respectively) is compared. The results show that a competition between the base (KOtBu) and iPrOH for complexation with the Mn center is highly likely. However, the iPrO adduct was generally found to be more stable than the OH adduct for all Mn-pincers studied here. This concludes that KOH may represent a better choice for the systems where IPA dehydrogenation is important. Consistent with our observation, Beller and co-workers observed that switching from KOH to KOtBu lead to catalyst deactivation in dehydrogenation of methanol catalyzed by a Mn-PNP pincer complex.[52]

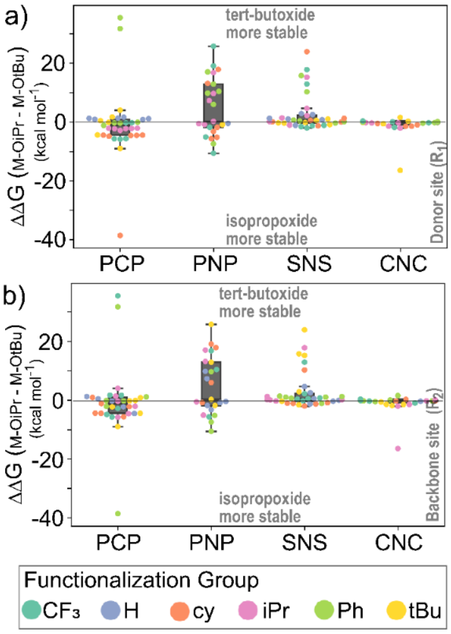


Figure 8. Comparison of isopropxide stabilities with tert-butoxide per catalytic ligand group. Swarmplots are added indicaton the functionalization on a) catalyst donor site R1 and b) catalyst backbone site R2.

Conclusions

In this investigation, we carried out a computational analysis of the effect of functionalization on different Mn-adducts of five types of pincer catalysts. Here, a data augmented approach was employed using fast xTB optimizations to analyze stabilities of metal adducts that can play a significant role in the catalytic cycle. The xTB results were compared to DFT calculations, which showed qualitative agreement and helped identifying the accuracy boundary of the accelerated xTB methodologies for studying Mn(I)-pincer complexes We identified linear scaling relation between Gibbs free energy for formation of different adducts, which can be used for rapid screening purposes.

Functionalization of the donor site directly affects the metal center activation, as illustrated by the changes in the computed CO stretching frequencies. Increased electronic density at the metal center and geometric strain both have destabilizing effects on the formation of alkoxides, hydroxide, bromides and hydride adducts. Comparison of relative stabilities of the iso-propoxide adduct with hydroxide and tert-butoxide adducts showed that KO^{t}Bu can have a poisoning effect during iPrOH dehydrogenation, and that KOH would be a more suitable base.

In the outlook, this work is a first step in mechanism based high throughput screening of pincer ligand based catalysts. Development of data-augmented approaches to screen and design highly active homogeneous catalysts is an ongoing effort in our laboratory.

Author contributions

A. K. carried out calculations and analysed the results. V. S. conceived and supervised the project. A. V. K. generated functionalized geometries of pincer complexes, coordinated in carrying out xTB and DFT calculations and compiling datasets from calculations. E. A. P. played a supervisory role and directed the project. All the authors discussed the results and wrote the manuscript.

Acknowledgements

V. S. acknowledges the ARC-CBBC project 2016.008 for funding. A.K. and E.A.P. acknowledges the financial support from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement no. 725686). The authors thank the Netherlands Organization for Scientific Research (NWO) for the access to SURFsara computational facilities and SURF Cooperative for support. The authors thank Mr. Rowin Borst for carrying out some of the xTB calculations. This work was carried out on the Dutch national e-infrastructure.

Keywords: data-driven catalysis • linear scaling relations • catalyst design • (de)hydrogenation catalysis • catalyst deactivation

(1) Werkmeister, S.; Junge, K.; Beller, M. Catalytic Hydrogenation of Carboxylic Acid Esters, Amides, and Nitriles with Homogeneous Catalysts. *Org. Process Res. Dev.* **2014**, *18* (2), 289–302. https://doi.org/10.1021/op4003278.

(2) Maser, L.; Vondung, L.; Langer, R. The ABC in Pincer Chemistry – From Amine- to Borylene- and Carbon-Based Pincer-Ligands. *Polyhedron* **2018**, *143*, 28–42. https://doi.org/10.1016/j.poly.2017.09.009.

(3) Werkmeister, S.; Neumann, J.; Junge, K.; Beller, M. Pincer-Type Complexes for Catalytic (De)Hydrogenation and Transfer (De)Hydrogenation Reactions: Recent Progress. *Chem. - A Eur. J.* **2015**, *21* (35), 12226–12250.

WILEY-VCH

https://doi.org/10.1002/chem.201500937.

(4) Dub, P. A.; Gordon, J. C. The Role of the Metal-Bound N–H Functionality in Noyori-Type Molecular Catalysts. *Nat. Rev. Chem.* **2018**, *2* (12), 396–408. https://doi.org/10.1038/s41570-018-0049-z.

(5) Kuriyama, W.; Matsumoto, T.; Ogata, O.; Ino, Y.; Aoki, K.; Tanaka, S.; Ishida, K.; Kobayashi, T.; Sayo, N.; Saito, T. Catalytic Hydrogenation of Esters. Development of an Efficient Catalyst and Processes for Synthesising (R)-1,2-Propanediol and 2-(l -Menthoxy)Ethanol. *Org. Process Res. Dev.* **2012**, *16* (1), 166–171. https://doi.org/10.1021/op200234j.

(6) Otsuka, T.; Ishii, A.; Dub, P. A.; Ikariya, T. Practical Selective Hydrogenation of α-Fluorinated Esters with Bifunctional Pincer-Type Ruthenium(II) Catalysts Leading to Fluorinated Alcohols or Fluoral Hemiacetals. *J. Am. Chem. Soc.* **2013**, *135* (26), 9600–9603. https://doi.org/10.1021/ja403852e.

(7) Aoki, W.; Wattanavinin, N.; Kusumoto, S.; Nozaki, K. Development of Highly Active Ir–PNP Catalysts for Hydrogenation of Carbon Dioxide with Organic Bases. *Bull. Chem. Soc. Jpn.* **2016**, *89* (1), 113–124. https://doi.org/10.1246/bcsj.20150311.

(8) Garbe, M.; Junge, K.; Walker, S.; Wei, Z.; Jiao, H.; Spannenberg, A.; Bachmann, S.; Scalone, M.; Beller, M. Manganese(I)-Catalyzed Enantioselective Hydrogenation of Ketones Using a Defined Chiral PNP Pincer Ligand. *Angew. Chemie Int. Ed.* **2017**, *56* (37), 11237–11241. https://doi.org/10.1002/anie.201705471.

(9) Kallmeier, F.; Kempe, R. Manganese Complexes for (De)Hydrogenation Catalysis: A Comparison to Cobalt and Iron Catalysts. *Angew. Chemie Int. Ed.* **2018**, *57* (1), 46–60. https://doi.org/10.1002/anie.201709010.

(10) Alig, L.; Fritz, M.; Schneider, S. First-Row Transition Metal (De)Hydrogenation Catalysis Based On Functional Pincer Ligands. *Chem. Rev.* **2019**, *119* (4), 2681–2751. https://doi.org/10.1021/acs.chemrev.8b00555.

(11) Filonenko, G. A.; van Putten, R.; Hensen, E. J. M.; Pidko, E. A. Catalytic (de)Hydrogenation Promoted by Non-Precious Metals – Co, Fe and Mn: Recent Advances in an Emerging Field. *Chem. Soc. Rev.* **2018**, *47* (4), 1459–1483. https://doi.org/10.1039/C7CS00334J.

(12) Agbossou-Niedercorn, F.; Michon, C. Bifunctional Homogeneous Catalysts Based on First Row Transition Metals in Asymmetric Hydrogenation. *Coord. Chem. Rev.* **2020**, *425*, 213523. https://doi.org/10.1016/j.ccr.2020.213523.

(13) Schneider, S.; Meiners, J.; Askevold, B. Cooperative Aliphatic PNP Amido Pincer Ligands - Versatile Building Blocks for Coordination Chemistry and Catalysis. *Eur. J. Inorg. Chem.* **2012**, *2012* (3), 412–429. https://doi.org/10.1002/ejic.201100880.

(14) Evans, K. J.; Mansell, S. M. Functionalised N-Heterocyclic Carbene Ligands in Bimetallic Architectures. *Chem. – A Eur. J.* **2020**, *26* (27), 5927–5941. https://doi.org/10.1002/chem.201905510.

(15) Friederich, P.; dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine Learning Dihydrogen Activation in the Chemical Space Surrounding Vaska's Complex. *Chem. Sci.* **2020**, *11* (18), 4584–4601. https://doi.org/10.1039/D0SC00445F.

(16) Janet, J. P.; Liu, F.; Nandy, A.; Duan, C.; Yang, T.; Lin, S.; Kulik, H. J. Designing in the Face of Uncertainty: Exploiting Electronic Structure and Machine Learning Models for Discovery in Inorganic Chemistry. *Inorg. Chem.* **2019**, *58* (16), 10592–10606. https://doi.org/10.1021/acs.inorgchem.9b00109.

(17) Wodrich, M. D.; Sawatlon, B.; Solel, E.; Kozuch, S.; Corminboeuf, C. Activity-Based Screening of Homogeneous Catalysts through the Rapid Assessment of Theoretically Derived Turnover Frequencies. *ACS Catal.* **2019**, *9* (6), 5716–5725. https://doi.org/10.1021/acscatal.9b00717.

(18) Sawatlon, B.; Wodrich, M. D.; Corminboeuf, C. Unraveling Metal/Pincer Ligand Effects in the Catalytic Hydrogenation of Carbon Dioxide to Formate. *Organometallics* **2018**, *37* (24), 4568–4575. https://doi.org/10.1021/acs.organomet.8b00490.

(19) Wodrich, M. D.; Busch, M.; Corminboeuf, C. Expedited Screening of Active and Regioselective Catalysts for the Hydroformylation Reaction. *Helv. Chim. Acta* **2018**, *101* (9), e1800107. https://doi.org/10.1002/hlca.201800107.

(20) Schneck, F.; Finger, M.; Tromp, M.; Schneider, S. Chemical Non-Innocence of an Aliphatic PNP Pincer Ligand. *Chem. - A Eur. J.* **2017**, *23* (1), 33–37. https://doi.org/10.1002/chem.201604407.

(21) Zhang, J.; Balaraman, E.; Leitus, G.; Milstein, D. Electron-Rich PNP- and PNN-Type Ruthenium(II) Hydrido Borohydride Pincer Complexes. Synthesis, Structure, and Catalytic Dehydrogenation of Alcohols and Hydrogenation of Esters. *Organometallics* **2011**, *30* (21), 5716–5724. https://doi.org/10.1021/om200595m.

(22) Glatz, M.; Stöger, B.; Himmelbauer, D.; Veiros, L. F.; Kirchner, K. Chemoselective Hydrogenation of Aldehydes under Mild, Base-Free Conditions: Manganese Outperforms Rhenium. *ACS Catal.* **2018**, *8* (5), 4009–4016. https://doi.org/10.1021/acscatal.8b00153.

(23) Bertini, F.; Glatz, M.; Stöger, B.; Peruzzini, M.; Veiros, L. F.; Kirchner, K.; Gonsalvi, L. Carbon Dioxide Reduction to Methanol Catalyzed by Mn(I) PNP Pincer Complexes under Mild Reaction Conditions. *ACS Catal.* **2019**, *9* (1), 632–639. https://doi.org/10.1021/acscatal.8b04106.

(24) Kostera, S.; Peruzzini, M.; Kirchner, K.; Gonsalvi, L. Mild and Selective Carbon Dioxide Hydroboration to Methoxyboranes Catalyzed by Mn(I) PNP Pincer Complexes. *ChemCatChem* **2020**, *12* (18), 4625–4631. https://doi.org/10.1002/cctc.202000469.

(25) Zhang, L.; Wang, Z.; Han, Z.; Ding, K. Manganese-Catalyzed Anti -Selective Asymmetric Hydrogenation of A-Substituted B-Ketoamides. *Angew. Chemie Int. Ed.* **2020**, *59* (36), 15565–15569. https://doi.org/10.1002/anie.202006383.

(26) Ling, F.; Chen, J.; Nian, S.; Hou, H.; Yi, X.; Wu, F.; Xu, M.; Zhong, W. Manganese-Catalyzed Enantioselective Hydrogenation of Simple Ketones Using an Imidazole-Based Chiral PNN Tridentate Ligand. *Synlett* **2020**, *31* (03), 285–289. https://doi.org/10.1055/s-0039-1690783.

(27) Kumar, A.; Daw, P.; Espinosa-Jalapa, N. A.; Leitus, G.; Shimon, L. J. W.; Ben-David, Y.; Milstein, D. CO 2 Activation by Manganese Pincer Complexes through Different Modes of Metal–Ligand Cooperation. *Dalt. Trans.* **2019**, *48* (39), 14580–14584. https://doi.org/10.1039/C9DT03088C.

(28) Zhang, L.; Tang, Y.; Han, Z.; Ding, K. Lutidine-Based Chiral Pincer Manganese Catalysts for Enantioselective Hydrogenation of Ketones. *Angew. Chemie Int. Ed.* **2019**, *58* (15), 4973–4977.

https://doi.org/10.1002/anie.201814751.

(29) Filonenko, G. A.; Hensen, E. J. M.; Pidko, E. A. Mechanism of CO_2 Hydrogenation to Formates by Homogeneous Ru-PNP Pincer Catalyst: From a Theoretical Description to Performance Optimization. *Catal. Sci. Technol.* **2014**, *4* (10), 3474–3485. https://doi.org/10.1039/C4CY00568F.

(30) Biswas, N.; Sharma, R.; Srimani, D. Ruthenium Pincer Complex Catalyzed Selective Synthesis of C-3 Alkylated Indoles and Bisindolylmethanes Directly from Indoles and Alcohols. *Adv. Synth. Catal.* **2020**, *362* (14), 2902–2910. https://doi.org/10.1002/adsc.202000326.

(31) Schörgenhumer, J.; Zimmermann, A.; Waser, M. SNS-Ligands for Ru-Catalyzed Homogeneous Hydrogenation and Dehydrogenation Reactions. *Org. Process Res. Dev.* **2018**, *22* (7), 862–870. https://doi.org/10.1021/acs.oprd.8b00142.

(32) Chen, X.; Jing, Y.; Yang, X. Unexpected Direct Hydride Transfer Mechanism for the Hydrogenation of Ethyl Acetate to Ethanol Catalyzed by SNS Pincer Ruthenium Complexes. *Chem. - A Eur. J.* **2016**, *22* (6), 1950–1957. https://doi.org/10.1002/chem.201504058.

(33) Spasyuk, D.; Smith, S.; Gusev, D. G. Replacing Phosphorus with Sulfur for the Efficient Hydrogenation of Esters. *Angew. Chemie Int. Ed.* **2013**, *52* (9), 2538–2542. https://doi.org/10.1002/anie.201209218.

(34) Koizumi, T.; Teratani, T.; Okamoto, K.; Yamamoto, T.; Shimoi, Y.; Kanbara, T. Nickel(II) Complexes Bearing a Pincer Ligand Containing Thioamide Units: Comparison between SNS- and SCS-Pincer Ligands. *Inorganica Chim. Acta* **2010**, *363* (11), 2474–2480. https://doi.org/10.1016/j.ica.2010.04.012.

(35) Das, U. K.; Daifuku, S. L.; Iannuzzi, T. E.; Gorelsky, S. I.; Korobkov, I.; Gabidullin, B.; Neidig, M. L.; Baker, R. T. Iron(II) Complexes of a Hemilabile SNS Amido Ligand: Synthesis, Characterization, and Reactivity. *Inorg. Chem.* **2017**, *56* (22), 13766–13776. https://doi.org/10.1021/acs.inorgchem.7b01802.

(36) Wu, X.; Ji, L.; Ji, Y.; Elageed, E. H. M.; Gao, G. Hydrogenation of Ethylene Carbonate Catalyzed by Lutidine-Bridged N-Heterocyclic Carbene Ligands and Ruthenium Precursors. *Catal. Commun.* **2016**, *85*, 57–60. https://doi.org/10.1016/j.catcom.2016.07.015.

(37) Hernández-Juárez, M.; López-Serrano, J.; Lara, P.; Morales-Cerón, J. P.; Vaquero, M.; Álvarez, E.; Salazar, V.; Suárez, A. Ruthenium(II) Complexes Containing Lutidine-Derived Pincer CNC Ligands: Synthesis, Structure, and Catalytic Hydrogenation of C□N Bonds. *Chem. - A Eur. J.* **2015**, *21* (20), 7540–7555. https://doi.org/10.1002/chem.201406040.

(38) Filonenko, G. A.; Cosimi, E.; Lefort, L.; Conley, M. P.; Copéret, C.; Lutz, M.; Hensen, E. J. M.; Pidko, E. A. Lutidine-Derived Ru-CNC Hydrogenation Pincer Catalysts with Versatile Coordination Properties. *ACS Catal.* **2014**, *4* (8), 2667–2671. https://doi.org/10.1021/cs500720y.

(39) Naziruddin, A. R.; Kuo, C.-L.; Lin, W.-J.; Lo, W.-H.; Lee, C.-S.; Sun, B.-J.; Chang, A. H. H.; Hwang, W.-S. Ruthenium Complexes Bearing Unsymmetric CNC′ Pincer Ligands: Molecular Structures and Electronic Properties. *Organometallics* **2014**, *33* (10), 2575–2582. https://doi.org/10.1021/om500205p.

(40) Gründemann, S.; Albrecht, M.; Loch, J. A.; Faller, J. W.; Crabtree, R. H. Tridentate Carbene CCC and CNC Pincer Palladium(II) Complexes: Structure, Fluxionality, and Catalytic Activity. *Organometallics* **2001**, *20* (25), 5485–5488. https://doi.org/10.1021/om010631h.

(41) Danopoulos, A. A.; Tsoureas, N.; Wright, J. A.; Light, M. E. N-Heterocyclic Pincer Dicarbene Complexes of Iron(II): C-2 and C-5 Metalated Carbenes on the Same Metal Center. *Organometallics* **2004**, *23* (2), 166–168. https://doi.org/10.1021/om0341911.

(42) Gagliardo, M.; Chase, P. A.; Brouwer, S.; van Klink, G. P. M.; van Koten, G. Electronic Effects in PCP-Pincer Ru(II)-Based Hydrogen Transfer Catalysis. *Organometallics* **2007**, *26* (9), 2219–2227. https://doi.org/10.1021/om060874f.

(43) Gruver, B. C.; Adams, J. J.; Warner, S. J.; Arulsamy, N.; Roddick, D. M. Acceptor Pincer Chemistry of Ruthenium: Catalytic Alkane Dehydrogenation by (CF 3 PCP)Ru(Cod)(H). *Organometallics* **2011**, *30* (19), 5133–5140. https://doi.org/10.1021/om200354y.

(44) Tang, S.; von Wolff, N.; Diskin-Posner, Y.; Leitus, G.; Ben-David, Y.; Milstein, D. Pyridine-Based PCP-Ruthenium Complexes: Unusual Structures and Metal–Ligand Cooperation. *J. Am. Chem. Soc.* **2019**, *141* (18), 7554–7561. https://doi.org/10.1021/jacs.9b02669.

(45) Adams, J. J.; Arulsamy, N.; Roddick, D. M. Acceptor PCP Pincer Iridium(I) Chemistry: Stabilization of Nonmeridional PCP Coordination Geometries. *Organometallics* **2011**, *30* (4), 697–711. https://doi.org/10.1021/om100864g.

(46) Murugesan, S.; Kirchner, K. Non-Precious Metal Complexes with an Anionic PCP Pincer Architecture. *Dalt. Trans.* **2016**, *45* (2), 416–439. https://doi.org/10.1039/C5DT03778F.

(47) Spasyuk, D.; Smith, S.; Gusev, D. G. Replacing Phosphorus with Sulfur for the Efficient Hydrogenation of Esters. *Angew. Chemie* **2013**, *125* (9), 2598–2602. https://doi.org/10.1002/ange.201209218.

(48) Liu, C.; van Putten, R.; Kulyaev, P. O.; Filonenko, G. A.; Pidko, E. A. Computational Insights into the Catalytic Role of the Base Promoters in Ester Hydrogenation with Homogeneous Non-Pincer-Based Mn-P,N Catalyst. *J. Catal.* **2018**, *363*, 136–143. https://doi.org/10.1016/j.jcat.2018.04.018.

(49) Larmier, K.; Liao, W.-C.; Tada, S.; Lam, E.; Verel, R.; Bansode, A.; Urakawa, A.; Comas-Vives, A.; Copéret, C. CO 2 -to-Methanol Hydrogenation on Zirconia-Supported Copper Nanoparticles: Reaction Intermediates and the Role of the Metal-Support Interface. *Angew. Chemie Int. Ed.* **2017**, *56* (9), 2318–2323. https://doi.org/10.1002/anie.201610166.

(50) Wang, W.-H.; Himeda, Y.; Muckerman, J. T.; Manbeck, G. F.; Fujita, E. CO 2 Hydrogenation to Formate and Methanol as an Alternative to Photo- and Electrochemical CO 2 Reduction. *Chem. Rev.* **2015**, *115* (23), 12936–12973. https://doi.org/10.1021/acs.chemrev.5b00197.

(51) Yang, H.; Zhang, C.; Gao, P.; Wang, H.; Li, X.; Zhong, L.; Wei, W.; Sun, Y. A Review of the Catalytic Hydrogenation of Carbon Dioxide into Value-Added Hydrocarbons. *Catal. Sci. Technol.* **2017**, *7* (20), 4580–4598. https://doi.org/10.1039/C7CY01403A.

(52) Andérez-Fernández, M.; Vogt, L. K.; Fischer, S.; Zhou, W.; Jiao, H.; Garbe, M.; Elangovan, S.; Junge, K.; Junge, H.; Ludwig, R.; Beller, M. A Stable Manganese Pincer Catalyst for the Selective Dehydrogenation of Methanol. *Angew. Chemie Int. Ed.* **2017**, *56* (2),

WILEY-VCH

559–562. https://doi.org/10.1002/anie.201610182.

(53) Daley, C. J. A.; Bergens, S. H. The First Complete Identification of a Diastereomeric Catalyst−Substrate (Alkoxide) Species in an Enantioselective Ketone Hydrogenation. Mechanistic Investigations. *J. Am. Chem. Soc.* **2002**, *124* (14), 3680–3691. https://doi.org/10.1021/ja0102991.

(54) Hamilton, R. J.; Bergens, S. H. An Unexpected Possible Role of Base in Asymmetric Catalytic Hydrogenations of Ketones. Synthesis and Characterization of Several Key Catalytic Intermediates. *J. Am. Chem. Soc.* **2006**, *128* (42), 13700–13701. https://doi.org/10.1021/ja065460s.

(55) Hamilton, R. J.; Bergens, S. H. Direct Observations of the Metal−Ligand Bifunctional Addition Step in an Enantioselective Ketone Hydrogenation. *J. Am. Chem. Soc.* **2008**, *130* (36), 11979–11987. https://doi.org/10.1021/ja8034812.

(56) Zimmer-De Iuliis, M.; Morris, R. H. Kinetic Hydrogen/Deuterium Effects in the Direct Hydrogenation of Ketones Catalyzed by a Well-Defined Ruthenium Diphosphine Diamine Complex. *J. Am. Chem. Soc.* **2009**, *131* (31), 11263–11269. https://doi.org/10.1021/ja9043104.

(57) Passera, A.; Mezzetti, A. Mn(I) and Fe(II)/PN(H)P Catalysts for the Hydrogenation of Ketones: A Comparison by Experiment and Calculation. *Adv. Synth. Catal.* **2019**, *361* (20), 4691–4706. https://doi.org/10.1002/adsc.201900671.

(58) Pham, J.; Jarczyk, C.; Reynolds, E.; Kelly, S.; Kim, T.; He, T.; Keith, J.; Chianese, A. . The Key Role of the Latent N-H Group in Milstein’s Catalyst for Ester Hydrogenation. **2021**. https://doi.org/10.26434/chemrxiv.13618988.v1.

(59) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended <scp>tight-binding</Scp> Quantum Chemistry Methods. *WIREs Comput. Mol. Sci.* **2021**, *11* (2). https://doi.org/10.1002/wcms.1493.

(60) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-XTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15* (3), 1652–1671. https://doi.org/10.1021/acs.jctc.8b01176.

(61) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All Spd-Block Elements (Z = 1–86). *J. Chem. Theory Comput.* **2017**, *13* (5), 1989–2009. https://doi.org/10.1021/acs.jctc.7b00118.

(62) 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. Gaussian16 Revision C.01. 2016.

(63) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7* (18), 3297. https://doi.org/10.1039/b508541a.

(64) Bühl, M.; Kabrede, H. Geometries of Transition-Metal Complexes from Density-Functional Theory. *J. Chem. Theory Comput.* **2006**, *2* (5), 1282–1290. https://doi.org/10.1021/ct6001187.

(65) Jensen, K. P.; Roos, B. O.; Ryde, U. Performance of Density Functionals for First Row Transition Metal Systems. *J. Chem. Phys.* **2007**, *126* (1), 014103. https://doi.org/10.1063/1.2406071.

(66) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113* (18), 6378–6396. https://doi.org/10.1021/jp810292n.

(67) Adamo, C.; Barone, V. Toward Reliable Density Functional Methods without Adjustable Parameters: The PBE0 Model. *J. Chem. Phys.* **1999**, *110* (13), 6158–6170. https://doi.org/10.1063/1.478522.

(68) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100. https://doi.org/10.1103/PhysRevA.38.3098.

(69) Caldeweyher, E.; Bannwarth, C.; Grimme, S. Extension of the D3 Dispersion Coefficient Model. *J. Chem. Phys.* **2017**, *147* (3), 034112. https://doi.org/10.1063/1.4993215.

(70) The Geometries of Functionalized Complexes Were Generated Using an In-House Developed Automated Python Based Workflow Named “ChemSpaX”. The Related Manuscript Is Currently in Preparation.

(71) Sinha, V.; Laan, J. J.; Pidko, E. A. Accurate and Rapid Prediction of p K a of Transition Metal Complexes: Semiempirical Quantum Chemistry with a Data-Augmented Approach. *Phys. Chem. Chem. Phys.* **2021**, *23* (4), 2557–2567. https://doi.org/10.1039/D0CP05281G.

(72) Hansch, C.; Leo, A.; Taft, R. W. A Survey of Hammett Substituent Constants and Resonance and Field Parameters. *Chem. Rev.* **1991**, *91* (2), 165–195. https://doi.org/10.1021/cr00002a004.

4 Summary & outlook

4.1. Summary

In this thesis, a workflow for the exploration of the local chemical space of a structure was presented. In this workflow, *ChemSpaX*, substituents are placed on specific sites of an input structure. A data-driven approach for calculating and correlating descriptors by using chemical intuition to place these substituents in combination xTB and DFT calculations was investigated.

In the first application of *ChemSpaX*, 'ChemSpax: Explorationof chemical space by automated functionalization of molecular scaffold', various complexes were investigated by extracting relevant descriptors and correlating them. First a ruthenium-based and manganese-based pincer complexes were investigated. The Ru based intermediates were derived from a functionalized PNP ligand. For the manganese-based pincers, the functionalized variants of four pincer type ligand scaffolds derived from PCP, CNC, PNP and SNS ligands were used. Here, a nearly linear scaling of ΔE and ΔG (R^2 = 0.99 for Ru- and the Mn-pincers) was found. Which indicates that ΔG can be replaced by ΔE in high-throughput screening. Then the GFN2-xTB calculated HOMO-LUMO gap was compared against the DFT calculated HOMO-LUMO gap and a reasonable correlation was found for the Ru-based structures (R^2 = 0.74 and RMSE = 0.4 eV).

Then cobalt porhpyrins were investigated. These Co porphyrins showed a linear increase in hRMSD upon serial functionalization on the same structure which can be used to predict when a higher-level optimization method is needed to optimize the structure. Clustering was observed in the data when the hRMSD was compared against the number of atoms in a structure. Using linear regression via OLS, easily computable descriptors extracted from GFN2-xTB were used to predict the DFT calculated HOMO-LUMO gap. A reasonable correlation was found on the test set (R^2 = 0.71 and RMSE = 0.12 eV).

To conclude, this first application of *ChemSpaX* showed that *ChemSpaX* can be used in the generation of many structures with a quality reasonably close to GFN2-xTB optimization for high throughput calculations. The potential of GFN2-xTB for high-throughput screening applications was also shown.

In the second application of *ChemSpaX*, 'Metal-ligand cooperative activation of HX (X=H, Br, OR) bond on Mn based pincer complexes', the thermodynamic stability of different adducts in various functionalized Mn based pincer complexes was investigated [116]. Here, DFT and xTB calculations were used to analyze the same functionalized Mn-based pincer complexes as presented in the previous application. The ΔG_{HX} of > 700 (M(X)-L(H)) intermediates were investigated and linear scaling relations were found between the ΔG_{HX} of various species. These scaling relations can be useful in HTS of catalysts. It was found that the CO stretching frequencies would change upon functionalization of the donor site, which indicates that this functionalization affects the metal center activation. Increased electron density at the metal center and geometric strain were found to have destabilizing effects on the formation of alkoxides, hydroxide, bromides and hydride adducts. Finally, it was found that KOH would be a more suitable base compared to KOtBu by comparison of relative stabilities of the isopropoxide adduct with the hydroxide and tert-butoxide adducts.

4.2. Outlook

The future of computational catalyst discovery is bright, but if we want to combine computational screening with experimental work, extensive benchmarking is needed. Benchmarking means that various levels of theory should be compared compared against each other, as well as comparing theoretical calculations against experimental results [117]. For DFT various examples can be given. So are the dispersion corrections and hydrogen-bonding interactions assessed [118–120]. And there are efforts made on the development of standardized and unbiased methods for benchmarking [121]. For xTB various benchmarks for structural and thermochemical properties including (transition-)metal systems are done against benchmarking sets [72]. Although computational chemistry is rapidly evolving, theoretical calculations on their own can be insufficient to predict reaction mechanisms [122]. For comparisons of theoretical calculations against experimental results, standardization of sharing methods, models and code is necessary [117]. This would increase the chance that other researchers start using this data and would simplify benchmarking. Futhermore, theoretical results for simple model systems should be compared with results of more accurate theoretical methods or with experimental results from well-characterized materials such as molecular catalyst [117]. An issue arises for TM complexes since there is not much high-quality theoretical/experimental data available for these complexes, even though these compounds play an essential role in catalysis. More research and development is needed on this front.

Next to this, good descriptors of catalytic activity are needed which can be used as target in predictive models. This would accelerate virtual screening (VS) methods in which both the descriptors and catalytic activity or selectivity are computed. Additionally, this would aid in the usage of methods like the volcano plot, which relate descriptors to catalytic performance and can distinguish structures based descriptors on catalytic performance using Sabatier's principle [123–126].

As explained in '*ChemSpax*: Exploration of chemical space by automated functionalization of molecular scaffold', a molecular projection scheme to map the electronic structure and geometry to relevant reaction descriptors is needed. This would enable ML methods for screening. Several representations like the autoencoders that map molecular structures into a continuous latent space and the sorted Coulom Matrix are mentioned [123, 127]. However, more research, benchmarking and standardization is needed in this area.

Several new projects and improvements based on the research reported in this thesis can be envisioned. In *ChemSpaX* the user is currently required to manually check their input geometry, which includes checking bond lengths/angles. In the *utilities_scripts* folder in the *ChemSpaX* repository on ISE's Github page (https://github.com/EPiCs-group/), *get_neighbour_distance_search.py* can be used to find the bond lengths between two atoms. A caveat is that the atoms should have a bond in the graph representation, which is prone to errors when *XYZ* files are used. If the approach used in this script is improved to create correct graph representations on its own, it can be used to calculate bond lengths between atoms and compare them against a quantum chemical database. A similar approach using Openbabel's methods can be used for bond angles. Using this, an automated bond/angle check can be done before functionalizations are done.

Currently, *ChemSpaX* uses Openbabel's force field methods for optimization of newly placed substituents. The end goal is to make *ChemSpaX* easily usable by the scientific community, implementing a standalone optimization method for newly placed substituent would lift the dependency on Openbabel. This optimization method could be based on Openbabel's force field methods or Spartan's PM3tm semi-empirical method. Having less dependencies on external packages would greatly improve the ease of installation and would thus improve the willingness of people to use the program.

Ideally, *ChemSpaX* would be implemented in a workflow where automated data-driven property calculation/prediction is done. Work is being done on a workflow that enables these automated high-throughput descriptor calculations, which is called *epic_dna*, and will be hosted on ISE's Github page (https://github.com/EPiCs-group/). Eventually, the computational workflow of *epic_dna* should be integrated with an experimental workflow. Work is required to make the integration as understandable as possible for researches from various disciplines. In the advances that are made with *epic_dna*, this will mean that experimental researchers can draw their synthesized catalyst structure in programs like Marvinsketch. The SMILES string generated from this drawing will be used together with experimental data for generating a database of molecular and reaction descriptors. After the generation of this database is realized, property prediction with ML combined with local chemical space exploration done by *ChemSpaX* can be researched. Using this workflow together with chemical intuition, a larger subset of the chemical space can then be covered in future research.

Extensive benchmarking is needed for *ChemSpaX*. This can be done as presented in this thesis, by calculating the same property for two structures that were optimized with a different optimization method and comparing the property afterwards. By using higher level DFT methods as 'standard' for comparison, the workflow can be improved and the bar can be raised.

ChemSpaX and related automation tools are meant to be open source and easy to use for the scientific community. By sharing data and code, more progress can be made in the exploration of chemical space, of which humanity has just been scratching the surface. The help of young researchers is needed to push this interdisciplinary young field forward. Unfortunately, this shows an educational challenge, since research programmes in chemistry, physics and computer science need to be tightly interwoven [128]. The coursework in conventional curriculae of chemistry, materials science, physics, computer science or biology would need to be adapted for students to reach a level by which they can have a meaningful contribution to this line of research [128]. These adaptations can also be done in the master's programme in chemical engineering at the TU Delft, where subjects like data science, object-oriented programming and machine learning currently do not get the attention that they deserve.

Acknowledgements

This project was performed as the final thesis for the master Chemical Engineering at the Delft University of Technology. The research was performed at the Inorganic Systems Engineering group. This project was performed fully online due to the ongoing Covid-19 pandemic, but that never spoiled the fun that I had while performing the research.

I would like to start by thanking Professor Evgeny Pidko for giving me the opportunity to work on this project and for the guidance during the past months. Thanks to you I was able to gain skills and gather experiences that a real researcher needs. This included (but was not limited to) attending symposia/workshops, gathering more info and networking at those events, and doing graphical design for a video about my research. Thank you for this wonderful all-in-1 learning experience.

Furthermore, I would like to thank my supervisor Dr. Vivek Sinha for the daily supervision and wise lessons. Even though he was very busy with his own projects and his newborn son, he always found the time to talk to me. I was happy to work with someone that also had very messed up working hours during the pandemic, this enabled us to have meetings at the weirdest times. I must say that I admire your (very unhealthy) work ethic, Vivek, and I think that I would never have learned as much as I have without your guidance. You were always kind enough to answer my questions, teach me how to write functional code (preferably in Bash) and you introduced me to the process of turning scientific research into a publication while I had not even thought of publishing a paper yet. You are always able to push me to do better, so I am very thankful that I was able to work under your supervision.

I am thankful for the bachelor, master and PhD students with whom I was able to exchange ideas and share a (digital) laugh. I also want to thank the rest of the ISE staff and everyone else that contributed to the pleasant working environment and this project in any imaginable way.

Normally I would spend my days at the university with other students, in the past year I spent most of my time at home with my parents and brothers. I would like to thank them for the joyful moments we shared. Most importantly, I would like to thank all of my family members for always believing in me and supporting me, especially my mother. As a little kid, she gave me a children's encyclopedia to let me learn about the world around us. This practically marked the start of my academic career. She has always motivated me to become a better version of myself and told me that I should never stop learning as long as I enjoy it. Thank you mom, I could not have done this without you.

I would like to finish this thesis with something that my grandpa (nana) used to tell me and I have always remembered this: "My grandson, even though I am just a simple farmer, I am proud that I was able to have grandchildren that are educated on a level that I could only dream of. I hope that you will all be able to make this world a better place."

- Adarsh Kalikadien
Delft, April 2021

Bibliography

[1] C Richard Catlow, Matthew Davidson, Christopher Hardacre, and Graham J Hutchings. Catalysis making the world a better place. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2061):20150089, feb 2016. ISSN 1471-2962. doi: 10.1098/rsta.2015.0089. URL https://pubmed.ncbi.nlm.nih.gov/26755766https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4707691/.

[2] Piet van Leeuwen. *Homogeneous Catalysis – Understanding the Art*. Springer Netherlands, apr 2004. doi: 10.1007/1-4020-2000-7.

[3] Stephen Mccord, Cynthia Labrake, and David Vanden Bout. Chemistry 302 University of Texas, 2014. URL http://ch302.cm.utexas.edu/kinetics/catalysts/catalysts-all.php.

[4] T.-C. Ong, R. Verel, and C. Copéret. Solid-state nmr: Surface chemistry applications. In John C. Lindon, George E. Tranter, and David W. Koppenaal, editors, *Encyclopedia of Spectroscopy and Spectrometry (Third Edition)*, pages 121–127. Academic Press, Oxford, third edition edition, 2017. ISBN 978-0-12-803224-4. doi: https://doi.org/10.1016/B978-0-12-409547-2.12130-4. URL https://www.sciencedirect.com/science/article/pii/B9780124095472121304.

[5] Hans Ulrich Blaser, Benoît Pugin, Felix Spindler, and Lionel A Saudan. Hydrogenation, nov 2017. URL https://doi.org/10.1002/9783527651733.ch9.

[6] Johannes de Vries and Cornelis Elsevier. *The Handbook of Homogeneous Hydrogenation*. Wiley Online Books. WILEY-VCH Verlag GmbH & Co. KGaA, oct 2006. ISBN 9783527619382. doi: https://doi.org/10.1002/9783527619382.fmatter. URL https://doi.org/10.1002/9783527619382.fmatter.

[7] David Benito-Garagorri and Karl Kirchner. Modularly Designed Transition Metal PNP and PCP Pincer Complexes based on Aminophosphines: Synthesis and Catalytic Applications. *Accounts of Chemical Research*, 41(2):201–213, feb 2008. ISSN 0001-4842. doi: 10.1021/ar700129q. URL https://doi.org/10.1021/ar700129q.

[8] Chidambaram Gunanathan and David Milstein. Metal–Ligand Cooperation by Aromatization–Dearomatization: A New Paradigm in Bond Activation and "Green" Catalysis. *Accounts of Chemical Research*, 44(8):588–602, aug 2011. ISSN 0001-4842. doi: 10.1021/ar2000265. URL https://doi.org/10.1021/ar2000265.

[9] Chidambaram Gunanathan and David Milstein. Bond Activation and Catalysis by Ruthenium Pincer Complexes. *Chemical Reviews*, 114(24):12024–12087, dec 2014. ISSN 0009-2665. doi: 10.1021/cr5002782. URL https://doi.org/10.1021/cr5002782.

[10] Jongwook Choi, Amy H Roy MacArthur, Maurice Brookhart, and Alan S Goldman. Dehydrogenation and Related Reactions Catalyzed by Iridium Pincer Complexes. *Chemical Reviews*, 111(3):1761–1779, mar 2011. ISSN 0009-2665. doi: 10.1021/cr1003503. URL https://doi.org/10.1021/cr1003503.

[11] Hussein A Younus, Nazir Ahmad, Wei Su, and Francis Verpoort. Ruthenium pincer complexes: Ligand design and complex synthesis. *Coordination Chemistry Reviews*, 276:112–152, 2014. ISSN 0010-8545. doi: https://doi.org/10.1016/j.ccr.2014.06.016. URL https://www.sciencedirect.com/science/article/pii/S0010854514001830.

[12] Nicklas Selander and Kálmán J Szabó. Catalysis by Palladium Pincer Complexes. *Chemical Reviews*, 111 (3):2048–2076, mar 2011. ISSN 0009-2665. doi: 10.1021/cr1002112. URL https://doi.org/10.1021/cr1002112.

[13] Marcel Garbe, Kathrin Junge, and Matthias Beller. Homogeneous Catalysis by Manganese-Based Pincer Complexes. *European Journal of Organic Chemistry*, 2017(30):4344–4362, aug 2017. ISSN 1434-193X. doi: https://doi.org/10.1002/ejoc.201700376. URL https://doi.org/10.1002/ejoc.201700376.

[14] A B Gallo, J R Simões-Moreira, H K M Costa, M M Santos, and E Moutinho dos Santos. Energy storage in the energy transition context: A technology review. *Renewable and Sustainable Energy Reviews*, 65:800–822, 2016. ISSN 1364-0321. doi: https://doi.org/10.1016/j.rser.2016.07.028. URL https://www.sciencedirect.com/science/article/pii/S1364032116303562.

[15] S Oldenhof, J I van der Vlugt, and J N H Reek. Hydrogenation of CO2 to formic acid with iridiumIII(bisMETAMORPhos)(hydride): the role of a dormant fac-IrIII(trihydride) and an active trans-IrIII(dihydride) species. *Catalysis Science & Technology*, 6(2):404–408, 2016. ISSN 2044-4753. doi: 10.1039/C5CY01476J. URL http://dx.doi.org/10.1039/C5CY01476J.

[16] Björn Loges, Albert Boddien, Felix Gärtner, Henrik Junge, and Matthias Beller. Catalytic Generation of Hydrogen from Formic acid and its Derivatives: Useful Hydrogen Storage Materials. *Topics in Catalysis*, 53 (13):902–914, 2010. ISSN 1572-9028. doi: 10.1007/s11244-010-9522-8. URL https://doi.org/10.1007/s11244-010-9522-8.

[17] Ferenc Joó. Breakthroughs in Hydrogen Storage—Formic Acid as a Sustainable Storage Material for Hydrogen. *ChemSusChem*, 1(10):805–808, oct 2008. ISSN 1864-5631. doi: 10.1002/cssc.200800133. URL https://doi.org/10.1002/cssc.200800133.

[18] Philip G Jessop. Homogeneous Hydrogenation of Carbon Dioxide, oct 2006. URL https://doi.org/10.1002/9783527619382.ch17.

[19] Christopher Zhou, William Grumbles, and Thomas Cundari. Using Machine Learning to Predict the pKa of C–H Bonds. Relevance to Catalytic Methane Functionalization, jul 2020. URL https://doi.org/10.26434/chemrxiv.12646772.v1.

[20] Carl Poree and Franziska Schoenebeck. A Holy Grail in Chemistry: Computational Catalyst Design: Feasible or Fiction? *Accounts of Chemical Research*, 50(3):605–608, mar 2017. ISSN 0001-4842. doi: 10.1021/acs.accounts.6b00606. URL https://doi.org/10.1021/acs.accounts.6b00606.

[21] David Balcells and Feliu Maseras. Computational approaches to asymmetric synthesis. *New Journal of Chemistry*, 31(3):333–343, 2007. ISSN 1144-0546. doi: 10.1039/B615528F. URL http://dx.doi.org/10.1039/B615528F.

[22] David Balcells, Eric Clot, and Odile Eisenstein. C—H Bond Activation in Transition Metal Species from a Computational Perspective. *Chemical Reviews*, 110(2):749–823, feb 2010. ISSN 0009-2665. doi: 10.1021/cr900315k. URL https://doi.org/10.1021/cr900315k.

[23] David L Davies, Stuart A Macgregor, and Claire L McMullin. Computational Studies of Carboxylate-Assisted C–H Activation and Functionalization at Group 8–10 Transition Metal Centers. *Chemical Reviews*, 117(13):8649–8709, jul 2017. ISSN 0009-2665. doi: 10.1021/acs.chemrev.6b00839. URL https://doi.org/10.1021/acs.chemrev.6b00839.

[24] Pietro Vidossich, Agustí Lledós, and Gregori Ujaque. First-Principles Molecular Dynamics Studies of Organometallic Complexes and Homogeneous Catalytic Processes. *Accounts of Chemical Research*, 49 (6):1271–1278, jun 2016. ISSN 0001-4842. doi: 10.1021/acs.accounts.6b00054. URL https://doi.org/10.1021/acs.accounts.6b00054.

[25] Marco Foscato and Vidar R Jensen. Automated in Silico Design of Homogeneous Catalysts. *ACS Catalysis*, 10(3):2354–2377, feb 2020. doi: 10.1021/acscatal.9b04952. URL https://doi.org/10.1021/acscatal.9b04952.

[26] Seihwan Ahn, Mannkyu Hong, Mahesh Sundararajan, Daniel H Ess, and Mu-Hyun Baik. Design and Optimization of Catalysts Based on Mechanistic Insights Derived from Quantum Chemical Reaction Modeling. *Chemical Reviews*, 119(11):6509–6560, jun 2019. ISSN 0009-2665. doi: 10.1021/acs.chemrev.9b00073. URL https://doi.org/10.1021/acs.chemrev.9b00073.

[27] K N Houk and Paul Ha-Yeon Cheong. Computational prediction of small-molecule catalysts. *Nature*, 455(7211):309–313, 2008. ISSN 1476-4687. doi: 10.1038/nature07368. URL https://doi.org/10.1038/nature07368.

[28] Althea S.-K. Tsang, Italo A Sanhueza, and Franziska Schoenebeck. Combining Experimental and Computational Studies to Understand and Predict Reactivities of Relevance to Homogeneous Catalysis. *Chemistry – A European Journal*, 20(50):16432–16441, dec 2014. ISSN 0947-6539. doi: https://doi.org/10.1002/chem.201404725. URL https://doi.org/10.1002/chem.201404725.

[29] Jeremy N Harvey, Fahmi Himo, Feliu Maseras, and Lionel Perrin. Scope and Challenge of Computational Methods for Studying Mechanism and Reactivity in Homogeneous Catalysis. *ACS Catalysis*, 9(8):6803–6813, aug 2019. doi: 10.1021/acscatal.9b01537. URL https://doi.org/10.1021/acscatal.9b01537.

[30] Theresa Sperger, Italo A Sanhueza, Indrek Kalvet, and Franziska Schoenebeck. Computational Studies of Synthetically Relevant Homogeneous Organometallic Catalysis Involving Ni, Pd, Ir, and Rh: An Overview of Commonly Employed DFT Methods and Mechanistic Insights. *Chemical Reviews*, 115(17): 9532–9586, sep 2015. ISSN 0009-2665. doi: 10.1021/acs.chemrev.5b00163. URL https://doi.org/10.1021/acs.chemrev.5b00163.

[31] Yexenia Nieves-Quinones and Daniel A Singleton. Dynamics and the Regiochemistry of Nitration of Toluene. *Journal of the American Chemical Society*, 138(46):15167–15176, nov 2016. ISSN 0002-7863. doi: 10.1021/jacs.6b07328. URL https://doi.org/10.1021/jacs.6b07328.

[32] Dieter Cremer. Density functional theory: coverage of dynamic and non-dynamic electron correlation effects. *Molecular Physics*, 99(23):1899–1940, dec 2001. ISSN 0026-8976. doi: 10.1080/00268970110083564. URL https://doi.org/10.1080/00268970110083564.

[33] Ana G Maldonado and Gadi Rothenberg. Predictive modeling in homogeneous catalysis: a tutorial. *Chemical Society Reviews*, 39(6):1891–1902, 2010. ISSN 0306-0012. doi: 10.1039/B921393G. URL http://dx.doi.org/10.1039/B921393G.

[34] Jos A. Hageman, Johan A. Westerhuis, Hans-Werner Frühauf, and Gadi Rothenberg. Design and Assembly of Virtual Homogeneous Catalyst Libraries –Towards in silico Catalyst Optimisation. *Advanced Synthesis & Catalysis*, 348(3):361–369, feb 2006. ISSN 1615-4150. doi: https://doi.org/10.1002/adsc.200505299. URL https://doi.org/10.1002/adsc.200505299.

[35] Enrico Burello, David Farrusseng, and Gadi Rothenberg. Combinatorial Explosion in Homogeneous Catalysis: Screening 60,000 Cross-Coupling Reactions. *Advanced Synthesis & Catalysis*, 346(13-15):1844–1853, dec 2004. ISSN 1615-4150. doi: https://doi.org/10.1002/adsc.200404170. URL https://doi.org/10.1002/adsc.200404170.

[36] Enrico Burello and Gadi Rothenberg. In Silico Design in Homogeneous Catalysis Using Descriptor Modelling, 2006.

[37] Nasser Goudarzi and Mohammad Goodarzi. Prediction of the acidic dissociation constant (pKa) of some organic compounds using linear and nonlinear QSPR methods. *Molecular Physics*, 107(14):1495–1503, jul 2009. ISSN 0026-8976. doi: 10.1080/00268970902950394. URL https://doi.org/10.1080/00268970902950394.

[38] Xuanyi Li, Yinqiu Xu, Hequan Yao, and Kejiang Lin. Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors. *Journal of Cheminformatics*, 12(1):42, 2020. ISSN 1758-2946. doi: 10.1186/s13321-020-00446-3. URL https://doi.org/10.1186/s13321-020-00446-3.

[39] Pascal Friederich, Gabriel dos Passos Gomes, Riccardo De Bin, Alán Aspuru-Guzik, and David Balcells. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chemical Science*, 11(18):4584–4601, 2020. ISSN 2041-6520. doi: 10.1039/D0SC00445F. URL http://dx.doi.org/10.1039/D0SC00445F.

[40] Malini Ravi, Anton J Hopfinger, Robert E Hormann, and Laurence Dinan. 4D-QSAR Analysis of a Set of Ecdysteroids and a Comparison to CoMFA Modeling. *Journal of Chemical Information and Computer Sciences*, 41(6):1587–1604, nov 2001. ISSN 0095-2338. doi: 10.1021/ci010076u. URL https://doi.org/10.1021/ci010076u.

[41] B T Luke. Comparison of three different QSAR/QSPR generation techniques. *Journal of Molecular Structure: THEOCHEM*, 468(1):13–20, 1999. ISSN 0166-1280. doi: https://doi.org/10.1016/S0166-1280(98)00492-8. URL http://www.sciencedirect.com/science/article/pii/S0166128098004928.

[42] Pierre Bruneau. Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets. *Journal of Chemical Information and Computer Sciences*, 41(6):1605–1616, nov 2001. ISSN 0095-2338. doi: 10.1021/ci010363y. URL https://doi.org/10.1021/ci010363y.

[43] Alan R Katritzky, Ruslan Petrukhin, Douglas Tatham, Subhash Basak, Emilio Benfenati, Mati Karelson, and Uko Maran. Interpretation of Quantitative Structure-Property and -Activity Relationships. *Journal of Chemical Information and Computer Sciences*, 41(3):679–685, may 2001. ISSN 0095-2338. doi: 10.1021/ci000134w. URL https://doi.org/10.1021/ci000134w.

[44] Viviana Consonni, Roberto Todeschini, Manuela Pavan, and Paola Gramatica. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies. *Journal of Chemical Information and Computer Sciences*, 42(3):693–705, may 2002. ISSN 0095-2338. doi: 10.1021/ci0155053. URL https://doi.org/10.1021/ci0155053.

[45] Mohammad Goodarzi and Matheus P. Freitas. Augmented Three-mode MIA-QSAR Modeling for a Series of Anti-HIV-1 Compounds. *QSAR & Combinatorial Science*, 27(9):1092–1097, sep 2008. ISSN 1611-020X. doi: https://doi.org/10.1002/qsar.200810030. URL https://doi.org/10.1002/qsar.200810030.

[46] G Krenkel, E A Castro, and A A Toropov. Improved molecular descriptors to calculate boiling points based on the optimization of correlation weights of local graph invariants. *Journal of Molecular Structure: THEOCHEM*, 542(1):107–113, 2001. ISSN 0166-1280. doi: https://doi.org/10.1016/S0166-1280(00)00822-8. URL http://www.sciencedirect.com/science/article/pii/S0166128000008228.

[47] Stefan H Unger. Molecular connectivity in structure–activity analysis. By Lemont B. Kier and Lowell H. Hall. Wiley: New York. 1986. xvii + 262 pp. ISBN 0-471-90983-1. $59.95. *Journal of Pharmaceutical Sciences*, 76(3):269–270, mar 1987. ISSN 0022-3549. doi: https://doi.org/10.1002/jps.2600760325. URL https://doi.org/10.1002/jps.2600760325.

[48] Mohammad Goodarzi, Matheus P Freitas, and Richard Jensen. Feature Selection and Linear/Nonlinear Regression Methods for the Accurate Prediction of Glycogen Synthase Kinase-3β Inhibitory Activities. *Journal of Chemical Information and Modeling*, 49(4):824–832, apr 2009. ISSN 1549-9596. doi: 10.1021/ci9000103. URL https://doi.org/10.1021/ci9000103.

[49] E Schrödinger. An Undulatory Theory of the Mechanics of Atoms and Molecules. *Physical Review*, 28(6):1049–1070, dec 1926. doi: 10.1103/PhysRev.28.1049. URL https://link.aps.org/doi/10.1103/PhysRev.28.1049.

[50] Vivek Sinha. *The molecular basis of clean energy*. PhD thesis, University of Amsterdam, 2019. URL https://hdl.handle.net/11245.1/ff73daec-b6f3-4a5f-b2b4-c85261131fce.

[51] Wolfram Koch and Max Holthausen. *A Chemist's Guide to Density Functional Theory*. Wiley Online Books, jul 2001. ISBN 9783527600045. doi: https://doi.org/10.1002/3527600043.part1. URL https://doi.org/10.1002/3527600043.part1.

[52] A Szabo and N S Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. McGraw-Hill, 1989. ISBN 9780070627390. URL https://books.google.nl/books?id=tFOxQgAACAAJ.

[53] D R Hartree. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(1):89–110, 1928. ISSN 0305-0041. doi: DOI:10.1017/S0305004100011919. URL https://www.cambridge.org/core/article/wave-mechanics-of-an-atom-with-a-noncoulomb-central-field-part-i-theory-and-methods/C9417AC1CEC84B934C1EA4C4B8401FEF.

[54] P Rushton. *Towards a non-local density functional description of exchange and correlation.* PhD thesis, Durham University, 2002.

[55] Frank Neese. Prediction of molecular properties and molecular spectroscopy with density functional theory: From fundamental theory to exchange-coupling. *Coordination Chemistry Reviews*, 253 (5):526–563, 2009. ISSN 0010-8545. doi: https://doi.org/10.1016/j.ccr.2008.05.014. URL https://www.sciencedirect.com/science/article/pii/S0010854508001197.

[56] Adrian E Feiguin. Phys 5870: Modern Computational Methods in Solids, 2009. URL https://web.northeastern.edu/afeiguin/phys5870/phys5870/phys5870.html.

[57] A D Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, 38(6):3098–3100, sep 1988. doi: 10.1103/PhysRevA.38.3098. URL https://link.aps.org/doi/10.1103/PhysRevA.38.3098.

[58] John P Perdew. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Physical Review B*, 33(12):8822–8824, jun 1986. doi: 10.1103/PhysRevB.33.8822. URL https://link.aps.org/doi/10.1103/PhysRevB.33.8822.

[59] Carlo Adamo and Vincenzo Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *The Journal of Chemical Physics*, 110(13):6158–6170, mar 1999. ISSN 0021-9606. doi: 10.1063/1.478522. URL https://doi.org/10.1063/1.478522.

[60] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Physical Review Letters*, 77(18):3865–3868, oct 1996. doi: 10.1103/PhysRevLett.77.3865. URL https://link.aps.org/doi/10.1103/PhysRevLett.77.3865.

[61] Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305, 2005. ISSN 1463-9076. doi: 10.1039/B508541A. URL http://dx.doi.org/10.1039/B508541A.

[62] Zaheer Ul-Haq. Introduction to geometry optimization, 2016. URL https://th.fhi-berlin.mpg.de/sitesub/meetings/dft-workshop-2016/uploads/Meeting/May_6_Qasmi.pdf.

[63] Jin Zhang, Haiyang Zhang, Tao Wu, Qi Wang, and David van der Spoel. Comparison of Implicit and Explicit Solvent Models for the Calculation of Solvation Free Energy in Organic Solvents. *Journal of Chemical Theory and Computation*, 13(3):1034–1043, mar 2017. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b00169. URL https://doi.org/10.1021/acs.jctc.7b00169.

[64] Jacopo Tomasi and Maurizio Persico. Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent. *Chemical Reviews*, 94(7):2027–2094, nov 1994. ISSN 0009-2665. doi: 10.1021/cr00031a013. URL https://doi.org/10.1021/cr00031a013.

[65] Jacopo Tomasi, Benedetta Mennucci, and Roberto Cammi. Quantum Mechanical Continuum Solvation Models. *Chemical Reviews*, 105(8):2999–3094, aug 2005. ISSN 0009-2665. doi: 10.1021/cr9904009. URL https://doi.org/10.1021/cr9904009.

[66] Modesto Orozco and F Javier Luque. Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems. *Chemical Reviews*, 100(11):4187–4226, nov 2000. ISSN 0009-2665. doi: 10.1021/cr990052a. URL https://doi.org/10.1021/cr990052a.

[67] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics*, 132(15):154104, apr 2010. ISSN 0021-9606. doi: 10.1063/1.3382344. URL https://doi.org/10.1063/1.3382344.

[68] W Kohn and L J Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review*, 140(4A):A1133–A1138, nov 1965. doi: 10.1103/PhysRev.140.A1133. URL https://link.aps.org/doi/10.1103/PhysRev.140.A1133.

[69] Robert G Parr. Density Functional Theory of Atoms and Molecules. In Kenichi Fukui and Bernard Pullman, editors, *Horizons of Quantum Chemistry*, pages 5–15, Dordrecht, 1980. Springer Netherlands. ISBN 978-94-009-9027-2.

[70] Fernand Spiegelman, Nathalie Tarrat, Jérôme Cuny, Leo Dontot, Evgeny Posenitskiy, Carles Martí, Aude Simon, and Mathias Rapacioli. Density-functional tight-binding: basic concepts and applications to molecules and clusters. *Advances in Physics: X*, 5(1):1710252, jan 2020. ISSN null. doi: 10.1080/23746149.2019.1710252. URL https://doi.org/10.1080/23746149.2019.1710252.

[71] G Seifert. Tight-Binding Density Functional Theory: An Approximate Kohn-Sham DFT Scheme. *The Journal of Physical Chemistry A*, 111(26):5609–5613, jul 2007. ISSN 1089-5639. doi: 10.1021/jp069056r. URL https://doi.org/10.1021/jp069056r.

[72] Christoph Bannwarth, Eike Caldeweyher, Sebastian Ehlert, Andreas Hansen, Philipp Pracht, Jakob Seibert, Sebastian Spicher, and Stefan Grimme. Extended tight-binding quantum chemistry methods. *WIREs Computational Molecular Science*, n/a(n/a):e01493, aug 2020. ISSN 1759-0876. doi: 10.1002/wcms.1493. URL https://doi.org/10.1002/wcms.1493.

[73] M Elstner, D Porezag, G Jungnickel, J Elsner, M Haugk, Th. Frauenheim, S Suhai, and G Seifert. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Physical Review B*, 58(11):7260–7268, sep 1998. doi: 10.1103/PhysRevB.58.7260. URL https://link.aps.org/doi/10.1103/PhysRevB.58.7260.

[74] Michael Gaus, Qiang Cui, and Marcus Elstner. DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *Journal of Chemical Theory and Computation*, 7(4):931–948, apr 2011. ISSN 1549-9618. doi: 10.1021/ct100684s. URL https://doi.org/10.1021/ct100684s.

[75] Yang, Haibo Yu, Darrin York, Qiang Cui, and Marcus Elstner. Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method: Third-Order Expansion of the Density Functional Theory Total Energy and Introduction of a Modified Effective Coulomb Interaction. *The Journal of Physical Chemistry A*, 111(42):10861–10873, oct 2007. ISSN 1089-5639. doi: 10.1021/jp074167r. URL https://doi.org/10.1021/jp074167r.

[76] Eike Caldeweyher, Sebastian Ehlert, Andreas Hansen, Hagen Neugebauer, Sebastian Spicher, Christoph Bannwarth, and Stefan Grimme. A generally applicable atomic-charge dependent London dispersion correction. *The Journal of Chemical Physics*, 150(15):154122, apr 2019. ISSN 0021-9606. doi: 10.1063/1.5090222. URL https://doi.org/10.1063/1.5090222.

[77] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671, mar 2019. ISSN 1549-9618. doi: 10.1021/acs.jctc.8b01176. URL https://doi.org/10.1021/acs.jctc.8b01176.

[78] Jeng-Da Chai. Density functional theory with fractional orbital occupations. *The Journal of Chemical Physics*, 136(15):154104, apr 2012. ISSN 0021-9606. doi: 10.1063/1.3703894. URL https://doi.org/10.1063/1.3703894.

[79] Christof Köhler, Gotthard Seifert, and Thomas Frauenheim. Density functional based calculations for Fen (n ≤ 32). *Chemical Physics*, 309(1):23–31, 2005. ISSN 0301-0104. doi: https://doi.org/10.1016/j.chemphys.2004.03.034. URL https://www.sciencedirect.com/science/article/pii/S0301010404002563.

[80] Andreas Klamt. The COSMO and COSMO-RS solvation models. *WIREs Computational Molecular Science*, 1(5):699–709, sep 2011. ISSN 1759-0876. doi: https://doi.org/10.1002/wcms.56. URL https://doi.org/10.1002/wcms.56.

[81] Andrew R Leach and Michael M Hann. The in silico world of virtual libraries. *Drug Discovery Today*, 5 (8):326–336, 2000. ISSN 1359-6446. doi: https://doi.org/10.1016/S1359-6446(00)01516-6. URL https://www.sciencedirect.com/science/article/pii/S1359644600015166.

[82] Ingo Vogt and Jürgen Bajorath. Design and Exploration of Target-Selective Chemical Space Representations. *Journal of Chemical Information and Modeling*, 48(7):1389–1395, jul 2008. ISSN 1549-9596. doi: 10.1021/ci800106e. URL https://doi.org/10.1021/ci800106e.

[83] Alexandre Varnek, Denis Fourches, Dragos Horvath, Olga Klimchuk, Cedric Gaudin, Philippe Vayer, Vitaly Solovev, Frank Hoonakker, Igor Tetko, and Gilles Marcou. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Current Computer - Aided Drug Design*, 4:191–198, sep 2008. doi: 10.2174/157340908785747465.

[84] Igor I Baskin, Vitaly P Solov'ev, Alexander A Bagatur'yants, and Alexandre Varnek. Predictive cartography of metal binders using generative topographic mapping. *Journal of Computer-Aided Molecular Design*, 31 (8):701–714, 2017. ISSN 1573-4951. doi: 10.1007/s10822-017-0033-6. URL https://doi.org/10.1007/s10822-017-0033-6.

[85] V P Solov'ev, A Varnek, and G Wipff. Modeling of Ion Complexation and Extraction Using Substructural Molecular Fragments. *Journal of Chemical Information and Computer Sciences*, 40(3):847–858, may 2000. ISSN 0095-2338. doi: 10.1021/ci9901340. URL https://doi.org/10.1021/ci9901340.

[86] Alexandre Varnek and Alex Tropsha. *Chemoinformatics approaches to virtual screening*. Royal Society of Chemistry, 2008. ISBN 0854041443.

[87] Kong T. Nguyen, Lorenz C. Blum, Ruud van Deursen, and Jean-Louis Reymond. Classification of Organic Molecules by Molecular Quantum Numbers. *ChemMedChem*, 4(11):1803–1805, nov 2009. ISSN 1860-7179. doi: https://doi.org/10.1002/cmdc.200900317. URL https://doi.org/10.1002/cmdc.200900317.

[88] Jean-Louis Reymond, Ruud van Deursen, Lorenz C Blum, and Lars Ruddigkeit. Chemical space as a source for new drugs. *MedChemComm*, 1(1):30–38, 2010. ISSN 2040-2503. doi: 10.1039/C0MD00020E. URL http://dx.doi.org/10.1039/C0MD00020E.

[89] Dmitry S Karlov, Sergey Sosnin, Igor V Tetko, and Maxim V Fedorov. Chemical space exploration guided by deep neural networks. *RSC Advances*, 9(9):5151–5157, 2019. doi: 10.1039/C8RA10182E. URL http://dx.doi.org/10.1039/C8RA10182E.

[90] Jon Paul Janet and Heather J Kulik. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure–Property Relationships. *The Journal of Physical Chemistry A*, 121(46): 8939–8954, nov 2017. ISSN 1089-5639. doi: 10.1021/acs.jpca.7b08750. URL https://doi.org/10.1021/acs.jpca.7b08750.

[91] Tohid N Borhani, Salvador García-Muñoz, Carla Vanesa Luciani, Amparo Galindo, and Claire S Adjiman. Hybrid QSPR models for the prediction of the free energy of solvation of organic solute/solvent pairs. *Physical Chemistry Chemical Physics*, 21(25):13706–13720, 2019. ISSN 1463-9076. doi: 10.1039/C8CP07562J. URL http://dx.doi.org/10.1039/C8CP07562J.

[92] Kai S Exner. A Universal Descriptor for the Screening of Electrode Materials for Multiple-Electron Processes: Beyond the Thermodynamic Overpotential. *ACS Catalysis*, 10(21):12607–12617, nov 2020. doi: 10.1021/acscatal.0c03865. URL https://doi.org/10.1021/acscatal.0c03865.

[93] J S Griffith and L E Orgel. Ligand-field theory. *Quarterly Reviews, Chemical Society*, 11(4):381–393, 1957. ISSN 0009-2681. doi: 10.1039/QR9571100381. URL http://dx.doi.org/10.1039/QR9571100381.

[94] Qi Yuan, Alejandro Santana-Bonilla, Martijn A Zwijnenburg, and Kim E Jelfs. Molecular generation targeting desired electronic properties via deep generative models. *Nanoscale*, 12(12):6744–6758, 2020. ISSN 2040-3364. doi: 10.1039/C9NR10687A. URL http://dx.doi.org/10.1039/C9NR10687A.

[95] Jutta Rogal. *Stability, Composition and Function of Palladium Surfaces in Oxidizing Environments*. PhD thesis, Freie Universitat Berlin, 2006. URL http://dx.doi.org/10.17169/refubium-5529.

[96] Catherine E Housecroft and A G Sharpe. *Housecroft inorganic chemistry*. Prentice Hall, Harlow, 3 edition, 2007. ISBN 9780131755536 0131755536.

[97] Jun-ichi Fujisawa. An unusual mechanism for HOMO–LUMO gap narrowing in a minimal near-IR dye generated by the deprotonation of bis(dicyanomethylene)indan. *Chemical Physics Letters*, 608: 355–359, 2014. ISSN 0009-2614. doi: https://doi.org/10.1016/j.cplett.2014.06.012. URL https://www.sciencedirect.com/science/article/pii/S0009261414005065.

[98] Daan Frenkel and Berend Smit. Chapter 4 - Molecular Dynamics Simulations. In Daan Frenkel and Berend B T Smit, editors, *Understanding Molecular Simulation (Second Edition)*, pages 63–107. Academic Press, San Diego, 2002. ISBN 978-0-12-267351-1. doi: https://doi.org/10.1016/B978-012267351-1/50006-7. URL https://www.sciencedirect.com/science/article/pii/B9780122673511500067.

[99] A K Rappe, C J Casewit, K S Colwell, W A Goddard, and W M Skiff. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*, 114(25):10024–10035, dec 1992. ISSN 0002-7863. doi: 10.1021/ja00051a040. URL https://doi.org/10.1021/ja00051a040.

[100] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, jul 2004. ISSN 0192-8651. doi: https://doi.org/10.1002/jcc.20035. URL https://doi.org/10.1002/jcc.20035.

[101] Scott J Weiner, Peter A Kollman, David A Case, U Chandra Singh, Caterina Ghio, Guliano Alagona, Salvatore Profeta, and Paul Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784, feb 1984. ISSN 0002-7863. doi: 10.1021/ja00315a051. URL https://doi.org/10.1021/ja00315a051.

[102] Scott J Weiner, Peter A Kollman, Dzung T Nguyen, and David A Case. An all atom force field for simulations of proteins and nucleic acids. *Journal of Computational Chemistry*, 7(2):230–252, apr 1986. ISSN 0192-8651. doi: https://doi.org/10.1002/jcc.540070216. URL https://doi.org/10.1002/jcc.540070216.

[103] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, may 1995. ISSN 0002-7863. doi: 10.1021/ja00124a002. URL https://doi.org/10.1021/ja00124a002.

[104] Junmei Wang, Piotr Cieplak, and Peter A Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, 21(12):1049–1074, sep 2000. ISSN 0192-8651. doi: https://doi.org/10.1002/1096-987X(200009)21:12<1049::AID-JCC3>3.0.CO;2-F. URL https://doi.org/10.1002/1096-987X(200009)21:12%3C1049::AID-JCC3%3E3.0.COhttp://2-f.

[105] Larry A Curtiss, Krishnan Raghavachari, Gary W Trucks, and John A Pople. Gaussian-2 theory for molecular energies of first- and second-row compounds. *The Journal of Chemical Physics*, 94(11):7221–7230, jun 1991. ISSN 0021-9606. doi: 10.1063/1.460205. URL https://doi.org/10.1063/1.460205.

[106] Jimmy Charnley Kromann. Calculate Root-mean-square deviation (RMSD) of Two Molecules Using Rotation, GitHub, v1.3.2, 2020. URL https://github.com/charnley/rmsd/releases/tag/rmsd-1.3.2.

[107] W Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, sep 1976. ISSN 0567-7394. URL https://doi.org/10.1107/S0567739476001873.

[108] Marco Taboga. "Cross-covariance matrix", Lectures on probability theory and mathematical statistics, Third edition. Kindle Direct Publishing. Online appendix., 2017. URL https://www.statlect.com/glossary/cross-covariance-matrix.

[109] Reza Bagheri. Understanding Singular Value Decomposition and its Application in Data Science, 2020. URL https://towardsdatascience.com/understanding-singular-value-decomposition-and-its-application-in-data-science-388a54be95d.

[110] Michael W Walker, Lejun Shao, and Richard A Volz. Estimating 3-D location parameters using dual number quaternions. *CVGIP: Image Understanding*, 54(3):358–367, 1991. ISSN 1049-9660. doi: https://doi.org/10.1016/1049-9660(91)90036-O. URL http://www.sciencedirect.com/science/article/pii/104996609190036O.

[111] Evangelos A Coutsias, Chaok Seok, and Ken A Dill. Using quaternions to calculate RMSD. *Journal of Computational Chemistry*, 25(15):1849–1857, nov 2004. ISSN 0192-8651. doi: https://doi.org/10.1002/jcc.20110. URL https://doi.org/10.1002/jcc.20110.

[112] Illustr - Wikimedia Commons. Image of regular tetrahedron ABCD and its circumscribed sphere, 2014. URL https://commons.wikimedia.org/wiki/File:%D0%92%D0%BF%D0%B8%D1%81%D0%B0%D0%BD%D0%BD%D1%8B%D0%B9_%D1%82%D0%B5%D1%82%D1%80%D0%B0%D1%8D%D0%B4%D1%80.svg.

[113] Jur van den Berg. Calculate Rotation Matrix to align Vector A to Vector B in 3d? Mathematics Stack Exchange, 2016. URL https://math.stackexchange.com/q/476311.

[114] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. The Open Babel Package, version 2.4.1, 2016. URL https://openbabel.org/.

[115] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, 2011. ISSN 1758-2946. doi: 10.1186/1758-2946-3-33. URL https://doi.org/10.1186/1758-2946-3-33.

[116] Annika Krieger, Vivek Sinha, Adarsh Kalikadien, and Evgeny A Pidko. Metal-ligand cooperative activation of HX (X=H, Br, OR) bond on Mn based pincercomplexes. *Zeitschrift für anorganische und allgemeine Chemie*, 2021. doi: inpress.

[117] Thomas Bligaard, R Morris Bullock, Charles T Campbell, Jingguang G Chen, Bruce C Gates, Raymond J Gorte, Christopher W Jones, William D Jones, John R Kitchin, and Susannah L Scott. Toward Benchmarking in Catalysis Science: Best Practices, Challenges, and Opportunities. *ACS Catalysis*, 6(4):2590–2602, apr 2016. doi: 10.1021/acscatal.6b00183. URL https://doi.org/10.1021/acscatal.6b00183.

[118] Stephan N Steinmann and Clemence Corminboeuf. Comprehensive Benchmarking of a Density-Dependent Dispersion Correction. *Journal of Chemical Theory and Computation*, 7(11):3567–3577, nov 2011. ISSN 1549-9618. doi: 10.1021/ct200602x. URL https://doi.org/10.1021/ct200602x.

[119] Biswajit Santra, Angelos Michaelides, and Matthias Scheffler. On the accuracy of density-functional theory exchange-correlation functionals for H bonds in small water clusters: Benchmarks approaching the complete basis set limit. *The Journal of Chemical Physics*, 127(18):184104, nov 2007. ISSN 0021-9606. doi: 10.1063/1.2790009. URL https://doi.org/10.1063/1.2790009.

[120] Noa Marom, Alexandre Tkatchenko, Mariana Rossi, Vivekanand V Gobre, Oded Hod, Matthias Scheffler, and Leeor Kronik. Dispersion Interactions with Density-Functional Theory: Benchmarking Semiempirical and Interatomic Pairwise Corrected Density Functionals. *Journal of Chemical Theory and Computation*, 7(12):3944–3951, dec 2011. ISSN 1549-9618. doi: 10.1021/ct2005616. URL https://doi.org/10.1021/ct2005616.

[121] Martin Korth and Stefan Grimme. "Mindless" DFT Benchmarking. *Journal of Chemical Theory and Computation*, 5(4):993–1003, apr 2009. ISSN 1549-9618. doi: 10.1021/ct800511q. URL https://doi.org/10.1021/ct800511q.

[122] R Erik Plata and Daniel A Singleton. A Case Study of the Mechanism of Alcohol-Mediated Morita Baylis–Hillman Reactions. The Importance of Experimental Observations. *Journal of the American Chemical Society*, 137(11):3811–3826, mar 2015. ISSN 0002-7863. doi: 10.1021/ja5111392. URL https://doi.org/10.1021/ja5111392.

[123] Benjamin Meyer, Boodsarin Sawatlon, Stefan Heinen, O Anatole von Lilienfeld, and Clémence Corminboeuf. Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. *Chemical Science*, 9(35):7069–7077, 2018. ISSN 2041-6520. doi: 10.1039/C8SC01949E. URL http://dx.doi.org/10.1039/C8SC01949E.

[124] Roger Parsons. The rate of electrolytic hydrogen evolution and the heat of adsorption of hydrogen. *Transactions of the Faraday Society*, 54(0):1053–1063, 1958. ISSN 0014-7672. doi: 10.1039/TF9585401053. URL http://dx.doi.org/10.1039/TF9585401053.

[125] H Gerischer. Mechanismus der Elektrolytischen Wasserstoffabscheidung und Adsorptionsenergie von Atomarem Wasserstoff. *Bulletin des Sociétés Chimiques Belges*, 67(7-8):506–527, jan 1958. ISSN 0037-9646. doi: https://doi.org/10.1002/bscb.19580670714. URL https://doi.org/10.1002/bscb.19580670714.

[126] Paul Sabatier. *La Catalyse en chimie organique, par Paul Sabatier ...* C. Beranger, Paris; Liege, 1913.

[127] Thomas Blaschke, Marcus Olivecrona, Ola Engkvist, Jürgen Bajorath, and Hongming Chen. Application of Generative Autoencoder in De Novo Molecular Design. *Molecular Informatics*, 37(1-2):1700123, jan 2018. ISSN 1868-1743. doi: https://doi.org/10.1002/minf.201700123. URL https://doi.org/10.1002/minf.201700123.

[128] O Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. Exploring chemical compound space with quantum-based machine learning. *Nature Reviews Chemistry*, 4(7):347–358, 2020. ISSN 2397-3358. doi: 10.1038/s41570-020-0189-9. URL https://doi.org/10.1038/s41570-020-0189-9.

[129] Weinan E, Jiequn Han, and Linfeng Zhang. Integrating Machine Learning with Physics-Based Modeling. *arXiv e-prints*, page arXiv:2006.02619, jun 2020.

[130] Andrea Grisafi and Michele Ceriotti. Incorporating long-range physics in atomic-scale machine learning. *The Journal of Chemical Physics*, 151(20):204105, nov 2019. ISSN 1089-7690. doi: 10.1063/1.5128375. URL http://dx.doi.org/10.1063/1.5128375.

[131] Sebastian Dick and Marivi Fernandez-Serra. Machine learning accurate exchange and correlation functionals of the electronic density. *Nature Communications*, 11(1):3509, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17265-7. URL https://doi.org/10.1038/s41467-020-17265-7.

[132] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017. ISSN 0036-8075. doi: 10.1126/science.aag2302. URL https://science.sciencemag.org/content/355/6325/602.

[133] Tom Westerhout, Nikita Astrakhantsev, Konstantin S. Tikhonov, Mikhail I. Katsnelson, and Andrey A. Bagrov. Generalization properties of neural network approximations to frustrated magnet ground states. *Nature Communications*, 11(1), Mar 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-15402-w. URL http://dx.doi.org/10.1038/s41467-020-15402-w.

[134] James Stokes, Josh Izaac, Nathan Killoran, and Giuseppe Carleo. Quantum natural gradient. *Quantum*, 4:269, May 2020. ISSN 2521-327X. doi: 10.22331/q-2020-05-25-269. URL http://dx.doi.org/10.22331/q-2020-05-25-269.

[135] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.*, 120: 143001, Apr 2018. doi: 10.1103/PhysRevLett.120.143001. URL https://link.aps.org/doi/10.1103/PhysRevLett.120.143001.

[136] Kristof T Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8(1):13890, 2017. ISSN 2041-1723. doi: 10.1038/ncomms13890. URL https://doi.org/10.1038/ncomms13890.

[137] Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, David Duvenaud, Dougal Maclaurin, Martin A Blood-Forsythe, Hyun Sik Chae, Markus Einzinger, Dong-Gwang Ha, Tony Wu, Georgios Markopoulos, Soonok Jeon, Hosuk Kang, Hiroshi Miyazaki, Masaki Numata, Sunghan Kim, Wenliang Huang, Seong Ik Hong, Marc Baldo, Ryan P Adams, and Alán Aspuru-Guzik. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials*, 15(10):1120–1127, 2016. ISSN 1476-4660. doi: 10.1038/nmat4717. URL https://doi.org/10.1038/nmat4717.

[138] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, Nov 2020. ISSN 2632-2153. doi: 10.1088/2632-2153/aba947. URL http://dx.doi.org/10.1088/2632-2153/aba947.

Appendices contents

A

Supporting information: *ChemSpaX*: Exploration of chemical space by automated functionalization of molecular scaffold

SUPPORTING INFORMATION

ChemSpaX: exploration of chemical space by automated functionalization of molecular scaffold

Adarsh V. Kalikadien, Evgeny A. Pidko* and Vivek Sinha*

Inorganic Systems Engineering Group, Department of Chemical Engineering, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, The Netherlands

Corresponding authors: Vivek Sinha (V.Sinha@tudelft.nl) and Evgeny A. Pidko (E.A.Pidko@tudelft.nl)

Table of Contents

S1. Remarks

With regards to computational methods following nomenclature is adopted:

GFN2-xTB(GAS): optimization using Grimme's xTB (6.3.3) package.

GFN2-xTB(THF): optimization and hessian calculation in THF using the GBSA solvation model using Grimme's xTB (6.3.3) package.

BP86(GAS): DFT calculations were performed at various levels of theory. Geometry optimizations were performed in the gas phase using BP86 XC functional and def2-SVP basis set. Free energy corrections were obtained via hessian calculations within the harmonic approximation at the same level of theory. This method is denoted as BP86(GAS).

PBE1PBE(thf) (or PBE0(THF)) and BP86(THF): To include the solvent effects the electronic energies were further refined via single point energy calculations using SMD solvation method with THF as solvent. Two XC functionals namely BP86 and PBE1PBE were used with a triple zeta quality basis set (def2-TZVP). The free energy corrections obtained via hessian calculations at BP86(gas) level of theory were added to the electronic energies obtained during the singlepoint energy calculations. These methods are denoted as PBE0(THF) and BP86(THF).

We have calculated pearson correlation coefficient (R^2) related to a linear fitting ($\widehat{y_i} = ax_i + b$) for all case where two methods/quantities (x_i, y_i) were compared in a scatter plot. We computed the root mean squared error (RMSE) related to the linear fit as :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\widehat{y_i} - y_i)^2}{N}}$$

The XYZ files of structures and datasets used for this publication are attached.

ChemSpaX will be made publicly available after publication on our group's Github page (https://github.com/EPiCs-group/) together with a manual.

S2. Observed issues with FF optimization

When functionalizing a complex recursively, it was observed that using Openbabel’s universal force field optimization (UFF) only for geometry optimization would cause hydrogens on the complex to interfere. Which is illustrated in figure 1.

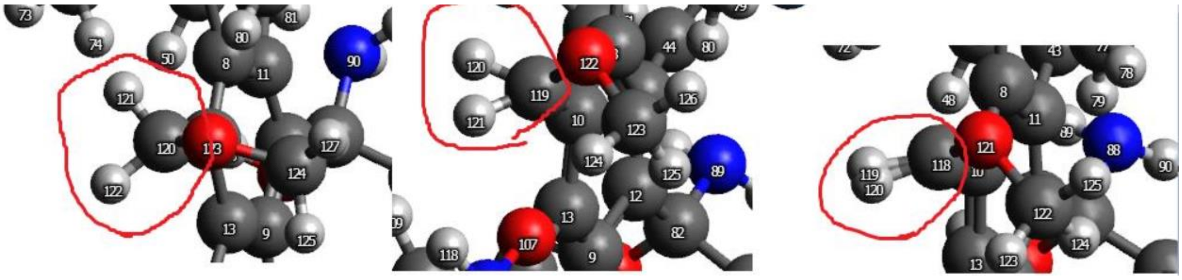
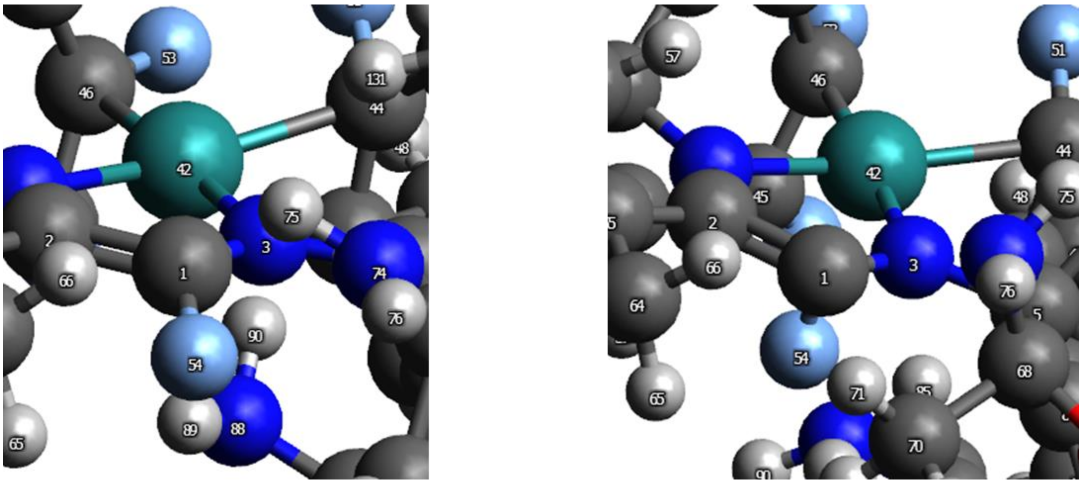


Figure 1. Hydrogen interference upon serial functionalization and optimization with UFF.

In the same scenario, but using Openbabel’s GAFF only for geometry optimization, bonds between carbon and a halogen would have an incorrect angle. Which is shown in figure 2, with the incorrect angle on the right side and the correct angle on the left side of the figure.



Using a combination of GAFF and UFF for geometry optimization resulted in a highly increased probability of an error-free geometry.

S3. Comparison of calculated energies of reaction using xTB or DFT for pincer complexes

RuPNP

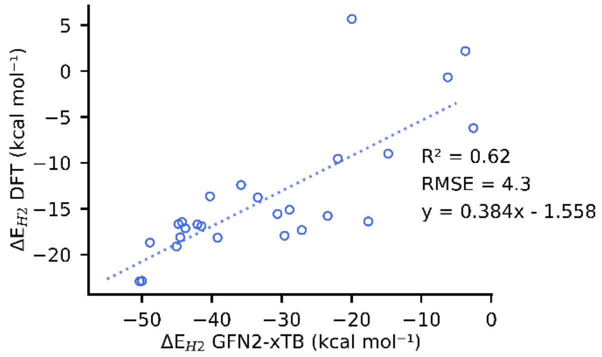


Figure 2. Comparison of ΔE calculated by BP86(GAS) vs GFN2-xTB(GAS).

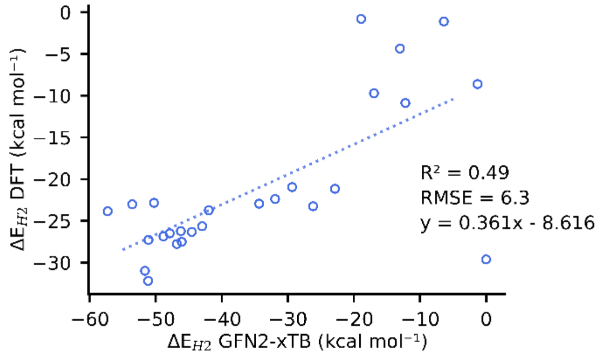


Figure 3. Comparison of ΔE calculated by PBE0(THF) vs GFN2-xTB(THF).

Mn-pincers

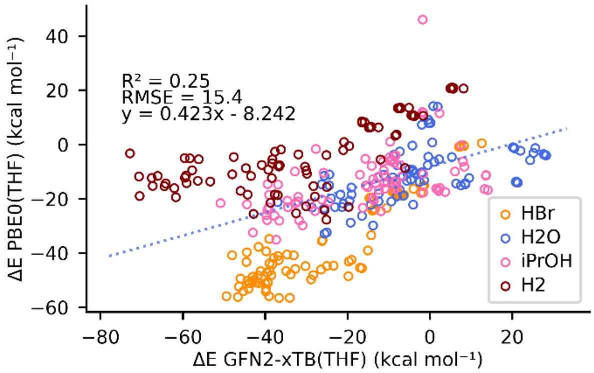


Figure 4. Comparison of ΔE calculated by PBE0(THF) vs GFN2-xTB(THF) for various adducts. The correlation is show for the total x and y dataset.

Linear scaling of ΔE and ΔG for Mn-pincers

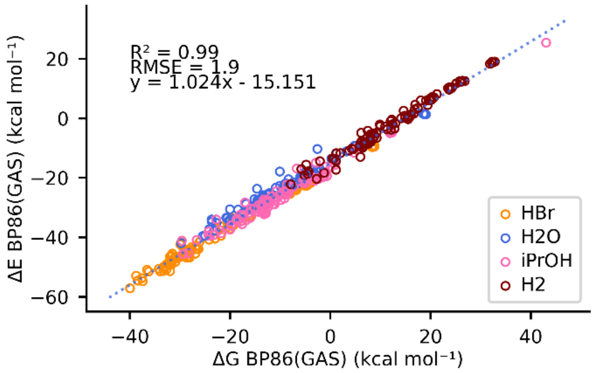


Figure 5. Linear scaling of ΔE against ΔG calculated by BP86(GAS) for various adducts. The correlation shown is for the total x and y dataset, thus the scaling relation holds regardless of the adduct.

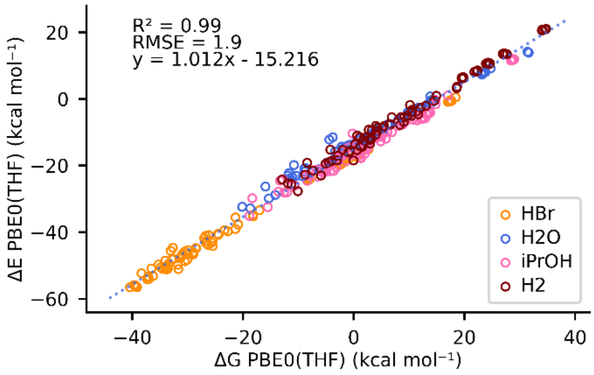


Figure 6. Linear scaling of ΔE against ΔG calculated by PBE0(THF) for various adducts. The correlation shown is for the total x and y dataset, thus the scaling relation holds regardless of the adduct.

S4. Distribution of Gibbs free energy for RuPNP

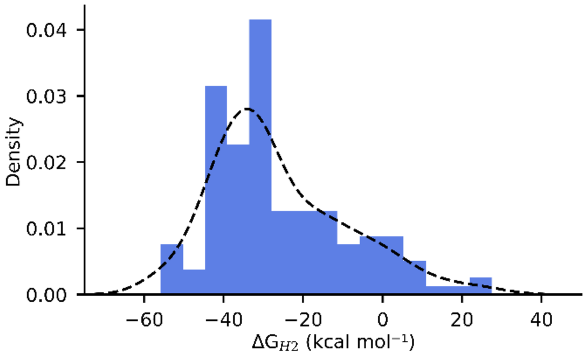


Figure 7. ΔG calculated by GFN2-xTB for all generated RuPNP geometries

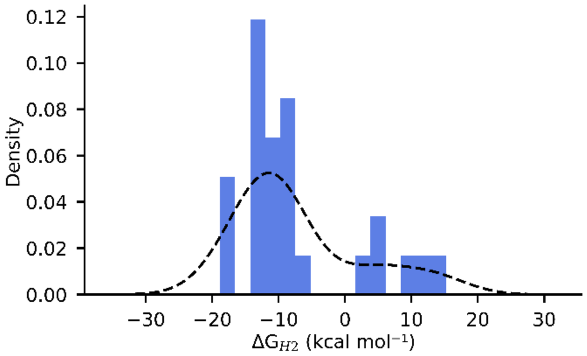


Figure 8. ΔG calculated using PBE0(THF) for 27 selected RuPNP geometries

S5. Distribution of total energy for Mn-pincers

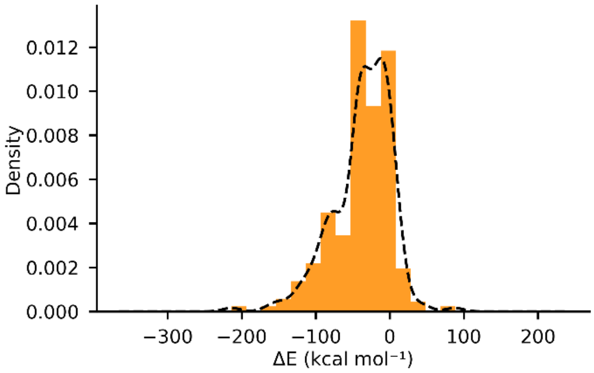


Figure 9. ΔE calculated using BP86(GAS) for geometries generated by ChemSpaX

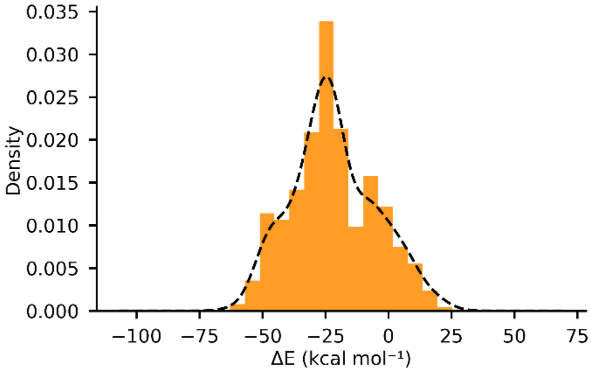


Figure 10. ΔE calculated using BP86(GAS) after full DFT optimization.

S6. Distribution of hRMSD for pincer complexes

RuPNP

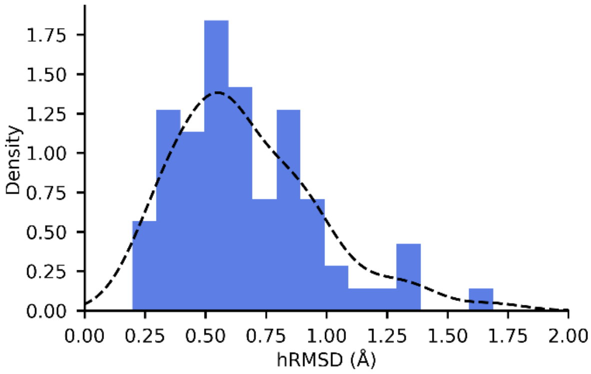


Figure 11. Distribution of hRMSD for ChemSpaX generated structures (newly placed substituents optimized with FF) compared against DFT (BP86) optimized structures.

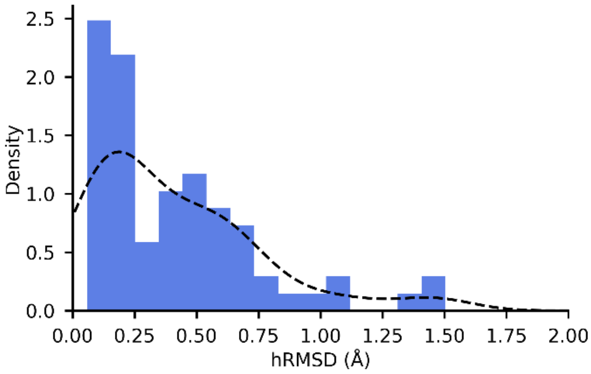


Figure 12. Distribution of hRMSD for GFN2-xTB optimized structures compared against DFT (BP86) optimized structures.

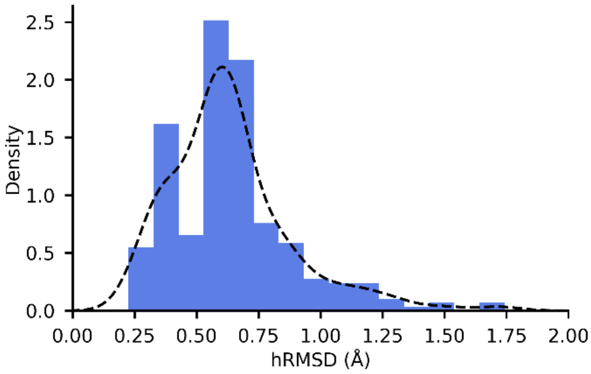


Figure 13. Distribution of hRMSD for ChemSpaX generated structures (newly placed substituents optimized with FF) compared against GFN2-xTB optimized structures.

Mn-pincers

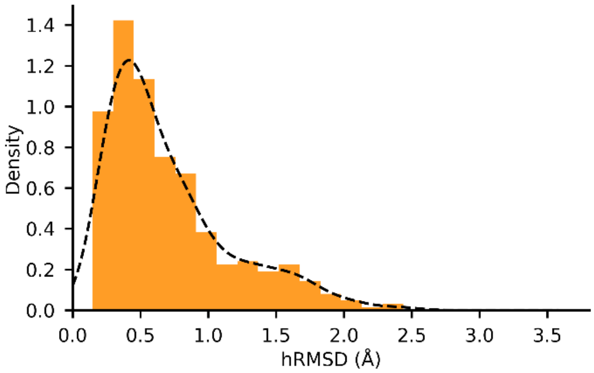


Figure 14. Distribution of hRMSD for ChemSpaX generated structures (newly placed substituents optimized with FF) compared against DFT (BP86) optimized structures. Plotted for all datapoints.

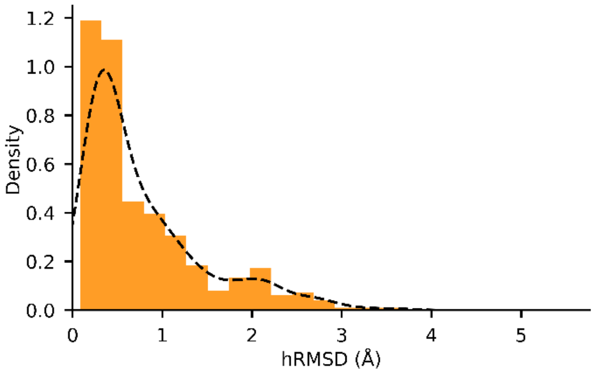


Figure 15. Distribution of hRMSD for GFN2-xTB optimized structures compared against DFT (BP86) optimized structures. Plotted for all datapoints.

S7. Comparison of DFT and xTB calculated HOMO-LUMO gap for Pincer complexes

RuPNP

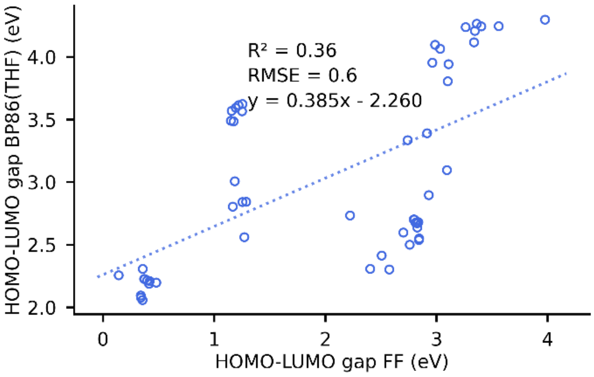


Figure 16. Comparison of HOMO-LUMO gap calculated using a BP86 SP on a FF optimized geometry and a fully DFT optimized geometry.

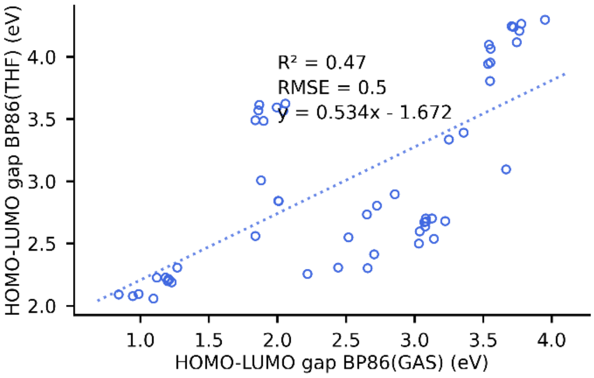


Figure 17. Comparison of HOMO-LUMO gap calculated using BP86(THF) and BP86(GAS), which shows the effect of solvation on the HOMO-LUMO gap.

Mn-pincers

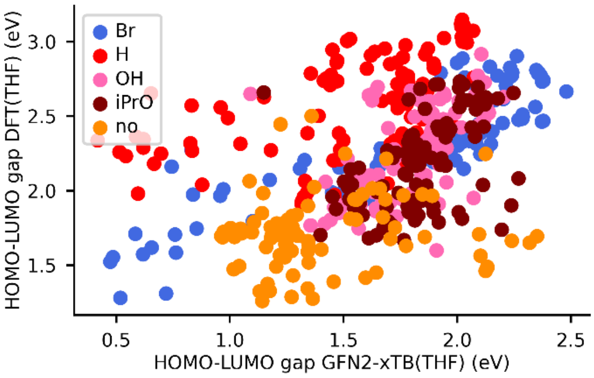


Figure 18. Comparison of HOMO-LUMO gap calculated by BP86(THF) against GFN2-xTB(THF) for various adducts on the metal site.

Table 1. Pearson's correlation coefficient and RMSE of HOMO-LUMO gap comparison for various adducts on the metal site.

Adduct on metal site	R^2	RMSE (eV)
Br	0.74	0.18
H	0.26	0.28
OH	0.32	0.25
iPrO	0.23	0.27
no	0.008	0.23

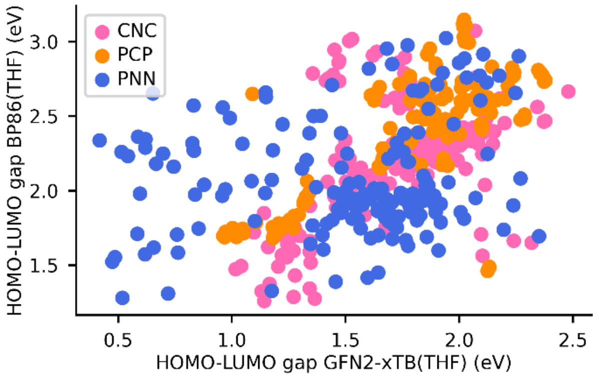


Figure 19. Comparison of HOMO-LUMO gap calculated by BP86(THF) against GFN2-xTB(THF) for various ligand backbones.

Table 2. Pearson's correlation coefficient and RMSE of HOMO-LUMO gap comparison for various ligand backbones

Ligand backbone	R^2	RMSE (eV)
PCP	0.58	0.26
PNN	0.28	0.37
CNC	0.25	0.34

B

Supporting information: Metal-ligand cooperative activation of HX (X=H, Br, OR) bond on Mn based pincer complexes

SUPPORTING INFORMATION

Metal-ligand cooperative activation of HX (X=H,OR) bond on Mn based pincer complexes

Annika M. Krieger, Vivek Sinha*, Adarsh V. Kalikadien, Evgeny A. Pidko*

Inorganic Systems Engineering Group, Department of Chemical Engineering, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, The Netherlands

Corresponding authors: Vivek Sinha (V.Sinha@tudelft.nl) and Evgeny A. Pidko (E.A.Pidko@tudelft.nl)

Table of Contents

S1. Remarks

With regards to computational methods following nomenclature is adopted:

xTB: optimization and hessian calculation in THF using the GBSA solvation model using Grimme's xTB (6.3.3) package.

bp86(gas): DFT calculations were performed at various levels of theory. Geometry optimizations were performed in the gas phase using BP86 XC functional and def2-SVP basis set. Free energy corrections were obtained via hessian calculations within the harmonic approximation at the same level of theory. This method is denoted as bp86(gas).

pbe0(thf) and bp86(thf): To include the solvent effects the electronic energies were further refined via single point energy calculations using SMD solvation method with THF as solvent. Two XC functionals namely BP86 and PBE0 were used with a triple zeta quality basis set (def2-TZVP). The free energy corrections obtained via hessian calculations at BP86(gas) level of theory were added to the electronic energies obtained during the singlepoint energy calculations. These methods are denoted as pbe0(thf) and bp86(thf).

We have calculated pearson correlation coefficient (R^2) related to a linear fitting ($\hat{y}_i = ax_i + b$) for all case where two methods/quantities (x_i, y_i) were compared in a scatter plot. We computed the root mean squared error (RMSE) related to the linear fit as :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}}$$

S2. Computational remarks on xTB optimization

Catalyst structures showing hemilability (bond length between metal and ligand > 2.8 A) in the activated catalyst structure were excluded from the analysis. One entire data set of ligand backbones (PNN) was excluded for this reason. The nature and related chemical reactivity arising from hemilability was not investigated further. A selection of the catalyst structures that show hemilability is depicted in Figure S1.

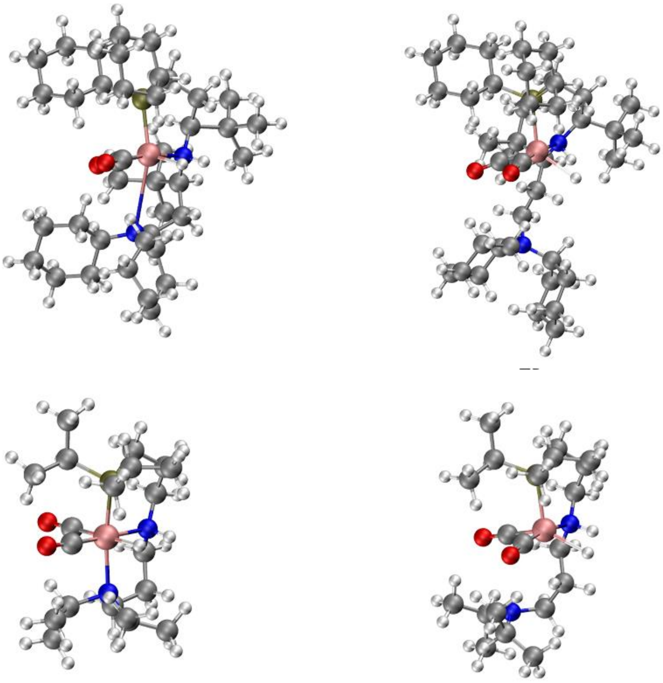
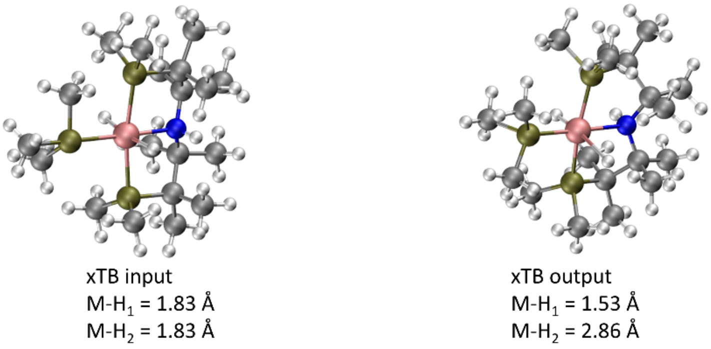


Figure S1. Structure of catalysts showing hemilability after GFN2-xTB optimization.

Additionally, it was observed that some structures showed an elongated M-H bond after geometry optimization with GFN2-xTB. This was not only observed for Mn metal centers from the current research, but also for Ru metal centers from a different research done by our group. Examples of these elongated M-H bonds are shown in Figure S2.



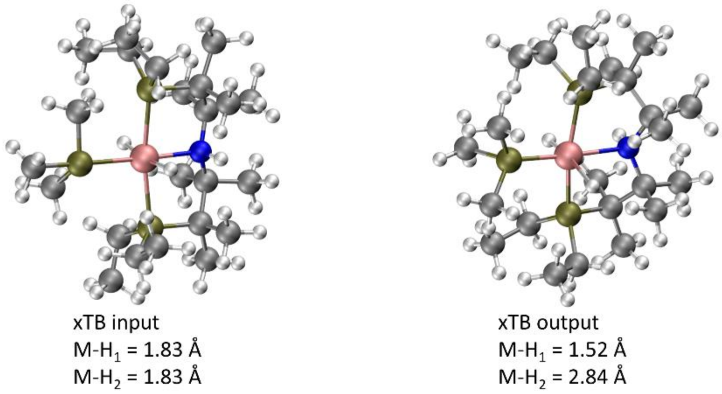


Figure S2. Structure of catalysts showing elongated M-H bonds after GFN2-xTB optimization.

S3. Computed Gibbs free energy for formation of various adducts (xTB)

The Gibbs free energy of formation of the Manganese adduct from the activated catalyst is summarized. The ligand, donor (R_1) and backbone (R_2) sites are given accompanied by the ΔG of formation in kcal mol^{-1} of the different metal adduct species. All structures were optimized using xTB.

Table S1. xTB computed Gibbs free energy of adduct formation for the Mn-PCP complexes (kcal mol^{-1})

Ligand	Donor	Backbone	Br	OH	OMe	OEt	OiPr	OtBu
PCP	CF3	CF3	16.36	20.93	14.91	15.68	15.68	21.41
PCP	CF3	cy	12.88	14.86	12.34	14.17	14.17	19.83
PCP	CF3	iPr	13.72	13.67	12.79	14.31	14.31	20.15
PCP	CF3	ph	13.51	13.98	12.99	14.66	14.66	14.94
PCP	CF3	tBu	13.47	13.84	12.92	14.53	14.53	14.78
PCP	cy	CF3	14.56	39.70	4.58	8.65	8.65	13.53
PCP	cy	cy	9.15	41.81	34.94	9.65	9.65	14.13
PCP	cy	iPr	9.85	42.29	35.39	9.67	9.67	14.28
PCP	cy	H	10.21	42.65	35.37	9.54	9.54	14.09
PCP	cy	ph	10.79	42.73	35.50	9.69	9.69	48.31
PCP	cy	tBu	10.44	43.06	35.94	10.46	10.46	14.93
PCP	H	CF3	5.00	21.77	19.39	20.68	20.68	19.06
PCP	H	cy	2.15	22.35	18.85	18.82	18.82	17.92
PCP	H	iPr	2.31	22.24	17.80	19.48	19.48	18.13
PCP	H	H	2.13	21.53	18.41	18.76	18.76	17.54
PCP	H	ph	2.63	20.75	20.63	18.98	18.98	17.87
PCP	H	tBu	2.38	21.92	19.03	18.76	18.76	17.95
PCP	iPr	CF3	13.25	37.72	4.60	6.25	6.25	9.10
PCP	iPr	cy	8.54	40.10	6.31	8.40	8.40	10.62

PCP	iPr	iPr	9.25	35.89	34.58	8.33	8.33	10.45
PCP	iPr	H	9.40	40.63	5.68	7.80	7.80	10.07
PCP	iPr	ph	9.53	36.25	5.64	7.86	7.86	9.99
PCP	iPr	tBu	8.78	35.72	6.44	8.21	8.21	10.50
PCP	ph	CF3	5.38	31.76	29.35	32.52	32.52	-2.94
PCP	ph	cy	1.13	33.67	26.98	29.41	29.41	30.58
PCP	ph	iPr	1.95	34.12	27.22	30.53	30.53	30.95
PCP	ph	H	2.07	34.61	27.97	30.55	30.55	30.88
PCP	ph	ph	1.88	34.30	28.06	30.20	30.20	-1.54
PCP	ph	tBu	1.61	33.48	27.44	30.41	30.41	30.52
PCP	tBu	CF3	28.29	10.51	9.96	13.81	13.81	12.13
PCP	tBu	cy	23.04	11.94	11.44	18.15	18.15	22.58
PCP	tBu	iPr	23.39	12.28	11.35	14.14	14.14	10.14
PCP	tBu	H	23.94	11.35	10.91	13.64	13.64	13.21
PCP	tBu	ph	24.61	12.27	11.27	14.16	14.16	14.31
PCP	tBu	tBu	23.44	12.22	11.24	14.09	14.09	23.09

Table S2. xTB computed Gibbs free energy of adduct formation for the Mn-PNP complexes

Ligand	Donor	Backbone	Br	OH	OMe	OEt	OiPr	OtBu
PNP	CF3	cy	-22.99	-14.89	-16.42	-15.82	-11.88	-30.95
PNP	CF3	iPr	-20.10	-14.12	-16.39	-14.66	-10.52	-5.40
PNP	CF3	H	-29.10	-25.02	-27.43	-26.79	-23.94	-22.13
PNP	CF3	ph	-22.84	-16.79	-18.22	-17.24	-13.47	-2.83
PNP	CF3	tBu	-23.69	-15.66	-15.97	-14.79	-10.49	-36.25
PNP	cy	CF3	-22.76	-1.00	-6.24	-4.20	-1.46	4.23
PNP	cy	cy	-29.75	-2.57	-7.54	-22.04	-2.62	-20.48
PNP	cy	iPr	-34.15	-6.49	-12.05	-10.67	-8.13	-21.25
PNP	cy	H	-31.44	1.21	-5.80	-2.87	-1.70	1.56
PNP	cy	ph	-31.53	-1.38	-7.62	-6.75	-4.41	0.81
PNP	cy	tBu	-31.56	-8.58	-10.32	-8.89	-27.92	-26.12
PNP	H	H	-34.40	-12.25	-17.08	-17.01	-15.72	-15.25
PNP	iPr	CF3	-17.41	4.88	2.63	2.25	5.86	-10.90
PNP	iPr	cy	-29.95	-0.10	-21.71	-24.02	-22.96	-22.08
PNP	iPr	iPr	-24.41	4.62	-1.39	1.36	-20.17	-19.58
PNP	iPr	H	-27.18	4.92	0.60	1.85	3.18	-4.09
PNP	iPr	ph	-25.93	4.94	-0.70	0.71	2.99	-6.68
PNP	iPr	tBu	-40.93	-14.90	-39.35	-15.12	-38.00	-38.23
PNP	ph	CF3	-12.03	7.60	3.27	3.39	6.24	-4.16
PNP	ph	cy	-26.49	-0.23	-7.69	-2.39	-0.24	-6.20
PNP	ph	iPr	-25.21	-0.72	-6.84	-5.06	-3.46	-20.48
PNP	ph	H	-27.57	-1.60	-2.68	-1.29	0.82	-8.98
PNP	ph	ph	-26.50	-0.51	-5.38	-0.74	-0.98	6.47
PNP	ph	tBu	-24.68	-4.50	-6.78	-5.92	-2.79	-15.66
PNP	tBu	H	-22.64	-0.27	-11.82	-10.60	-9.69	-8.79

Table S3. xTB computed Gibbs free energy of adduct formation for the Mn-SNS complexes

Ligand	Donor	Backbone	Br	OH	OMe	OEt	OiPr	OtBu
SNS	CF3	CF3	-2.57	11.33	-5.62	-7.27	5.57	-7.33
SNS	CF3	cy	-15.34	2.83	-2.98	-2.63	-3.15	-1.12
SNS	CF3	iPr	-15.16	3.78	-21.03	-1.87	-1.75	-19.56
SNS	CF3	H	-24.67	-6.13	-11.70	-11.80	-11.74	-10.42
SNS	CF3	ph	-19.67	-2.04	-7.11	-6.89	-6.98	-5.84

SNS	CF3	tBu	-9.79	7.87	2.29	2.46	2.62	4.20
SNS	cy	CF3	-2.95	13.03	0.89	2.31	3.77	3.29
SNS	cy	cy	-21.37	6.35	-10.19	-0.47	0.70	2.01
SNS	cy	iPr	-24.82	-0.82	-12.52	-10.95	-9.96	-10.01
SNS	cy	H	-22.25	-2.98	-14.14	-10.73	-12.85	-12.81
SNS	cy	ph	-24.27	-3.02	-10.01	-12.73	-11.25	-12.48
SNS	cy	tBu	-12.85	12.99	7.08	7.85	8.37	-15.54
SNS	H	CF3	3.42	6.28	-5.69	-5.08	-4.33	-5.08
SNS	H	cy	-18.79	-10.09	-21.85	-20.15	-22.26	-21.12
SNS	H	iPr	-13.00	-3.00	-15.51	-14.76	-13.78	-15.01
SNS	H	H	-20.52	-5.57	-28.89	-15.28	-12.73	-15.24
SNS	H	ph	-20.03	-8.79	-21.23	-29.08	-26.30	-27.29
SNS	H	tBu	-9.39	-6.76	-20.70	-19.54	-16.48	-19.32
SNS	iPr	CF3	5.69	18.12	5.66	6.63	9.34	7.18
SNS	iPr	cy	-19.94	7.37	-8.86	1.01	1.90	2.96
SNS	iPr	iPr	-17.83	1.43	-12.75	-10.85	-9.84	-9.98
SNS	iPr	H	-22.95	-1.01	-9.03	-6.70	-6.60	-11.21
SNS	iPr	ph	-17.19	4.07	-7.83	-6.09	-4.18	-5.56
SNS	iPr	tBu	-12.07	13.98	-5.92	8.07	8.93	-6.27
SNS	ph	CF3	-2.39	11.49	-0.92	-0.04	0.98	0.62
SNS	ph	cy	-20.91	6.34	-10.96	-9.74	0.98	-9.27
SNS	ph	iPr	-21.22	-1.12	-16.81	-11.33	-15.66	-15.76
SNS	ph	H	-23.58	-1.53	-12.98	-12.13	-11.93	-11.55
SNS	ph	ph	-21.08	-1.60	-13.09	-11.67	-11.23	-11.89
SNS	ph	tBu	-15.71	9.95	-11.24	-10.38	5.41	-10.34
SNS	tBu	CF3	1.31	22.89	6.29	6.88	8.59	8.79
SNS	tBu	cy	-14.01	13.16	-7.15	-6.14	-5.07	-4.16
SNS	tBu	iPr	-19.94	7.74	-7.86	2.74	3.40	3.71
SNS	tBu	H	-22.37	0.14	-11.83	-9.90	-7.98	-8.47
SNS	tBu	ph	-20.42	3.34	-9.45	-8.27	-6.27	-7.28
SNS	tBu	tBu	-9.05	18.07	11.97	12.98	13.71	14.10

Table S4. xTB computed Gibbs free energy of adduct formation for the Mn-CNC complex

Ligand	Donor	Backbone	Br	OH	OMe	OEt	OiPr	OtBu
CNC	CF3	CF3	-3.43	8.72	18.99	19.73	20.67	20.34
CNC	CF3	cy	-26.77	7.37	1.53	2.56	2.81	2.88
CNC	CF3	iPr	-28.14	4.82	-1.20	-0.63	0.94	1.21
CNC	CF3	ph	-30.05	3.54	-2.63	-2.39	-0.24	-0.30
CNC	CF3	tBu	-27.56	4.01	-1.39	-0.79	0.19	2.01
CNC	cy	CF3	-0.11	16.02	20.03	20.84	21.08	22.07
CNC	cy	cy	-23.27	12.03	3.41	4.86	6.20	6.36
CNC	cy	iPr	-24.90	7.69	0.89	2.33	3.45	4.91
CNC	cy	ph	-25.23	8.46	1.44	2.89	4.08	4.59
CNC	cy	tBu	-25.28	7.15	0.24	1.76	2.86	4.33
CNC	iPr	CF3	-9.56	7.10	10.43	10.94	12.39	12.73
CNC	iPr	cy	-31.25	1.97	-4.59	-2.88	-1.11	-0.82
CNC	iPr	iPr	-24.97	7.57	0.99	2.18	3.58	5.66
CNC	iPr	ph	-22.51	10.64	3.93	5.19	6.44	6.69
CNC	iPr	tBu	-36.15	-2.85	-9.50	-7.94	-5.77	-4.33
CNC	ph	CF3	-3.36	13.84	20.48	21.18	22.29	22.35
CNC	ph	cy	-32.52	3.97	-2.99	-2.55	-0.20	0.04
CNC	ph	iPr	-22.46	12.77	5.22	6.96	8.80	8.69

CNC	ph	H	-29.09	7.43	0.14	1.57	3.36	3.77
CNC	ph	ph	-26.67	9.01	1.95	3.89	5.11	5.36
CNC	ph	tBu	-22.17	13.00	6.34	7.56	9.70	9.88
CNC	tBu	CF3	0.36	11.86	18.99	19.62	20.77	20.92
CNC	tBu	cy	-32.15	1.53	-5.52	-4.71	-2.19	-2.45
CNC	tBu	iPr	-25.63	5.20	-1.67	-0.57	-14.05	2.41
CNC	tBu	ph	-22.58	9.16	1.88	1.55	4.53	2.99
CNC	tBu	tBu	-21.68	9.13	2.84	3.90	6.37	6.99

S4. Computed Gibbs free energy for formation of various adducts (DFT)

The Gibbs free energy of formation of the Manganese adduct from the activated catalyst is summarized. The ligand, donor (R_1) and backbone (R_2) sites are given accompanied by the ΔG of formation in kcal mol^{-1} of the different metal adduct species. All structures were optimized by DFT.

Table S5. Computed Gibbs free energy (pbe0(thf) for addition of HBr on selected Mn pincers with backbone functionalization R_2=H. All energy values in kcal mol^{-1}.

Ligand	Donor	Backbone	ΔG_{HBr}
PCP	H	H	-8.21
PCP	Ph	H	-2.70
PCP	iPr	H	0.27
PCP	Cy	H	-1.54
PCP	tBu	H	17.52
CNC	CF_3	H	-33.81
CNC	Ph	H	-36.98
CNC	iPr	H	-34.50
CNC	cy	H	-37.11
PNN	Ph	H	-36.54
PNN	iPr	H	-31.67
PNN	tBu	H	-26.74

Table S6. Computed Gibbs free energy (pbe0(thf)) for addition of HBr on selected Mn pincers with donor site functionalization R_1=^{i}Pr. All energy values in kcal mol^{-1}.

Ligand	Donor	Backbone	ΔG_{HBr}
PCP	iPr	CF3	1.29
PCP	iPr	H	0.27
PCP	iPr	Ph	0.62
PCP	iPr	iPr	0.63
PCP	iPr	cy	0.60
CNC	iPr	CF3	-28.53
CNC	iPr	H	-34.50
CNC	iPr	Ph	-32.79
CNC	iPr	iPr	-30.96
CNC	iPr	cy	-38.34
PNN	Ph	CF3	-28.98

PNN	iPr	H	-31.67
PNN	iPr	Ph	-31.51
PNN	iPr	iPr	-28.97
PNN	iPr	cy	-34.86
PNN	iPr	tBu	-23.32

Table S7. DFT computed Gibbs free energy of adduct formation for the Mn-PCP complexes (bp86(gas))

Ligand	Donor	Backbone	Br	OH	OiPr	H
PCP	CF3	cy	-4.267692109	19.10076167	-12.8670823	15.8069643
PCP	CF3	tBut	-3.886166334	18.69162548	-13.12812625	15.57290326
PCP	cy	CF3	-7.004888548	-6.173438461	-3.18774826	23.7198591
PCP	cy	cy	-5.542163904	-3.364078429	-0.797564574	25.84523378
PCP	cy	iPr	-6.782122676	-5.042666342	-2.174320417	24.30469795
PCP	cy	H	-7.028733909	-5.229036663	-2.22201114	24.1490756
PCP	cy	ph	-7.614200273	-5.875371448	-5.476275407	23.49270066
PCP	cy	tBut	-7.473010635	-5.426074646	-2.864580867	23.85979372
PCP	H	CF3	-9.596502783	-13.40423043	-14.33294449	19.10640926
PCP	H	cy	-8.953933056	-12.18184192	-12.13917128	20.6745555
PCP	H	iPr	-9.027979176	-12.14607388	-12.17305679	20.44174947
PCP	H	H	-9.474138431	-12.59160563	-13.41489809	19.97111735
PCP	H	ph	-9.171051342	-12.43723829	-12.51191192	20.08406906
PCP	H	tBut	-9.144695943	-12.20756981	-13.04090243	20.36331078
PCP	iPr	CF3	-4.787269975	-3.04969617	-2.452934635	25.12736291
PCP	iPr	cy	-4.554463951	-3.310112612	-0.672690184	26.83920882
PCP	iPr	iPr	-4.647962866	-2.520078152	-1.276354323	26.37798934
PCP	iPr	H	-5.200171226	-2.809987541	-1.335967725	25.86907914
PCP	iPr	ph	-4.714478874	-2.774219499	-1.44264434	25.88413937
PCP	iPr	tBut	-4.745226839	-2.226403706	-1.178462841	26.24872238
PCP	ph	iPr	-8.030239071	-13.78575621	-12.87712245	20.22525869
PCP	ph	H	-8.445022851	-14.53876761	-13.56048029	19.78223699
PCP	ph	ph	-8.179586332	-14.51805979	-9.394444725	20.16501778
PCP	tBut	CF3	8.800820737	6.287017681	12.28663601	31.86053984
PCP	tBut	cy	8.424315037	6.817890717	11.89569759	32.42969096
PCP	tBut	iPr	8.773837829	7.107172597	12.22827763	32.83568961
PCP	tBut	H	8.372859259	6.347886102	12.30357877	32.52820995
PCP	tBut	ph	8.832823722	7.398337005	43.07036955	32.22261282
PCP	tBut	tBut	8.253004944	6.845501135	12.12285603	32.32426936

Table S8. DFT computed Gibbs free energy of adduct formation for the Mn-PCP complexes (bp86(thf))

Ligand	Donor	Backbone	Br	OH	OiPr	H
PCP	CF3	cy	0.643761997	26.68688253	-0.24737805	13.38907847
PCP	CF3	tBut	1.045248849	26.32569433	-0.469008131	13.26648193
PCP	cy	CF3	-1.804842823	7.717891575	9.774086717	20.18115972
PCP	cy	cy	-1.805991167	10.29342902	11.28328472	21.07851085
PCP	cy	iPr	-2.121164087	9.405408947	10.72264263	20.5565861
PCP	cy	H	-2.301642095	9.124378817	10.51519426	20.29803964
PCP	cy	ph	-2.868465151	8.578451828	6.7312053	19.72084385
PCP	cy	tBut	-2.688219322	9.110284954	10.20399975	20.15947299
PCP	H	CF3	-8.62195543	-2.729583117	-4.928511069	13.68727098
PCP	H	cy	-8.589651241	-1.48131612	-2.496980782	14.74630621
PCP	H	iPr	-8.544652534	-1.37699894	-2.44665452	14.65578797

PCP	H	H	-8.936632619	-1.919531104	-4.121935458	14.1574513
PCP	H	ph	-8.502647048	-1.649338063	-2.778042287	14.43228163
PCP	H	tBut	-8.750218372	-1.47567481	-3.67226215	14.49329438
PCP	iPr	CF3	-0.579850153	11.43438561	7.688383192	21.6004356
PCP	iPr	cy	-0.788258609	10.86945136	9.563099198	22.91445305
PCP	iPr	iPr	-0.822991259	12.52213568	8.967580132	22.51081384
PCP	iPr	H	-1.29007168	12.2542958	8.86205186	22.05125725
PCP	iPr	ph	-0.954548626	12.20366834	8.68106557	21.98171038
PCP	iPr	tBut	-0.944665351	12.91474328	9.184503891	22.45167107
PCP	ph	iPr	-4.171425877	-0.792706019	-1.497784855	17.06401883
PCP	ph	H	-4.688010523	-1.814040482	-2.462768964	16.41033591
PCP	ph	ph	-4.433266766	-1.815684557	1.355262387	16.79073216
PCP	tBut	CF3	13.32389039	20.46052619	24.74971389	28.17562447
PCP	tBut	cy	12.56861995	21.13737048	24.38433393	28.63872648
PCP	tBut	iPr	12.95733071	21.36606005	24.64682743	29.20279477
PCP	tBut	H	12.66134703	20.54137451	24.77438128	28.91929226
PCP	tBut	ph	12.86880167	21.96306631	52.84775069	28.2982712
PCP	tBut	tBut	12.24545884	20.99089722	24.49261697	28.47406172

Table S9. DFT computed Gibbs free energy of adduct formation for the Mn-PCP complexes (pbe0(thf))

Ligand	Donor	Backbone	Br	OH	OiPr	H
PCP	CF3	cy	2.212836951	31.65212574	1.358788363	18.63228925
PCP	CF3	tBut	2.721640478	31.42314752	1.230010864	18.69006405
PCP	cy	CF3	-0.786802787	8.715636951	13.18173878	24.65709581
PCP	cy	cy	-2.154591519	10.474088	13.43200839	24.13035179
PCP	cy	iPr	-1.483112428	10.50041202	13.85572791	24.63252254
PCP	cy	H	-1.541495912	10.33139861	13.75078949	24.49207336
PCP	cy	ph	-2.119237633	9.800368697	9.34342632	23.91057286
PCP	cy	tBut	-1.9750234	10.25686931	13.36952727	24.31194676
PCP	H	CF3	-7.498004339	-2.071314105	-2.542645276	19.51828874
PCP	H	cy	-8.09716296	-1.326705058	-0.644617291	19.84780025
PCP	H	iPr	-8.006500387	-1.17393787	-0.549662554	19.82607587
PCP	H	H	-8.212593333	-1.533212159	-2.065593729	19.52077995
PCP	H	ph	-7.848079339	-1.37110763	-0.769472856	19.74909301
PCP	H	tBut	-8.162511799	-1.272902394	-1.813912219	19.72777651
PCP	iPr	CF3	1.294508173	13.48133942	11.46684944	26.94276152
PCP	iPr	cy	0.597451794	12.06737227	12.90860899	27.71161753
PCP	iPr	iPr	0.62813701	13.74324306	12.35800706	27.36566527
PCP	iPr	H	0.270933502	13.60219775	12.38846009	27.03626671
PCP	iPr	ph	0.622552175	13.54145483	12.22767333	26.99839023
PCP	iPr	tBut	0.492149426	14.16644175	12.60451161	27.29308752
PCP	ph	iPr	-2.249885112	-0.071755084	2.025987839	22.40813941
PCP	ph	H	-2.697675891	-1.020844377	1.160834217	21.8321861
PCP	ph	ph	-2.459881166	-1.092399285	4.362004905	22.20155701
PCP	tBut	CF3	18.45073713	23.20331273	28.86796398	34.18161071
PCP	tBut	cy	17.2662502	23.58145623	28.42006025	34.14178895
PCP	tBut	iPr	17.71803821	23.82030517	28.76797034	34.79262926
PCP	tBut	H	17.51746735	23.09261378	29.03614908	34.60827324
PCP	tBut	ph	17.7319752	24.76512113	63.69604645	34.02534829
PCP	tBut	tBut	17.02416331	23.45076483	28.64203546	34.09260476

Table S10. DFT computed Gibbs free energy of adduct formation for the Mn-PNN complexes (bp86(gas))

Ligand	Donor	Backbone	Br	OH	OiPr	H

PNN	CF3	CF3	-29.75336294	-23.6734234	-29.62221346	-4.80672277
PNN	CF3	iPr	-37.41086137	-25.463708	-29.29779105	-2.720881192
PNN	CF3	ph	-26.75449504	-20.17003786	-19.79102212	11.3347041
PNN	CF3	tBut	-28.69224438	-24.96546546	-28.32954389	-1.236821225
PNN	cy	CF3	-32.65496687	-23.6207126	-26.65974111	4.450297374
PNN	cy	cy	-38.3966788	-18.38540084	-22.12284742	0.690887959
PNN	cy	iPr	-33.29314403	-18.14945727	-17.58783627	7.458577917
PNN	cy	ph	-28.07665756	-12.93297079	-12.37950742	12.81248897
PNN	H	CF3	-26.11820041	-9.162893719	-11.73066259	12.05822255
PNN	H	iPr	-31.95843133	-12.84574697	-14.47099658	8.076674774
PNN	H	ph	-37.62609713	-17.4115061	-19.79729722	4.588976974
PNN	iPr	CF3	-29.37246468	-14.82804949	-18.60251913	8.312618346
PNN	iPr	cy	-30.92178563	-16.78525162	-18.00262005	10.96070844
PNN	iPr	iPr	-26.29327556	-12.83507931	-15.97639187	12.10340324
PNN	iPr	H	-31.97035401	-23.21910652	-23.17329833	1.140812271
PNN	iPr	ph	-30.70905991	-15.6613821	-16.18221499	5.875371449
PNN	iPr	tBut	-19.74584144	-2.577181516	-8.655238533	18.05846839
PNN	ph	CF3	-29.8958076	-20.01504301	-22.78675247	-5.77245989
PNN	ph	cy	-31.98353171	-22.68760597	-18.6866054	-7.884029358
PNN	ph	iPr	-20.17756797	-10.78061321	-8.743089864	7.65498839
PNN	ph	H	-36.56874362	-29.75650049	-24.28273512	-5.148087938
PNN	ph	ph	-29.228765	-20.57917405	-17.28349416	8.481418402
PNN	ph	tBut	-31.69864239	-20.75738675	-24.03926144	-4.359308497
PNN	tBut	CF3	-19.46408967	-10.07905759	-6.682976175	16.65347462
PNN	tBut	cy	-20.5490536	-8.869846783	-14.31035415	14.12272881
PNN	tBut	H	-21.88502132	-15.1330191	-15.83896729	14.90272312
PNN	tBut	ph	-21.67794319	-10.87348462	-15.12548899	12.1404263

Table S11. DFT computed Gibbs free energy of adduct formation for the Mn-PNN complexes (bp86(thf))

Ligand	Donor	Backbone	Br	OH	OiPr	H
PNN	CF3	CF3	-29.2155622	-17.15602025	-21.70076069	-8.771250099
PNN	CF3	iPr	-33.27884937	-16.14997192	-18.92755775	-6.584097112
PNN	CF3	ph	-23.8285563	-10.21447251	-9.361250727	9.61173483
PNN	CF3	tBut	-27.43972287	-13.78905527	-17.18478189	-4.971868329
PNN	cy	CF3	-29.93284948	-16.35086281	-17.726925	0.488123206
PNN	cy	cy	-36.2536209	-8.367530071	-10.92293989	-3.705246679
PNN	cy	iPr	-31.62827975	-7.992285665	-7.718129651	1.295382802
PNN	cy	ph	-25.67151405	-2.119279928	-2.247093949	8.575603694
PNN	H	CF3	-28.09116558	-1.130469283	-3.31590578	3.016547234
PNN	H	iPr	-31.90166054	-3.325880295	-5.329464539	0.519561431
PNN	H	ph	-35.69495547	-6.95881517	-9.211301057	-1.092460273
PNN	iPr	CF3	-26.51314234	-4.358064396	-7.524360993	3.161790583
PNN	iPr	cy	-28.88359085	-6.190812568	-8.72830072	5.552068395
PNN	iPr	iPr	-24.08604227	-1.840326855	-6.732644807	8.399932409
PNN	iPr	H	-30.48144966	-13.88021358	-14.3957971	-3.310951084
PNN	iPr	ph	-28.21728871	-5.101794931	-6.750848858	1.739308123
PNN	iPr	tBut	-18.12977857	2.397922509	-0.380165336	12.63087129
PNN	ph	CF3	-29.98589286	-11.69580908	-14.89195631	-10.9178095
PNN	ph	cy	-31.28877829	-11.53948392	-10.03009428	-10.99474216
PNN	ph	iPr	-18.09889255	0.812187177	1.312723519	5.005576646
PNN	ph	H	-35.07973888	-20.54450252	-15.91133922	-10.14545198
PNN	ph	ph	-26.7650874	-9.560086776	-7.090256196	4.438690839
PNN	ph	tBut	-29.36361679	-9.801050424	-14.21357461	-9.118783685
PNN	tBut	CF3	-18.7907406	-3.032343901	0.333614171	13.00706951
PNN	tBut	cy	-19.86640484	-0.189154832	-4.062322055	8.660110398
PNN	tBut	H	-22.68340166	-7.421578326	-8.270695692	9.06315975
PNN	tBut	ph	-22.42290364	-3.225911756	-6.742139025	7.949844945

Table S12. DFT computed Gibbs free energy of adduct formation for the Mn-PNN complexes (pbe0(thf))

Ligand	Donor	Backbone	Br	OH	OiPr	H
PNN	CF3	CF3	-32.64837175	-15.93151176	-18.2864633	-8.393528348
PNN	CF3	iPr	-39.12931496	-18.67391653	-18.82764632	-10.06675105
PNN	CF3	ph	-26.55222363	-9.16507055	-5.612497052	11.39420562
PNN	CF3	tBut	-32.60191722	-15.31063504	-15.51168532	-5.88356565
PNN	cy	CF3	-28.42014371	-11.81199975	-13.57564281	3.96761814
PNN	cy	cy	-39.69377858	-11.46100231	-10.59861851	-2.831064512
PNN	cy	iPr	-31.74096163	-5.767978845	-3.594075094	3.031173163
PNN	cy	ph	-25.44031445	0.966766864	2.579413568	11.67816622
PNN	H	CF3	-29.87985003	0.697226433	0.463206805	4.305651232
PNN	H	iPr	-37.28023899	-4.827800186	-4.977074657	-1.901095433
PNN	H	ph	-40.40039821	-7.734945024	-8.376287343	-0.700594458
PNN	iPr	CF3	-28.97966255	-2.997486752	-4.913953476	4.356837182
PNN	iPr	cy	-34.85529756	-8.731586987	-8.590776365	2.093015085
PNN	iPr	iPr	-28.97165553	-2.88144142	-4.166903416	6.694008865
PNN	iPr	H	-31.66901767	-14.15188227	-12.02425734	-0.022470042
PNN	iPr	ph	-31.51031424	-3.599029909	-4.67251292	2.768711039
PNN	iPr	tBut	-23.3199222	-3.765351302	-0.917326645	10.19014039
PNN	ph	CF3	-33.05866256	-11.67694088	-13.29506448	-12.93164568
PNN	ph	cy	-35.22811978	-11.03019194	-7.946286458	-11.25315189
PNN	ph	iPr	-18.15218066	4.723327385	6.944150022	8.38461986
PNN	ph	H	-36.53505264	-20.13593836	-11.19345359	-8.767931892
PNN	ph	ph	-30.54648475	-8.802175531	-5.425687473	5.15925868
PNN	ph	tBut	-34.01862032	-10.36239633	-12.57029101	-11.84792422
PNN	tBut	CF3	-21.34304764	-4.381709958	-0.191342079	14.56982412
PNN	tBut	cy	-24.51399747	-1.639531094	-1.385906814	6.585261469
PNN	tBut	H	-26.73780956	-8.910697024	-6.617353114	12.15051145
PNN	tBut	ph	-26.16852667	-3.389504487	-4.112115562	9.071868264

Table S13. DFT computed Gibbs free energy of adduct formation for the Mn-CNC complexes (bp86(gas))

Ligand	Donor	Backbone	Br	OH	OiPr	H
CNC	CF3	cy	-31.99545439	-12.18121441	-16.76328878	3.807727646
CNC	CF3	iPr	-19.28650448	0.222138363	0.031375475	14.63665909
CNC	CF3	H	-29.91839794	-7.312995713	-10.67519161	1.163402613
CNC	CF3	ph	-31.30393892	-9.12336062	-13.77257851	-4.516813381
CNC	CF3	tBut	-15.72476056	4.80672277	0.994602558	16.94526654
CNC	cy	CF3	-21.83231052	-2.041288404	-5.010663357	16.08244098
CNC	cy	iPr	-28.38539223	-5.608679911	-9.144695944	8.76003262
CNC	cy	H	-32.1617444	-8.468240702	-12.48932158	6.679211118
CNC	cy	ph	-31.63589144	-7.727151983	-13.04278496	6.705566517
CNC	cy	tBut	-28.65710385	-6.468367926	-9.860684283	7.0149287
CNC	H	CF3	-26.53298419	-4.500498134	-8.618842983	13.00199684
CNC	H	cy	-23.00763582	0.881023338	-3.332075445	15.42920359
CNC	H	iPr	-32.93671864	-8.426825075	-13.3954453	5.991460706
CNC	H	tBut	-24.16037077	0.28865437	-5.542791414	12.27471333
CNC	iPr	CF3	-28.90308757	-7.810610747	-10.22212976	8.641433324
CNC	iPr	cy	-31.50599698	-8.340228764	-10.73229498	6.163398309
CNC	iPr	iPr	-27.69324925	-4.83182315	-7.272207596	7.473010635
CNC	iPr	H	-30.13677125	-5.415406985	-9.552577119	8.545424371
CNC	iPr	ph	-28.79892099	-6.012796029	-9.367461816	9.956065727
CNC	ph	CF3	-26.80093075	-5.630015234	-8.971503322	12.44288588
CNC	ph	cy	-34.21119043	-14.56575051	-17.87021554	4.297812565
CNC	ph	iPr	-30.88288004	-8.648963439	-11.46334355	6.278860057
CNC	ph	ph	-33.22913806	-7.890304453	-13.44250851	6.457700265
CNC	tBut	CF3	-17.23768597	1.228036091	-1.040410751	17.45731429

CNC	tBut	cy	-19.68309049	-5.023213547	-5.071531779	16.84298249
CNC	tBut	iPr	-15.90485579	3.232928944	4.802957713	15.93121119
CNC	tBut	ph	-27.99508132	-5.407249361	-8.887417048	7.845751279
CNC	tBut	tBut	-13.8252893	5.006898301	5.411641928	14.88389783

Table S14. DFT computed Gibbs free energy of adduct formation for the Mn-CNC complexes (bp86(thf))

Ligand	Donor	Backbone	Br	OH	OiPr	H
CNC	CF3	cy	-31.87666684	-3.165488867	-4.455161828	-4.037757688
CNC	CF3	iPr	-22.21543017	6.510858226	6.654297736	4.726121565
CNC	CF3	H	-29.42890916	2.131306273	0.439700927	-4.11755807
CNC	CF3	ph	-30.32641717	0.607725757	-2.344765803	-8.804062571
CNC	CF3	tBut	-18.86299205	10.13564803	9.109761235	7.078133848
CNC	cy	CF3	-21.39369393	6.653240132	4.842539757	9.637054838
CNC	cy	iPr	-26.95921374	4.082452937	3.004614632	3.293881332
CNC	cy	H	-32.1847489	1.018160896	-1.435341385	-0.109642344
CNC	cy	ph	-31.03526445	1.181495344	-2.435309149	0.266901006
CNC	cy	tBut	-27.32640093	3.212599268	2.140013218	1.853577602
CNC	H	CF3	-24.49182129	5.210005932	1.986191814	6.892033356
CNC	H	cy	-23.64241815	9.0290915	6.093869002	8.867489737
CNC	H	iPr	-34.70229192	-2.18088881	-5.731804805	-4.222176455
CNC	H	tBut	-26.61138523	5.887916995	2.003931508	3.051461863
CNC	iPr	CF3	-25.49975215	4.15338661	2.627362196	5.099012811
CNC	iPr	cy	-31.84363474	0.358617036	-1.146636814	-1.041832807
CNC	iPr	iPr	-26.93798509	4.414719217	4.506696673	1.320251003
CNC	iPr	H	-30.65516939	4.05458524	0.69485884	0.854551103
CNC	iPr	ph	-28.23645913	3.992022543	1.011807613	2.197477909
CNC	ph	CF3	-23.85225105	6.193394894	3.808705306	8.497039504
CNC	ph	cy	-35.33888775	-5.901367535	-8.008779502	-4.153922246
CNC	ph	iPr	-30.26680377	1.325766052	-0.284426211	0.031572394
CNC	ph	ph	-32.09163277	2.472822043	-1.308935871	0.232463285
CNC	tBut	CF3	-14.89848743	12.21678956	11.16395123	12.78424088
CNC	tBut	cy	-19.81076984	4.300123432	4.169761722	11.12395115
CNC	tBut	iPr	-15.97900231	12.74049644	13.808521	9.990060307
CNC	tBut	ph	-27.12814557	4.885439193	2.285746025	2.480057987
CNC	tBut	tBut	-13.87020643	14.45873668	15.59244893	8.733635686

Table S15. DFT computed Gibbs free energy of adduct formation for the Mn-CNC complexes (pbe0(thf))

Ligand	Donor	Backbone	Br	OH	OiPr	H
CNC	CF3	cy	-36.23821554	-3.825278459	-3.114312706	-4.044605134
CNC	CF3	iPr	-26.67109903	6.124098017	8.562829602	4.02407517
CNC	CF3	H	-33.80514158	2.305909787	1.882601919	-2.968482817
CNC	CF3	ph	-33.38076946	1.202170778	0.306122352	-7.368654732
CNC	CF3	tBut	-22.70483738	10.26256709	11.439421	6.907261693
CNC	cy	CF3	-21.32940559	10.49328351	10.69289177	13.07319514
CNC	cy	iPr	-30.06991011	4.84621257	5.852722748	3.305407364
CNC	cy	H	-37.11435058	0.466622967	-0.28715023	-0.941489081
CNC	cy	ph	-33.56826302	2.368961942	0.664957385	1.541352658
CNC	cy	tBut	-30.50218885	3.800969996	4.91094394	1.725846726
CNC	H	CF3	-25.64892998	8.014966118	6.130006646	9.113949051
CNC	H	cy	-26.31247735	10.41092917	8.954044125	10.45856386
CNC	H	iPr	-39.33953064	-2.620284517	-5.059429003	-5.051400201
CNC	H	tBut	-29.95821342	6.419102782	4.009733622	3.254284166
CNC	iPr	CF3	-28.52769884	4.899852082	4.873412597	5.522749833

CNC	iPr	cy	-38.33692107	-2.175731686	-2.054505636	-4.20553622
CNC	iPr	iPr	-30.96188349	3.98818118	6.078030034	0.651581837
CNC	iPr	H	-34.49563421	4.316538077	2.578064423	1.230929983
CNC	iPr	ph	-32.78621048	3.478706217	2.01283524	1.122941874
CNC	ph	CF3	-25.43570225	8.817193088	7.936782199	10.52206782
CNC	ph	cy	-39.49787012	-6.17653773	-6.499080123	-4.49183743
CNC	ph	iPr	-33.8048843	1.933495449	2.15352287	0.291290983
CNC	ph	ph	-34.05373575	4.103542527	2.9167564	2.994526608
CNC	tBut	CF3	-17.00801121	13.8614784	14.83945572	14.94422767
CNC	tBut	cy	-24.42082486	3.015930511	4.649912539	10.55862025
CNC	tBut	iPr	-20.84223858	11.77352839	14.79944572	9.720430708
CNC	tBut	ph	-31.00921025	5.161084288	3.893976945	3.243453351
CNC	tBut	tBut	-18.21308673	13.99587839	17.04070885	8.584111405

S5. Plots of Relative Gibbs Free Energy

In the following section figures displaying thermodynamic trends are displayed.

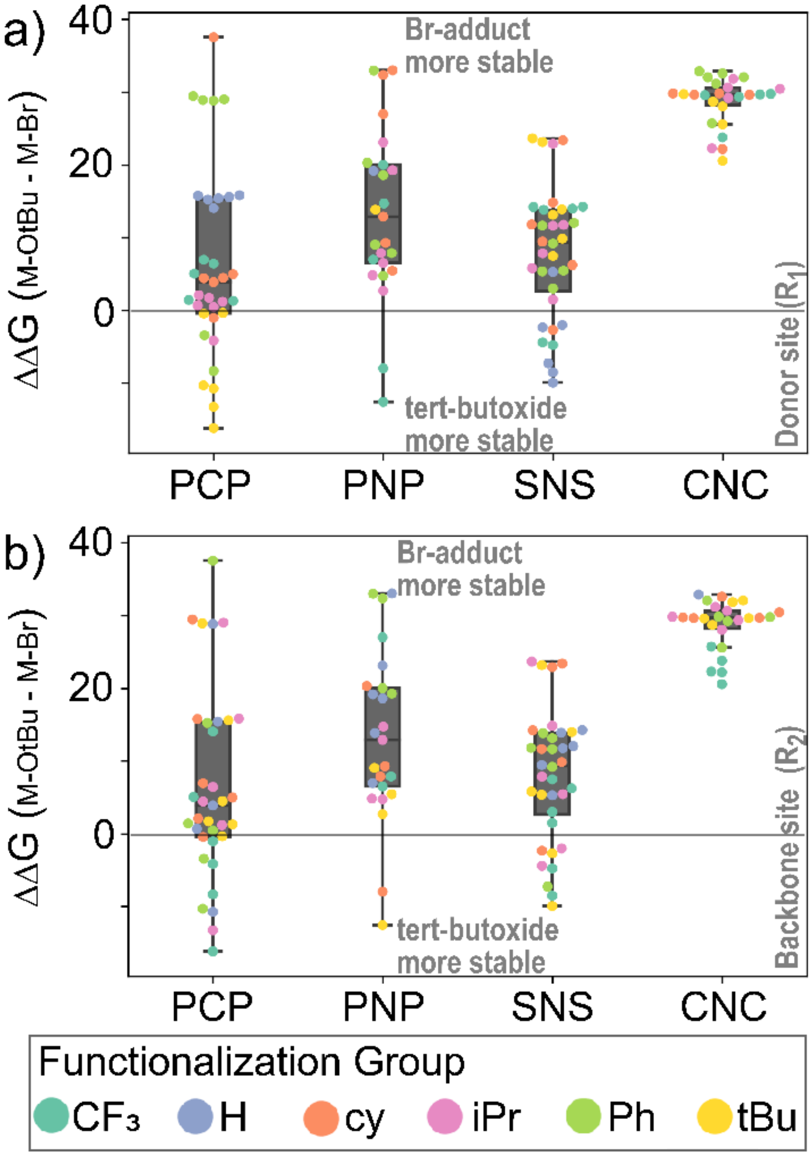


Figure S3. Boxplot and swarmplot of the change in xTB computed Gibbs free energy for the addition of ^{t}BuOH and HBr on various Mn-pincers. Individual data points are sorted by ligand backbones, and color coded based on functionalization of the donor and backbones sites. All energy values in kcal mol^{-1}.

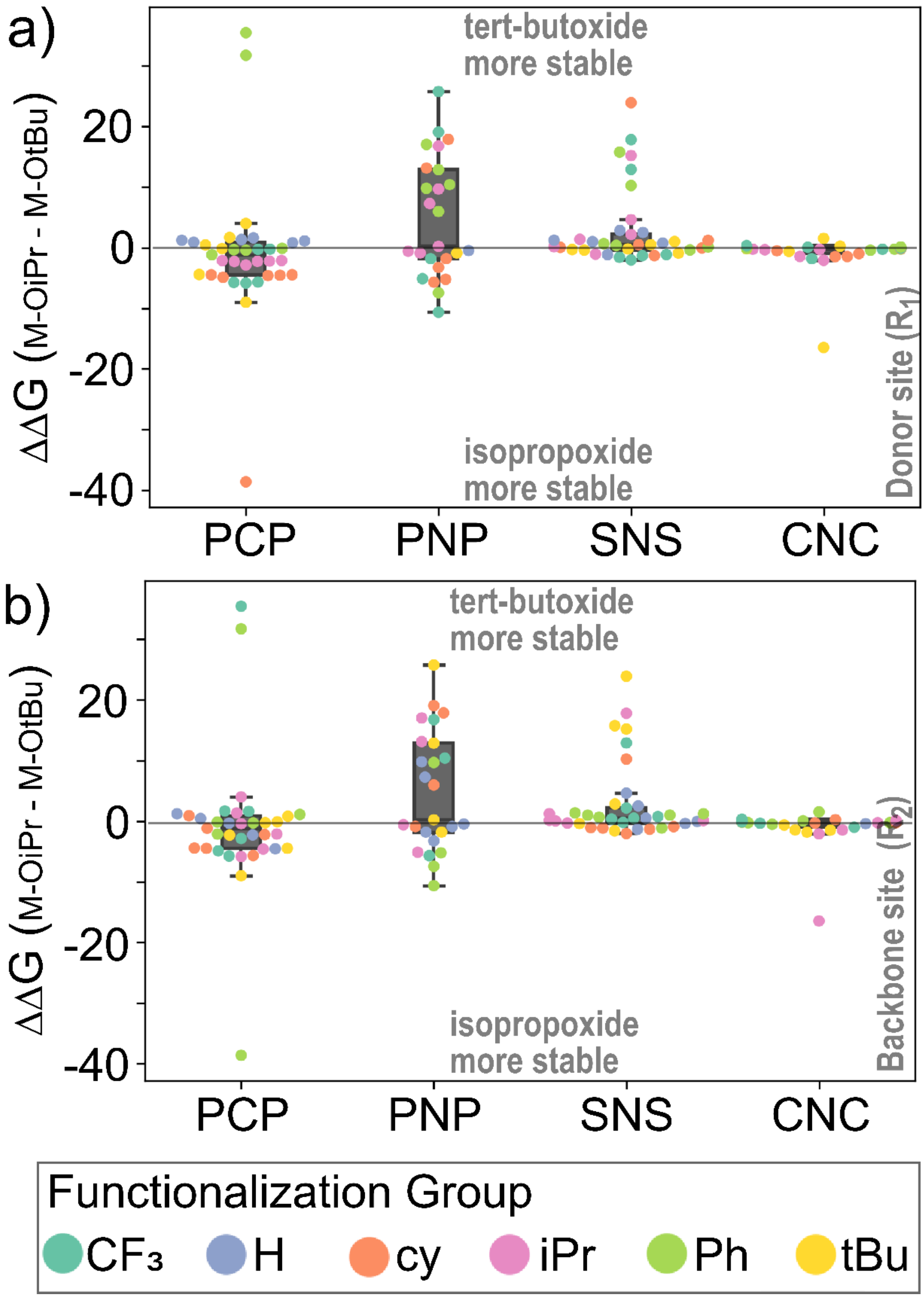


Figure S4. Boxplot and swarmplot of the change in xTB computed Gibbs free energy for the addition of ^{t}BuOH and ^{i}PrOH on various Mn-pincers. Individual data pints are sorted by ligand backbones, and color coded based on functionalization of the donor and backbones sites. All energy values in kcal mol^{-1}.

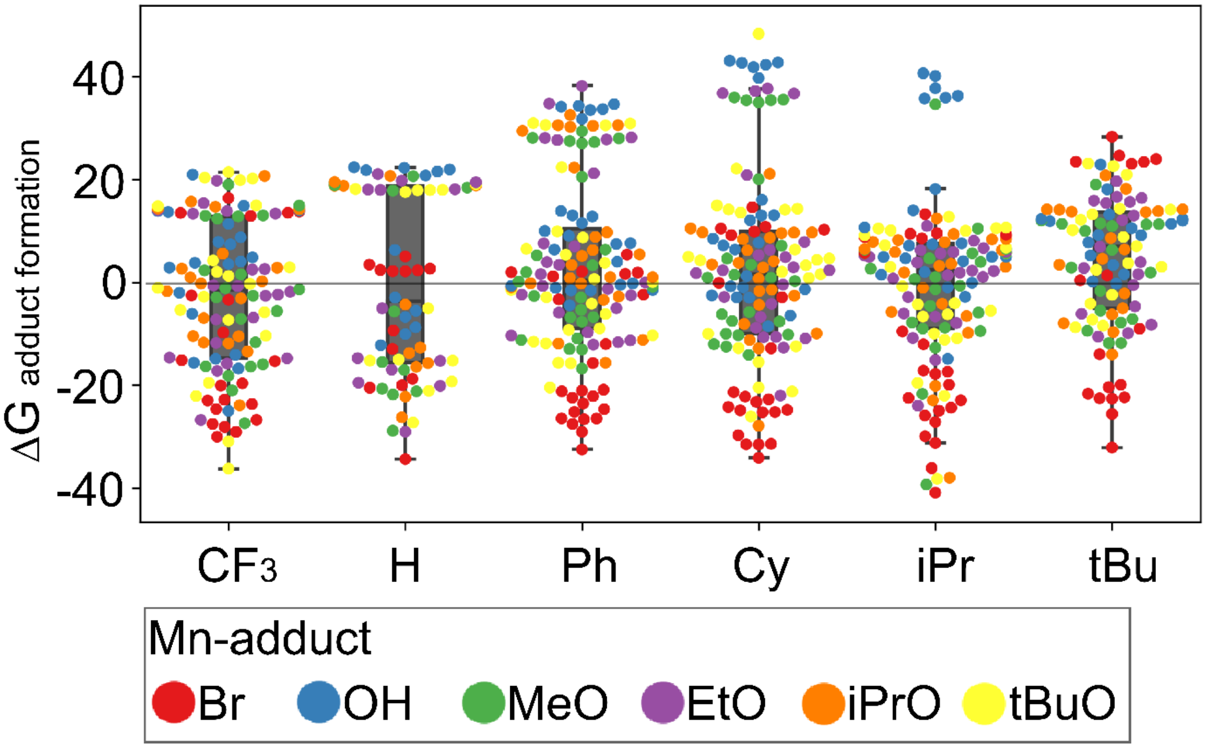


Figure S5. Boxplot and swarmplot of xTB computed Gibbs free energy of addition of various substrates (HX = HBr, H_2O, MeOH, EtOH, ^{i}PrOH and ^{t}BuOH) on Mn-pincers. Data points are sorted by donor site functionalization groups and color coded by the substrate. All energy values in kcal mol^{-1}.

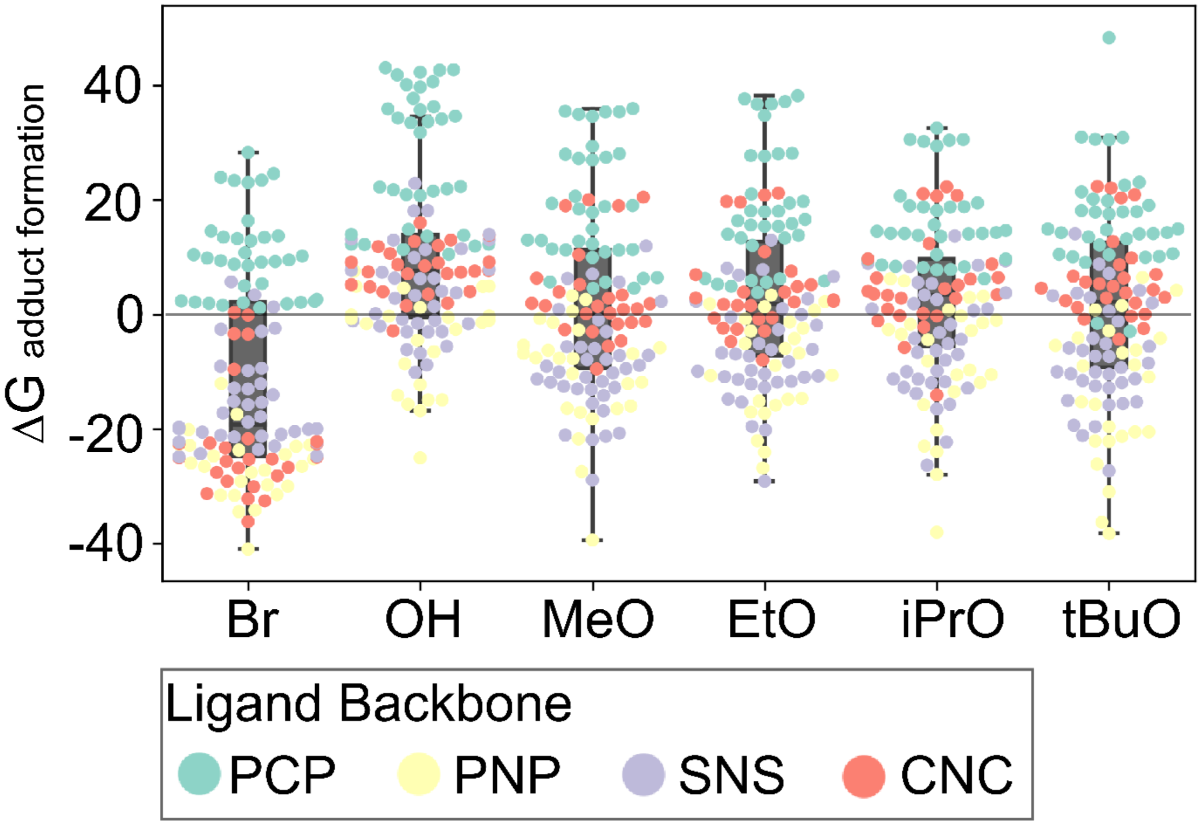


Figure S6. Boxplot and swarmplot of xTB computed Gibbs free energy of addition of various substrates (HX = HBr, H_2O, MeOH, EtOH, ^{i}PrOH and ^{t}BuOH) on Mn-pincers. Data points are sorted by the substrate and color coded by the ligand backbone. All energy values in kcal mol^{-1}.

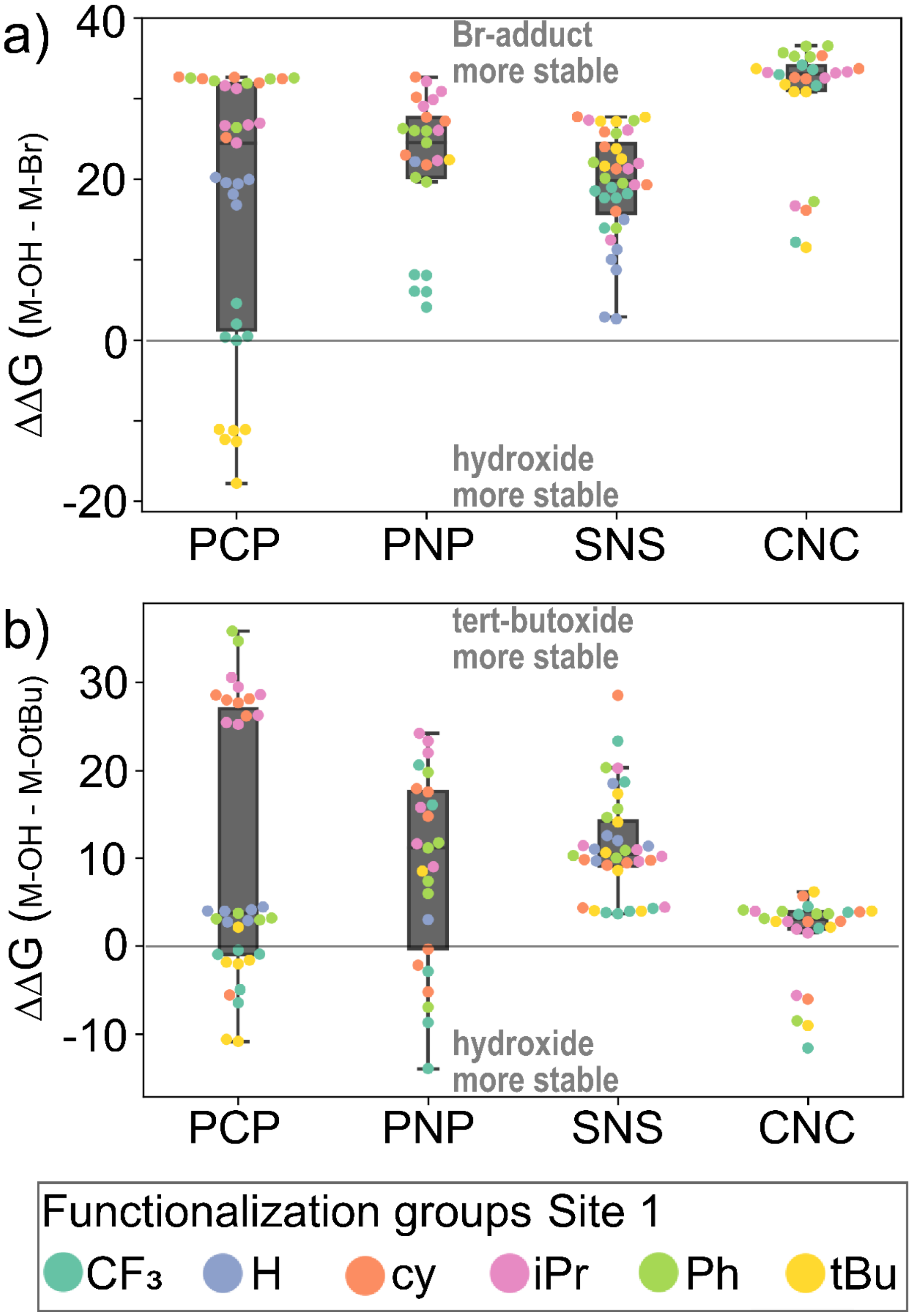


Figure S7. Boxplot and swarmplot of the change in xTB computed Gibbs free energy for the addition of a) H_2O and HBr and b) H_2O and ^tBuOH on various Mn-pincers. Individual data points are sorted by ligand backbones, and color coded based on functionalization of the donor site (R_1). All energy values in kcal mol^{-1}.

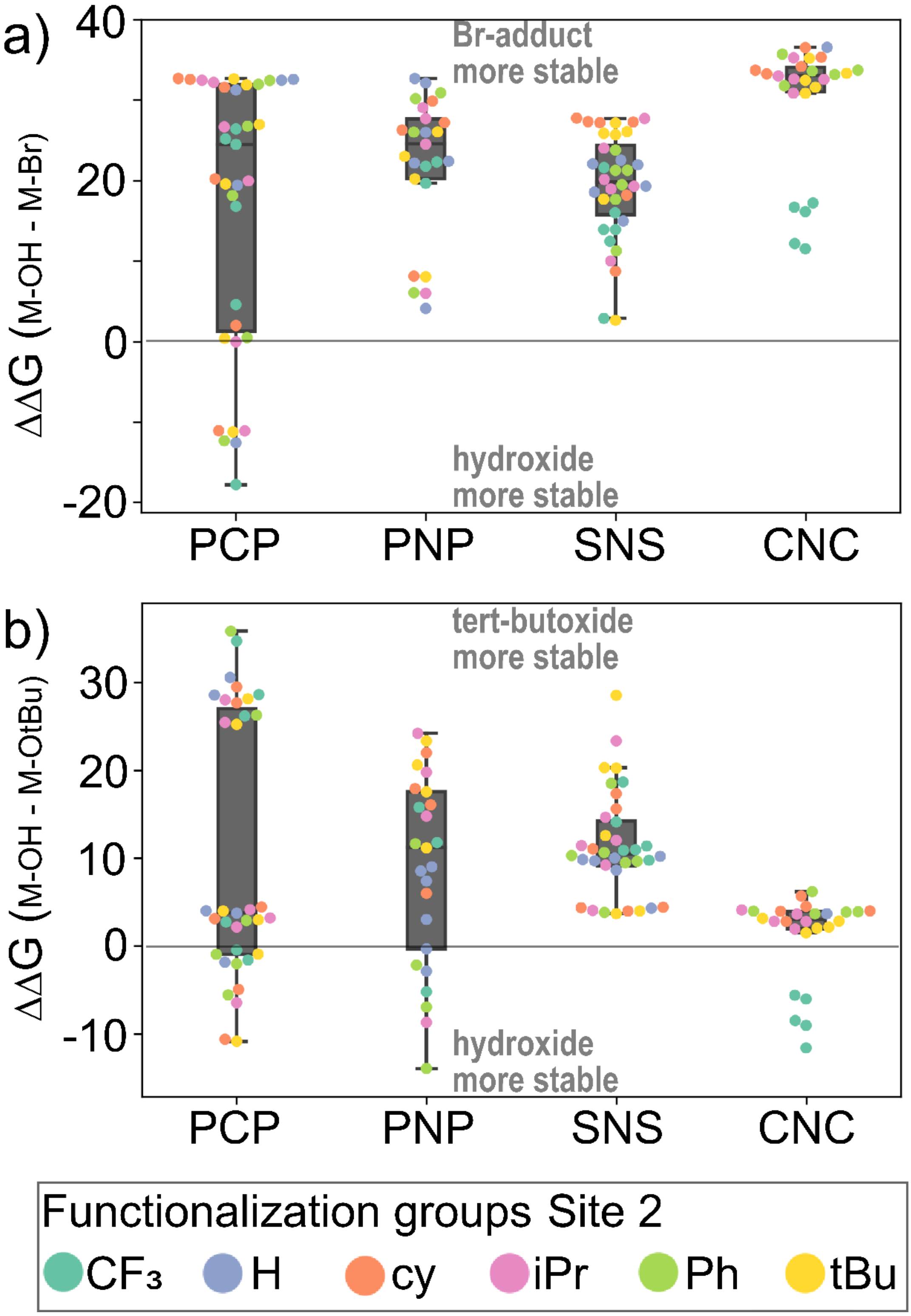


Figure S8. Boxplot and swarmplot of the change in xTB computed Gibbs free energy for the addition of a) H_2O and HBr and b) H_2O and ^{t}BuOH on various Mn-pincers. Individual data points are sorted by ligand backbones, and color coded based on functionalization of the ligand backbone site (R_2). All energy values in kcal mol^{-1}.

S6. Heatmaps of computed Gibbs free energies

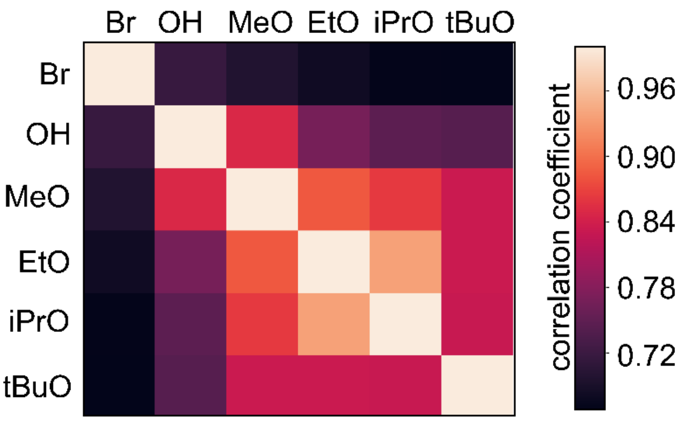


Figure S9. Correlation matrix of Gibbs free energy of formation in all ligand backbones investigated by xTB of the different metal adducts

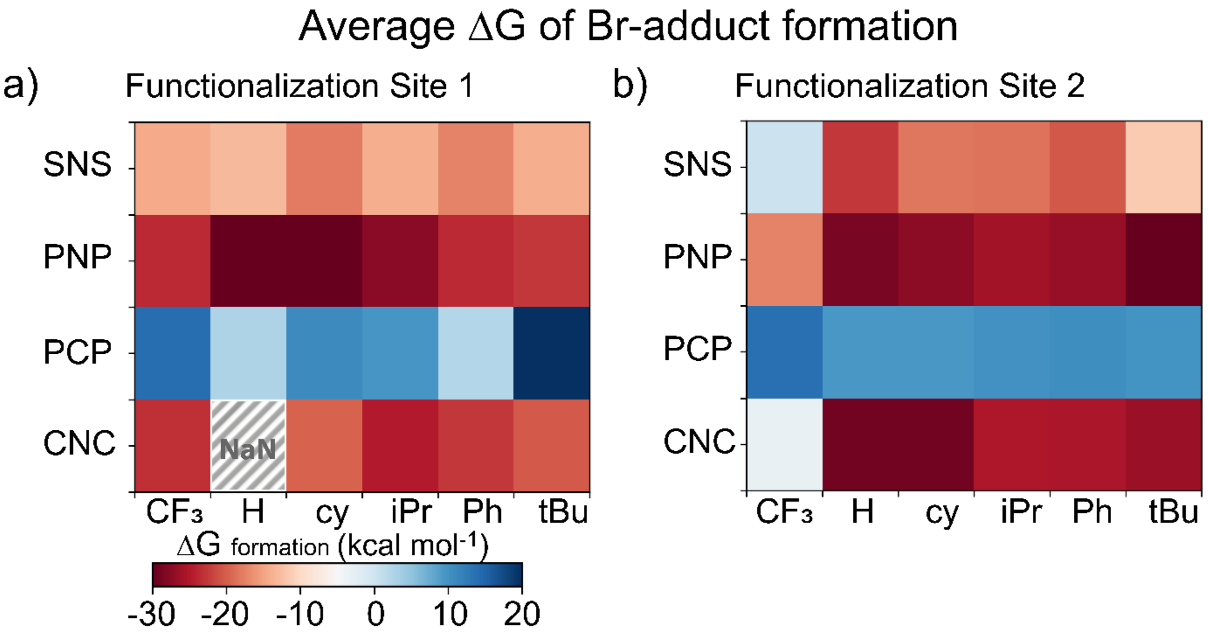


Figure S10. Heat maps of DFT (pbe0(thf)) computed average Gibbs free energies of formation of the Bromide adduct species in different ligand backbones a) as functionalization of donor site (R_1) and b) functionalization on the backbone site (R_2).

S7. Linear scaling plots

In the following section figures displaying scaling relationships between computed ΔG_{HX} for addition of various substrates to Mn-pincers using DFT, xTB, and between both methods are illustrated.

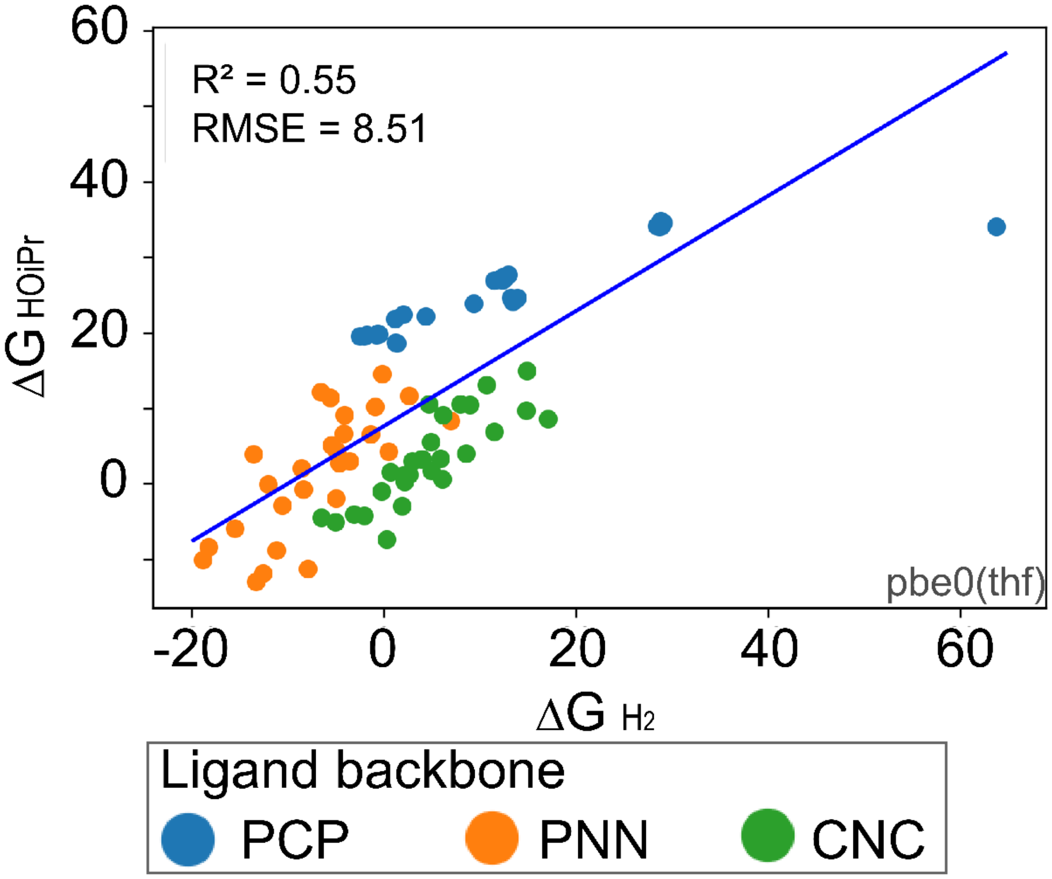


Figure S11. Comparison of DFT (pbe0(thf)) computed Gibbs free energies for the formation of ^{i}PrO and hydride complexes. The data for catalysts with different backbones are differentiated by colors. All energy values in kcal mol^{-1}.All energy values in kcal mol^{-1}.

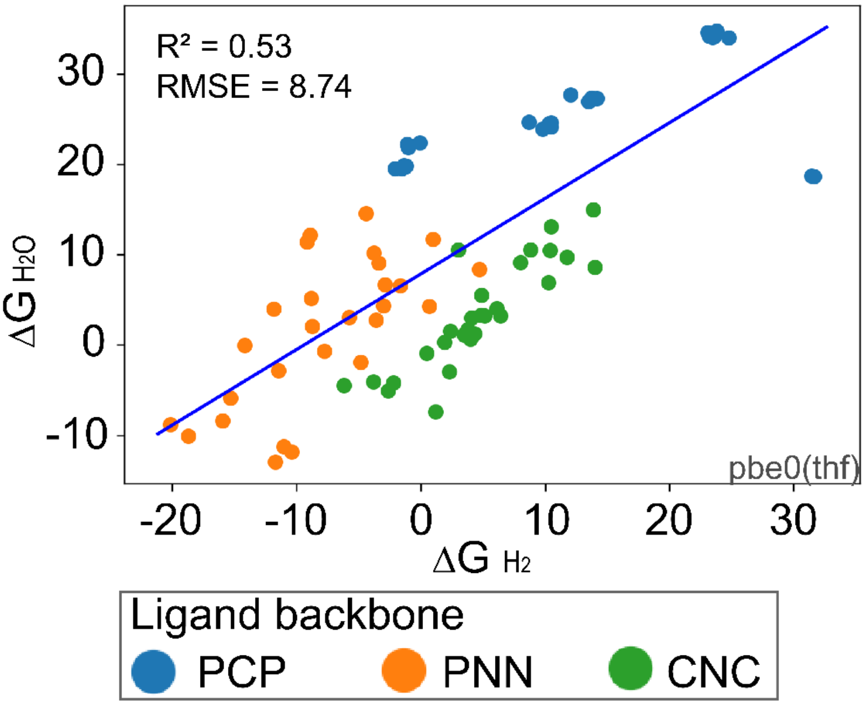


Figure S12. Comparison of DFT (pbe0(thf)) computed Gibbs free energies for the formation of hydroxide and hydride complexes. The data for catalysts with different backbones are differentiated by colors. All energy values in kcal mol^{-1}.

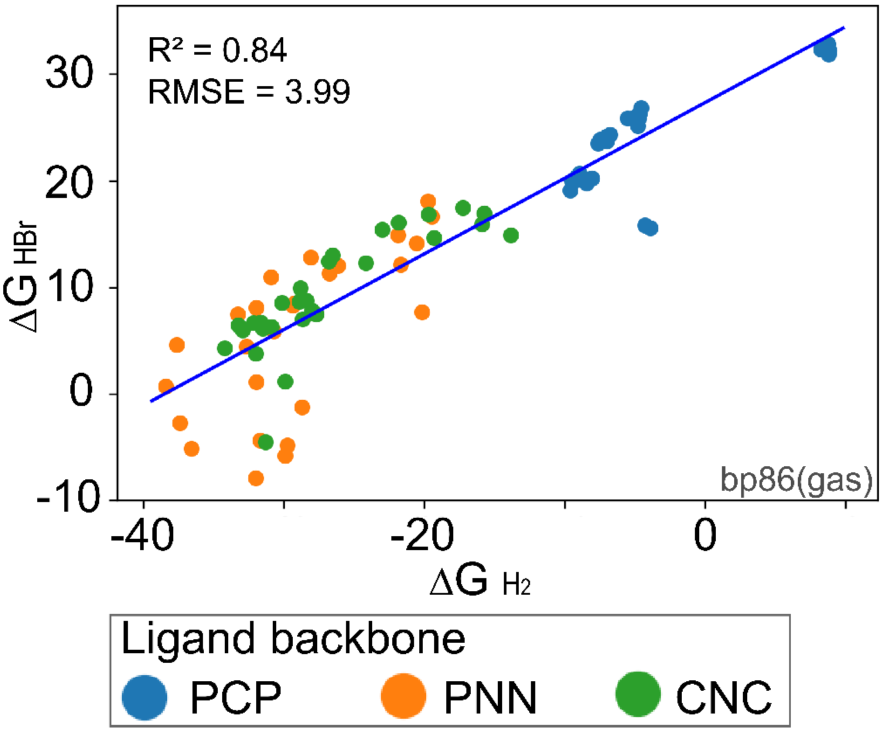


Figure S13. Comparison of DFT (bp86(thf)) computed Gibbs free energies for the formation of bromide and hydride complexes. The data for catalysts with different backbones are differentiated by colors. All energy values in kcal mol^{-1}.All energy values in kcal mol^{-1}.

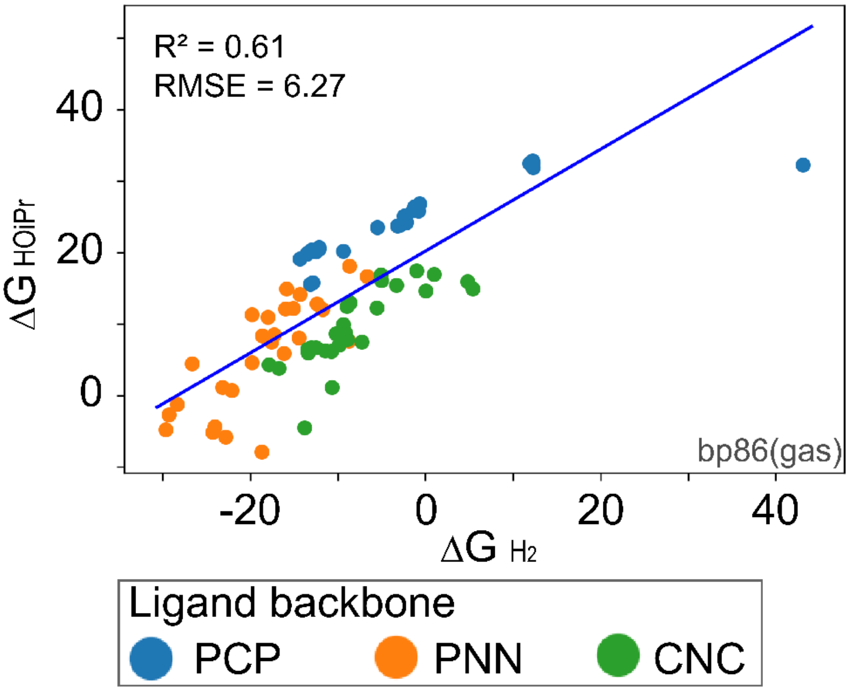


Figure S14. Comparison of DFT (bp86(gas)) computed Gibbs free energies for the formation of ^{i}PrO and hydride complexes. The data for catalysts with different backbones are differentiated by colors. All energy values in kcal mol^{-1}. All energy values in kcal mol^{-1}.

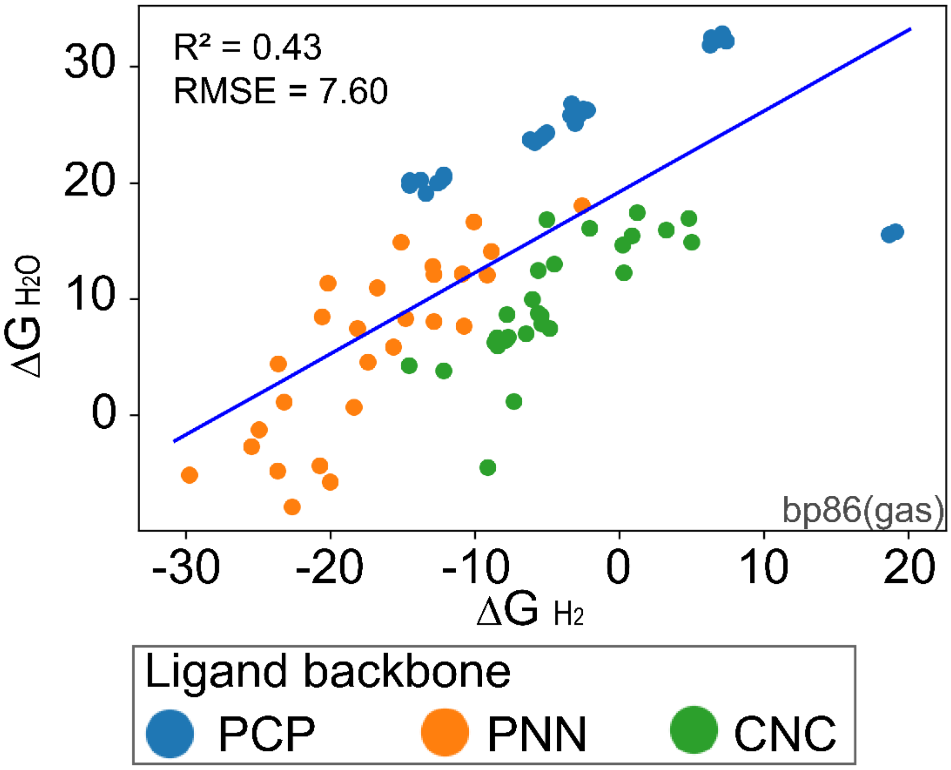


Figure S15. Comparison of DFT (bp86(gas)) computed Gibbs free energies for the formation of hydroxide and hydride complexes. The data for catalysts with different backbones are differentiated by colors. All energy values in kcal mol^{-1}.

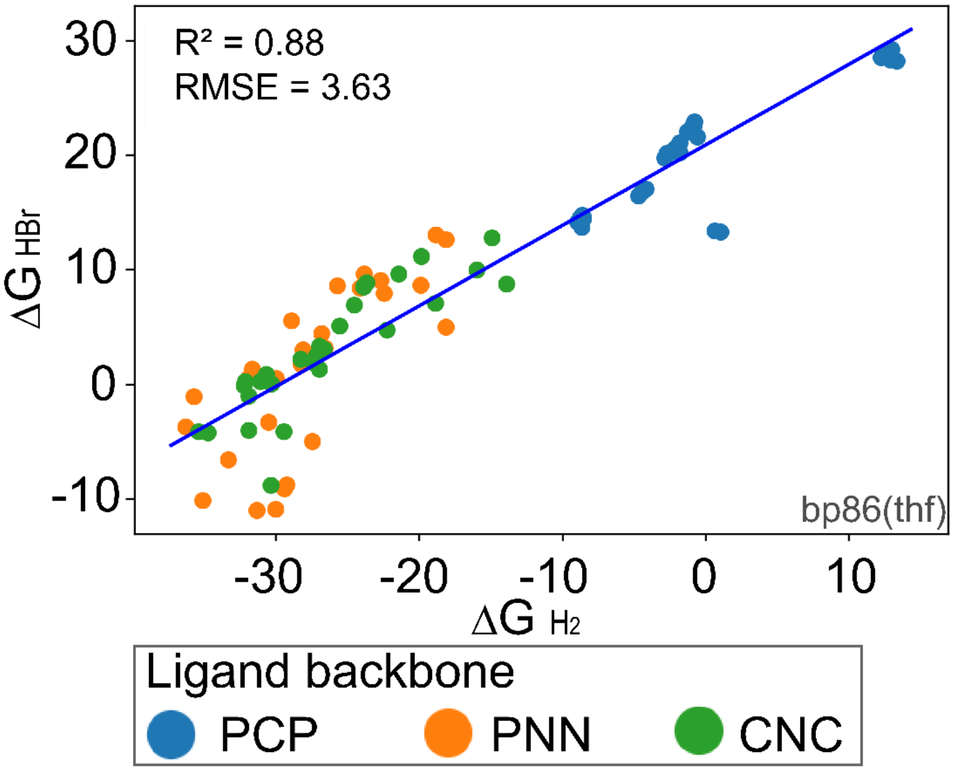


Figure S16. Comparison of DFT (bp86(thf)) computed Gibbs free energies for the formation of bromide and hydride complexes. The data for catalysts with different backbones are differentiated by colors. All energy values in kcal mol^{-1}.

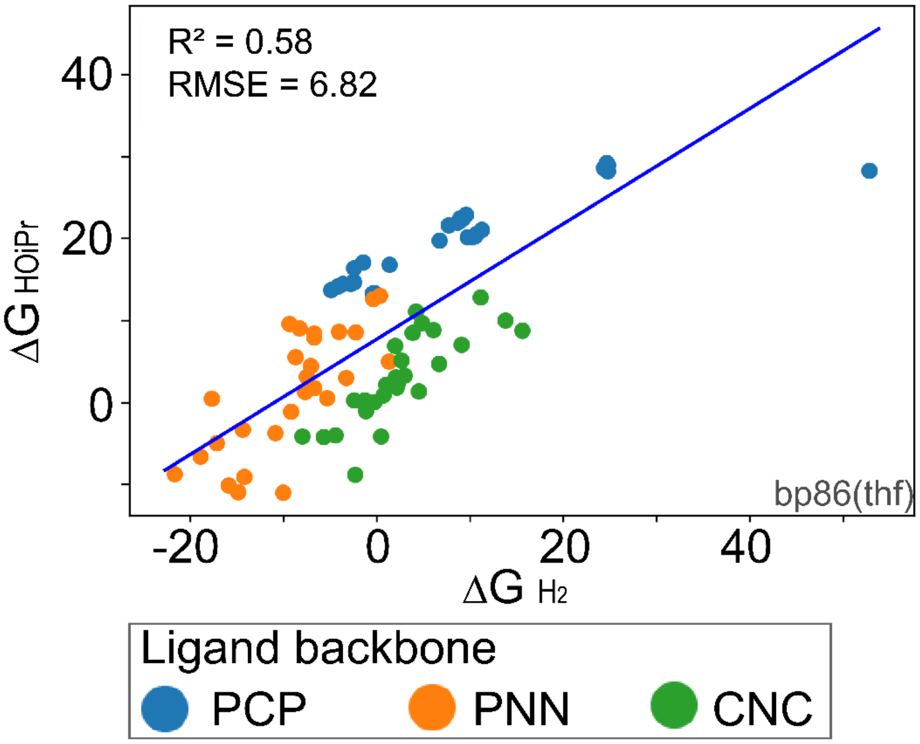


Figure S17. Comparison of DFT (bp86(thf)) computed Gibbs free energies for the formation of ^{i}PrO and hydride complexes. The data for catalysts with different backbones are differentiated by colors. All energy values in kcal mol^{-1}. All energy values in kcal mol^{-1}.

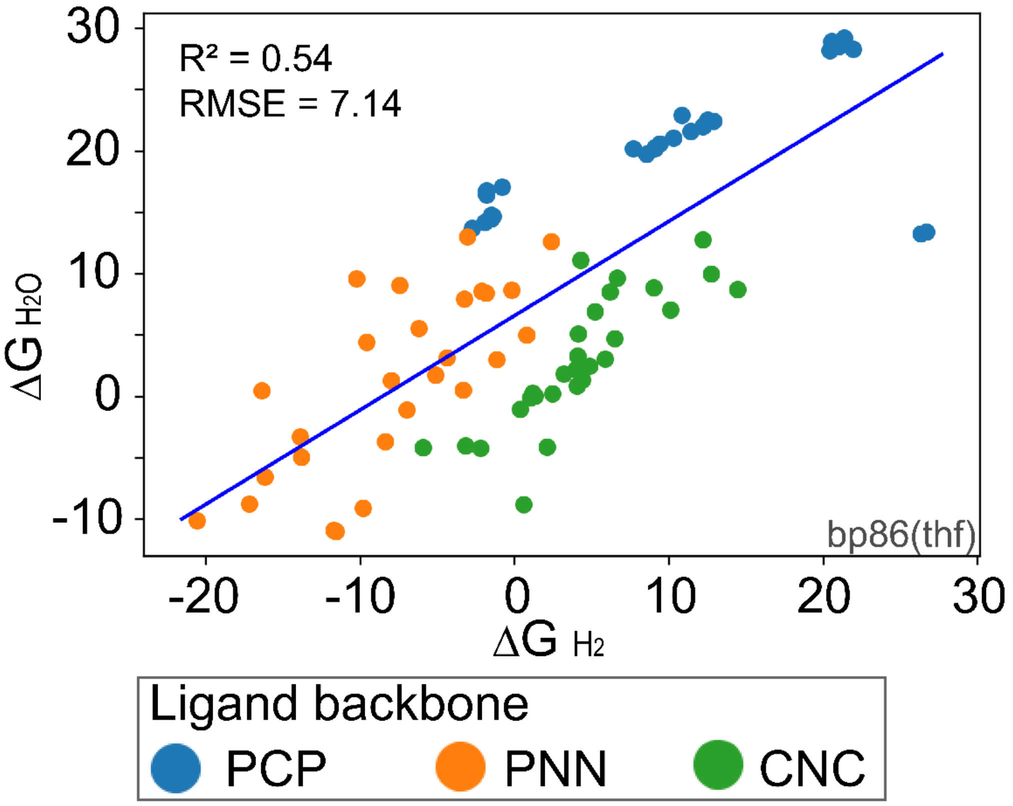


Figure S18. Comparison of DFT (bp86(thf)) computed Gibbs free energies for the formation of hydroxide and hydride complexes. The data for catalysts with different backbones are differentiated by colors. All energy values in kcal mol^{-1}.

Impact of solvation on Gibbs free energies

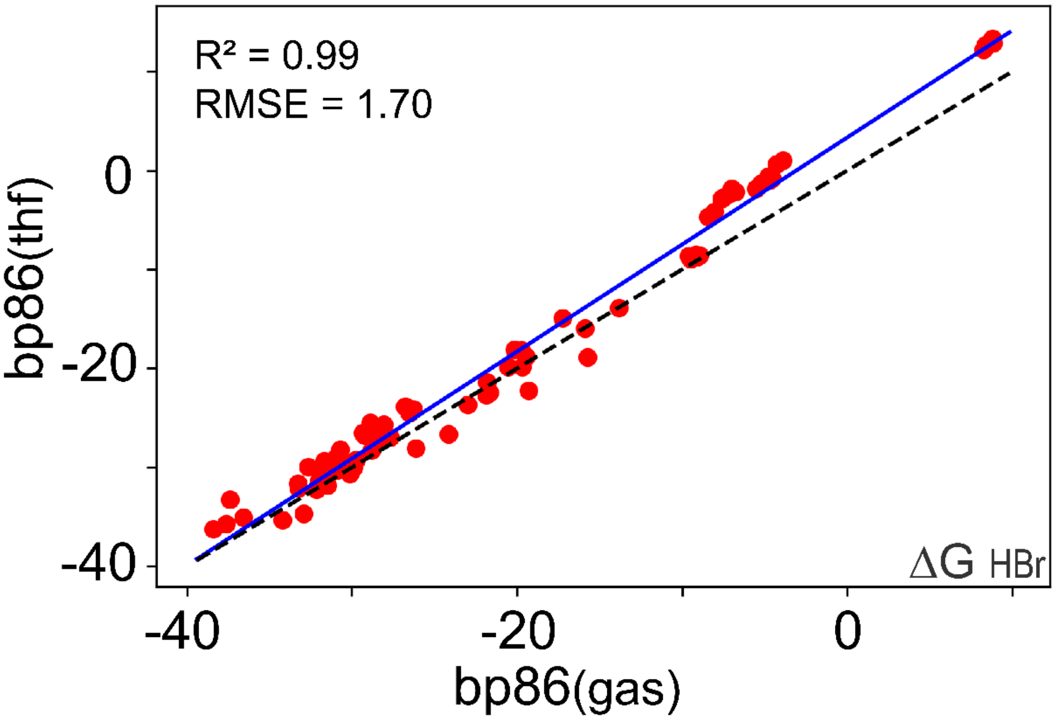


Figure S19. Comparison of Gibbs free energies for the formation of bromide complexes in gas and solvated phase. The blue line shows a linear fit of the data. The dashed black line is the y = x line. Points below the y = x line show stabilization of the bromide adduct upon solvation while points above this line show a destabilizing effect of the solvent.. All energy values in kcal mol^{-1}.

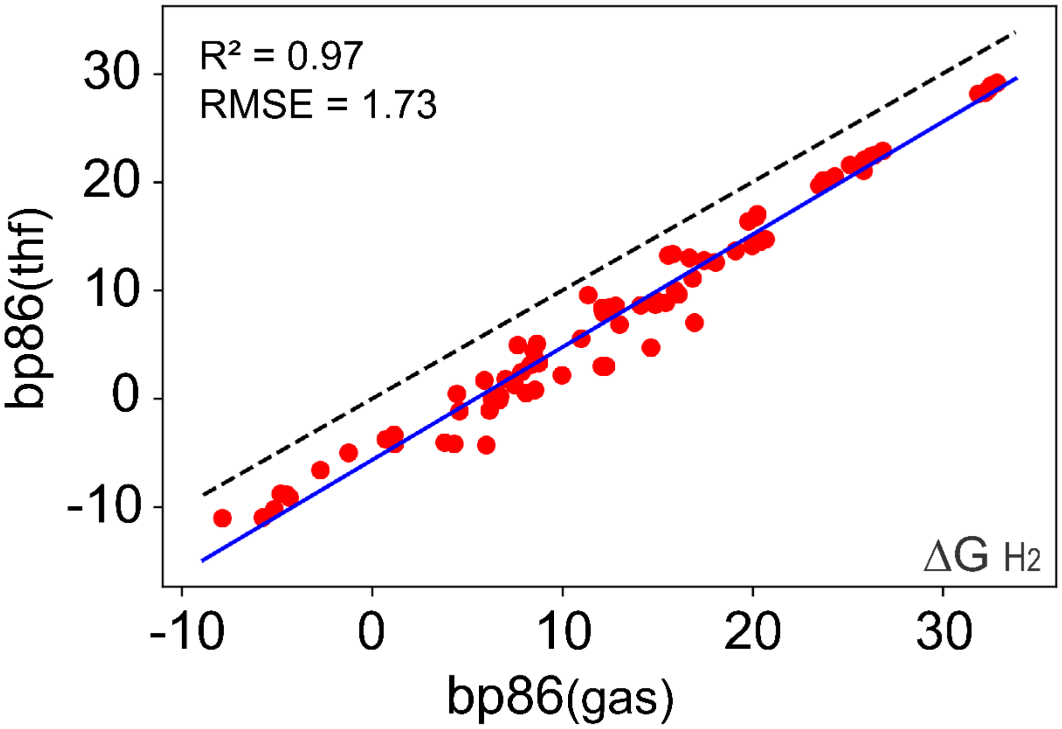


Figure S20. Comparison of Gibbs free energies for the formation of hydride complexes in gas and solvated phase. The blue line shows a linear fit of the data. The dashed black line is the y = x line. Points below the y = x line show stabilization of the hydride adduct upon solvation while points above this line show a destabilizing effect of the solvent. All energy values in kcal mol^{-1}.

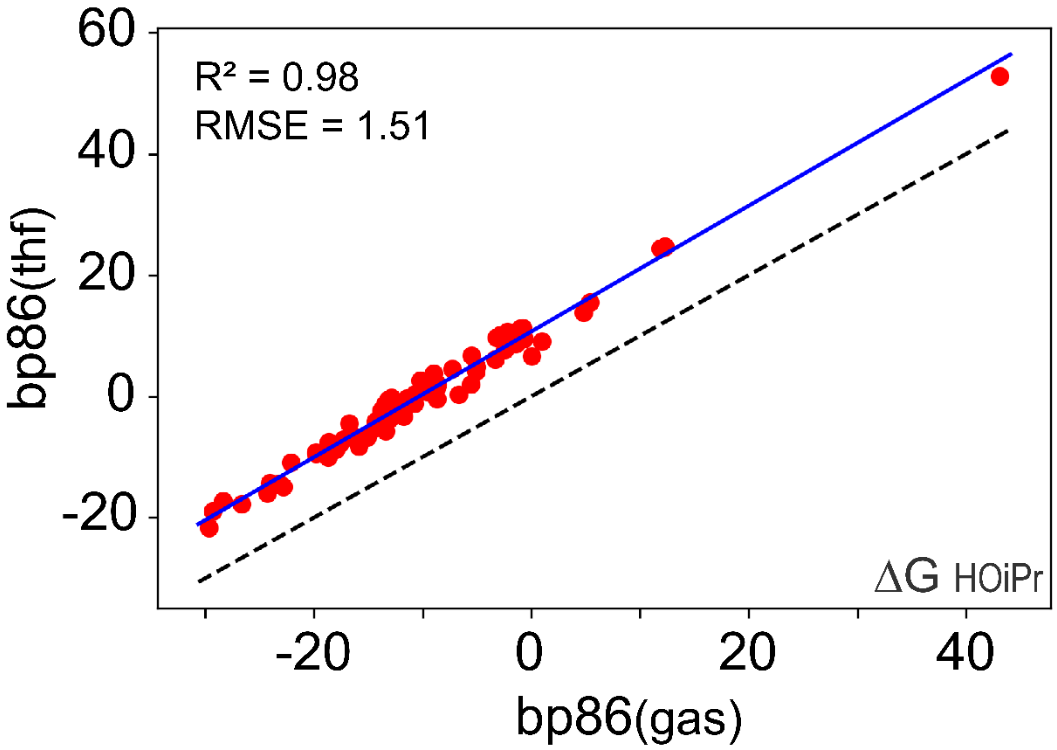


Figure S21. Comparison of Gibbs free energies for the formation of ^{i}PrO complexes in gas and solvated phase. The blue line shows a linear fit of the data. The dashed black line is the y = x line. Points below the y = x line show stabilization of the ^{i}PrO adduct upon solvation while points above this line show a destabilizing effect of the solvent. All energy values in kcal mol^{-1}.

Figure S19: The linear fit to the data is close to the y = x line but generally above it indicating that the addition of HBr complexes seems to be slightly less favoured in THF solvent compared to the gas phase.

Figure S20: The linear fit to the data is clearly below the y = x line indicating that addition of H_2 is favoured in THF solvent compared to the gas phase.

Figure S21: The linear fit to the data is clearly above the y = x line indicating that the addition of ^{i}PrOH is less favoured in THF solvent with respect to the gas phase.

Impact of XC functional on Gibbs free energies

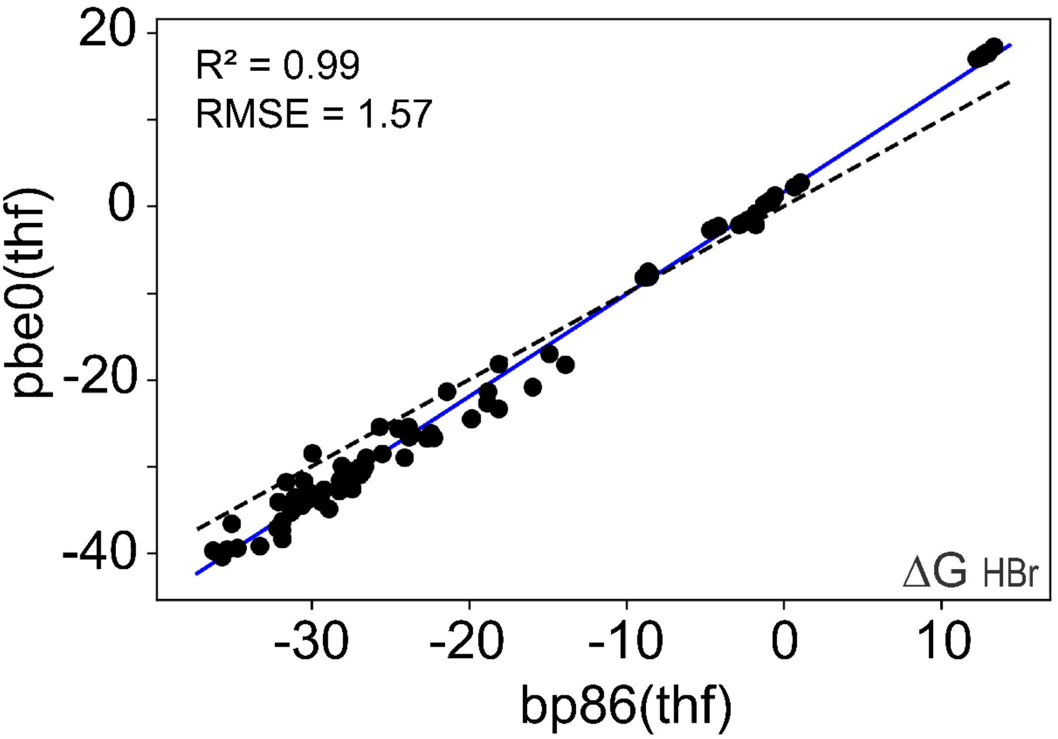


Figure S22. Comparison of Gibbs free energies for the formation of bromide complexes obtained using BP86 and PE0 XC functionals. The blue line shows a linear fit of the data. The dashed black line is the y = x line. Points below the y = x line show stabilization of the bromide adduct with PBE0 functional while points above this line show that BP86 overestimates the stabilization. All energy values in kcal mol^{-1}.

Figure S22 shows that pbe0(thf) and bp86(thf) agree reasonably well. The GGA functional BP86 seems to mostly underestimate the stability of bromide complexes.

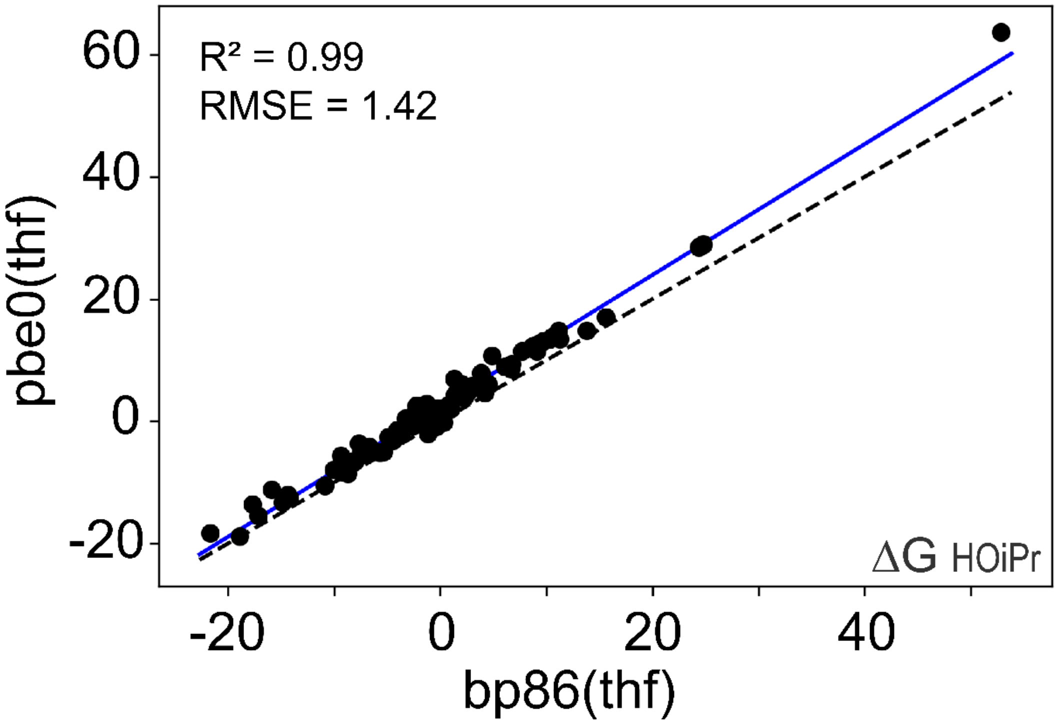


Figure S23. Comparison of Gibbs free energies for formation of ^{i}PrO complexes obtained using BP86 and PE0 XC functionals. The blue line shows a linear fit of the data. The dashed black line is the y = x line. Points below the y = x line show stabilization of the ^{i}PrO adduct with PBE0 functional while points above this line show that BP86 overestimates the stabilization. All energy values in kcal mol^{-1}.

Figure S23 shows that pbe0(thf) and bp86(thf) agree reasonably well. The GGA functional BP86 seems to mostly overeestimate the stability of ^{i}PrO complexes.

Comparison of xTB and DFT computed Gibbs free energies

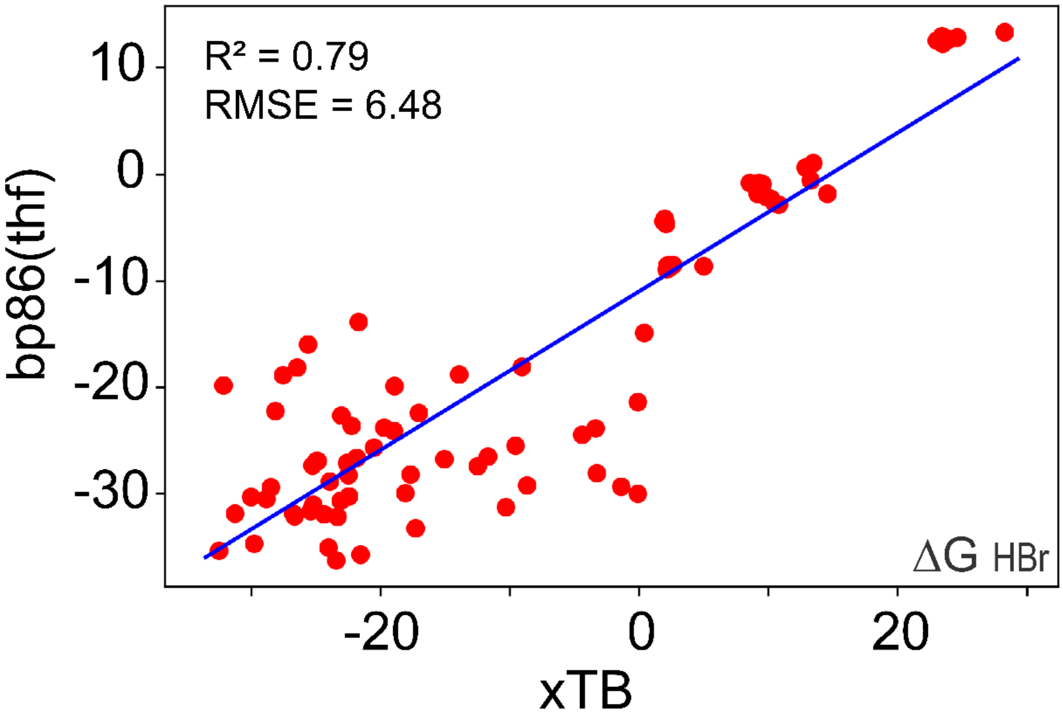


Figure S24. Comparison of xTB and DFT-computed formation energies of bromide complexes. All energy values in kcal mol^{-1}.

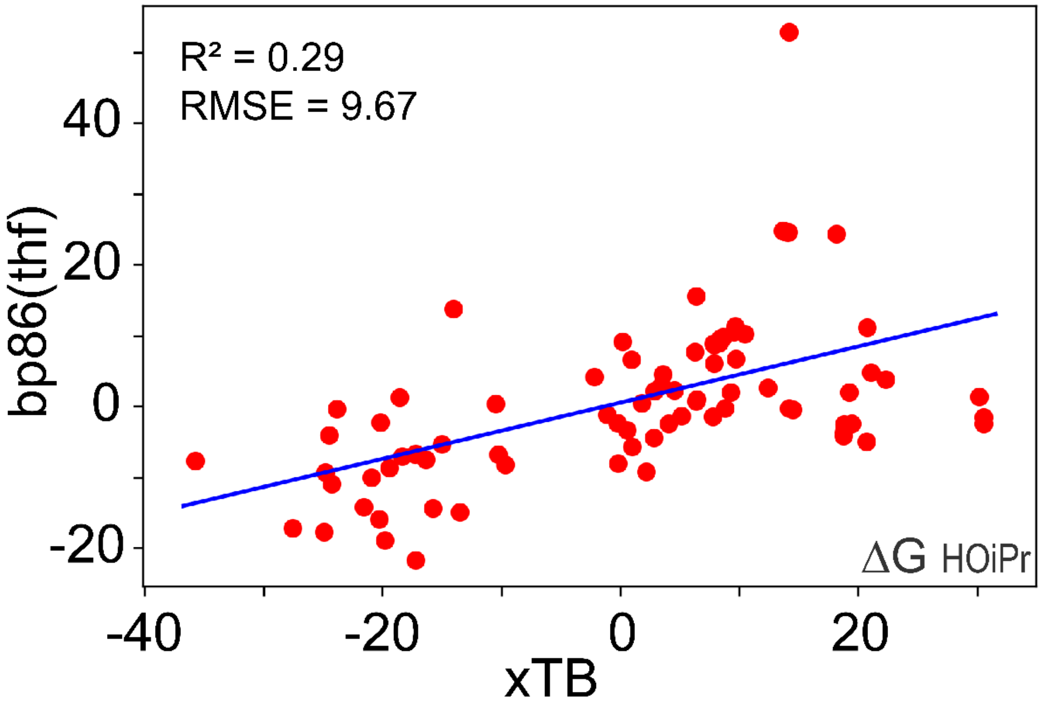


Figure S25. Comparison of xTB and DFT-computed formation energies of ^{i}PrO complexes. All energy values in kcal mol^{-1}. All energy values in kcal mol^{-1}.

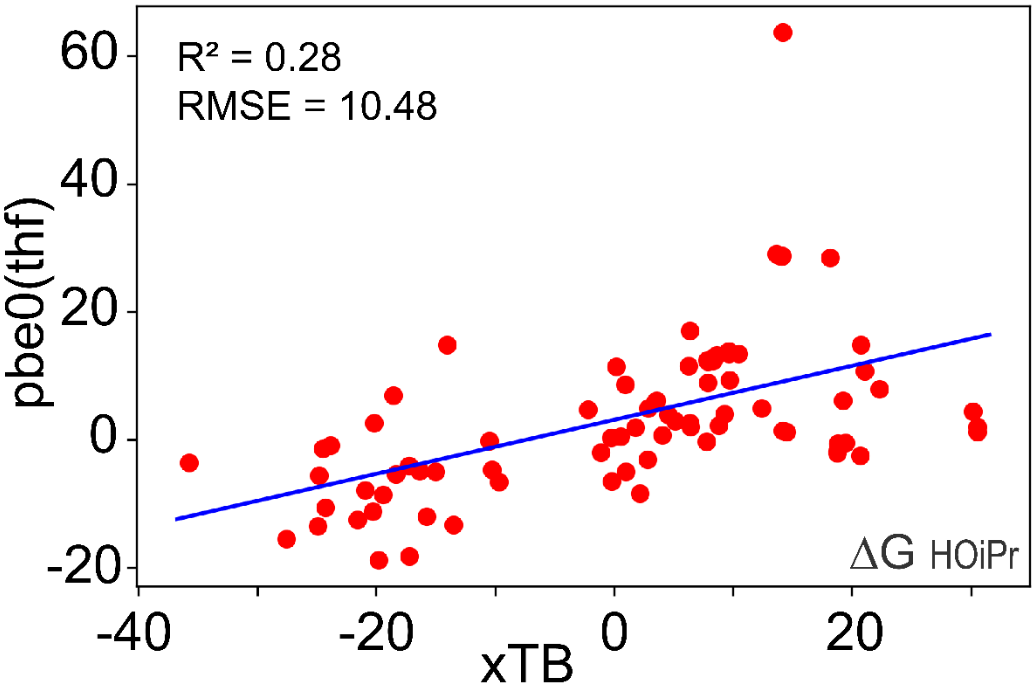


Figure S26. Comparison of xTB and DFT-computed formation energies of ^{i}PrO complexes. All energy values in kcal mol^{-1}. All energy values in kcal mol^{-1}.

S8. CO stretching frequencies

In this section figures displaying the CO stretching frequencies calculated by DFT (bp86(gas)) and/or xTB are illustrated.

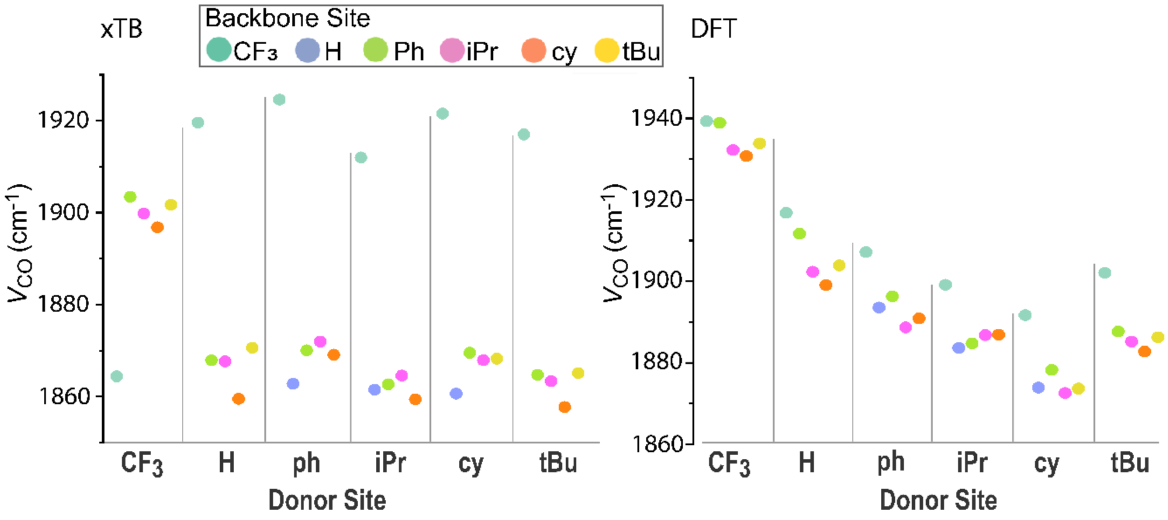


Figure S27. xTB (left) and DFT (right) computed CO stretching frequencies in the Mn-CNC complexes.

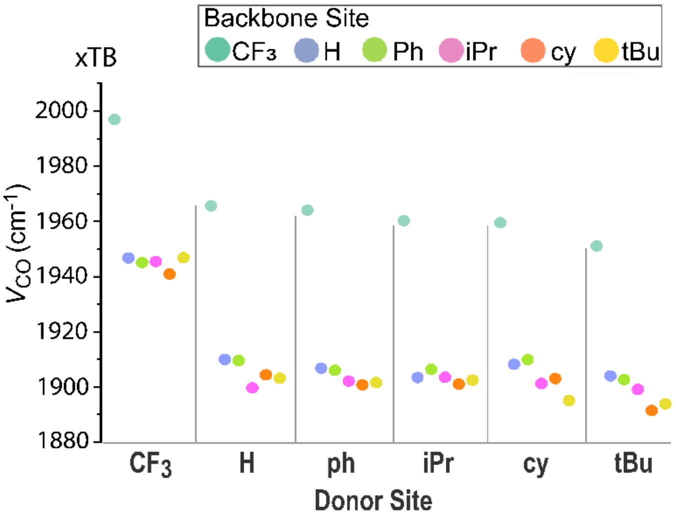


Figure S29. xTB computed CO stretching frequencies of Mn-SNS complexes.

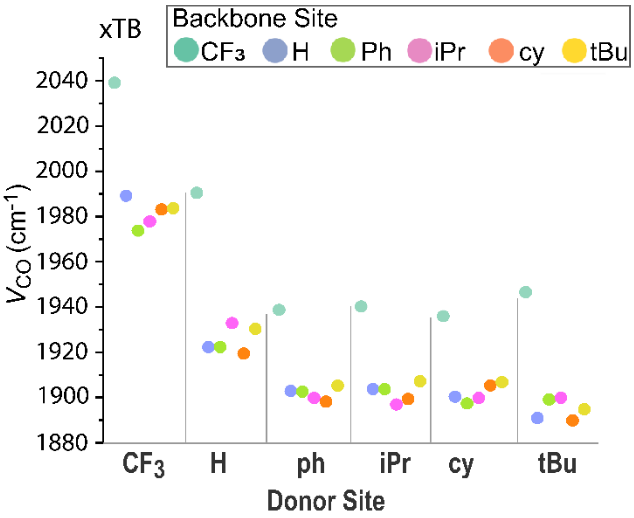


Figure S30. xTB computed CO stretching frequencies of Mn-PNP complexes.

C

Conferences/workshops/symposia logbook

C.1. Workshop Princeton University July 2020

On 13 and 14 July 2020, before the start of Adarsh's master project, an online workshop on molecular simulation with machine learning organized by Princeton University was attended. More information and, eventually, a recording of the session can be found on the website of the workshop and the website of the Computational Chemical Science Center. A summary of the attended lecture sessions of the workshop is given in this chapter.

C.1.1. Morning session 1, 13-7-2020

The first presentation was given by prof. Weinan E, on machine learning assisted modelling (MLAM). Neural networks offer an accurate approach to model potential-energy surfaces. These models work since the analytical option, a discrete fourier transform, does not work in higher dimensions. Instead a probability and expectation with low error rate are used.

Several methods were explained, a DNN-SGD and the Behler Parinello architecture being the main ones. The Behler Parinello architecture was explained in more detail, here each atom is a sub-network and the total energy of each atom needs to be fitted. The user needs to pay attention to the preservation of symmetry otherwise the same interaction will have a different energy.

Deep potentials in deep-potential generator (DP-GEN, a software package that was explained in the next presentation) were generated with concurrent learning using the exploration-labeling-training algorithm. The DeepMD package was used and 0.005% of the configurations given in this package were labeled [129]. It was concluded that MLAM can be a powerful tool for solving multi-scale problems.

The second presentation was given by prof. Roberto Car, on molecular simulation with the deep potential method. It was explained that the usage of DNN's has multiple benefits; Performance can be boosted, the 'curse' of dimensionality can be overcome and a molecule can be followed more accurately. The potential energy, polarization and polarizability can be used as observables. The mathematical details of fitting a DNN for this use-case were given, however at the time of writing the used presentation slides were not uploaded by the workshop organisation yet, so these details will be skipped. Using DNN's with DP-GEN and DeepMD were shown, multiple DNN's were fit to represent a vector $\vec{o}$ that contains a symmetry preserving continuous function f. The dependency is non-linear so each initialization leads to a new DNN ensemble. The cost of this method scales linearly with the system size and it was concluded that DeepMD is more efficient than DFT based ab initio molecular dynamics.

The third presentation, titled 'the art of non-boltzmann sampling', was given by dr. Pablo Piaggi. In molecular dynamics simulations a lot of the simulation time is spent on waiting for a rare event to occur. By using enhanced sampling the probability of the rare event is enhanced. A non-physical distribution $Q(R)$ can be chosen using two methods: the generalized ensemble method or the collective variable method. In the generalized ensemble method a random walk in potential energy space is realized and the simulation can avoid the multiple-minima problem. In the collective variable method, the large number of degrees of freedom of a physical system is reduced into a few ’collective variables’. These methods were explained in further mathematical detail and were shown to be available in the PLUMED plugin. A public repository which contains data to reproduce results of PLUMED-enhanced molecular dynamics simulations called PLUMED-NEST was shared as well.

Dr. Gary Grest presented use cases and several functions of LAMMPS, which is a large scale molecular dynamics simulation software that focuses on material modeling. It can make accurate descriptions of > 1000 atoms. The software package can be downloaded from its own website.

Prof. Thomas Kuhne presented the use cases of CP2K, this is a quantum chemistry and solid state physics software package that can perform atomistic simulations of solid state, liquid, molecular, periodic, material, crystal, and biological systems. It also provides a general framework for different modeling methods such as DFT using the mixed Gaussian and plane waves approaches GPW and GAPW. This software package can also be found on its own website.

C.1.2. Morning session 2, 14-7-2020

Dr. Michele Ceriotti gave a presentation on equivariant representations for atomistic machine learning. Atomic structures are mapped to mathematical representations, however basic physical symmetries should be included. This is hard to do in cartesian coordinates. Radial functions and spherical harmonics might be an outcome and the experimental Smooth Overlap of Atomic Position (SOAP) can be used for this. The representation of the structure is an additive property because the representation is local and it can be proved that by combining representations of the system that are sensitive to long-range effects, this method can outperform current machine learning methods, and provides a conceptual framework to incorporate non-local physics into atomistic machine learning [130].

PhD candidate Sebastian Dick’s presentation was titled ’Machine learning XC potentials in SIESTA: NeuralXC’. Spanish Initiative for Electronic Simulations with Thousands of Atoms (SIESTA) software package and method to perform electronic structure calculations and ab initio molecular dynamics simulations. As explained, ML might be used to develop density functionals. For the approaches that were considered in SIESTA, 2 choices have to be made, a grid based vs. a basis set expansion and a problem specific solution vs. universal functional.

2 packages for developing the functional were explained. ML-electron needs a small dataset, is based on unsupervised training and 4 functionals are obtained: LDA, GGA, meta-GGA and near region approximation (NRA).

In NeuralXC the electron density is projected on atom centered basis functions. The total energy is summed from atomic contributions. The potential is obtained by taking the functional derivative of this energy. The loss of the neural network is optimized by taking the RMSE, keeping ρ fixed while the machine learned parameters are being optimized and afterwards recalculating ρ using the self consistent field (SCF) [131]. Related to NeuralXC is the Deep post Hartree-Fock method (DeepHF), where eigenstates and eigenvalues are projected instead of a functional. The workshop was concluded with a panel discussion between all presenters, the audience could submit questions and topics to discuss.

C.2. NWO ACOS symposium October 2020

On 28 october 2020, while Adarsh was doing his master’s project and Covid-19 ravaged the world, an online symposium from Applied Computational Sciences (ACOS) organized by the NWO was attended. More information the session can be found on the website of ACOS. The sessions are posted on youtube. Summaries of the attended symposium sessions are given in this chapter.

C.2.1. Benjamin Sanderse - Uncertainty quantification meets machine learning

Dr. Sanderse from the scientific computing group of Centrum Wiskunde en Informatica (CWI) Amsterdam gave a presentation. He combines physics based modelling with data-driven methods.

Uncertainty quantification (UQ) and machine learning (ML) go hand in hand, an example was drawn with the modelling of sloshing of a liquid in a tank. Machine learning is used to reduce the complexity problem, PDE’s are used in the navier stokes equation and uncertainty quantification gives the margin in found parameters. Currently, uncertainty quantification is also used by CWI to determine the uncertainty in Covid-19 models used by the dutch government. Generally, machine learning emphasizes simplifying complex problems and is more data driven while uncertainty quantification focuses on applications in the real world and is more physics driven. Model reduction, regression and inference are used in ML and UQ, but nomenclature of the methods used is sometimes different.

If we have input z, PDE(u,z) and output u(z), we can say that the input and output have a probability distribution, ML comes into play in approximating the exact PDE, this combines physics driven and data driven methods.

An example was given with real time wind-turbine control, this requires unsteady flow models. The dimensions are reduced using projection (POD-Galerkin). In the reduced order model (ROM), the PDE is discretized, a matrix (snapshot) is gathered of results, singular value decomposition (SVD) is and the results are truncated. After some iterations this returns a ROM. We feed the input to the returned ROM, but we need to preserve the structure of the PDE in the ROM to have a stable solution. Another example was shown with shear-stress of a liquid with periodic boundary conditions and the order was reduced with O(100-1000) from 40000 to 16.

An example of a multi-level neural network was given. Here dimension reduction via grid coarsening is used, we want to get to a parametric circuit model without doing to many expensive computations. We divide the grid into coarse gridpoints and fine gridpoints. The error structure between the grid levels is learned to use a similar approach to multi-level monte-carlo methods. Use convolutional neural network (CNN) to go towards a MLP, train the model on the coarse grid and use it on the fine grid. It was shown that when the number of levels of the network is increased the error decreases fast.

C.2.2. Giuseppe Carleo - Machine learning in many-body quantum science

Prof. Carleo from EPFL Switzerland gave a presentation on the application of ML techniques on many-body quantum mechanics problems.

There is a problem in quantum science, if we look at a qubit it can have 2 states, we know that we need a multi-dimensional vector in space. This becomes very complex if we go to multi-body systems with many qubits for example, we go to a 2^n complex vector space. Since 2019 we can save the wave function of 54 qubits on the SUMMIT supercomputer, which is the current world record.

If we only focus on physical states in Hilbert spaces we don't have to store all the complexity of these wavefunctions. We want to find a way to encode this 'corner of the Hilbert space' with some vector which is parametrized by a large set of conventional parameters, using the Rayleigh quotient. If we try this with classic variational states we use our GPU, CPU etc. (conventional computer) but when approaching it with a quantum variational state we need a quantum computer.

In the classical setting we compute variational parameters by using a neural network (NN) [132]. Almost all variational representation and entanglement can be modeled using this method. There is a loss function but there is no dataset, it is closer to a reinforcement learning (RL) method.

This method is tested on frustrated 2D spins in a J1-J2 model. It is shown that making the network larger (adding more hidden layers?) the performance is increased [132]. A key insight was shown how many samples are needed to learn these wavefunctions, they found if you increase the number of states that the amount of samples increase [133].

For the quantum method a QPU is used with a variational quantum eigensolver. The same approach is taken as the classical approach but encoding now happens on the QPU and it's a composition of multiple layer of grids. The comparison between classical and quantum methods is made and shows how the quantum method is inspired by the classical method [134]. The group is working on Netket which is a python library that uses the classical method.

C.2.3. Tess Smidt - Neural Networks with Euclidean Symmetry for Physical Sciences

Dr. Smidt focuses on the use of neural networks to design crystal structures and gave a presentation on neural networks with euclidean symmmetry. She has posted her slides on a google drive.

Deep learning is a subset of machine learning which in turn is a subset of artificial intelligence. Our model is the function f(x,w) = y and the evaluation is done with loss = mean((y-ytrue))^2. Dr. Smidt used to have a atomic structure and use quantum mechanical methods to calculate properties, she wanted to use ML to make these calculations more efficient. She wanted to use inverse design to use properties and based on that design a good molecule while also mapping properties to a structure.

We have freedom to choose our starting coordinates in a molecule which is called euclidean symmetry E(3). We use 3 methods to modify our coordinates: translation, rotation and inversion. The magnitude of a vector is invariant, however the direction can be variant if you translate it.

We want to make a model that understands the symmetry of a structure. Generally there are 3 approaches:

1. Data-augmentation
2. Take a system and use invariant parameters to describe it
3. Make a model that uses equivariant and uses euclidean symmetry preserving

For 3D data the data-augmentation can get very expensive, you need a lot of training data. With the third approach of equivariant models, you would only need 1 example for the model to get it correctly. Dr. Smidt focuses on these equivariant models.

Euclidean models are similar to CNNs, but there are 2 important differences:

1. We use equivariant convolutional filters which are based on learned radial functions and spherical harmonics
2. Geometric tensor algebra allow us to generalize scalar operations to more complex geometric tensors

We give an input to the system, but we also need a representation list to show what changes when we apply translation, rotation or inversion.

Examples were shown, if the system is given a complex and it's rotated copy the force predicted will always be the same, the network thus learns well. For crystals represented as primitive unit cells, conventional unit cells, and supercells of the same crystal produce the same output, assuming periodic boundary conditions. The network can also can predict molecular Hamiltonians in any orientationfrom seeing a single example, so less data is needed. The code was also applied to molecular dynamics to show its data efficiency [135].

An advantage found by accident is that the input for these networks are only geometric tensors. Another advantage is that the outputs have equal or higher symmetry than the inputs.

The work of Dr. Smidt's research group on these E(3) neural networks is posted on github: e3nn: a modular PyTorch framework for Euclidean neural networks.

C.3. WWU Munster mini-symposium Molecular Machine Learning January 2021

On 14 January 2021, Adarsh was still working on his master's project and a new mutation of Covid-19 was ravaging the world. An online symposium from the WWU Munster was held, unfortunately the session was not recorded. The attended sessions are summarized in the next sections.

C.3.1. Abby Doyle - Machine learning for experimental synthetic chemists

Abby Doyle from Princeton University is focusing her work on data-driven machine learning. Toward these efforts, her group has utilized high-throughput experimentation (HTE) for the generation of multi-dimensional datasets and developed tools to automate the parameterization of reaction components using computationally-derived descriptors that can be correlated with physical behavior and chemical reactivity. Experimental synthetic chemists generate a lot of data related to reactions, yield, selectivity, but also analytical data such as structures of catalysts. A fraction of the data is saved in electronic notebooks and only a small part of that data is published.

ML can help in reaction prediction and in mechanistic understanding. To model experimental data, datasets from literature or *de novo* datasets need to be generation. *de novo* datasets are mostly smaller and may be more biased. The next step is featurization, how do you best describe the structure? Is there a universal way for this? After featurization, modelling needs to happen, there are various ways to do this (lin. reg. etc.).

An example is shown by finding new phospine ligands for a Ni catalyst. A student found out that only 2 commercially available pohsphine worked. It turned out that only phosphines that were functionalized on the meta site worked. These ligands were named after dinosaurs.

A multivariate linear model was made that could predict new phosphines using the cone angle, radius and buried volume (V_{bur}). The cone angle had a positive coefficient and the V_{bur} a negative coefficient. After analysis it turned out that a low buried volume and high cone angle is optimal for phosphine ligands on Ni catalysts.

Capturing 4D data on phosphine ligands can be important for activity-structure relations. The smallest conformational buried volume is also the smallest possible structure, a treshold (about 58 cubic angstrom) was established which shown that after a certain size no binding would happen. Another example was given with Pd catalysts, it is known that these are sensitive to speciation. It was observed that the treshold for Pd catalysts was similar to the one of the Ni catalysts (58 cubic angstrom).

Reaction optimization is being researched using bayesian optimization. There needs to be statistical evaluation for how the bayesian optimization (BO) works for synthetic chemists. The system needs to be tuned to various types of parameters. Chemistry knowledge needs to be included together with the selection of how many experiments are optimized at once (a batch). Search space, surrogate model and acquisition function are key to BO. auto-Qchem was used to automate DFT featurization.

50 chemists could compete against the BO system. BO was able to find the global optimum every time in only 5 batches of experiments. The human participants were able to get higher yields in some experiments, but the BO performed better on average. Ultimately the human experimentalists were outperformed because they thought they had reached the optimal yield and did not continue with experimentation.

C.3.2. Klaus Muller - Machine learning for the sciences, towards understanding

Klaus is an computer scientist, he is working on applying ML in quantum chemistry applications and he is trying to understand the black box of ML and AI. His opinion is very respected even with political people.

2 major workhorses in ML, one is kernel networks and the other is neural networks. ML is not about fitting data, but about generalizing a model. He explained NNs using a convolutional neural network (CNN). Error-backpropagation is the exact opposite, you take the output of the model and propagate it back to iterate and approximate a decision function. This theory is called the deep taylor decomposition. The relevance of each layer should be conserved in this process. Datasets can contain artifacts (copyright links etc.) that lets the model say the right thing but for the wrong reason.

ML for chemical space exploration, it was suggested that we use DFT optimized structures and learn a function from that to predict new molecules. The coulomb matrix was used first, where the interaction between the ith and jth matrix is shown. Using a kernel ridge regression a simple equation was left to solve. With deep neural networks an atomistic representation can be 'learned'. This is a newer and improved representation over the coulomb matrix [136]. With other ML methods like reinforcement learning (RL), a scanning probe microscope can be used to move multiple molecules instead of only 1 molecule. A video example showing how the RL model learns was shown.

Explaining and interpretation are important when bringing ML into chemistry because you have to know how and why something (does not) work.

C.3.3. Alan aspuru Guzik - There is no time for science as usual: Materials Acceleration Platforms

Alan is one of the most famous players in the field. He is currently focusing on self-driven laboratories and quantum chemical applications.

There is a certain time pressure for finding new materials (climate change, Covid-19 etc.). A typical material takes more than 10 years to market. Instead of the current paradigm, we need a closed loop using automation and AI (generative and backpropagation).

For photomaterials AI-discovery of materials was used, in 2014 they started working on this. 100.000 DFT complexes were used as training and other 300.000 candidates were predicted and synthesized [137].

For generative models autoencoders were used to encode a SMILES string and decode it again. They used a method that can relate this latent space to a function which makes it a differentiable. You can relate a property to latent space (backward searching). So you can find from properties 'which material shows this certain property?'.

Complicated generative models for reticular materials (MOFs for example) were published from their group. SELFIES is a representation of molecules that claims to be right in 100% of the cases where [138].

Deep molecular dreaming was also mentioned. A direct gradient-based molecule optimization that applies inceptionism techniques from computer vision was described. This exploits the use of gradients by directly reversing the learning process of a neural network, which is trained to predict real-valued chemical properties. This method uses SELFIES.

C.4. Workshop SCM Chemistry & Materials Modeling workshop for TU Delft January 2021

All material for this hands-on workshop was uploaded to the workshop website.

C.5. PAC symposium March 2021

On 4 March 2021, Adarsh was in the process of finalizing his project while the vaccination plan against Covid-19 in the Netherlands had finally started. The online PAC symposium, organised by 4 chemistry study associations (ACD, CDL, U.S.S Proton and VCSVU), was held. Various speakers from dutch universities and companies gave a talk and there was a panel discussion about the future of earth and the role of young chemists and engineers in this future. Nobel prize winner Prof. Dr. M. Stanley Whittingham presented his ground-breaking research on lithium-ion batteries.

Since the symposium was held online this year, the poster competition was replaced with a video-pitch competition. In this competition the participants had to make a video of 1.5 minutes explaining their thesis work. The maker of the best video was invited as speaker to the PAC symposium. Adarsh's submission was posted to linkedin (link to post). Unfortunately no notes were made of this symposium because Adarsh was present as one of the speakers thanks to his video-pitch submission :).