Transferability of descriptors for *in silico* catalyst screening

Aydin Najl Hossaini

TU Delft

Transferability of descriptors for in silico catalyst screening

by

Aydin Najl Hossaini

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday April 21, 2023 at 3:30 PM.

Performed at:
Inorganic Systems Engineering
Faculty of Applied Sciences

Under supervision of:
Prof. Dr. E. A. Pidko
A.V. Kalikadien, M. Sc.

Student number:	4752589	
Project duration:	August 21, 2022 – April 21, 2023	
Thesis committee:	Prof. Dr. E. A. Pidko,	TU Delft, supervisor
	Dr. A. Bansode,	TU Delft
	Dr. B. Bera,	TU Delft

Abstract

Homogeneous transition metal-based (TM) catalysts are crucial to producing chemically pure drugs, stemming from their ability to obtain high product selectivity. However, experimental screening of TM-based complexes is expensive, so computational methods are leveraged instead. Especially machine learning (ML) approaches show promise due to being efficient as well as unbiased. ML of homogeneous TM-based catalysts is based on physiochemical properties named descriptors. Descriptors are dependent on the method of simulation and the simulated complex itself. Methods with a higher level of theory are more accurate, but also more resource intensive. Similarly, larger complexes simply demand more computational resources. Two general methods to minimize the number of resources needed are: 1) using the lowest level of theory containing reasonable accuracy and 2) using the simplest representative complex.

In this thesis, possible simplifications were investigated for a homogeneous TM-based catalyst screening workflow. Objective 1 was investigating the effect of levels of theory for geometry optimization on descriptors. Structures were optimized for four levels of theory relevant to this workflow, namely: MACE, GFN-FF, GFN2-xTB and DFT. Subsequently, xTB level descriptors were calculated for the first three levels of theory and were then correlated against xTB level descriptors of the benchmark, DFT. In addition, it was investigated how descriptors obtained from xTB and DFT single-point calculations differ. Objective 2 was investigating the effect of the chemical structure on descriptors. To do so, a set of octahedral complexes and a set of simplified structures were generated and descriptors of both sets were correlated against each other.

Regarding objective 1, it was observed that solely descriptors from the GFN2-xTB level of theory correlated well with DFT, at least for the majority of descriptors. Next to that, it was found that GFN2-xTB geometries more or less coincide with DFT geometries. Regarding objective 2, it was found that the bidentate ligands in the model set deform towards the metal centre, which leads to decreased correlations among the majority of the descriptors. Additionally, it was found that clustering occurred due to the presence of two different ligand classes in the dataset.

The primary conclusion of this research was that geometries originating from GFN2-xTB geometry optimization are structurally comparable to geometries originating from DFT geometry optimization. However, descriptors obtained from GFN2-xTB single-point calculation are not comparable to descriptors obtained from DFT single-point calculation. As such, to accurately extract descriptors, DFT single-point calculations are necessitated.

Contents

List of Figures

Abbreviations

DFT	Density functional theory
DFTB	Density functional based tight-binding
FF	Force field
GGA	Generalized gradient approximation
HOMO	Highest occupied molecular orbital
hRMSD	Cartesian heavy-atom root-mean-square deviation
IQR	Interquartile range
KS-DFT	Kohn-Sham density functional theory
LDA	Local-density approximation
LFER	Linear free energy relation
LUMO	Lowest unoccupied molecular orbital
ML	Machine learning
MM	Molecular mechanics
NBO	Natural bond orbital
PES	Potential energy surface
QC	Quantum chemistry
QM	Quantum mechanics
RMSD	Root mean square deviation of atomic positions
RMSE	Root mean square error
SASA	Solvent accessible surface area
SCF	Self-consistent field
SM	Statistical mechanics
SMILES	Simplified Molecular-input line-entry system
TM	Transition metal
UFF	Universal force field

1

Introduction

Whether knownst or unbeknownst to the average Joe, catalysis is omnipresent in modern society. Behind the scenes, it serves as the foundation for modern materials and as the driver for high-standard living conditions. Some might be familiar with catalysis as a concept from their cars' catalytic converters, but many are oblivious to the scale catalysis encompasses. Catalysts, the substances that induce the process of catalysis, underpinned 30% of the Gross Domestic Product of European countries in 2016 and are used in approximately 80% of all manufactured products [1]. Use cases are ample, ranging from the medicines we take to the fabrics that keep us warm, to the fuels that power the vehicles that transport us [2].

The current definition of a catalyst is: "a substance which increases the rate at which a chemical reaction approaches equilibrium without becoming itself permanently involved" [3]. In other words, a catalyst gently provides an alternative pathway for creating and breaking chemical bonds. This alternative pathway possesses a lower activation energy, therefore increasing the reaction rate and speeding up the reaction.
In general, three major types of catalysts exist: heterogeneous, homogeneous, and bio-catalysts.

Heterogeneous catalysts exist in a different aggregation state than the reactants. Having catalysts in a distinct phase is an enormous advantage since separation and re-utilization becomes simple and cheap [4]. In heterogeneous catalysis the reaction mechanism usually goes as follows: the reactants are adsorbed on the surface of the catalysts, the chemical reaction occurs and then the product gets desorbed from the surface. Due to their general applicability and low cost, heterogeneous catalysts are mostly used in large-scale chemical engineering contexts and are closely connected to the fields of surface science and solid-state chemistry [5].

Homogeneous catalysts, on the other hand, exist in the same aggregation state as the reactants. Biocatalysts technically also fall under this definition, but usually are considered a distinct type and are generally labelled enzymes [6]. Homogeneous catalysts have superb selectivity and their respective mechanisms are generally better understood than heterogeneous ones [4]. Nonetheless, heterogeneous catalysts are usually preferred over their homogeneous counterpart, due to (expensive) separation difficulties.

In the context of pharmaceuticals, however, the utilization of homogeneous catalysts is pervasive, owing to the industry's target aim of developing safe and effective drugs for human usage. The presence of impurities, e.g. different stereoisomers, may result in the absence of a reaction or even toxicity in human consumption.

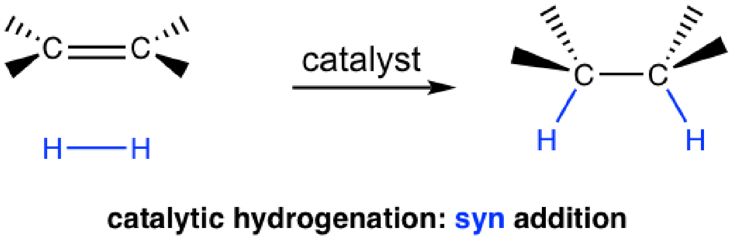


Figure 1.1: Simplest form of a hydrogenation reaction. Ethene is converted to ethane through a (TM-based organometallic) catalyst in a syn addition mechanism, where both hydrogen atoms are facing the same side. Schematic adapted from [7].

One example is the antidepressant drug Citalopram. The drug is sold as a mix of both the **S** and **R** enantiomers, however, only the **S** enantiomer is responsible for its beneficial effects [8], making the drug possessing a redundant compound. Another example is the drug **D**-penicillamine, which is used to treat rheumatoid arthritis. **L**-penicillamine, **D**-penicillamine's enantiomer, inhibits the function of an essential B vitamin [9] and is very toxic to humans, therefore the reaction of producing **L**-penicillamine needs to be pure.

For reactions concerning stereoselective drugs, organometallic complexes are often utilized [5]. These organometallic complexes consist of a metal centre connected to (in)organic molecules named ligands. Especially transition metal (TM) based organometallic complexes have been studied extensively in the past decades for their use in obtaining high selectivities [10].

One of the first successful applications of TM-based homogeneous catalysts is the reduction of unsaturated organic molecules by means of molecular hydrogen, a reaction known as hydrogenation. This reaction, which can be seen in Figure 1.1, consists of the addition of a hydrogen molecule to an unsaturated moiety, e.g. $C = C$, $C = N$, $C \equiv C$, $C = O$, etc.

Breaking double bonds of this moiety can lead to asymmetry of the product. In Figure 1.1 both hydrogen atoms are facing the same side (syn-addition), however, by utilizing a different catalyst, one hydrogen can face backwards and the other hydrogen face frontwards (anti-addition). As mentioned before, it is important to get stereochemically pure products in asymmetric catalysis and thus in asymmetric hydrogenation. For asymmetric hydrogenation extensive research has been done to obtain useful and pure stereoisomers through the noble-metal (Rh, Ir, Ru) catalysts with chiral ligands (phosphines) [11–16]. The ISE group and their pharmaceutical partner are currently continuing on the previous research by investigating the asymmetric hydrogenation of enamines with a rhodium metal centre and asymmetric hydrogenation of amines with an iridium metal centre. Their research serves as the inspiration for the work done in this thesis.

Selectivity originates from the ligands of the homogeneous catalyst and because of that, for the catalyst to have high regio-, diastereo- and/or enantioselectivity, appropriate ligands need to be selected [10]. Often, experimentalists use chelating ligands, i.e. ligands that bind to multiple sites (monodentate = 1, bidentate = 2, etc.) to get fitting selectivities. This chelating effect makes the ligand bind more tightly to the metal centre, therefore, creating more steric hindrance, leading to higher selectivities.

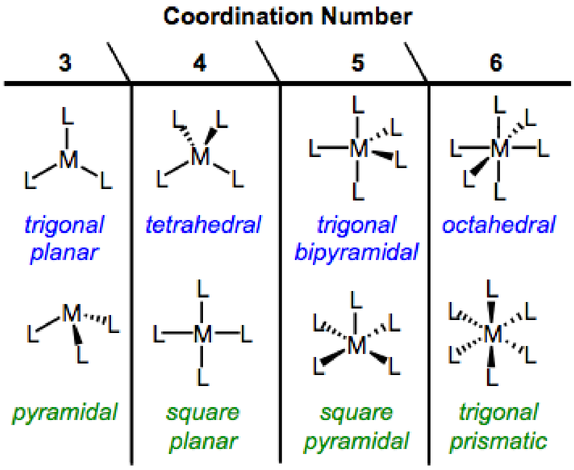


Figure 1.2: Different geometries for an organometallic complex. The coordination number refers to the number of ligands connected to the metal centre. Schematic adapted from [17].

Possible combinations of ligands and metal centres scale factorially, therefore making the chemical space close to infinite [18]. Solely by changing the number of binding sites the ligand has, e.g. monodentate, bidentate or pincer ligand, the complex's characteristics become completely different. Additionally, metals have multiple oxidation states, which allow for a different number of bonds. These different oxidation states give rise to different geometries with different amounts of ligands connected to the metal centre, see Figure 1.2. For example, the reactant will interact with an octahedral complex in a completely different manner than a square planar complex, either through steric hindrance or the different electronic properties of the metal. In a nutshell, catalyst design from the chemical space gets profoundly difficult.

Recent advances in analytical chemistry accelerated the exploration of this extensive chemical space, however. One such method is high-throughput experimentation (HTE), which leverages robotics to automate reactions. As such, for one specific target, many reactions can be done in parallel with varying conditions [19–21].

Still, experimental exploration of the chemical space remains costly. Therefore, current research also leverages *in silico* design, i.e. computational methods, to screen catalysts. These *in silico* methods to rationally design catalysts have been referred to as the "Holy Grail" of chemistry, showcasing its significance [22]. In recent years, many methods to design and select catalysts have been tried and proposed to find this elusive "Holy Grail" [23–27]. In Figure 1.3, three general approaches are outlined.

The first approach is automated *in silico* catalyst design, which is based on virtual screening methodologies. In this approach, the user enters a substrate in a workflow, then complexes are automatically generated through an exhaustive ligand library. These complexes are analysed for transition states, and subsequently, different conformers are sought and stereoisomers are found.

The second approach is rational design, which uses a chemist's knowledge to design new catalysts and subsequently model them to validate the catalysts' use. A complete computational study is subsequently carried out to find catalyst activity/selectivity.

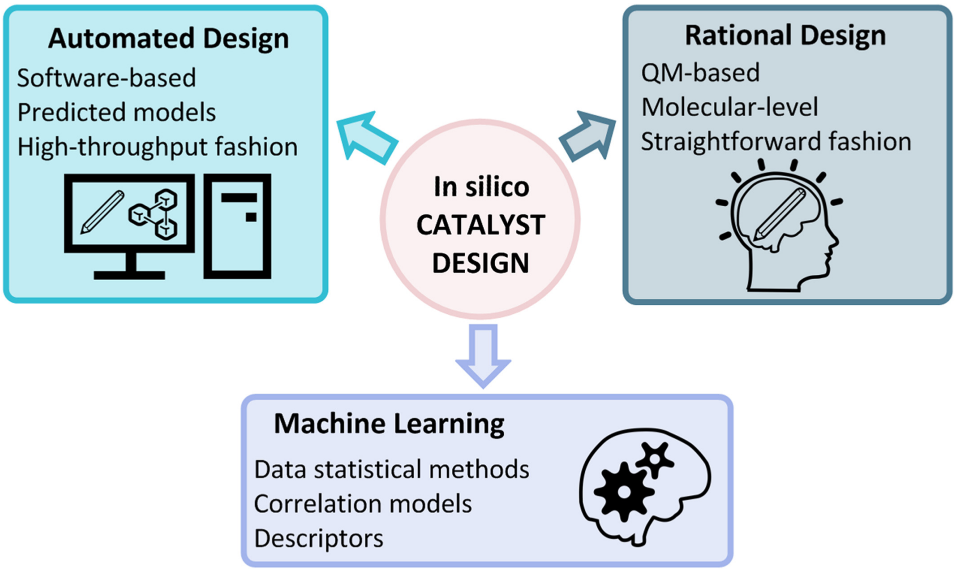


Figure 1.3: Overview of different approaches for computationally-led organocatalyst design. Figure adapted from [26].

The last approach is machine learning, which is based on physiochemical properties and mathematical relations to make predictive models over the chemical space. This approach especially shows promise as it is characterized by a faster rate of execution compared to the other two methods while remaining bias-free.

The methods of relating physiochemical properties to useful variables are called quantitative structure-activity or structure-property relationships (QSAR/QSPRs) [28]. Through QSAR/QSPRs, these physiochemical properties, called *descriptors*, can be related to mathematical equations, e.g. with linear free energy relationships (LFERs) [29, 30]. One example of LFERs is the Hammet equation, which relates the rate of a reaction to the electronic properties of the substitutes [31]. Another example is the Brønsted equation which relates, the Gibbs free energy for proton dissociation to the activation energy of the catalytic step [32].

QSAR/QSPRs are the relationships between molecular descriptors and the chemical space. By utilizing these relations, novel catalysts can be identified and analyzed. However, finding descriptors that represent structural properties properly is a major bottleneck in these models [33]. Descriptors can be based on simple information, such as the amount of X-type atoms in the structure, but also on higher level information based on the 3D geometry of the structure, e.g. the bite angle between the ligands and the metal centre. Many 3D descriptors originate from Tolman's work on homogeneous TM-based bidentate phosphorus complexes [34]. Some examples of these 3D descriptors include Tolman's own cone angle, polarity or flexibility, which in turn can relate to product conversion, product selectivity or turnover number [35].

The challenge with machine learning stems from obtaining adequate higher-level descriptors in organometallics. 2-D representations of organometallics, without mechanistic knowledge, are simply insufficient to obtain adequate descriptors. As such, 3D structures need to be created from scratch. Using a chemist's knowledge is one possibility, but this is slow and inherently biases the search for the chemical space through the chemist's intuition. The ISE group, therefore, strives to create a purely data-driven workflow to generate and screen molecules to omit this astutely titled expert bias. Descriptors themselves, however, are dependent on which computational or quantum chemical (QC) method is used to simulate the system. Methods that utilize more computational power, usually give more realistic structures and geometries, therefore creating more accurate values for the descriptors. These methods which give more realistic structures are usually referred to as methods of a higher level of theory. Ideally, one calculates descriptors on the highest level of theory at all times, but this would require large quantities of computational power. A balance needs to be found in-between levels of theory and computational efficiency. In this thesis, descriptors on different levels of theory, e.g. DFT and xTB-GFN2 are being compared to each other, to find this balance.

In order to obtain screening data of complexes on a sufficient level of theory, it is typically required to use high-performance computing, specifically supercomputers. Although computational power has increased exponentially in the past decade, computational experiments are still surprisingly not very cost- or ecologically friendly.

Computational time is measured in standard billing units (SBUs), which is the time taken for the calculation times the number of cores, compared to a benchmark calculation. In the ISE group, for organometallic complexes with 100-150 atoms, DFT calculations take 1132 SBUs on average. 1 SBU is approximately equal to 1 euro cent, hence calculating a single complex costs 11.32 euros. A real-life experiment of one complex at the ISE group costs about 3 euros, withholding all costs of lab space, electricity and labour. However, computational methods have the benefit of being easily restartable and being much faster than experimental experiments, whilst not producing any chemical pollution. These reasons make computational methods still very favourable. Regardless, for the amount of power a supercomputer uses on a yearly basis, a small city of 50,000 people can be sustained [36]. By using such large quantities of power, a huge carbon footprint naturally follows, which, for all supercomputers worldwide, comes down to 100 megatons of CO_2 emissions per year. Despite the anticipation of a 2- to 9-fold growth in the next decade, the current number is already comparable to the complete aviation section of the United States [37–39]. With the now irreversible effect of climate change, massive steps need to be made to lower emissions as much as possible [40].

One approach to lowering computational time, thus lowering cost and carbon footprint, is by simplifying the to-be-calculated system. By excluding atoms that are not part of the relevant chemical sites, the complexity of a system, i.e. degrees of freedom, gets reduced a multitude of times, leading to reduced computing time. A quote from C.J Cramer, a pillar of the computational chemistry field, encapsulates this concept perfectly:

> The talent of the well-trained computational chemist is knowing how to maximize the accuracy of a prediction while minimizing the investment of such resources. - *Cramer 2004* [41]

In other words, it is paramount for a computational chemist to minimize resources by using the simplest system with the lowest level of theory possible.

Therefore, the goal of this thesis was to compare possible simplification methods which could be made in the ISE group's homogeneous catalyst screening workflow. This was done by simplifications on:

1. The level of structure optimization and descriptor calculation

2. The chemical structure that digitally represents a catalyst structure in the virtual screening process

2 Theory

In this chapter, applicable computational and quantum mechanical (QM) background information is given for the computational methods used in this thesis. After that, in Section 2.5, the parameters (descriptors) to computationally represent, compare and analyse our TM complexes are outlined. The majority of information on QM computational methods has been extracted from the following two books:

- "Essentials of Computational Chemistry" - C.J. Cramer [41]
- "Computational quantum chemistry: molecular structure and properties in silico" - J.J.W. Mcdouall [42]

2.1. Computational Chemistry

Computational chemistry seeks to understand chemical systems. Computational screening, however, seeks to explore the chemical space. The potential energy surface (PES) is monumental in exploring this chemical space. The PES is the relationship, whether mathematical or graphical, between a molecule's energy and its geometry. The most intuitive model to represent molecules in chemical space is to consider molecules as balls (atoms) held together by springs (bonds). Usually, springs are in an equilibrium state, but through "grasping" atoms and therefore stretching or compressing the bonds, the potential energy of the molecule changes. This model is motionless thus the potential energy remains in the system. In a simple diatomic molecule, the PES can be expressed as the potential energy versus the bond length, see figure Figure 2.1A. However, for polyatomic molecules, the dimensionality increases, since the PES is a collection of all possible atomic arrangements. The PES, therefore, becomes a hypersurface and has $3N-6$ dimensions for $N \geq 3$ where N is the number of atoms, stemming from the three-dimensional nature of Cartesian space. Drawing hypersurfaces is impossible in our three-dimensional space, thus we take slices through potential energy surfaces that involve one or two coordinates (e.g. bond lengths). See Figure 2.1B . The PES is useful since by setting the first-order partial derivatives to 0, $\frac{\partial E}{\partial q_1} = \frac{\partial E}{\partial q_2} = \ldots = 0$ (where E is the energy and q_1 and q_2 is the bond length of arbitrary bond 1 and bond 2), the minima (stationary point) of the PES is given. This stationary point corresponds to the lowest energy state respective to that bond. The lowest energy state here corresponds with the most stable geometric structure.

If the second-order partial derivative is bigger than 0 for all variables, $\frac{\partial^2 E}{\partial q_1^2} = \frac{\partial^2 E}{\partial q_2^2} = \ldots > 0$, this means the stationary point is truly a minimum. However, if the second-order partial derivative is smaller than 0 in the direction of the most likely reaction pathway, this means the stationary point is a saddle point. Saddle points correspond to the transition states of the molecule. In this way, we can (intuitively) calculate the reaction pathway with its corresponding transition states and most stable geometry.

The PES is essentially a plot of the molecular energy versus the molecular geometry. The concept of molecular geometry is only valid because nuclei can be regarded as stationary with respect to electrons, since they are 2000x heavier, as proven by the Born-Oppenheimer approximation[44, 45]. The Born-Oppenheimer approximation states that the Schrödinger equation for a molecule can be separated into an electronic and a nuclear equation. Consequently, a molecule has a defined shape [41, 42, 46].

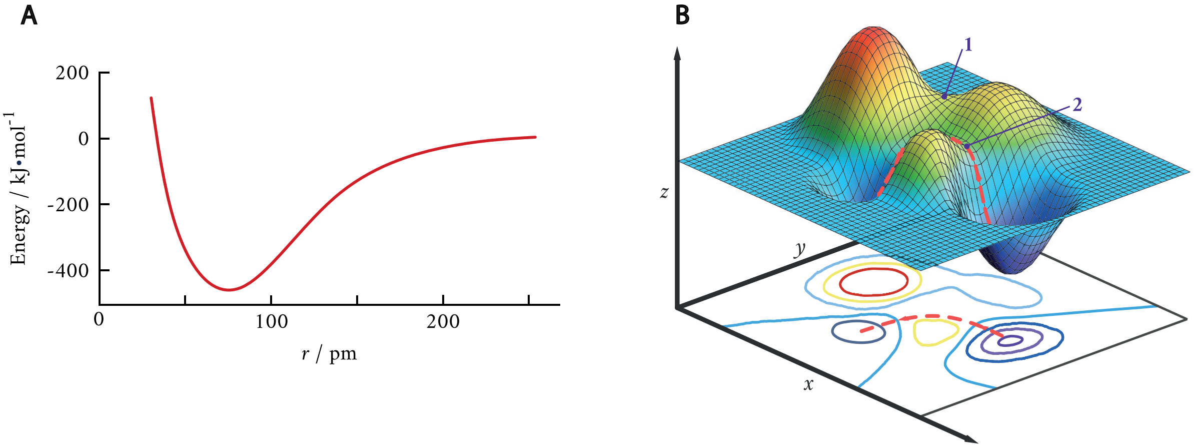


Figure 2.1: A: 2-D potential energy surface of a Hydrogen molecule (H_2) where r is the distance between the Hydrogen nuclei; B: Hypothetical 3-D PES with its corresponding contour plot for an endothermic reaction. Images reproduced from [43].

.

2.2. Molecular Mechanics

The basis of molecular mechanics (MM) utilizes the same model as the PES: a molecule is a collection of balls (atoms) held together by springs (bonds). Springs resist being distorted therefore creating different energies from different geometries. MM is thus a purely mathematical model to find minimum-energy geometries. The form of this mathematical expression for energy adds up to a force field, which is why MM methods are frequently called force-field methods. By nature, this method omits the electronic properties of a system and needs specification at what angle the bonds are drawn and how strong they are. Furthermore, vibrational energies are not taken into account with the MM's mechanical model, whilst atomic vibrations do in fact affect the spring (bond) length.

From the normal spring lengths and the angle between springs, the energy of this collection can be calculated with Equation (2.1).

$$E = \sum_{bonds} E_{stretch} + \sum_{angles} E_{bend} + \sum_{dihedrals} E_{torsion} + \sum_{pairs} E_{nonbond} \tag{2.1}$$

While the individual terms and their calculations are out of scope for this work, in every term a proportionality constant (k) is present. The process of finding values for k is called parameterizing the forcefield. Usually, experimental values or more accurate methods of representing the chemical space are used to get good fits for k. However, there are many interactions between atoms creating many different k values. Considering only the first 100 elements on the periodic table and making no distinction between single, double or triple bonds would lead to more than 10^8 parameters necessary to describe the force field. As such, these parameters are minimized either through chemical knowledge or by only considering a subset of all elements. Many force field methods also assign atom 'types', consisting of element (atom number) plus hybridization, oxidation state or another state altering atom behaviour. Furthermore, general force fields such as DREIDING, UFF and VALBOND decrease the scope by making almost all parameters, which depend on more than one atom, functions of single-atom-specific parameters. A more modern generic MM approach is GFN-FF, which seeks to model all complexes even beyond the 1000-atom size regime [47]. Approaches such as MM2 and MM3 introduce quantum mechanical (QM) enhancement in the form of π bond orders to the mechanical model, to more accurately calculate k values. Ultimately, different approaches in MM are catered to different systems and goals.

The process of exploring the PES with MM, by changing the bond lengths and bond angles, is called geometry optimization. Through iterative methods the lowest energy thus the most stable geometry is going to be found. These iterative methods are carried out by a minimization algorithm, usually some form of Newton-Raphson method. This process happens iteratively until the lowest energy is found. Note that this could be a local minimum and not necessarily the most stable state of the whole PES. The advantages of MM are that it is fast and hardware undemanding, while still being relatively accurate. One drawback is that it empirically estimates electronic properties with parameterization since electrons are completely ignored in the calculation.

2.3. Density Functional Theory

Density Functional Theory (DFT) is one of the cornerstones of computational chemistry and constitutes widespread use, even in biology and geosciences [48, 49]. Explaining DFT starts with quantum mechanics (QM) [50].

2.3.1. Quantum Chemistry

The fundamental equation (in its general form) of quantum chemistry (QC) is the time-dependent Schrödinger equation given in Equation (2.2)

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{r},t) = \hat{H}\Psi(\mathbf{r},t) \tag{2.2}$$

For stationary states, the equation can be simplified to its time-independent form, as seen in Equation (2.3).

$$\hat{H}\psi = E\psi \tag{2.3}$$

The Hamiltonian operator ($\hat{H}$) consists of a system's kinetic- and potential energy, combining for the total energy, see Figure 2.2. In this figure, a system of N electrons and M nuclei is described, where index i runs over electrons (N) and A/A' iterates over nuclei (M). ∇ is the Laplacian operator over i electrons. M_A denotes the mass of nucleus A. Z_A denotes the charge of nucleus A. r_{iA} is the distance between electron i and nucleus A. r_{ij} is the distance between electron i and j. $R_{AA'}$ is the distance between nucleus A and A'.

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

Figure 2.2: Hamiltonian operator in the non-relativistic, time-independent Schrödinger equation, figure reproduced from [51, 52].

.

In the case of very small molecular systems, e.g. $H_2{}^+$, Equation (2.3) can be solved by using the Hamiltonian operator described in Figure 2.2. The Hamiltonian describes pairwise interactions for all particles, implying that all particles are interdependent on each other. Thus to solve larger systems approximations are necessary. The first approximation is again the Born-Oppenheimer approximation. As mentioned in Section 2.1, nuclei are regarded as stationary compared to electrons. In doing so, the electronic wavefunction and the nuclei wavefunction can be solved independently, reducing complexity a great deal [53]. One method of solving the Hamiltonian for larger molecules is with the Hartree-Fock Self-Consistent Field (HF-SCF) method, where the secular equation is solved iteratively through an initial guess. Its main drawback, however, is that it uses one-electron operators, therefore ignoring all electron correlations.

Density Functional Theory (DFT) circumvents this issue by using electron density to solve the Hamiltonian, reducing the degrees of freedom from 3N to 3. The manner in which the Hamiltonian is solved is through the two Hohenberg-Kohn (HK) theorems [54].

The first HK theorem states that the energy of an atomic system, and all other observables, is unambiguously determined by the electronic density of the system. The second HK theorem states that only the ground-state electronic density, will minimize the total energy of the system [55].

Kohn and Sham realised a method to obtain the electron density by considering a set of non-interacting electrons [56]. This Hamiltonian can be expressed as a sum of one-electron operators, has eigenfunctions which are Slater determinants of the individual one-electron eigenfunctions and has eigenvalues that are the sum of the one-electron eigenvalues [41]. The starting point is a fictitious system of non-interacting electrons that have the same overall ground-state density as a real interest where electrons do interact. The energy functional (a function that takes another function as input) can therefore be expressed as in Equation (2.4).

$$E[\rho(\mathbf{r})] = T_{ni}[\rho(\mathbf{r})] + V_{ne}[\rho(\mathbf{r})] + V_{ee}[\rho(\mathbf{r})] + \Delta T[\rho(\mathbf{r}] + \Delta V_{ee}[\rho(\mathbf{r})] \tag{2.4}$$

Where ρ is the electron density, T_{ni} the kinetic energy of non-interacting electrons, V_{ne} the nuclear-electron interaction, V_{ee} the classical electron-electron repulsion, ΔT the correction to the kinetic energy derived from interacting electrons, ΔV_{ee} all non-classical correction to the electron-electron repulsion energy.

2.3.2. Exchange-correlation potential

The last two terms from Equation (2.4) are commonly added together to form the exchange-correlation energy (E_{XC}), which includes effects of QM exchange and correlation, correction for classical self-interaction energy and correction for the difference in kinetic energy between the fictitious and real system.

To solve E_{XC}, an exchange-correlation *functional* ($v_{XC}(\mathbf{r})$) is used. However, this functional is unknown and can only be approximated. Approximations generally are dealt with as an exchange term and a correlation term [57]. Due to this, functionals are frequently still called by a combination of both terms, e.g. PBE1PBE, instead of its respective general name PBE0 [58, 59]. Devising good functionals is the main problem in DFT since all difficult-to-calculate terms have been swept into the functional. Functionals form a hierarchical structure where the level of sophistication increases in the following order: a) the local density approximation (LDA), b) the generalized gradient approximation (GGA), d) meta-GGA, d) hybrid GGA and hybrid meta-GGA, e) random phase approximation (RPA). This hierarchical structure is also referred to as DFT Jacob's ladder, analogous to the biblical ladder reaching up to heaven, that culminates into the divine functional [60].

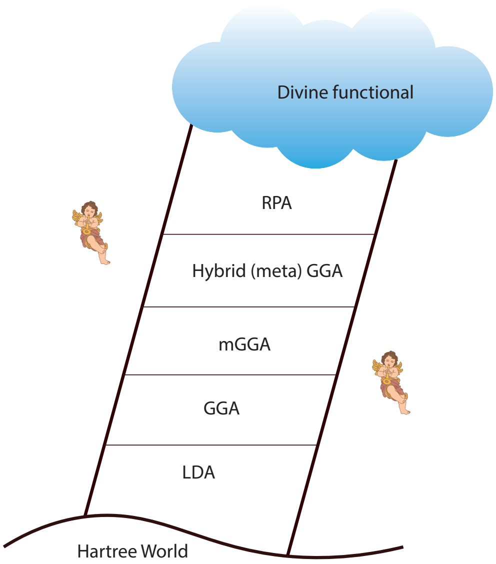


Figure 2.3: Jacob's ladder metaphor for DFT exchange-correlation functionals. The Hartree world represents Hartree approximation theory, while every subsequent rung represents a higher level of theory. The cloud represents the divine functional which solves every system exactly.

The simplest approximation is LDA, which solely depends on the value of the electronic density at each point in space to determine E_{XC}. LDA serves as a base for the construction of more sophisticated approximations such as GGA. In GGA, instead of taking the electron density as uniform in every point in space, it takes the gradient of electron density, in other words, where the density is locally changing. This approach is more generally applicable, since in a molecular system the electron density is not spatially uniform. Meta-GGA functionals improve on GGA functionals by adding a second derivative term to the electron density function (whereas GGA functionals only use the first derivative). Currently, Hybrid (meta) GGA functionals are also being used. HGGA functionals add the Hartree-Fock electron exchange, calculated from the Kohn-Sham wavefunction of noninteracting electrons. RPA is the most novel method and the fifth rung of accuracy. RPA is an approximate, but fully nonlocal method to determine E_{XC}. Here, one assumes that electrons only respond to the total electric potential. A dielectric function can be calculated which then adequately predicts properties of the electron gas [61, 62].

2.3.3. Basis sets

The set of functions making up the wave function of an atom in DFT is called a basis set. Basis sets are mathematical approximations of the actual wave function, which are theoretically most accurate when infinite, but larger sets take up more computational power. One method to create basis sets is to use the Linear Combination of Atomic Orbitals (LCAO) method. Atomic orbital basis sets are usually atom centred, meaning the same as real orbitals, they are centred in the core of an atom. Using the LCAO method, the best types of basis sets are formed with Slater-Type Orbitals (STO), which are based on the exact solution of the single electron Hydrogen atom system. However, STO basis sets are computationally expensive, so instead they are approximated by Gaussian-Type Orbitals (GTO). Multiplying GTOs of atoms has been shown to give a complete set of functions for molecules [63].

The main drivers of bonding are the valence electrons. As such, the inner-shell AOs and valence AOs are commonly split. The inner AOs are described by a single basis function while the valence AOs are described by two or more basis functions. This method is dubbed split-valence polarization (SVP). To increase accuracy further, basis sets include polarization functions to account for the polarization of the electron density of the atoms in molecules. Lastly, the addition of diffuse functions is common to account for the parts of the AO that are far away from the nucleus [64–66].

2.3.4. Solvation

With the currently described methods, only the non-interacting (gas phase) molecule has been taken into account. In real-world experiments, solvent molecules are often present, therefore they have to be considered during in-silico research as well. Solvation is included either explicitly (discreet) or implicitly (continuum). Explicit solvation models include solvent molecules explicitly, creating a physically solved description of the solvent, but interactions between molecules increase a great deal, increasing computational load. Implicit solvation models use a homogeneous polarizable medium for the solvent, which is much more computationally efficient, but fail to take local fluctuations of the solvent density into account [22, 67–69].

2.3.5. Dispersion corrections

In larger molecules, London dispersion forces become more apparent. Therefore, accounting for these forces becomes paramount to reaching chemical accuracy in Kohn-Sham DFT [70, 71]. Multiple methods have been developed to tackle dispersion forces, such as atom pairwise sum over $C_6 R^{-6}$ potentials (DFT-D), dispersion-corrected atom-centred potentials (DCAP) and "pure" density functionals (DFs) [72]. Grimme refined the DFT-D method to gain higher accuracy, a broader range of applicability, and less empiricism, to get the D3(BJ) and its novel successor D4 methods [73]. Grimme's D3 and especially D4 have unprecedented accuracy and range of applicability to many systems, including TM complexes.

2.4. Density Functional Tight Binding

Semi-empirical DFT combines ab initio (Hartree-Fock theory) calculations with empirical parameters for the correlation functional, whether from a higher level of theory or experimental parameters, to obtain a molecule's energy [74]. Density Functional Tight Binding (DFTB) is one such method, where the parameters are obtained by calculating the Hamiltonian and orbital overlap out of AO's [75, 76]. This method originates from the tight-binding model of bands, describing the electronic structure [77]. The exchange-correlation functional is then described a Taylor expansion of the KS-DFT total energy around a reference energy density [78].

Much research has been done in recent decades on DFTB methods, giving rise to the extended Tight-Binding method (xTB), which was introduced by Grimme et al. The extended refers to parameter availability to nearly the full set of elements of the periodic table ($Z \leq 86$). The theory originates from DFT perturbation expansion (adding successive corrections) of the electron density in fluctuation terms to various orders similar to the original DFTB model.

This xTB method is designed for calculations on Geometries, (vibrational) Frequencies and Non-covalent interactions, hence the full family of methods are dubbed GFNn-xTB, with $n = 0, 1, 2$ [79]. n refers to different versions that exist within the xTB family. In this manner, GFN1-xTB is the antecedent of the GFN family and is in turn based on DFTB3. Like DFTB3 GFN1-xTB uses a Taylor-expansion up until the third term to approximate the system's energy, but GFN1-xTB does not depend on atom pair-wise parameterization [79, 80]. Instead, it uses element-specific empirical fitting to enable consistent parameterization of a large part of the periodic table.

GFN1-xTB and DFTB3 both use monopole-type, spherically symmetric descriptions of the atom pairwise electrostatic interactions, which impairs describing non-covalent interactions. GFN2-xTB was created to solve this issue by firstly including anisotropic (direction-dependent) second-order density fluctuation effects via short-range damped interactions of cumulative atomic multipole moments, therefore it includes electrostatic interactions and exchange-correlation effects with greater accuracy [81]. This method increases the physical basis for calculating the Hamiltonian. Secondly, improvements stem from the dispersion model used. GFN1-xTB uses the older D3 dispersion model with Becke-Johnson damping, while GFN2-xTB uses a modified D4 dispersion model. D4 is modified by taking atomic partial charges from a Mulliken population and are solved self-consistently, which allows a large three-body term to be dropped from the dispersion energy equation [82]. Lastly, D4 does not use any element-pair-specific parameters and corrections, while D3 does have necessary H-H and halogen bond corrections. Consequently, GFN2-xTB uses only global and element-specific parameters, which is an improvement compared to its predecessor. GFNn-xTB methods use a polar continuum (implicit) solvation model in the form of a generalized Born model. Here a molecule is considered as a continuous region with a dielectric constant ϵ_{in} surrounded by infinite solvent with a dielectric constant ϵ_{out} [83]. Additionally, a nonpolar contribution is added to the solvation energy in the form of the solvent-accessible surface area (SASA). Lastly, an additional shift is included depending on the chosen reference state of the solution to ultimately reproduce COSMO-RS16 solvation-free energies [81, 84]. In the GFN family, improvements are still ongoing. GFN0-xTB recently came to live, but only currently exists as a preliminary, proof-of-principle version. In GFN0 electrostatics are treated classically and only keep QM contributions to the electronic structure of the first order, leading to speed-ups of factor 2-20 [85].

The GFN family's main advantage is that almost any chemically interesting species can be computed due to having parameters for almost all elements of the periodic table (including transition metals). By having all these parameters, a good balance is found between accuracy and computational efficiency. Furthermore, it is designed to be able to handle systems of up to a couple of thousand atoms.

2.5. Descriptors

Molecular descriptors represent each structure uniquely. Descriptors are subsequently used to predict experimental catalyst activities via quantitative structure-activity or structure-property relationships (QSAR/QSPRs) [28]. QSAR/QSPRs are mathematically quantified forms of the molecular structure of catalyst compounds and as such relate descriptors to the figure of merit, i.e. a quantity that describes a catalyst's usefulness [86].

Multiple methods exist to correlate descriptors (also called features in ML context) to figures of merit, for example (multivariate) linear regression or machine learning [25, 87, 88]. Molecular descriptors generally fall in one of four categories [89]:

- 0D descriptors (atom type, molecular weight, bond types)
- 1D descriptors (counts of atom types, fingerprints, one-hot encoding)
- 2D descriptors (topology and connectivity-based descriptors)
- 3D descriptors (QC descriptors)

If the dimensionality of the descriptors increases, so does the necessary computational time, since higher dimensionality leads to higher degrees of freedom. Therefore, 0D, 1D and 2D are very useful to screen large amounts of structures quickly. However, this advantage is offset by several limitations. Firstly, they ignore the conformational space. Secondly, they are unable to treat chirality, which is central to asymmetric reactions and thus selectivity. Lastly, they lack mechanistic interpretation [28, 90]. Whilst for large biological systems these concessions are acceptable, in homogeneous catalysis, differences between isomer energies are minuscule enough to necessitate 3D descriptors [28].

Descriptors are then ranked and correlated to the figure of merit. Although this approach cannot guarantee that the model includes all the important parameters. One method to solve this is to include as many descriptors as possible and then rank the best descriptors using selection algorithms or chemical intuition. Many descriptors exist, however, and all of them describe the chemical space in a slightly different way [91]. An overview of the descriptors used in this research is given below and given a particular category. Note that to a certain degree, categories are arbitrarily defined. In some descriptor cases, arguments could be made to move the descriptor to another bin.

2.5.1. Geometric descriptors

Geometric descriptors describe a structure's (bond) angles and are independent of any electronic calculation, instead are solely dependent on Cartesian coordinates.

Bite angle

The bite angle is the angle (denoted by β) between the central metal atom of the catalyst and its two chelating ligand atoms, as can be seen in Figure 2.4A. In this simplified illustration, purple refers to the metal, the bidentate phosphine is denoted with orange, the connection between the phosphines are some undefined atoms and the blue spheres denote rest groups. The bite angle is only valid for bidentate ligands and thus for bidentate phosphine ligands specifically, their use is widespread, due to their ability to adopt a wide range of bite angles. The bite angle is a useful parameter to explain observed rates and selectivities [92].

Cone angle

The Tolman cone angle (denoted by θ), from here on simply referred to as the cone angle, is defined as the apex angle of a cylindrical cone centred at a distance of 2.28 Å and extended to touch the van der Waals radii of the outermost atoms of the ligand [34, 93].

Donor-metal bond length

The donor-metal bond length is the Euclidean distance between the central metal atom of the catalyst and the chelating atom(s) of the ligand. For example, in the case of a P-P bidentate ligand, two descriptors would be calculated, the $M - P_1$ bond length and the $M - P_2$ bond length. Bond length is useful as a descriptor since the bond length is generally correlated to bond strength. Bond length's dependence on bond strength originates from the potential energy terms of the Hamiltonian which are Coulombic, and are consequently inversely proportional to distance [94]. On the other side, bond-length bond-strength correlation is not a law but simply an empirical correlation and bond length is influenced by many other factors such as strain, steric effects, dispersion stabilization, hybridization defects and so on [95]. One or all of these effects can void this empirical correlation.

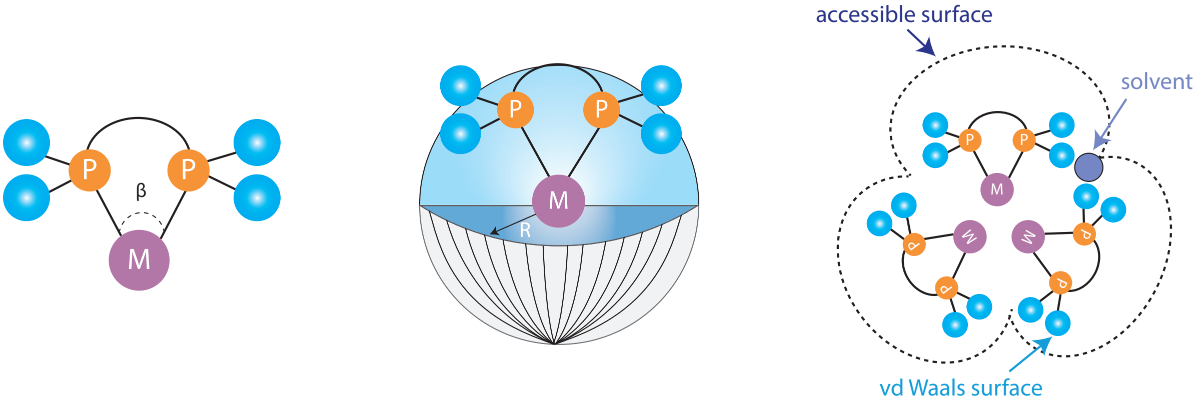


Figure 2.4: Schematic representation of 3 descriptors. Left: Bite angle (β) is the angle between the bidentate chelating phosphorus atoms and metal centre. Middle: 3D representation of the percent buried volume (%V_{bur}. Grey is empty space of the buried volume sphere, blue is how much of the ligand occupies the sphere. Right: The solvent accessible surface area is how much the solvent (dark blue) can roll over the van der Waals radii of the molecules. Colour code of atoms: light blue = rest groups, purple = metal centre, orange = phosphorus.

2.5.2. Steric descriptors

Enantioselectivity often comes down to very small energy differences in the order of ±3 kcal/mol [96]. These energy differences are caused by chiral ligands and their steric hindrance. As such steric descriptors try to modulate steric hindrance and are key in differentiating between enantioselective and non-enantioselective ligands.

Buried volume

The per cent buried volume (%V_{bur}), from here on referred to simply as buried volume, is a metric for steric hindrance in TM complexes similar to Tolman's cone angle [97]. It is the per cent of the total volume of a sphere centred around the metal that a ligand occupies, see Figure 2.4B [98, 99]. In this illustration R denotes the radius of the sphere, the blue part is how much of the ligand occupies the sphere and the grey part is how much of the sphere is empty. As such, $\%V_{bur}$ determines what percentage of a metal-centred sphere of the defined radius is occupied by the ligand [91]. The sphere radius and metal-donor distance affect this parameter but have been shown to correlate highly with cone angles [97]. Though initially developed for N-heterocyclic carbenes, which could not be described by cone angles, its use is also common in bidentate phosphine ligands [100]. Computationally, $\%V_{bur}$ is calculated as atoms either being present in the sphere or adding their respective volume cumulatively. That volume term is then divided by the total volume of the sphere to obtain the *per cent* buried volume.

Quadrant and octant buried volume

In asymmetric catalysis, chirality and asymmetry are crucial. Reactions happen on specific sections of the catalyst to favour one reaction giving rise to an excess of one enantiomer over the other. To this extent, the quadrant and octant per cent buried volume descriptors were developed. The approach is exactly the same as for $\%V_{bur}$, see Figure 2.4B but the sphere is cut twice vertically to obtain quadrants and once horizontally to obtain octants, see Figure 2.5. These regions can highlight optimal reaction pathways for enantioselectivity [101].

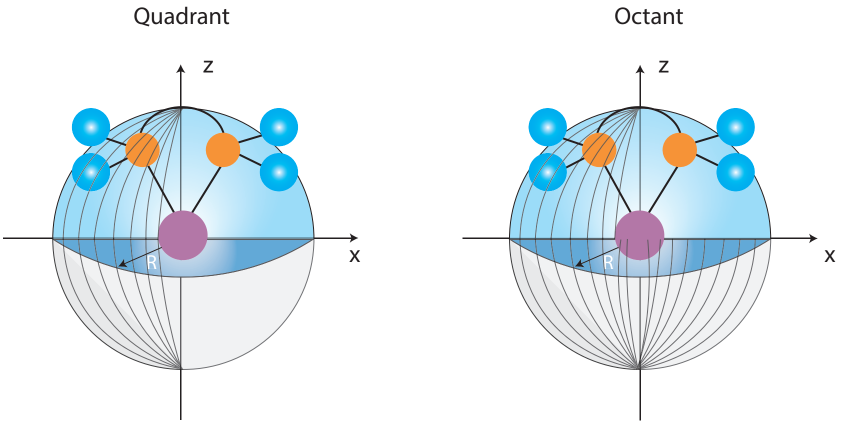


Figure 2.5: Quadrant (left) and octant (right) $\%V_{bur}$ schematic. Curved lines showcase the excluded volumetric parts. Colour code of atoms: light blue = rest groups, purple = metal centre, orange = phosphorus.

.

SASA

The Solvent-Accessible Surface Area (SASA) is the surface area of a molecule that can be accessed by the solvent, see Figure 2.4C. In this illustration, the solvent (light blue) "rolls" along the van der Waals surfaces of the molecules, creating the accessible surface which is drawn with dashed lines. This accessible surface serves as the metric for steric hindrance.

2.5.3. Electronic descriptors

After approximating the Schrödinger equation, the total energy of the system is known, but there is still a need to relate this energy in the form of physical descriptors to catalyst activity. Electronic descriptors stem from the LCAO-molecular orbitals (LCAO-MO) theory. In this model, every AO is a basis set for the MO. From linearly combining one AO, two MOs are formed: a bonding- and an anti-bonding orbital. This can be seen in Figure 2.6, where every black horizontal line represents a molecular orbital.

Electrons (black arrows) fill these orbitals in a bottoms-up manner. Bonding orbitals are lower in energy, as such correspond to the lower three (electron-filled) MOs. The upper three MOs are anti-bonding orbitals and each anti-bonding orbital corresponds with its respective bonding orbital.

Many electronic descriptors are then calculated from either filled or empty MOs, again see Figure 2.6.

Dispersion

The dispersion descriptor (P_{int}) is a universal quantitative descriptor of London dispersion interaction potentials [72, 99]. The London dispersion force is a force emerging from attractions between instantaneously

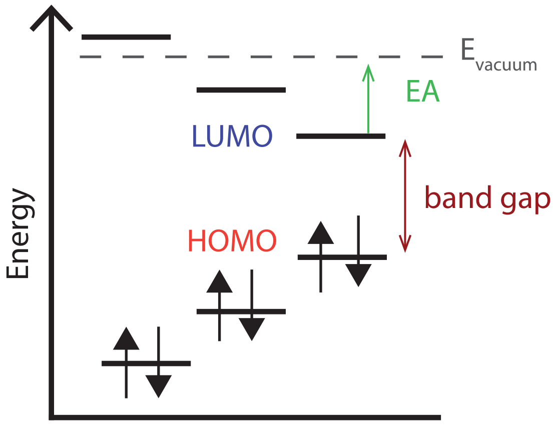


Figure 2.6: Schematic of 4 electronic descriptors as acquired from the LCAO-MO method. Electron affinity (green), lowest unoccupied molecular orbital (blue), highest occupied molecular orbital (red), band gap (brown).

induced dipoles on neighbouring atoms, see Figure 2.7 [102]. In this schematic, an instantaneous dipole moment in the atom gives an unsymmetrical distribution of charge, creating partially charged atoms. Similarly, this will affect the neighbouring atom by creating an attractive force on behalf of two (local) charge differentials. The dispersion on an atom is calculated as the average dispersion interaction energy of an atom with respect to its van der Waals surface. While London dispersion forces have been ignored by the scientific community until recently, increasing effects of dispersion are being noticed in organometallic and inorganic compounds [102, 103]. As such behaviour of such complexes needs to be modelled and often has a defining structural role. Dispersion is usually calculated as the dispersion on the metal centre but can be calculated on every atom individually.

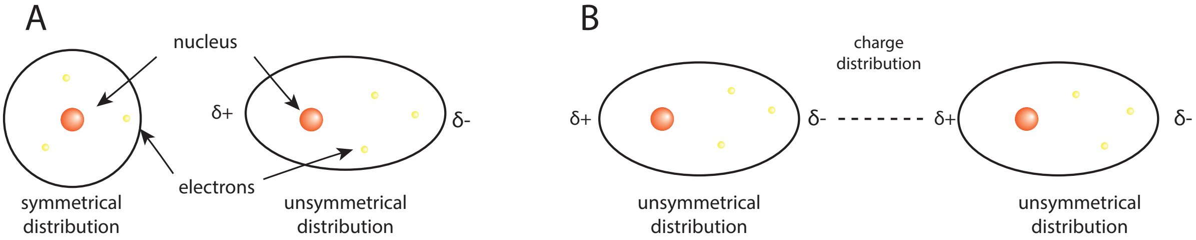

A
nucleus
δ+
δ-
electrons
symmetrical distribution
unsymmetrical distribution
B
charge distribution
δ+
δ-
δ+
δ-
unsymmetrical distribution
unsymmetrical distribution

Figure 2.7: Schematic representation of London dispersion forces. A: Unsymmetrical distribution of atoms creates instantaneous dipole moment across atoms. B: Unsymmetrical distribution of atoms 1 and 2 creates charge distribution across molecules.

Dipole

The molecular dipole is an effect arising from the uneven distribution of the electrons between two atoms in a molecule, due to different electronegativity values of atoms [104]. Dipole is then calculated as the length of a vector that represents the distribution of charges over a molecule, see Equation (2.5).

$$\vec{\mu} = \sum_i q_i \vec{r_i} \tag{2.5}$$

In this equation, $\vec{\mu}$ is the dipole moment vector, q is the magnitude of charge i, and $\vec{r_i}$ is the vector representing the position of charge i. Convention is that the vector of dipole points from positive to negative. The larger the difference in electronegativity between atoms, the larger this vector becomes.

HOMO-LUMO gap

The highest occupied molecular orbital (HOMO) is the highest filled orbital in the LCAO-MO theory, whilst the lowest unoccupied molecular orbital (LUMO) is the lowest empty orbital, see Figure 2.6. In this schematic, the black arrows represent electrons. In line with the definition, electrons are only filled until the HOMO. The energy difference between these two is the HOMO-LUMO gap, also known as the band gap. This difference in energy is imperative for optoelectronic materials since photovoltaic effects originate from exciting electrons through this barrier. In TM complexes HOMO-LUMO gap has been used to predict the strength and stability of the respective complex [105]. The HOMO serves as a potential place where the electron could attack and the LUMO as the potential place where nucleophiles could attack [106].

HOMOs and LUMOs are paramount in QC methods, since their levels themselves are a parent metric for many descriptors, due to the influence of Koopmans' theorem [107, 108].

Ionisation potential & Electron affinity

The ionisation potential (IP), also known as ionisation energy, is the amount of energy required to remove an electron from the isolated molecule or atom. Quantitatively it can be expressed as follows:

$$\mathrm{X(g)} + \mathrm{energy} \rightarrow \mathrm{X^{+}(g)} + \mathrm{e^{-}} \tag{2.6}$$

The electron affinity (EA) is the exact opposite of the IP, instead, it is the amount of energy required to add an electron to the isolated molecule or atom [109]. It is thus the energy required to excite an electron to vacuum (infinitely away) from the LUMO, as can be seen in Figure 2.6. Quantitatively it can therefore be expressed as follows:

$$\mathrm{X(g)} + \mathrm{e^{-}} \rightarrow \mathrm{X^{-}(g)} + \mathrm{energy} \tag{2.7}$$

From Koopmans' theorem, it can be stated that IP and EA are equivalent to the negative of the HOMO and LUMO, respectively [107]. For (semi-empirical) DFT, this technique is used to simplify calculations by skipping solving the Hamiltonian for an excited state.

Nucleophilicity & Electrophilicity

A nucleophile is an atom or functional group that has an electron pair available for bonding. It can then donate this pair to form a new bond either with an electron-poor atom, which is called an electrophile. Nucleophilicity (NP) refers to the ability of a nucleophile to donate electrons, whilst electrophilicity (EP) refers to the ability of an electrophile to accept electrons [110, 111].

The formula for electrophilicity from DFT is given in Equation (2.8) [112].

$$\frac{(\frac{1}{2} * \mathrm{LUMO} + \mathrm{HOMO})^2}{2 * (\frac{1}{2}(\mathrm{LUMO} - \mathrm{HOMO})} \tag{2.8}$$

Nucleofugality & Electron fugality

Nucleofugality (NF) is the ability of a chemical species to depart from the bonding electron pair. On the other hand, electrofugality (EF) describes the ability of a chemical species to depart and leave behind the bonding electrons [113, 114].

Electronegativity

Electronegativity is a fundamental measure of the ability of an atom to attract electrons and can be used to predict the polarity of a bond and the distributions of electrons in a molecule. The equation for electronegativity as defined by Mulliken is given in Equation (2.9) [112, 115].

$$-\frac{1}{2} * ((\mathrm{LUMO} + \mathrm{HOMO}) \tag{2.9}$$

Hardness & Softness

Lewis acids/bases are species that donate/accept a pair of electrons to form a coordinate bond. The concept of hardness (η) and softness (σ) originates from the hard and soft acids and bases (HSAB) principle.

According to the HSAB principle, Lewis acids and bases can be divided into hard or soft types [116].

- Hard Lewis acids are characterized by small ionic radii, high positive charge and high-energy LUMOs.
- Soft Lewis acids are characterized by large ionic radii, low positive charge and low-energy LUMOs.
- Hard Lewis bases are characterized by small ionic radii, high electronegativity and high-energy HOMOs.
- Soft Lewis bases are characterized by large ionic radii, intermediate electronegativity and low-energy HOMOs.

While the HSAB principle is qualitative, DFT calculations found a theoretical basis from the chemical potential to quantify the hardness and softness of a system [117]. The equation for hardness is given in 2.10.

$$\eta = \frac{\mathrm{IP} - \mathrm{EA}}{2} = \frac{\mathrm{LUMO} - \mathrm{HOMO}}{2} \tag{2.10}$$

Where IP is the ionisation potential and EA is the electron affinity. Note that by utilizing Koopmans' theorem once again, the IP and EA can be translated to the LUMO and HOMO energies.

The softness is simply the reciprocal of the hardness: $\sigma = \frac{1}{\eta}$

Natural bond orbital

Natural bond orbitals (NBOs) are localized charge orbitals on a specific atom and represent how much of an atom orbital is part of the bond [118]. A natural population analysis (NPA) is done to obtain NBOs. From the NPA, AOs are transformed into natural atomic orbitals (NAOs). NAOs list which atoms an NAO is attached to, the orbital type, the orbital occupancy and the orbital energy. Natural hybrid orbitals (NHOs) are combined then with NAOs to obtain NBOs. This is done by examining all possible interactions between filled (donor) Lewis-type NBOs and empty (acceptor) non-Lewis-type NBOs and estimating their importance through 2nd-order perturbation theory. Finally, a model is obtained where interactions of orbitals are quantified.

Mulliken population analysis

Mulliken populations characterize the electronic charge distribution of a molecule, as well as the bonding, antibonding or nonbonding character of MOs for pairs of atoms. Mulliken analysed the probability density of the charge distribution of two normalized AOs, see Equation (2.11).

$$1 = c_{ij}^2 + c_{ik}^2 + 2c_{ij}c_{ik}S_{jk} \tag{2.11}$$

Where ij and ik are the respective electrons, c_{ij} is the first AO, c_{ik} is the second AO, and S_{jk} is the overlap integral of the two AOs.

Here, both electrons contribute a square term, c_{ij}^2 and $c_{ik}2$, to the electronic charge. The overlap integral is called the overlap population and is > 0 for a bonding MO, < 0 for an anti-bonding MO and exactly 0 for a non-bonding MO. For every MO these values can be transformed into matrix form to obtain the population matrix.

2.5.4. Root mean-square deviation

The Root Mean Square Deviation of atomic positions (RMSD) is a metric to quantify how much geometries of the same complex differ from each other. RMSD is widely used to compare enzymes [119]. The RMSD is calculated by the minimum distance between the positions of the same atoms' Cartesian coordinates, see Equation (2.12):

$$RMSD(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((A_{ix} - B_{ix})^2 + (A_{iy} - B_{iy})^2 + (A_{iz} - B_{iz})^2)} \tag{2.12}$$

In this equation, A is complex 1, B is complex 2, N is the number of atoms and xyz refers to the Cartesian coordinates in their respective dimension.

hRMSD

One variation on the standard RMSD is heavy-atom RMSD (hRMSD), where all atoms except hydrogen are considered. It has been used in previous literature for comparing TM complexes [79]. Hydrogen atoms "corrupt" the RMSD by having wildly different positions compared to the central atoms. In this way, the chemical structure (and thus activity) does not differ much, but RMSD values are relatively quite high. Omitting them can be justified in this manner.

In this thesis specifically, hRMSDs were calculated with a Python script by J.C. Kromann that uses the Kabsch algorithm to rotate and align complexes and calculate their respective minimum hRMSD [120, 121]. No reordering or reflection methods were included, nor were any molecules ignored.

The script for complex 1 and complex 2 works as follows:

1. Recenters both molecules to the centroid (mean position of all coordinates in the molecule

2. An optimal rotation matrix which minimizes the RMSD, is calculated through the Kabsch algorithm [121]

3. The second molecule is rotated to the centre on the first molecule coordinates and the true minimal RMSD is calculated

3 Methods

In this chapter, the methods and workflow are described to calculate descriptors of the TM complexes. In Section 3.1, supercomputers that were used in this research are showcased and credited. In Section 3.2, the in-house workflow that generates complexes and calculates descriptors is outlined. In Section 3.3, the complexes and their method of generation are presented.

3.1. High performance computing

For the work in this thesis, three different supercomputers were used for running the in-house workflow, described in Section 3.2. These high-performance computers (HPCs) were:

- DelftBlue - Technical University Delft's supercomputer [122]
- Tetralith - National Supercomputer Centre at Linköping University [123]
- Snellius - Dutch National Supercomputer [124]

3.2. OBeLiX

Open Bidentate Ligand eXploration tool (OBeLiX) is the fully Python-based workflow used to generate non-biased 3-D representations of TM complexes, which as discussed in Chapter 1, is necessary for proper exploration of the chemical space. OBeLiX has been developed in parallel to the research done in this thesis with as goal automated *in silico* high-throughput TM catalyst screening. Furthermore, OBeLiX is open-source and can be found on Github: https://github.com/EPiCs-group/obelix. After generating 3-D TM complexes, the OBeLiX workflow automatically calculates descriptors on these structures. OBeLiX consists of a combination of multiple in-house and open-source tools of which the relevant tools and their function will be discussed below. Note that the module of ChemSpaX, which automatically substitutes (bidentate) ligands in TM complexes, was not used in this research and therefore will not be discussed, but is a core part of the OBeLiX workflow [125].

3.2.1. MACE

MetAl Complexes Embedding (MACE) is an open-source library for automated screening and discovery of metal complexes. It is able to discover all possible configurations for square-planar and octahedral TM-based complexes. MACE utilizes RDKit Uniform Force Field Molecular Mechanics (UFF-MM) to optimize geometries so that 3D structures are generated that are suitable for further computation. MACE as a standalone package is currently freely available on Github: https://github.com/EPiCs-group/obelix.

MACE carries out an exhaustive conformational search, where the lowest energy conformer is then chosen so that only a single structure retains. However, next to conformers, MACE also creates isomers, which is done in two ways: 1. By putting the substituents on different parts of the ligand 2. By putting the ligands on different parts of the metal centre. In an octahedral TM complex, ligands can then be placed either axially or equatorially leading to many different isomers. However, in a simpler complex, atoms can also be placed in different places on the ligand itself, creating isomers of the ligand. When placing the ligand subsequently on the metal centre, isomers of the complex are generated.

3.2.2. GFN family

The GFN (Geometry Frequency Non-covalent) family consist of methods developed by Grimme et al. [79, 126]. In this research, from the GFN family, the GFN-FF (Force Field) and GFN2-xTB methods were used for geometry optimization [47, 81].

GFN-FF is the non-electronic generic force field-based GFN approach. It is the computationally most efficient member of the GFN family and is useful for (metallo-) proteins, supramolecular assemblies and metal-organic frameworks. For the MM calculations, the GFN-FF package version 1.0.1 was used.

xTB is a QC semi-empirical based method that extends the original density functional tight binding model and is primarily useful for fast calculation for molecular systems (including metals) up to 1000 atoms [79]. GFN2 is the novel approach of this semi-empirical method which uses Grimme's D4 dispersion model [82]. For the semi-empirical calculations, the GFN2-xTB (Geometry Frequency Non-covalent eXtended Tight Binding) package version 6.5.1 was used.

For the metal centre atoms, Rh (I) and Ir (III) at their ground state were chosen. All calculations were performed in the gas phase to create the simplest model with the least interactions possible. No solvent was present either. However, omitting solvent was a big simplification. For other results, which did include solvent, see Appendix B.

3.2.3. Density Functional Theory

DFT calculations were performed using Gaussian 16 C.01 suite [127]. In this research, DFT was the "golden" standard for benchmarking geometries and descriptors. DFT use has been ubiquitous in both the ISE group and literature, therefore served as a reasonable choice for high-level benchmark [125, 128–134] GD3BJ was used for dispersion correction [135]. Natural bond order NBO calculations were included to obtain phosphorus lone pair NBO descriptors. The basis set of choice was def2-SVPP which contained both the standard basis set of def2-SVP and an additional set of polarization functions [64]. def2-SVPP was less accurate than def2-TZVP but was computationally cheaper. Furthermore, the PBE1PBE (in literature also known as PBE0) hybrid exchange-correlation functional was used [136–139]. PBE0 has been proven in the ISE group to be adequate for TM complexes and was in good agreement with experimental data [140]. An ultrafine grid was used for increased accuracy against small additional computational costs. All geometry optimizations and single-point calculations were done in the gas phase. The combination of this functional and basis set gave good geometries with modest computational costs for organometallic complexes [141, 142]. A (nigh) perfect correlation between Gibbs free energy and electronic energy has been found previously in the ISE group [140]. As such, Gibbs free energy could be estimated through electronic energy, thus for screening purposes the Hessian matrix calculations were omitted.

3.2.4. Morfeus

Morfeus (MOleculaR FEatUreS for machine learning) is a tool developed by K. Jorner at AstraZeneca UK in collaboration with the Aspuru-Guzik, Sigman and Gensch groups [143] for calculating descriptors. It had been used before to study monodentate phosphorus ligands and calculate descriptors on those complexes [130]. The research of Gensch et al. served as inspiration for the research on bidentate phosphorus complexes of this thesis. Morfeus was therefore ingrained in the OBeLiX workflow to calculate steric and electronic GFN2-xTB descriptors.

Morfeus calculated descriptors through single-point GFN2-xTB calculations and subsequently extracted descriptors. Some descriptors contained slight adaptations from the definition in Section 4.2.2. One such descriptor was the SASA. Morfeus calculated the SASA based on the CRC Handbook of Chemistry and Physics van der Waals radii and atomic positions as well as a modified version of the Shrake and Rupley algorithm [144, 145]. Another adaptation was regarding the dispersion descriptor. A modified version of dispersion, which used electron densities instead of an electron density isosurface, was used to save computational costs. A third adaption to Morfeus originated from the OBeLiX workflow. Regarding the $\%V_{\mathrm{bur}}$, in literature as well as in the default setting of Morfeus, the radius of the buried volume sphere was 3.5 Å. In OBeLiX, however, the radius had been increased iteratively to find an optimum value, similar to the approach of Gensch et al. [130].

3.3. Structure generation & descriptor calculation

As mentioned in Chapter 1, the ISE group cooperated with an industrial partner to research Ir and Rh organometallic complexes. For the research done in this thesis, the industrial partner presented an (experimental) set of ligands and substrates to combine with either the Rh or Ir metal centres. These ligands served as the basis for computational modelling and descriptor calculation.

Actual molecular geometries (.xyz files) of the ligands have not been issued publicly yet, however, will be able to be found on the OBeLiX Github: https://github.com/EPiCs-group/obelix.

For the research done in this thesis, MACE was used to generate Rh and Ir-based structures from 2D Simplified Molecular-Input Line-Entry System (SMILES) [146].

Investigation of geometry optimization level

To investigate objective 1, structures needed to be generated. As such, 186 complexes containing only Rh (I) and the bidentate ligand were generated through MACE. These complexes are hereby referred to as *pristine* complexes. As mentioned above, some of the generated pristine complexes were isomers of each other, as such only 99 complexes were unique structures. In addition, 1 complex wrongly contained monodentate ligands. Therefore, 98 unique pristine complexes possessed a bidentate ligand. From these 98 complexes, the majority (88) were P,P bidentate ligands, whilst the rest were P,N bidentate ligands.

The next step in investigating objective 1, was to look at a single descriptor on four levels of geometry optimization: MACE (UFF MM), GFN-FF, GFN2-xTB and DFT, in order of predicted accuracy. Subsequently, geometry optimization was executed on MACE complexes for GFN-FF, GFN2-xTB and DFT, as can be seen in Figure 3.1. Since MACE had already been automatically optimized with UFF MM, the optimization level is referred to as MACE. DFT was chosen as the benchmark geometry since it possessed the highest accuracy and highest computational cost.

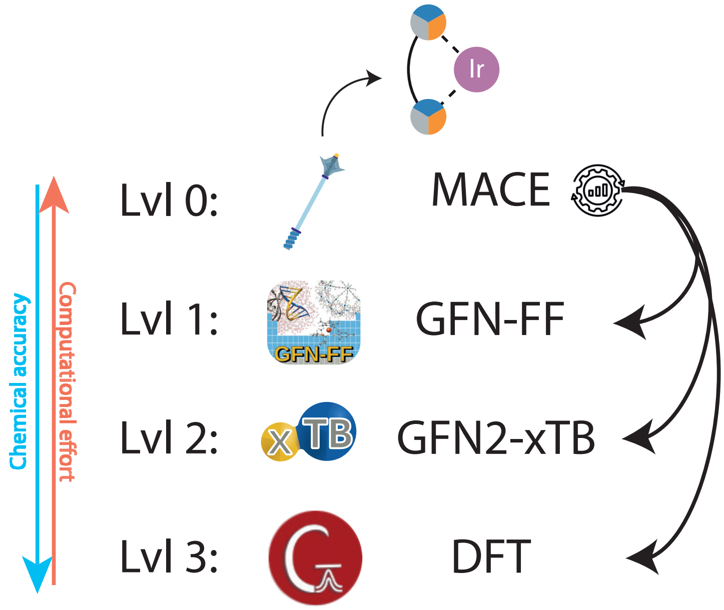


Figure 3.1: Illustration of the workflow to generate optimized structures for different levels of theory. The levels refer to the level of theory of the geometry optimization methods. Higher levels have increased accuracy but a higher computational cost. Pristine complexes consisting of an Ir metal centre and a bidentate ligand, are generated through MACE. Geometry optimization is subsequently done on the MACE-generated structures for every level of theory.

Next, descriptors were calculated through the *descriptor_calculator.py* script of the OBeLiX workflow. This script utilized Morfeus, which performed xTB single-point calculations to extract descriptors. To test the results, the HOMO-LUMO gap was used as a preliminary descriptor, since it had been used earlier in the ISE group [125, 147].

However, extracting descriptors through Morfeus contained one major flaw: all descriptors are extracted from xTB single-point calculations. For every level theory up until GFN2-xTB this could be permitted since xTB was considered a higher level of theory. However, for DFT this method was scientifically unsound, due to DFT being theoretically more accurate in representing structures than xTB. This effect was therefore investigated by extracting descriptors from single-point DFT calculations on both xTB and DFT-optimized geometries.

In the case of geometric and steric descriptors, it was deduced that comparing those would lead to little gain. First of all, for these descriptors correlations between different levels of theory were quite consistent. Second of all, the calculation of angles was independent of the level of theory, i.e. only dependent on geometry. However, for electronic descriptors, this logic did not hold. Especially for the dipole moment and the HOMO-LUMO gap which had low correlations, a multilevel comparison was deemed appropriate. Ultimately 3 electronic descriptors were correlated against each other for xTB and DFT level descriptors and geometries.

Investigation of structure representation

For objective 2, two sets of complexes were generated via MACE, which were an Ir (III)-based octahedral (OH) and an Ir (III)-based pristine set that served as model structures for OH complexes. The bidentate ligands used in the Ir (III) set were identical to the ligands used in the Rh (I) set. The model structure is shown in Figure 3.2 on the left. In this schematic, the three colours refer to a combination of atoms contained in the ligands. The bidentate ligands consisted either of the P-P or P-N class and the C groups were the atoms attached to the chelating ligand. Note that a pristine bidentate complex could have never existed in practice and only served as a model structure to calculate descriptors on. The OH structure is shown on the right in Figure 3.2. Note the addition of the hydrides and the acetonitrile.

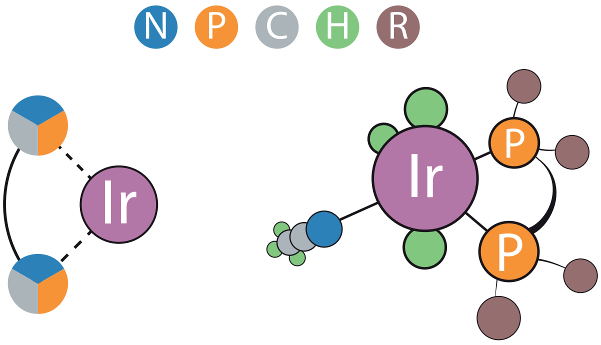


Figure 3.2: Schematic overview of the complexes used to calculate descriptors on. On the left, the bidentate pristine Iridium complex in 2D with the circle representing different kinds of ligands is depicted. On the right, the octahedral Iridium complex in 3D is depicted. Color codes: purple = TM centre, blue = nitrogen, orange = phosphorus, grey = carbon, green = hydrogen, and brown = rest group.

For the pristine set, 217 complexes were generated, of which 111 were isomers of some form. For the OH set, 435 complexes were generated of which 360 were isomers. Note that the mismatch in the number of unique complexes (106 vs 75) stemmed from converging issues during MACE. To match the two datasets, a single isomer of every complex was needed, due to the mismatch of isomers present. As such, the lowest energy isomer was chosen in both sets. Note, however, that this isomer does not necessarily need to be the same isomer across both sets.

Subsequently, DFT geometry optimization was done for all unique complexes.

To find correlations between the two sets, the datasets needed to be merged. Note that 1. Some complexes in one set could have failed during MACE generation but could have been successfully generated in the other set. 2. Some complexes could have failed in one set during the geometry optimization but could have remained successful in the other set, therefore the sets did not contain the exact same structures. Descriptors were then calculated through an updated version of the *descriptor_calculator.py* script of the OBeLiX workflow, which extracted both xTB and DFT level descriptors. Again, during xTB single-point calculation some complexes could have failed in one set but could have ran successfully in the other set. As such, during the merging of both datasets, some complexes were excluded and ultimately, only 57 complexes, of which 5 P-N ligands, were left to correlate descriptors with.

4
Results & Discussion

In this chapter, the main results of this thesis are presented and discussed. All descriptors in this chapter are either made bold or made abbreviated for comprehensibility. In Section 4.1, MACE, GFN-FF, and GFN2-xTB are correlated to a benchmark level of theory (DFT), for a single electronic descriptor (**HOMO-LUMO gap**). In Section 4.2, the same computational methods are correlated to DFT for 22 additional descriptors, which are subdivided into the following categories: geometric, steric and electronic. In Section 4.3, xTB and DFT geometries are compared to each other, through hRMSDs. In Section 4.4, geometries obtained from xTB and DFT geometry optimization are compared to descriptors obtained from xTB and DFT single-point calculations. In Section 4.5, descriptors obtained from xTB and DFT single-point calculations are correlated to each other for simplified model structures and octahedral complexes.

4.1. Level of theory: Simple regression analysis

Simple correlation plots for the **HOMO-LUMO gap** can be seen in Figure 4.1. In this figure, hardly any correlation between MACE (UFF) and DFT geometry-optimized structures is observed ($R^2 = 0.10$ and $RMSE = 0.028$ a.u.). R^2 in this context is determined as the coefficient of determination.

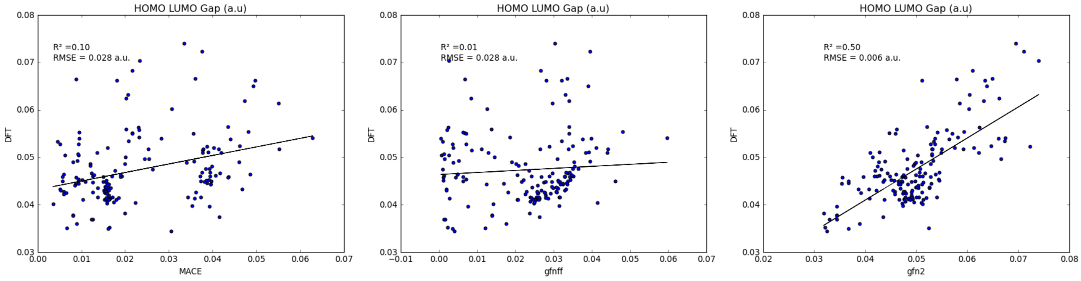


Figure 4.1: HOMO-LUMO gap correlations (R^2) and spread (RMSE) for different levels of theory. Scatter plots showcasing the correlation of MACE versus DFT (left), GFN-FF versus DFT (middle) and GFN2-xTB versus DFT (right).

Correlation is observed to be completely absent for GFN-FF ($R^2 = 0.01$ and $RMSE = 0.028$ a.u.). The RMSE indicates that there is a high level of spread in both methods as well. Since both methods use force field optimization, it is theorized that the results are similar due to force field optimization.

In the case of xTB versus DFT correlation is observed, but is not conclusive by any means ($R^2 = 0.50$ and $RMSE = 0.006$ a.u.). However, since the spread is significantly smaller, outlier analysis is applicable. Manual inspection of structures with a high standard deviation from the set did not show consistency among chemical classes. Some ligands of the complexes with high standard deviation contained large steric hindered ligands, others very small ligands. Similarly, some ligands contained polar atoms, e.g. oxygen and nitrogen, while other ligands were completely apolar, e.g. due to benzene rings. Another method of extracting outliers is the interquartile range (IQR), see Appendix A.2 for a detailed description of the IQR method. Inspecting the complexes marked as outliers through IQR also did not show any consistency among the chemical classes of the ligands. Therefore, no discernible trend was found among outlier ligands.

4.2. Correlation matrix of descriptors

The same correlation plots as in the previous section have been done for an additional 22 descriptors, see Appendix A.1. Descriptors were split into three groups: geometric, steric and electronic, based on a similar method done by Sigman's group [130, 131]. The 23 descriptors were subdivided as follows: 4 are geometric: **bite angle**, **cone angle** and **Rh-Donor length** (from Rh to donor atom 1 and Rh to donor atom 2). 8 are steric: **SASA** and 7x **buried volume** with different radii and centring. 11 are electronic descriptors: **dispersion** on Rh and donors, **dipole**, **ionisation potential** (IP), **electron affinity** (EA), **electrofugality** (EF), **nucleofugality** (NF), **nucleophilicity** (NP), **electrophilicity** (EP), and **HOMO-LUMO gap**.

Now with increased dimensionality originating from increased descriptor count, a heatmap is more suitable to visualize the correlations, see Figure 4.2. This heatmap will be discussed in the coming sections. Every value corresponds to a correlation between a descriptor and optimization level, and the lighter the colour, the better the correlation. Note however, that r^2 here refers to the Pearson correlation coefficient squared, whereas in Section 4.1, R^2 referred to the coefficient of determination. While sometimes r^2 and R^2 correspond, the coefficient of determination (R^2), and the Pearson correlation coefficient squared, r^2, are not equal [148, 149].

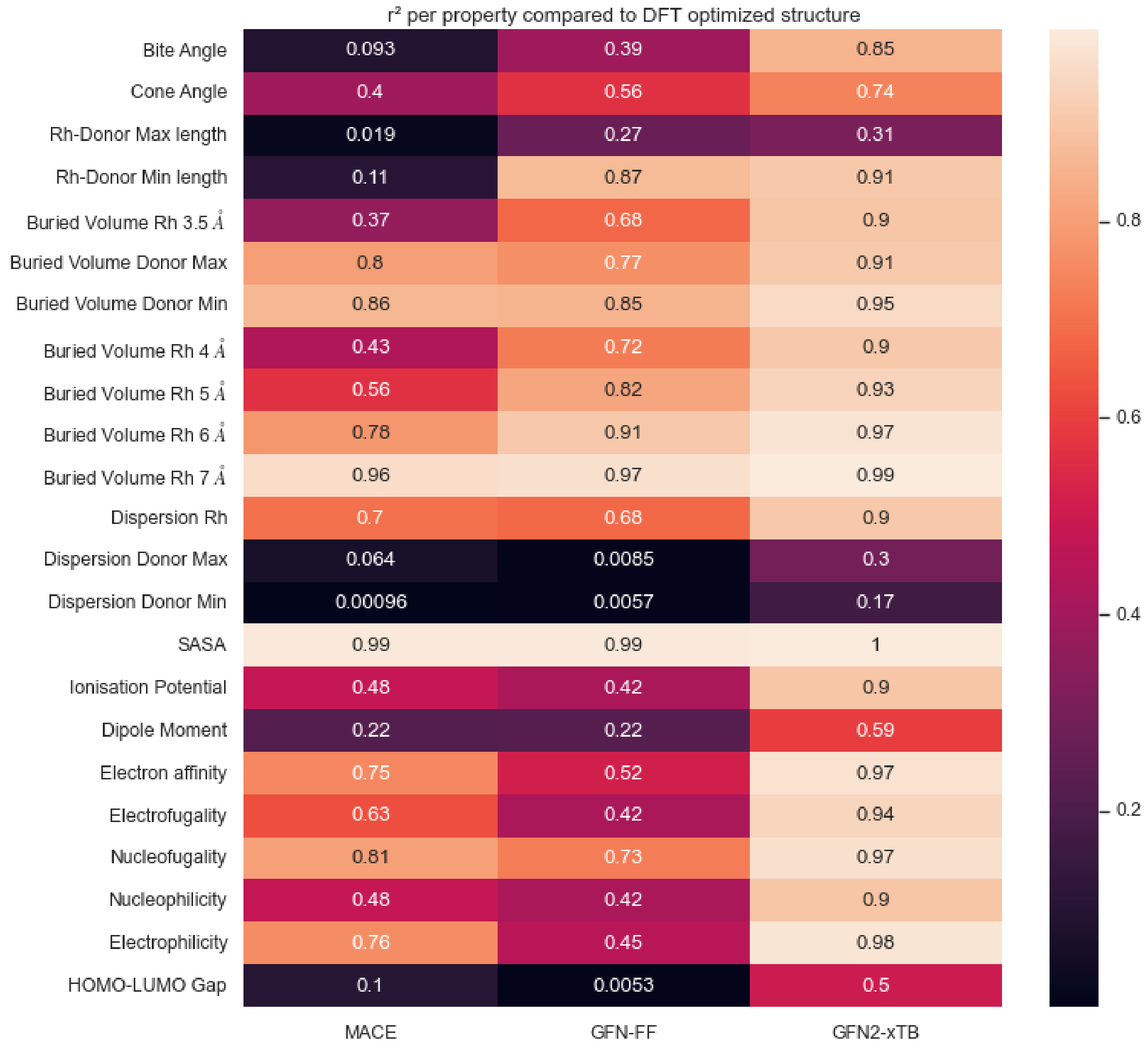


Figure 4.2: Correlation matrix for all calculated descriptors on MACE, GFN-FF and GFN2-xTB level of theory against DFT. Values represent Pearson correlation coefficients squared. Lighter boxes signify a higher correlation.

4.2.1. Geometric descriptors

To begin with, in Figure 4.2, a very low correlation is observed for geometric descriptors calculated on MACE structures. **Rh-Donor max length** and **Rh-Donor min length**, descriptors which were included to signify bond strength, contain no correlation compared to DFT structures. Similarly, for the **bite angle** no correlation was observed either. While there is a certain degree of correlation with the **cone angle** ($r^2 = 0.40$), it does not provide conclusive evidence.

MACE (i.e. UFF MM) is relatively crude in finding geometries on the PES, due to being the lowest level of theory. Based on these geometric descriptors, the initial impression is that geometries found by MACE are not representative of the chosen benchmark of DFT. Further analysis in comparing geometries quantitatively is done in Section 4.3.

On the GFN-FF level structures, geometric descriptors show increased correlation. With the **bite angle** descriptor a certain degree of correlation is observed ($r^2 = 0.39$) and the **cone angle** shows slightly higher correlation than its MACE counterpart ($r^2 = 0.56$), though again not sufficient to be conclusive. **Rh-Donor max length** shows very low correlation, however, **Rh-Donor min length** abruptly shows excellent correlation ($r^2 = 0.87$). This result is peculiar and warrants extra inspection. Upon inspection, the distribution of **Rhodium-Donor min length** data points, see Appendix A.1, shows 10 big outliers clustered together at smaller bond lengths. The other data points are clustered 0.3 Å higher. Since this trend is both present on DFT and GFN-FF levels, the trendline goes through the middle of both clusters. The correlation of **Rhodium-Donor min length** therefore ends up being quite high. After the removal of the outliers, clusters are reduced in size and no correlation is present anymore ($r^2 = 0.27$).

Closer inspection of the outlier complexes on the GFN-FF level shows that all outliers in the smaller cluster contain P-N bidentate ligands. In this specific case, the minimum donor thus refers to a nitrogen atom, which possesses a smaller shell radius, hence the bond length can be smaller.

Inspecting the **Rhodium-Donor min length** descriptor for MACE structures interestingly shows no clustering, see Appendix A.1. In the case of MACE structures, the data distribution is much more evenly spread than with GFN-FF and therefore there are no clusters to be considered.

Two of the outlier complexes, L184 and L127, form a "tridentate" complex with a carbon atom being close enough to the metal centre to signify a bond, 2.11 & 2.08 Å, respectively. L184 connects to a carbon in an ether group and L127 connects to a carbon in a phenyl group. Lastly, two other complexes, L81 and L73 form a "quadridentate" complex with both having two hydrogen atoms close to the metal centre (1.77, 1.78; 1.80, 1.79 Å). L81 and L73 have the highest distance to the metal centre of the cluster of P-N bidentate complexes (0.02 Åhigher), which could be attributed to having formed this "quadridentate" complex.

In the case of GFN2-xTB structures, correlations of geometric descriptors are much higher than in the previous two methods. **Bite** and **cone angle** show enough correlation to assert that they are comparable to DFT structures. Scatterplots of **bite** and **cone angle** on xTB level versus DFT, which showcase the distribution of the data, can again be found in Appendix A.1. For the descriptor **Rh-Donor max length** correlation is observed to be low, similar to the other two levels of theory. It can be discerned that **Rh-Donor max length** is very sensitive to different geometries and therefore contains no correlation. **Rh-Donor min length** shows good correlation, same as with MACE and GFN-FF structures. Further inspection, shows that two clusters are formed once more, with the smaller cluster containing a nitrogen atom as the donor min.

4.2.2. Steric descriptors

Correlation for SASA is excellent for all levels of theory ($r^2 \geq 0.99$). It is known that the SASA depends solely on the size of the system, i.e. the number of electrons, and not on the atomic positions [150]. As such, values across different optimization levels should correlate well, which is validated by this result. However, with correlations being consistently close to 1, the question should be raised if the SASA is suitable as a descriptor for novel catalyst design. SASA will always have very good correlations and therefore it does not quantify the steric hindrance of the system, which is the primary objective of steric descriptors.

In the case of MACE structures, the correlation of $\%V_{bur}\ Rh$ is observed to be relatively low up until 5 Å. As can be discerned from Figure 4.2, for increasing radii of $\%V_{bur}$ correlation increases: $r^2 = 0.37$ for $\%V_{bur}$ on 3.5 Å, $r^2 = 0.43$ on 4 Å, $r^2 = 0.56$ on 5 Å, $r^2 = 0.78$ on 6 Å, and $r^2 = 0.96$ on 7 Å. Increasing the radius of the sphere by definition means that the sphere encompasses more of the complex. As such, if the radius goes to infinity the correlation would go to 1, since all atoms are found in the sphere. Logically, at 7 Å most of the complex is included in the sphere, therefore creating a good correlation.

One adaptation on the standard model of percent buried volume was to calculate the $\%V_{bur}$ centred on either donor atom instead of the metal centre, again with 3.5 Å. Correlations of $\%V_{bur}$ on the donor atoms are found to be quite good ($r^2 = 0.80 \& 0.86$), contrary to $\%V_{bur}$ on the metal centre with 3.5 Å ($r^2 = 0.37$). It is theorized that this is due to the centring of the sphere on the phosphorus or nitrogen atoms. In that case, the sphere does not need to be as large to fully encompass all molecules, because in the pristine model, the ligand is exactly in the middle of the complex, not the metal centre.

For GFN-FF level descriptors, a trend across correlations is observed, similar to MACE level descriptors. With larger radii of $\%V_{bur}$, correlations increase. However, the lowest radius (3.5 Å) of GFN-FF level $\%V_{bur}$ shows better correlation than MACE ($r^2 = 0.68$ versus $r^2 = 0.37$, respectively). For donor-centred $\%V_{bur}$ correlation is about equal, likely due to the reasons mentioned above. However, it is sensible to say that GFN-FF seems to describe steric hindrances better than MACE.

xTB is better than the previous two levels by quite a large margin. All correlations are quite good ($r^2 \geq 0.9$). A trend across correlations with increasing radii of $\%V_{bur}$ is present again, reinforcing the previously stated theory. At 7 Å xTB is almost as good as DFT in describing buried volume ($r^2 = 0.99$). It is assumed that this very high correlation is due to the previously stated theory as well as the fact that xTB structures are similar to DFT ones.

Correlations of xTB-level donor atom centred $\%V_{bur}$ ($r^2 = 0.91 \& 0.95$, for max and min respectively) are better than their MACE ($r^2 = 0.8 \& 0.86$) and GFN-FF ($r^2 = 0.77 \& 0.85$ counterparts. One possible explanation for higher xTB correlations is that xTB geometries are qualitatively more similar to DFT geometries than their respective MACE and GFN-FF counterparts. Geometry comparison will be discussed more in-depth in Section 4.3.

4.2.3. Electronic descriptors

Correlation for **dispersion Rh** is found to be quite good ($r^2 = 0.7$) for MACE-level structures. Correlations of **dispersion Donor max** and **dispersion Donor min** are observed to be completely absent, however. The primary reason for the absence of correlation is due to 8 data points/complexes having a value of 0 at either MACE or DFT level. Filtering out these data points gives marginally better correlations for **dispersion Donor max** and **dispersion Donor min** ($r^2 = 0.21$ & $r^2 = 0.18$, respectively), but these correlations are far too low to be conclusive.

In the case of GFN-FF level structures results are similar to MACE level structures. The correlation of **dispersion on Rh** is found to be quite good, but no correlation is found in the case of **dispersion Donor max** and **Dispersion Donor min**. Once more, outliers with a value of 0 decrease the correlation to being almost completely absent. Even after filtering these values out, no significant correlations are observed.

In the case of xTB level structures, the correlation for **dispersion Rh** is found to be excellent. Correlation for **dispersion Donor Max** and **dispersion Donor min** is once again low. By filtering the 0 points out, the correlation increases again for both **dispersion Donor max** and **dispersion Donor min** ($r^2 = 0.49$ & $r^2 = 0.52$, respectively). In line with expectations, correlation is best for dispersion on xTB-level structures.

The **dipole** moment shows a poor correlation for both MACE and GFN-FF. For xTB level structures correlation is present. By inspecting the scatter plots, trends are discernible, see Appendix A.1, but there is high variance in the data. It is theorized that this is due to the level of theory used to calculate descriptors, which is discussed more in detail in section 4.4.

For the HOMO-LUMO gap results and discussion, see the previous section, Section 4.1.

The descriptors **electrophilicity** (EP), **nucleophilicity** (NP), **electrofugality** (EF) and **nucleofugality** (NF) are all calculated directly from the **ionisation potential** (IP) and the **electron affinity** (EA). Note that IP and EA themselves are dependent on the HOMO and the LUMO levels, due to Koopmans' theorem. These 6 descriptors are very codependent on each other. To showcase this interdependence, the correlations between descriptors calculated on DFT structures are shown in Figure A.2. In that figure, correlations are calculated between descriptors and shown in a codependence matrix.

As can be seen in Figure 4.2, **nucleophilicity** and **ionisation potential** does not follow the same trend as the other 4 descriptors. In Morfeus' source code, NP is calculated as the negative of the IP, therefore IP and NP contain the exact same correlation. The other 3 descriptors are both dependent on the EA as well as the IP.

The correlation of IP in MACE structures is mediocre and by virtue of it, so is the correlation of NP. Surprisingly, for EA, EF, NF, and EP an increased correlation is found, perhaps due to the EA dependency term. For GFN-FF structures correlation of all 6 electronic descriptors is slightly lower. On xTB level structures, correlations for all 6 electronic descriptors are excellent ($r^2 \geq 0.9$), alluding to the fact that xTB describes electronic structure better than MACE and GFN-FF.

Interestingly, IP correlation is the lowest for all 3 levels, while NF correlation is among the highest. It is assumed that the discrepancy in the correlation of EA, EF, NF and EP is simply due to standard deviation. From statistics, it follows that if two variables are strongly positively correlated, the resulting standard deviation will be larger than simply the product of both standard deviations. Both variables tend to vary in the same direction, which increases the variability of the product. Therefore, higher values of EF compared to IP are theorized to be simply statistical variations. It should be noted that the magnitude of the variation is not considerable and that a strong correlation among all the aforementioned electronic descriptors exists at the xTB level.

Filtration methods

Lastly, the IQR method has also been applied to this dataset to improve correlation, however, the correlations did not change in a meaningful manner. Descriptors with low correlation continued to have low correlation and descriptors with good correlations continued to have good correlation. Though this method has been used successfully before on a previous dataset, see Figure B.1, it showcases that data filtration methods are not always applicable, but instead are system specific. This result highlights the pitfalls of data massaging. Filtration methods should be applied to data only if need be. Next to that, if filtration methods are applied, then it is crucial to pick an appropriate filtration technique for the dataset.

4.3. hRSMD

One problem with the previous approach of comparing levels of theory is that all descriptors are calculated on GFN2-xTB level, regardless of the level of geometry optimization. RMSD has been used before to measure structures between different levels of theory [47, 79, 151]. In literature, it has been accepted that hRMSD values < 0.6 Å signify that structures are more or less equal [152].

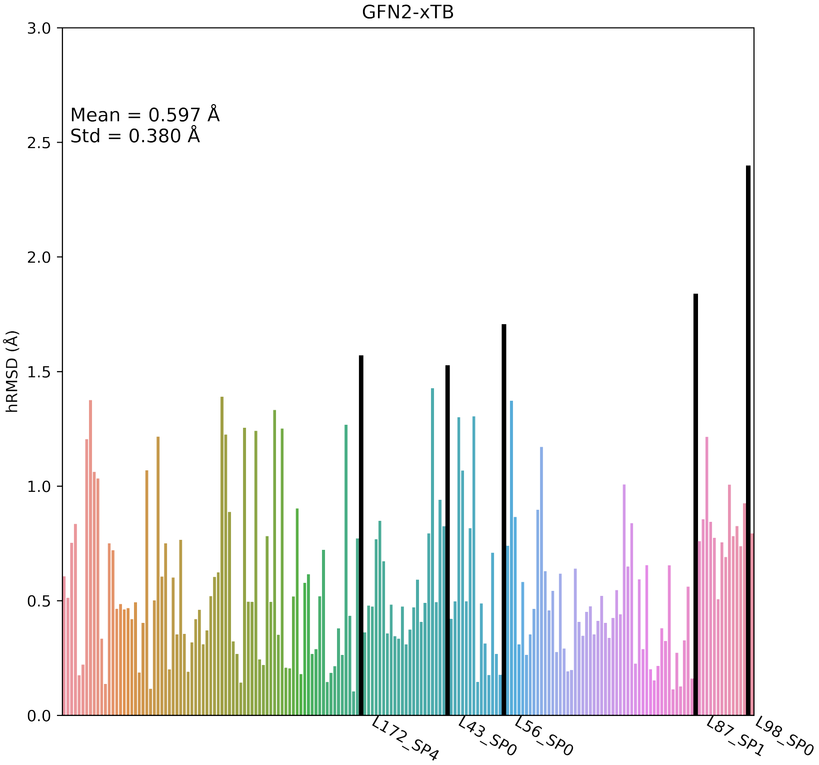


Figure 4.3: hRMSDs (versus DFT) of all GFN2-xTB complexes. Bars dyed black are complexes classified as outliers by the IQR filtering method. Std refers to the standard deviation of the set.

As can be seen in Figure 4.3, the mean hRMSD is quite low (0.597 ± 0.380 Å), thus xTB geometry optimization is qualitatively quite good compared to DFT. Furthermore, since the mean is lower than 0.6 Å, on average xTB geometry-optimized structures more or less coincide with DFT geometry-optimized structures. Similar hRMSD figures of MACE and GFN-FF geometries versus DFT geometries have also been generated and can be found in Figure A.3. The mean hRMSD of xTB level structures is 0.4 Å lower than MACE (0.884 ± 0.433

Å) and GFN-FF (0.925 ± 0.655 Å). The spread of the xTB dataset is also lower than both MACE and GFN-FF, suggesting that the xTB level of theory is not only better on average but is also the most consistent method of the three geometry optimization methods considered in this thesis. Interestingly, the mean of the MACE dataset is slightly lower than the mean of the GFN-FF dataset (0.884 versus 0.925), alluding to GFN-FF being the worst method for geometry optimization. Possibly this is due to TM elements not being quite as well parameterized as is the case with UFF. The spread of the MACE dataset is also lower than the GFN-FF dataset (0.433 versus 0.655), making GFN-FF also the least consistent method in finding good geometries.

Outlier hRMSDs are denoted in black. Outlier detection was done based on the IQR method, see Appendix A.5. Subsequently, outlier structures were visualized and compared using structure overlay plots, see Figure 4.4. Inspecting the outlier structures does not show similarities among ligands. Most complexes contain a single phenyl group and are at least in some way polar, however, this is the case with most ligands in the dataset, due to their steric nature.

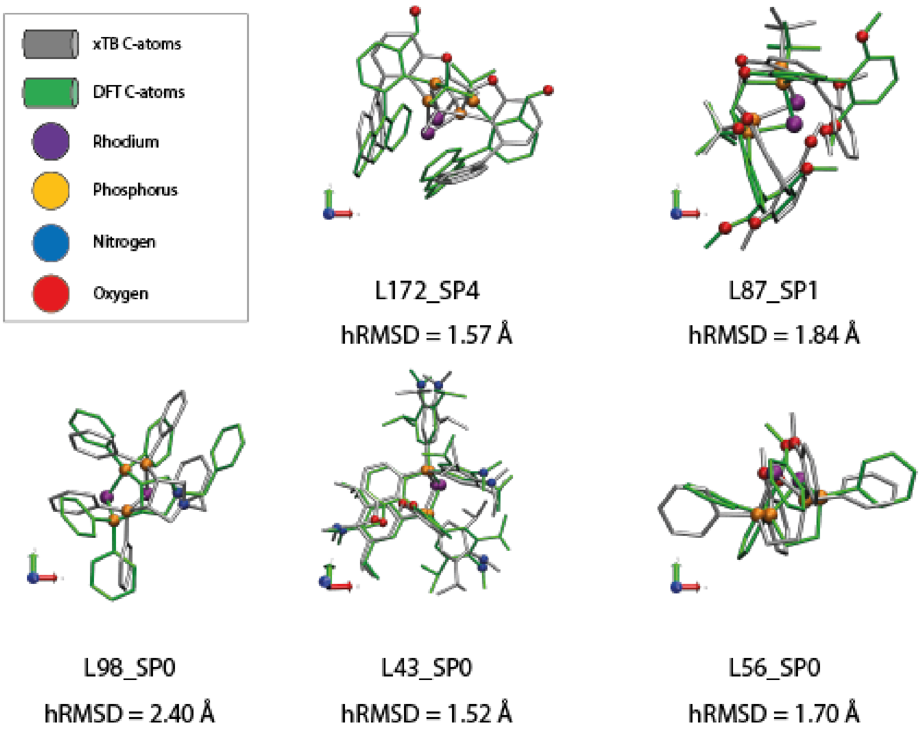


Figure 4.4: Structure overlay plot of all complexes classified as outliers by IQR method. GFN2-xTB geometry optimized structure is overlaid on DFT geometry optimized structure. xTB atoms are denoted in silver, whilst DFT atoms are denoted in green. Rhodium (purple), phosphorus (orange), nitrogen (blue), and oxygen (red) are the same colour for both levels.

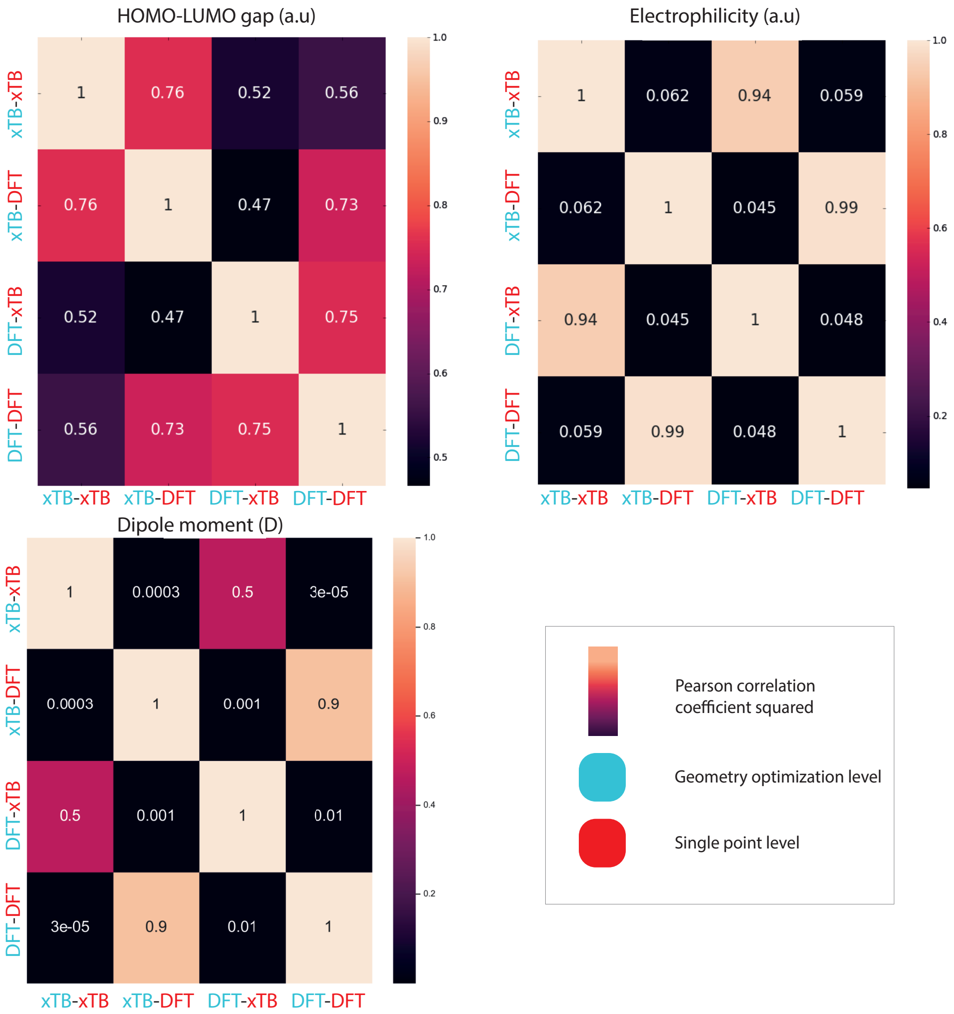
L98 has the highest hRMSD by quite a substantial margin, as seen in Figure 4.3 as the first bar from the right. The cause of the inflated hRMSD can be attributed to the fact that the xTB structure forms a haptic binding site between Rh and a phenyl group which is bound to the nitrogen. DFT on the other hand does not form this haptic binding site, as is confirmed by the structure of L98 which can be seen in Figure 4.4 on the bottom left. All DFT atoms, even the Rh centre, are in a significantly different Cartesian position than their xTB counterparts. Interestingly, this formation of a haptic binding site also occurs for L172, but is present on both levels of theory, as such hRMSD is much lower. It can be inferred, however, that the creation of haptic binding sites, therefore folding the ligand over, creates larger hRMSDs. This effect would originate from a larger degree of freedom of the ligands in pristine complexes, creating much more possibilities for local minima on the PES. It is then possible that DFT finds a lower local minimum than xTB.

4.4. Multiple levels of theory comparison

For the comparison between descriptors on multiple levels of theory, single-point xTB and DFT calculations were done on both xTB and DFT optimized structures leading to 4 combinations:

- Structures generated from xTB geometry optimization and descriptors obtained from xTB single-point calculations
- Structures generated from xTB geometry optimization and descriptors obtained from DFT single-point calculations
- Structures generated from DFT geometry optimization and descriptors obtained from xTB single-point calculations
- Structures generated from DFT geometry optimization and descriptors obtained from DFT single-point calculations

In Figure 4.5, the first term (e.g. xTB) refers to the level of geometry optimization, while the second term (e.g. DFT) refers to the level of single-point calculation in which the descriptors are calculated. For example, the combination of xTB-DFT would refer to DFT single-point calculations on DFT geometry-optimized structures.



HOMO-LUMO gap (a.u)

	xTB-xTB	xTB-DFT	DFT-xTB	DFT-DFT
xTB-xTB	1	0.76	0.52	0.56
xTB-DFT	0.76	1	0.47	0.73
DFT-xTB	0.52	0.47	1	0.75
DFT-DFT	0.56	0.73	0.75	1

Electrophilicity (a.u)

	xTB-xTB	xTB-DFT	DFT-xTB	DFT-DFT
xTB-xTB	1	0.062	0.94	0.059
xTB-DFT	0.062	1	0.045	0.99
DFT-xTB	0.94	0.045	1	0.048
DFT-DFT	0.059	0.99	0.048	1

Dipole moment (D)

	xTB-xTB	xTB-DFT	DFT-xTB	DFT-DFT
xTB-xTB	1	0.0003	0.5	3e-05
xTB-DFT	0.0003	1	0.001	0.9
DFT-xTB	0.5	0.001	1	0.01
DFT-DFT	3e-05	0.9	0.01	1

1.0
0.9
0.8
0.7
0.6
0.5
1.0
0.8
0.6
0.4
0.2
1.0
0.8
0.6
0.4
0.2
Pearson correlation coefficient squared
Geometry optimization level
Single point level

Figure 4.5: Descriptors calculated on xTB & DFT level for three electronic descriptors: **electrophilicity**, **HOMO-LUMO gap**, and **dipole moment**. The first term (blue) is the geometry optimization level and the second term (red) is the descriptor extraction level, i.e. the single-point calculation level. Values represent Pearson correlation coefficients squared. Lighter boxes represent higher correlations.

HOMO-LUMO gap

The most interesting investigation for objective 1, is to compare all geometry optimizations (marked with red labels) and single-point descriptor calculation (marked with blue labels) combinations to the benchmark level of DFT-DFT. As can be seen in Figure 4.5 on the top left, the correlation of the **HOMO-LUMO gap** exists for xTB-xTB versus DFT-DFT ($r^2 = 0.56$) but is very not significant. This is in line with expectations since xTB-xTB has the least overlap with the combination of DFT-DFT due to having neither geometry optimization nor single-point calculations on the same level of theory. In the case of xTB-DFT versus DFT-DFT, the correlation is substantially better ($r^2 = 0.73$). As seen from Section 4.3, xTB structures are similar to DFT structures for most complexes. It is expected that doing single-point calculations for descriptors would lead to similar values as well. In the case of DFT-xTB versus DFT-DFT, the correlation is similar ($r^2 = 0.75$), showcasing that an xTB single-point calculation is quite decent at calculating the **HOMO-LUMO gap**.

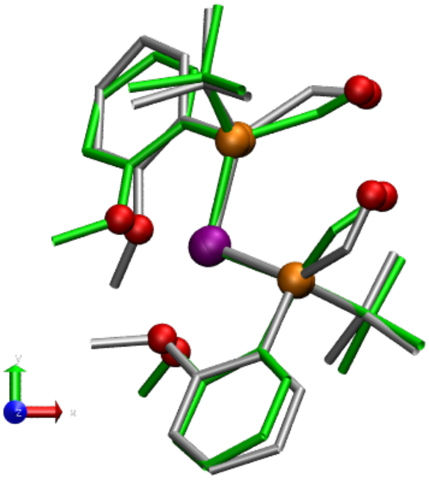


Figure 4.6: Structure overlay plot of L138_SP4. xTB atoms are denoted in silver, whilst DFT atoms are denoted in green. Rhodium (purple), phosphorus (orange), nitrogen (blue), and oxygen (red) are the same colour for both levels.

In the case of xTB-xTB versus DFT-xTB, a low correlation ($r^2 = 0.52$) is found, showcasing that xTB single-point calculations are substantially dependent on the geometry, contrary to DFT single-point calculations. For xtB-xTB versus xTB-DFT, a good correlation ($r^2 = 0.76$) is observed, which can possibly be attributed to the fact that DFT single-point calculations are more accurate, regardless of structure. In the case of xTB-DFT versus DFT-xTB, the lowest correlation is found ($r^2 = 0.47$), which can be attributed to the fact that neither calculation is done on the same level. Assuming there is an error margin while comparing levels, this would propagate most into DFT-xTB and xTB-DFT.

An in-depth inspection of the data shows one large outlier, L138 on both xTB and DFT levels. This anomaly can be explained since L138 has two oxygen atoms on the edges of the ligand, see Figure 4.6. These oxygen atoms have a lone pair and can coordinate to the metal centre, creating a square planar structure, which has different geometry and electronic structure. Removing this ligand from the dataset does not substantially alter the correlation, however.

Electrophilicity

Electrophilicity is inherently linked to the **HOMO-LUMO gap** by being dependent on the HOMO and the LUMO level. As can be seen in Figure 4.5 on the top right, correlation for **electrophilicity** is absent for xTB-xTB versus DFT-DFT ($r^2 = 0.06$). In the case of DFT-xTB versus DFT-DFT, the correlation is equally bad ($r^2 = 0.05$). For xTB-DFT versus DFT-DFT however, the correlation is excellent ($r^2 = 0.99$). This can be attributed to the fact that both combinations utilize single-point DFT calculations to extract descriptors. Even if the level of geometry optimization is slightly different, doing the same DFT single-point calculations gives very similar results. This result is also validated by xTB single-point calculations. In the case of xTB single-point calculation, xTB-xTB correlates excellently with DFT-xTB ($r^2 = 0.94$). As such, it is more important to have the same level of theory for descriptor extraction than it is for geometry optimization. This is confirmed by xTB-DFT versus DFT-xTB and xTB-xTB versus xTB-DFT (($r^2 = 0.05$ & $r^2 = 0.06$) where correlation are absent for both.

Accordingly, it can be discerned that if the single-point calculations' level of theory differs, the **electrophilicity** values will be dissimilar, even if their geometries are identical.

Dipole

As can be seen in Figure 4.5 on the bottom left, dipole moment correlation supports the findings of 4.4. Correlations are absent for combinations where the level of theory for single-point calculations differ. Only xTB-DFT versus DFT-DFT contains great correlation ($r^2 = 0.90$), again due to both doing DFT single-point calculations to extract descriptors.

In essence, this means single-point calculations are not comparable across the two levels.

4.5. Molecular geometry comparison

In Figure 4.7, a scatter plot is shown for the correlation between all octahedral complex and pristine complex descriptors.

4.5.1. Geometric xTB descriptors

Starting on the left with quadrant and octant $\%V_{\text{bur}}$, no correlation is found between descriptors. This can be attributed to the fact that steric hindrance does not apply in the pristine model. Ligands have unrestricted movement, therefore it is illogical to divide the $\%V_{\text{bur}}$ into sections.

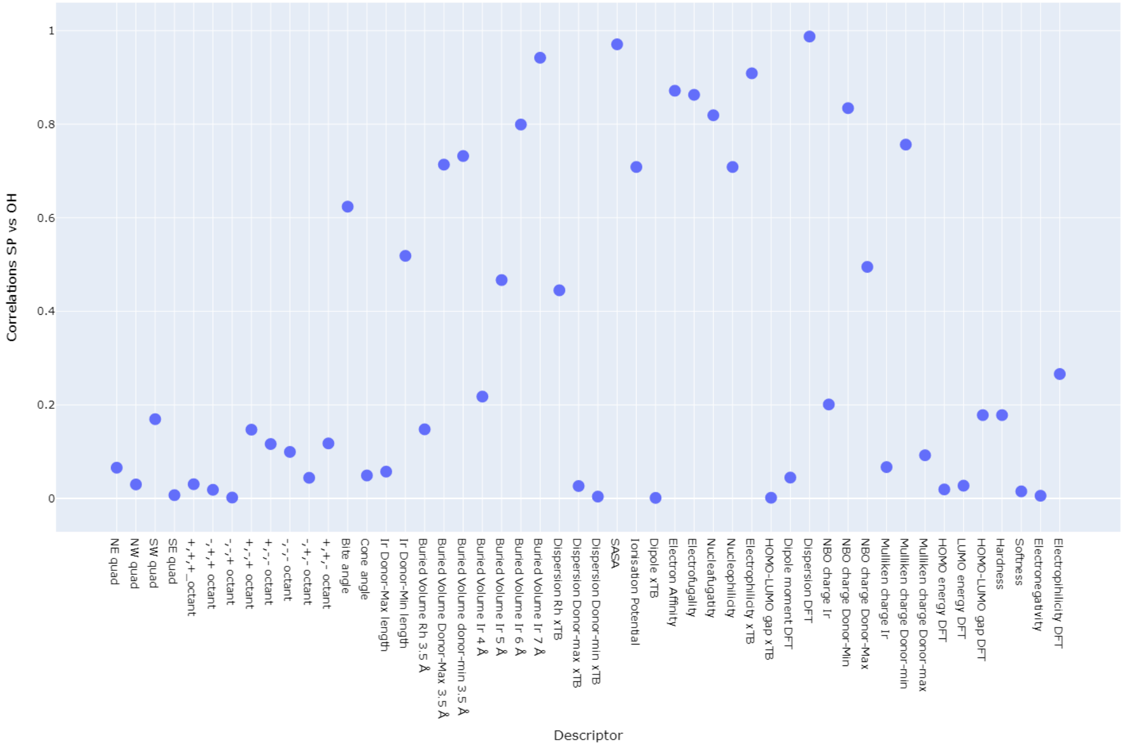


Figure 4.7: Scatter plot of the Pearson correlation coefficients of octahedral- versus pristine complex descriptors. Both xTB and DFT descriptors are included. Every point is a correlation of 57 complexes.

The **bite angle** does show some correlation ($r^2 = 0.62$), although it is lower than initially anticipated. One explanation for the lower correlation could perhaps be due electronic effects of the hydrides and ACN. In this scenario, the hydrides and ACN repulse the phosphorus/nitrogen and thus lower the **bite angle**. This is supported further by the means of both sets, which are 94.95° and 92.57° for the pristine and OH set, respectively. Another contribution could be the steric hindrance introduced in OH complexes due to the ACN substrate. In pristine complexes, the ligand can coordinate to the metal centre and fold itself, lowering the **bite angle**.

Correlation of the **cone angle** is surprisingly completely absent. The **cone angle** is theoretically only dependent on the bidentate ligand, so at least some correlation was initially expected. Further inspection of the data shows that **cone angle** values of the pristine set are often very high (> 270°). The mean of the pristine set is also much higher than the mean of the OH set, 291° versus 233°, respectively. The reason for this discrepancy is the lack of steric hindrance in the pristine complexes, which allows ligands to fold over and move towards the metal centre, even without forming bonds. The **cone angle** is defined as the cone towards the van der Waals radii of the last atoms of the ligand, see Section 4.2.1. As such, if the ligand folds over then the **cone angle** will be much larger than its non-folded counterpart. **Cone angle** is therefore not a suitable descriptor for pristine complexes.

The correlation between **Ir-Donor min length** exhibits an acceptable degree of correlation, which is significantly superior to correlations observed between **Ir-Donor max length**. This discrepancy is attributed to the considerably smaller bond length of nitrogen atoms in comparison to phosphorus atoms. As 10% of the data constitutes nitrogen, the resultant distribution is clustered. While a regression line can be drawn through the centre of the two clusters, indicating an acceptable correlation, the resulting correlation is deceptive.

4.5.2. Steric xTB descriptors

In **buried volume Donor min** and **buried volume Donor max** good correlation is observed (r^2 = 0.73 & 0.71, respectively), which can be attributed to the fact that most of the atoms in the ligand are being included in the circle sphere, similarly to Section 4.2.2, even though the OH set now additionally contains the hydrides and ACN ligands.

The descriptors of **buried volume Ir** follow the same trend as **buried volume Rh** in Section 4.2. The higher the radius of the buried volume sphere, the higher the correlation. Again, this is due to a bigger sphere encompassing all the atoms of the complex. The default value of 3.5 Å shows no correlation, however.

The dispersion on Rh exhibits some correlation (r^2 = 0.44) however with dispersion on either donor min or donor max, correlation is absent. For the metal centre, P_{int} in a pristine complex is only the interaction with the bidentate ligand. For an OH complex, interactions of the hydrides and ACN are added. Although these interactions are constant for every ligand, this dissimilarity could explain the lower correlation, by virtue of geometric variations. For the donor atoms, the first shell atoms to the donor atom consist again of the metal, but also a lot of distinct classes ($CH_2, O, C-O, C_5H_6,$etc. with high variation in polarity between them. These different groups change the geometry significantly as well, which can theoretically change the P_{int} by large amounts. As such, a correlation on dispersion is absent for the donor atoms.

SASA exhibits very good correlation (r^2 = 0.97), similar to Section 4.2. Again, it is generally known that **SASA** scales with the number of electrons in the system [150]. The amount of electrons scales linearly in both the pristine and the OH complexes since the only difference is the constant electrons obtained from the hydrides and ACN.

4.5.3. Electronic xTB descriptors

Dipole xTB, as determined by the dipole vector in Morfeus, contains no correlation. The dipole moment is a measure of a molecule's polarity, and it is hypothesized that the dipole values diverge due to the higher polarity of the OH complexes. This hypothesis is supported by the observation that, on average, the **dipole xTB** of OH complexes is twice as large as their pristine counterparts (3.04 versus 1.54 for OH and pristine, respectively).

For all electronic HOMO/LUMO dependant xTB descriptors (EA, IP, NF, NP, EF, EP) good correlation is observed ($r^2 \geq 0.7$), similar to in Section 4.2. IP and NP once again exhibit the lowest correlation (r^2 = 0.71), while for EP the highest correlation is found (r^2 = 0.91). This is in line with the hypothesis presented in Section 4.2.3, where differences in correlation are solely due to statistical deviations.

Correlation of **HOMO-LUMO gap xTB** is completely absent. Initially, it was theorized that the method of finding MO orbitals might not be exact enough. In other words, a higher level of theory single-point calculation would be needed to accurately determine the HOMO and LUMO levels. However, the correlation of **HOMO-LUMO gap DFT** is also completely absent. In addition, the descriptors **HOMO** and **LUMO** also show no correlation. This finding is interesting, since **HOMO-LUMO gap** has been used in previous studies as a useful electronic descriptor [125]. In an online database of descriptors, the tmQM database, **HOMO-LUMO gap** has been calculated on a higher accuracy DFT (TPSSh-D3BJ/def2-SVP). In that research, Balcells and Skjelstad compared **HOMO-LUMO gap** to other electronic descriptors for 86k TM complexes [153]. A low correlation was found between electronic properties. As such, in general, electronic descriptors, and especially HOMO and LUMO levels are very sensitive to geometry changes. Nonetheless, a plethora of reasons is possible why HOMO and LUMOs do not correlate on pristine versus OH complexes.

4.5.4. Electronic DFT descriptors

Correlation on **dipole moment DFT**is completely absent, similar to **dipole moment xTB**. Means here are very dissimilar, 30.12 and 6.22D for the pristine and OH sets, respectively. In the pristine set, values range from -10 to 60 Debye, whilst for the OH sets values are within 0 and 13 Debye. The pristine complexes are models and thus very chemically unstable, which could lead to very volatile dipole moment values. These values will then not correspond with the much more stable OH complexes.

Dispersion DFT exhibits extremely good correlation ($R^2 = 0.99$). Dispersion energy from DFT refers to the length of the vector of dispersion on the whole system, contrary to $DipsersionxTB$. This value is thus dependent on the size of the system. Adding extra ligands to the metal centres in the form of ACN and hydrides changes the size of the system with a constant value. Thus it is logical that the energies will correspond.

NBO charge Donor min exhibits strong correlation ($r^2 = 0.83$), whereas for **NBO charge Ir** and **NBO charge Donor min** low correlations are found. The observed strong correlation of the minimum donor is attributed to the formation of two clusters containing either N or P donor atoms, which distorts the correlation coefficient. By excluding the complexes with a nitrogen minimum donor, the correlation becomes non-existent and similar to the correlation between the metal centre and the maximum donor.

Correlation for **Mulliken charge Donor min** is observed to be good ($R^2 = 0.76$), but correlation for **Mulliken charge Ir** and **Mulliken charge Donor max** is absent. Once again, this is due to the same reason as mentioned above. Clusters are formed and skew the correlation coefficient. Clusters themselves contain no correlation.

For the descriptors **hardness**, **softness**, **electronegativity**, and **electrophilicity DFT** no correlation is observed. The lack of correlation can be attributed to the fact that HOMO and LUMO levels themselves do not show a correlation either. The above descriptors are again all fully dependent on the HOMO and the LUMO level, whether linearly or quadratically, due to Koopmans' theorem. Since HOMO and LUMO levels do not correlate, correlations remains absent for these descriptors as well.

5
Conclusion & outlook

5.1. Conclusion

The aim of this research was to compare possible simplification methods for *in silico* homogeneous TM-based catalyst screening workflow.

Objective 1 was to compare geometries and descriptors of 3 lower levels of theory versus DFT. The 3 levels of theory considered were: MACE (UFF-MM), GFN-FF, and GFN2-xTB. Geometry optimizations have been done, for each level of theory, on a set of chemically relevant pristine rhodium complexes. On these complexes, geometric, steric and electronic descriptors have been extracted from xTB single-point calculations. In general, descriptors from MACE and GFN-FF optimized structures lacked correlation versus DFT descriptors, except for SASA and high radii $\%V_{\text{bur}}$. It was concluded that a good correlation of SASA originated because of dependence system size and good correlation of $\%V_{\text{bur}}$ due to all atoms being found in larger spheres. However, most descriptors obtained from xTB-optimized structures correlated very well with their DFT counterpart. Possibly this was due to descriptors being extracted from xTB single-point calculations.

As such, to further investigate objective 1, hRMSDs were computed for the same set of complexes for each level of theory against DFT. By using hRMSD, purely the geometries of the structures are being matched and possible bias originating from single-point xTB calculations is omitted. The hRMSDs mean and standard deviation of the xTB set were much lower than the mean and standard deviation of both the MACE and the GFN-FF set. Outlier inspection of the xTB dataset via the IQR method showed that some ligands formed bonds with the metal centre, therefore creating a polydentate ligand. These ligands, however, do not correspond with realistic structures. By considering the small mean value and unrealistic outlier structures, it can be concluded that xTB geometry optimization is adequate for obtaining structures with comparable levels of precision to DFT geometry-optimized structures.

To again omit possible bias originating from single-point xTB calculation, additional DFT single-point calculations were done for both xTB and DFT geometry-optimized structures. From these DFT single-point calculations, 3 electronic descriptors were extracted and correlated to their xTB counterpart for both xTB and DFT geometries. From these results, it was deduced that xTB single-point descriptors do not correlate well with DFT descriptors, even when considering the exact same geometries. On the other hand, DFT single-point descriptors do correlate well with xTB and DFT geometries. For objective 1 it can be concluded that xTB geometry optimization is adequate in comparison to DFT geometry optimization, xTB single-point level descriptors are not comparable to DFT single-point descriptors however. Thus, the workflow ideally should be:

1. Generate structures with MACE
2. Optimize structures with GFN2-xTB geometry optimization
3. Extract descriptors with single-point DFT calculation

Objective 2 of this research was to simplify the digitally represented structures. For this reason, two sets of complexes were generated. The first set contained realistic octahedral iridium bidentate complexes based on precatalyst structure. The second set contained a simplified version of the octahedral set, consisting of just the iridium metal centre and the bidentate ligands, which were named pristine complexes. These pristine complexes served as the model structure for the octahedral complexes, which contained additional acetonitrile and hydrides bonded to the metal centre. DFT geometry optimization was done on both sets of complexes, and both xTB level descriptors and DFT level descriptors were extracted for both sets. Subsequently, the descriptors were subjected to intercorrelation analysis. In most cases, correlations were not observed, which could be attributed to the descriptor not being suited for ligands capable of deforming towards the metal centre. In the case of high $\%V_{\mathrm{bur}}$ radii and SASA correlations were excellent, similar to the results obtained from the rhodium set. Electronic descriptors that depend on HOMO/LUMO level, extracted from xTB single-point calculations, generally exhibited strong correlations as well. However, it should be noted that these correlations do not extend to DFT descriptors, as per the previous paragraph. High correlations were also observed for descriptors centred on the min donor, however, upon inspection these high correlations were observed to be misleading. The presence of two distinct bidentate ligand classes in the dataset, namely P-P and P-N, formed two separate clusters. Although a trendline through both clusters yielded good correlations, no correlations were present within the clusters.

This finding, however, raises the question of the applicability of doing regression methods on a set of chemically dissimilar complexes, at least for atom-centred atoms. Instead, clustering methods might be more suitable to describe different classes.

5.2. Outlook

From the research in this thesis, a concrete conclusion can be drawn: **Do xTB geometry optimization and get descriptors from single-point DFT calculations**. This change can be implemented in the workflow of OBeLiX, see Figure 5.1, to obtain faster screening with similar accuracy.

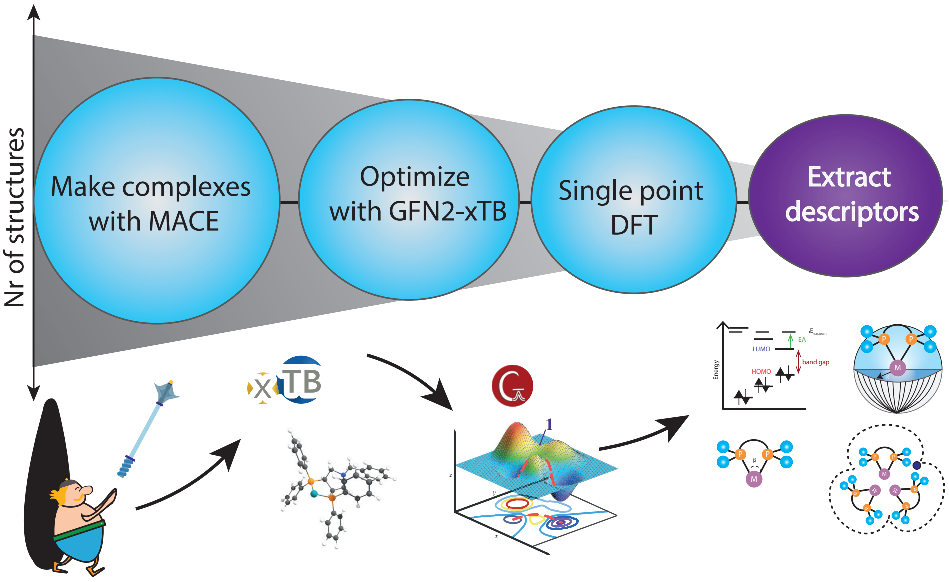


Figure 5.1: Possible improvement to the current workflow of OBeLiX

Next to that implementation, several other possible paths of research can be envisioned. In the dataset obtained from the pharmaceutical partner, approximately 40% of the complexes possessed ligands containing ferrocenes. Ferrocenes are not able to be accurately generated from 2D SMILES notation, as such cannot be automated. However, ferrocenes are a key part of the homogeneous catalysis field, and the dataset used in this research is thus not truly representative of the TM catalysts used experimentally. As such, it should be considered to use the OBeLiX workflow to calculate descriptors for the whole dataset, including (hand-made and DFT optimized) ferrocenes.

It is known that good ligands on a complex have similar descriptors to each other. By varying the metal centre, it can be investigated if descriptors remain similar even across different metal centres, which could potentially lead to novel catalyst structures. Further research on the transferability of descriptors can thus be done by changing the TM centre and running the OBeLiX workflow. Whether the metal centre needs to be changed by elements in the same group (vertical) or series (horizontal) on the periodic table, e.g. Ir → Rh or Ir → Mn, is something that needs to be validated. This undoubtedly needs to be preceded by obtaining extensive (experimental) data on the stable pre-catalyst structure, so that the multiplicity and oxidation state of the complex is known.

For future comparison of descriptors, it should be carefully evaluated if regression models are applicable. As mentioned in Section 5.1, chemically dissimilar complexes show clusters for descriptors on atoms which skew the correlation coefficient. It might be better to use other methods to handle these clusters better, e.g. logistic regression, dimensionality reduction techniques followed by clustering descriptors, or deep learning. Another possible approach to keep using regression methods is modifying descriptors to average out over multiple atoms, which would create a continuous distribution of data. However, if one applies this method, valuable chemical insight gets lost, since in chemistry interesting findings lie within the outliers. Another additional approach is to calculate descriptors on structures of different geometry optimization levels, exactly as has been done in this research, but followed by principal component analysis to reduce the dimensionality of all the descriptors. Subsequently, by plotting the first two or three PCs in a scatter plot, then colour code/-mark the level of theory and the structures, both clustering and overall differences in the descriptor data can be found.

In general, since the field of homogeneous *in silico* catalyst design is quite new, benchmarking studies are mandatory to advance the field. The goal of benchmarking studies is to identify the most accurate and efficient methods for predicting new catalysts. Using a lower level of theory can speed up calculations by dozens of times. If the same accuracy is kept, this can lower the amount of resources used a great deal.

Benchmarking needs to be done on DFT programs, DFT basis sets and DFT methods (functionals). For programs, the in-this-research-used Gaussian needs to be compared to other DFT programs such as Turbo-Mole, ORCA or VASP. Gaussian is the most well-known and is traditionally used the most, but newer DFT programs provide options for newer basis sets (TubroMole), use plane-wave basis sets (VASP), or are simply free to use (ORCA). DFT basis sets provide the basis for orbitals and the electronic structure. Generally, using larger basis sets increases accuracy, but comes with an increased computational cost. Having said that, the list of basis sets is enormous, in which some are atom-based and some are molecule-based, and others are combinations of the two. Then polarization and diffusion functions are sometimes added, increasing the list. To determine whether the use of widely used basis sets, such as STO-3G or cc-pVDZ, or the exploration of the effectiveness of other basis sets developed by Aldrich et al. (Def2SVP, Def2TZVP or QZVP), could lead to improved results, benchmarking studies are required. Lastly, DFT functionals can be varied to be either pure (combination of exchange + correlation functionals) or hybrid. It can be investigated to use simplified pure functionals, such as GGA (TPSS) or LDA, and comparing energies or descriptors. These calculations will cost much less than a PBE0 calculation. It is also possible to consider other prevalent hybrid functionals, such as B3LYP or MO6-L, but theoretical considerations need to be made since no best functional exists and they are all system specific.

Benchmarking studies also extend towards semi-empirical methods. Possible options to try, include but are not limited to: GFN0-xTB, the most contemporary GFN method, MOPAC, AM1 and PM6. Again, methods should be chosen on a theoretical basis. For example, MOPAC has been tested extensively for biological systems, not TM-based systems.

Taking all these aspects into consideration, the field of *in silico* homogeneous TM catalyst design is uncharted though fresh and exciting.

The ~~DFT functional heaven~~ sky is the limit.

Acknowledgements

The research in this thesis was performed at the Inorganic Systems Engineering group for the Msc Chemical Engineering at the Technical University Delft.

First and foremost, I would like to express my gratitude towards my daily supervisor Msc. Adarsh Kalikadien. Next to being extremely helpful in computational chemistry, data science and plain science, you also were always very much a delight to work with. I will be looking forward to seeing which company (or continent?!) will have you as their CTO. Second, I'd like to thank my supervisor Professor Pidko for a multitude of reasons. Firstly, for welcoming me into the ISE group instantaneously, quickly taking away my fear of never finding an MEP. Secondly, allowing us Msc students to experience the luxury that is a scientific conference. And thirdly, for barging into our office with some random topic, which always ended up sparking conversation. These experiences made my journey much more memorable and authentic. Third, I'd like to thank my fellow master's student Andrian Mirza. Even though your work ethic and working hours still shock me to the bone, you have opened my eyes to what a person can do, if only they put their mind to it. Thanks of course also go out to the many other unique personalities in the ISE group: Dr Sreenithya Avadakkam, Dr Christophe Rebreyend, Dr Alexander Kolganov and last but very much not least, Dr Georgy Filonenko.

Lastly, I would like to thank my family and friends for supporting me throughout these 7 months. Matthijs for joining me on an unforgettable adventure to the end of the world. All my other friends for their unconditional support, which made all the difference in the world.My brother and my dad for bearing with me when I was lost in the matrix. My mom and stepmom for being kind, patient and inquiring. My sister for her high school drama distractions. And everybody else, for guiding, supporting and cooking food for me.

My gratitude especially goes out to some exceptionally kind people, whose names need not be mentioned for they know who they are, since I could not have persevered near the end of my journey without them.

- Aydin Najl Hossaini
Delft, April 2023

Bibliography

[1] C Richard Catlow, Matthew Davidson, Christopher Hardacre, and Graham J Hutchings. Catalysis making the world a better place. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2061):20150089, 2016.

[2] Calvin H Bartholomew and Robert J Farrauto. *Fundamentals of industrial catalytic processes*. John Wiley & Sons, 2011.

[3] Piet WNM Van Leeuwen. *Homogeneous catalysis: understanding the art*. Springer Science & Business Media, 2006.

[4] Erica Farnetti, Roberta Di Monte, and Jan Kašpar. Homogeneous and heterogeneous catalysis. *Inorganic and bio-inorganic chemistry*, 2(6):50–86, 2009.

[5] Christophe Copéret, Mathieu Chabanas, Romain Petroff Saint-Arroman, and Jean-Marie Basset. Homogeneous and heterogeneous catalysis: bridging the gap through surface organometallic chemistry. *Angewandte Chemie International Edition*, 42(2):156–181, 2003.

[6] Elizabeth L Bell, William Finnigan, Scott P France, Anthony P Green, Martin A Hayes, Lorna J Hepworth, Sarah L Lovelock, Haruka Niikura, Sílvia Osuna, Elvira Romero, et al. Biocatalysis. *Nature Reviews Methods Primers*, 1(1):46, 2021.

[7] Jennifer Lew (UCD). Reaction of alkenes- hydrogenation, 2023. URL https://chem.libretexts.org/Bookshelves/Organic_Chemistry/Organic_Chemistry_I_28Liu29/103A_Alkenes_and_Alkynes/10.053A_Reaction_of_Alkenes-_Hydrogenation.

[8] Ulla Lepola, Alan Wade, and Henning Friis Andersen. Do equivalent doses of escitalopram and citalopram have similar efficacy? a pooled analysis of two positive placebo-controlled studies in major depressive disorder. *International clinical psychopharmacology*, 19(3):149–155, 2004.

[9] Silas W Smith. Chiral toxicology: it's the same thing... only different. *Toxicological sciences*, 110(1):4–30, 2009.

[10] Ivano Bertini. *Inorganic and Bio-Inorganic Chemistry-Volume II*, volume 6. EOLSS Publications, 2009.

[11] Ryoji Noyori. Asymmetric hydrogenation. *Acta Chemica Scandinavica*, 50:380–390, 1996.

[12] Billy D Vineyard, William S Knowles, Milton J Sabacky, G_ L_ Bachman, and D_ J Weinkauff. Asymmetric hydrogenation. rhodium chiral bisphosphine catalyst. *Journal of the American Chemical Society*, 99(18):5946–5952, 1977.

[13] Michel van den Berg, Robert M Haak, Adriaan J Minnaard, André HM de Vries, Johannes G de Vries, and Ben L Feringa. Rhodium/monophos-catalysed asymmetric hydrogenation of enamides. *Advanced Synthesis & Catalysis*, 344(9):1003–1007, 2002.

[14] Stephen J Roseblade and Andreas Pfaltz. Iridium-catalyzed asymmetric hydrogenation of olefins. *Accounts of chemical research*, 40(12):1402–1411, 2007.

[15] Tetsuo Ohta, Hidemasa Takaya, Masato Kitamura, Katsunori Nagai, and Ryoji Noyori. Asymmetric hydrogenation of unsaturated carboxylic acids catalyzed by binap-ruthenium (ii) complexes. *The Journal of Organic Chemistry*, 52(14):3174–3176, 1987.

[16] Jian-Hua Xie, Shou-Fei Zhu, and Qi-Lin Zhou. Transition metal-catalyzed enantioselective hydrogenation of enamines and imines. *Chemical Reviews*, 111(3):1713–1760, 2011. doi: 10.1021/cr100218m. URL https://doi.org/10.1021/cr100218m. PMID: 21166392.

[17] Michael Evans. Predicting the geometry of organometallic complexes, 2019. URL https://chem.libretexts.org/Courses/Johns_Hopkins_University/030.356_Advanced_Inorganic_Laboratory/02%3A_Lab_BCD-_Four_Coordinate_Nickel_Complexes-_Ligand_Effects_and_Organometallic_Catalysis/2.02%3A_Predicting_the_Geometry_of_Organometallic_Co

[18] Peter Kirkpatrick and Clare Ellis. Chemical space. *Nature*, 432(7019):823–824, 2004.

[19] Steven M Mennen, Carolina Alhambra, C Liana Allen, Mario Barberis, Simon Berritt, Thomas A Brandt, Andrew D Campbell, Jesús Castañón, Alan H Cherney, Melodie Christensen, et al. The evolution of high-throughput experimentation in pharmaceutical development and perspectives on the future. *Organic Process Research & Development*, 23(6):1213–1242, 2019.

[20] Shane W Krska, Daniel A DiRocco, Spencer D Dreher, and Michael Shevlin. The evolution of chemical high-throughput experimentation to address challenging problems in pharmaceutical synthesis. *Accounts of chemical research*, 50(12):2976–2985, 2017.

[21] Joshua A Selekman, Jun Qiu, Kristy Tran, Jason Stevens, Victor Rosso, Eric Simmons, Yi Xiao, and Jacob Janey. High-throughput automation in chemical process development. *Annual review of chemical and biomolecular engineering*, 8:525–547, 2017.

[22] Carl Poree and Franziska Schoenebeck. A holy grail in chemistry: Computational catalyst design: Feasible or fiction?, mar 2017. ISSN 15204898.

[23] David Balcells and Feliu Maseras. Computational approaches to asymmetric synthesis. *New Journal of Chemistry*, 31:333, 2007. ISSN 1144-0546. doi: 10.1039/b615528f.

[24] 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*, 9:6803–6813, 8 2019. ISSN 2155-5435. doi: 10.1021/acscatal.9b01537.

[25] Marco Foscato and Vidar R. Jensen. Automated in silico design of homogeneous catalysts. *ACS Catalysis*, 10:2354–2377, 2 2020. ISSN 2155-5435. doi: 10.1021/acscatal.9b04952.

[26] Iñigo Iribarren, Marianne Rica Garcia, and Cristina Trujillo. Catalyst design within asymmetric organocatalysis. *WIREs Computational Molecular Science*, 12, 11 2022. ISSN 1759-0876. doi: 10.1002/wcms.1616.

[27] Qi An, Yidi Shen, Alessandro Fortunelli, and William A. Goddard. Qm-mechanism-based hierarchical high-throughput in silico screening catalyst design for ammonia synthesis. *Journal of the American Chemical Society*, 140:17702–17710, 12 2018. ISSN 0002-7863. doi: 10.1021/jacs.8b10499.

[28] Enrico Burello and Gadi Rothenberg. In silico design in homogeneous catalysis using descriptor modelling. *International Journal of Molecular Sciences*, 7:375–404, 9 2006. ISSN 1422-0067. doi: 10.3390/i7090375.

[29] Kaid C Harper, Elizabeth N Bess, and Matthew S Sigman. Multidimensional steric parameters in the analysis of asymmetric catalytic reactions. *Nature chemistry*, 4(5):366–374, 2012.

[30] Wendy L Williams, Lingyu Zeng, Tobias Gensch, Matthew S Sigman, Abigail G Doyle, and Eric V Anslyn. The evolution of data-driven modeling in organic chemistry. *ACS central science*, 7(10):1622–1637, 2021.

[31] HH Jaffe. A reexamination of the hammett equation. *Chemical Reviews*, 53(2):191–261, 1953.

[32] Pilar López Cornejo, Rafael Jiménez, María Luisa Moyá, Francisco Sánchez, and John Burgess. Use of the brønsted equation in the interpretation of micellar effects in kinetics. *Langmuir*, 12(21):4981–4986, 1996.

[33] 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, 2009.

[34] Chadwick A Tolman. Steric effects of phosphorus ligands in organometallic chemistry and homogeneous catalysis. *Chemical reviews*, 77(3):313–348, 1977.

[35] Ana G. Maldonado and Gadi Rothenberg. Predictive modeling in homogeneous catalysis: A tutorial. *Chemical Society Reviews*, 39:1891–1902, 5 2010. ISSN 14604744. doi: 10.1039/b921393g.

[36] Wu-chun Feng and Kirk Cameron. The green500 list: Encouraging sustainable supercomputing. *Computer*, 40(12):50–55, 2007.

[37] Nicola Jones et al. The information factories. *Nature*, 561(7722):163–6, 2018.

[38] Yanan Liu, Xiaoxia Wei, Jinyu Xiao, Zhijie Liu, Yang Xu, and Yun Tian. Energy consumption and emission mitigation prediction based on data center traffic and pue for global data centers. *Global Energy Interconnection*, 3(3):272–282, 2020.

[39] LOÏC LANNELONGUE. Carbon footprint, the (not so) hidden cost of high performance computing, 2021. URL https://www.bcs.org/articles-opinion-and-research/carbon-footprint-the-not-so-hidden-cost-of-high-performance-computing/.

[40] Joana Setzer and Catherine Higham. Global trends in climate change litigation: 2022 snapshot. *Grantham Research Institute*, 2022.

[41] Christopher J Cramer. *Essentials of Computational Chemistry*. Wiley Editorial Offices, 2004.

[42] Joseph JW McDouall. *Computational quantum chemistry: molecular structure and properties in silico*. Royal Society of Chemistry, 2013.

[43] Stephen Lower and Tom Neils, 2022. URL https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Physical_Chemistry_%28LibreTexts%29/30%3A_Gas-Phase_Reaction_Dynamics/30.10%3A_The_Potential-Energy_Surface_Can_Be_Calculated_Using_Quantum_Mechanics.

[44] Max Born. Born-oppenheimer approximation. *Ann. Phys*, 84:457–484, 1927.

[45] Max Born and W Heisenberg. Zur quantentheorie der molekeln. *Original Scientific Papers Wissenschaftliche Originalarbeiten*, pages 216–246, 1985.

[46] Errol G Lewars and Errol G Lewars. *An Outline of What Computational Chemistry Is All About*. Springer, 2011.

[47] Sebastian Spicher and Stefan Grimme. Robust atomistic modeling of materials, organometallic, and biochemical systems. *Angewandte Chemie International Edition*, 59:15665–15673, 9 2020. ISSN 1433-7851. doi: 10.1002/anie.202004239.

[48] John P. Brodholt and L. Voĉadlo. Applications of density functional theory in the geosciences. *MRS Bulletin*, 31(9):675–680, Sep 2006. ISSN 1938-1425. doi: 10.1557/mrs2006.176. URL https://doi.org/10.1557/mrs2006.176.

[49] Daniel J Cole and Nicholas DM Hine. Applications of large-scale density functional theory in biology. *Journal of Physics: Condensed Matter*, 28(39):393001, 2016.

[50] P. Geerlings, F. De Proft, and W. Langenaeker. Conceptual density functional theory. *Chemical Reviews*, 103(5):1793–1874, 2003. doi: 10.1021/cr990029p. URL https://doi.org/10.1021/cr990029p. PMID: 12744694.

[51] Adarsh Kalikadien. Automated data-driven exploration of chemical space for catalysts. master's thesis. Master's thesis, Delft University of Technology, April 2021.

[52] Attila Szabo and Neil S Ostlund. *Modern quantum chemistry: introduction to advanced electronic structure theory*. Courier Corporation, 2012.

[53] Ira N Levine, Daryle H Busch, and Harrison Shull. *Quantum chemistry*, volume 6. Pearson Prentice Hall Upper Saddle River, NJ, 2009.

[54] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964. doi: 10.1103/PhysRev.136.B864. URL https://link.aps.org/doi/10.1103/PhysRev.136.B864.

[55] Junlei Zhao and Flyura Djurabekova. Computational modeling of nanoparticles in inert environment. *Frontiers of Nanoscience*, 17:5–26, 2020.

[56] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965. doi: 10.1103/PhysRev.140.A1133. URL https://link.aps.org/doi/10.1103/PhysRev.140.A1133.

[57] Pragya Verma and Donald G Truhlar. Status and challenges of density functional theory. *Trends in Chemistry*, 2(4):302–318, 2020.

[58] 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, 1999.

[59] Matthias Ernzerhof and Gustavo E Scuseria. Assessment of the perdew–burke–ernzerhof exchange-correlation functional. *The Journal of chemical physics*, 110(11):5029–5036, 1999.

[60] John P Perdew and Karla Schmidt. Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conference Proceedings*, 577(1):1–20, 2001.

[61] David Bohm and David Pines. A collective description of electron interactions. i. magnetic interactions. *Physical Review*, 82(5):625, 1951.

[62] A Heßelmann and A Görling. Random-phase approximation correlation methods for molecules and solids. *Molecular Physics*, 109(21):2473–2500, 2011.

[63] S Francis Boys. Electronic wave functions-i. a general method of calculation for the stationary states of any molecular system. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 200(1063):542–554, 1950.

[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, 2005. ISSN 1463-9076. doi: 10.1039/b508541a.

[65] Ansgar Schäfer, Hans Horn, and Reinhart Ahlrichs. Fully optimized contracted gaussian basis sets for atoms li to kr. *The Journal of Chemical Physics*, 97(4):2571–2577, 1992.

[66] Jingjing Zheng, Xuefei Xu, and Donald G Truhlar. Minimally augmented karlsruhe basis sets. *Theoretical Chemistry Accounts*, 128:295–305, 2011.

[67] Manuel Pérez, Torren M Peakman, Alexander Alex, Paul D Higginson, John C Mitchell, Martin J Snowden, and Iñaki Morao. Accuracy vs time dilemma on the prediction of nmr chemical shifts: A case study (chloropyrimidines). *The Journal of Organic Chemistry*, 71(8):3103–3110, 2006.

[68] Anoop Kumar Kushwaha. A brief review of density functional theory and solvation model. *T.B.D.*, 2022.

[69] 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, 1994.

[70] 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, 2011.

[71] Erin R Johnson, Iain D Mackie, and Gino A DiLabio. Dispersion interactions in density-functional theory. *Journal of Physical Organic Chemistry*, 22(12):1127–1135, 2009.

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

[73] Eike Caldeweyher, Jan-Michael Mewes, Sebastian Ehlert, and Stefan Grimme. Extension and evaluation of the d4 london-dispersion model for periodic systems. *Physical Chemistry Chemical Physics*, 22(16): 8499–8512, 2020.

[74] Michael E. Foster and Karl Sohlberg. Empirically corrected dft and semi-empirical methods for non-bonding interactions. *Phys. Chem. Chem. Phys.*, 12:307–322, 2010. ISSN 1463-9076. doi: 10.1039/B912859J.

[75] G. Seifert. Tight-binding density functional theory: An approximate kohn-sham dft scheme. *The Journal of Physical Chemistry A*, 111:5609–5613, 7 2007. ISSN 1089-5639. doi: 10.1021/jp069056r.

[76] 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*, 5:1710252, 1 2020. ISSN 2374-6149. doi: 10.1080/23746149.2019.1710252.

[77] Pekka Koskinen and Ville Mäkinen. Density-functional tight-binding for beginners. *Computational Materials Science*, 47:237–253, 11 2009. ISSN 09270256. doi: 10.1016/j.commatsci.2009.07.013.

[78] Marcus Elstner and Gotthard Seifert. Density functional tight binding. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2011):20120483, 2014. doi: 10.1098/rsta.2012.0483. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2012.0483.

[79] Christoph Bannwarth, Eike Caldeweyher, Sebastian Ehlert, Andreas Hansen, Philipp Pracht, Jakob Seibert, Sebastian Spicher, and Stefan Grimme. Extended tight-binding quantum chemistry methods. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 11, 3 2021. ISSN 17590884. doi: 10.1002/wcms.1493.

[80] 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*, 7:931–948, 4 2011. ISSN 1549-9618. doi: 10.1021/ct100684s.

[81] 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*, 15:1652–1671, 3 2019. ISSN 1549-9618. doi: 10.1021/acs.jctc.8b01176.

[82] 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*, 150:154122, 4 2019. ISSN 0021-9606. doi: 10.1063/1.5090222.

[83] Alexey V Onufriev and David A Case. Generalized born implicit solvent models for biomolecules. *Annual review of biophysics*, 48:275–296, 2019.

[84] Andreas Klamt. The cosmo and cosmo-rs solvation models. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5):699–709, 2011.

[85] Philipp Pracht, Eike Caldeweyher, Sebastian Ehlert, and Stefan Grimme. A robust non-self-consistent tight-binding quantum chemistry method for large molecules. *T.B.D.*, 2019.

[86] Mati Karelson, Victor S. Lobanov, and Alan R. Katritzky. Quantum-chemical descriptors in qsar/qspr studies. *Chemical Reviews*, 96:1027–1044, 1 1996. ISSN 0009-2665. doi: 10.1021/cr950202r.

[87] Giovanni Occhipinti, Hans-René Bjørsvik, and Vidar R. Jensen. Quantitative structure-activity relationships of ruthenium catalysts for olefin metathesis. *Journal of the American Chemical Society*, 128:6952–6964, 5 2006. ISSN 0002-7863. doi: 10.1021/ja060832i.

[88] Mandana Saebi, Bozhao Nan, John E. Herr, Jessica Wahlers, Zhichun Guo, Andrzej M. Zurański, Thierry Kogej, Per-Ola Norrby, Abigail G. Doyle, Nitesh V. Chawla, and Olaf Wiest. On the use of real-world datasets for reaction yield prediction. *Chemical Science*, 2023. ISSN 2041-6520. doi: 10.1039/D2SC06041H. URL http://xlink.rsc.org/?DOI=D2SC06041H.

[89] 5.3: Molecular Descriptors — chem.libretexts.org. https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics/05%3A_5._Quantitative_Structure_Property_Relationships/5.03%3A_Molecular_Descriptors, 2022. [Accessed 29-Mar-2023].

[90] Balakumar Chandrasekaran, Sara Nidal Abed, Omar Al-Attraqchi, Kaushik Kuche, and Rakesh K. Tekade. *Computer-Aided Prediction of Pharmacokinetic (ADMET) Properties*. Elsevier, 2018. doi: 10.1016/B978-0-12-814421-3.00021-X.

[91] Derek J. Durand and Natalie Fey. Computational ligand descriptors for catalyst design. *Chemical Reviews*, 119:6561–6594, 6 2019. ISSN 15206890. doi: 10.1021/ACS.CHEMREV.8B00588/ASSET/IMAGES/MEDIUM/CR-2018-00588P_M008.GIF. URL https://pubs-acs-org.tudelft.idm.oclc.org/doi/full/10.1021/acs.chemrev.8b00588.

[92] Peter Dierkes and Piet W. N. M. van Leeuwen. The bite angle makes the difference: a practical ligand parameter for diphosphine ligands. *Journal of the Chemical Society, Dalton Transactions*, pages 1519–1530, 1999. ISSN 03009246. doi: 10.1039/a807799a.

[93] Natalie Fey, James A. S. Howell, Jonathan D. Lovatt, Paul C. Yates, Desmond Cunningham, Patrick McArdle, Hugo E. Gottlieb, and Simon J. Coles. A molecular mechanics approach to mapping the conformational space of diaryl and triarylphosphines. *Dalton Transactions*, 11(46):5464, 2006. ISSN 1477-9226. doi: 10.1039/b610123b.

[94] Linus Pauling. The dependence of bond energy on bond length. *The Journal of Physical Chemistry*, 58 (8):662–666, 1954.

[95] Martin Kaupp, David Danovich, and Sason Shaik. Chemistry is about energy and its changes: A critique of bond-length/bond-strength correlations. *Coordination Chemistry Reviews*, 344:355–362, 8 2017. ISSN 00108545. doi: 10.1016/j.ccr.2017.03.002.

[96] Mihai Burai Patrascu, Joshua Pottel, Sharon Pinus, Michelle Bezanson, Per Ola Norrby, and Nicolas Moitessier. From desktop to benchtop with automated computational workflows for computer-aided design in asymmetric catalysis. *Nature Catalysis*, 3:574–584, 7 2020. ISSN 25201158. doi: 10.1038/s41929-020-0468-3.

[97] Hervé Clavier and Steven P. Nolan. Percent buried volume for phosphine and n-heterocyclic carbene ligands: steric properties in organometallic chemistry. *Chemical Communications*, 46:841, 2010. ISSN 1359-7345. doi: 10.1039/b922984a.

[98] Laura Falivene, Raffaele Credendino, Albert Poater, Andrea Petta, Luigi Serra, Romina Oliva, Vittorio Scarano, and Luigi Cavallo. Sambvca 2. a web tool for analyzing catalytic pockets with topographic steric maps. *Organometallics*, 35:2286–2293, 7 2016. ISSN 0276-7333. doi: 10.1021/acs.organomet.6b00371.

[99] Robert Pollice and Peter Chen. A universal quantitative descriptor of the dispersion interaction potential. *Angewandte Chemie International Edition*, 58:9758–9769, 7 2019. ISSN 1433-7851. doi: 10.1002/anie.201905439.

[100] Kevin Wu and Abigail G Doyle. Parameterization of phosphine ligands demonstrates enhancement of nickel catalysis via remote steric effects. *Nature chemistry*, 9(8):779–784, 2017.

[101] Giuseppe Antinucci, Busra Dereli, Antonio Vittoria, Peter H.M. Budzelaar, Roberta Cipullo, Georgy P. Goryunov, Pavel S. Kulyabin, Dmitry V. Uborsky, Luigi Cavallo, Christian Ehm, Alexander Z. Voskoboynikov, and Vincenzo Busico. Selection of low-dimensional 3-d geometric descriptors for accurate enantioselectivity prediction. *ACS Catalysis*, pages 6934–6945, 2022. ISSN 21555435. doi: 10.1021/ACSCATAL.2C00976/SUPPL_FILE/CS2C00976_SI_002.XYZ. URL https://pubs.acs.org/doi/full/10.1021/acscatal.2c00976.

[102] David J Liptrot and Philip P Power. London dispersion forces in sterically crowded inorganic and organometallic molecules. *Nature Reviews Chemistry*, 1(1):0004, 2017.

[103] Sándor Kristyán and Péter Pulay. Can (semi) local density functional theory account for the london dispersion forces? *Chemical physics letters*, 229(3):175–180, 1994.

[104] Anna V Gubskaya and Peter G Kusalik. The total molecular dipole moment for liquid water. *The Journal of chemical physics*, 117(11):5290–5302, 2002.

[105] J. S. Griffith and L. E. Orgel. Ligand-field theory. *Quarterly Reviews, Chemical Society*, 11:381, 1957. ISSN 0009-2681. doi: 10.1039/qr9571100381.

[106] Catherine E Housecroft and Alan G Sharpe. *Inorganic chemistry*, volume 1. Pearson Education, 2008.

[107] Takao Tsuneda, Jong-Won Song, Satoshi Suzuki, and Kimihiko Hirao. On koopmans' theorem in density functional theory. *The Journal of Chemical Physics*, 133:174101, 11 2010. ISSN 0021-9606. doi: 10.1063/1.3491272.

[108] Ji Luo, Zeng Quan Xue, Wei Min Liu, Jin Lei Wu, and Zhong Qin Yang. Koopmans' theorem for large molecular systems within density functional theory. *The Journal of Physical Chemistry A*, 110:12005–12009, 11 2006. ISSN 1089-5639. doi: 10.1021/jp063669m.

[109] Hiroyuki Yoshida, Kazuto Yamada, Jun'ya Tsutsumi, and Naoki Sato. Complete description of ionization energy and electron affinity in organic solids: Determining contributions from electronic polarization, energy band dispersion, and molecular orientation. *Physical Review B*, 92(7):075145, 2015.

[110] Robert G. Parr and Ralph G. Pearson. Absolute hardness: companion parameter to absolute electronegativity. *Journal of the American Chemical Society*, 105:7512–7516, 12 1983. ISSN 0002-7863. doi: 10.1021/ja00364a005.

[111] R. Contreras, J. Andres, V. S. Safont, P. Campodonico, and J. G. Santos. A theoretical study on the relationship between nucleophilicity and ionization potentials in solution phase. *The Journal of Physical Chemistry A*, 107:5588–5593, 7 2003. ISSN 1089-5639. doi: 10.1021/jp0302865.

[112] Chang-Guo Zhan, Jeffrey A Nichols, and David A Dixon. Ionization potential, electron affinity, electronegativity, hardness, and electron excitation energy: molecular properties from density functional theory orbital energies. *The Journal of Physical Chemistry A*, 107(20):4184–4195, 2003.

[113] Paul W. Ayers, James S. M. Anderson, and Libero J. Bartolotti. Perturbative perspectives on the chemical reaction prediction problem. *International Journal of Quantum Chemistry*, 101:520–534, 2005. ISSN 0020-7608. doi: 10.1002/qua.20307.

[114] Paola R. Campodónico, Claudio Pérez, Margarita Aliaga, Marcela Gazitúa, and Renato Contreras. Electrofugality index for benzhydryl derivatives. *Chemical Physics Letters*, 447:375–378, 10 2007. ISSN 00092614. doi: 10.1016/j.cplett.2007.09.042.

[115] Robert S Mulliken. A new electroaffinity scale; together with data on valence states and on valence ionization potentials and electron affinities. *The Journal of Chemical Physics*, 2(11):782–793, 1934.

[116] Ralph G Pearson. The hsab principle—more quantitative aspects. *Inorganica Chimica Acta*, 240(1-2): 93–98, 1995.

[117] Ralph Pearson. Chemical hardness and density functional theory. *Journal of Chemical Sciences*, 117: 369–377, 09 2005. doi: 10.1007/BF02708340.

[118] Frank Weinhold and Clark R Landis. *Valency and bonding: a natural bond orbital donor-acceptor perspective*. Cambridge University Press, 2005.

[119] Oliviero Carugo. How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared. *Journal of Applied Crystallography*, 36:125–128, 2 2003. ISSN 0021-8898. doi: 10.1107/S0021889802020502.

[120] J. C. Kromann. Calculate root-mean-square deviation (rmsd) of two molecules using rotation, github, v1.3.2, 2020. URL http://github.com/charnley/rmsd.

[121] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32:922–923, 9 1976. ISSN 0567-7394. doi: 10.1107/S0567739476001873.

[122] Delft High Performance Computing Centre (DHPC). DelftBlue Supercomputer (Phase 1). https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase1, 2022.

[123] National Supercomputer Centre Linköping University & Swedish National Infrastructure for Computing NSC-LIU. Tetralith national supercomputer. https://www.nsc.liu.se/systems/tetralith/, 2023.

[124] Snellius: de Nationale Supercomputer. Snellius: de nationale supercomputer. https://www.surf.nl/en/dutch-national-supercomputer-snellius, 2023.

[125] Adarsh V. Kalikadien, Evgeny A. Pidko, and Vivek Sinha. <i>chemspax</i> : exploration of chemical space by automated functionalization of molecular scaffold. *Digital Discovery*, 1:8–25, 2022. ISSN 2635-098X. doi: 10.1039/D1DD00017A.

[126] 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 (<i>z</i> = 1–86). *Journal of Chemical Theory and Computation*, 13:1989–2009, 5 2017. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b00118.

[127] 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. Gaussian Inc. Wallingford CT.

[128] Annika M. Krieger, Vivek Sinha, Adarsh V. Kalikadien, and Evgeny A. Pidko. Metal-ligand cooperative activation of hx (x=h, br, or) bond on mn based pincer complexes. *Zeitschrift für anorganische und allgemeine Chemie*, 647:1486–1494, 7 2021. ISSN 0044-2313. doi: 10.1002/zaac.202100078.

[129] Yury Minenkov, Åsmund Singstad, Giovanni Occhipinti, and Vidar R. Jensen. The accuracy of dft-optimized geometries of functional transition metal compounds: a validation study of catalysts for olefin metathesis and other reactions in the homogeneous phase. *Dalton Transactions*, 41:5526, 2012. ISSN 1477-9226. doi: 10.1039/c2dt12232d.

[130] Tobias Gensch, Gabriel Dos Passos Gomes, Pascal Friederich, Ellyn Peters, Théophile Gaudin, Robert Pollice, Kjell Jorner, Akshatkumar Nigam, Michael Lindner-D'Addario, Matthew S. Sigman, and Alán Aspuru-Guzik. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *Journal of the American Chemical Society*, 144(3):1205–1217, January 2022. ISSN 15205126. doi: 10.1021/jacs.1c09718.

[131] Jordan J Dotson, Lucy van Dijk, Jacob C Timmerman, Samantha Grosslight, Richard C Walroth, Francis Gosselin, Kurt Püntener, Kyle A Mack, and Matthew S Sigman. Data-driven multi-objective optimization tactics for catalytic asymmetric reactions. *T.B.D.*, 2022.

[132] Gabriel dos Passos Gomes, Robert Pollice, and Alán Aspuru-Guzik. Navigating through the maze of homogeneous catalyst design with machine learning. *Trends in Chemistry*, 3:96–110, 2021. ISSN 2589-5974. doi: https://doi.org/10.1016/j.trechm.2020.12.006. URL https://www.sciencedirect.com/science/article/pii/S2589597420303166.

[133] Rubén Laplaza, Jan Grimo Sobez, Matthew D. Wodrich, Markus Reiher, and Clémence Corminboeuf. The (not so) simple prediction of enantioselectivity - a pipeline for high-fidelity computations. *Chemical Science*, 13(23):6858–6864, may 2022. ISSN 20416539. doi: 10.1039/d2sc01714h.

[134] Christopher J. Cramer and Donald G. Truhlar. Density functional theory for transition metals and transition metal chemistry. *Physical Chemistry Chemical Physics*, 11:10757, 2009. ISSN 1463-9076. doi: 10.1039/b907148b.

[135] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry*, 32(7):1456–1465, May 2011. ISSN 01928651. doi: 10.1002/jcc.21759.

[136] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Physical Review Letters*, 77(18):3865–3868, October 1996. ISSN 0031-9007. doi: 10.1103/PhysRevLett.77.3865.

[137] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple [phys. rev. lett. 77, 3865 (1996)]. *Physical Review Letters*, 78:1396–1396, 2 1997. ISSN 0031-9007. doi: 10.1103/PhysRevLett.78.1396.

[138] Matthias Ernzerhof and Gustavo E. Scuseria. Assessment of the perdew–burke–ernzerhof exchange-correlation functional. *The Journal of Chemical Physics*, 110:5029–5036, 3 1999. ISSN 0021-9606. doi: 10.1063/1.478401.

[139] 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, April 1999. ISSN 0021-9606. doi: 10.1063/1.478522.

[140] Vivek Sinha, Jochem J Laan, and Evgeny A Pidko. Accurate and rapid prediction of p k a of transition metal complexes: semiempirical quantum chemistry with a data-augmented approach. *Physical Chemistry Chemical Physics*, 23(4):2557–2567, 2021.

[141] Michael Bühl, Christoph Reimann, Dimitrios A. Pantazis, Thomas Bredow, and Frank Neese. Geometries of third-row transition-metal complexes from density-functional theory. *Journal of Chemical Theory and Computation*, 4:1449–1459, 9 2008. ISSN 1549-9618. doi: 10.1021/ct800172j.

[142] Kasper P. Jensen, Björn O. Roos, and Ulf Ryde. Performance of density functionals for first row transition metal systems. *The Journal of Chemical Physics*, 126:014103, 1 2007. ISSN 0021-9606. doi: 10.1063/1.2406071.

[143] Kjell Jorner and Lukas Turcani. kjelljorner/morfeus: v0.7.2, August 2022. URL https://doi.org/10.5281/zenodo.7017599.

[144] William M Haynes. *CRC handbook of chemistry and physics*. CRC press, 2016.

[145] A. Shrake and J.A. Rupley. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of Molecular Biology*, 79:351–371, 9 1973. ISSN 00222836. doi: 10.1016/0022-2836(73)90011-9.

[146] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28:31–36, 2 1988. ISSN 1549-9596. doi: 10.1021/ci00057a005.

[147] Mahesh Sundararajan, Vivek Sinha, Tusar Bandyopadhyay, and Swapan K Ghosh. Can functionalized cucurbituril bind actinyl cations efficiently? a density functional theory based investigation. *The Journal of Physical Chemistry A*, 116(17):4388–4395, 2012.

[148] Krishna Rao, 2020. URL https://towardsdatascience.com/r%C2%B2-or-r%C2%B2-when-to-use-what-4968eee68ed3#:~:text=3.,the%20strength%20of%20a%20model.

[149] Nico JD Nagelkerke et al. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692, 1991.

[150] Mark Heezen, E A Pidko, and A V Kalikadien. Optimisation of in silico techniques for homogeneous transition metal based catalysis research. Bachelor's thesis, Delft University of Technology, July 2022.

[151] 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*, 11, 3 2021. ISSN 1759-0876. doi: 10.1002/wcms.1493.

[152] Markus Bursch, Andreas Hansen, and Stefan Grimme. Fast and reasonable geometry optimization of lanthanoid complexes with an extended tight binding quantum chemical method. *Inorganic Chemistry*, 56:12485–12491, 10 2017. ISSN 0020-1669. doi: 10.1021/acs.inorgchem.7b01950.

[153] David Balcells and Bastian Bjerkem Skjelstad. tmqm dataset—quantum geometries and properties of 86k transition metal complexes. *Journal of Chemical Information and Modeling*, 60(12):6135–6146, 2020. doi: 10.1021/acs.jcim.0c01041. URL https://doi.org/10.1021/acs.jcim.0c01041. PMID: 33166143.

Appendices contents

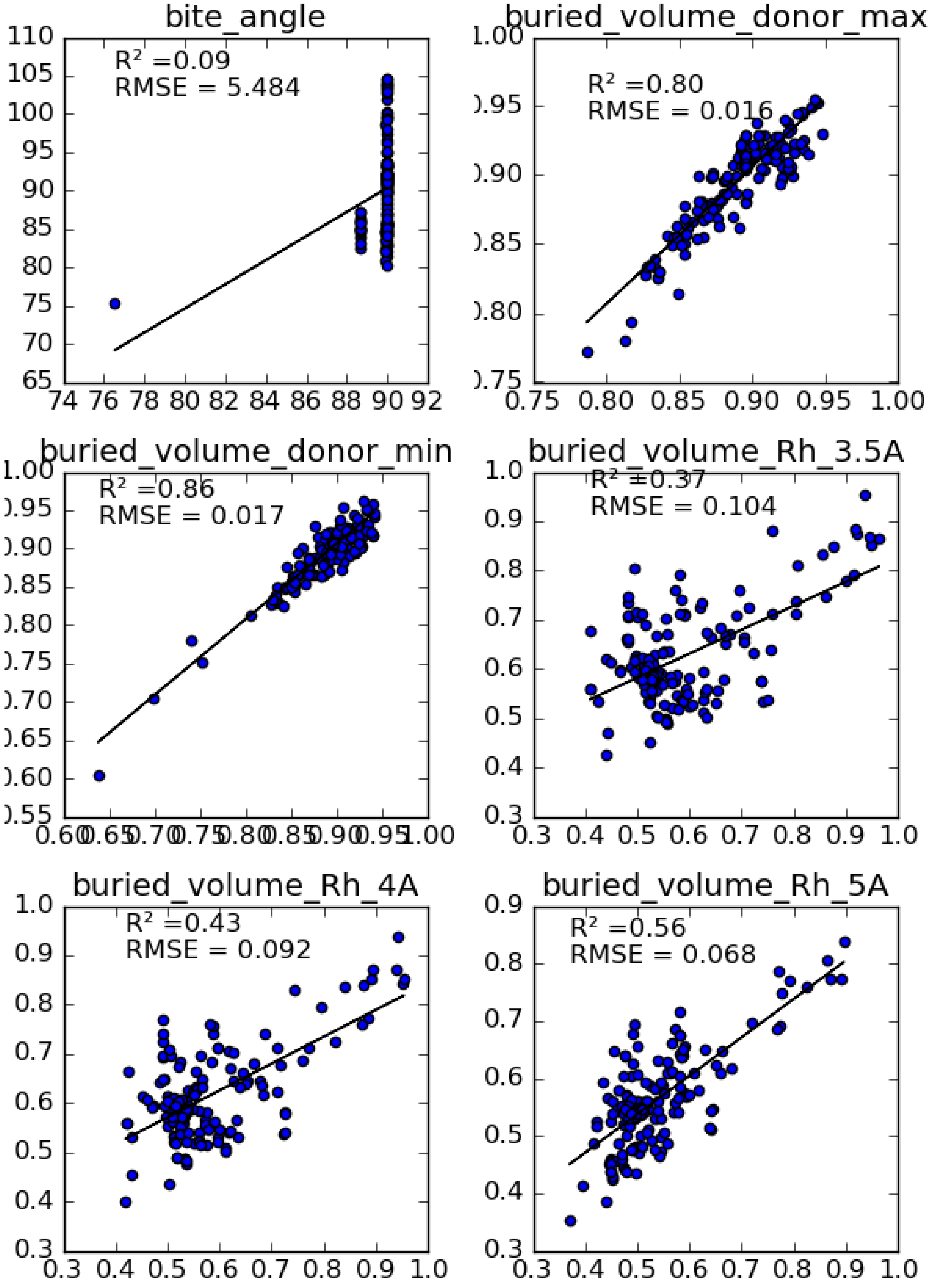
A

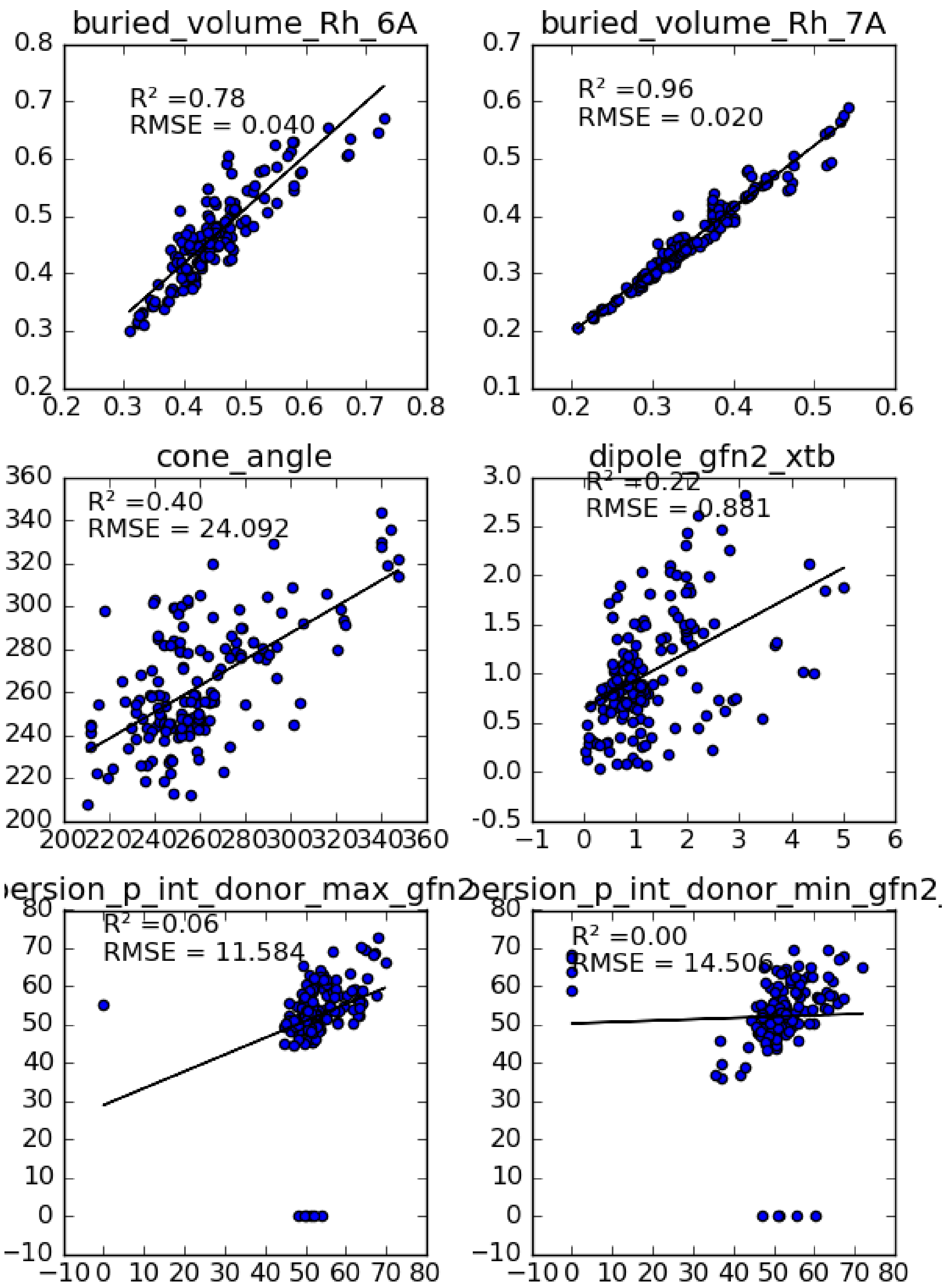
Supporting plots of Rh (I) bidentate complexes

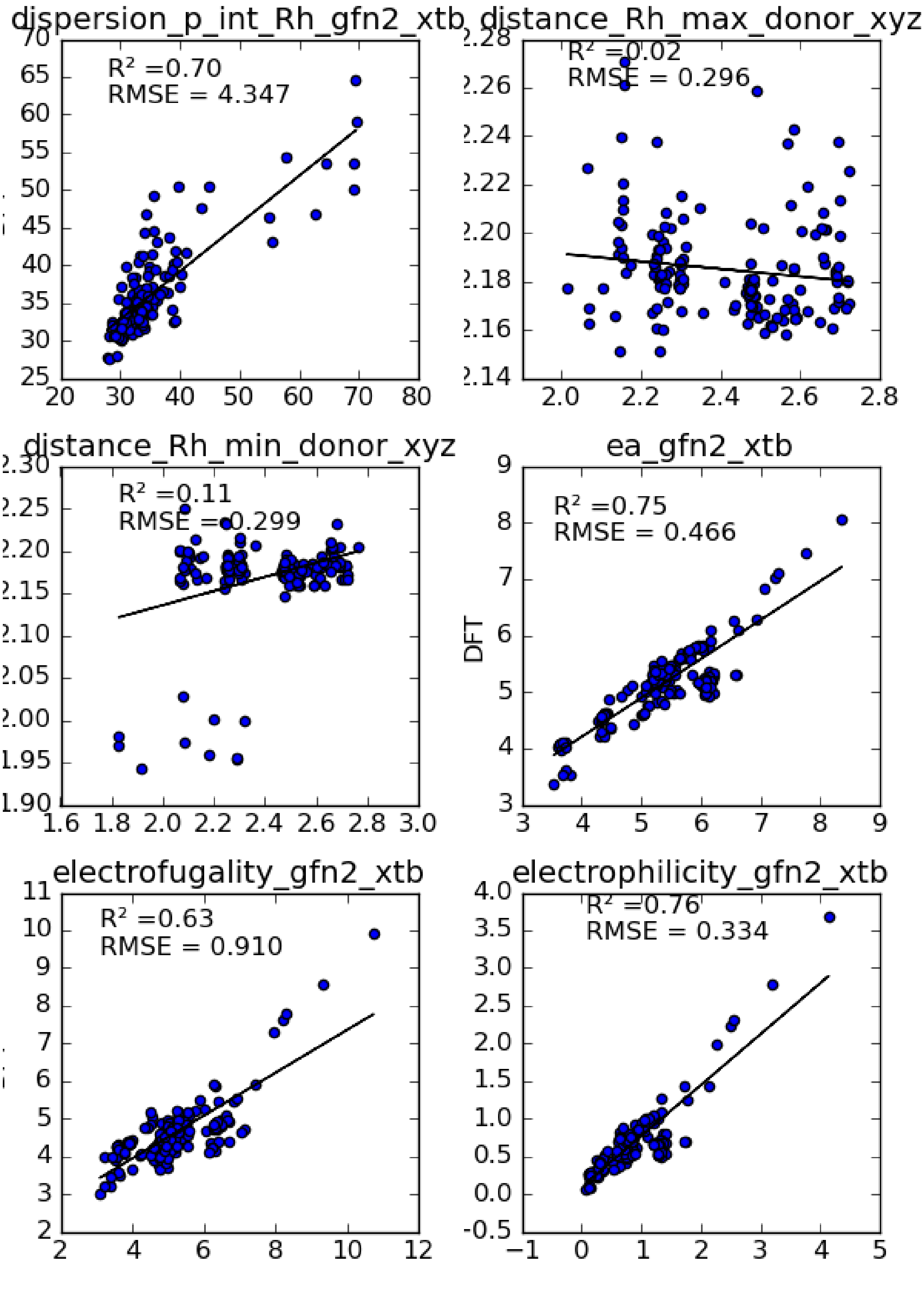
A.1. Detailed regression plots

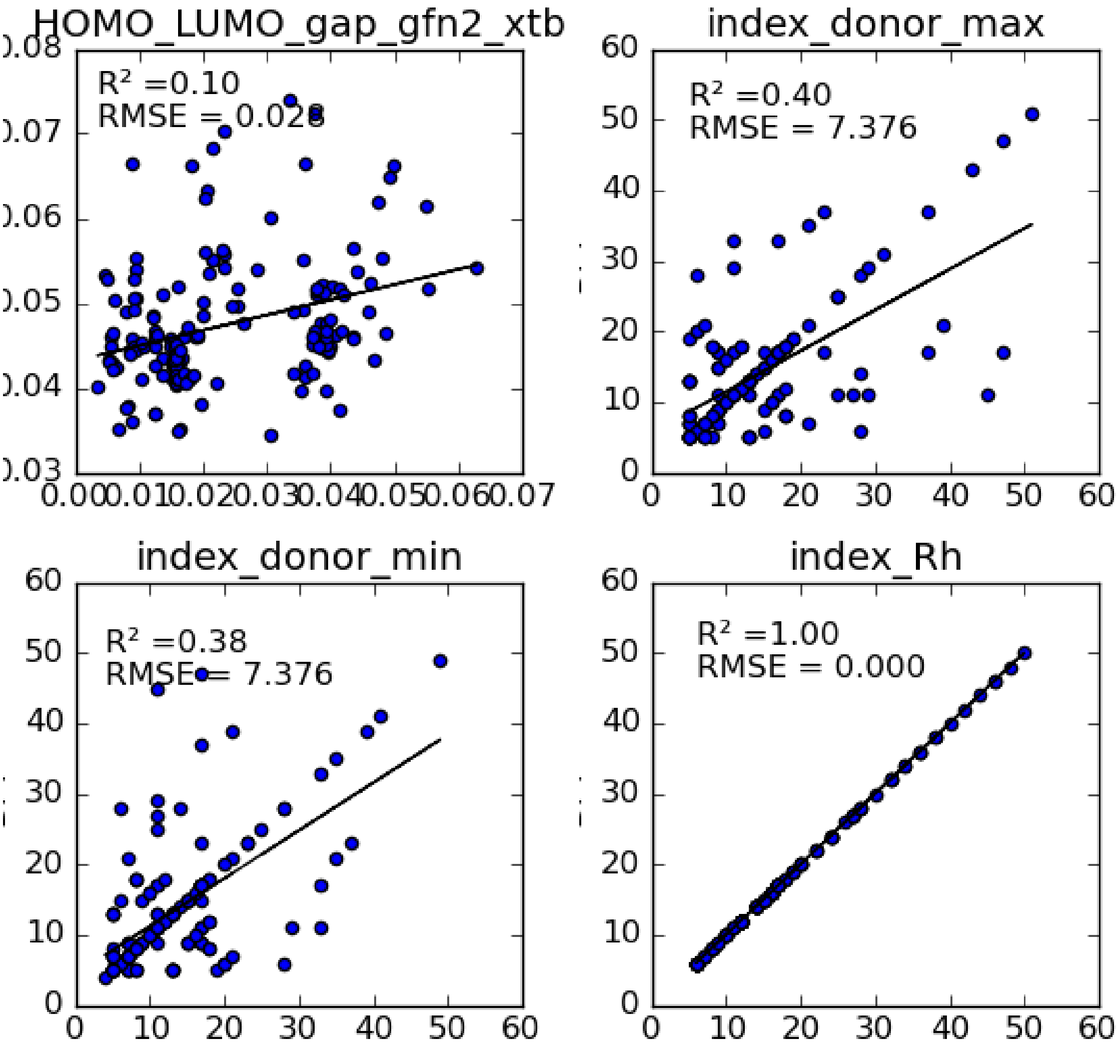
Regression plots in the form of scatter matrices of all descriptors calculated for pristine Rhodium complexes.

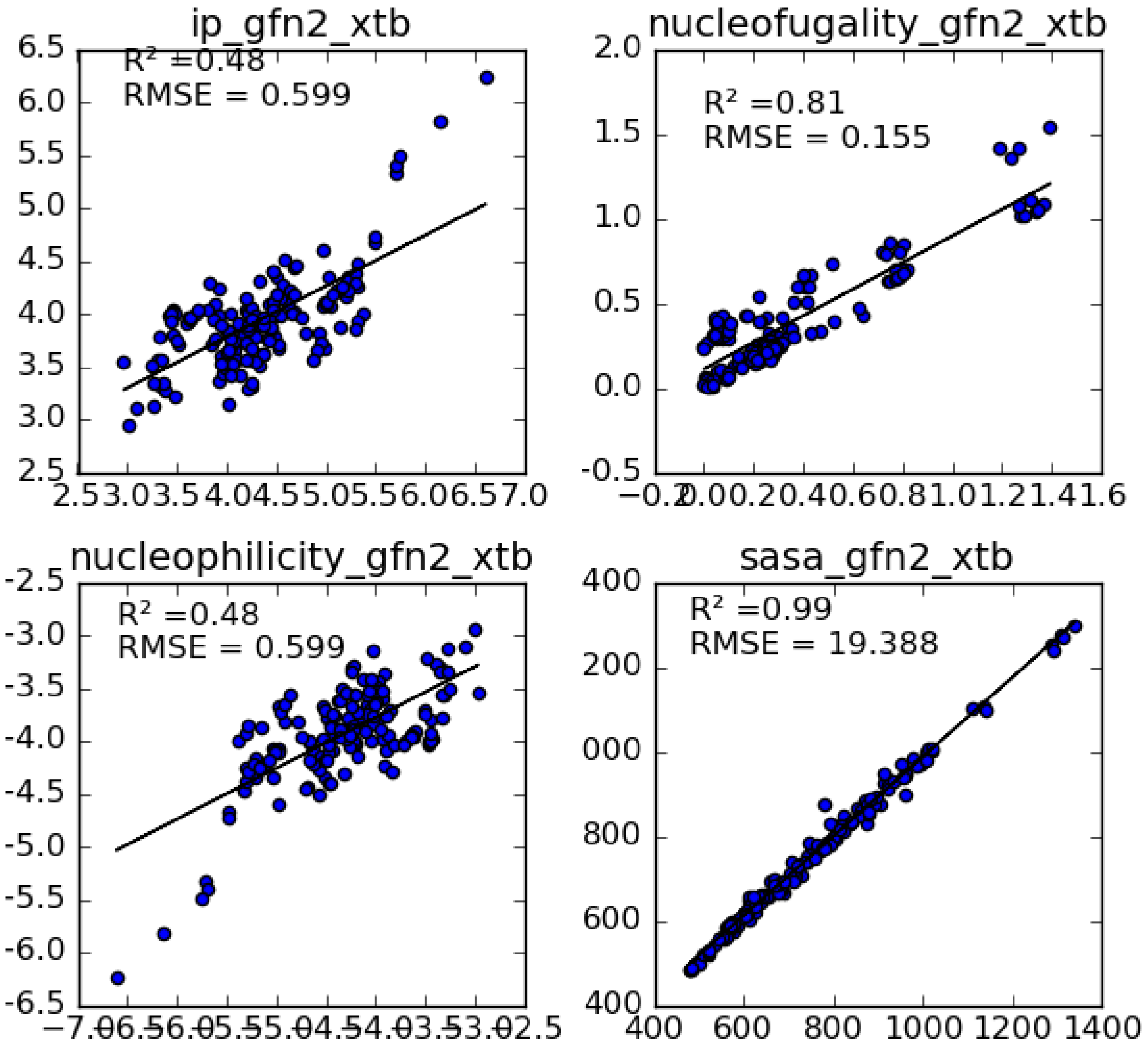
MACE unfiltered



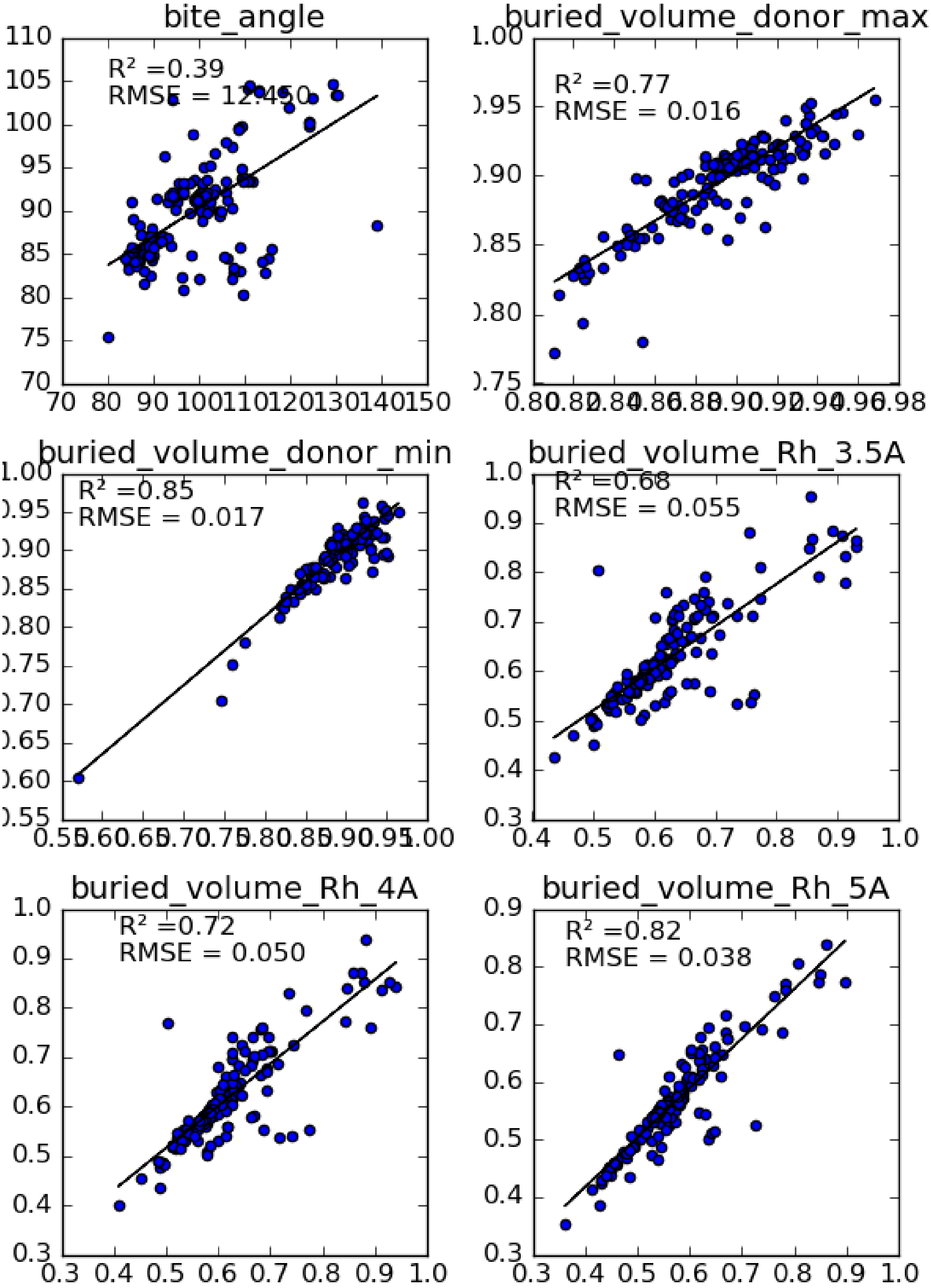

buried_volume_Rh_6A
R² =0.78
RMSE = 0.040
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.2 0.3 0.4 0.5 0.6 0.7 0.8
buried_volume_Rh_7A
R² =0.96
RMSE = 0.020
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.2 0.3 0.4 0.5 0.6
cone_angle
R² =0.40
RMSE = 24.092
360
340
320
300
280
260
240
220
200
200220240260280300320340360
dipole_gfn2_xtb
R² =0.22
RMSE = 0.881
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
−1 0 1 2 3 4 5 6
ersion_p_int_donor_max_gfn2
R² =0.06
RMSE = 11.584
80
70
60
50
40
30
20
10
0
−10
−10 0 10 20 30 40 50 60 70 80
ersion_p_int_donor_min_gfn2_
R² =0.00
RMSE = 14.506
80
70
60
50
40
30
20
10
0
−10
−10 0 10 20 30 40 50 60 70 80

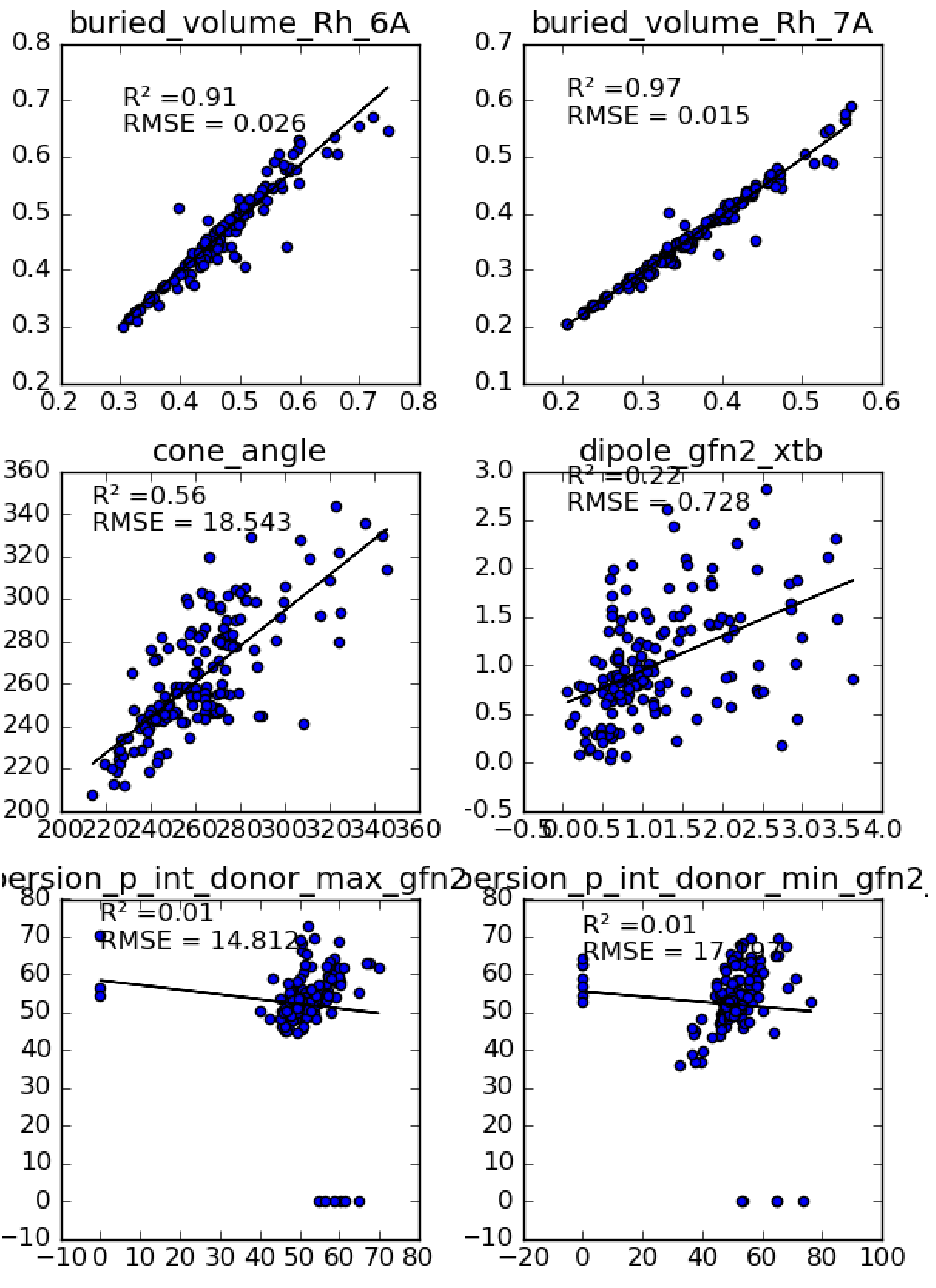

dispersion_p_int_Rh_gfn2_xtb
R² =0.70
RMSE = 4.347
25
30
35
40
45
50
55
60
65
70
20
30
40
50
60
70
80
distance_Rh_max_donor_xyz
R² =0.02
RMSE = 0.296
2.14
2.16
2.18
2.20
2.22
2.24
2.26
2.28
2.0
2.2
2.4
2.6
2.8
distance_Rh_min_donor_xyz
R² =0.11
RMSE = 0.299
1.90
1.95
2.00
2.05
2.10
2.15
2.20
2.25
2.30
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
ea_gfn2_xtb
R² =0.75
RMSE = 0.466
DFT
3
4
5
6
7
8
9
3
4
5
6
7
8
9
electrofugality_gfn2_xtb
R² =0.63
RMSE = 0.910
2
3
4
5
6
7
8
9
10
11
2
4
6
8
10
12
electrophilicity_gfn2_xtb
R² =0.76
RMSE = 0.334
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
−1
0
1
2
3
4
5

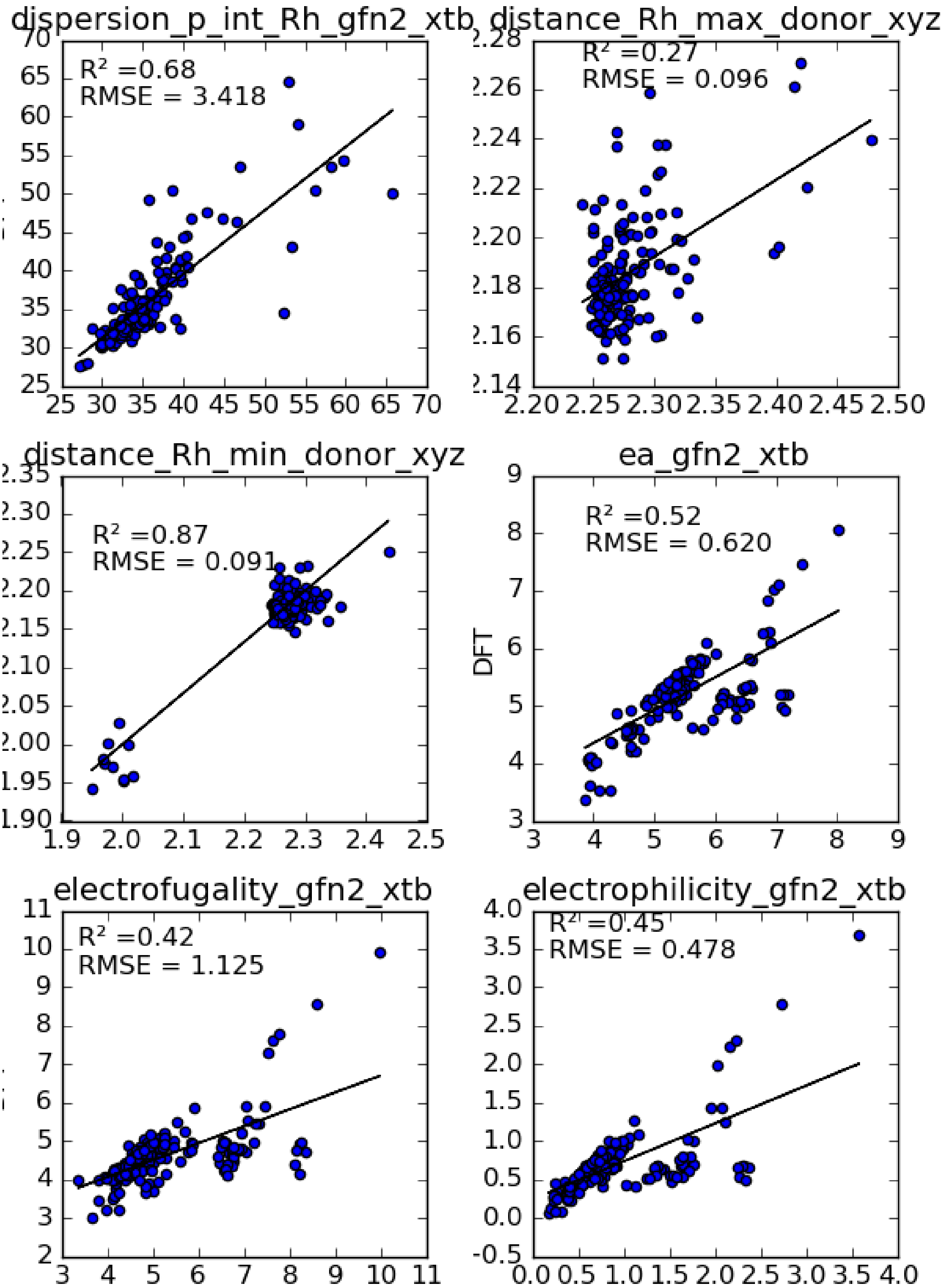

HOMO_LUMO_gap_gfn2_xtb
R² =0.10
RMSE = 0.028
0.08
0.07
0.06
0.05
0.04
0.03
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
index_donor_max
R² =0.40
RMSE = 7.376
0
10
20
30
40
50
60
index_donor_min
R² =0.38
RMSE = 7.376
0
10
20
30
40
50
60
index_Rh
R² =1.00
RMSE = 0.000
0
10
20
30
40
50
60

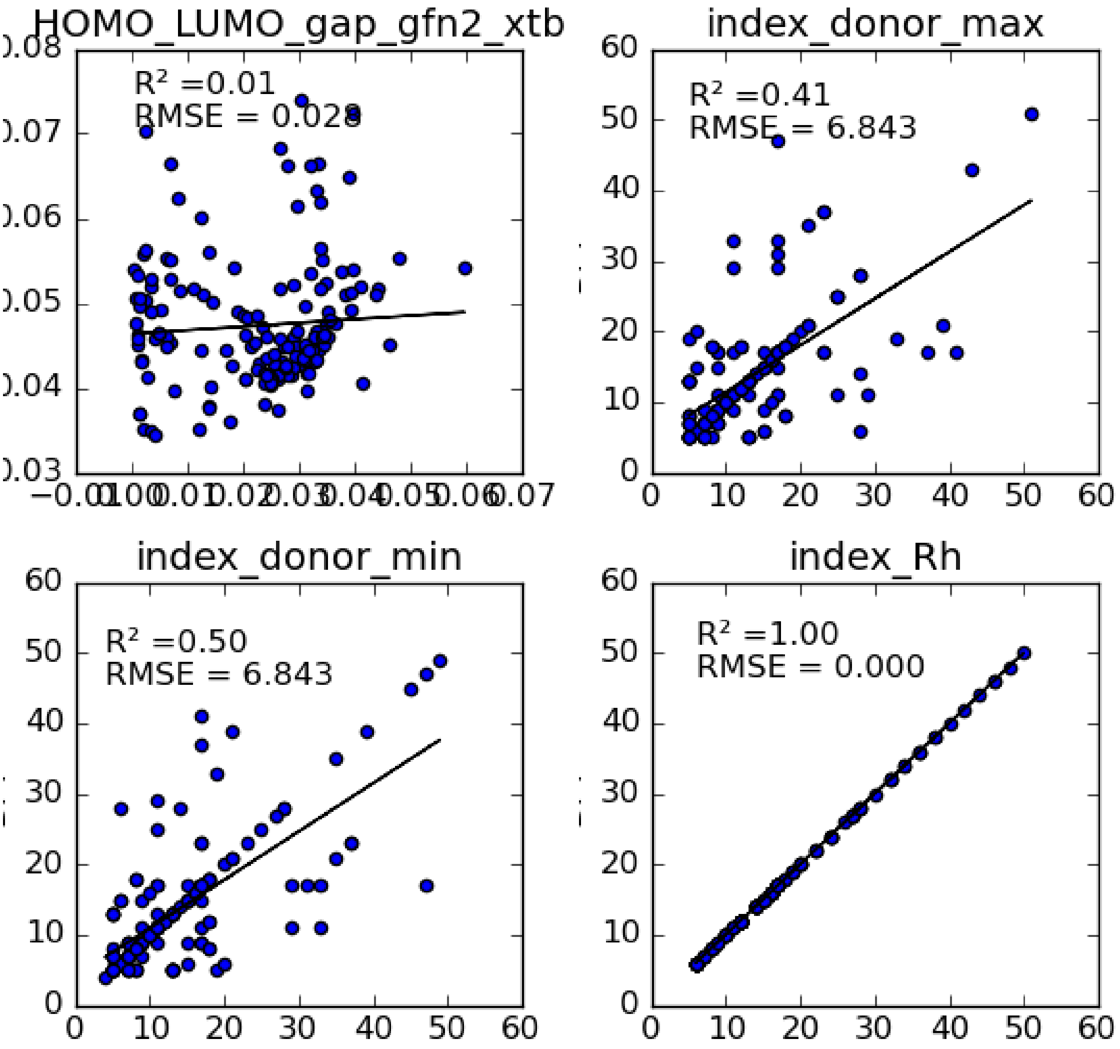

ip_gfn2_xtb
R² =0.48
RMSE = 0.599
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.53.03.54.04.55.05.56.06.57.0
nucleofugality_gfn2_xtb
R² =0.81
RMSE = 0.155
2.0
1.5
1.0
0.5
0.0
-0.5
−0.2
0.00.20.40.60.81.01.21.41.6
nucleophilicity_gfn2_xtb
R² =0.48
RMSE = 0.599
-2.5
-3.0
-3.5
-4.0
-4.5
-5.0
-5.5
-6.0
-6.5
−7.0−6.5−6.0−5.5−5.0−4.5−4.0−3.5−3.0−2.5
sasa_gfn2_xtb
R² =0.99
RMSE = 19.388
400
200
000
800
600
400
400 600 800 1000 1200 1400

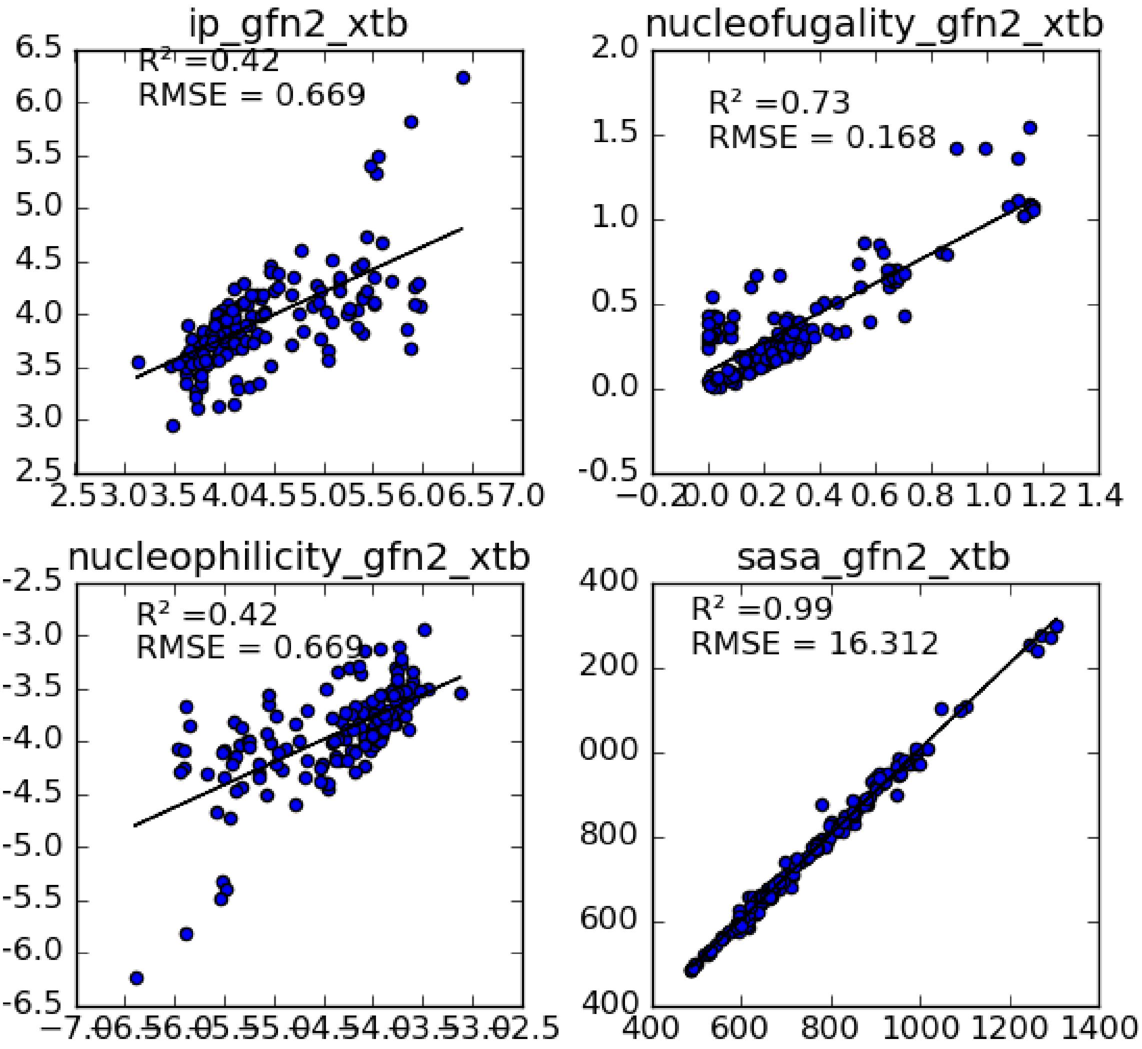
GFN-FF unfiltered



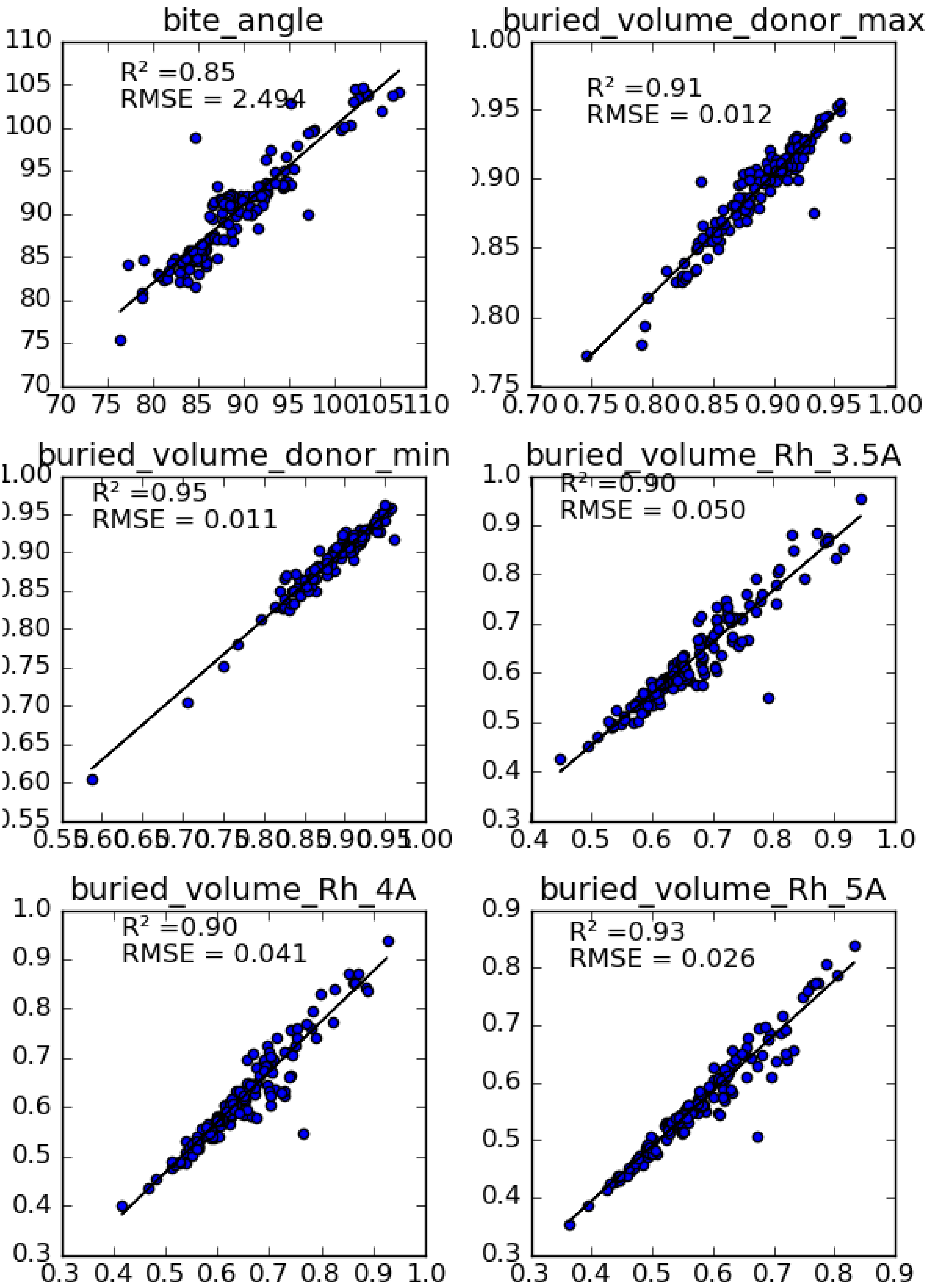

buried_volume_Rh_6A
R² =0.91
RMSE = 0.026
0.2
0.3
0.4
0.5
0.6
0.7
0.8
buried_volume_Rh_7A
R² =0.97
RMSE = 0.015
0.1
0.2
0.3
0.4
0.5
0.6
0.7
cone_angle
R² =0.56
RMSE = 18.543
200
220
240
260
280
300
320
340
360
dipole_gfn2_xtb
R² =0.22
RMSE = 0.728
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
ersion_p_int_donor_max_gfn2
R² =0.01
RMSE = 14.812
−10
0
10
20
30
40
50
60
70
80
ersion_p_int_donor_min_gfn2
R² =0.01
RMSE = 17
−20
0
20
40
60
80
100

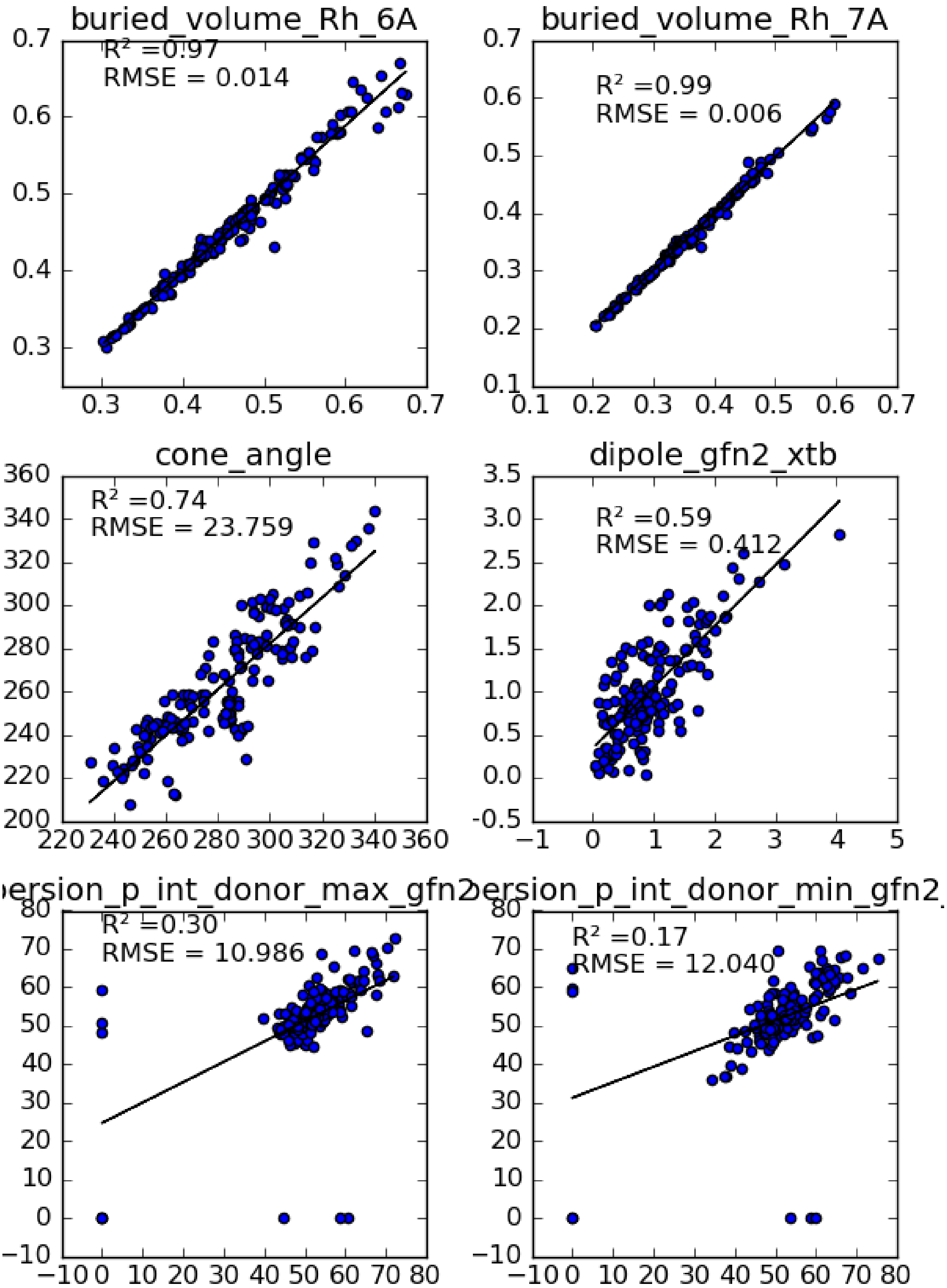

dispersion_p_int_Rh_gfn2_xtb
R² =0.68
RMSE = 3.418
25 30 35 40 45 50 55 60 65 70
25 30 35 40 45 50 55 60 65 70
distance_Rh_max_donor_xyz
R² =0.27
RMSE = 0.096
2.20 2.25 2.30 2.35 2.40 2.45 2.50
2.14 2.16 2.18 2.20 2.22 2.24 2.26 2.28
distance_Rh_min_donor_xyz
R² =0.87
RMSE = 0.091
1.9 2.0 2.1 2.2 2.3 2.4 2.5
1.90 1.95 2.00 2.05 2.10 2.15 2.20 2.25 2.30 2.35
ea_gfn2_xtb
R² =0.52
RMSE = 0.620
DFT
3 4 5 6 7 8 9
3 4 5 6 7 8 9
electrofugality_gfn2_xtb
R² =0.42
RMSE = 1.125
3 4 5 6 7 8 9 10 11
2 3 4 5 6 7 8 9 10 11
electrophilicity_gfn2_xtb
R² =0.45
RMSE = 0.478
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
-0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0

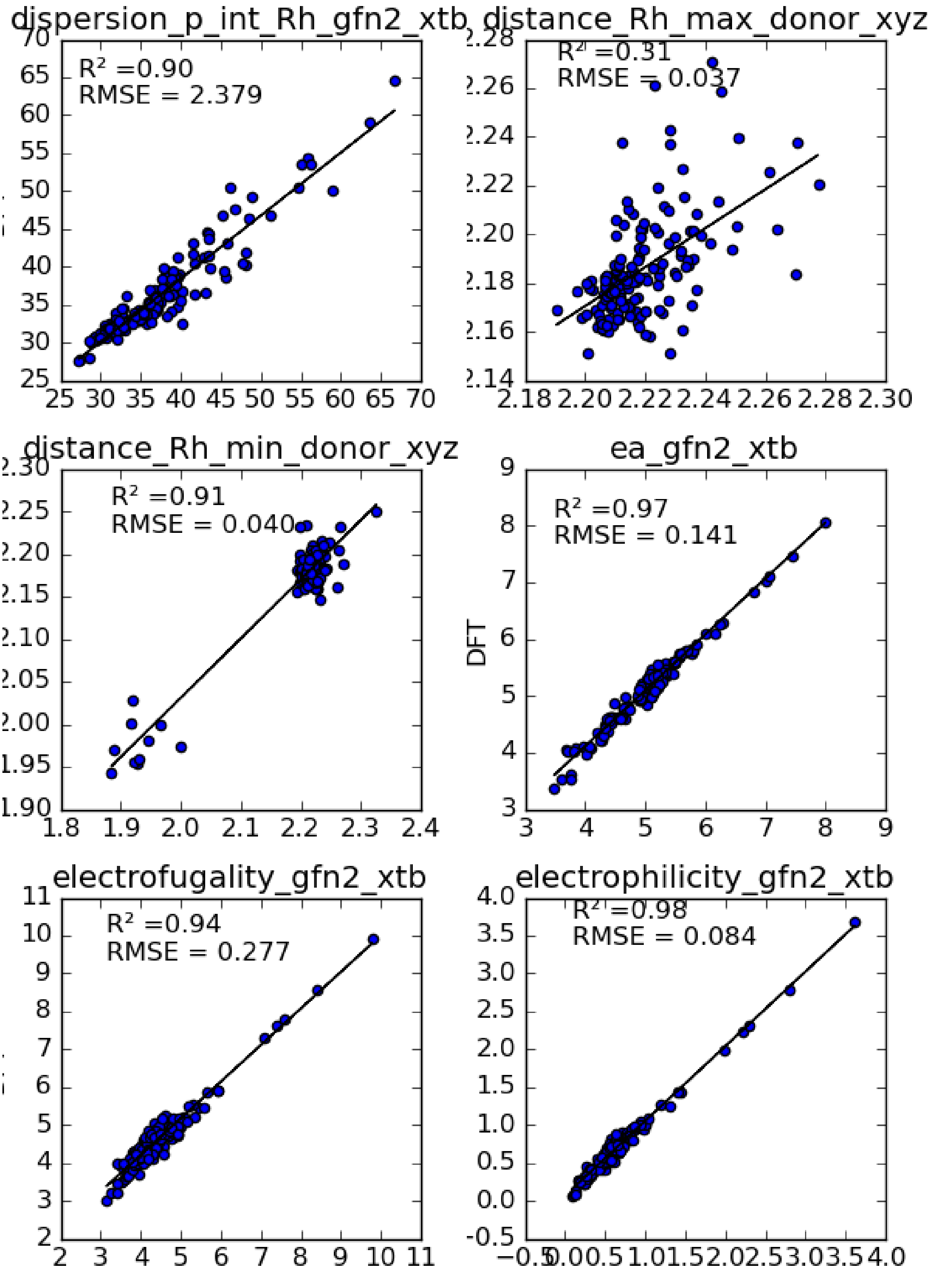

HOMO_LUMO_gap_gfn2_xtb
R² =0.01
RMSE = 0.028
0.08
0.07
0.06
0.05
0.04
0.03
−0.01
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
index_donor_max
R² =0.41
RMSE = 6.843
0
10
20
30
40
50
60
index_donor_min
R² =0.50
RMSE = 6.843
0
10
20
30
40
50
60
index_Rh
R² =1.00
RMSE = 0.000
0
10
20
30
40
50
60

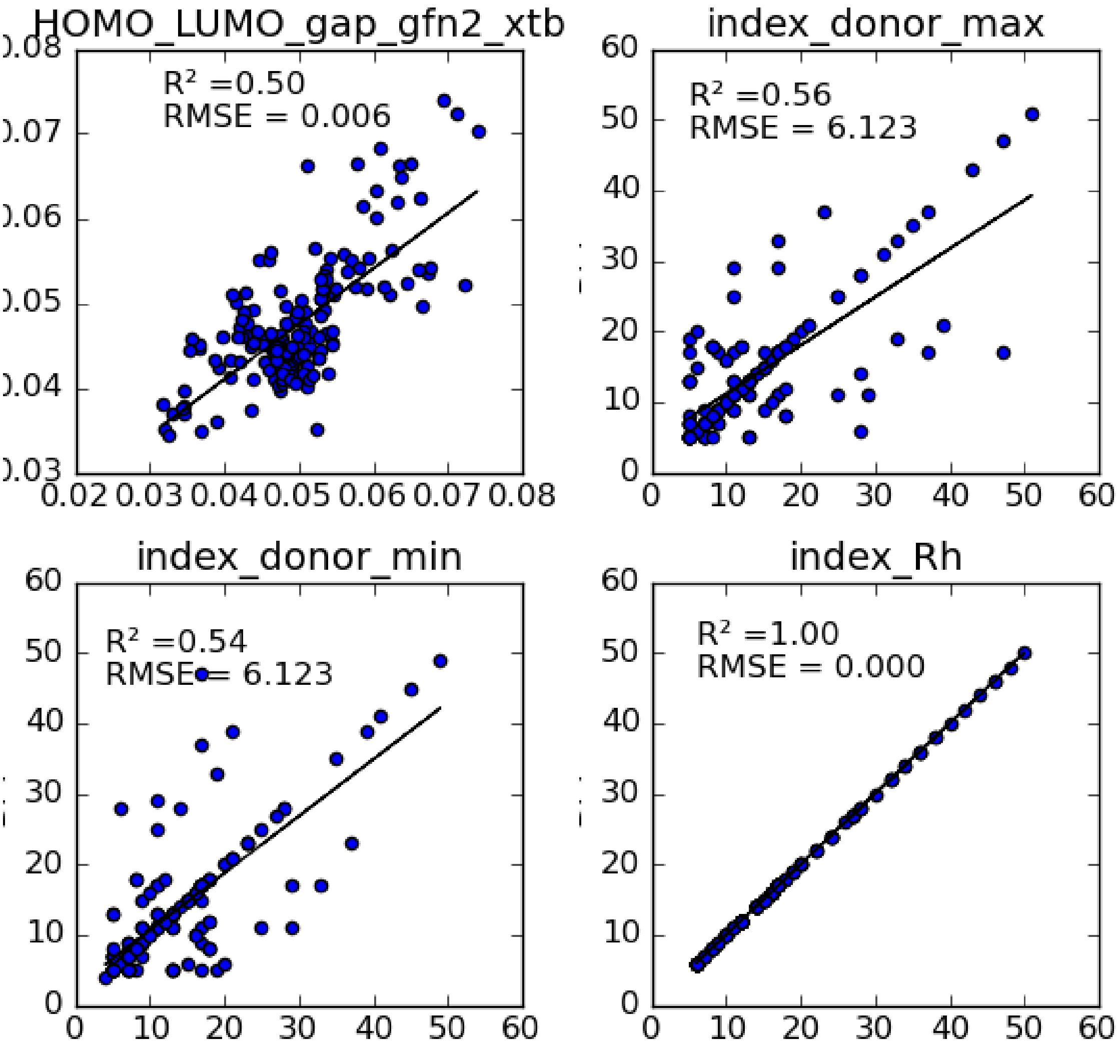

ip_gfn2_xtb
R² =0.42
RMSE = 0.669
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.53.03.54.04.55.05.56.06.57.0
nucleofugality_gfn2_xtb
R² =0.73
RMSE = 0.168
2.0
1.5
1.0
0.5
0.0
-0.5
−0.20.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
nucleophilicity_gfn2_xtb
R² =0.42
RMSE = 0.669
-2.5
-3.0
-3.5
-4.0
-4.5
-5.0
-5.5
-6.0
-6.5
−7.06.56.05.55.04.54.03.53.02.5
sasa_gfn2_xtb
R² =0.99
RMSE = 16.312
400
200
000
800
600
400
400 600 800 1000 1200 1400

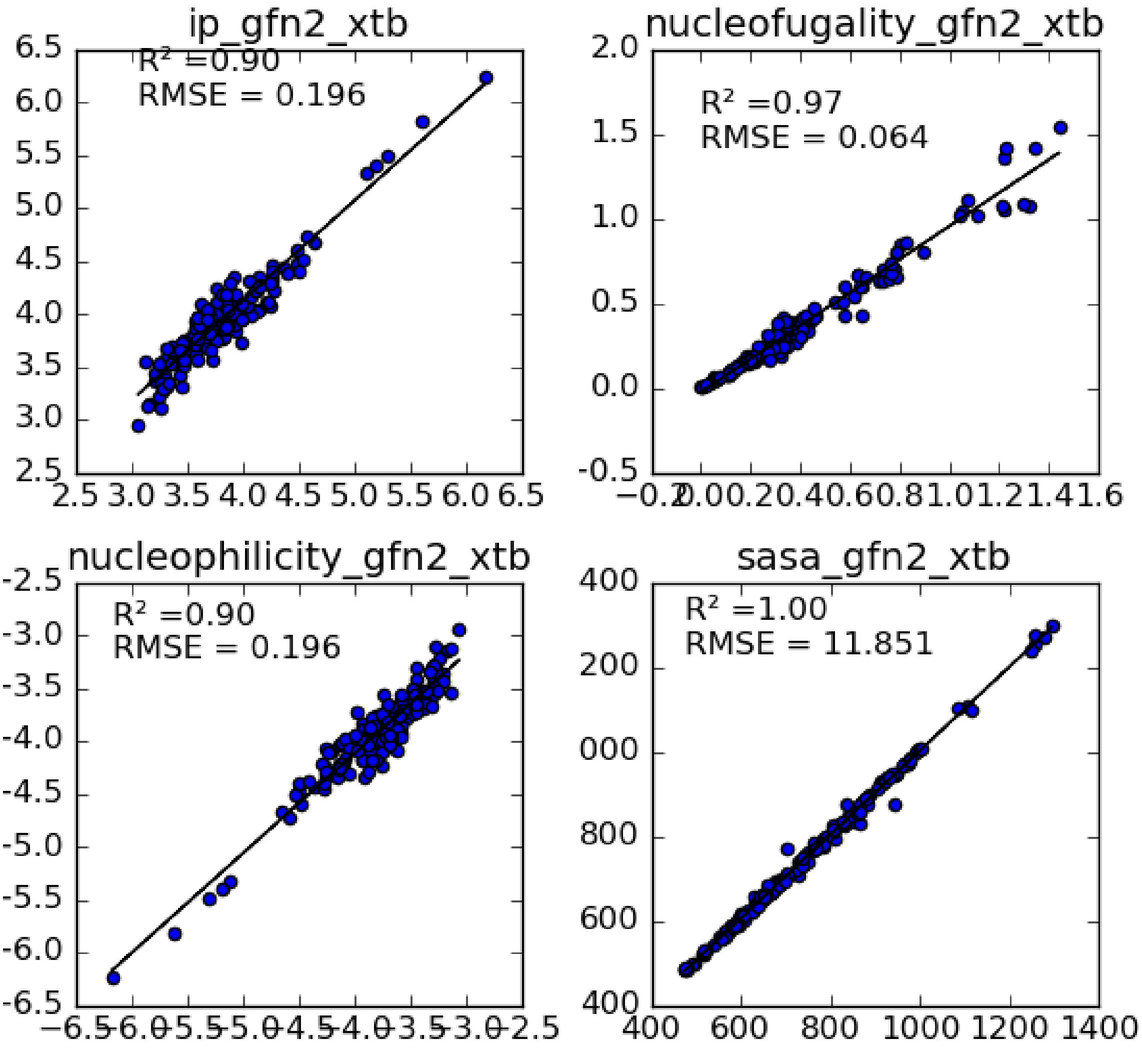
GFN2 unfiltered



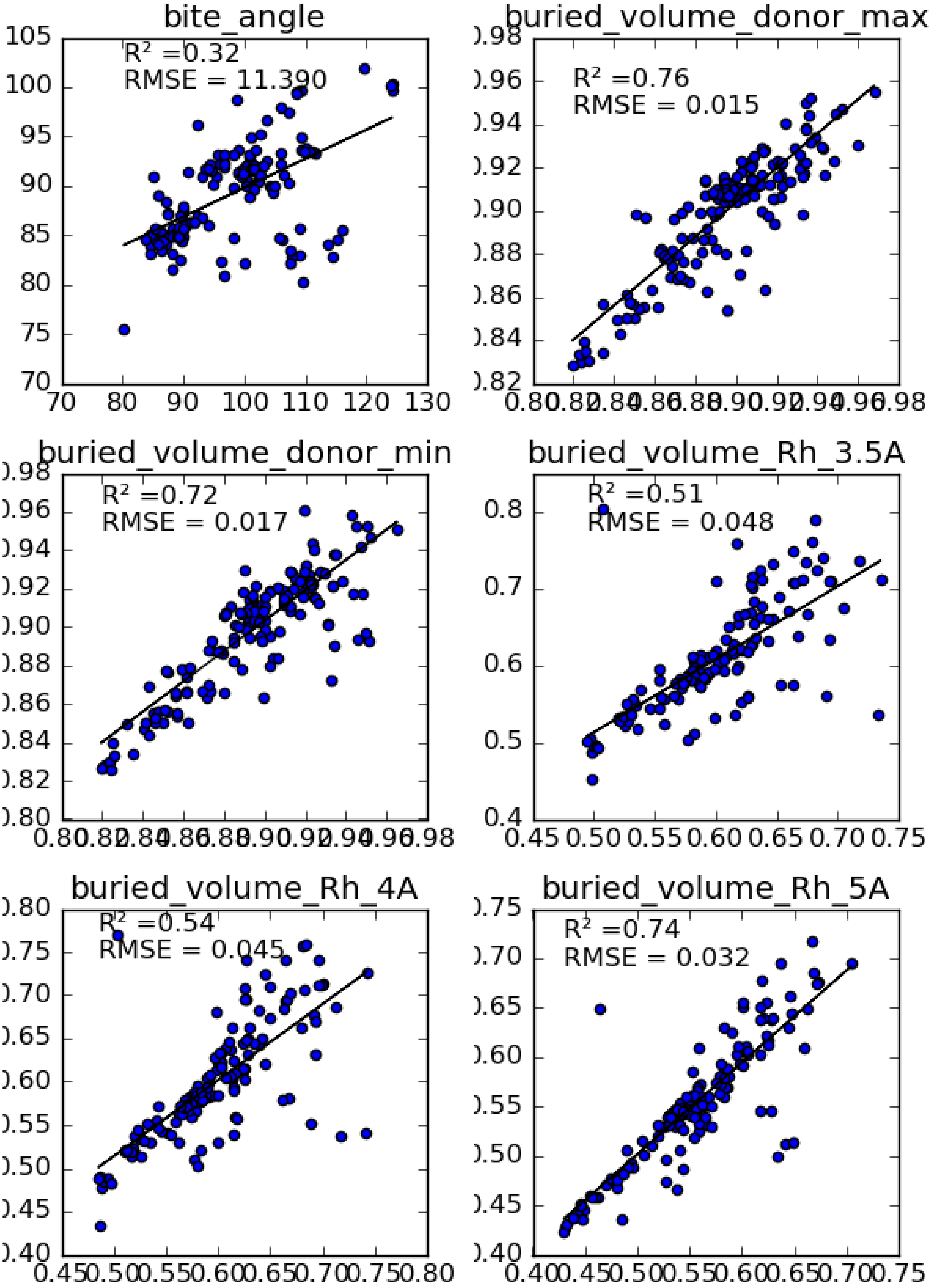

buried_volume_Rh_6A
R² =0.97
RMSE = 0.014
0.3
0.4
0.5
0.6
0.7
buried_volume_Rh_7A
R² =0.99
RMSE = 0.006
0.1
0.2
0.3
0.4
0.5
0.6
0.7
cone_angle
R² =0.74
RMSE = 23.759
200
220
240
260
280
300
320
340
360
dipole_gfn2_xtb
R² =0.59
RMSE = 0.412
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
−1
0
1
2
3
4
5
ersion_p_int_donor_max_gfn2
R² =0.30
RMSE = 10.986
ersion_p_int_donor_min_gfn2
R² =0.17
RMSE = 12.040
−10
0
10
20
30
40
50
60
70
80

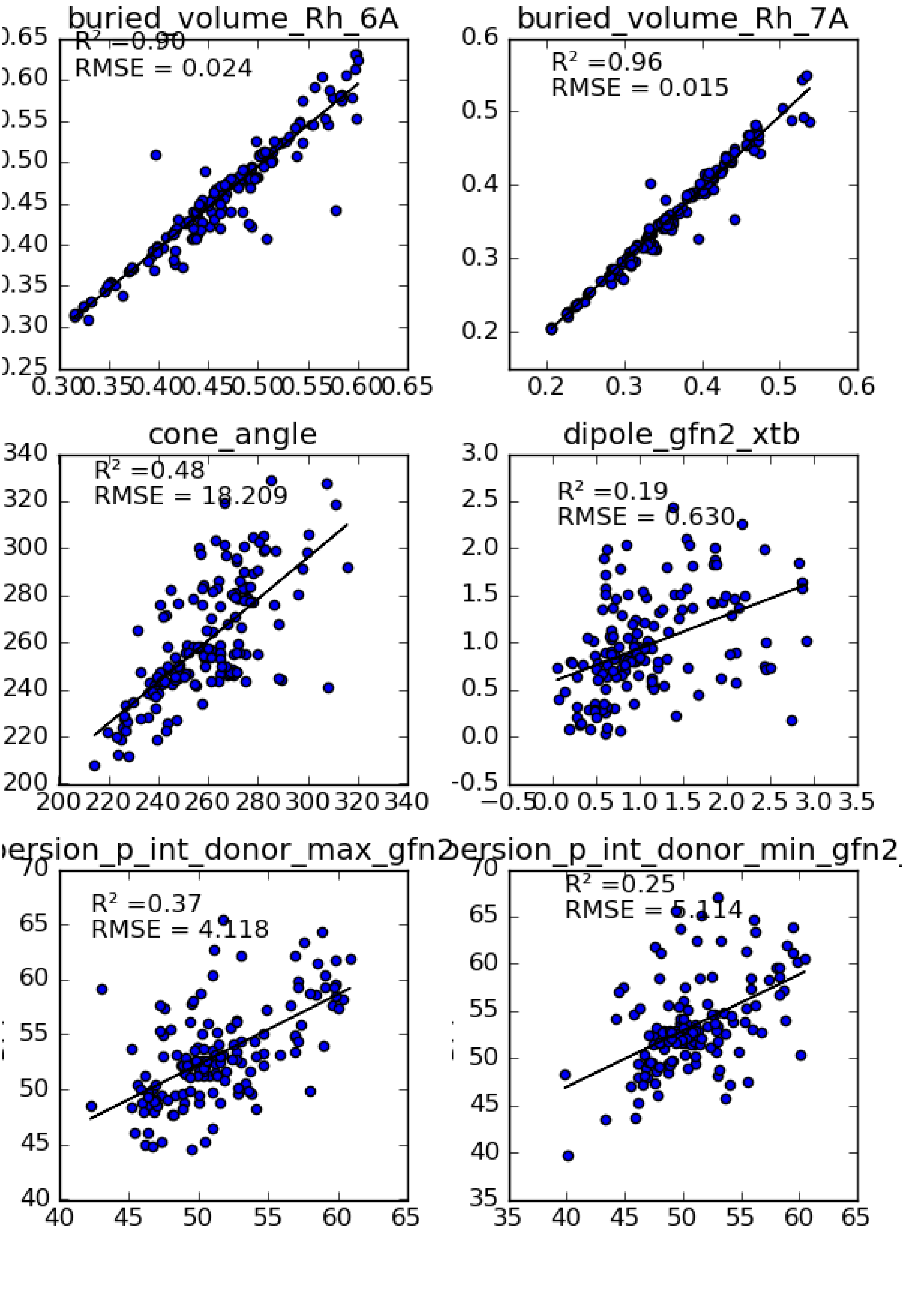

dispersion_p_int_Rh_gfn2_xtb
R² =0.90
RMSE = 2.379
70
65
60
55
50
45
40
35
30
25
25 30 35 40 45 50 55 60 65 70
distance_Rh_max_donor_xyz
R² =0.31
RMSE = 0.037
2.28
2.26
2.24
2.22
2.20
2.18
2.16
2.14
2.18 2.20 2.22 2.24 2.26 2.28 2.30
distance_Rh_min_donor_xyz
R² =0.91
RMSE = 0.040
2.30
2.25
2.20
2.15
2.10
2.05
2.00
1.95
1.90
1.8 1.9 2.0 2.1 2.2 2.3 2.4
ea_gfn2_xtb
R² =0.97
RMSE = 0.141
DFT
9
8
7
6
5
4
3
3 4 5 6 7 8 9
electrofugality_gfn2_xtb
R² =0.94
RMSE = 0.277
11
10
9
8
7
6
5
4
3
2
2 3 4 5 6 7 8 9 10 11
electrophilicity_gfn2_xtb
R² =0.98
RMSE = 0.084
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
−0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0

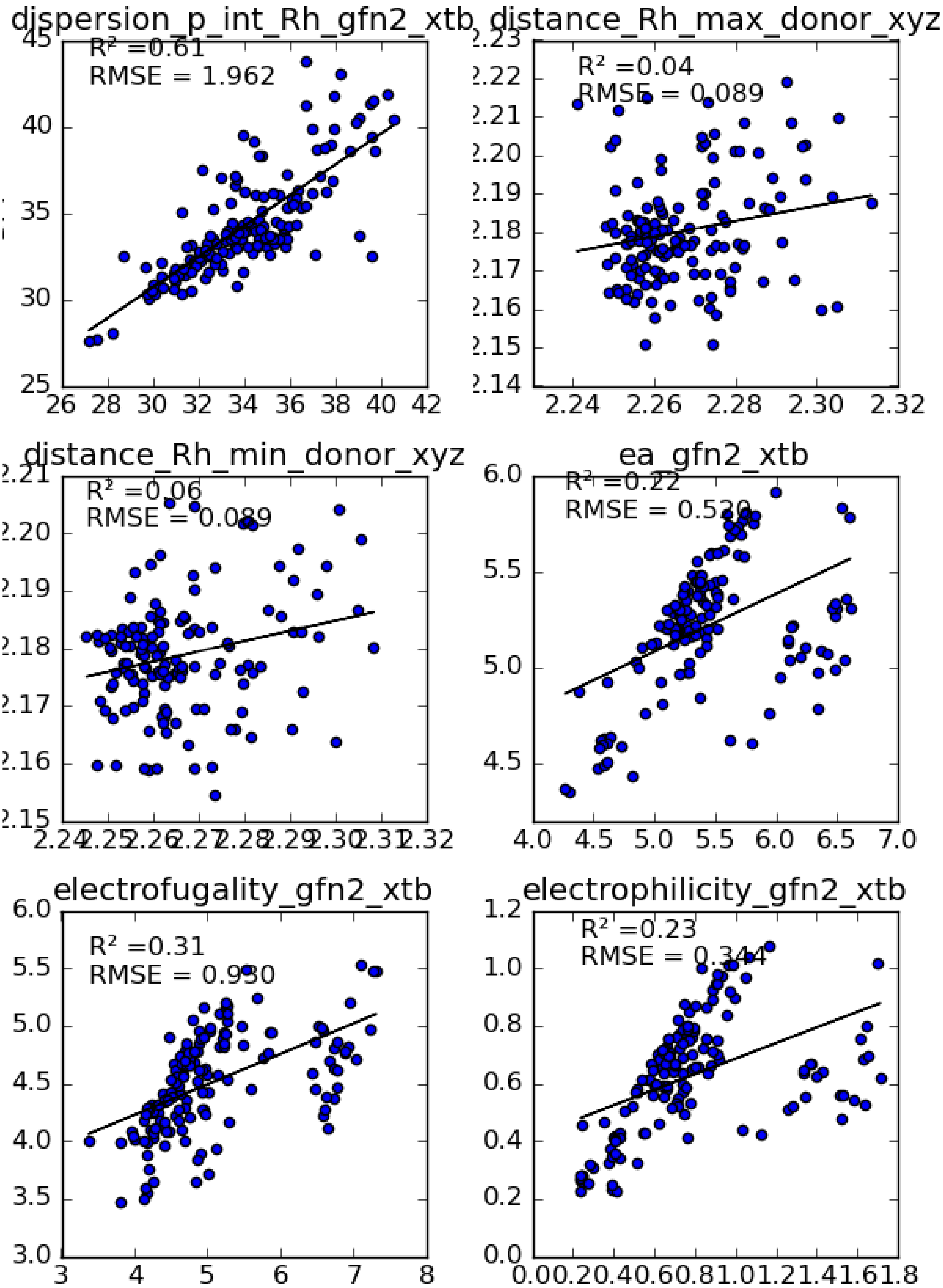

HOMO_LUMO_gap_gfn2_xtb
R² =0.50
RMSE = 0.006
0.02 0.03 0.04 0.05 0.06 0.07 0.08
).03).04).05).06).07).08
index_donor_max
R² =0.56
RMSE = 6.123
0 10 20 30 40 50 60
index_donor_min
R² =0.54
RMSE = 6.123
0 10 20 30 40 50 60
index_Rh
R² =1.00
RMSE = 0.000
0 10 20 30 40 50 60


ip_gfn2_xtb
R² =0.90
RMSE = 0.196
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5
nucleofugality_gfn2_xtb
R² =0.97
RMSE = 0.064
2.0
1.5
1.0
0.5
0.0
-0.5
−0.2 0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6
nucleophilicity_gfn2_xtb
R² =0.90
RMSE = 0.196
-2.5
-3.0
-3.5
-4.0
-4.5
-5.0
-5.5
-6.0
-6.5
−6.5 −6.0 −5.5 −5.0 −4.5 −4.0 −3.5 −3.0 −2.5
sasa_gfn2_xtb
R² =1.00
RMSE = 11.851
400
200
000
800
600
400
400 600 800 1000 1200 1400

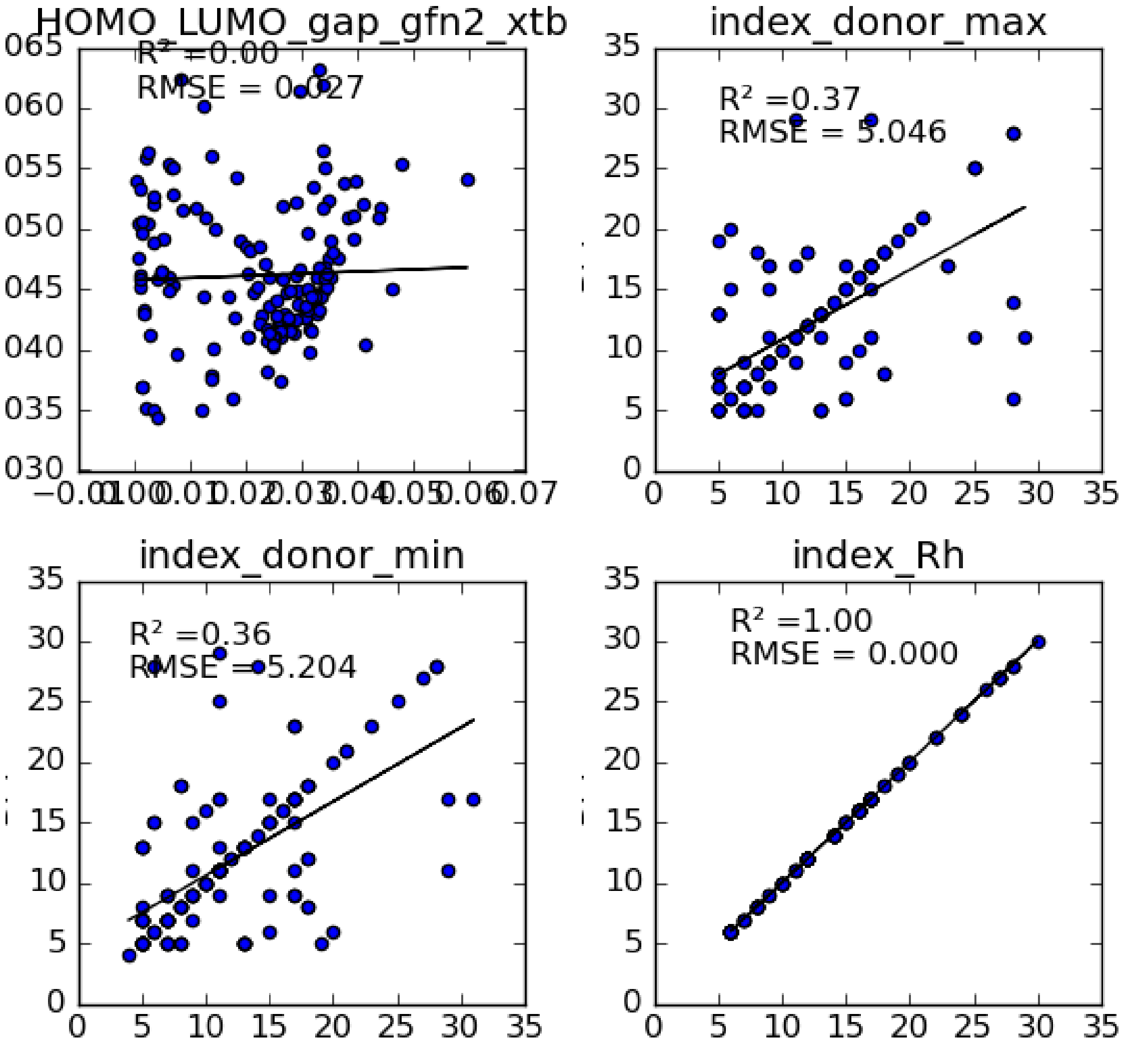
MACE filtered

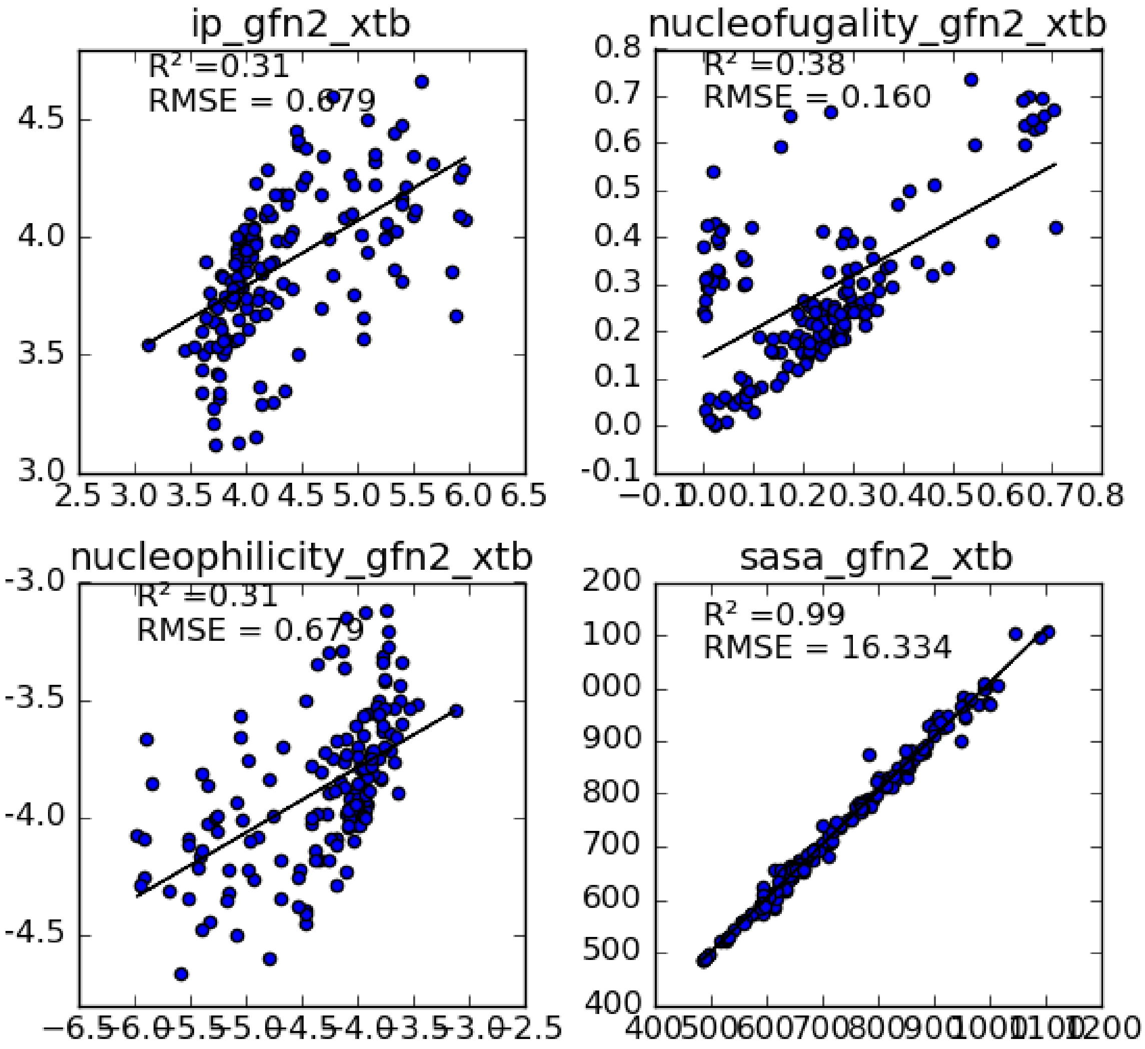



buried_volume_Rh_6A
R² =0.90
RMSE = 0.024
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65
buried_volume_Rh_7A
R² =0.96
RMSE = 0.015
0.6
0.5
0.4
0.3
0.2
0.2 0.3 0.4 0.5 0.6
cone_angle
R² =0.48
RMSE = 18.209
340
320
300
280
260
240
220
200
200 220 240 260 280 300 320 340
dipole_gfn2_xtb
R² =0.19
RMSE = 0.630
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
−0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
ersion_p_int_donor_max_gfn2
R² =0.37
RMSE = 4.118
70
65
60
55
50
45
40
40 45 50 55 60 65
ersion_p_int_donor_min_gfn2
R² =0.25
RMSE = 5.114
70
65
60
55
50
45
40
35
35 40 45 50 55 60 65

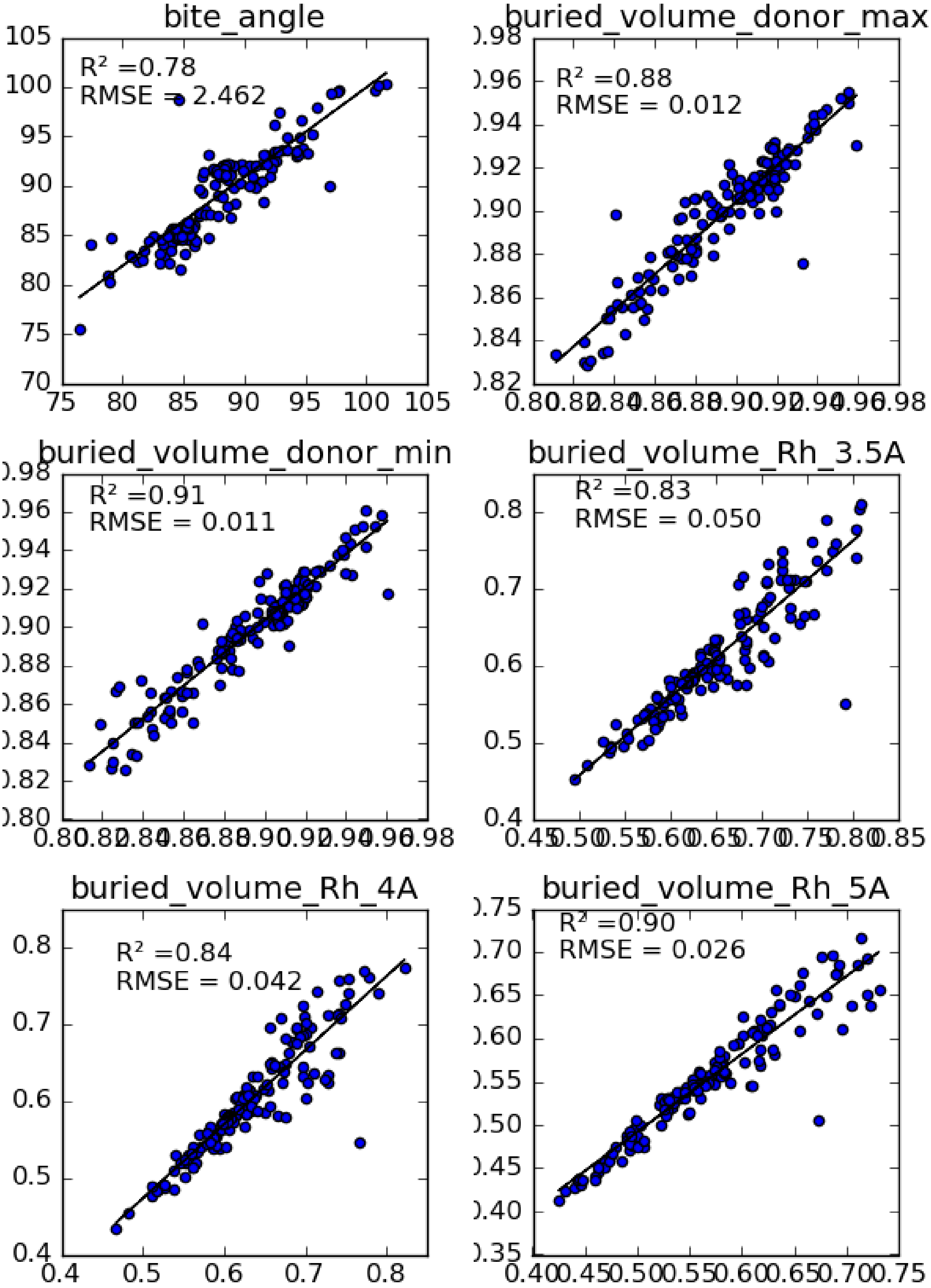

dispersion_p_int_Rh_gfn2_xtb
R² =0.61
RMSE = 1.962
45
40
35
30
25
26 28 30 32 34 36 38 40 42
distance_Rh_max_donor_xyz
R² =0.04
RMSE = 0.089
2.23
2.22
2.21
2.20
2.19
2.18
2.17
2.16
2.15
2.14
2.24 2.26 2.28 2.30 2.32
distance_Rh_min_donor_xyz
R² =0.06
RMSE = 0.089
2.21
2.20
2.19
2.18
2.17
2.16
2.15
2.24 2.25 2.26 2.27 2.28 2.29 2.30 2.31 2.32
ea_gfn2_xtb
R² =0.22
RMSE = 0.520
6.0
5.5
5.0
4.5
4.0 4.5 5.0 5.5 6.0 6.5 7.0
electrofugality_gfn2_xtb
R² =0.31
RMSE = 0.930
6.0
5.5
5.0
4.5
4.0
3.5
3.0
3 4 5 6 7 8
electrophilicity_gfn2_xtb
R² =0.23
RMSE = 0.344
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8

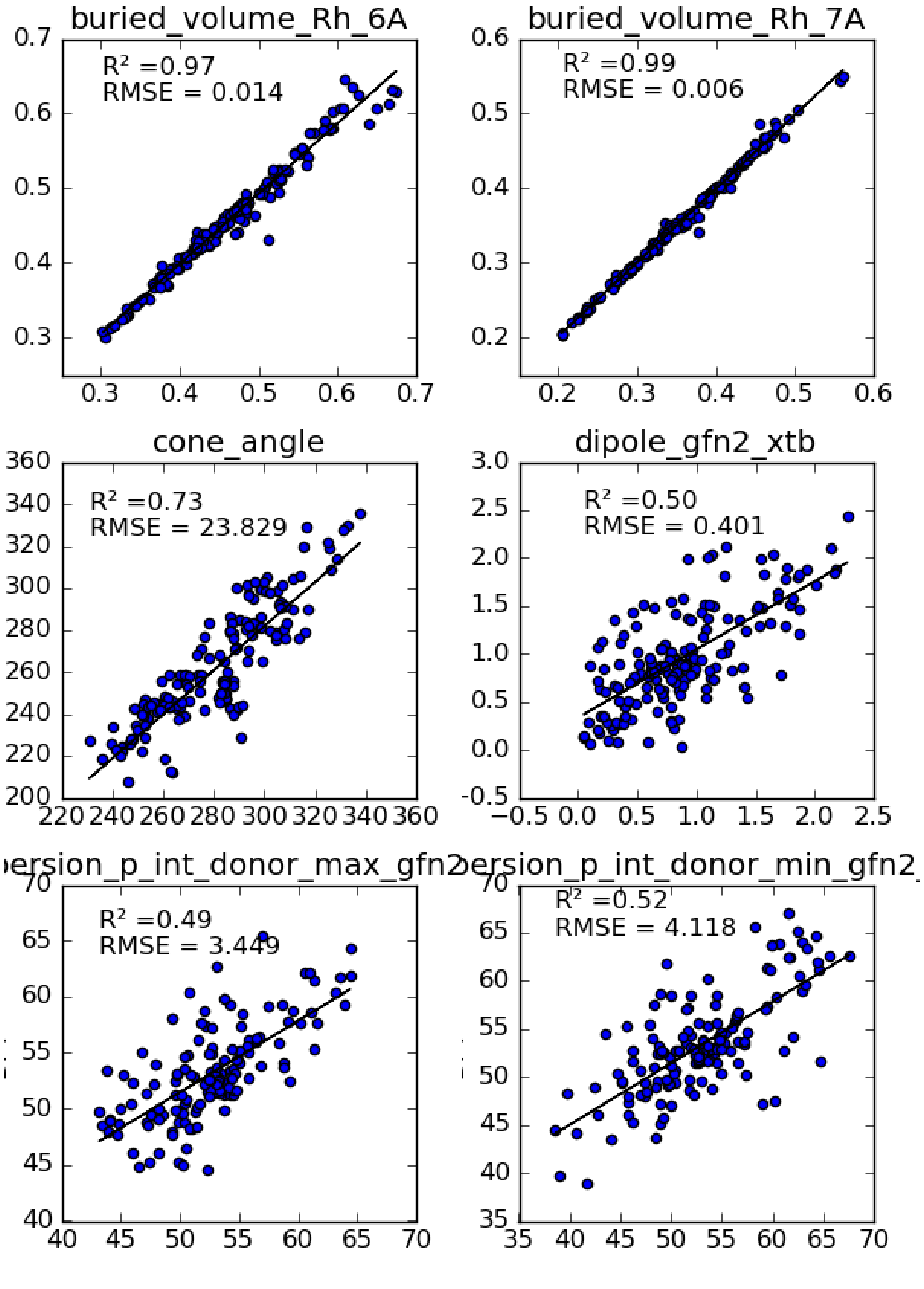
GFN-FF filtered

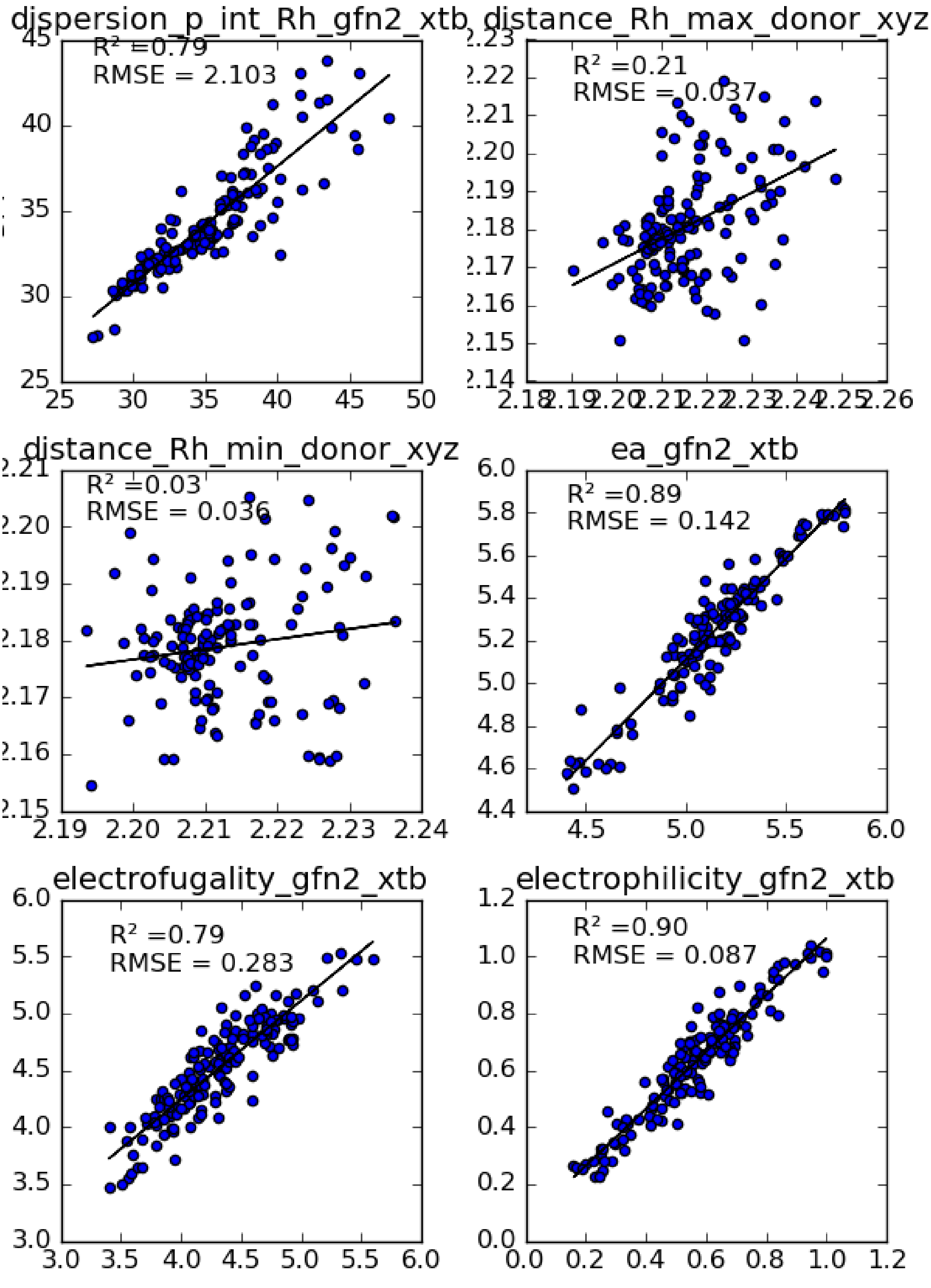


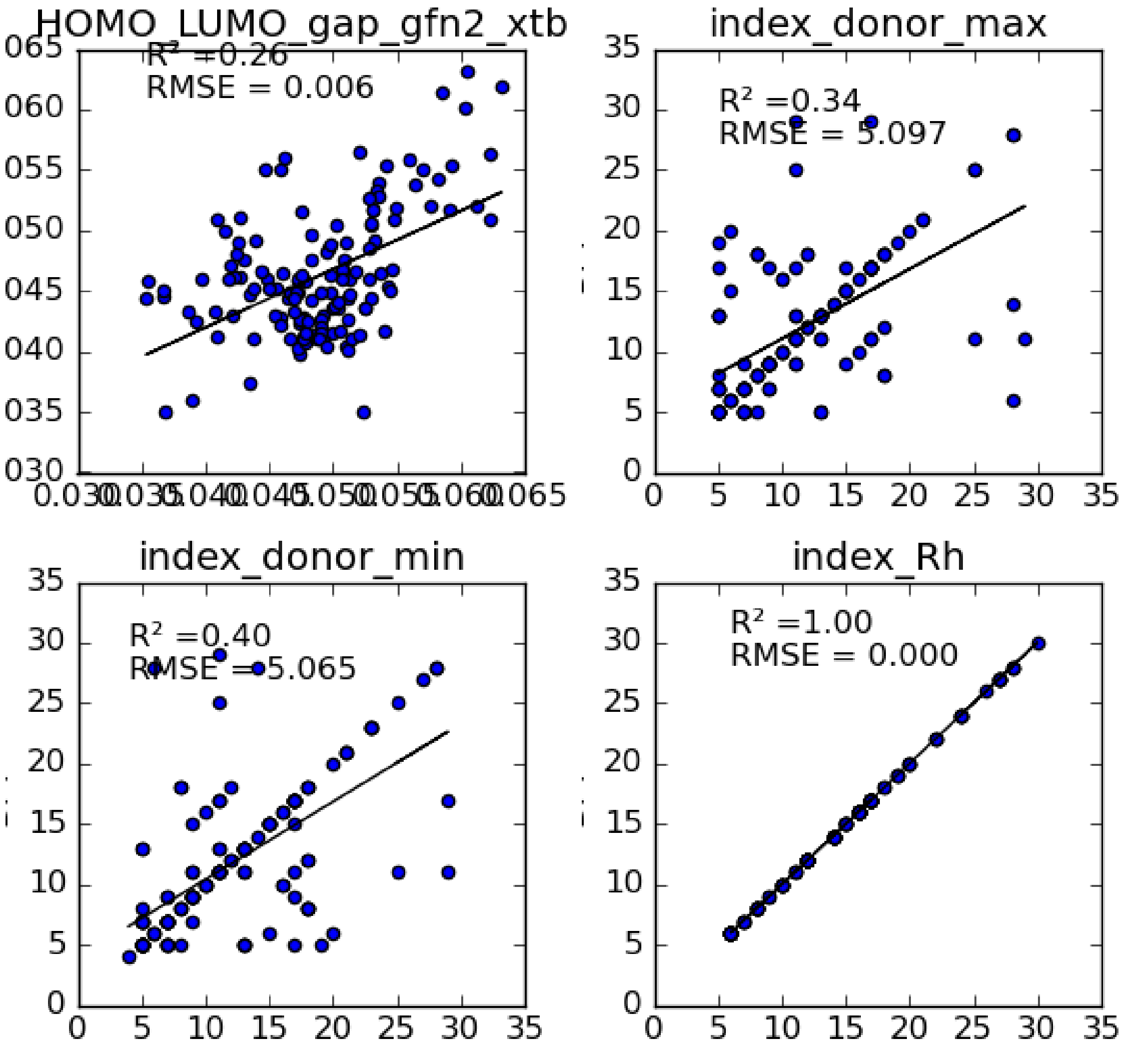

ip_gfn2_xtb
R² =0.31
RMSE = 0.679
4.5
4.0
3.5
3.0
2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5
nucleofugality_gfn2_xtb
R² =0.38
RMSE = 0.160
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
−0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
nucleophilicity_gfn2_xtb
R² =0.31
RMSE = 0.679
-3.0
-3.5
-4.0
-4.5
−6.5 −6.0 −5.5 −5.0 −4.5 −4.0 −3.5 −3.0 −2.5
sasa_gfn2_xtb
R² =0.99
RMSE = 16.334
200
100
000
900
800
700
600
500
400
400 500 600 700 800 900 1000 1100 1200

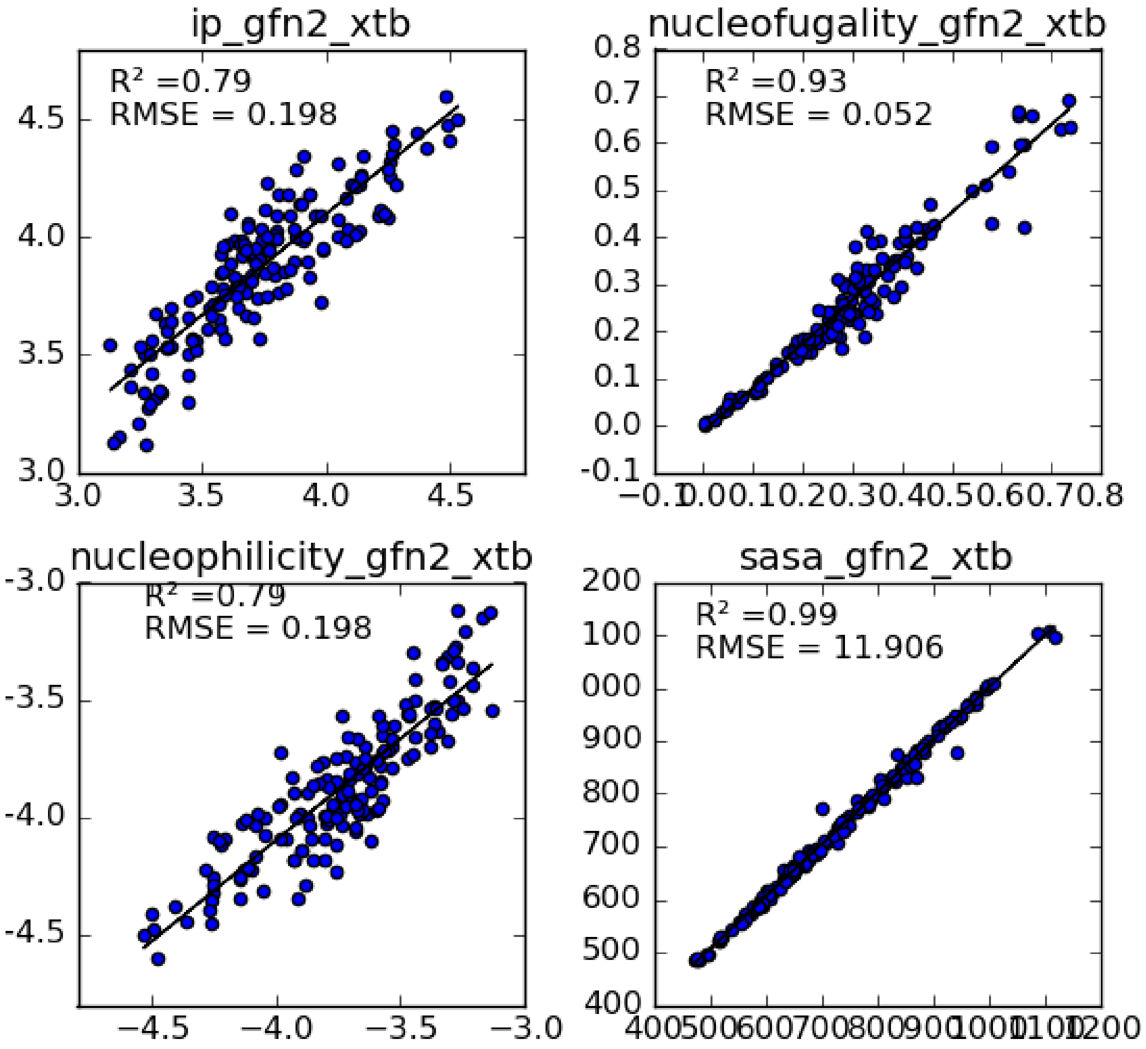
GFN2-xTB filtered




buried_volume_Rh_6A
R^2 =0.97
RMSE = 0.014
0.3 0.4 0.5 0.6 0.7
0.3 0.4 0.5 0.6 0.7
buried_volume_Rh_7A
R^2 =0.99
RMSE = 0.006
0.2 0.3 0.4 0.5 0.6
0.2 0.3 0.4 0.5 0.6
cone_angle
R^2 =0.73
RMSE = 23.829
200 220 240 260 280 300 320 340 360
220 240 260 280 300 320 340 360
dipole_gfn2_xtb
R^2 =0.50
RMSE = 0.401
-0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0
−0.5 0.0 0.5 1.0 1.5 2.0 2.5
ersion_p_int_donor_max_gfn2
R^2 =0.49
RMSE = 3.449
40 45 50 55 60 65 70
40 45 50 55 60 65 70
ersion_p_int_donor_min_gfn2
R^2 =0.52
RMSE = 4.118
35 40 45 50 55 60 65 70
35 40 45 50 55 60 65 70


dispersion_p_int_Rh_gfn2_xtb
R² =0.79
RMSE = 2.103
45
40
35
30
25
25
30
35
40
45
50
distance_Rh_max_donor_xyz
R² =0.21
RMSE = 0.037
2.23
2.22
2.21
2.20
2.19
2.18
2.17
2.16
2.15
2.14
2.18
2.19
2.20
2.21
2.22
2.23
2.24
2.25
2.26
distance_Rh_min_donor_xyz
R² =0.03
RMSE = 0.036
2.21
2.20
2.19
2.18
2.17
2.16
2.15
2.19
2.20
2.21
2.22
2.23
2.24
ea_gfn2_xtb
R² =0.89
RMSE = 0.142
6.0
5.8
5.6
5.4
5.2
5.0
4.8
4.6
4.4
4.5
5.0
5.5
6.0
electrofugality_gfn2_xtb
R² =0.79
RMSE = 0.283
6.0
5.5
5.0
4.5
4.0
3.5
3.0
3.0
3.5
4.0
4.5
5.0
5.5
6.0
electrophilicity_gfn2_xtb
R² =0.90
RMSE = 0.087
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2


HOMO_LUMO_gap_gfn2_xtb
R² =0.26
RMSE = 0.006
065
060
055
050
045
040
035
030
0.030
0.035
0.040
0.045
0.050
0.055
0.060
0.065
index_donor_max
R² =0.34
RMSE = 5.097
index_donor_min
R² =0.40
RMSE = 5.065
index_Rh
R² =1.00
RMSE = 0.000
0
5
10
15
20
25
30
35


ip_gfn2_xtb
R² =0.79
RMSE = 0.198
4.5
4.0
3.5
3.0
3.0
3.5
4.0
4.5
nucleofugality_gfn2_xtb
R² =0.93
RMSE = 0.052
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
−0.10.00.10.20.30.40.50.60.70.8
nucleophilicity_gfn2_xtb
R² =0.79
RMSE = 0.198
-3.0
-3.5
-4.0
-4.5
−4.5
−4.0
−3.5
−3.0
sasa_gfn2_xtb
R² =0.99
RMSE = 11.906
200
100
000
900
800
700
600
500
400
40050060070080090010001100200

-

A.2. Interquartile range method

The interquartile range (IQR) is a measure of the spread of the data. When the data is ordered from high to low it can be divided into 4 segments, called quartiles. The point exactly in between the second and third quartile is then by definition the mean. IQR only uses the second and third quartiles as data and filters the extreme values (both high and low) out.

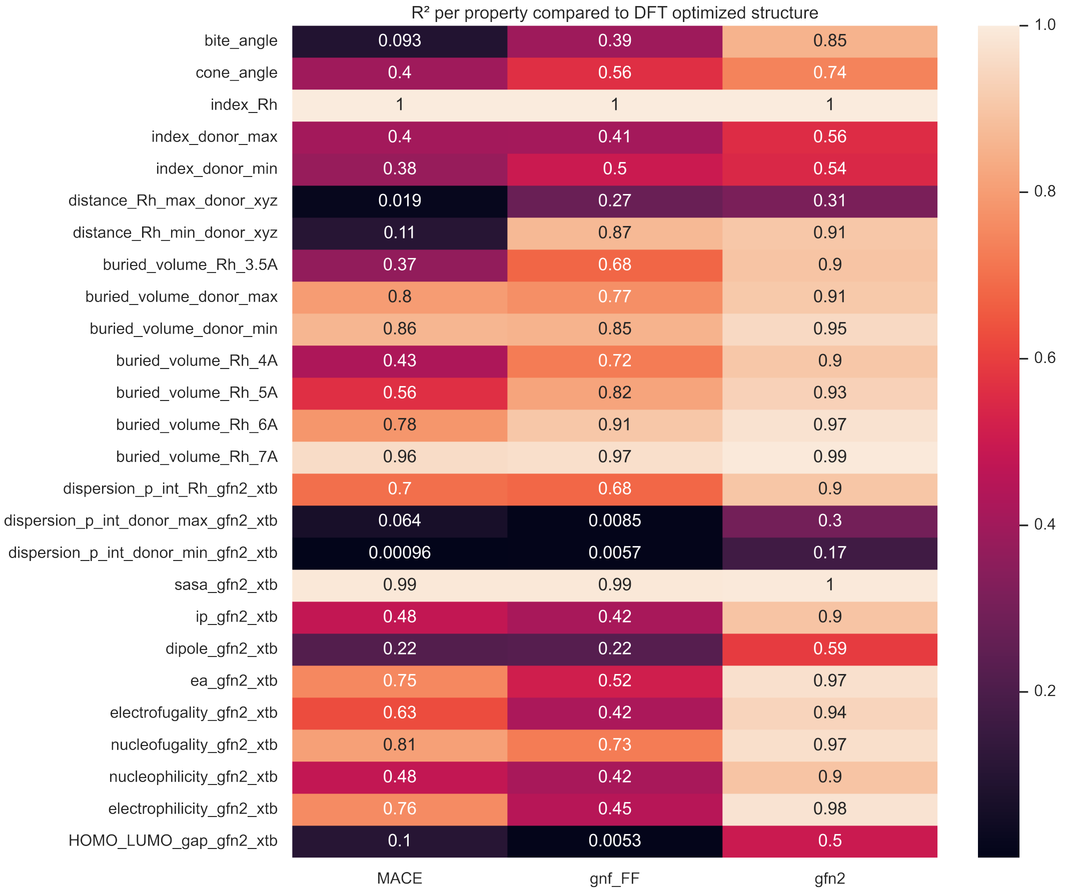
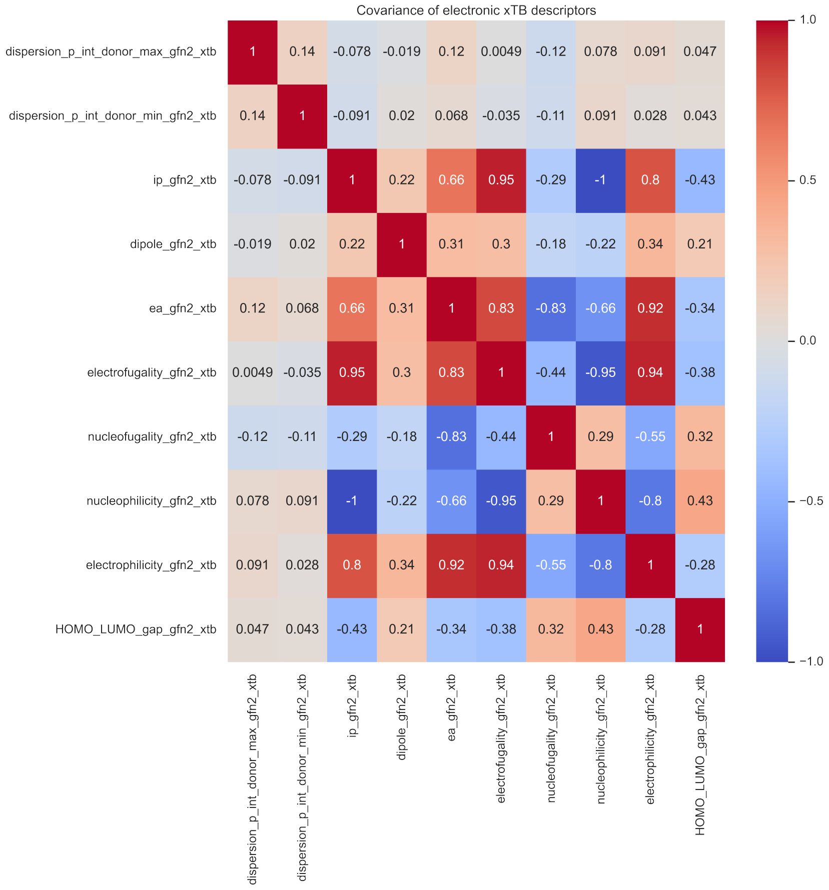


Figure A.1: Heatmap of the descriptors with IQR filtering

A.3. Filtered heatmap

A.4. Covariance of descriptors



	dispersion_p_int_donor_max_gfn2_xtb	dispersion_p_int_donor_min_gfn2_xtb	ip_gfn2_xtb	dipole_gfn2_xtb	ea_gfn2_xtb	electrofugality_gfn2_xtb	nucleofugality_gfn2_xtb	nucleophilicity_gfn2_xtb	electrophilicity_gfn2_xtb	HOMO_LUMO_gap_gfn2_xtb
dispersion_p_int_donor_max_gfn2_xtb	1	0.14	-0.078	-0.019	0.12	0.0049	-0.12	0.078	0.091	0.047
dispersion_p_int_donor_min_gfn2_xtb	0.14	1	-0.091	0.02	0.068	-0.035	-0.11	0.091	0.028	0.043
ip_gfn2_xtb	-0.078	-0.091	1	0.22	0.66	0.95	-0.29	-1	0.8	-0.43
dipole_gfn2_xtb	-0.019	0.02	0.22	1	0.31	0.3	-0.18	-0.22	0.34	0.21
ea_gfn2_xtb	0.12	0.068	0.66	0.31	1	0.83	-0.83	-0.66	0.92	-0.34
electrofugality_gfn2_xtb	0.0049	-0.035	0.95	0.3	0.83	1	-0.44	-0.95	0.94	-0.38
nucleofugality_gfn2_xtb	-0.12	-0.11	-0.29	-0.18	-0.83	-0.44	1	0.29	-0.55	0.32
nucleophilicity_gfn2_xtb	0.078	0.091	-1	-0.22	-0.66	-0.95	0.29	1	-0.8	0.43
electrophilicity_gfn2_xtb	0.091	0.028	0.8	0.34	0.92	0.94	-0.55	-0.8	1	-0.28
HOMO_LUMO_gap_gfn2_xtb	0.047	0.043	-0.43	0.21	-0.34	-0.38	0.32	0.43	-0.28	1

Figure A.2: Covariance heatmap of electronic descriptors

A.5. hRMSD of MACE and GFN-FF vs DFT

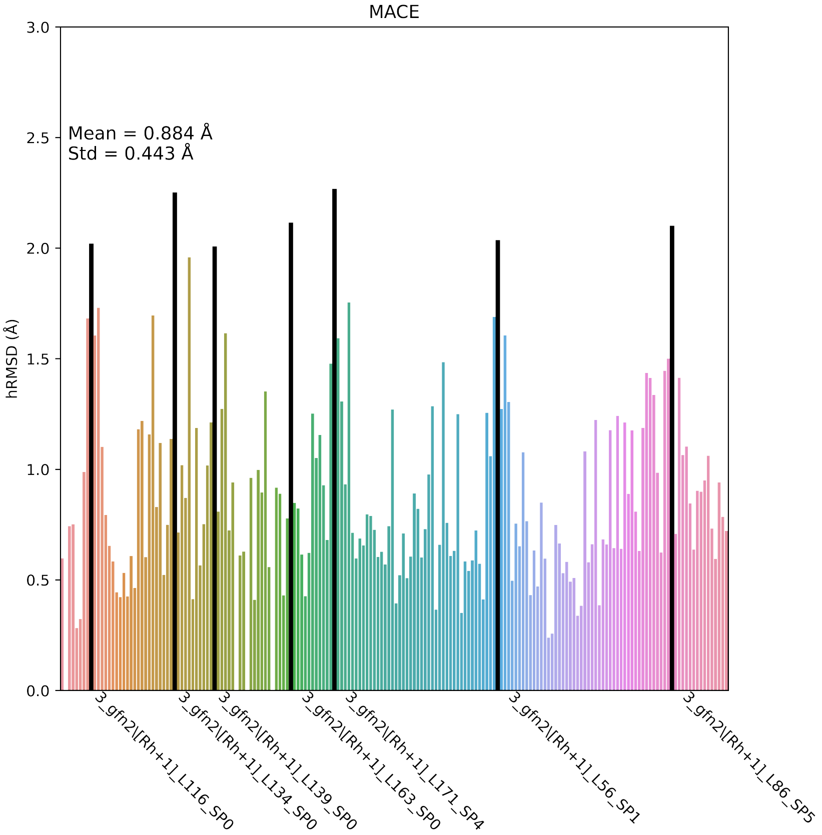


Figure A.3: hRMSD representation of MACE

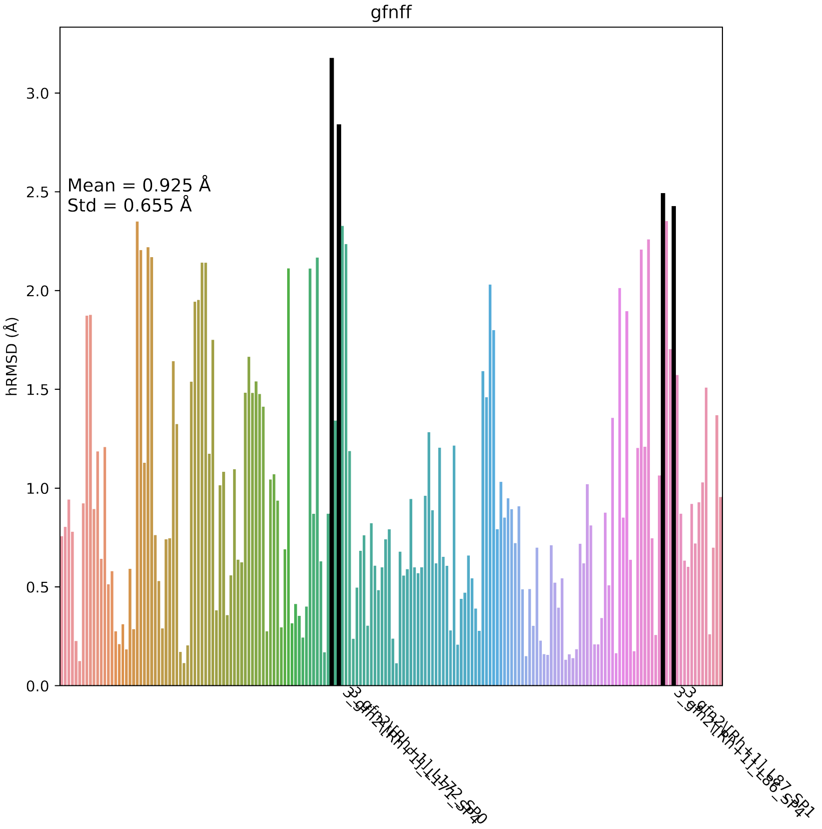
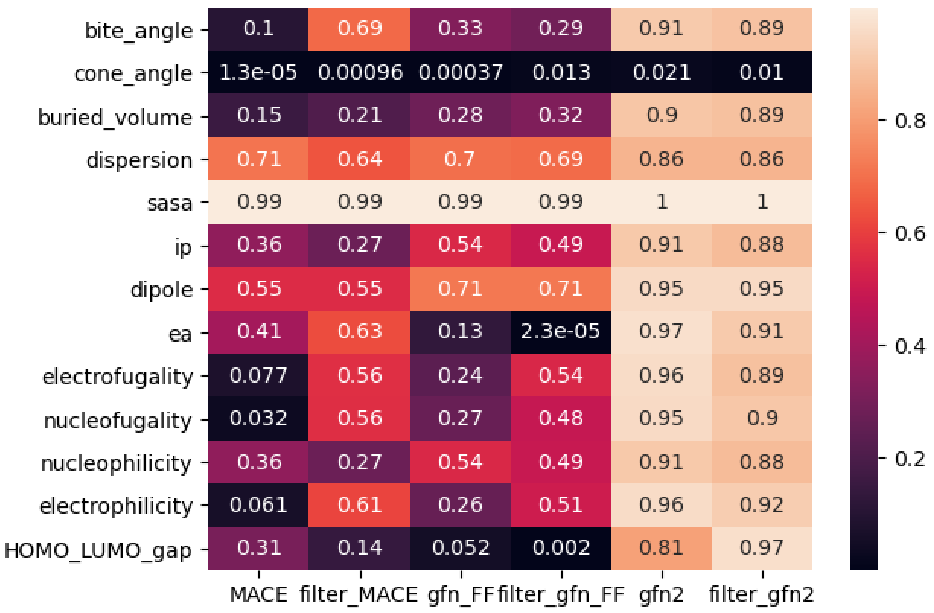


Figure A.4: hRMSD representation of GFN-FF

B

Intermediary dataset

During this research, correlations were investigated for another (preliminary) dataset obtained from the pharmaceutical partner. This dataset contained some of the same ligands, but also ligands outside of the main dataset. IQR filtering has been done upon all descriptors, to filter out large (chemically interesting) outliers.



	MACE	filter_MACE	gfn_FF	filter_gfn_FF	gfn2	filter_gfn2
bite_angle	0.1	0.69	0.33	0.29	0.91	0.89
cone_angle	1.3e-05	0.00096	0.00037	0.013	0.021	0.01
buried_volume	0.15	0.21	0.28	0.32	0.9	0.89
dispersion	0.71	0.64	0.7	0.69	0.86	0.86
sasa	0.99	0.99	0.99	0.99	1	1
ip	0.36	0.27	0.54	0.49	0.91	0.88
dipole	0.55	0.55	0.71	0.71	0.95	0.95
ea	0.41	0.63	0.13	2.3e-05	0.97	0.91
electrofugality	0.077	0.56	0.24	0.54	0.96	0.89
nucleofugality	0.032	0.56	0.27	0.48	0.95	0.9
nucleophilicity	0.36	0.27	0.54	0.49	0.91	0.88
electrophilicity	0.061	0.61	0.26	0.51	0.96	0.92
HOMO_LUMO_gap	0.31	0.14	0.052	0.002	0.81	0.97

Figure B.1: Both filtered and unfiltered descriptor values of the old dataset.