TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Urmas Pitsi 192028IAPM

Conformational Analysis of an Organometallic Compound with Data Science Inspired Workflow

Master's thesis

Supervisors:	prof. Toomas Tamm
	Juhan-Peep Ernits Phd

Tallinn 2023

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Urmas Pitsi 192028IAPM

Üleminekumetalliühendi konformatsioonianalüüs andmeteadusest inspireeritud töövooga

Magistritöö

Juhendajad: prof. Toomas Tamm
Juhan-Peep Ernits Phd

Tallinn 2023

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Urmas Pitsi

01.01.2023

Abstract

Conformational analysis is one of the central topics in computational chemistry. In addition to being a general problem with exponential complexity, it is well known that conformational analysis of transition metal compounds is a more challenging task, mainly due to the difficult nature of accurately modelling the forces between atoms that include (transition)metals. In this thesis we present the results of a conformational analysis by which we found a set of conformers for a titanium tartrate complex. This complex is known from “Sharpless epoxidation” - a work leading to a Nobel Prize by T.Katsuki and K.B.Sharpless. Conformers that we found are a significant step forward in the particular research, as the last notable work at TalTech on the subject was carried out in 2011. In addition to conformational analysis, this thesis presents (1) an overview of a workflow with analysis and suggestions that could help future research in this area and (2) an open-source software “Molli”, that was written by the author, and which was used extensively in preparing numerical and visual analysis presented in this thesis.

This thesis is written in english and is 54 pages long, including 6 chapters, 3 figures and 14 tables.

Annotatsioon

Konformatsioonianalüüs kuulub arvuskeemia võtmeteemade hulka. Lisaks sellele, et tegemist on eksponentsiaalset keerukust omava üldprobleemiga, on hästi teada, et üleminekumetalliühendite konformatsioonianalüüs pakub veel suuremat väljakutset, kuna aatomite vaheliste jõudude täpne modelleerimine (ülemineku)metalli aatomite puhul lisab märkimisväärselt keerukust antud probleemile. Käesolevas magistritöös esitame konformatsioonianalüüsi tulemused, mille käigus leidsime konformeeride hulga titaan-tartraat ühendile. Antud ühend on tuntud “Sharpless epoksüdatsioonist” - Nobeli preemiani viinud tööst, mille autoriteks T.Katsuki ja K.B.Sharpless. Meie leitud konformeerid on oluline samm edasi antud uurimissuunal, kuna eelnev töötulemus nimetatud ühendiga pärineb aastast 2011. Lisaks konformatsioonianalüüsi tulemustele esitab käesolev magistritöö (1) ülevaate töövoost koos analüüsi ja ettepanekutega, mis loodetavasti aitavad sarnast tööd tulevikus paremini läbi viia, (2) käesoleva magistritöö autori loodud vaba tarkvara “Molli”, mida autor kasutas ulatuslikult numbrilise ja visuaalse analüüsi koostamisel antud töös.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 54 leheküljel, 6 peatükki, 3 joonist, 14 tabelit.

List of abbreviations and terms

Ab Initio	Latin: “from the first principles”
CENSO	An open-source software for evaluating and sorting molecular geometries
Conformation	Energy-optimized molecular geometry (local minimum)
Conformer	Low-energy conformation of a particular molecule
CREST	Conformer–Rotamer Ensemble Sampling Tool
DFT	Density Functional Theory
GFN	Geometries, Frequencies and non-covalent interactions
GFN-FF	GFN force field method
GFN2-xTB	GFN extended tight binding method
PES	Potential energy surface
TargetMol	The titanium tartrate complex this thesis works with
xTB	Extended Tight Binding

Table of contents

List of figures

List of tables

1 Introduction

Conformational analysis is one of the central topics in computational chemistry, entailing the process of sampling molecular configurations from the conformational space. The complexity of this problem is exponential, $O(c^n)$, where n is the number of rotatable bonds and $c > 0$. In addition to the high combinatorial complexity, we need to do expensive energy calculations at each configuration which makes a conformational analysis of transition metal compounds even more challenging task because of the difficult nature of accurately modelling the forces between atoms that include (transition) metals [1], [2], [3], [4], [5], [6]. Our research subject was a titanium tartrate complex [7], catalyst from famous Sharpless epoxidation - a work leading to a Nobel prize by T. Katsuki and K.B. Sharpless with chemical formula $C_{24}H_{44}O_{16}Ti_2$ and molecular structure as shown in Figure 1.

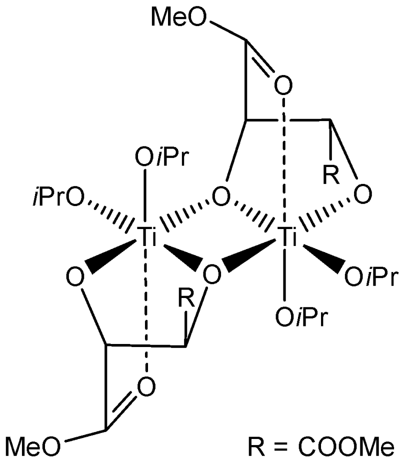


Figure 1. Molecular structure of the titanium tartrate complex.

For the sake of simplicity, we refer to the compound we are working with as "TargetMol" throughout this thesis. Our aim is to make progress in the research with the TargetMol, a medium sized organometallic compound containing 86 atoms including 2 titanium atoms, as the last notable work with the TargetMol at TalTech was in 2011. Among the characteristics of the TargetMol is the Ti_2O_2 core, forming nearly a symmetric planar rhombus (as in [7]) and methyl groups instead of ethyl groups. The main objective of the research was to find good quality conformers for the TargetMol with the additional research goal of investigating whether we can achieve our main objective with CREST [8] software. The major open question was whether force field and semiempirical methods, underlying the calculations of CREST, are capable of modeling TargetMol correctly, as the results of CREST would be the inputs to calculations at the level of Density Functional Theory (DFT), which was our ultimate goal. As calculating everything at a higher level of DFT would be computationally prohibitive, we needed to apply data analysis tools to find computational shortcuts.

1.1 Main contributions

Main contributions of this thesis could be grouped into following three categories:

- We found new conformer ensembles for a particular medium sized organometallic molecule, referenced in this work as “TargetMol”. The TargetMol contains 86 atoms including 2 titanium atoms. A smaller set of conformers was fully optimized at the theory level of PBE0/cc-pVTZ and a broader set of conformers was optimized at the level of PBE0/def2-SV(P) and BP86/def2-SV(P). With this result we make significant progress furthering the conformational analysis of the TargetMol. The best conformer found has a relative energy more than 112 kJ/mol lower than the best one known so far at the theory level of PBE0/cc-pVTZ. Altogether we report 205 additional conformers within a relative conformational energy threshold of <67 kJ/mol, all fully optimized at the theory level of PBE0/def2-SV(P). We further show that it is highly likely that these conformers have their relative conformational energies in the same range also at the PBE0/cc-pVTZ level. Current work serves as a proof of concept for further analysis in this area and especially with the TargetMol. By utilizing CREST software for sampling conformational space and by further refinement of the results with desired DFT, one can extend the research to obtain even more conformers.
- We propose and implement a workflow on how to practically approach similar tasks in the future, using conformational analysis of the TargetMol as a use case. Proposed workflow helps expert computational chemists to: (1) generate candidate geometries (CREST), (2) pre-screen, analyze and filter intermediate results and (3) select final geometries for further analysis.

- We created an open-source tool Molli, written in Python, that allows to perform various helpful tasks a computational chemist faces, e.g., to process and extract information from log files generated by the Gaussian program, to analyze geometries of molecules, analyze and compare optimization trajectories etc.

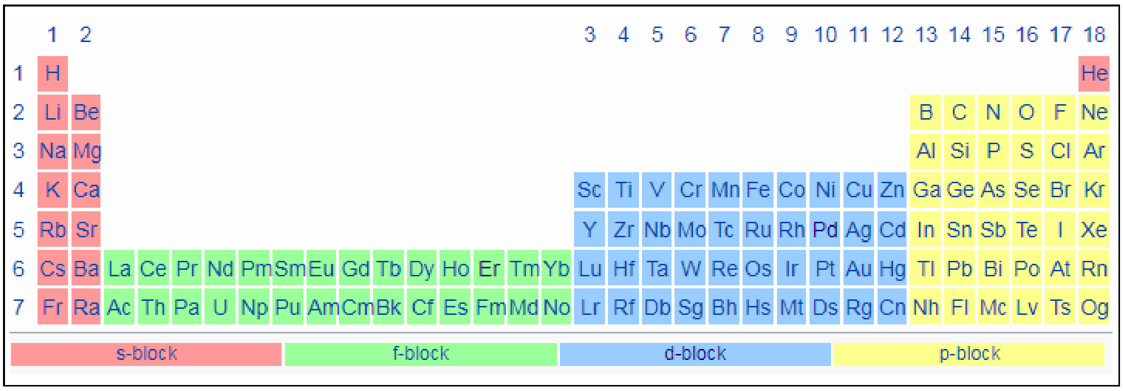
2 Theoretical background

The main aim of this chapter is to give a brief overview of the theoretical background and key concepts. Our aim is not to explain everything in a very technical and detailed manner, rather to explain the key concepts, understandings and core principles of the modern theory. For a more curious reader there are a multitude of great textbooks on the topic of computational chemistry [9], [10], [11], [12]. As the result of reading this chapter a reader should get an idea of the following:

- How does modern science view molecular structure?
- What are some of the central topics in computational chemistry?
- What is the importance and motivation of the Density Functional Theory?
- What is the importance and motivation of conformational analysis?

2.1 Elementary Chemistry: molecules, atoms, electrons

Everything material around us consists of molecules. A molecule, being a typical unit of abstraction in chemistry, is a group of atoms held together by interatomic forces. Atoms consist of a nucleus and one or more electrons. In modern chemistry the major fundamental unit is an atom, as molecules are viewed as collections of atoms, chemical formulas represent the counts of the atoms in a molecule, molecular structures are visually depicted as atoms which are connected by lines representing chemical bonds. However, in computational chemistry, electrons play an even more important role as a fundamental unit. The reason being that from the probability density of electrons we can derive all physical properties of a molecule. That is, if we could find out for all points in the space what is the probability of an electron being at that point, then we would know everything there is to know about that space. This remarkable discovery [13] in 1960-s led to the creation of Density Functional Theory (DFT) and was later awarded the Nobel prize. We will describe DFT in more detail in the following chapters. The periodic table depicts chemical elements arranged in the order of their atomic number, which is defined as the number of protons in the nucleus of an atom. In a way we could interpret the atomic number as the complexity of a chemical element, starting from the simplest, element no.1 Hydrogen(H) until element no.118 Oganesson(Og). Although definitions of the term "transition metal" vary, they are generally regarded as those elements shown in blue in the d-block as shown in Figure 2.



	1	2															3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	H																															He
2	Li	Be																									B	C	N	O	F	Ne
3	Na	Mg																									Al	Si	P	S	Cl	Ar
4	K	Ca															Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
5	Rb	Sr															Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
6	Cs	Ba	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
7	Fr	Ra	Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr	Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg	Cn	Nh	Fl	Mc	Lv	Ts	Og

s-block f-block d-block p-block

Source: https://en.wikipedia.org/wiki/Block_(periodic_table)#d-block

Figure 2. Block periodic table.

Transition metal compounds are an important class of molecules, both in basic chemical research as well as in industrial and pharmaceutical chemistry. In addition, transition metal complexes are crucial for numerous biological processes [3]. In biology transition metals are some of the key elements in life and evolution, e.g.:

- Iron: without iron, oxygen wouldn't make it to the brain and life would not exist. Helps transporting oxygen to the brain and muscles inside hemoglobin.
- Cobalt: component of vitamin B12.

In material science transition metal compounds play an important role in the production of coloured paints, semiconductors, solar-cells, batteries etc. Another very important role of transition metals is their use as catalysts.

2.2 Computational Chemistry

Computational chemistry as the name suggests, is literally chemistry coupled with computer science. All natural sciences used to be mostly experimental sciences. With the advent of computers natural sciences grew new branches with computers involved. The same happened with chemistry. Instead of conducting relatively expensive "in vitro" (physical experiments in a laboratory) experiments, we could do experiments virtually using computers, with the added benefit of algorithms and methods from computer science. Computational methods are a complement to the experiments, not a replacement. No matter how powerful and exact the methods of computational chemistry are, the ground truth about nature is finally decided by the experiment.

Motivation

By using the tools and methods of computational chemistry we can calculate many important physical molecular properties. That leads us to better understanding of these properties so that we could design new and better ones e.g.:

- Drugs in the Pharmaceutical industry.
- Materials, plastics, fuels, component materials in electronics and batteries etc in Materials Science.
- etc.

Main methods of Computational Chemistry

According to [9] we could categorize the main tools into five broad classes:

- Molecular Mechanics: collection of atoms in a molecule is modelled based on classical Newtonian mechanics. Atoms are modelled as balls connected to each other with elastic connections. By knowing the lengths of each connection and the energy needed to bend or stretch these, we could calculate the energy of a given molecule. Calculations in this paradigm are “fast”.
- Ab Initio: (*ab initio*, Latin: “from the start”, i.e., “from first principles”) Ab Initio calculation is based only on quantum mechanics (deriving from quantum physics, the best-known description of the matter in our Universe) and is in this sense “from first principles”. Ab Initio calculations are based on the Schrödinger equation, which is one of the fundamental equations of modern physics and describes, among other things, how the electrons in a molecule behave. The Ab Initio method solves the Schrödinger equation for a molecule and gives us an energy and a wavefunction. The wavefunction is a mathematical function that can be used to calculate the electron distribution (and, in theory at least, anything else about the molecule). Because the Schrödinger equation cannot be solved exactly for any molecule with more than one electron, we need to use

approximations [9], meaning that any practical Ab Initio method approximates the Schrödinger equation. “True” Ab initio calculations are computationally “impossible” on any practically relevant system.

- Semiempirical methods: empirical here means experimental and by combining theory with experiment gives us the name “semiempirical”. The idea is that we approximate Ab Initio calculations with some empirical data to speed up the calculations at the expense of accuracy. It is based on the Schrödinger equation but parameterized with experimental (or high-level theoretical or computational) values [9]. Calculations are slower than Molecular Mechanics but much faster than Ab Initio.
- Density Functional Methods based on Density Functional Theory (DFT): a functional is a function that takes a function as input and returns a value, same as “higher-order function” in computer science, e.g., energy functional might take wave function representing atomic orbitals as input and return a value corresponding to the energy. DFT methods are based on the Schrödinger equation, however, theoretical DFT does not calculate a wavefunction, but rather derives the electron distribution (electron density function) directly. Calculations are faster than Ab Initio, but slower than semiempirical [9]. In practice DFT methods are approximations to the Schrödinger equation, by using a set of simpler wave functions to achieve computational speed-up.
- Molecular Dynamics: applying the laws of motion to the atoms under the influence of a force field. A force field could be generated with any method described above: Ab Initio, Semiempirical or Density Functional Methods. It is important to distinguish between Molecular Dynamics and Molecular Mechanics, first of which describes “motion” while the latter describes a “mechanical” treatment of molecules [9].

2.3 Schrödinger Equation

Famous Austrian physicist Erwin Schrödinger, after whom the equation is named, made a key discovery in quantum mechanics in 1925 by postulating the equation which describes fundamental particles and their forces, earning him a Nobel Prize in physics in 1933. Schrödinger equation involves a wave function which gives a precise quantum mechanical description about the evolution of a physical system over time. By separating variables we arrive at time-independent Schrödinger equation (2.1),

$$H\Psi \ = \ E\Psi \qquad (2.1)$$

where: Ψ is wave function, E is the energy of the system and H is called Hamiltonian operator, a function acting upon Ψ returning the observable property of the system, a scalar value of energy in this case. More technically, if equation (2.1) holds Ψ is called an eigenfunction and E an eigenvalue [11]. Another important property of the wave function Ψ is that while being a complex variable function, the square of its modulus, $|\Psi|^2$, represents a function of the probability density. Thus, the probability that a chemical system will be found within some region of multi-dimensional space is equal to the integral over that region of space [11].

2.4 Density Functional Theory (DFT)

Motivation: To calculate accurate molecular properties.

How to achieve: Create a mathematical model that maps electron density to the energy of the system.

As mentioned above, the computational problem is how to efficiently calculate molecular properties so that they would be accurate enough, in compliance with theories of quantum mechanics. DFT is based on the two Hohenberg-Kohn theorems, which state that the ground-state properties of an atom or molecule are determined by its electron density function, and that a trial electron density must give an energy greater than or equal to the true energy (the latter theorem is true only if the exact functional could be used) [9]. Origin of the Density Functional Theory dates to 1964 with the now famous publication by P. Hohenberg and W. Kohn, “Inhomogeneous Electron Gas” [13]. Walter Kohn (1923-2016) received the Nobel Prize in chemistry 1998 “for his development of the density-functional theory”. As is known from Quantum Mechanics the Schrödinger equation is well defined, but impossible to solve in practice. DFT offers an alternative approach to the solution of the Schrödinger equation by searching for a (universal) functional that maps electron density to the energy of the system (e.g., molecule). During the last decades DFT has become a leading methodological framework, achieving its position largely due to the excellent accuracy over the computational cost ratio as well as the broad applicability across system classes of today's density functional approximations (DFAs) [14], [15], [16] [17], [18], [19]. The strengths of density functional theory are practicality, universality for all electronic ground states, and a sound theoretical foundation [20]. Based on Quantum Mechanics and Coulomb's law for the electron-electron interaction, we know almost everything we need, in principle, for the description of atoms, molecules, and solids. By being grounded in experiment the underlying principles of quantum mechanics and Coulomb's law are accepted as universally valid and basic. Starting from these principles, we can prove that the ground-state exchange-correlation energy is a functional of the total electron density [20]. The caveat is that although we can prove that the functional exists, neither don't we know this functional exactly nor its systematic approximation that would always converge to the exact answer as given by the fully ab initio theory. To the question whether density functional theory is ab initio or semiempirical, one could argue that if the functionals are constructed without empirical fitting then it can fall in between as a nonempirical theory. [20]. Modern density functional approximations (DFAs) try to further improve on this balance between the accuracy and computational cost in several ways, e.g., by efficient technical implementation in modern programs and by sophisticated design and parameterization [21].

2.5 Density Functionals and Basis sets

In this context a functional is a function that takes a function as input and returns a value. In computer science, in functional programming particularly it is known as higher-order function.

A density functional is a function that maps the electron density to the value of energy, where electron density is modeled with the collection of basis functions.

A basis set is a set of mathematical functions (basis functions) and the linear combinations of basis functions yield molecular orbitals [9]. Molecular orbital is a function that describes the location and behaviour of an electron in a molecule. The core idea is to approximate these orbitals with a set of functions as closely as possible. A desired property to have is that when the basis set size is increased then it would approach theoretical limits. Usually, several basis functions describe the electron distribution around an atom. The combination of these atomic basis functions describes the electron distribution in the molecule as a whole. The typical size of a basis set ranges from hundreds to thousands of basis functions. In this thesis we mainly used basis sets ranging from 696 to 1952 basis functions.

2.6 Complexity of DFT: Theoretical vs practical

In computational complexity theory, a complexity class relates a set of problems to the resources needed to effectively solve them by a computer. To simplify, the easiest problems in this scale belong to category P (Polynomial time complexity), the problems that could be solved by computers, i.e., in a "reasonable" time. NP is the complexity class in which problems are "hard" to solve but "easy" to check the solution. The term NP comes from the term "non-deterministic polynomial" complexity. For a decision problem that means we can check the correctness of the answer in polynomial time, e.g., the decision version of the travelling salesman problem: given a route, it is trivial to verify whether the length of the route is smaller than x. P and NP are for classical computers, the analogous classes for quantum computers are called BQP and QMA correspondingly. The authors of [22] claim to have proved that to find a universal functional for DFT is QMA-hard, meaning that by finding an efficient approximation to the universal DFT functional would imply that QMA=P. That would collapse most of the complexity classes into class P (including NP would be equal to P), which as we know contains problems that are easily solvable by today's computers. It remains to be seen, but more likely than not there exists no universal functional for DFT, which in practice means that the research of designing new and better functionals is never ending, explaining the broadness and variety of today's "zoo" of available functionals and basis sets. The mere fact that in practice we get very accurate results with reasonable time is quite astonishing and shows that there are no limits for the ingenuity of scientists in inventing clever algorithmic shortcuts and implementation tricks. Authors of [22] also show that Kohn-Sham DFT, the practical approximation to DFT that is used in most algorithms for DFT calculations, as well as Hartree-Fock method are both NP-hard problems, while Hartree-Fock method is also shown to be NP-complete (meaning by solving this problem, you solve all NP problems). In practical implementations the DFT algorithms have complexity typically in range $O\left(n^{2.5}\right)..O\left(n^{3}\right)$. Considering the exponential complexity, $O(c^n)$, where $c > 0$, of the Schrödinger equation, the sub-cubic scaling in practice is a very large speedup.

2.7 Molecular Geometry and Potential Energy Surface (PES)

Molecular geometry is a description of a molecule by the 3D coordinates of its constituent atoms. It is essentially equivalent to the notion of representing a molecule as a three-dimensional point cloud, where the points are the atoms of the molecule.

According to Born-Oppenheimer approximation the nuclei are essentially stationary compared to the electrons in a molecule. This seemingly simple statement has surprisingly profound implications. It gives meaning to the potential energy surface (PES), a central concept in computational chemistry. PES is a relation between energy and geometry of a molecule.

Molecular geometry optimization means finding a local energy minimum on the hypersurface mentioned above - the Potential Energy Surface (PES). Optimization is done by gradually changing individual positions of the atoms, so that the energy after the change is smaller than before the change. Optimization arrives at its final geometry if no further adjustments of atomic positions are possible that would lead to smaller energy.

2.8 Conformational analysis

It is rather obvious that the properties of any chemical compound depend on its chemical structure, the chemical elements that the molecule is made of and how these elements are bonded together. But that is not all. In the physical world all chemical compounds have a specific 3D geometry, a particular configuration in space, specified by the x,y,z-coordinates of all its constituent atoms. As it turns out, two otherwise identical molecules may have different properties depending on their 3D-geometrical configuration. From that basic idea arises the motivation for conformational analysis - finding appropriate, “the best”, three-dimensional configurations of a compound. As each molecular geometry of a compound has its corresponding energy value, by connecting all possible configurations with the corresponding energy values we get an extremely high dimensional hyper surface called Potential Energy Surface (PES).

Conformational analysis is the study by which we analyse how molecular geometries are related to their corresponding energies. The lower the energy the more stable the structure is and vice versa. Conformational analysis is a major area of study in computational chemistry. The core challenge is how to sample chemical space, i.e., to identify structures that best describe the system under investigation [8]. Once the geometries have been generated or found, they need to be energy-optimized to a nearby minimum on the PES, these are called conformations. The best of the conformations, the low-energy structures with the lowest minima on the PES are called conformers. The procedure of adjusting a geometry, the optimization, to find its local minimum on the PES is the most compute intensive part, especially in case of using DFT methods if we are aiming at a higher chemical accuracy.

Computational complexity

We mentioned above the computational complexity of DFT. In conformational analysis we must combine this with the combinatorial complexity of possible configurations that atoms could occupy within a molecule. A sample of different 3D-geometries could be achieved by rotating parts of a molecule around interatomic bonds or just shifting atoms around in the space according to molecular dynamics. This gives rise to the combinatorial explosion of possible searchable geometries. Combinatorial complexity of molecular geometry configurations, if modeled with the molecular mechanics principles, scales with the number of interatomic bonds, as bonds determine the degrees of freedom in this type of modelling. Exhaustive enumeration of all the possible rotations around every bond scales exponentially with the number of rotatable bonds. That means the generation of conformers ranked by energy is computationally very demanding. To summarize, finding conformers is a two-step process:

- Generate a candidate 3D geometry.
- Calculate conformer energy.

Generating candidate geometries scales exponentially, so does the energy calculation if maximum theoretical accuracy is needed, according to the Schrödinger equation. In practice more precise methods than DFT could scale up to $O\left(n^7\right)$, depending on the chemical accuracy, where n is the system size in terms of number of electrons. The DFT algorithms that are implemented in practical applications have usually the complexity in range $O\left(n^{2.5}\right)..O\left(n^{3}\right)$, as higher scaling would be too slow to calculate any reasonable molecular system.

Due to the vast combinatorial search space and slow energy calculation, the common shortcuts are:

- To explore conformational space very sparsely through a combination of pre-defined distances and stochastic samples.
- To calculate conformer energies with faster and more inaccurate methods, e.g. force field or semiempirical methods.

Practical scaling in the case of TargetMol in terms of CPU time with different levels of DFT was as follows:

- Changing functional from BP86 to PBE0 increased CPU time on average 1.5 times.
- Changing the basis set from def2-SV(P) to cc-pVTZ increased CPU time on average 13.7 times.
- Changing functional and basis set from BP86/def2-SV(P) to PBE0/cc-pVTZ increased CPU time on average 20.7 times.

2.9 Key takeaways

- One of the main motivations is a search for molecules that have better properties to design better drugs, materials etc.
- The key question is how to efficiently calculate molecular properties so that they would be accurate enough, in compliance with quantum mechanical theories?
- The main computational bottleneck: according to the Schrödinger equation we need to calculate space integrals which are computationally extremely demanding and by designing suitable algorithms scientists try to find a good balance between accuracy and the speed of computation by approximating the space integrals.
- Density Functional Theory (DFT): Everything can be derived from electron density, and it is computationally possible in practice with high accuracy and

reasonable speed. DFT is a leading theory that offers an approximation to the solution of the Shrödinger equation.

- Conformational analysis is used to find “better” molecular geometries, which helps in designing new types of chemical compounds with desired properties.

3 Aim of the thesis

The main objective of this thesis is to find conformers for a titanium tartrate complex, code named TargetMol, containing 86 atoms including 2 titanium atoms. The last notable work involving TargetMol at TalTech was carried out in 2011.

Our main goals are:

- RG1: To find conformers for TargetMol, using CREST software.
- RG2: To assess the suitability of CREST software in our context.
- RG3: To establish a workflow that could assist similar research (conformational analysis) in the future, as well as to find practical suggestions and "short-cuts" by utilizing tools and methods of data science.

The additional research questions were:

- RQ1: To estimate the applicability of GFN2-xTB and GFN-FF methods for geometry optimization in our context, by answering the question: are the resulting geometries structurally valid and reasonable?
- RQ2: To estimate the relative performance of GFN2-xTB and GFN-FF methods with comparison to DFT methods, by answering the question: are relative conformational energies that are calculated by GFN-family methods, good predictors of relative conformational energies that have been calculated at DFT level?
- RQ3: Could a force field method GFN-FF be considered at all as an alternative method in our context or is GFN2-xTB, a semiempirical quantum mechanical method, the only serious candidate?
- RQ4: How well does the conformational sampling of CREST cover the search space and how well do the resulting CREST conformers cover the search space?

4 Methods

Our main aim was to use well established theoretical methods and apply these using large scale computational resources that are available for a working scientist in TalTech.

4.1 Generating geometries with CREST

To generate candidate geometries for conformer ensembles we utilized open-source software CREST [8] [30]. CREST is a conformer sampling program, named after the abbreviation from Conformer-Rotamer Ensemble Sampling Tool. For energy calculations and geometry optimizations, CREST uses either force field or semiempirical methods. We were particularly interested in two methods: GFN-FF and GFN2-xTB, where GFN stands for Geometries, Frequencies and Non-Covalent Interactions, FF stands for Force Field, and xTB stands for Extended Tight Binding, where "extended" emphasizes the parameter availability for almost the entire periodic table of elements ($Z < 87$). Motivation for using CREST:

- CREST promises to give state-of-art results in terms of output geometries.
- Applicable out-of-the-box for compounds covering most of the periodic table (elements up to 86, Radon).
- Very good computational efficiency.
- CREST authors claim that their software is more generic than analogous competing software.
- CREST, xtb software is public, free, actively maintained by authors.

CREST stands on two main pillars:

- Conformational search by extensive metadynamic sampling.
- Fast semiempirical and force field energy calculation by GFN family methods (GFN2-xTB, GFN-FF).

CREST provides an efficient balance between speed and accuracy due to fast energy calculators by GFN family methods. The key conformational search workflow implemented in CREST generates conformer/rotamer ensembles (CREs) by efficient sampling of the vast conformational space. The methods use either GFN parameterized xTB Hamiltonian (GFN2-xTB) or GFN parameterized force field (GFN-FF) for computational modelling of molecular structures, to explore the potential energy landscape. Authors say that GFN-FF is approaching the accuracy of semiempirical QM methods, in some cases even reaching DFT accuracy. The key premise of CREST is the computational efficiency of generating conformers with sufficiently high chemical accuracy. Although GFN family methods provide reasonably accurate results, in the overall workflow CREST is an efficient pre-screening tool and the results of CREST should be further refined at desired DFT level [8] [30]. Details of CREST experiments settings are presented in the table Appendix 3.

4.2 Density Functionals and Basis sets

As already mentioned, geometry optimization using DFT methods for larger molecules can be computationally infeasible. One must find an optimal balance between calculation speed and accuracy, meaning more exact DFT methods would require considerably more computational resources with much longer calculation times while producing more accurate results. Selecting appropriate DFT functional means also selecting an appropriate basis set. An expert chemist must know, guess or experiment to find out which combination works best in the context, by choosing among tens of DFT functionals and basis sets, whether to choose “pure” functional or “hybrid” etc. A “hybrid” DFT functional might have been parameterized with a set of chemical compounds, making it crucial to know its applicability in the context, while a “pure” DFT functional might not be parameterized and lack that kind of “bias”, but the general applicability issue remains - one must validate that the selected functional produces chemically accurate results in the context [20]. As for the choice of a basis set, with a too small basis set there is a risk of having too low accuracy, and with a too large basis set, the computations may become too expensive. Again, there is a need to find a good balance in between and it is a topic where one should do a thorough research beforehand to have up-to-date information of which density functionals and basis sets to consider. Fortunately, there are quite many research papers published on this very topic that could be extremely helpful [6], [20], [24], [26], [27], [28], [38]. Another interesting direction in the development DFT functionals are extensions to Density Functional Approximation (DFA) methods which are claimed to be much faster than “classical” DFT functionals and with the accuracies comparable to a higher level of DFT theory [14], [21], [39]. Unfortunately, none of these methods were implemented in the

Gaussian software at the time of our research. DFA methods such as r2SCAN-3c [14] which is implemented in the default workflow of CENSO, would be among the suggestions for candidates of density functionals to try in future research.

Our choice of DFT functionals and basis sets was based on experience and literature [6], [20], [26], [27], [28]. Most of the experiments were done using a “pure” BP86 and a “hybrid" PBE0 [23], [24] functionals (Gaussian keyword for PBE0 is PBE1PBE) and from the basis sets we used cc-pVTZ and def2-SV(P) [25] (Gaussian keyword corresponding to def2-SV(P) is Def2SVPP). We used density fitting as provided in the Gaussian software to speed up the calculations. We used PBE0/cc-pVTZ as “ground truth” while BP86/def2-SV(P) and PBE0/def2-SV(P) were the main workhorses with which we conducted a much broader range of experiments. Motivation for our selection was mainly driven by the experience and by the suggestions of experts in the field [24], [26], [27], [28]. The specification of functionals and basis sets that we used, as defined by Gaussian software is shown in Table 1.

Table 1. DFT functionals and basis sets specification.

DFT Functional	**Basis functions**	**Primitive gaussians**	**Cartesian basis functions**
BP86/def2-SV(P)	696	1318	740
PBE0/def2-SV(P)	696	1318	740
PBE0/cc-pVTZ	1952	3536	2228

4.3 Data driven analysis

To answer the research questions stated above, our aim was to perform various data analysis tasks. We can split data analysis into two categories by the purpose.

1. Predictive analytics, an analysis with which we aim to find shortcuts to the expensive DFT calculations, answering research question RQ2, e.g.:
   - correlation analysis of relative conformational energies to get useful information about the filtering of conformers and to assess the relative performance of GFN family methods in comparison with DFT methods.
   - prediction of conformers ranking.
2. Descriptive analysis by which we assess the quality of the conformers, answering research questions RQ1, RQ4, e.g.:

- analyze geometries to get better understanding of the distribution of certain structural properties, e.g., Ti-O-C, Ti-O-Ti angles and selected dihedral angles in the case of conformers generated by CREST with GFN2-xTB method as well as to assess the coverage of the search space by CREST conformers.
- analyze structural validity of the conformers.

4.4 Software

CREST [8]**:** One of the main goals of the thesis was to test the applicability and goodness of CREST software in the conformer generation process, especially in our context with a fairly complex transition metal compound.

Gaussian

All DFT level quantum chemical computations were done using Gaussian 16, Revision C.02 [31]. Gaussian software provides state-of-the-art capabilities for electronic structure modeling. It is a respected software in the computational chemistry community and has been used for many years.

ASE

The Atomic Simulation Environment (ASE) [35] is a set of tools and Python modules for setting up, manipulating, running, visualizing and analyzing atomistic simulations. ASE is a fantastic open-source and free library for scientific computing, more specifically an atomistic simulation environment - exactly what we needed. Many useful general chemistry and quantum chemistry functions and analysis tools etc. have been implemented. On top of that ASE has very broad coverage of API connections to pretty much all well-known quantum chemistry software providers (Gaussian, Turbomole, Orca, xtb, CP2K, QuantumEspresso, LAMMPS, Fleur, Psi4 etc).

Molli

A brand new open-source library on top of ASE written by the author as the side product of this thesis. The source code is available on GitHub [40]. All results and analysis presented in this thesis were made with the help of Molli. It was an essential companion for exploring and analyzing atomic structures, optimization trajectories, conformer ensembles, processing gaussian log files etc. In the GitHub repository there are some examples illustrating the capabilities and functionality of Molli. Hopefully Molli can be useful for future researchers. Further details about Molli are presented in Appendix 4.

Psi4 - Open-Source Quantum Chemistry

Psi4 [41] is an excellent quantum chemistry resource providing a wide variety of DFT implementations for single point energy calculations and geometry optimizations among other functionalities. We did quite some work with the Psi4 running a multitude of experiments in the AI-Lab compute centre. Documentation and code are well understandable and implementation, usability through Python API is quite user friendly. Psi4 has quite broad DFT functionals and basis set coverage. Parallelisation on 64 CPU-s seemed to work very well. The only problem for us was that we couldn't compile it properly to perform geometry optimization. Therefore, we had to abandon Psi4. As for single point energy calculation, Psi4 is a very good resource, especially considering it is an open-source and free product.

4.5 Hardware

Most of the conducted experiments required a lot of computational resources and these were done using the resources of TalTech High Performance Computing Centre (HPC Centre) [42] and TalTech AI-LAB [43]. AI-Lab is a sandbox environment consisting of GPU equipped workstations that provides a stepping stone for students and staff to efficiently use the order of magnitude larger resources of the HPC Centre. Usually, our single jobs utilized 64-240 CPU threads each. Some calculations performed with Gaussian 16c02, also included the use of NVidia A100 GPU-s, but these did not seem to offer much of a speed advantage (order of magnitude 10-20%) for TargetMol, not to mention considerably higher electricity consumption.

TalTech High Performance Computing Centre (HPC Centre) develops and manages the compute resources for scientific use. HPC provides following compute resources (as of Jan 1, 2023):

- 32 nodes with: 2 x Intel Xeon Gold 6148 2.40 GHz (40 cores, 80 threads per node), 96GB RAM, 25 Gb/s Ethernet, 800GB local scratch space
- GPU server AMP.HPC.TALTECH.EE: 2 x AMD EPYC 7742 (128 cores, 256 threads per server), 8 x NVidia A100 with 40GB RAM GPU, 1TB memory, 100Gb/s Ethernet.
- GPU server AMP2.HPC.TALTECH.EE: 2 x AMD EPYC 7713 (128 cores, 256 threads per server), 8 x NVidia A100 with 80GB RAM GPU, 2TB memory, 100Gb/s Ethernet.

Taltech AI-LAB is an environment to learn how to use modern computational resources to solve various problems that require substantial computational resources. AI-Lab hosts a varying number of workstations using the 24 and 32 core (48/64 thread) 3rd generation AMD ThreadRipper processors (3960X or 3970X), or a 10-core 10th generation Intel Core i9-10900X processor, 128 GB of memory and NVidia or AMD graphics cards for CUDA or ROCM computations.

5 Results and Analysis

We started with a single molecular geometry of the target molecule, a titanium tartrate complex consisting of 86 atoms, including 2 titanium atoms. The geometry was a local optimum geometry at the theory level of BP86/def2-SV(P) calculated with Turbomole software. Our aim was to find geometries that would be at a better local optimum. Our approach was the following: use CREST software to produce as many candidate geometries as possible, then perform geometry optimization at a higher-level theory, ultimately at the DFT level of PBE0/cc-pVTZ. As this level is computationally too expensive, we estimated that we could get a broader coverage of geometries optimizing also at the levels of PBE0/def2-SV(P) and BP86/def2-SV(P). Initially we considered using STO-3G as the basis set to cut down calculation budget even more, but our preliminary experiments using PBE0/STO-3G and BP86/STO-3G were not too promising. Although being considerably faster, they seemed to lack accuracy in our context. Considering this and the limited available time we decided not to focus too much on the experiments with STO-3G basis set. cc-pVTZ basis set, however more accurate theoretically, was obviously too large for performing geometry optimizations on a larger scale. The def2-SV(P) basis set seemed to offer a balanced sweet spot between speed and accuracy. Combined with the two well-known and widely used functionals BP86 and PBE0 became our main workhorses for quantum chemical computations.

Our goal was twofold:

- To test how capable is CREST software in our context.
- To find a better set of conformers than the currently available single geometry that we used as main input in our experiments.

5.1 Overview of main results

Timing report in Table 2 gives the overview of computational resources used and the number of experiments concluded. For a comparison we performed full optimization on selected geometries also with other well-known functionals that were suggested in the literature (TPSS, B3P86) [3], [4], [18], [26] and by prof. Tamm (wB97XD).

Table 2. Timing report of geometry optimization.

DFT functional	Optimization step, CPU Hours	Total time, CPU Days	Total Wall Time, Days	Number of experiments
TPSS/STO-3G	0.75	82.2	1.4	38
PBE0/STO-3G	0.78	90	1.6	69
BP86/STO-3G	0.86	225.9	4	288
TPSS/def2-SV(P)	2.66	20.7	0.4	4
B3P86/def2-SV(P)	3.62	38.9	0.6	4
BP86/def2-SV(P)	3.79	721.8	12.8	318
PBE0/def2-SV(P)	4.26	2438.8	40.1	269
wB97XD/def2-SV(P)	4.89	41.2	0.7	4
PBE0/cc-pVTZ	78.41	2490.9	11.6	10

Main results of the conformational analysis are presented in the Table 3, which include 10 geometries fully optimized at the level of PBE0/cc-pVTZ. Conformer “original” refers to the single input geometry of TargetMol that we started with and arguably it has been fully optimized at the level of BP86/def2-SV(P). Naming convention of conformers is explained in chapter 5.2 and full details of CREST experiments are presented in Appendix 3. Conformer “ex16_c5_def2svpp_step10” is the “head start” version of “ex16_c5”, meaning that it was first optimized with 10 steps at the level of PBE0/def2-SV(P) and then continued until convergence with PBE0/cc-pVTZ. Analogously “ex0a_c10_bp86def2svpp_step10”, but with the 10 step pre optimizing at the level of BP86/def2-SV(P). In this case the pre optimization leads to an even better conformer. This seemingly minor fact has quite an interesting implication. Namely not only can pre optimization be used as a tool for making an end-to-end process considerably faster, but it could also be used as a tool for generating variety in high quality conformers. Column “Energy delta” is presented to give an indication of how far the input geometries were from the fully optimized geometries at the start of the optimization. Energy delta is defined as the energy at the end of optimization minus the energy at the start of the optimization (starting energy corresponds to energy of CREST output). We can make following observations:

- 10-step pre optimization covers 91% (PBE0, ex16_c5) and 76% (BP86, ex0a_c10) of the energy delta compared to full optimization.
- Rather large energy delta values indicate that CREST output geometries are pretty far from optimal by DFT methods.
- CREST geometries optimized with GFN Force Field (GFN-FF) method have approximately double the energy delta value compared to the GFN2-xTB method.

Table 3. Conformers optimized at the theory level of PBE0/cc-pVTZ.

Index	Conformer	Energy Difference to best, kJ/mol	Final Energy, hartree	Energy delta, optimization start to end, kJ/mol
1	ex0ff_c6	0.00	-3842.08132304	-606
2	ex16_c5	5.58	-3842.07919842	-242
3	ex16_c5_def2svpp_step10	5.58	-3842.07919827	-21
4	ex0a_c10_bp86def2svpp_step10	6.06	-3842.07901627	-64
5	ex0a_c10	6.35	-3842.07890493	-273
6	ex19_c23	8.78	-3842.07797753	-242
7	ex0a_c24	9.09	-3842.07786203	-261
8	ex0a_c26	11.95	-3842.07677058	-268
9	ex0b_c21	26.92	-3842.07106838	-241
10	original	112.77	-3842.03837075	-57

In addition (including above-mentioned geometries), we fully optimized 205 conformers at PBE0/def2-SV(P) and 34 conformers at the theory level of BP86/def2-SV(P). For detailed results see Appendix 2.

Table 4 presents the general statistics of the fully converged experiments including the information about sample size, average number of steps it took to converge and average change in the energy across all converged experiments.

Table 4. General statistics of the fully converged experiments.

DFT Functional	Sample size	Avg steps to converge	Avg energy delta, kJ/mol
PBE0/cc-pVTZ	8	90	-274
PBE0/def2-SV(P)	205	63	-311
BP86/def2-SV(P)	34	92	-422

Stationary points and convergence

Although all found conformers have been optimized until convergence, we cannot claim that these stationary points are also true local minima. For this purpose, frequency analysis would be needed and could be performed by later research if deemed necessary. To be more specific, all our geometry optimizations using Gaussian software were run with the additional command option *"CalcFC"*. From the Gaussian official website one can read the following: "CalcFC specifies that the force constants are computed at the first point. Alternatively, keyword CalcAll specifies that the force constants are to be computed at every point. In this case the vibrational frequency analysis is automatically done at the converged structure" [44] and "In a geometry optimization, an estimated Hessian is used unless you explicitly request a computed one using the CalcFC or CalcAll option to the Opt keyword. As is well known, frequency and thermochemistry results are based on a harmonic analysis that is only valid at true stationary points. Accordingly, some results will be incorrect at non-stationary points" [45].

Numerical instability

Quantum Chemical calculations require high numerical precision. During geometry optimization the significant digits can very well be 8 decimal places. Arguably even different hardware could produce different results (e.g., CPU vs GPU, software compiling parameters etc.). We faced a similar problem at least once (of what we are aware of), namely two otherwise identical geometry optimization experiments using Gaussian software, one using GPU-s, the other not. One experiment converged, the other didn't. By investigating and comparing Gaussian log files, we noted that the divergence started after the first optimization step with the numerical difference of a single parameter occurring at the 8th decimal place. From there on the divergence amplified with each successive step.

Observations

Experiments suggest that it is reasonable to optimize CREST outputs at first with faster DFT up to some number of steps to pre-screen. Further refinement could be done either by continuing pre-screened optimizations to convergence at the level of desired DFT or by calculating single point energies at the level of desired DFT on fully optimized geometries obtained with a faster DFT method. This common approach helped us significantly reduce the amount of computational resources needed. By comparing the results of medium and higher levels of DFT, we saw that input geometries converged not only by the measure of final energy but also by the measure of achieving near identical final geometries, which is a good sign and indicates that in our case different DFT levels “understand” the Potential Energy Surface (PES) quite similarly. Obviously, we cannot draw any fundamental conclusions, as any research situation must be assessed case by case. Finding shortcuts is rather a practical suggestion for the future, especially if these shortcuts lead to identical experimental results. In our experiments the resource savings of a single optimization run were up to 50% (i.e., 2 times less), 42% reduction in wall time (16 hours instead of 28), corresponding to saving approximately 125 CPU-days per single experiment (pre-optimizing by 10 steps with def2-SV(P) basis set and continuing to the full convergence with larger cc-pVTZ basis set).

5.2 Searching conformers with CREST

To generate candidate geometries for conformer ensembles we utilized open-source software CREST [8]. Our first results from experiments with xtb were not very promising. It seemed that the Force Field method, GFN-FF, morphed the chemical structure of our TargetMol by quite a bit. Semi-empirical GFN2-xTB was more promising in that respect, however we were not convinced. Reason being that certain Ti-O-C angles were bent too much. As it turned out, this fact has been observed in other publications [5]. We continued two-fold:

- Let GFN methods work as is.
- Constrain Ti-O-C angles, namely the ligands connected to titanium atoms to their original value, as well as the angles of O-Ti-O, that form the central metal core of the molecule.

At first we focused only on the second alternative, constraining Ti-O-C angles. However, later on we proceeded with both paths, ignoring the seemingly distorted nature of the intermediary results. The important caveat is to properly validate that geometries are structurally correct and correspond to the structure of our TargetMol (see chapter Validation of structural correctness). Some general observations regarding our experiments with CREST:

- GFN-FF vs GFN2-xTB: otherwise identical experiments had wall time differences 30-70 times in favour of GFN-FF, e.g., 12 minutes vs 7.5 hours on a 64 CPU server (more details about timings regarding CREST experiments can be found in Appendix 3). GFN-FF should be regarded seriously as an option when using CREST for conformer search as it is very fast.
- Re-running CREST experiments would find different geometries. Conformational search is implemented as a (partly) stochastic process. It is not clear whether the user could control all the necessary variables seeding the randomness inside CREST. However not being able to control the randomness is not inherently a bad thing per se. Re-running the same experiment multiple times would give even a large sample of possible geometries for further analysis.

Setup strategies for conformational search by executing CREST:

1. With default parameters multiple times using original input.
2. With angle-constrained parameters multiple times using original input.
3. With angle-constrained parameters multiple times using some of the best results from previous strategy as input.

Appendix 3, Table A3.1. gives more details. From there we can see that experiment 19 (ex19) uses CREST-s 5th best result from experiment 16 as input geometry. Experiments with the prefix “ex0” mean experiments with default settings in CREST. This kind of recursive way of generating new sets of conformers might be a good option to try out if there is a need for ever more geometries. The underlying assumption here is that the configuration space of geometries can be considered infinite for all practical purposes and the same goes for the immense size of the Potential Energy Surface (PES). As the process of searching for conformers is partially guided by randomness then we are practically guaranteed to get a distinctly new set of geometries each time.

To get an indication about the variety of generated conformers we performed a statistical analysis of selected angles and dihedrals. The results are presented in Appendix 3, Figure A3.2. This kind of analysis should give a preliminary idea whether generated conformers cover all “interesting” areas of the search space.

“noreftopo” - a useful keyword in CREST

As defined in CREST documentation for the keyword *“noreftopo”*: “Turn off only the initial topology check prior to the conformational search.” [46]. As it appeared then our TargetMol topology changed while performing initial energy optimization either with GFN2-xTB or GFN-FF method. The change in topology caused CREST to not start the conformational search. The question of what is the correct topology and whether the concept of topology is even relevant in the context is up to a researcher to decide as molecular topology entails predefined definitions of bonding properties of different chemical elements. In our case we let the CREST ignore whatever it thinks the topology is and whether it changes. We tied 3 alternative solutions for the “change in topology” error:

- Fix absolute positions of certain atoms. Result: no need for the use of *“noreftopo”* keyword at the expense of restricting the conformational search space too much. The centre of the molecule was essentially fixed in space and possible conformations were very limited.
- Use the *“noreftopo”* keyword and fix certain Ti-O-C angles as given by the input structure. Result: worked well, however the search space became limited most probably, but a much better result than described in previous point with fixing of absolute positions.
- Use *“noreftopo”* keyword and let CREST work with otherwise default settings as it comes “out-of-the-box” after the installation. Results: worked well, larger variety of possible conformations as the search space is not further constrained by the user.

5.3 Data driven approach for selecting the right subset

Quantum chemical calculations produce vast amounts of intermediary data which conveys useful information that could be analyzed to get better insight. This in turn could lead us to discover better solutions.

Our main goal is:

- To get as many “good” geometries as possible.
- Try to gradually filter out more promising geometries.

The important caveat here is obviously how small or large is the sample size, i.e., number of candidate geometries. An end-to-end workflow in a conformational analysis will most probably be an “expert-in-the-loop” system in the foreseeable future. There is no golden rule that would apply for every situation, unless we achieve computational power big enough that we could always calculate everything at an arbitrary level of theory. The latter is extremely unlikely to happen in the next 5-10 years to stay on the modest side. In our case we had generated close to 300 candidate geometries in total. If it were 3000 then we should revisit the specific details of the workflow, however the core principle of selecting the right subset would stay intact. Exhaustive full optimization on all candidates is computationally prohibitive because the estimated average time for a single full optimization in our experience was approximately 15 hours with 240 CPU threads. Even performing a single point energy calculation on every geometry could be computationally impossible in practice due to the constraints on affordable resources. So, the question is how to select a good subset from a large set of candidate geometries? Selecting the “right” subset of geometries for the final refinement as conformers from a potentially huge number of candidate geometries is a major computational problem. To gradually filter candidate geometries produced by CREST we needed to develop a heuristic that would allow us to accurately group geometries as fast as possible. The first natural thing to do is to pre-screen a larger sample with a faster and more inaccurate DFT so that we could cover a broader number of geometries. Pre-screening with a faster DFT relies on the correlation assumption, namely that there is a correlation between relative conformational energies between different DFT functionals. If the correlation assumption holds then we could gradually filter out more relevant geometries much faster. Why is there a need to do expensive

DFT calculation, if the geometries are already of good quality, as promised by CREST authors? The answer is: no, you don’t necessarily need to further fine-tune the geometries at a higher level of DFT. But this applies only in case if the accuracies provided by the semiempirical GFN(x)-xTB and/or molecular force field based GFN-FF methods are good enough for one’s research. It is known that transition metal complexes are extremely challenging to model and the problem transcends also to GFN methods with published reports that GFN methods might systematically distort the geometries of transition metal compounds [3], [5]. These are the main reasons why we should first refine CREST generated geometries with a more accurate DFT method. We observed immediately that GFN2-xTB and GFN-FF optimized geometries tend to be quite far from DFT optimized ones. As we had approximately 300 geometries, we assessed that we could optimize every single geometry with a DFT up to a certain number of steps to transform them closer to their more accurate DFT optimized structure. If there would be thousands or even more candidate geometries, then it would be practical to rank candidates based on some descriptor that could be calculated more efficiently, e.g., rely on GFN2-xTB energy and/or calculate single point energies at a faster DFT level or combine these two measures etc. There are various ways to approach this in practice. Grimme lab [47] has published an excellent software CENSO [48]. CENSO is a threshold-based sorting algorithm where the thresholds have been determined for typical drug-like organic molecules up to 200 atoms. CENSO has implemented a workflow that allows a researcher to filter conformer candidates by gradually ranking and refining at an increasing level of theory. Although we didn’t use CENSO in this research, we feel that CENSO is a very good candidate for this kind of task. Another alternative for ranking would be a trained machine learning model that could predict the difference between GFN2-xTB energy and DFT level energy, e.g. DelFTa, an open-source toolbox for the prediction of electronic properties of drug-like molecules at the density functional (DFT) level of theory [32]. Unfortunately, both above mentioned solutions didn’t seem to fit our needs and we opted to implement our own rather simple and down-to-earth workflow for filtering.

Main reasons for us not choosing CENSO were:

- There is no Gaussian software backend for DFT calculations.

- CENSO is targeted for sorting a considerably larger size of conformer set and therefore uses two steps of “cheap pre-screening”, meaning calculation of single point energies on outputs from CREST. Our initial tests showed that there is no correlation between relative conformational energies calculated by GFN and DFT methods and therefore in our case it is crucial to optimize CREST outputs at least to some extent with a DFT method to get a reasonable indication of how good the conformers are relative to each other. However later calculations on a broader set of CREST conformers with subsequent full optimization at level of PBE0/def2-SV(P) revealed a significant correlation of 0.75 between GFN2-xTB and DFT results, while GFN-FF results showed insignificant correlation of 0.22, and mixed together the correlation was -0.1. More details about the correlations are given in the next chapter.

Reasons for not pursuing a Machine Learning based approach, i.e., use a prediction model trained by machine learning:

- The field is just emerging, and the available research and models cover a very narrow selection of chemical compounds. They either work as a proof-of-concept or even if their claimed performance is excellent, they cover compounds consisting of only the most common chemical elements used in organic chemistry, e.g., Delfta [32] covers H, C, O, N, F, S, P, Cl, Br, I, which is probably in the high end of the element coverages among analogous models.
- Machine learning models as far as we came across, are all restricted to the representation of inputs (most importantly internal representation) using SMILES strings [49]. SMILES strings contain the information that can be converted into two-dimensional structure diagrams. This fixed topology of a 2-dimensional structure is usually implemented through Graph Neural Networks (GNNs) [50] in a typical library based on Machine Learning. It may be due to inexperience, but we never obtained expert accepted structural graphs from SMILES strings generated by Avogadro [33] from xyz-file and processed later by either OpenBabel [34] or RDKit [51] to draw the structural graph of TargetMol. OpenBabel and RDKit being two of the most popular open-source and free software for the conversion in between different formats. The fundamental problem here seems to be that an implementation of a SMILES

string generator may output a chemical structure that does not match with your understanding of the chemical structure corresponding to the input. The inclusion of titanium atoms may have been a major cause of ambiguity.

- Probably there will be more capable machine learned models off the shelf in a couple of years that are suitable for more challenging chemical compounds.

5.4 Correlation between DFT functionals

Filtering based on partial optimization or single point energy thresholds is a common practice. Our workflow as well as CENSO and DelFTa are based on the same fundamental assumption that there is a correlation between some properties or descriptors between calculations at different levels of theory. If the correlation assumption doesn't hold then the filtering results would be close to random. From that we formed our first hypotheses to be tested. Namely, as our input data was the result of the optimizations using different DFT functional / basis set combinations, there were two main ideas we wanted to test: (i) whether geometries obtained by different DFT functionals tend to converge/diverge, (ii) how strong is the correlation between relative conformational energies that are calculated using different DFT functionals. We define "relative conformational energy" as the difference in the final energy between a conformation and the best conformer in the ensemble and ground truth in our case means values calculated by DFT functional PBE0/cc-pVTZ.

- **Hypothesis 0:** There is a strong correlation between relative conformational energies calculated by different DFT functionals on the same inputs. It is the fundamental assumption that must hold before anything else. If true, then it means there is an inherent correlation between different DFT functionals, implying that given arbitrary geometries, fast DFT would rank them very similarly to ground truth. Meaning that by performing single point energy calculation with fast DFT gives similar ranking as with ground truth. Basically, the same assumption is used in CENSO in its fast pre-screening phases.
- **Hypothesis 1:** Relative conformational energies during the early steps of the optimization with fast DFT correlate well with the relative conformational energies after the final step of full optimization with the same fast DFT. We compare relative conformational energies across optimization runs with the

same DFT. If true, we could predict the ranking after full optimization, by observing the relative conformational energies in the early stages of the optimization without necessarily performing full optimization with a fast DFT.

- **Hypothesis 2:** Relative conformational energies during the early steps of fast DFT calculations correlate well with the relative conformational energies after the final step of a full optimization with the ground truth. We compare relative conformational energies across different DFT. If true, we could predict the ranking of the full optimization with the ground truth, by observing the relative conformational energies in the early stages of fast DFT.
- **Hypothesis 3:** Differences between geometries during the early steps of fast DFT calculations correlate well with the differences between geometries after the final step of full optimisation with the ground truth. We compare geometries across different DFT methods. If true, we could assert that fast DFT and ground truth both end up in the same local optimum geometry or very close.

Some clarifications:

- It is important to note that grounded on the correlation assumption, we do not predict the final energies nor the differences in the final energy. We test how well do the relative conformational energies correlate between different DFT functionals and also between different optimization steps. Filtering could done by ranking or by energy threshold. By using different energy thresholds, we could guide the filtering, with smaller thresholds we would keep a smaller sample of very best and with larger threshold we would keep a larger sample making sure that we would not lose any good candidates.
- Obviously, the above-mentioned hypotheses cannot be rigorously tested nor proven in the current context because of a small sample size. If we were to collect a large sample, then this whole exercise would lose its meaning because then we could just calculate everything and always use ground truth results. The main point here is merely to get an insight into whether it would be plausible to assume that different DFT functionals produce well correlated results in our case.
- The analysis of relative performances between different DFT methods is quite common in validation of the applicability of a particular DFT method in

particular context [2], [21], [27] etc. However it must be emphasized that our results may not be necessarily taken as a broad generalization beyond our context.

- Problem of finding a good approximating predictor is very similar to the hyperparameter search in machine learning, which is often known to resemble more art than science, mainly because there exists no provably golden solution.

Correlation throughout this work is defined as *Pearson correlation coefficient*, calculated by formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \qquad (5.1)$$

where:

- n is sample size.
- x_i , y_i are the individual sample points indexed with *i*.
- $\overline{x}$, $\overline{y}$ are the sample means correspondingly.

Tables 5-7 give the overview of the correlations between different DFT methods, but most importantly the data gives strong evidence for the support of Hypothesis 0.

Hypothesis 0 holds: different DFT functionals have strong intrinsic correlation. Some observations:

- There is a near perfect positive correlation between all DFT methods “At start”: energy calculation in the beginning of the optimization run, i.e., on CREST output. The implication: all selected DFT methods “understand” the chemical structure of our TargetMol almost identically. That is very good news, because we were certain that the STO-3G basis set would be too small to accurately model the structure of the TargetMol and we stopped experimenting with STO-3G in quite early stages. Based on this data BP86/STO-3G would be a solid candidate for cheap pre-screening based on single point energies.
- There is a near perfect positive correlation after 25 steps of optimization with either PBE0/def2-SV(P) and BP86/def2-SV(P) against PBE0/cc-pVTZ. The

implication: PBE0/def2-SV(P) and BP86/def2-SV(P) are both very solid candidates for pre-screening with 25-step partial optimization. The data indicates that the relative conformational energies after 25 steps of optimization correlate near perfectly with the ground truth results after full optimization to convergence.

- There is a near perfect positive correlation after full optimization with either PBE0/def2-SV(P) and BP86/def2-SV(P) against PBE0/cc-pVTZ. The implication: not only does it reiterate the previous point, but it also gives a strong indication about the absolute quality of a conformer. Meaning that by performing a full optimization at the level of PBE0/def2-SV(P) or BP86/def2-SV(P) should give a very good prediction whether a particular conformer ends up with better or worse relative conformational energy if calculated at the level of PBE0/cc-pVTZ. All this indicates strongly that in our context much of the computational work in pre-screening phases could be done with PBE0/def2-SV(P) and/or BP86/def2-SV(P)
- Substantially degrading correlation in optimizing with BP86/STO-3G. The implication: BP86/STO-3G is probably not a good candidate for geometry optimization in our case.
- Drop in correlation after 10 steps while optimizing with BP86/def2-SV(P). The implication: 10 steps seem to be too early as the cut-off point for optimization.

Table 5. Correlation of relative conformational energies for fully optimized cases vs PBE0/cc-pVTZ. Comparison is done at the same step nr or at convergence.

DFT Functional	At start	After 10 steps	After 25 steps	Converged
PBE0/def2-SV(P)	0.999	0.994	0.994	0.992
BP86/def2-SV(P)	0.930	0.718	0.998	0.995
Sample size	8	8	8	8

Table 6. Correlation of relative conformational energies, BP86/def2-SV(P) vs PBE0/def2-SV(P). Comparison is done at the same step nr or at convergence.

DFT Functional	At start	After 10 steps	After 25 steps	Converged
BP86/def2-SV(P)	0.983	0.426	0.957	0.984
Sample size	55	55	34	31

Table 7. Correlation of relative conformational energies vs PBE0/def2-SV(P). Comparison is done at the same step nr.

DFT Functional	At start	After 10 steps	After 25 steps
BP86/def2-SV(P)	0.966	0.850	0.960
BP86/STO-3G	0.982	0.727	0.544
Sample size	33	33	28

Table 8 shows the correlation of relative conformational energies between early steps vs fully optimized structures, providing evidence that Hypothesis 1 holds at 25-step and fails for shorter pre-optimizations: 10-step and single point energy at-start. More detailed results for BP86/def2-SV(P) are provided in Appendix 2, table A2.3.

Table 8. Correlation of relative conformational energies against converged structures. Comparison is done between values at particular step numbers vs values at convergence.

DFT Functional / source	Converged	At start	10 steps	25 steps	Avg steps in optimization
BP86/def2-SV(P)	1.00	-0.36	0.33	0.89	
PBE0/def2-SV(P) / GFN2-xTB	1.00	0.75	0.62	0.89	51
PBE0/def2-SV(P) / GFN-FF	1.00	0.26	0.02	0.41	110
PBE0/def2-SV(P)	1.00	-0.10	0.25	0.77	

Hypothesis 1 holds at 25-steps, some observations:

- Correlation “At start” vs “Converged” is significant if CREST conformers have been optimized with GFN2-xTB. This is an important finding and confirms the correlation assumption underlying CENSO, which uses GFN2-xTB single point energies for pre-screening.
- Correlation “At start” vs “Converged” is insignificant for GFN-FF method or if conformers from both methods are mixed together. It shows that we cannot use energy numbers directly coming from optimization with GFN-FF method or if conformers from GFN-FF and GFN2-xTB methods are mixed together. It is extremely important to know which method was used in the optimization of conformers.

- Correlation “25-steps” vs “Converged” is 89% for GFN2-xTB and 41% for GFN-FF for PBE0/def2-SV(P), from which we could conclude that 25 steps is not enough for an adequate decision in case of conformers from GFN-FF method. This might be explained with the fact that 25 steps covers on average 50% of the full optimization steps in case of GFN2-xTB, as opposed to 23% in case of GFN-FF conformers. Finding an optimal pre-optimization schedule, means finding a suitable sweet spot for each case.

In addition to correlation between different optimization steps, let's assess the predictive power of the rank prediction. Table 9 shows that with a larger sample size, after 25 steps of optimization with PBE0/def2-SV(P) we could predict correctly 87.5% of the top 50% of the fully optimized conformers, meaning that 87.5% of the top 50% conformers in final ranking were included in the top 50% conformers after 25 steps. Considerably lower number for BP86/def2-SV(P) might be explained by a smaller sample size.

Table 9. Rank prediction: coverage of top 50% conformers compared to fully optimized top 50%.

	At start	**10 steps**	**25 steps**	**Sample size**
PBE0/def2-SV(P)	51.6%	54.7%	87.5%	128
BP86/def2-SV(P)	64.7%	70.6%	70.6%	34

Table 10 shows the results of correlation analysis between PBE0/def2-SV(P) and converged structures of PBE0/cc-PVTZ, corresponding to our Hypothesis 2: fast DFT vs ground truth. Similar table but with BP86/def2-SV(P) data is given in Appendix 2, Table A2.4.

Table 10. Relative conformational energies at different steps of optimization, correlation vs “ground truth”. Ground truth means fully optimized at PBE0/cc-PVTZ, units kJ/mol except correlation.

Label	**Ground Truth**	**PBE0/def2-SV(P) 10-steps**	**PBE0/def2-SV(P) 25-steps**	**PBE0/def2-SV(P) Converged**
ex0ff_c6	0.00	0.00	0.00	0.00
ex16_c5	5.58	-62.05	-10.37	-1.29
ex0a_c10	6.35	-59.73	-13.81	-5.77
ex19_c23	8.78	-61.55	-13.73	-4.57
ex0a_c24	9.09	-54.96	-9.92	-1.54
ex0a_c26	11.95	-44.97	-1.61	0.48
ex0b_c21	26.92	-32.49	16.18	20.45

original	112.77	40.61	91.44	91.10
Correlation	**1.0000**	**0.7964**	**0.9804**	**0.9917**

Hypothesis 2 holds at 25 steps, some observations:

- Unexpectedly high correlation between 25 steps of fast DFT vs ground truth might be explained by the small and lucky sample. As shown in Table 8 above, the internal correlation of PBE0/def2-SV(P) was in range 0.41-0.89, suggesting that we shouldn't expect correlation above 0.89, as predicting itself should always be more precise than predicting something else.

Hypothesis 3 holds as tested geometries converge.

To test Hypothesis 3, we set up two experiments by comparing the geometries of fully optimized structures at PBE0/cc-pVTZ. We took two geometries and fully optimized them at PBE0/cc-pVTZ, forming a ground truth set. On the other hand, we optimized the same two initial geometries as follows: first geometry optimized 10 steps with PBE0/def2-SV(P) and continued with PBE0/cc-pVTZ until convergence. Second geometry optimized 10 steps with BP86/def2-SV(P) and continued with PBE0/cc-pVTZ until convergence. From the Table 3 above we already saw that from an energy standpoint these experiments ended in pretty much identical results. Experiments under discussion are:

- ex16_c5 vs ex16_c5_def2svpp_step10
- ex0a_c10 vs ex0a_c10_bp86def2svpp_step10

We analyzed further the final geometries and found that these are nearly identical. Convergence of optimization trajectories can be seen on Figure 3. The blue line represents ground truth and the orange line represents the pre optimized version. Orange line starts from a lower point because it is closer to the destination - distance from the fully optimized ground truth structure. From that we concluded that Hypothesis 3 holds.

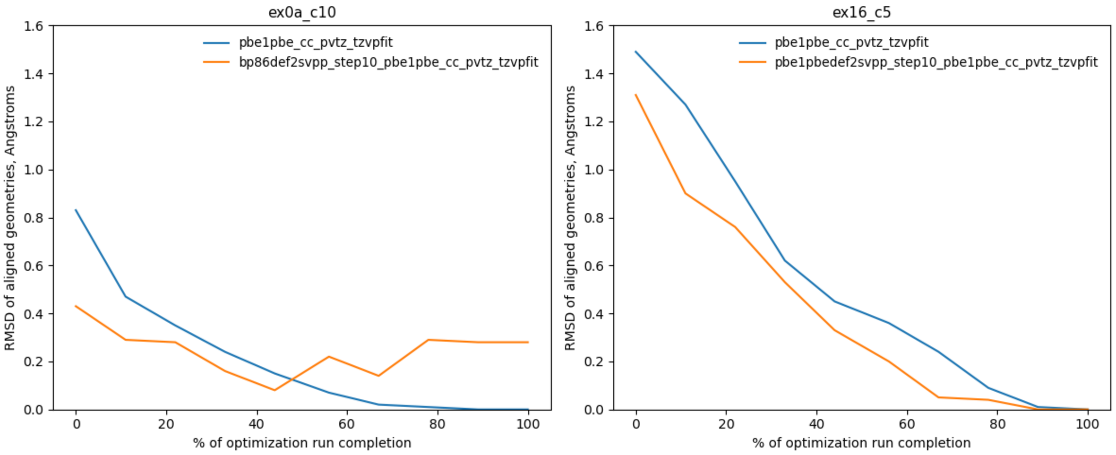


Figure 3. Comparison of optimization trajectories: end-to-end optimized with PBE0/cc-pVTZ (blue line) vs 10-step pre optimized with BP86/def2-SV(P) (case: ex0a_c10) or PBE0/def2-SV(P) (case: ex16_c5) and continued with PBE0/cc-pVTZ until convergence.

5.5 Partial optimization strategy

Previous section presented the high correlation between the calculation results between chosen DFT functionals. That in turn gives rise to the idea of the partial optimization strategy. The main idea is to minimize end-to-end time of full optimization with the desired accurate and slow DFT while not sacrificing the accuracy and the quality of the results. To emphasize once more, this is purely a suggestion for resource optimization, not a revolutionary idea that qualitatively improves the results themselves. By comparing head start optimization where we pre-optimize with fast DFT some number of steps with the full optimization at the desired slow DFT, the idea relies on a claim that both optimizations end at the same geometry, while starting from scratch with the desired slow DFT is considerably more time consuming. As already mentioned above, in our experiments where we tested this idea, the resource savings of a single optimization run were up to 50% (i.e. two times less), 42 % reduction in wall time (16 hours instead of 28), corresponding to saving approximately 125 CPU-days per single experiment (pre-optimizing 10 steps with def2-SV(P) basis set and continuing to the full convergence with larger cc-pVTZ basis set). Saving 125 CPU-days per single optimization is impressive, but our data indicates that we could possibly do even better, namely by starting with a 25-step pre-optimized structure. Correlation analysis above indicates clearly that 25-step pre optimization correlates even better with the end result than 10-step pre optimization. Unfortunately, we didn’t have time to test this in practice. It might even be that the optimal way is to start optimizing at the desired slow DFT with the structure that is fully optimized or close to full convergence by a fast DFT. These kinds of decisions must be made on a case-by-case basis, because there is no general answer that satisfies all scenarios. Based on the results above, to perform a DFT optimization on a broader sample starting from CREST conformers, our suggestions are as follows.

Group conformers based on the optimization method they were calculated:

- e.g. GFN2-xTB, GFN-FF methods, as discussed above, have completely different correlation profiles compared to DFT methods. Pre-optimization schedules must be adjusted accordingly.
- In case of conformers optimized with GFN2-xTB, one can consider filtering directly using single point energies calculated with GFN2-xTB, as does CENSO. As shown above, for the sample size around 200, the correlation between relative conformational energies calculated with GFN2-xTB and PBE0/def2-SV(P) was 0.75, which is rather significant.

If resources allow:

- 25-step optimization with PBE0/def2-SV(P).
- Make final selection.

If less resources or larger initial sample:

- 10-step optimization with BP86/def2-SV(P).
- 25-step optimization with BP86 or PBE0/def2-SV(P).
- Make final selection.

Even less resources or much larger initial sample:

- 10-step optimization with BP86/def2-SV(P).
- 10-15-step continuation from previous step with BP86/def2-SV(P).
- 10-15-step continuation from previous step with PBE0/def2-SV(P).
- Make final selection.

5.6 Validation of structural correctness of molecular geometries

Based on the TargetMol we defined a set of criteria that had to be satisfied to be validated as a correct structure, e.g., certain interatomic bonds that had to be retained.

The general validation criteria were that all Ti-O, C-O, C-C and C-H bonds should stay intact, and neither removal nor creation of any such bonds should occur. The atomic neighbour lists and corresponding interatomic bonds were determined using covalent radii as implemented in ASE software [35], which are based on [36].

Summary of the results of structural validation is presented in Table 11 and Table 12.

Table 11. Structural validation of fully optimized conformers: (A) CREST conformers, (B) Fully DFT optimized conformers, % of valid structures.

A. CREST conformers

Optimization method	At start	Sample size
GFN-FF	100%	218
GFN2	43%	172

B. Fully DFT optimized conformers

DFT Functional	At start	After 10 steps
BP86/def2-SV(P)	28%	100%
PBE0/def2-SV(P)	30%	100%
PBE0/cc-pVTZ	40%	100%

Table 12. Structural validation of all conformers, % of valid structures.

DFT Functional	At start	After 10 steps	After 25 steps	After convergence
BP86/def2-SV(P)	59%	100%	100%	100%
PBE0/def2-SV(P)	56%	100%	100%	100%
PBE0/cc-pVTZ	40%	100%	100%	100%
BP86/STO-3G	62%	37%	21%	
PBE0/STO-3G	6%	6%	3%	

Validation errors are all caused by a single interatomic distance that becomes too short and gets caught by the validator: Atoms O(31) and C(42) with the original distance 3Å vs 1.6-2Å in conformers. While atom O(31) belongs to a ligand connected to one titanium atom and atom C(42) belongs to a ligand connected to the other titanium atom, the shortening of the above mentioned distance connected to separate ligands, creating a different chemical structure. Based on the validation results we made the following observations:

- Although CREST force field method GFN-FF produces 100% structurally valid structures by our validation protocol, a visual inspection of the geometries reveals remarkable distortions in the central part of the molecule.
- It appears as if def2-SV(P) or a larger basis set repairs all validation errors. STO-3G validation results even degrade during DFT optimization. Possible implications: probably a basis set larger than STO-3G is more suitable in our context.
- We either need a more strict structural validation protocol or we can't always trust a human expert's conclusions based on visual inspection of the geometry.

5.7 Duplicates and near-similar geometries

Hypothesis is that different optimization runs could converge into the same final geometry even if they started from different input geometries. To perform that check we need to define a metric and threshold by which we could determine whether two molecular geometries are duplicate or nearly similar. Commonly used metrics here are Mean Absolute Distance (MAD), Root Mean Squared Distance (RMSD), Maximum distance amongst many possibilities. There is no one answer which metric is the right one and we could argue in favour of each one of them. It is recommended to monitor various metrics and not rely on a single one. Our choice was RMSD and Max Distance in addition to visually inspect the structures as we don't have that many. After the metric has been chosen, there is a question about the threshold value. Obviously, there is no single answer to that. It must be noted that a typical quantum chemistry software has some pre-set defaults in its geometry optimization routines, defining when to stop, e.g., Gaussian [31] has RMSD and Max Distance with the threshold values 0.0012A and 0.0018A correspondingly, meaning essentially identical geometries (this the point of the stopping criteria for an optimizer). Further analysis of near similars is needed to decide whether the geometries in question are unique conformers or not. Probably this needs to be done by an expert through manual inspection of the geometries. Further details about the distances between near-similar geometries are presented in Appendix 2 in the sub section Comments to Table A2.1. Figures A2.1 and A2.2 present distance matrices with RMSD values between a selection of conformers. Figure A2.1 shows two distance matrices, presenting distances between conformers in conformer group 6 kJ/mol (A), where RMSD values are in range 0.95-2.18 Å and in conformer group 10 kJ/mol (B) where RMSD values are in range 0.82-2.61 Å. These results suggest that both groups contain unique conformers, but for a definitive conclusion further analysis is needed. As a contrast to intra-group distances, Figure A.2.2 presents distances between representatives from different conformer groups. It shows RMSD values between representatives from the first ten conformer groups, 0-13 kJ/mol, with values in range 1.15-2.56 Å. RMSD, MAD and Max Distance metrics provide a quick indication of how similar geometries are and in case of duplicates we can detect these immediately. However, in case of very similar geometries that are not identical, as most confromers in the same group are, a further analysis is needed to determine their status. CREST workflow includes thorough sorting and filtering of geometries making sure that resulting conformers are unique. Based on this fact and also considering that the Potential Energy Surface is infinite in all practical purposes, we shouldn’t collide on duplicates very often.

As a technical sidenote: before calculating similarity metric between two molecular geometries, the geometries must be aligned: a general problem that in applied mathematics is called Wahba's problem [52], seeks to find a rotation matrix between two coordinate systems. The general idea is to calculate the optimal rotation matrix that minimizes the RMSD (root mean squared deviation) between two paired sets of points (see Kabsch algorithm [53]). Algorithmic solution to the alignment problem is an independent research topic. In this work we used the implementation in the ASE software, which has implemented an algorithm for finding the optimal rotation matrix using quaternions (algorithm itself based on [37]).

5.8 Descriptive analysis of geometries

As already discussed, GFN family methods tend to compress the Ti-O-C ligands by making the angles narrower compared to DFT calculations. Table 13 shows the statistics of Ti-O-C angles before and after full optimization at PBE0/def2-SV(P) level. It can be seen that certain Ti-O-C angles get straightened by the DFT method by as much as 12.4-18.9 degrees. Where Ti(7)-O(20)-C(18), Ti(7)-O(31)-C(29), Ti(61)-O(56)-C(48), Ti(61)-O(68)-C(72) represent pure ligands not directly constrained by any other covalent bonding. Additional details about the distribution angles are presented in Appendix 3 Figures A3.4 and A3.5.

Table 13. Ti-O-C angles in ligands connected to the central core of the TargetMol. "At start" corresponds to the starting geometry and "Converged" corresponds to the fully optimized geometry at the level of PBE0/def2-SV(P), all values in degrees.

	At start		**Converged**		**Difference**
Angle	**mean**	**stdev**	**mean**	**stdev**	**in mean**
Ti(7)-O(6)-C(4)	107.5	1.1	114.2	2.4	6.7
Ti(7)-O(20)-C(18)	139.7	11.5	154.4	5.3	14.7
Ti(7)-O(31)-C(29)	127.5	5.5	146.4	4.3	18.9
Ti(7)-O(60)-C(53)	120.1	10.6	126.9	1.7	6.7
Ti(61)-O(56)-C(48)	134.4	8.9	150.4	5.1	16.0
Ti(61)-O(59)-C(54)	118.9	8.1	121.4	0.6	2.5
Ti(61)-O(60)-C(53)	92.6	4.1	108.7	2.3	16.1
Ti(61)-O(68)-C(72)	130.6	6.9	143.0	5.7	12.4
Ti(7)-O(8)-C(10)	111.4	2.9	115.0	0.8	3.6

In addition to ligands connected to central titanium atoms, the analysis of the characteristic Ti_2O_2 core of the TargetMol reveals that at sufficiently accurate DFT level, the central core always forms a nearly perfect symmetric planar rhombus, as was described in the original publication [7]. Planarity of the rhombus cannot be deduced from the results in Table 14, but can be confirmed by visual inspection of the geometries. However, near zero variability (stdev) among converged geometries supports the claim of nearly perfect symmetry of the rhombus. Additional details about the distribution angles are presented in Appendix 3 Figures A3.4 and A3.5.

Table 14. Ti-O-Ti angles in the central Ti_2O_2 core of the TargetMol. "At start" corresponds to the starting geometry and "Converged" corresponds to the fully optimized geometry at the level of PBE0/def2-SV(P), all values in degrees.

	At start		**Converged**		**Difference**
Angle	**mean**	**stdev**	**mean**	**stdev**	**in mean**
Ti(7)-O(8)-Ti(61)	109.9	7.8	108.3	0.7	-1.5
Ti(7)-O(60)-Ti(61)	97.0	5.9	108.9	0.5	11.9

6 Summary

In this work we found a set of conformers for a titanium tartrate complex TargetMol with the use of an open-source software CREST, achieving the main goal of the work. The found conformers present a significant step forward in the analysis of TargetMol. We also found answers to most of the research questions that we posed in chapter 3:

- CREST is an excellent tool for generating conformers, answering our research goals RG1 and RG2.
- Answering our research questions RQ1 and RQ2, we found that GFN2-xTB and GFN-FF methods are applicable in our context, however GFN-optimized geometries cannot be always considered directly as the assessment of the final conformer ensemble. All geometries should be optimized at the level of DFT by at least 25 steps to get an adequate estimate of their relative conformational energies. In the case where conformers were generated with the GFN-FF method, 25 steps seemed insufficient to make adequate decisions.
- The GFN-FF method is a solid alternative to use in the CREST workflow. In the case of a larger molecule or in case of a more constrained computational resources, GFN-FF method becomes the first suggestion, directly answering our research question RQ3.
- Answering research question RQ4, we found that results about the search space coverage by CREST remain open. We performed preliminary analysis on the conformers found by CREST and reported distribution of selected angles and dihedrals, giving an indication of the search space coverage among found conformers. Further work is needed to assess the coverage of search space by the conformational sampling of CREST as the current thesis did not investigate that topic, apart from collecting all necessary input data and building appropriate software (Molli) to do this kind of analysis.
- Research goal RG3 was also achieved. By answering above-mentioned points we gained a lot of insight into how to orchestrate a similar workflow more efficiently and published open-source software Molli to support similar research.

References

[1] Spicher, S., Grimme, S. Robust atomistic modeling of materials, organometallic and biochemical systems Angew. Chem. Int. Ed. 2020, accepted article, DOI: 10.1002/anie.202004239.

[2] Minenkov Y, Sharapa DI, Cavallo L (2018) Application of Semiempirical Methods to Transition Metal Complexes: Fast Results but Hard-to-Predict Accuracy. Journal of Chemical Theory and Computation. Available: http://dx.doi.org/10.1021/acs.jctc.8b00018.

[3] Markus Bursch, Andreas Hansen, Philipp Pracht, Julia T. Kohn, Stefan Grimme, Theoretical study on conformational energies of transition metal complexes, DOI: 10.1039/d0cp04696e.

[4] Markus Bursch, Hagen Neugebauer, and Stefan Grimme, Structure Optimisation of Large Transition-Metal Complexes with Extended Tight-Binding Methods, DOI: 10.1002/anie.201904021.

[5] V. Sinha, J. J. Laan and E. A. Pidko, Accurate and rapid prediction of pKa of transition metal complexes: semiempirical quantum chemistry with a data-augmented approach, Phys. Chem. Chem. Phys., 2021, 23, 2557 DOI: 10.1039/D0CP05281G.

[6] Yury Minenkov, Edrisse Chermak, and Luigi Cavallo, Troubles in the Systematic Prediction of Transition Metal Thermochemistry with Contemporary Out-of-the-Box Methods, Journal of Chemical Theory and Computation 2016 12 (4), 1542-1560, DOI: 10.1021/acs.jctc.5b01163

[7] Ian D. Williams, Steven F. Pedersen, K. Barry Sharpless, Stephen J. Lippard, Crystal Structures of Two Titanium Tartrate Asymmetric Epoxidation Catalysts, J. Am. Chem. Soc. 1984, 106, 6430-6431.

[8] Philipp Pracht, Fabian Bohle, Stefan Grimme, Automated exploration of the low-energy chemical space with fast quantum chemical methods, doi.org/10.1039/C9CP06869D.

[9] Errol G. Lewars, Computational Chemistry Introduction to the Theory and Applications of Molecular and Quantum Mechanics, Third Edition 2016, DOI 10.1007/978-3-319-30916-3, ISBN 978-3-319-30914-9.

[10] Levine, Ira N., Quantum chemistry / Ira N. Levine Seventh edition 2014. ISBN-13: 978-0-321-80345-0.

[11]Cramer, Christopher J., Essentials of computational chemistry : theories and models 2nd ed 2004, ISBN 0-470-09181-9, ISBN 0-470-09182-7.

[12] Jensen, Frank, Introduction to computational chemistry, Third edition. Chichester, UK; Hoboken, NJ : John Wiley & Sons, 2017, ISBN 9781118825983.

[13] P. Hohenberg and W. Kohn, “Inhomogeneous Electron Gas,” Phys. Rev., 136 (1964) B864-B71. DOI: 10.1103/PhysRev.136.B864.

[14] Sebastian Ehlert, Uwe Huniar, Jinliang Ning, James W. Furness, Jianwei Sun, Aaron D. Kaplan, John P. Perdew, Jan Gerit Brandenburg “r2SCAN-D4: Dispersion corrected meta-generalized gradient approximation for general chemical applications”, J. Chem. Phys. 154, 061101 (2021); doi: 10.1063/5.0041008.

[15] R. G. Parr and W. Yang, Density-Functional Theory of Atoms and Molecules, Oxford University Press, Oxford, 1989.

[16] W. Kohn, “Electronic structure of matter - Wave functions and density functionals,” Rev. Mod. Phys. 71, 1253–1266 (1998).

[17] K. Burke, “Perspective on density functional theory,” J. Chem. Phys. 136, 150901 (2012).

[18] A. D. Becke, “Perspective: Fifty years of density-functional theory in chemical physics,” J. Chem. Phys. 140, 18A301 (2014).

[19] R. J. Maurer, C. Freysoldt, A. M. Reilly, J. G. Brandenburg, O. T. Hofmann, T. Bjöorkman, S. Lebègue, and A. Tkatchenko, “Advances in density-functional calculations for materials modeling,” Annu. Rev. Mater. Res. 49, 1–30 (2019).

[20] John P. Perdew, Adrienn Ruzsinszky, Jianmin Tao, et al, Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits, J. Chem. Phys. 123, 062201 (2005); https://doi.org/10.1063/1.1904565.

[21] Stefan Grimme, Andreas Hansen, Sebastian Ehlert, et al, “r2SCAN-3c: A “Swiss army knife” composite electronic-structure method, https://doi.org/10.1063/5.0040021.

[22] Norbert Schuch, Frank Verstraete, 2010, Computational Complexity of interacting electrons and fundamental limitations of Density Functional Theory.

[23] C. Adamo and V. Barone, “Toward reliable density functional methods without adjustable parameters: The PBE0 model,” J. Chem. Phys., 110 (1999) 6158-69. DOI: 10.1063/1.478522.

[24] Ernzerhof, M.; Scuseria, G. E., “Assessment of the Perdew-Burke-Ernzerhof exchange-correlation functional,” The Journal of ChemicalPhysics, 1999, 110, 5029-36, DOI: 10.1063/1.478401.

[25] F. Weigend and R. Ahlrichs, “Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy,” Phys. Chem. Chem. Phys., 7 (2005) 3297-305. DOI: 10.1039/B508541A.

[26] Bursch M., Mewes J.M., Hansen A., Grimme S., Best-Practice DFT Protocols for Basic Molecular Computational Chemistry, First published: 14 September 2022 https://doi.org/10.1002/anie.202205735.

[27] Michael Bühl and Hendrik Kabrede, Geometries of Transition-Metal Complexes from Density-Functional Theory, J. Chem. Theory Comput. 2006, 2, 1282-1290.

[28] Mark P. Waller, Heiko Braun, Nils Hojdis, Michael Bühl, Geometries of Second-Row Transition-Metal Complexes, from Density-Functional Theory J. Chem. Theory Comput. 2007, 3, 2234-2242.

[29] Michael Bühl, Christoph Reimann, Dimitrios A. Pantazis, Thomas Bredow, Frank Neese, Geometries of Third-Row Transition-Metal Complexes from Density-Functional Theory, J. Chem. Theory Comput. 2008, 4, 1449–1459.

[30] Stefan Grimme, Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations, J. Chem. Theory Comput. 2019, 15, 2847−2862.

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

Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2016.

[32] Kenneth Atz, Clemens Isert, Markus N. A. Böcker, José Jiménez-Luna, Gisbert Schneider, Δ-Quantum machine-learning for medicinal chemistry, DOI: 10.1039/D2CP00834C.

[33] Avogadro: an open-source molecular builder and visualization tool. Version 1.2. http://avogadro.cc/

[34] O'Boyle, N.M., Banck, M., James, C.A. et al. Open Babel: An open chemical toolbox. J Cheminform 3, 33 (2011). https://doi.org/10.1186/1758-2946-3-33.

[35] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E. Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N. Groves, Bjørk Hammer, Cory Hargus, Eric D. Hermes, Paul C. Jennings, Peter Bjerre Jensen, James Kermode, John R. Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S. Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, Karsten Wedel Jacobsen The Atomic Simulation Environment - A Python library for working with atoms J. Phys.: Condens. Matter Vol. 29 273002, 2017.

[36] Covalent radii revisited, Beatriz Cordero, Verónica Gómez, Ana E. Platero-Prats, Marc Revés, Jorge Echeverría, Eduard Cremades, Flavia Barragán and Santiago Alvarez, Dalton Trans., 2008, 2832-2838 doi: 10.1039/B801115J.

[37] Melander et al., Removing External Degrees of Freedom from Transition State Search Methods using Quaternions, J. Chem. Theory Comput., 2015, 11, 1055.

[38] J. Grant Hill, Gaussian Basis Sets for Molecular Applications, DOI: 10.1002/qua.24355.

[39] Markus Bursch, Hagen Neugebauer, Sebastian Ehlert, et al. Dispersion corrected r2SCAN based global hybrid functionals: r2SCANh, r2SCAN0, and r2SCAN50 J. Chem. Phys. 156, 134105 (2022); https://doi.org/10.1063/5.0086040.

[40] https://github.com/urmaspitsi/molli, accessed on 29.dec.2022.

[41] https://psicode.org, accessed on 20.dec.2022.

[42] https://taltech.ee/en/itcollege/hpc-centre, accessed on 1.jan.2023.

[43] https://ai-lab.pages.taltech.ee, accessed on 1.jan.2023.

[44] https://gaussian.com/opt/, accessed on 20.dec.2022.

[45] https://gaussian.com/faq3/, accessed on 22.dec.2022.

[46] https://crest-lab.github.io/crest-docs/page/documentation/keywords.html#noreftopo, accessed on 29.dec.2022.

[47] https://github.com/grimme-lab, accessed on 29.dec.2022.

[48] https://xtb-docs.readthedocs.io/en/latest/CENSO_docs/censo.html, accessed on 29.dec.2022.

[49] https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system, accessed on 20.dec.2022.

[50] https://pytorch-geometric.readthedocs.io/en/latest/, accessed on 22.dec.2022.

[51] https://www.rdkit.org, accessed on 22.dec.2022.

[52] https://en.wikipedia.org/wiki/Wahba%27s_problem, accessed on 20.dec.2022.

[53] https://en.wikipedia.org/wiki/Kabsch_algorithm, accessed on 20.dec.2022.

[54] https://www.kaggle.com, accessed on 29.dec.2022.

[55] Axelrod, S. and Gomez-Bombarelli, R., 2020. GEOM: Energy-annotated molecular conformations for property prediction and molecular generation. arXiv preprint arXiv:2006.05531. arXiv: 2006.05531.

[56] Balcells D, Skjelstad BB. The tmQM Dataset - Quantum Geometries and Properties of 86k Transition Metal Complexes. ChemRxiv. Cambridge: Cambridge Open Engage; 2020; This content is a preprint and has not been peer-reviewed.

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I Urmas Pitsi

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis “Conformational Analysis of an Organometallic Compound with Data Science Inspired Workflow”, supervised by prof. Toomas Tamm and Juhan-Peep Ernits Phd.
   - 1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   - 1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

[1] The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 - Conformers

Table A2.1. Conformers fully optimized with PBE0/def2-SV(P)

Index	Conformer	Energy difference to best, kJ/mol	Final Energy, hartree	energy delta kJ/mol
1	ex21_c12	0.00	-3839.38363489	-581
2	ex21_c43	1.42	-3839.38309480	-565
3	ex21_c4	1.93	-3839.38289836	-571
4	ex21_c2	5.52	-3839.38153264	-575
5	ex21_c45	5.65	-3839.38148335	-575
6	ex0ff_c4	5.97	-3839.38136096	-647
7	ex21_c59	6.32	-3839.38122896	-583
8	ex21_c44	6.32	-3839.38122888	-546
9	ex21_c57	8.18	-3839.38051917	-543
10	ex21_c58	8.30	-3839.38047335	-512
11	ex19_c61	9.27	-3839.38010312	-236
12	ex0ff_c3	9.99	-3839.37982882	-704
13	ex23_c25	10.42	-3839.37966586	-242
14	ex23_c26	10.42	-3839.37966583	-239
15	ex19_c44	10.42	-3839.37966567	-232
16	ex19_c40	10.42	-3839.37966563	-236
17	ex21_c41	10.94	-3839.37946958	-515
18	ex21_c16	10.94	-3839.37946951	-544
19	ex21_c14	10.94	-3839.37946909	-547
20	ex21_c32	11.04	-3839.37942992	-512
21	ex0a_c3	11.24	-3839.37935309	-265
22	ex0a_c10	11.29	-3839.37933581	-278
23	ex0a_c21	11.29	-3839.37933563	-283
24	ex0a_c13	11.29	-3839.37933562	-277
25	ex0a_c16	11.29	-3839.37933551	-278
26	ex23_c30	11.37	-3839.37930614	-233
27	ex19_c52	11.37	-3839.37930613	-239
28	ex19_c45	11.37	-3839.37930612	-241
29	ex19_c48	11.37	-3839.37930612	-234
30	ex23_c29	11.37	-3839.37930609	-250
31	ex19_c49	11.37	-3839.37930608	-242

Index	Conformer	Energy difference to best, kJ/mol	Final Energy, hartree	energy delta kJ/mol
32	ex19_c47	11.37	-3839.37930606	-232
33	ex23_c31	11.37	-3839.37930605	-235
34	ex19_c51	11.37	-3839.37930605	-239
35	ex19_c46	11.37	-3839.37930605	-246
36	ex23_c27	11.37	-3839.37930604	-238
37	ex19_c50	11.37	-3839.37930602	-236
38	ex21_c56	11.58	-3839.37922351	-578
39	ex19_c56	11.74	-3839.37916500	-240
40	ex19_c57	11.74	-3839.37916477	-248
41	ex21_c13	12.41	-3839.37890643	-518
42	ex23_c13	12.48	-3839.37888102	-243
43	ex19_c36	12.48	-3839.37888096	-242
44	ex19_c34	12.48	-3839.37888087	-238
45	ex23_c20	12.48	-3839.37888082	-234
46	ex19_c42	12.48	-3839.37888078	-239
47	ex19_c23	12.48	-3839.37888068	-244
48	ex19_c16	12.48	-3839.37888056	-237
49	ex23_c22	12.48	-3839.37888055	-244
50	ex19_c39	12.48	-3839.37888051	-223
51	ex23_c15	12.48	-3839.37888050	-244
52	ex23_c16	12.48	-3839.37888046	-236
53	ex19_c29	12.48	-3839.37888043	-236
54	ex19_c26	12.48	-3839.37888040	-239
55	ex19_c24	12.48	-3839.37888034	-234
56	ex19_c19	12.48	-3839.37888026	-241
57	ex19_c22	12.48	-3839.37888006	-249
58	ex19_c17	12.48	-3839.37888000	-238
59	ex19_c38	12.48	-3839.37887997	-229
60	ex19_c27	12.48	-3839.37887996	-231
61	ex19_c18	12.48	-3839.37887984	-229
62	ex23_c17	12.48	-3839.37887978	-232
63	ex23_c23	12.48	-3839.37887961	-243
64	ex23_c21	12.49	-3839.37887952	-230
65	ex19_c20	12.49	-3839.37887944	-239

Index	Conformer	Energy difference to best, kJ/mol	Final Energy, hartree	energy delta kJ/mol
66	ex23_c14	12.49	-3839.37887932	-244
67	ex23_c19	12.49	-3839.37887916	-234
68	ex19_c21	12.49	-3839.37887916	-240
69	ex19_c33	12.49	-3839.37887859	-225
70	ex21_c26	12.77	-3839.37877211	-524
71	ex23_c33	13.14	-3839.37862918	-233
72	ex0a_c6	13.14	-3839.37862856	-275
73	ex0a_c20	13.16	-3839.37862078	-251
74	ex19_c60	13.18	-3839.37861372	-233
75	ex0a_c4	13.18	-3839.37861361	-254
76	ex0a_c11	13.27	-3839.37858101	-244
77	ex0a_c9	13.27	-3839.37858082	-250
78	ex19_c31	13.27	-3839.37857997	-231
79	ex0a_c5	13.58	-3839.37846437	-279
80	ex0a_c28	13.70	-3839.37841834	-269
81	ex21_c24	13.76	-3839.37839316	-519
82	ex21_c22	13.87	-3839.37835116	-544
83	ex21_c30	13.94	-3839.37832454	-492
84	ex0a_c7	14.10	-3839.37826483	-255
85	ex0a_c15	14.10	-3839.37826465	-273
86	ex21_c50	14.16	-3839.37824152	-498
87	ex23_c18	14.23	-3839.37821448	-234
88	ex19_c41	14.23	-3839.37821445	-237
89	ex19_c37	14.23	-3839.37821440	-243
90	ex19_c32	14.23	-3839.37821439	-235
91	ex19_c28	14.23	-3839.37821404	-233
92	ex19_c25	14.23	-3839.37821358	-225
93	ex19_c35	14.23	-3839.37821352	-231
94	ex23_c24	14.23	-3839.37821352	-230
95	ex19_c43	14.23	-3839.37821330	-233
96	ex19_c30	14.23	-3839.37821329	-229
97	ex16_c6	14.64	-3839.37805738	-243
98	ex0a_c12	14.64	-3839.37805713	-287
99	ex16_c7	14.64	-3839.37805694	-249

Index	Conformer	Energy difference to best, kJ/mol	Final Energy, hartree	energy delta kJ/mol
100	ex0a_c14	14.65	-3839.37805678	-255
101	ex0a_c8	15.05	-3839.37790317	-272
102	ex23_c32	15.05	-3839.37790315	-238
103	ex19_c10	15.41	-3839.37776701	-236
104	ex16_c4	15.41	-3839.37776696	-247
105	ex19_c55	15.41	-3839.37776693	-238
106	ex19_c7	15.41	-3839.37776691	-225
107	ex23_c2	15.41	-3839.37776689	-231
108	ex23_c6	15.41	-3839.37776688	-225
109	ex16_c1	15.41	-3839.37776687	-236
110	ex19_c11	15.41	-3839.37776687	-232
111	ex19_c5	15.41	-3839.37776686	-230
112	ex0a_c2	15.41	-3839.37776685	-250
113	ex19_c1	15.41	-3839.37776684	-230
114	ex23_c1	15.41	-3839.37776684	-235
115	ex23_c9	15.41	-3839.37776682	-231
116	ex19_c15	15.41	-3839.37776680	-228
117	ex16_c3	15.41	-3839.37776679	-240
118	ex19_c8	15.41	-3839.37776678	-226
119	ex23_c10	15.41	-3839.37776678	-224
120	ex23_c8	15.41	-3839.37776677	-230
121	ex23_c5	15.41	-3839.37776676	-238
122	ex23_c12	15.41	-3839.37776675	-221
123	ex16_c2	15.41	-3839.37776672	-240
124	ex23_c7	15.41	-3839.37776668	-232
125	ex19_c14	15.41	-3839.37776667	-225
126	ex19_c2	15.41	-3839.37776667	-231
127	ex19_c12	15.41	-3839.37776666	-230
128	ex19_c4	15.41	-3839.37776664	-228
129	ex19_c6	15.41	-3839.37776662	-226
130	ex19_c13	15.41	-3839.37776659	-224
131	ex19_c3	15.41	-3839.37776657	-227
132	ex23_c4	15.41	-3839.37776654	-232
133	ex19_c9	15.41	-3839.37776653	-225

Index	Conformer	Energy difference to best, kJ/mol	Final Energy, hartree	energy delta kJ/mol
134	ex23_c11	15.41	-3839.37776648	-233
135	ex23_c3	15.41	-3839.37776619	-237
136	ex0ff_c9	15.43	-3839.37775958	-660
137	ex0a_c24	15.52	-3839.37772520	-263
138	ex21_c33	15.54	-3839.37771639	-511
139	ex19_c54	15.76	-3839.37763127	-228
140	ex19_c53	15.76	-3839.37763121	-236
141	ex16_c5	15.76	-3839.37763108	-236
142	ex21_c11	16.08	-3839.37750989	-547
143	ex21_c38	16.34	-3839.37741247	-539
144	ex0a_c17	16.51	-3839.37734674	-236
145	ex0a_c18	16.51	-3839.37734607	-262
146	ex0ff_c6	17.05	-3839.37713962	-622
147	ex21_c46	17.12	-3839.37711321	-517
148	ex21_c47	17.23	-3839.37707160	-544
149	ex0a_c23	17.33	-3839.37703238	-262
150	ex21_c25	17.34	-3839.37702963	-536
151	ex0a_c26	17.53	-3839.37695637	-271
152	ex0a_c25	17.54	-3839.37695584	-274
153	ex21_c3	18.23	-3839.37668969	-560
154	ex0a_c19	18.29	-3839.37666936	-267
155	ex21_c20	18.80	-3839.37647384	-512
156	ex21_c40	19.32	-3839.37627788	-505
157	ex21_c18	19.40	-3839.37624543	-586
158	ex0a_c1	19.68	-3839.37613830	-252
159	ex21_c29	21.63	-3839.37539738	-544
160	ex15_c10	21.82	-3839.37532324	-665
161	ex15_c2	22.87	-3839.37492341	-673
162	ex0ff_c8	23.02	-3839.37486569	-598
163	ex0a_c29	23.99	-3839.37449816	-274
164	ex21_c48	24.02	-3839.37448548	-502
165	ex0ff_c1	24.22	-3839.37440996	-638
166	ex0a_c22	25.46	-3839.37393777	-276
167	ex0ff_c7	26.56	-3839.37351892	-598

Index	Conformer	Energy difference to best, kJ/mol	Final Energy, hartree	energy delta kJ/mol
168	ex15_c4	26.82	-3839.37342033	-656
169	ex0b_c22	35.64	-3839.37006213	-280
170	ex0b_c30	35.93	-3839.36994918	-266
171	ex0b_c34	35.97	-3839.36993275	-266
172	ex0b_c31	35.98	-3839.36992927	-263
173	ex0b_c29	36.01	-3839.36992020	-235
174	ex0b_c16	36.32	-3839.36980132	-292
175	ex0b_c21	37.50	-3839.36935231	-242
176	ex0b_c15	37.68	-3839.36928523	-237
177	ex0b_c2	37.86	-3839.36921374	-286
178	ex0b_c1	37.86	-3839.36921361	-288
179	ex0b_c20	37.90	-3839.36919950	-291
180	ex0b_c7	38.54	-3839.36895472	-286
181	ex0b_c28	38.94	-3839.36880208	-262
182	ex0b_c25	39.54	-3839.36857347	-272
183	ex0b_c27	39.78	-3839.36848275	-269
184	ex0b_c26	39.78	-3839.36848262	-266
185	ex0b_c9	40.70	-3839.36813285	-264
186	ex0b_c4	40.90	-3839.36805855	-257
187	ex0b_c3	40.90	-3839.36805845	-281
188	ex0b_c19	40.91	-3839.36805277	-276
189	ex0b_c11	41.13	-3839.36797092	-264
190	ex0b_c13	41.13	-3839.36797060	-270
191	ex0b_c10	41.13	-3839.36797053	-269
192	ex0b_c14	41.13	-3839.36797048	-272
193	ex0b_c12	41.13	-3839.36797038	-278
194	ex0b_c36	41.36	-3839.36788069	-282
195	ex0b_c6	41.54	-3839.36781168	-288
196	ex0b_c18	42.03	-3839.36762787	-272
197	ex0b_c17	42.03	-3839.36762780	-273
198	ex0b_c33	44.17	-3839.36681074	-284
199	ex0b_c35	58.43	-3839.36137937	-273
200	ex0b_c38	60.93	-3839.36042853	-225
201	ex0b_c39	61.93	-3839.36004788	-226

Index	Conformer	Energy difference to best, kJ/mol	Final Energy, hartree	energy delta kJ/mol
202	ex0b_c32	62.84	-3839.35970153	-227
203	ex0b_c23	66.65	-3839.35825023	-216
204	ex0b_c37	66.81	-3839.35818882	-228
205	original	108.15	-3839.34244253	-48

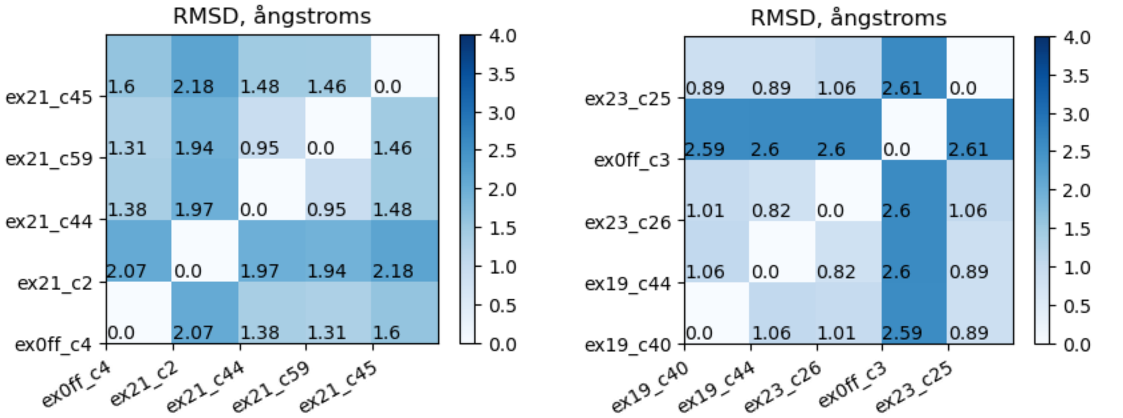
Table A2.2. Conformers groups, PBE0/def2-SV(P), grouped by rounded relative conformational energy, kJ/mol.

idx	First conformer in group	Threshold, kJ/mol	Number of conformers
1	ex21_c12	0	1
2	ex21_c43	1	1
3	ex21_c4	2	1
4	ex0ff_c4	6	5
5	ex21_c58	8	2
6	ex19_c61	9	1
7	ex19_c40	10	5
8	ex19_c50	11	21
9	ex19_c19	12	32
10	ex19_c31	13	9
11	ex19_c30	14	18
12	ex0a_c12	15	40
13	ex19_c54	16	7
14	ex21_c47	17	7
15	ex0a_c26	18	4
16	ex21_c18	19	3
17	ex0a_c1	20	1
18	ex21_c29	22	2
19	ex0ff_c8	23	2
20	ex0a_c29	24	3
21	ex0a_c22	25	1
22	ex0ff_c7	27	2
23	ex0b_c34	36	6
24	ex0b_c21	37	1
25	ex0b_c20	38	4

idx	First conformer in group	Threshold, kJ/mol	Number of conformers
26	ex0b_c28	39	2
27	ex0b_c25	40	3
28	ex0b_c13	41	10
29	ex0b_c18	42	3
30	ex0b_c33	44	1
31	ex0b_c35	58	1
32	ex0b_c38	61	1
33	ex0b_c39	62	1
34	ex0b_c32	63	1
35	ex0b_c23	67	2
36	original	108	1
	Total		**205**

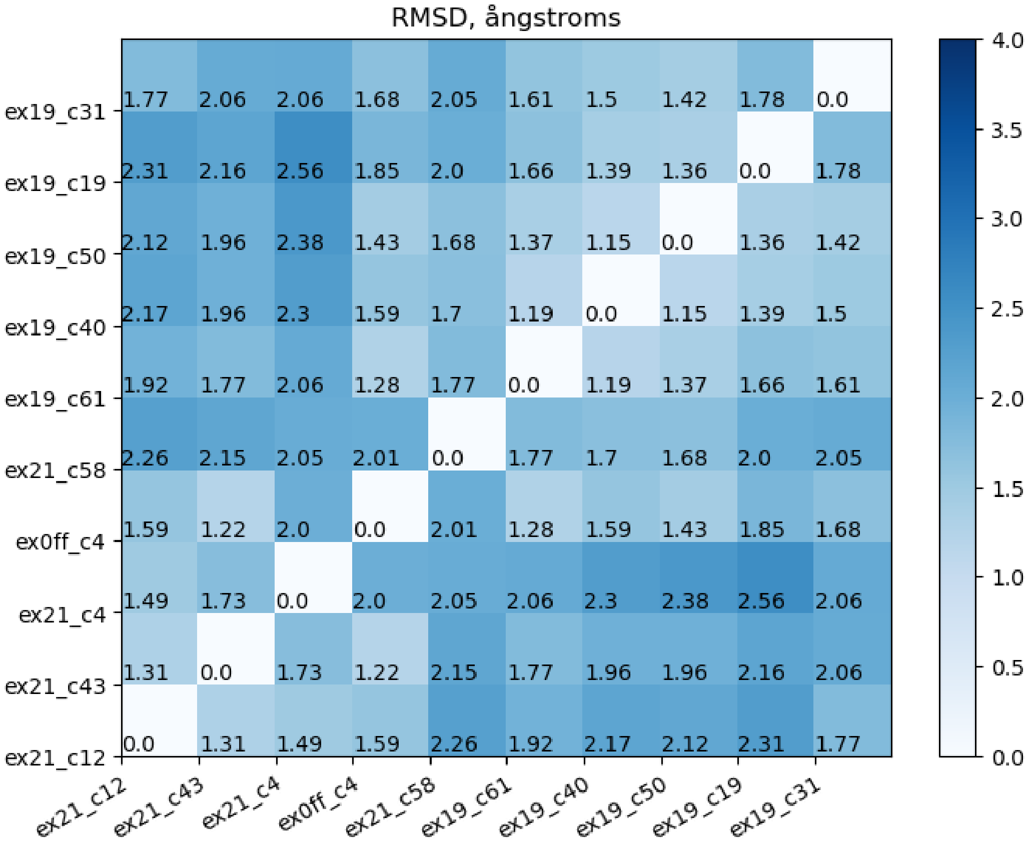
Comments to Table A2.1.

All conformers with very similar relative conformational energy must be analyzed further. RMSD and Max distance information is given below. All distances between geometries are given in ångstroms (Å).



(A) **(B)**

Figure A2.1. RMSD between conformers in group 6 kJ/mol (A) and group 10 kJ/mol (B).



	ex21_c12	ex21_c43	ex21_c4	ex0ff_c4	ex21_c58	ex19_c61	ex19_c40	ex19_c50	ex19_c19	ex19_c31
ex19_c31	1.77	2.06	2.06	1.68	2.05	1.61	1.5	1.42	1.78	0.0
ex19_c19	2.31	2.16	2.56	1.85	2.0	1.66	1.39	1.36	0.0	1.78
ex19_c50	2.12	1.96	2.38	1.43	1.68	1.37	1.15	0.0	1.36	1.42
ex19_c40	2.17	1.96	2.3	1.59	1.7	1.19	0.0	1.15	1.39	1.5
ex19_c61	1.92	1.77	2.06	1.28	1.77	0.0	1.19	1.37	1.66	1.61
ex21_c58	2.26	2.15	2.05	2.01	0.0	1.77	1.7	1.68	2.0	2.05
ex0ff_c4	1.59	1.22	2.0	0.0	2.01	1.28	1.59	1.43	1.85	1.68
ex21_c4	1.49	1.73	0.0	2.0	2.05	2.06	2.3	2.38	2.56	2.06
ex21_c43	1.31	0.0	1.73	1.22	2.15	1.77	1.96	1.96	2.16	2.06
ex21_c12	0.0	1.31	1.49	1.59	2.26	1.92	2.17	2.12	2.31	1.77

Figure A2.2. RMSD between the first conformer in each of the first 10 conformers groups, 0-13 kJ/mol.

Table A2.3. Conformers fully optimized with BP86/def2-SV(P)

Index	Conformer	Energy Difference to best, kJ/mol	Final Energy, hartree	Energy delta in optimization start to end, kJ/mol
1	ex0ff_c6	0.00	-3842.44270854	-650
2	ex0ff_c4	2.22	-3842.44186280	-663
3	ex0a_c27	8.02	-3842.43965316	-365
4	ex0a_c13	8.02	-3842.43965288	-363
5	ex0a_c10	8.02	-3842.43965259	-367
6	ex0a_c20	8.63	-3842.43942032	-335
7	ex0a_c9	9.54	-3842.43907391	-337
8	ex0a_c7	9.54	-3842.43907386	-342
9	ex16_c5	9.87	-3842.43894817	-327
10	ex0a_c1	10.17	-3842.43883519	-345
11	ex19_c23	10.17	-3842.43883497	-332
12	ex0a_c6	10.45	-3842.43872819	-365

Index	Conformer	Energy Difference to best, kJ/mol	Final Energy, hartree	Energy delta in optimization start to end, kJ/mol
13	ex0a_c26	11.49	-3842.43833132	-364
14	ex0a_c23	11.51	-3842.43832491	-355
15	ex15_c5	12.00	-3842.43813683	-693
16	ex0ff_c8	12.06	-3842.43811451	-619
17	ex16_c6	12.40	-3842.43798645	-330
18	ex0a_c24	12.67	-3842.43788436	-354
19	ex16_c2	13.13	-3842.43770884	-329
20	ex16_c1	13.14	-3842.43770252	-325
21	ex16_c4	13.15	-3842.43769824	-336
22	ex0ff_c1	14.02	-3842.43736721	-659
23	ex0ff_c7	16.12	-3842.43656932	-616
24	ex15_c10	16.84	-3842.43629344	-685
25	ex0a_c29	17.62	-3842.43599928	-362
26	ex15_c4	19.29	-3842.43535948	-678
27	ex15_c8	21.50	-3842.43452136	-704
28	ex0b_c34	29.21	-3842.43158332	-352
29	ex0b_c16	30.29	-3842.43117189	-380
30	ex0b_c21	32.64	-3842.43027744	-324
31	ex0b_c6	35.68	-3842.42911692	-377
32	ex0b_c19	35.85	-3842.42905524	-364
33	ex0b_c9	36.32	-3842.42887667	-345
34	original	97.30	-3842.40564860	-16

Table A2.4. Relative conformational energies at different stages of optimization, correlation vs "Converged". BP86/def2-SV(P), kJ/mol

Label	Converged	At start	10 steps	25 steps
ex0ff_c6	0.00	0.00	0.00	0.00
ex0ff_c4	2.22	14.81	31.67	-0.14
ex0a_c27	8.02	-277.15	-24.88	16.78
ex0a_c13	8.02	-279.07	-44.96	-5.56
ex0a_c10	8.02	-274.98	-49.10	-4.50
ex0a_c20	8.63	-306.37	-44.46	7.04
ex0a_c9	9.54	-303.30	-44.93	-2.98
ex0a_c7	9.54	-298.16	-40.71	-1.44
ex16_c5	9.87	-312.97	-48.72	-0.98
ex0a_c1	10.17	-294.49	-45.78	0.77
ex19_c23	10.17	-308.22	-48.39	-3.48
ex0a_c6	10.45	-274.96	-43.42	-2.33
ex0a_c26	11.49	-274.32	46.46	4.34
ex0a_c23	11.51	-283.39	-43.79	1.12
ex15_c5	12.00	55.00	56.02	26.59
ex0ff_c8	12.06	-18.67	-7.49	12.08
ex16_c6	12.40	-307.76	-47.20	-0.71
ex0a_c24	12.67	-283.27	-40.70	1.35
ex16_c2	13.13	-307.73	-46.76	0.10
ex16_c1	13.14	-311.45	-48.37	-0.06
ex16_c4	13.15	-300.83	-48.54	0.04
ex0ff_c1	14.02	23.26	59.62	2.35
ex0ff_c7	16.12	-18.11	-14.52	6.81
ex15_c10	16.84	51.64	73.34	41.75
ex0a_c29	17.62	-270.70	-33.41	4.35
ex15_c4	19.29	47.41	42.79	14.64
ex15_c8	21.50	75.33	25.94	25.47
ex0b_c34	29.21	-268.72	-11.71	21.52
ex0b_c16	30.29	-239.99	3.77	19.93
ex0b_c21	32.64	-293.53	-24.04	21.33
ex0b_c6	35.68	-237.61	-4.81	24.41
ex0b_c19	35.85	-250.26	30.12	26.27
ex0b_c9	36.32	-268.34	-16.20	27.11
original	97.30	-536.30	43.85	93.15
Correlation	**1.00**	**-0.36**	**0.33**	**0.89**

Table A2.5. Relative conformational energies at different stages of optimization, correlation vs "Converged". BP86/def2-SV(P). ground truth is fully optimized at PBE0/cc-pVTZ, kJ/mol

Label	Ground Truth	BP86/def2-SV(P) 10-steps	BP86/def2-SV(P) 25-steps	BP86/def2-SV(P) Converged
ex0ff_c6	0.00	0.00	0.00	0.00
ex16_c5	5.58	-48.72	-0.98	9.87
ex0a_c10	6.35	-49.10	-4.50	8.02
ex19_c23	8.78	-48.39	-3.48	10.17
ex0a_c24	9.09	-40.70	1.35	12.67
ex0a_c26	11.95	46.46	4.34	11.49
ex0b_c21	26.92	-24.04	21.33	32.64
original	112.77	43.85	93.15	97.30
Correlation	**1.0000**	**0.5849**	**0.9939**	**0.9953**

Appendix 3 - CREST Experiments

Table A3.1. Details of CREST experiments

Exp. Name	Opt. Method	Wall time	CREST Result: Energy, Number of conformers	CREST Command line input	Comment
ex0	GFN2	1h :35m	E lowest : -142.72799 1 structures remain within 5.00 kcal/mol window, unique 1	crest mol24_final.xyz --gfn2 -T 64 --opt normal --squick --nowr --noreftopo | tee stdout.txt	crest GFN2 run on original mol24_final.xyz to see how good/bad does it work out-of-the-box
ex0a	GFN2	14h :22m	E lowest : -142.74475 112 structures remain within 6.00 kcal/mol window, unique 29	crest mol24_final.xyz --gfn2 -T 64 --opt vtight --nowr --noreftopo | tee stdout.txt	crest GFN2 run on original mol24_final.xyz to see how good/bad does it work out-of-the-box vtight opt
ex0b	GFN2	7h :25m	E lowest : -142.73587 440 structures remain within 6.00 kcal/mol window, unique 39	crest mol24_final.xyz --gfn2 -T 64 --opt vtight --noreftopo --keepdir | tee stdout.txt	repeat ex0a with more verbose output, aim is to analyze intermediate results
ex0ff	GFN-FF	12m	E lowest : -13.47791 9 structures remain within 6.00 kcal/mol window, unique 9	crest mol24_final.xyz --gfnff -T 64 --opt vtight --noreftopo --keepdir | tee stdout.txt	crest GFN-FF run on original mol24_final.xyz to see how good/bad does it work out-of-the-box vtight opt, verbose full output
ex8	GFN2	12h :18m	E lowest : -142.74595 53 structures remain within 6.00 kcal/mol window, unique 53	crest final_constraints4.xyz --gfn2 -T 48 --opt vtight --nowr --cinp constraints4.txt	fix atoms: 2,4,7,8,45,60,61 (add O:45)
ex10	GFN-FF	30m	E lowest : -13.56466 4 structures remain within 6.00 kcal/mol window, unique 4	crest final_constraints4.xyz --gfnff -T 48 --opt vtight --nowr --cinp constraints4.txt	fix atoms: 2,4,7,8,45,60,61 (add O:45)
ex15	GFN-FF	14m	E lowest : -13.47818 158 structures remain within 6.00 kcal/mol window, unique 150	crest mol24_final_constraints6_angles.xyz --gfnff -T 64 --opt vtight --nowr --cinp mol24_constraints6_angles.txt --noreftopo | tee stdout.txt	Ti-O-C angles fixed as in original input: mol24_final.xyz, GFN-FF

Exp. Name	Opt. Method	Wall time	CREST Result: Energy, Number of conformers	CREST Command line input	Comment
ex16	GFN2	11h :41m	E lowest : -142.75207 7 structures remain within 6.00 kcal/mol window, unique 7	crest mol24_final_constraints6_angles.xyz --gfn2 -T 64 --opt vtight --nowr --cinp mol24_constraints6_angles.txt --noreftopo | tee stdout.txt	Ti-O-C angles fixed as in original input: mol24_final.xyz, GFN2
ex19	GFN2	2h :36m	E lowest : -142.75236 183 structures remain within 5.00 kcal/mol window, unique 61	crest mol24_ex16_gfn2_crest5_constr6.xyz --gfn2 -T 64 --opt normal --squick --nowr --mrest 3 --cinp mol24_constr6.txt --noreftopo | tee stdout.txt	start new CREST run from ex16 crest best 5
ex20	GFN2	2h :45m	E lowest : -142.73324 3 structures remain within 5.00 kcal/mol window, unique 3	crest mol24_final_constr7.xyz --gfn2 -T 64 --opt normal --squick --nowr --mrest 3 --cinp mol24_constr7.txt --noreftopo > stdout.txt	relax some angles from constr6
ex21	GFN-FF	2m	E lowest : -13.60953 60 structures remain within 5.00 kcal/mol window, unique 59	crest mol24_ex16_gfn2_crest5_constr6.xyz --gfnff -T 64 --opt normal --squick --nowr --mrest 3 --cinp mol24_constr6.txt --noreftopo > stdout.txt	repeat ex19 but with GFN-FF
ex21a	GFN-FF	14m	E lowest : -13.61704 150 structures remain within 6.00 kcal/mol window, unique 146	crest mol24_ex16_gfn2_crest5_constr6.xyz --gfnff -T 48 --opt vtight --nowr --cinp mol24_constr6.txt --noreftopo > stdout.txt	repeat ex21 but more thorough search
ex22	GFN2	1h :52m	E lowest : -142.74937 8 structures remain within 5.00 kcal/mol window, unique 1	crest mol24_ex20_gfn2_crest1_constr7.xyz --gfn2 -T 64 --opt normal --squick --nowr --mrest 3 --cinp mol24_constr7.txt --noreftopo > stdout.txt	repeat ex20 with its best 1
ex23	GFN2	1h :43m	E lowest : -142.75242 89 structures remain within 5.00 kcal/mol window, unique 33	crest mol24_ex16_gfn2_crest3_constr6.xyz --gfn2 -T 48 --opt normal --squick --nowr --mrest 3 --cinp mol24_constr6.txt --noreftopo | tee stdout.txt	repeat ex19 with source from ex16 crest best 3
ex24	GFN2	15h : 4m	E lowest : -142.75207 7 structures remain within 6.00 kcal/mol window, unique 4	crest mol24_final_constraints6_angles.xyz --gfn2 -T 64 --opt vtight --cinp mol24_constraints6_angles.txt --noreftopo --keepdir | tee stdout.txt	repeat ex16 with more verbose output, aim is to analyze intermediate results

Figure A3.2. Selection of Ti-O-C and Ti-O-Ti angles from conformers generated by CREST using GFN2-xTB method (experiments ex0a, ex0b).

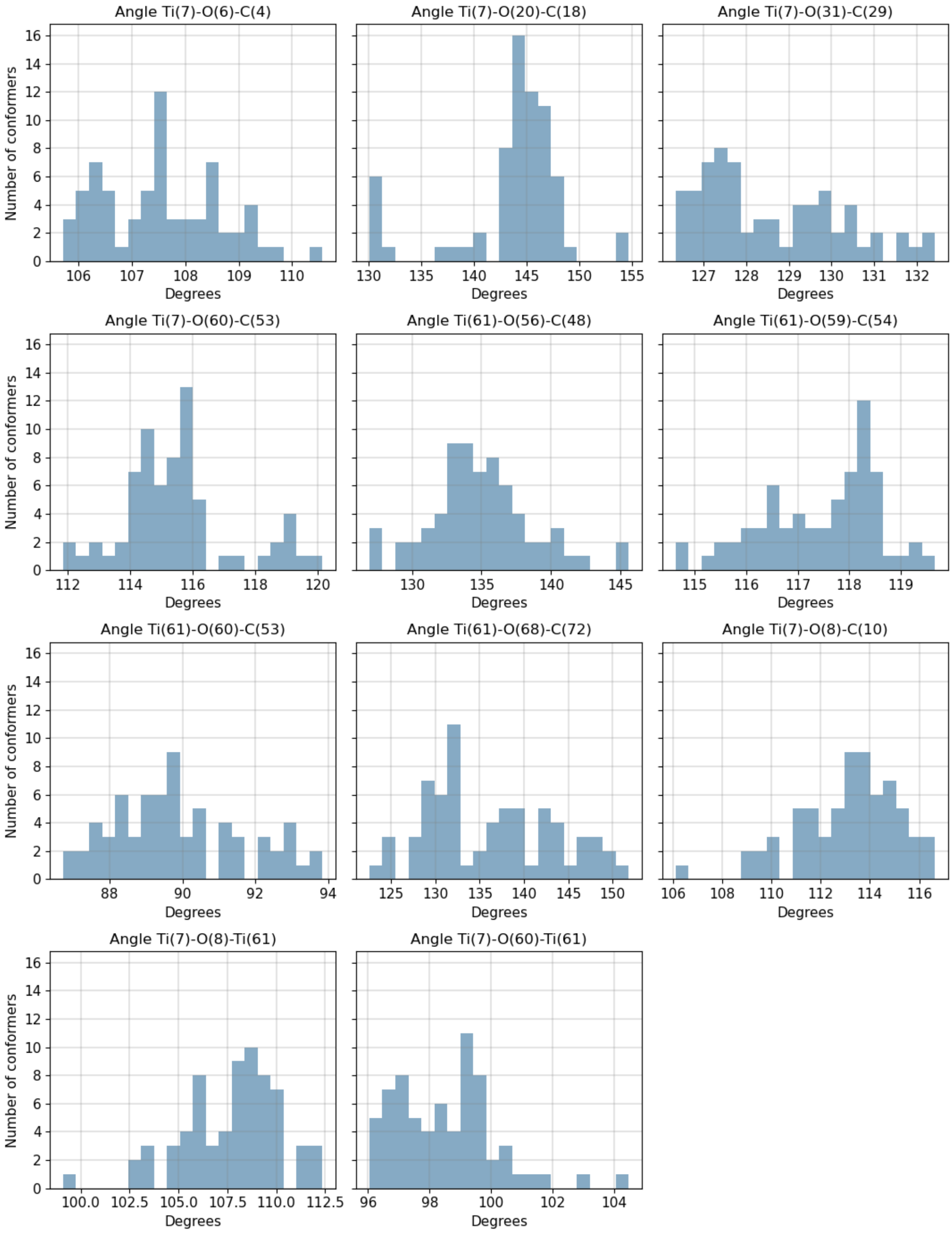
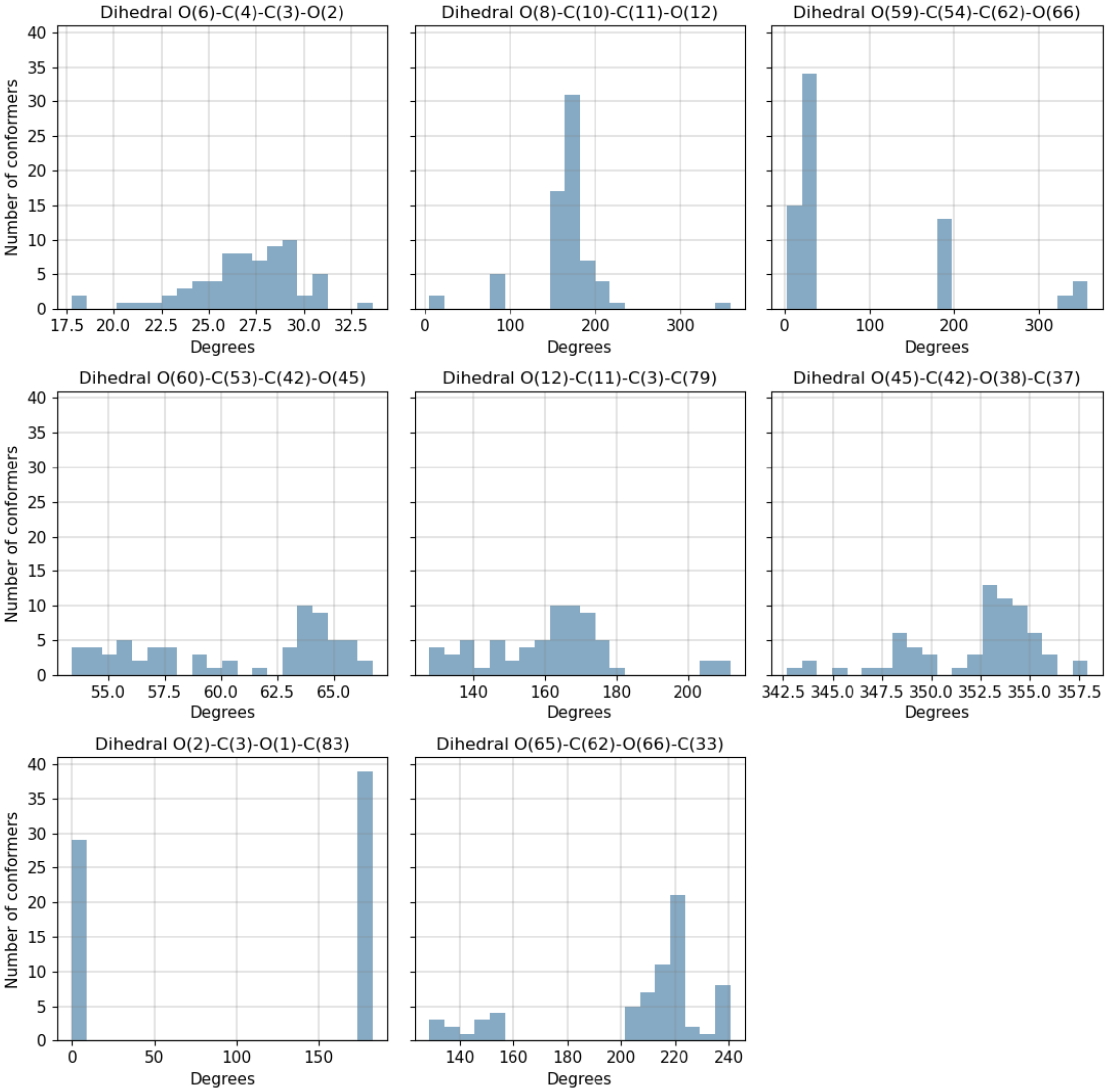


Figure A3.3. Selection of dihedral angles from conformers generated by CREST using GFN2-xTB mthod (experiments ex0a, ex0b).



Dihedral O(6)-C(4)-C(3)-O(2)

Dihedral O(8)-C(10)-C(11)-O(12)

Dihedral O(59)-C(54)-C(62)-O(66)

Dihedral O(60)-C(53)-C(42)-O(45)

Dihedral O(12)-C(11)-C(3)-C(79)

Dihedral O(45)-C(42)-O(38)-C(37)

Dihedral O(2)-C(3)-O(1)-C(83)

Dihedral O(65)-C(62)-O(66)-C(33)

Figure A3.4. Selection of Ti-O-C and Ti-O-Ti angles at the start of optimization with PBE0/def2-SV(P). Ti-O-C angles are significantly smaller compared to Figure A3.5. which shows the same angles, but at the end of the DFT optimization, showing how much GFN family optimization methods compress the Ti-O-C ligands compared to more accurate DFT methods.

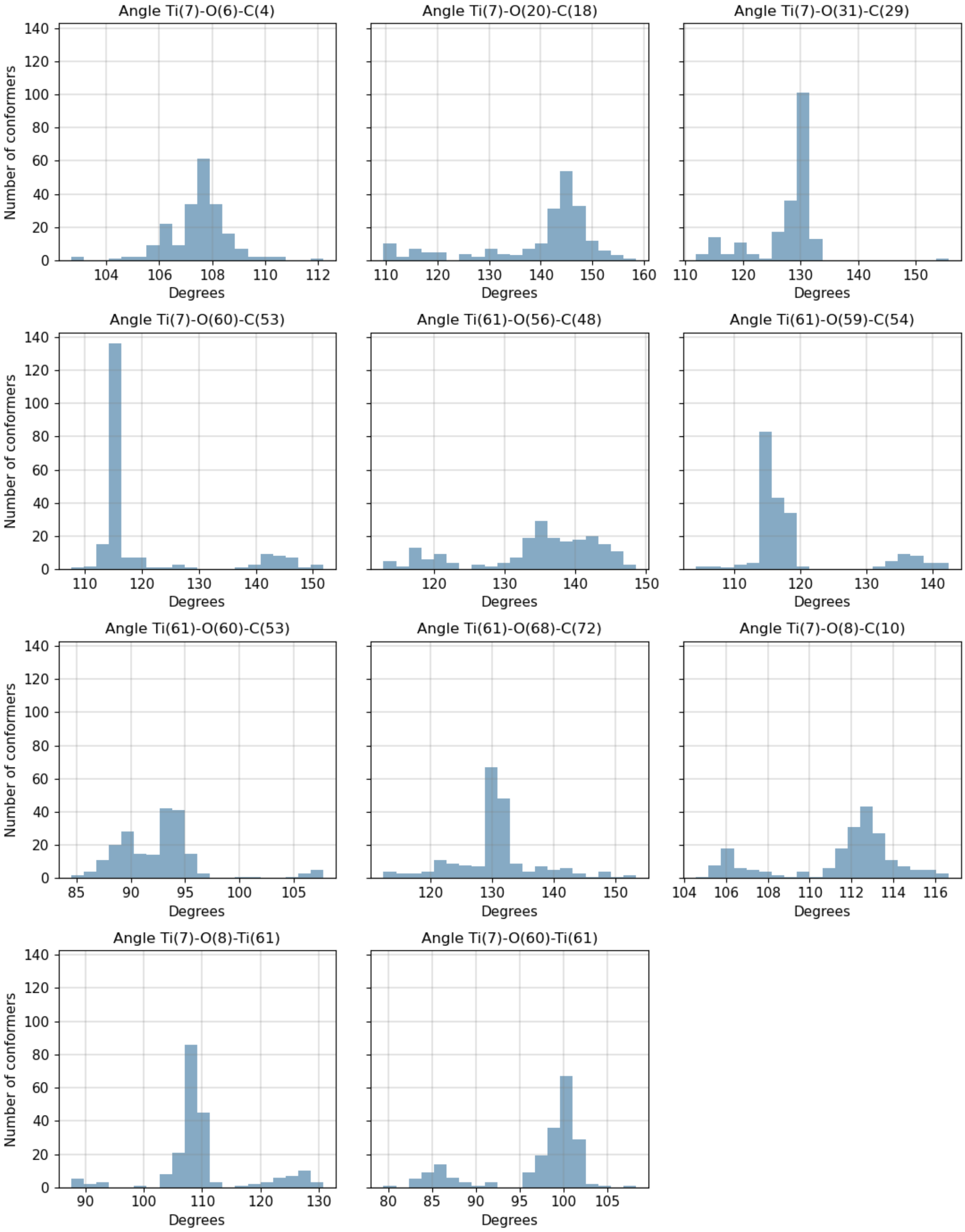


Figure A3.5. Selection of Ti-O-C and Ti-O-Ti angles at the end of full optimization with PBE0/def2-SV(P). Ti-O-C angles show a significant shift towards straightening compared to Figure A3.4. which shows the same angles, but at the start of the DFT optimization.

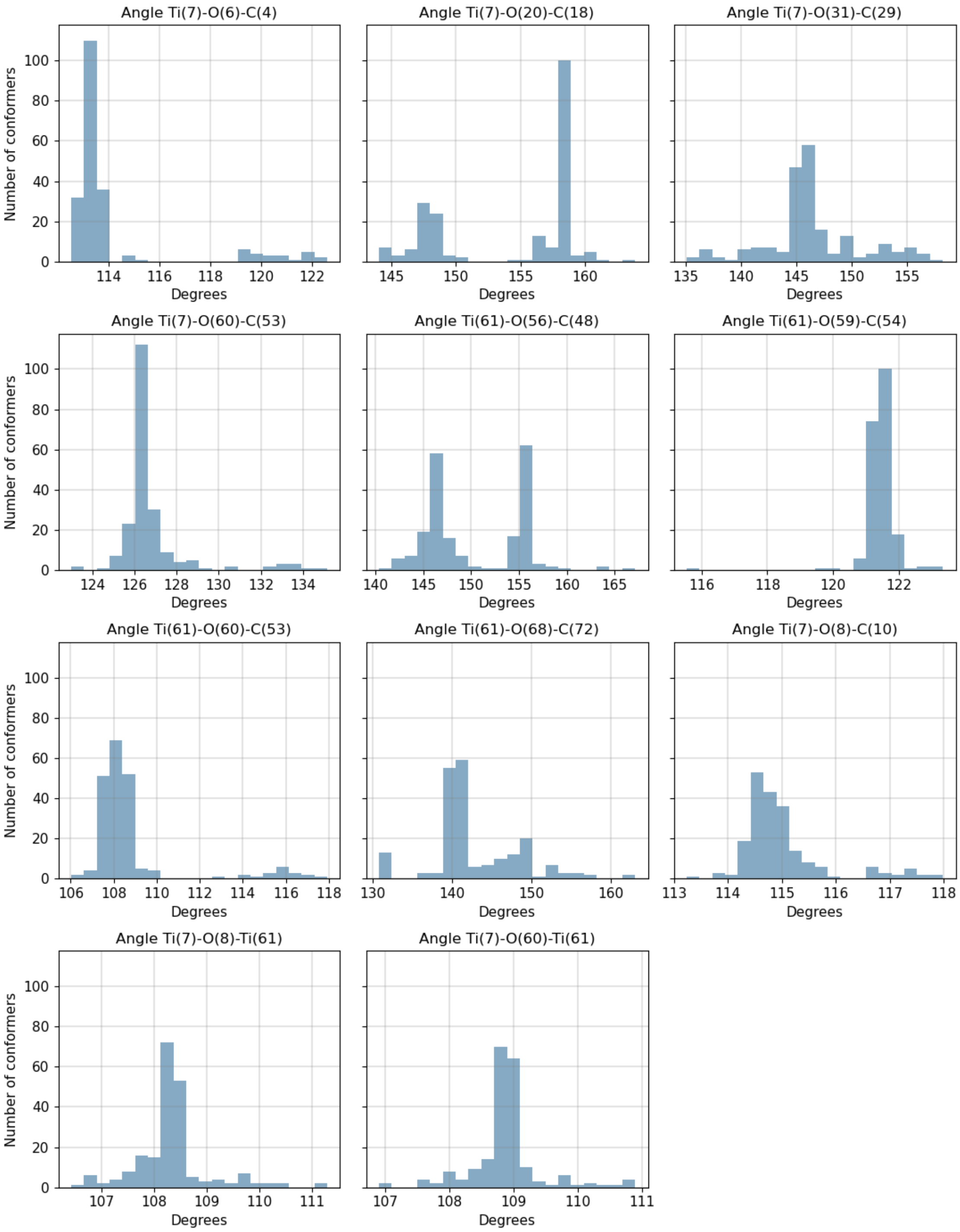


Figure A3.6. Selection of dihedral angles at the start of optimization with PBE0/def2-SV(P). Indicating the coverage of the search space by found conformers by CREST.

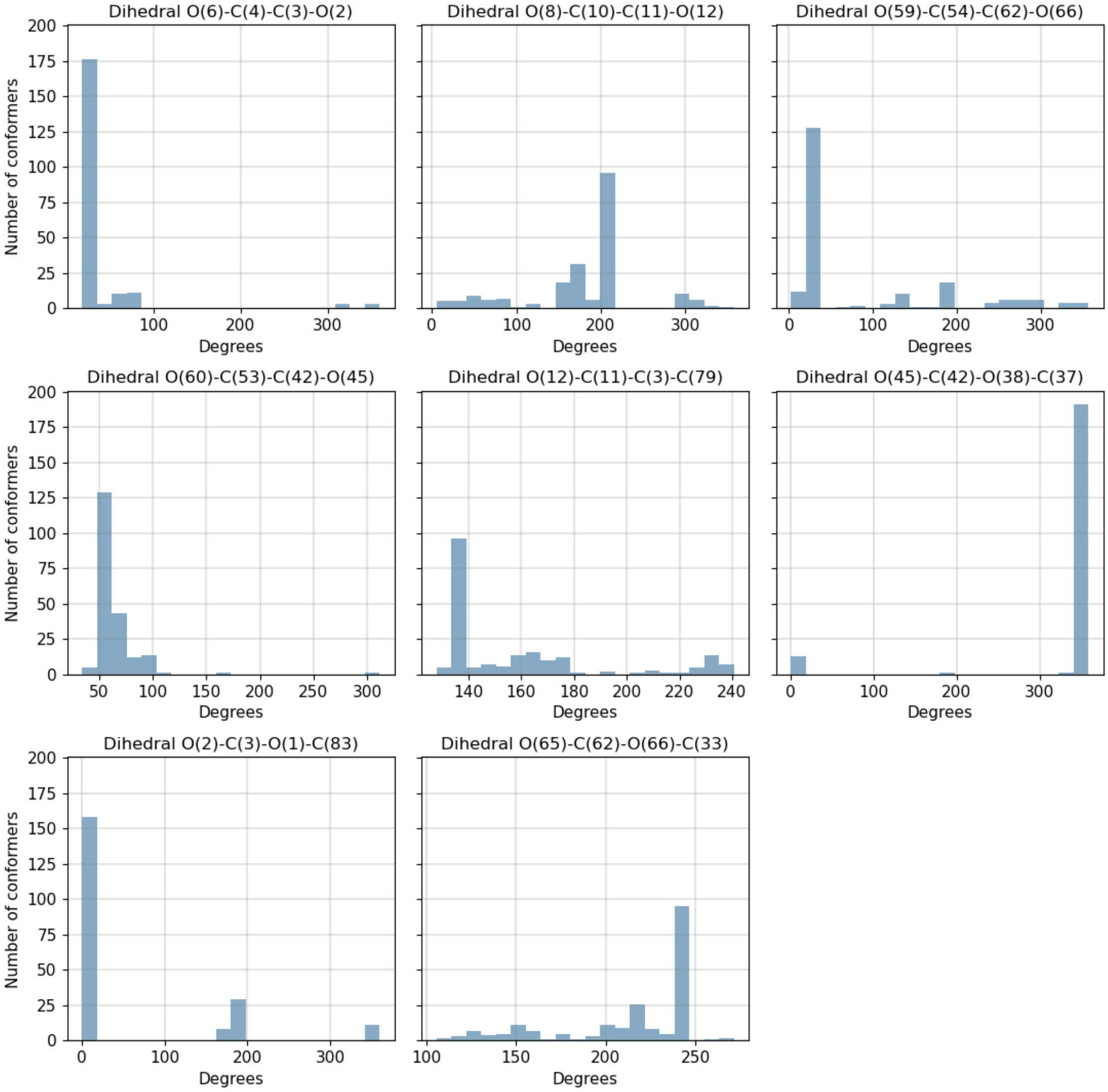
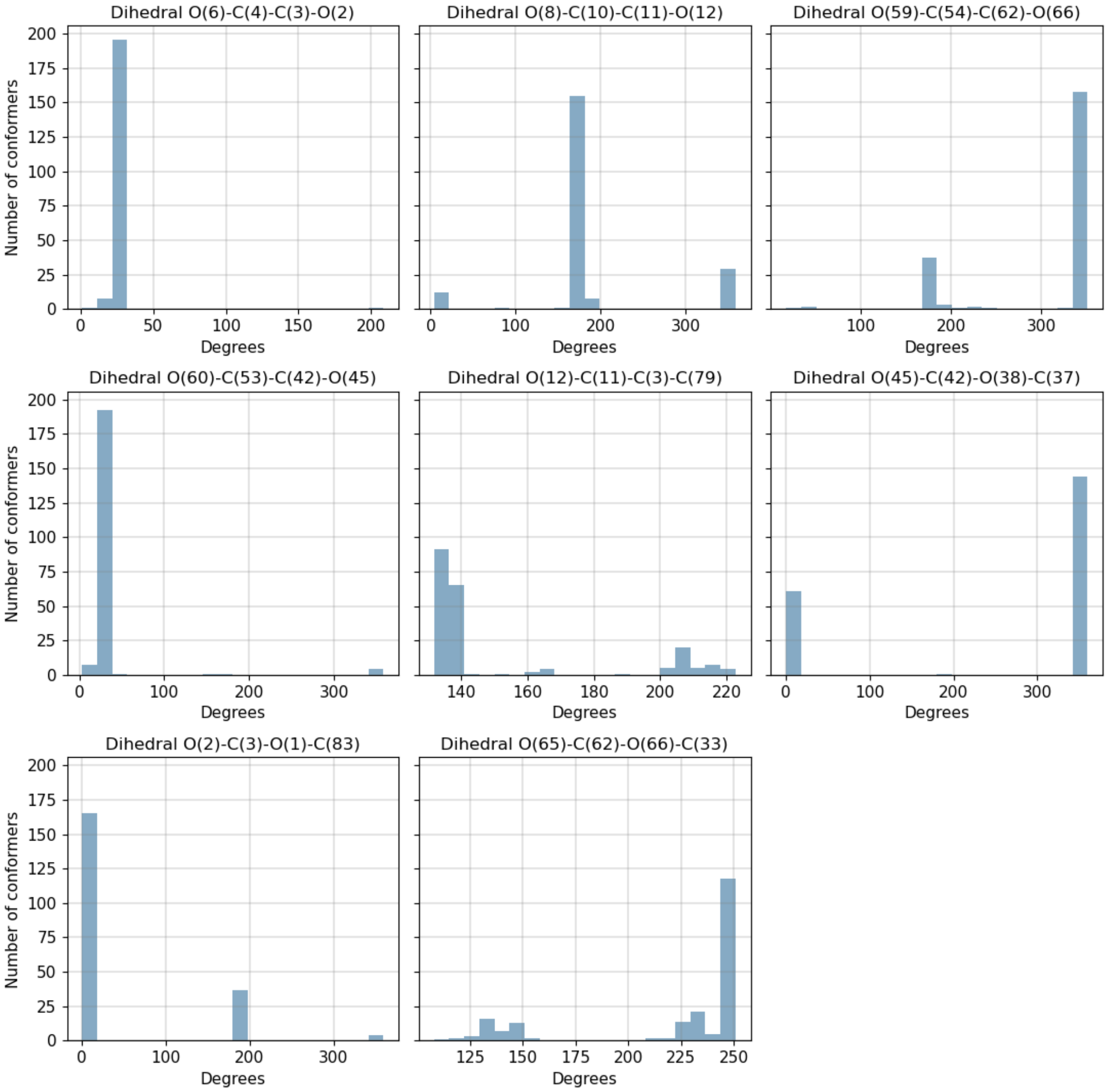


Figure A3.7. Selection of dihedral angles at the end of full optimization with PBE0/def2-SV(P). Indicating the coverage of the search space by fully optimized conformers.



Appendix 4 - Molli

Molli [40] is a newly created open-source library. Written in Python with the heavy lifting of computational chemistry related functionality based on an excellent open-source library called ASE [35].

Main workflows using Molli:

- EDA - Exploratory Data Analysis.
- Dataset generation.
- Analysis: molecules, optimization trajectories, ensembles of molecules etc.
- Helper functions.

A4.1. EDA - Exploratory Data Analysis

With the advent of machine learning competitions held by Kaggle [54] emerged a new kind of data science and visual analysis related subfield. It is quite a common situation that we have a lot of data, but we lack the idea of what the data presents. Are there any interesting patterns to be detected? How diverse is the data? Are there any common descriptors that allow grouping and clustering of the data? As you can imagine the number of questions one could ask is limitless. That is the main goal of the EDA - Exploratory Data Analysis. You start to poke around with the data, visualize various properties from different viewpoints, analyze various metrics etc. Readers of an EDA should get a quick overview of the main attributes and patterns found in the data. It doesn’t matter if none are found - then the reader knows that information and can draw his/her own conclusions about the data. EDA is not a definitive ruleset or a proof of how things are or should be. EDA is rather a nice informative visual guide to better understand the data.

A4.2. Dataset generation

Recent years have shown tremendous progress in machine learning. Although many contribute this success to the advances in hardware and to some extent to algorithms, one could argue that the role of datasets has played an equally important part. While performing long lasting DFT calculations, computational chemistry software produces huge amounts of valuable intermediate data that is mostly ignored as an artefact and only results are gathered and published. This intermediate data could potentially be a game changer for a machine learning pipeline in the future. Similarly high-quality dataset consisting of DFT calculations over some well-defined chemical structures could potentially be an invaluable asset for a broader audience of computational chemists. Especially concerning transition metal complexes. On a smaller scale one could compile a dataset within the context of experiments done in this thesis. Molli has been built keeping these considerations in mind. Molli helps to navigate, analyze and extract useful information easily to compile a dataset for testing particular hypotheses or even for more general purposes. There are already multiple good resources that host either datasets or references to datasets [55], [56].

A4.3. Analysis: molecules, optimization trajectories, ensembles of molecules etc.

As said before computational chemists drown in the abundance of the data. Molli comes to help here. Whether you want to sort molecules by the value of a certain angle between 3 atoms or by the value of some inter atomic distances etc. All this and much more can be easily done with Molli. Compare optimisation trajectories, etc.

A4.4. Helper functions for a computational chemist

Molli contains many practical and useful utility functions regarding parsing and processing several common file formats that are used in computational chemistry.

- Gaussian log file parser, extracting useful information from the log files produced by Gaussian software for further analysis.
- xyz-file parser allows reading and writing of xyz-files.
- Align molecular geometries so that their root mean squared distance is minimized.
- etc.