

# Exploring Molecular Conformational Space

im Fachbereich Physik der Freien Universität Berlin eingereichte

Dissertation zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM

vorgelegt von

Adriana Supady

Berlin, 2016

Erstgutachter (Betreuer): Prof. Dr. Matthias Scheffler  
Zweitgutachter: Prof. Dr. Roland Netz  
Tag der Disputation: 1. September 2016

## ABSTRACT

---

Flexible organic molecules and biomolecules can adopt a variety of energetically favorable conformations that differ in chemical and physical properties. The identification of such low-energy conformers is a fundamental problem in molecular physics and computational chemistry. Here we describe our efforts to develop methods exploring molecular conformational spaces at the first-principles level.

We present a genetic algorithm (GA) based search for sampling the conformational space of molecules. This GA is available in the Python library Fafoom and has been developed in this thesis. The GA search aims not only at finding the global minimum, but also at identifying all conformers within a certain energy window. The implementation of the GA search is designed to work with first-principles methods, facilitated by the incorporation of local optimization and blacklisting conformers to prevent repeated evaluations of very similar solutions. The performance of the GA search is evaluated for seven amino-acid dipeptides and eight drug-like molecules. The evaluation focuses on: (i) how well the GA search can reproduce the reference data; and (ii) how well the conformational space is covered. Our study shows that the GA search samples the conformational space of the evaluated molecules with high accuracy and efficiency.

For the purpose of the investigation of the dynamics of the conformational ensemble, we propose a strategy to construct a reduced energy surface from low-energy minima and selected transition states. The strategy selects pairs of conformers for the optimization of the transition states. The resulting energy barriers are then arranged into a barrier tree, a convenient representation of a high-dimensional energy surface. The method is evaluated for: (i) the alanine tetrapeptide, at the force-field level, where it matches the findings of free-energy simulations; and (ii) a synthetic peptide, employing first principles, where the resulting barrier-tree representation helps interpreting the experiment.

Accurate predictions of properties, e.g. catalytic activity, require identification of energetically favorable 3D structures. We investigate the relation between the adopted 3D structures and the catalytic activity in eight (thio)urea based compounds. The conformational preferences of the (thio)urea based compounds significantly differ between each other. The investigation of the interaction between an example thiourea catalyst and a model substrate reveals that only in its active form can the catalyst activate the substrate.

## ZUSAMMENFASSUNG

---

Flexible organische und biologische Moleküle können verschiedene 3D Konformationen annehmen, die unterschiedliche chemische und physikalische Eigenschaften aufweisen. Die Suche nach energetisch günstigen Konformeren ist ein fundamentales Problem der Molekülphysik und Computerchemie. In der vorliegenden Arbeit stellen wir Methoden vor, die der Untersuchung des molekularen Konformationsraumes dienen und die *ab initio*-Methoden verwenden.

Wir präsentieren eine Suchtechnik, die unter Verwendung eines genetischen Algorithmus (GA) den Konformationsraum durchsucht. Diese Suchtechnik wurde als Teil der Python-Bibliothek Fafoom implementiert, die im Rahmen dieser Doktorarbeit entwickelt wurde. Ziel der GA-basierten Suchtechnik ist es das Auffinden des globalen Minimums und aller lokalen Minima in einem bestimmten Energiefenster. Die effiziente Verwendung rechenintensiver *ab initio*-Methoden wird durch die Durchführung lokaler Optimierungen und das Vermeiden der Auswertung von bekannten Lösungen unterstützt. Die Suchtechnik wurde eingesetzt um den Konformationsraum von sieben Dipeptiden und acht Arzneistoff-ähnlichen Molekülen zu untersuchen. Im Anschluss wurden folgende Punkte überprüft: (i) wie gut kann die Suchtechnik die Referenzdaten reproduzieren; und (ii) wie gut ist der Konformationsraum erforscht worden. Unsere Studie zeigt, dass die GA-basierte Suchtechnik den Konformationsraum der untersuchten Moleküle mit hoher Genauigkeit und Effizienz probt.

Wir präsentieren eine Strategie, die eine vereinfachte Darstellung der Energiefläche bietet um eine Untersuchung der Vielfalt des Konformationsensembles zu ermöglichen. Die vereinfachte Darstellung besteht aus energetisch günstigen lokalen Minima und ausgewählten Übergangszuständen. Die resultierenden Energiebarrieren werden verwendet um die vieldimensionale Energiefläche in Form eines Energiebaumes anschaulich darzustellen. Folgende Moleküle wurden mit der Methode untersucht: (i) das Alanin-Tetrapeptid mit Hilfe von Molekülmechanik-Rechnungen und (ii) ein synthetisches Peptid unter Verwendung von ersten Prinzipien. Die für das Alanin-Tetrapeptid gewonnenen Resultate stimmen mit den Erkenntnissen aus Vergleichssimulationen überein. Das für das synthetische Peptid konstruierte Energie-Baumdiagramm unterstützt die Interpretation von experimentellen Daten.

Die Bestimmung von energetisch günstigen Konformeren ist zur korrekten Vorhersage von Eigenschaften notwendig. Wir untersuchen den Zusammenhang zwischen der 3D-Struktur und der katalytischen Aktivität von acht (Thio-)Harnstoffverbindungen. Die Unterschiede zwischen den strukturellen Präferenzen von den (Thio-)Harnstoffen sind signifikant. Die Untersuchung der Interaktion zwischen einem Thioharnstoff basierten Katalysator und einem Modellsubstrat hat ergeben, dass nur ein bestimmtes Konformer des Katalysators das Substrat aktivieren kann.

# CONTENTS

---

1	INTRODUCTION	1
2	INVESTIGATING MOLECULAR STRUCTURES	5
2.1	Flexible organic molecules	5
2.1.1	Describing molecular structures	5
2.1.2	Geometrical similarity of structures	6
2.1.3	Technical details	8
2.2	Theoretical methods for treatment of molecular structures	8
2.2.1	Force fields	9
2.2.2	Towards quantum mechanics	10
2.2.3	The Hartree-Fock approximation and beyond	11
2.2.4	Density-Functional Theory	12
2.2.5	Dispersion corrections	14
2.2.6	Solvent models	15
2.3	Energy landscapes	16
2.3.1	Characterizing the Potential-Energy Surface	16
2.3.2	Towards Free-Energy Surface	20
2.4	Packages for molecular simulations	21
2.4.1	FHI-aims	21
2.4.2	ORCA	21
2.4.3	Tinker	22
3	FLEXIBLE ALGORITHM FOR OPTIMIZATION OF MOLECULES	23
3.1	Motivation	23
3.2	Fafoom: implementation details	24
3.2.1	Structure of the package	24
3.2.2	Example of use: genetic algorithm based search	26
3.3	Fafoom: benchmark calculations	28
3.3.1	Amino acid dipeptides	29
3.3.2	Drug-like molecules	33
3.4	Fafoom: and beyond	36
3.4.1	Applications	36
3.4.2	Free choice of the external simulation package	38
3.4.3	Extensibility of the kind of optimized degrees of freedom	38
3.4.4	Parallelization	40
4	REDUCED MOLECULAR POTENTIAL-ENERGY LANDSCAPES	41
4.1	Motivation	41
4.2	Description of the method	43
4.3	The example of AcAla <sub>3</sub> NMe	45
4.3.1	Reference data for comparison	46
4.3.2	Sampling of the potential-energy surface	46
4.3.3	Towards the network of states	47
4.3.4	Calculations of the energy barriers	48
4.4	Reduced PES of a synthetic peptide from first principles	51
4.4.1	Sampling of the potential-energy surface	51
4.4.2	Towards a network of states	52
4.4.3	Calculations of the energy barriers	53

5	GLOBAL STRUCTURE SEARCH FROM FIRST PRINCIPLES: THIOUREA CATALYSTS	57
5.1	Motivation	57
5.1.1	Structure-dependent catalytic activity	57
5.1.2	Thioureas acting as organocatalysts	59
5.1.3	Studied problem	60
5.2	Molecular systems	60
5.3	Isolated molecules	61
5.3.1	Energetics	61
5.3.2	Energy barriers	67
5.4	Catalyst-substrate complexes	67
6	CONCLUSIONS AND OUTLOOK	71
6.1	Summary	71
6.1.1	Fafoom - Flexible algorithm for optimization of molecules	71
6.1.2	Reduced representation of molecular landscapes	71
6.1.3	Structure-dependent catalytic activity	72
6.2	Conclusions	72
6.3	Outlook	73
A	APPENDIX	75
A.1	Fafoom parameter	75
A.2	Fafoom benchmark: parameter lists	77
A.3	(Thio)urea molecules: impact of different dispersion corrections	78
	SELBSTÄNDIGKEITSERKLÄRUNG	79
	CURRICULUM VITAE	81
	PUBLICATIONS	83
	ACKNOWLEDGMENTS	85
	BIBLIOGRAPHY	87

## INTRODUCTION

---

The properties of a chemical compound are determined by its structure [1]. This general formulation is one of the most fundamental concepts in chemistry, physics, and materials science. The chemical structure is given by the molecular composition, i.e. type and count of atoms, and their spatial arrangement.

The interest in studying properties of a chemical compound drives the demand to identify its structure. Thus, a number of experimental techniques for determination of molecular structures have been developed. The most popular technique is X-ray diffraction, which allows for resolving the arrangement of atoms in a solid crystal. Other methods, e.g. neutron scattering and electron microscopy, can also be utilized to identify structures.

An alternative to experimental structure determination is theoretical structure prediction. The idea of structure prediction is to obtain a structure without prior experimental knowledge. Structure prediction is an integral part of diverse research areas, such as: (i) biophysics, e.g. in protein structure prediction [2] or molecular docking [3–5]; (ii) molecular physics, e.g. in the analysis of gas-phase spectra of biological molecules [6]; (iii) solid and surface science, e.g. in crystal structure prediction [7] or adsorbate-surface modeling [8]; (iv) catalyst research, e.g. in studying the environment-dependent stability of metal clusters [9].

The persistent increase in the availability of computational resources has opened the way for high-throughput calculations at large scale. This is not only facilitated by the availability of bigger and faster computers, but also by the advances in applications benefiting from distributed computing, such as the World Community Grid [10].

Parallel to the impressive infrastructure development, numerous advances on the algorithmic side have been achieved. Accuracy and performance of modern quantum chemistry software have been improved by, e.g. efficient parallelization, implementation of recent theoretical methods, and extension of the selection of available functionalities.

The amount of data that can be generated nowadays is immense. Nevertheless, independently from the number of performed calculations or applied postprocessing, the quality of the final results depends on the accuracy of the underlying model. This concerns many aspects, starting from the formulation of the research question, through the level of the utilized theory, to the numerical precision. This thesis addresses some of these aspects in the context of modeling small to medium-sized organic molecules.

The relation between the state of a system and its potential energy can be described with a multi-dimensional function, the potential-energy surface (PES). This state can be explicitly characterized by a set of parameters, referred to as degrees of freedom (DOF). Real-world systems usually have several degrees of freedom and it is difficult to calculate the potential energy for all possible states of such systems. Thus, the complete PES usually re-

mains uncharted, but it can be well-characterized by its stationary points, in particular minima and transition states.

The identification of PES minima that represent energetically favorable states is one of the most fundamental problems in computational chemistry. Thus, the PES minima are usually targeted first when investigating energy landscapes. This is usually approached by formulating an optimization problem with the goal to define and find such DOF values that minimize the energy function.

In addition to the states that minimize the energy, the transition states are relevant for the characterization of a system. Mathematically, a transition state is a first-order saddle point, i.e. a maximum in exactly one direction and a minimum in the remaining directions. Knowledge of transition states is required for the calculation of transition rates and further description of the dynamics of the system.

A number of approaches and tools have been developed for the exploration of different aspects of the PES, aiming at connecting the observable properties with the underlying landscape [6, 11]. Diverse sampling algorithms have been developed allowing for finding PES minima and several methods can be employed to identify the transition states. Moreover, diverse attempts to utilize the PES to derive the free-energy surfaces (FES) have been made [12]. Finally, convenient visualizations in the form of, e.g. disconnectivity graphs [13] and barrier trees [14] have been suggested. A detailed review of the energy landscapes field is given by Wales [15].

The approaches for the investigation of the PES utilize a theoretical model to determine the potential energy of a given state. Commonly, computationally feasible empirical models with fitted parameters, e.g. force fields are used. However, the performance of force fields critically depends on the initial parametrization and may lead to considerable rearrangements of the PES [16, 17]. An attractive alternative is offered by the density-functional theory (DFT) approximations which offer an improved accuracy at a reasonable computational cost. It should be noted that dispersion effects need to be included to correctly describe (bio)organic molecules [18–23].

This thesis focuses on the development of methods for the investigation of the PES of flexible organic molecules with DFT. In the following, the most essential foundations of the work are formulated.

Flexible organic molecules can usually adopt a variety of energetically favorable conformers. The propensity to adopt a certain conformation is strongly dependent on the environment and can also be induced by interactions with other molecules [24, 25]. Thus, instead of focusing on the global minimum of the PES, a coverage of the accessible conformational space is sought.

The chosen level of theory, DFT, offers an attractive balance of accuracy and cost. However, two factors need to be considered. First, the applied DFT approximation should be carefully chosen and the dispersion effects should be incorporated. Second, the computational cost of DFT calculations is significantly higher compared to the cost of force-field evaluations. Hence, it should be ensured that any redundant calculations are avoided.

On the technical side, the methods implemented in this thesis were designed for general use. This is accomplished by: (i) an implementation design with a clear separation of the energy evaluation from the handling of the



molecular structures and (ii) making the implementation publicly available and open for contributions.

This thesis is divided into four chapters:

1. The first chapter details the theoretical background concerning:
  - the characteristics and practical handling of the investigated systems
  - energy modeling
  - methods for PES exploration
2. The second chapter describes the implementation and performance of the Python package Fafoom - Flexible algorithm for optimization of molecules [26]. Fafoom is a user-friendly open-source tool for sampling the PES of organic molecules. The following features of Fafoom should be highlighted:
  - Fafoom can be interfaced to different simulation packages.
  - The great number of parameters facilitates the creation of user-tailored searches while ready-to-use sets of parameters are also provided.
  - The optimized degrees of freedom can be freely selected and defined.

Fafoom was released under the LGPL license in early 2015 and has already benefited from external contributions extending its functionality and further development is ongoing.

3. The third chapter presents a framework for the construction of energy barrier trees along with a proof of concept and an application to a real-world system. The main idea of the framework is to perform a smart guess of likely connected pairs of minima that limits the number of calculations but provides meaningful insights into the investigated system.
4. The last chapter investigates the relation between the adopted 3D structure and catalytic activity of eight (thio)urea based molecules. Further, the conformational preferences of one of the molecules are studied in the presence of a model substrate. This study demonstrates that prediction of properties requires a broad sampling of the PES.

Finally, the thesis summarizes the conclusions and presents an outlook.



INVESTIGATING MOLECULAR STRUCTURES

---

The aim of this chapter is to familiarize the reader with the studied systems and with the methods that were used to explore and describe the molecular energy surfaces. The overview presented here is by intention brief and the focus is given only to aspects that: (i) the author of this thesis considered critical for understanding the thesis or (ii) that were the subject of methodological development. None of the sections of this chapter aims for completeness, but provides starting points for further exploration to an interested reader. Besides this, the author of this thesis found the following resources remarkably helpful: books [15, 27], lecture scripts [28], and dissertations [29, 30].

## 2.1 FLEXIBLE ORGANIC MOLECULES

The systems investigated in this thesis are small to medium-sized, flexible organic molecules. We first present the setup for the handling of the molecules and discuss related issues.

2.1.1 *Describing molecular structures*

Figure 1 depicts popular chemical representations on the example of acetylsalicylic acid, also known as aspirin. The chemical formula stores only the composition of the molecule. The simplified molecular-input line-entry system (SMILES) [31] string is a convenient representation as it allows for encoding the connectivity, the bond type (single, double, triple, and aromatic), and the stereochemical information in a one-line notation. It should be noted that a number of valid SMILES codes can be constructed for the same molecule. Thus, the concept of a 'canonical SMILES' emerged to provide a single, unique SMILES string. However, the exact formulation of the canonical SMILES string of a given molecule depends on the utilized canonicalization algorithm [32]. The great advantage of the SMILES strings is that they are intuitive and can be easily read and written by a human. An alternative one-line representation is the International Chemical Identifier (InChI) [33], which provides the canonical representation of a chemical substance. Both SMILES or InChI can be used to generate a schematic, 2D representation of a molecule. Finally, the last missing piece of information, namely the spatial arrangement of atoms, is revealed in a 3D representation of a molecule. Two types of coordinates are commonly employed to represent a molecular 3D structure: Cartesian and internal coordinates. In Cartesian coordinates, each atom is represented as a point in 3D space. Cartesian coordinates are universal, intuitive and always relative to the origin of the coordinate system. The simplest internal coordinates are based on the 'Z-matrix coordinates', i.e. which include bond lengths, bond angles as well as dihedral angles (torsions). The alternative chemical representations are depicted in Figure 1. The main advantage of the internal coordinates is that they are orientation- and location-invariant,

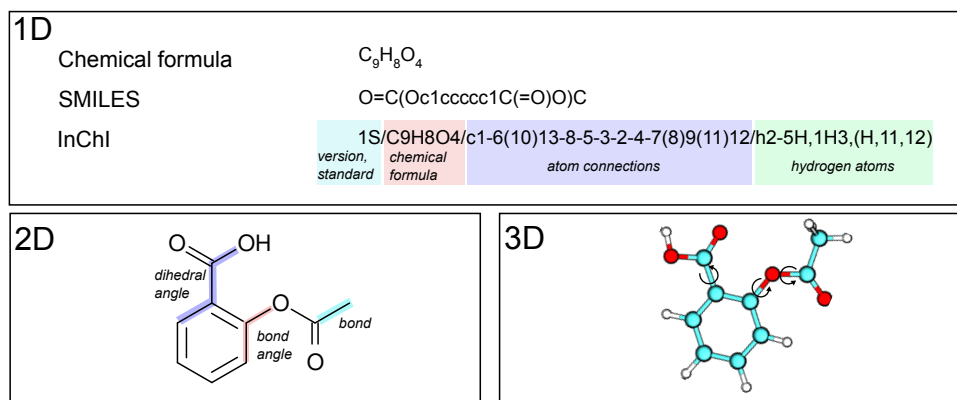


Figure 1: Alternative chemical representations on the example of acetylsalicylic acid.

i.e. they remain unchanged upon translation and rotation in 3D space. Bond lengths and bond angles usually have only one minimum and the energy will increase rapidly if these parameters adopt non-optimal values. On the other hand, there is no single minimum for the value of a dihedral angle and in most cases, diverse values can be adopted. The adopted values depend on the neighboring atoms/functional groups and on the steric interactions within the conformation.

In this thesis, the global searches for structures of small to medium sized organic molecules operate on significant internal coordinates, mostly dihedral angles. We refer to these significant coordinates as relevant degrees of freedom (DOF). The local optimizations of structures are performed under the consideration of all Cartesian coordinates.

We consider only single, non-ring bonds between non-terminal atoms to be fully rotatable bonds. Bonds attached to methyl groups that carry three identical substituents are excluded due to the three-fold rotational symmetry of such groups. Further, we utilize the *cis/trans* nomenclature for describing the relative orientation of functional groups within a molecule. In cases where the functional groups are oriented in the same direction we refer to such orientation as *cis*, whereas, when the groups are oriented in opposite directions, we refer to it as *trans*.

The full representation of a 3D structure is the list of its Cartesian coordinates. An alternative way to store 3D structures is to use a reduced representation that contains the SMILES and DOFs with the corresponding values. The difference between these two alternative representations is illustrated in Figure 2. The substantial advantage of the reduced representation is that the only stored data are simultaneously the DOFs for the optimization. This is extremely convenient, especially for larger systems. Nevertheless, one should keep in mind that the reduced representation stores no information about the bond lengths and bond angles as it assumes no substantial changes of these coordinates.

### 2.1.2 Geometrical similarity of structures

The quantification of the similarity between two molecules is a common problem that needs to be solved, e.g. in order to remove duplicates from a pool

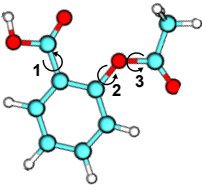
full				reduced
	O	2.25740	-2.41400	-2.13540
	C	2.07120	-1.74730	-1.12750
	O	2.99120	-0.92100	-0.48160
	C	4.38630	-0.93170	-0.71760
	C	4.94000	-1.29700	-1.94780
	C	6.32790	-1.31480	-2.10730
	C	7.16220	-0.95270	-1.05140
	C	6.61110	-0.56430	0.16970
	C	5.22080	-0.54100	0.34390
	C	4.65430	-0.13190	1.65510
	O	3.60630	-0.48440	2.15240
	O	5.45550	0.76400	2.27210
	C	0.81020	-1.77870	-0.31670
	H	4.32050	-1.57430	-2.79550
	H	6.75850	-1.61630	-3.05970
	H	8.24180	-0.96910	-1.17840
	H	7.27370	-0.28880	0.98710
	H	4.96850	1.00050	3.08880
	H	-0.00870	-2.16660	-0.92850
H	0.54150	-0.77440	0.01960	
H	0.96350	-2.43790	0.54200	
				<b>SMILES:</b> <chem>O=C(Oc1ccccc1C(=O)O)C</chem> and <b>Significant degrees of freedom:</b> rotatable bond 1 (C-C-C-O): 30 ° rotatable bond 2 (C-C-O-C): 29 ° rotatable bond 3 (C-O-C-C): 161 °

Figure 2: The comparison of a full and reduced representation of 3D structure of the acetylsalicylic acid. The full representation contains all atomic coordinates in an arbitrary Cartesian coordinate system. The reduced representation consists of the SMILES string and dictionary of the rotatable bonds together with the corresponding values.

of 3D structures. In this thesis, only the similarity of different 3D structures of the same chemical molecule is of interest. Similarity of chemical molecules having different chemical composition or chirality is beyond the scope of this work. The most popular method to quantify the similarity is the root-mean-square deviation RMSD, calculated for two sets of Cartesian coordinates.

**ROOT-MEAN-SQUARE DEVIATION (RMSD)** Given two 3D geometries of a molecule with  $N$  atoms, the formula for the RMSD is defined as follows:

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^N d_i^2}, \quad (1)$$

where  $d_i$  is the distance between the atomic positions. Although fast to calculate, the RMSD value is meaningless until the best superposition of the geometries is identified. The most popular algorithm for finding the best alignment of two sets of coordinates is the Kabsch algorithm [34]. After translating the centroids of the two sets of coordinates to the origin of the coordinate system, the Kabsch algorithm computes the optimal rotation matrix that minimizes the RMSD. Hydrogen atoms are commonly ignored in RMSD calculations. There are multiple advantages of using the Cartesian RMSD, e.g. it is a well-recognized metric, it is easy to calculate and reproduce, and it is available as a basic functionality in most of the modeling packages.

**TORSIONAL RMSD (tRMSD)** Instead of using Cartesian coordinates, the values of the significant torsional degrees of freedom, i.e. rotatable bonds, can be used. Analogically to the Cartesian RMSD, given two 3D geometries with  $m$  rotatable bonds, the formula for tRMSD reads:

$$\text{tRMSD} = \sqrt{\frac{1}{m} \sum_{i=1}^m \theta_i^2}, \quad (2)$$

where  $\theta_i$  is the angular difference between values of the corresponding dihedral angles. In contrast to the Cartesian RMSD, the calculation of the tRMSD

does not require any fitting. The value of the tRMSD is also easier to interpret. The major drawback of the tRMSD is the necessity to always provide a list of considered torsions in order to ensure reproducibility.

In this thesis, both similarity descriptors are utilized.

### 2.1.3 *Technical details*

A few technical details should be mentioned here. The first issue is the treatment of achiral molecules, i.e. molecules whose 3D structure is identical with the mirrored 3D structure. For achiral molecules, in order to determine the similarity of the target and reference structure, the calculation of the RMSD/tRMSD is performed twice: once for the target and the reference and once for the target and the mirror of the reference. The lower of the resulting values is then considered.

A further issue that may lead to an increased and incorrect RMSD/tRMSD value is the potential automorphisms of the functional groups. An example of an automorphic group is the phenyl group. If a phenyl group rotates by  $180^\circ$ , the structure remains unchanged. The indices of the atoms are swapped but the bond connectivity does not change. Thus, in order to ensure, that the obtained RMSD/tRMSD represents a meaningful value, it should always be verified if the utilized software handles the automorphism correctly, i.e. is 'symmetry corrected'.

In addition to SMILES strings, which are utilized for the representation of molecules, SMARTS strings are employed for substructure searches, e.g. in order to find rotatable bonds. SMARTS (SMiles ARbitrary Target Specification) strings allow to define patterns including logical operators. For more information please refer to the original tutorials [35].

In this thesis, two cheminformatics packages have been used in order to handle the molecular structures: Open Babel [36] (mostly for format conversions) and RDKit [37]. Both, Open Babel and RDKit compute 'symmetry corrected' RMSD values. For visualization purposes following packages have been used: Avogadro [38], VMD [39] and Jmol [40].

## 2.2 THEORETICAL METHODS FOR TREATMENT OF MOLECULAR STRUCTURES

Computational chemistry offers a variety of methods to tackle molecule structures. The concepts behind the methods differ in the underlying physics from very straightforward mechanistic descriptions (force fields) to complex coupled-cluster methods accounting for electron correlation. Prior to the decision on the method, the following aspects should be considered:

1. Characteristics of the investigated systems - in particular the number and kind of (heavy) atoms and the number of degrees of freedom.
2. The task to perform - in general, for searches in the molecular conformational space, computationally feasible methods should be utilized and followed by more accurate methods in the refinement step.
3. The utilized software and infrastructure - in particular their capabilities and limitations.

It is always advisable to verify the performance and accuracy of the utilized method by comparing to: (i) results obtained with higher level methods; (ii) experimental findings. Comparisons can be made with, e.g. S22 [41], a standard set of non-covalent interactions that has been developed for benchmarking.

In the following subsections we give a brief overview of some popular methods used to calculate the properties of molecules.

### 2.2.1 Force fields

The potential-energy surface of large molecules is often evaluated with force fields. A force field comprises both the functional form and the empirically derived parameters. The functional form of a classic force field can be written as follows:

$$E = \underbrace{E_{\text{bond}} + E_{\text{angle}} + E_{\text{dihedral}}}_{E_{\text{bonded}}} + \underbrace{E_{\text{electrostatic}} + E_{\text{van der Waals}}}_{E_{\text{non-bonded}}}, \quad (3)$$

and in more detail:

$$E = \sum_{\text{bonds}} k_r (r - r_q)^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{v_n}{2} \times [1 + \cos(n\varphi - \gamma)] + \sum_{i=1} \sum_{i < j} \frac{q_i q_j}{4\pi\epsilon_0 R_{ij}} + \sum_{i=1} \sum_{i < j} 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]. \quad (4)$$

The force field parameters are obtained by fitting quantum-mechanical calculations [42–49] and experiments [47–49]. The parametrization against experimental findings might utilize structures from 3D shape databases, values obtained with X-ray diffraction, and neutron diffraction. Force fields are very fast but strongly dependent on the molecule class used in the parametrization step. As a consequence, no good universal force field exists, and different force fields give significantly different results. They are typically used for different systems types: CHARMM [47], Amber [44, 45], OPLS-AA [43] are used for modeling of large biomolecules, e.g. proteins; small organic and drug-like molecules are often described by the Merck Molecular [42] force fields; while GLYCAM [46] force fields model oligosaccharides.

Several enhancements have been suggested in order to improve the accuracy of the force field and led to the development of polarizable and reactive force fields. If an appropriate force field is used, a reasonable insight into the general properties of a molecule can be obtained. However, considerable rearrangements of the true PES have been observed in a number of force-field based evaluations [16, 17, 50]. Furthermore, a significant disagreement of the predictions of the dynamic system properties obtained from simulations with different protein force fields was reported recently [51]. Because of these inaccuracies resulting from classical mechanics, it is clear the electrons need to be treated explicitly in order to obtain a more accurate description of the PES.

## 2.2.2 Towards quantum mechanics

The many-body problem is at the basis of modeling properties and functions of materials. The total energy of a system can be found by solving the time-independent Schrödinger equation:

Schrödinger  
equation (1926)

$$\hat{H}\psi(\mathbf{x}_i, \mathbf{R}_i) = E\psi(\mathbf{x}_i, \mathbf{R}_i), \quad (5)$$

where  $\hat{H}$  is the Hamiltonian,  $E$  the total energy, and  $\psi$  the wave function.  $\psi$  is the many-body wave function depending on the nuclear coordinates,  $\mathbf{R}_i$ , and the spatial coordinates and spin states of the electrons  $\mathbf{x}_i = (\mathbf{r}_i, \sigma_i)$ .

For a molecular system with  $N_{el}$  electrons and  $N_{at}$  atom nuclei, the Hamilton operator (Eq. 6) is defined as follows:

$$\hat{H} = \hat{T}_n + \hat{T}_{el} + \hat{V}_{ee} + \hat{V}_{ne} + \hat{V}_{nn}, \quad (6)$$

where the terms are defined as follows in atomic units:

$$\begin{aligned} \hat{T}_n &= -\sum_{\Lambda=1}^{N_{at}} \frac{1}{2M_{\Lambda}} \nabla_{\Lambda}^2 && \text{the kinetic energy of the nuclei,} \\ \hat{T}_{el} &= -\sum_{i=1}^{N_{el}} \frac{1}{2} \nabla_i^2 && \text{the kinetic energy of the electrons,} \\ \hat{V}_{ee} &= \sum_{i=1}^{N_{el}} \sum_{j>i}^{N_{el}} \frac{1}{r_{ij}} && \text{the Coulomb repulsion between the electrons,} \\ \hat{V}_{ne} &= -\sum_{i=1}^{N_{el}} \sum_{\Lambda=1}^{N_{at}} \frac{Z_{\Lambda}}{r_{i\Lambda}} && \text{the Coulomb attraction between the electrons} \\ &&& \text{and the nuclei,} \\ \hat{V}_{nn} &= \sum_{\Lambda=1}^{N_{at}} \sum_{B>\Lambda}^{N_{at}} \frac{Z_{\Lambda}Z_B}{r_{\Lambda B}} && \text{the Coulomb repulsion between the nuclei.} \end{aligned}$$

Born-  
Oppenheimer  
approximation  
(1927)

The atomic nuclei are 3-4 orders of magnitude heavier than the electrons. This results in a significant difference in the speed of movement of the electrons and nuclei and in turn often implies that this movement happens at different time scales. The Born-Oppenheimer approximation (BOA) [52] separates the dynamics of the electrons from that of the nuclei. With this, the Hamiltonian and the wave function can be split into a nuclear and electronic contribution part. It should be noted, that BOA is applicable only if an adiabatic system can be assumed, i.e. there are no interactions between different electronic states.

The BOA decouples the dynamics of the nuclei and electrons in terms of the kinetic energy operators. However, in order to tackle real-world problems further approximations need to be employed. Two of the most common are:

- methods based on the Hartee-Fock (HF) theory
- density-functional theory (DFT)



### 2.2.3 The Hartree-Fock approximation and beyond

The main assumption of the Hartree-Fock (HF) [53–55] approximation is that the electronic wave function can be approximated by a single Slater determinant of single particle wave functions. The Slater determinant reads:

Hartree-Fock method (1930)

$$\Phi^{\text{HF}} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \varphi_1(\mathbf{x}_1) & \cdots & \varphi_N(\mathbf{x}_1) \\ \vdots & \cdots & \vdots \\ \varphi_1(\mathbf{x}_N) & \cdots & \varphi_N(\mathbf{x}_N) \end{vmatrix}. \quad (7)$$

The best Slater determinant, i.e. the one yielding the lowest energy can be found by applying the variational principle. According to the variational principle, the energy expectation value for any trial wave function will be an upper bound to the ground-state energy.

The HF approximation is a mean-field approximation, i.e. it assumes that the electrons move in an average potential of the remaining electrons. Thus, electronic correlations are not captured. The resulting error is referred to as the correlation error and the difference between the HF energy  $E_{\text{HF}}$  and the exact total energy  $E_{\text{tot}}$  is defined as the correlation energy  $E_{\text{corr}} = E_{\text{tot}} - E_{\text{HF}}$ . Several approaches have been proposed to evaluate this difference and are referred to as post-Hartree-Fock methods.

Møller-Plesset perturbation theory (MP) [56] improves the HF theory by adding electron correlation effects in a perturbative manner. The computationally least expensive MP method is MP2, where the correlation energy correction is truncated after the second-order term (the first order term is already included in the HF). This truncation usually leads to an overestimation of the absolute value of the correlation energy. The correction can be improved if higher-order terms are included (MP3-MP5 methods). However, this also increases the computational cost.

Møller-Plesset perturbation theory (1934)

A further method to handle the electron correlation is coupled-cluster theory [57] that utilizes an exponential 'cluster operator',  $\hat{T}$ , that is a sum of  $\hat{T}_n$  operators:

Coupled Cluster methods (1960s)

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \dots + \hat{T}_N. \quad (8)$$

The operator  $\hat{T}_n$  incorporates excitations, i.e. such determinants, where electrons are excited from an occupied to an unoccupied state, up to an order  $n$ . With this,  $\hat{T}_1$  generates all single excitations,  $\hat{T}_2$  generates double excitations and so on.

A well known coupled-cluster method is the CCSD(T) method that allows for a full treatment of the single and double excitations and a perturbative treatment of the associated triple excitations. The CCSD(T) method is considered to be the 'gold standard' method because of its extremely good accuracy when a sufficiently large basis set is used. In addition, it should be mentioned that, e.g. bond dissociation in diatomic nitrogen is challenging for coupled-cluster methods [58, 59].

The high computational cost of the CCSD(T) limits its applicability to systems with less than 50 atoms [60]. However, system sizes of few hundreds of atoms can be evaluated if the domain based local pair natural orbital (DLPNO) approximation [61, 62] is used.

## 2.2.4 Density-Functional Theory

In contrast to the approaches based on wave functions, density-functional theory (DFT) focuses on the electron density as the fundamental quantity describing the system. DFT is based on the Hohenberg-Kohn theorem [63], which states that the ground-state properties of the system can be retrieved from the electron density.

The  
Hohenberg-  
Kohn theorem  
(1964)

First, the theorem states that there is one-to-one mapping between the non-degenerate ground-state wave functions of the Hamiltonians and the densities of the particles belonging to the non-degenerate ground state. With this, any ground-state property can be formulated as a functional of the electron density.

Second, the theorem states that the density that minimizes the total energy is the ground-state density. In consequence, any non-zero trial density will result in an energy greater than or equal to the ground-state energy. With this, the ground-state energy can be then reached with the variational principle.

Kohn-Sham  
ansatz (1965)

Kohn and Sham suggested an approach [64] that allows for practical use of the Hohenberg-Kohn theorem. This approach is based on involving a set of non-interacting particles that have the same electron density,  $n$ , and the same total energy as the real set of interacting particles. For a system of non-interacting particles,  $n$  is:

$$n(\mathbf{r}) = \sum_{i=1}^{N_{el}} |\varphi_i|^2. \quad (9)$$

The Kohn-Sham ansatz for the energy functional reads:

$$E_v[n] = \underbrace{T_s[n]}_{\text{non-interacting kinetic energy}} + \underbrace{\int v(\mathbf{r})n(\mathbf{r})d\mathbf{r}}_{\text{potential energy}} + \underbrace{E^H[n]}_{\text{Coulomb interaction}} + \underbrace{E^{xc}[n]}_{\text{exchange correlation energy}}, \quad (10)$$

where the Hartree-energy term <sup>1</sup> describes the Coulomb interaction:

$$E^H[n] = \frac{1}{2} \iint \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}d\mathbf{r}', \quad (11)$$

and the kinetic energy of the non-interacting electrons reads:

$$T_s[n] = -\frac{1}{2} \sum_i \langle \varphi_i | \nabla^2 | \varphi_i \rangle. \quad (12)$$

The formulation of the Kohn-Sham ansatz is in principle exact and its solution is consistent with solving the corresponding many-particle Schrödinger equation. However, in practice the formulation of the  $E^{xc}[n]$  remains unknown. In other words, DFT is exact but, to make it applicable, density-functional approximations (DFAs) for  $E^{xc}[n]$  have to be applied.

Exchange-  
correlation  
functionals

Numerous successful DFAs have been suggested by now. Perdew proposed a classification of the different DFAs depending on the underlying physics

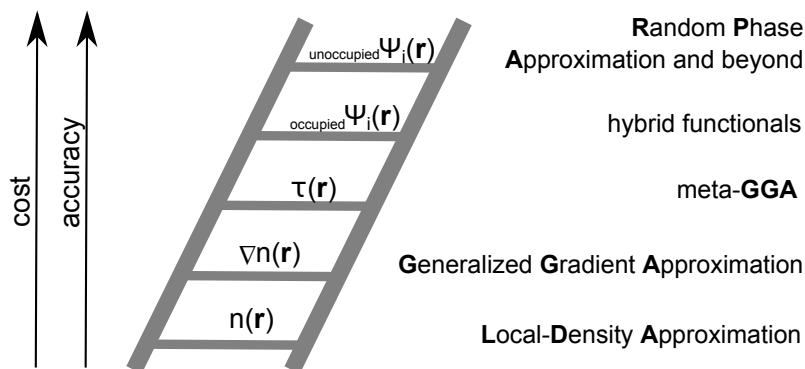


Figure 3: Jacob's Ladder of the DFT.

that arranges the approximations to the "Jacob's Ladder" of DFT [65] (Figure 3).

The simplest approximation of the  $E^{\text{xc}}[n]$  is the local-density approximation (LDA). LDA depends solely on the local electronic density and performs well for systems with nearly uniform electron density, e.g. many solids. However, LDA often fails to correctly describe systems with rapidly varying density, e.g. small molecules. LDA tends to underestimate the bond lengths and reaction barriers while the binding is being overestimated.

In order to account for rapid changes in the density, the generalized gradient approximations (GGAs) were suggested [66]. GGAs include, in addition to the local electron density, the corresponding gradients. A widely used GGA is the Perdew-Burke-Ernzerhof (PBE) [67] functional that provides a moderately accurate description of atoms, molecules, and solids [66].

Approximations including a further term,  $\tau$ , that accounts for the second derivative (Laplacian) of the density or the orbital kinetic-density, are referred to as meta-GGAs.

It needs to be noted the LDAs and GGAs suffer from self-interaction errors. In contrast, HF theory is self-interaction free. Therefore the hybrid functionals, which include a fraction of exact exchange, were suggested. A popular hybrid functional is the PBE0 functional [68], which mixes the PBE exchange energy (75%) with the HF exchange energy (25%).

The top rung of the Jacob's Ladder of DFT contains the random phase approximation (RPA), and advancements, such as the renormalized second order perturbation theory, rPT2 [69, 70] that add the unoccupied orbitals. The implementation of the RPAs combines the exact exchange with the correlation energy. A great advantage of RPA, in contrast to the semi-local and hybrid functionals, is that it accounts for long-range van der Waals interactions.

The computational cost of the introduced DFAs increases with the complexity of the implementation. It is therefore crucial to select an appropriate approximation for the physical problem one wants to solve. In this thesis, most of the problems are based on sampling the PES of molecules, requiring a vast number of optimizations of similar structures, i.e. identical in terms of the chemical composition but different 3D geometries. It is often advisable to use the GGAs at the search level to obtain reasonable geometries for a further

<sup>1</sup> Hartree atomic units are used

refinement with hybrid functionals. As already mentioned, these methods do not automatically describe the long-range van der Waals interactions and thus correction schemes are required.

### 2.2.5 Dispersion corrections

Van der Waals (vdW) interactions arise from correlated fluctuations of the electrons in materials [71]. The vdW interactions are ubiquitous, relatively weak, and always attractive forces that scale with the system size. These interactions play a major role in defining the structure and function of molecules and materials.

Several attempts to include the vdW interactions without the necessity to utilize high-level methods have been proposed: non-local functionals, e.g. Langreth-Lundqvist [72], highly-parametrized meta-GGAs, e.g. the Minnesota functionals [73], or dispersion corrections. In the following, we focus on two classes of dispersion corrections: the pairwise corrections and the many-body dispersion.

The basic idea of dispersion corrections is to formulate the total energy as:

$$E_{\text{total}} = E_{\text{DFT}} + E_{\text{vdW}}, \quad (13)$$

where:

$$E_{\text{vdW}} = - \sum_{i,j}^N \frac{C_6^{ij}}{R_{ij}^6}, \quad (14)$$

$R_{i,j}$  is the distance between the atoms  $i$  and  $j$ , and  $C_6^{ij}$  is the dispersion coefficient accounting for the interaction strength. For very short distances,  $E_{\text{vdW}}$  diverges to  $-\infty$ . Thus, in practice damping is required at short distances [74].

A number of pair-wise correction schemes have been developed, which mostly differ in the calculation of the  $C_6$  coefficients. The two most popular modern approaches are the: DFT-D<sub>3</sub> method [75] and the Tkatchenko-Scheffler (vdW<sup>TS</sup>) scheme [76]. Both schemes allow the  $C_6$  coefficients to vary with the environment of the atom.

**DFT-D<sub>3</sub>** The DFT-D<sub>3</sub> method [75] makes use of time-dependent DFT (TDDFT) with the PBE<sub>38</sub> functional for the calculation of the  $C_6$  terms. On top of that, the approximation is extended by a further term accounting for three-body interactions.

**Tkatchenko-Scheffler scheme** In the Tkatchenko-Scheffler (vdW<sup>TS</sup>) scheme [76], the  $C_6$  coefficients and the vdW radii are derived directly from the ground state electron density of the studied system. A study benchmarking the performance of the interatomic dispersion corrections on the S22 data set [41] reported a mean absolute error (MAE) of only 12 meV for the energies obtained with PBE+vdW when compared to the CCSD(T) energies [77]. Similarly, a MAE of 10 meV was obtained with BLYP+D<sub>3</sub> for the S22 data set [75].

Despite the huge success of the pairwise methods, some limitations, e.g. in describing molecular crystals or layered materials, have been observed. This is due to the fact that the pairwise dispersion corrections do not include the many-body effects. Thus, a more general approach, considering the full many-body response, has been suggested.

The methods accounting for many-body dispersion go beyond the pairwise interactions and include both the electrostatic screening and the many-body effects [78]. A recent method, the MBD@rsSCS [79] includes the range separation of the self-consistent screening and calculates the long-range contributions with a RPA based model. A single parameter is required for the range separation step. Benchmarking of the performance of the scheme proved it both efficient and accurate.

In this thesis, for calculations performed with FHI-aims [80], both correction schemes, i.e.  $\text{vdW}^{\text{TS}}$  and MBD, have been utilized. In the ORCA package [81], DFT-D3 correction is available and is used for all DFT calculations performed with this software.

Many-body  
dispersion  
scheme

### 2.2.6 Solvent models

The majority of the results presented in this thesis were obtained from calculations performed in vacuo. However, the structures of molecules are environment-dependent and solution is known to often influence the adopted structure. The correct modeling of solvent effects requires realistic models to describe the interaction of the molecule with the solvent. Two modeling approaches have been developed: explicit and implicit solvents.

**EXPLICIT SOLVENT MODELS** The explicit solvent treatment allows modeling of the direct interaction of the solute and the solvent molecules. This is achieved by considering a large number of solvent molecules and evaluating their degrees of freedom. The high computational cost of the explicit solvent models can be reduced if only a small fraction of the solvent molecules (e.g. within a certain cutoff distance from the solute) is treated explicitly. The effect of the remaining solvent can then be approximated with a simpler model or by using periodic boundary conditions. Nevertheless, applications of the explicit solvent model exclusively at the quantum-mechanical level are rare.

**IMPLICIT SOLVENT MODELS** Implicit solvent models aim at decreasing the number of degrees of freedom to consider. The solution is modelled as a polar continuum with a defined dielectric constant and the solute is placed in a medium cavity. Further, the dispersion and electrostatic interactions between the solute and the the medium need to be identified upon energy calculation. Examples of implicit solvent models are: Polarizable Continuum Model (PCM) [82] and Conductor-like Screening Model (COSMO) [83]. The COSMO model (accessed via ORCA) was utilized to model the solvent effects in Chapter 4 of this thesis. The implicit solvent models became very popular as they are stable, easy to use, and offer high quality to cost ratio. Nevertheless, several limitations have been noted, e.g. in computing binding affinities.

## 2.3 ENERGY LANDSCAPES

## 2.3.1 Characterizing the Potential-Energy Surface

When the dynamics of the electrons can be separated from that of the nuclei using the Born-Oppenheimer approximation (Section 2.2.2), the potential energy can be obtained for a defined set of nuclear positions. The potential-energy surface (PES) describes the relation between the geometry and the potential energy of a molecule [84] (Figure 4).

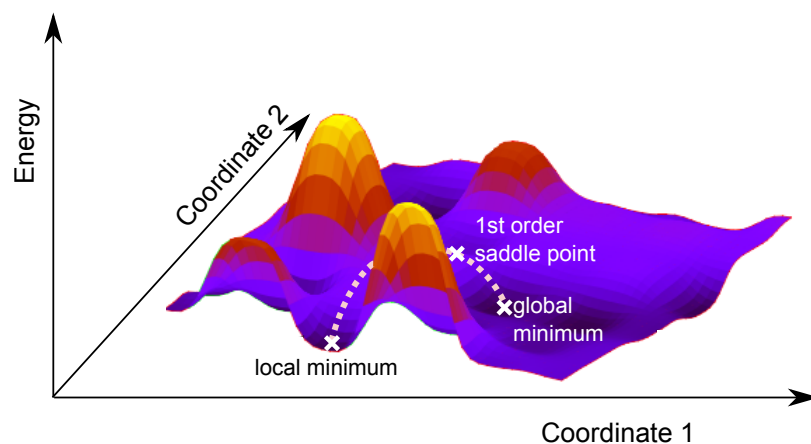


Figure 4: Schematic representation of a model PES representing the energy as a function of two coordinates.

Usually, a number of local minima with a global minimum among them exist on the PES of flexible organic molecules. If the system is in a minimum, any small displacement will increase the potential energy. Each of PES minimum corresponds to a different 3D geometry. As a consequence, a variety of low-energy conformers can be adopted by flexible molecules. Another type of stationary points located on the PES are the first-order saddle points. A first-order saddle point is a maximum in exactly one direction and a minimum in all other directions. The saddle points correspond to the transition state (TS) structures.

The PES is a multi-dimensional function of the DOFs and the calculation of the full PES is generally infeasible for real-world problems. However, a number of methods have been developed in order to investigate the PES without the need to calculate it everywhere. These methods allow sampling and characterization of the essential properties of the PES. With this, the presumably insignificant features and regions of PES are ignored.

In the following, methods for the investigation of the PES are described. *Local optimization* allows for finding the closest minimum for any given point. The goal of *global optimization* is to find the global minimum and all relevant local minima on the PES. Finally, with the *transition state search* the saddle points of the PES and the corresponding energy barriers can be identified.

### 2.3.1.1 Local optimization

The goal of local optimization is to find the closest local minimum from an arbitrary starting point. Efficient methods for local structure optimization are the gradient-based algorithms, which rely on the first (in some cases also the second) derivative of the energy [84]. Alternatively, gradient free methods, e.g. the downhill Simplex method [85], can be utilized if gradients are not available or extremely expensive to calculate.

In the steepest descent (SD) method, the optimization is performed step-wise following the direction of the negative gradient. This method is easy to implement but it might be very slowly converging in the vicinity of the minimum, leading to the 'zigzag' behavior. This happens if the gradient direction is nearly perpendicular to the direction to the sought minimum. The convergence of this method can be improved if an adaptive step size is used, but this in turn slows down the performance. On the other hand, using a small step size from the beginning is also inefficient.

The conjugate gradient (CG) method offers an improvement over the SD method. The main idea of the CG methods is to improve the selection of the step direction. In short, this is achieved by choosing a new direction based on both the current gradient and the previous directions. The number of steps needed for convergence is usually significantly lower compared to the SD method.

Nowadays though, the most popular local optimization methods are the quasi-Newton (QN) methods. The default formulation of the Newton method is:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \mathbf{J}^{-1}(\mathbf{x}_n)\mathbf{f}(\mathbf{x}_n), \quad (15)$$

which can be applied to find approximate solutions to  $\mathbf{f}(\mathbf{x}) = 0$ . Further, in order to find the stationary points of  $\mathbf{f}(\mathbf{x})$ , one needs to solve  $\mathbf{J}(\mathbf{x}) = 0$  and this can be in turn approximated with:

$$\mathbf{J}_{n+1} = \mathbf{J}_n - \mathbf{H}^{-1}(\mathbf{x}_n)\mathbf{J}(\mathbf{x}_n). \quad (16)$$

In contrast to the Newton method, which calculates the Hessian matrix  $\mathbf{H}$  directly, the QN methods replace it with an approximation. Several schemes have been proposed for updating the Hessian. The most popular is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) family of schemes [86–90]. The algorithm can be further enhanced by controlling the step size between the updates with, e.g. the trust-radius (TR) method [91].

If the potential energy can be approximated with a quadratic function, the QN methods are a sensible choice to tackle the local optimization. The performance of the QN methods can be greatly improved with a reasonable guess for the initial inverse Hessian. It was shown to be advantageous to utilize the knowledge about the structure and the connectivity of the molecule in the initial guess. Several successful models for the initialization have been suggested by Schlegel [92], Fischer and Almlöf [93], and Lindh [94].

### 2.3.1.2 Global optimization

The exploration of the a high-dimensional PES is a complex task. The solution space is huge and thus it is generally infeasible to tackle the search problem



in a deterministic way. A number of stochastic methods have been developed in order to efficiently sample the PES and to generate low-energy conformers. Some of the most popular methods are summarized in Table 1.

Table 1: Popular sampling approaches. Names of freely available programs are highlighted in boldface. Reprinted from [95].

Method	Description	Implemented, e.g., in
grid-based	based on grids of selected Cartesian or internal coordinates (e.g., grids of different torsional angle values of a molecule)	CAESAR [96], <b>Open Babel</b> [36], <b>Confab</b> [97], MacroModel [98], MOE [99]
rule/knowledge based	use known (e.g., from experiments) structural preferences of molecules	<b>ALFA</b> [100], <b>CONNECT</b> [101], CORINA and ROTATE [102, 103], <b>COSMOS</b> [104, 105], <b>OMEGA</b> [106]
population-based metaheuristic	improve candidate solutions in a guided search	<b>Balloon</b> [107], <b>Cyndi</b> [108]
distance geometry	based on a matrix with permitted distances between pairs of atoms	<b>RDKit</b> [37]
basin-hopping [109] / minima hopping [110]	based on moves across the PES combined with local relaxation	<b>ASE</b> [111], <b>GMIN</b> [112], <b>TINKER SCAN</b> [113]

In this thesis, we selected a genetic algorithm based search for sampling the PES. The genetic algorithm (GA) [114–116] is a metaheuristic optimization method and belongs to the family of evolutionary computing techniques. The concept of GA is to mimic evolution and follow the ‘survival of the fittest’ concept. The algorithm starts with a pool of random candidates for solutions. The best of the solutions are allowed to evolve while the unfavorable solutions are removed from the pool. With this, the algorithm uses the available information in order to explore promising candidates. GAs have found numerous applications in the field of 3D structure prediction, e.g: (i) conformational searches for molecules, e.g. unbranched alkanes [117] or polypeptide folding [118]; (ii) molecular design [119, 120]; (iii) protein-ligand docking [3, 4] (iv) cluster optimization [121–131]; (v) predictions of crystal structures [7, 132–134]; (vi) structure and phase diagram predictions [135]. In most applications, the fitness is a function of the total energy. In addition to that, an example of a GA, where the experimental information is included in the search process was suggested by Neiss and Schoos [136]. A meta-GA optimizing the algorithmic parameters of a GA conformer search was proposed by Brain and Addicoat [137].

Genetic algorithm based searches combined with local optimization are stable, easy to implement, and do not require any data on structural preferences of the optimized molecule. Further, GA searches can be performed without assumptions about the importance of specific degrees of freedom or combinations thereof. A clear disadvantage of GA based searches is the re-discovery of already known solutions. However, this can to some extent be prevented if only new candidates for solutions enter the local optimization.



A flexible, open-source implementation of a genetic algorithm for structure searches has been developed for this thesis. The implementation details together with benchmark results are presented and discussed in Chapter 3.

### 2.3.1.3 *Transition states search*

A path connecting two minima on the PES where any point on the path is an energy minimum in all directions perpendicular to the path is the minimum energy path (MEP). A transition state (TS) is the highest point on the MEP. Mathematically, a TS is a stationary point on the PES with all forces equal to zero and exactly one negative second order derivative.

The methods for TS searching can be loosely divided into surface-walking methods and chain-of-states methods. The surface-walking methods start at a single point at the PES and use the first or second derivative of the PES to move towards the transition state. Instead of optimizing a single initial structure, the chain-of-states methods aim at optimizing the complete MEP between the initial and the final state by representing the path as a set of points. In the following, the most popular double-ended methods: nudged elastic band, string, and growing string methods are briefly introduced.

**NUDGED ELASTIC BAND (NEB)** The NEB method [138] starts with a guess for the MEP by generating a series of images equally spaced between the initial and final minimum. The initial guess for the MEP is typically based on a linear combination of the initial and final minimum. The equal spacing is ensured by the artificially established connecting spring forces. In the optimization step, the spring forces act along the path and the potential forces act perpendicular to the path upon the energy minimization.

**STRING METHOD** The string method [139, 140] is similar to the NEB method, while the main difference is the mechanism that keeps the images equally spaced. In the string method, no spring forces are utilized and the images are relocated after each iteration step.

**GROWING STRING (GS) METHOD** In the GS method [141], two strings are generated first, starting from the two minima. Once the strings meet, the resulting string is optimized with the string method. The main advantage of this method is that it does not require an initial guess for the MEP.

It should be noted that the presented optimization methods merely generate an approximation for the MEP. As a consequence, for an exact determination of the structure associated to the TS, a further refinement and verification needs to be performed for the highest point on the resulting MEP. The optimization to the transition state can be performed with the eigenvector following (EF) method. A transition state can be verified with frequency calculations—if there is exactly one imaginary frequency, the transition state has been found.

The convergence of the mentioned method requires tens to thousands of iteration steps and is sensitive to the utilized optimizer. In this thesis, we exclusively use the string method, as implemented in the aimsChain tool [142]

that is interfaced to FHI-aims [80]. The selected optimizer is the BFGS algorithm enhanced with the TR method.

The success of the MEP optimizations is strongly dependent on the quality of the initial guess for the path. A bad guess will severely slow down the convergence or completely prevent it. The most straightforward way to generate the initial guess is to perform an interpolation of the Cartesian coordinates of the initial and final geometries. Unfortunately, this approach is inefficient for describing conformational changes within flexible organic molecules as they evolve via rotations around rotatable bonds. Internal coordinates are much better at representing rotations than Cartesian coordinates. Thus, an intuitive approach for constructing the initial paths is to perform the interpolation utilizing the relevant internal coordinates i.e. those coordinates that change most between the initial and final state. To this purpose, we developed a method generating a guess for a path between two geometries which will be published together with aimsChain.

The number of images between the initial and final geometry in the path should be chosen carefully. If the number of images is too low, the image with the highest energy might be a poor approximation for the transition state and the associated energy barrier might be highly inaccurate. Thus, a sufficient number of images should be used, ensuring a sensible result. In addition, if the conformational change is minor, e.g. if the initial and final geometry differ in the value for a peripheral torsion, the number of images can be set to a lower number compared to the situation where multiple torsions change by a significant amount. In this thesis we use 10-20 images per path.

### 2.3.2 *Towards Free-Energy Surface*

In order to capture the thermodynamics of real-world systems, the dependence on physical quantities, e.g. temperature and pressure needs to be considered. Thus, at realistic conditions, the relevant energy surface to consider is the free-energy surface (FES). FES is a low-dimensional representation in terms of collective variables that can be defined as coarse-grained order parameters [143].

Molecular dynamics (MD) simulations can be utilized to explore the FES. The MD trajectories are obtained by solving Newton's equations of motions. As the transitions between the different states are rare events, very long MD simulations need to be performed [144]. Alternatively, many independent MD simulations can be performed and the results can be combined and analyzed, e.g. with Markov-State models (MSM) [145, 146]. An attractive alternative to the classical MD simulations are enhanced sampling methods, e.g. replica-exchange MD (REMD) [147] or metadynamics [148]. These methods can be utilized to access interesting regions of conformational space.

The PES minima can be utilized as starting points for the investigation of the FES. This is based on the assumption that at least a subset of the PES minima corresponds to structures close to related FES minima. For a given molecular conformation, the harmonic approximations of the free energy can be computed.

## 2.4 PACKAGES FOR MOLECULAR SIMULATIONS

In the following, short descriptions of the software utilized for molecular calculations presented in this thesis are provided.

### 2.4.1 FHI-aims

The Fritz Haber Institute ab initio molecular simulations (FHI-aims) [80, 149] package is an all-electron, full-potential electronic structure code. FHI-aims offers a wide range of methods and can handle molecules and periodic systems. It is efficiently parallelized [150] and highly accurate [151].

FHI-aims utilizes numeric atom-centered orbitals [152] and comes with hierarchical, pre-defined basis sets for the elements 1-102 of the periodic table. FHI-aims offers three different level of species defaults: *light*, *tight*, and *really tight*. Species defaults define the utilized basis set, integration grids, and the value of the Hartree potential. The *light* settings can be used for structure searches and pre-relaxations. The *tight* settings allow for obtaining meV-level accuracy, and are well-suited for the refinement of the results obtained with the *light* settings. The *really tight* settings are designed to be used only for very specific tasks.

Apart from the defaults for the chemical elements, a range of settings, e.g. for SCF convergence, need to be considered for obtaining meaningful results. More details on the performance and on settings ensuring numerical convergence are given in the paper by Blum and colleagues [80] and in the FHI-aims manual [153].

In this thesis the *light* settings are used for the sampling of the molecular PES. If needed, the results are refined with the *tight* settings. The local optimization is set to terminate when the maximum residual force component per atom is equal to  $\max. 5 \cdot 10^{-3} \text{ eV/ \AA}$ .

### 2.4.2 ORCA

The program ORCA [81] offers a variety of quantum-chemical methods allowing for calculations with semi-empirical, DFT, and high-level wave-function methods. With ORCA, the standard Gaussian basis functions [154] are utilized. The convergence and accuracy of the ORCA results can be controlled with a wide range of parameters. For detailed descriptions please refer to the ORCA manual [155].

In this thesis, for the DFT calculations with ORCA, the cc-pVTZ Dunning's correlation consistent basis set is used. Further, *Tight* convergence criteria are selected for the convergence of the SCF calculation. The initial Hessian for the local optimization is generated with an approach developed by Fischer and Almlöf [93] and followed by BFGS optimization.

ORCA was also utilized to perform the reference MP2 calculations in Chapter 5. In order to reduce the high computational cost required by the convergence to the complete-basis-set (CBS) limit, a two-step procedure was conducted. First, a relaxation with MP2 and cc-pVTZ basis set was performed. Second, a 3-4 extrapolation [156] to the CBS limit was utilized to calculate the single-point energy of the final geometry.

### 2.4.3 *Tinker*

The TINKER molecular modeling software [113] allows for performing calculations at molecular mechanics (force field) level. A number of tasks can be performed with Tinker: geometry optimization, molecular dynamics, transition state searching, PES scanning, and more. Parameters from the following families can be utilized: Amber [44, 45], CHARMM [47], OPLS-AA [43], Merck Molecular FF [42], AMOEBA polarizable FF [49], and others.

In this thesis, the calculations with Tinker were performed in the proof of concept part of Chapter 4. The selected force field was Amber99SB [157].

## FLEXIBLE ALGORITHM FOR OPTIMIZATION OF MOLECULES

---

This chapter describes the Python package Fafoom (Flexible algorithm for optimization of molecules). In the first part details about the implementation, composition and features of Fafoom can be found. In the second part the performance of Fafoom is evaluated with two benchmark data sets. The third part of the chapter focuses on the extensibility and practical applications of Fafoom. A commented list of keywords is provided in Appendix [A.1](#). The content of this chapter is published in:

- Supady A., Blum V., and Baldauf C. First-Principles Molecular Structure Search with a Genetic Algorithm. *J. Chem. Inf. Model.*, **2015**, (55), 2338-2348. doi: 10.1021/acs.jcim.5b00243.
- Supady A. and Baldauf C. Fafoom: a Flexible Algorithm for Optimization of Molecules, in preparation.

### 3.1 MOTIVATION

The generation of low-energy conformers for a given molecule is a fundamental problem in computational chemistry [158] and an integral part of methods such as protein-ligand docking [3–5, 159] or 3D pharmacophore modeling [160]. Flexible organic molecules can adopt a variety of energetically favorable conformations that differ in chemical and physical properties. Thus, evaluating the properties of a single, arbitrarily generated conformer might lead to the wrong conclusions. The environment as well as interactions with other molecules can further influence the propensity of adopting a specific conformation [25]. Further, it has been shown that the bioactive conformation of drug-like molecules can be higher in energy than the respective global minimum [24]. In many practical applications, the PES minima are taken as starting points to explore the free-energy surface (FES). Thus, it is insufficient to focus just on the conformer representing the global minimum of the PES. Instead, a good coverage of the accessible conformational space should be targeted.

In the following sections, we present our implementation that allows for an exploration of a high-dimensional PES and yields diverse low-energy conformers. We aimed at developing an approach that utilizes electronic-structure methods for the entire search and does not rely on empirical force fields. Further, we intended the implementation to be both easy to use and flexible, i.e. facilitating creation of user-tailored searches. The implementation is open-source and can be interfaced to different electronic structure codes.

If not stated differently, I abbreviate "potential energy" with energy and "minima of the potential-energy surface" with energy minima.

## 3.2 FAFOOM: IMPLEMENTATION DETAILS

Fafoom is a Python package developed in this thesis for sampling the conformational space of organic molecules. It is implemented using Python 2.7 and employs the RDKit library [37]. RDKit is an open source collection of cheminformatics and machine-learning software distributed under the BSD license. Fafoom utilizes especially the 2D and 3D molecular operations implemented in RDKit.

Fafoom is distributed under the GNU Lesser GENERAL Public License [161] and is available from:

GITHUB: <https://github.com/adrianasupady/fafoom>

Fafoom performs the global search based on user-curated selection of degrees of freedom and conducts the local optimization on Cartesian coordinates with an external software.

## 3.2.1 Structure of the package

The structure of the Fafoom package is depicted in Figure 5.

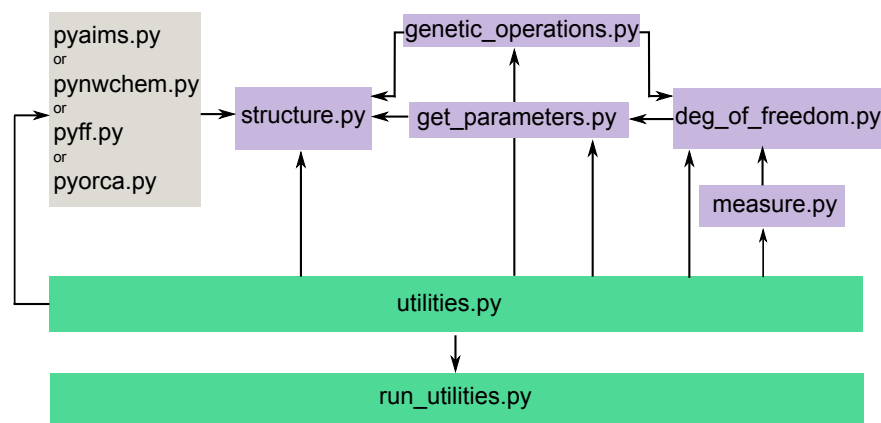


Figure 5: Fafoom: module overview. The modules responsible for handling the molecular structures, degrees of freedom and genetic operations are colored purple. Modules collecting repeatedly used support functions are colored green. The wrapper for the external simulation packaged is colored gray. The arrows depict the import direction, e.g. the `utilities.py` is imported by all other modules.

The following modules take care of the molecular structures, degrees of freedom and genetic operations:

- *structure.py* - the heart of the package. Two classes are defined by *structure.py*: (i) **MoleculeDescription** that initializes a molecule and handles its permanent attributes, i.e. the number of atoms and location of the user-defined degrees of freedom (DOFs) and (ii) **Structure** that initializes, compares and modifies the 3D structures.

- *get\_parameters.py* - a module responsible for identifying the permanent attributes of the molecule (e.g. location of the DOFs) and generating a template 3D geometry that is a starting geometry for all subsequent 3D structures.
- *deg\_of\_freedom.py* - a collection of classes that control the DOFs. Currently, the following DOFs can be handled: fully rotatable bonds, *cis/trans* bonds, and pyranose rings.
- *measure.py* - a set of functions for measuring and setting the values of the DOFs.
- *genetic\_operations.py* - a module collecting the functions required to perform a genetic algorithm based search.

The repeatedly used supporting functions are collected by the following modules:

- *utilities.py* - a collection of diverse supporting functions, i.e. vector operations and conversions between chemical formats.
- *run\_utilities.py* - a collection of functions to support a multi-step optimization process, i.e. convergence checks.

The following modules take care of the external simulation packages:

- *pyaims.py*, *pyff.py*, *pynwchem.py*, *pyorca.py* - wrappers for the supported simulation packages; responsible for launching and supervising the external optimization as well as for retrieving and storing the results.

Fafoom measures energies in eV, distances in Å, and includes an internal verification of the generated 3D structures. A sensible 3D structure needs to fulfill two requirements: (i) the shortest distance between a pair of non-bonded atoms must be longer than a defined threshold (*distance\_cutoff\_1*) and (ii) the longest distance between a pair of bonded atoms needs to be shorter than a chosen threshold (*distance\_cutoff\_2*).

A unique 3D structure is a structure that is not similar to already evaluated structures. There are two options in Fafoom that can be utilized to decide if 3D structures are similar or not:

- Cartesian RMSD - if the RMSD exceeds a certain cutoff, the structures are considered to be different.
- Degrees of freedom deviation (DOFd) - a measure of the variation of the values of the DOFs. DOfd is a list with one value (*True* or *False*) for each type of degree of freedom, e.g. for the type 'rotatable bond' the list will store 'True' if the tRMSD (see Section 2.1.2) value for the evaluated 3D structures does not exceed a certain cutoff. If the list contains at least one 'False', the two structures are considered to be different. In other words, a pair of similar structures is similar only if the values of all degrees of freedom under consideration are similar.



The 3D structures are internally encoded as strings in structure-data format (SDF). The SDF is a combination of the MDL Mol format that stores the information about the atoms, bonds, connectivity, and 3D coordinates with associated data. If needed, the strings are reformatted, e.g. for the needs of the selected external simulation package.

Fafoom keeps track of the generated conformers by maintaining a blacklist. The blacklist stores all structures that: (i) were subject to the local relaxation and (ii) resulted from converged local relaxations. The blacklist is consulted in order to evaluate the uniqueness of the newly generated structures.

### 3.2.2 Example of use: genetic algorithm based search

The main aim of the Fafoom package is to perform genetic algorithm based searches for sampling the conformational space. The operating principle of the algorithm is given by the following pseudocode (Algorithm 1):

```
# initialization
while i < popsize:
    x = random_sensible_geometry
    if x is not in the blacklist:
        blacklist.append(x)
        x = local_relaxation(x)
        blacklist.append(x)
        population.append(x)
    i+=1

# iteration
while j < iterations:
    population.sort(index=energy)
    (parent1, parent2) = population.select_candidates(2)
    (child1, child2) = sensible_crossover(parent1, parent2)
    repeat
        (child1, child2) = mutation(child1, child2)
    until child1 and child2 are sensible and are not in the blacklist
    blacklist.append(child1, child2)
    (child1, child2) = local_relaxation(child1, child2)
    blacklist.append(child1, child2)
    population.append(child1, child2)
    population.sort(index=energy)
    population.delete_high_energy_candidates(2)
    if convergence_criteria_met:
        break
    else:
        j+=1
```

**Algorithm 1** : Genetic algorithm for sampling the conformational space of molecules.

**INITIALIZATION** The algorithm starts by generating a random and sensible 3D structure directly from the SMILES code. The distance geometry method implemented in RDKit is used to obtain the initial coordinates. This initial structure acts as a template for the upcoming structures. During the initialization step, the DOFs are identified and located. In other words, the



number and location of rotatable bonds, *cis/trans* bonds, and pyranose rings are determined. For each identified DOF type, a list of random values is generated. The following values are utilized: integers in the range  $-179^\circ$  to  $180^\circ$  (rotatable bonds),  $0^\circ$  or  $180^\circ$  (*cis/trans* bonds) and integers from the range 0 to 37 (pyranose rings), each corresponding to one of the 38 sugar puckers. The collection of the 38 puckers contains: two chairs, six boats, six skew-boats, 12 half-chairs and 12 envelopes [162, 163]. If the resulting 3D geometry is sensible and unique, a local relaxation is performed.

A structure is unique if it is different from all structures stored in the blacklist. Once the local relaxation is completed, the values of the DOFs are updated, and the structure is added to the population and to the blacklist. The procedure (i.e. generating and optimizing a random structure) is repeated until the intended population size  $N$  is reached.

**ITERATION** The objective function (energy) is optimized as the population evolves over subsequent generations. An iteration begins with assigning fitness values,  $F$ , to the population members. For each individual,  $i$ , holds:

$$F_i = \frac{E_{\max} - E_i}{E_{\max} - E_{\min}} \quad (17)$$

where:

- $E_{\min(\max)}$  the lowest (highest) energy among the energies of the structures in the current population
- $E_i$  energy of structure  $i$

Consequently:  $F = 1$  for the structure with the lowest energy ('best', i.e. most stable) and  $F = 0$  for the structure with the highest energy ('worst', i.e. least stable and unlikely). In the case of populations with little variation in the energies values, i.e. if  $E_{\max} - E_{\min} < 0.001$  eV, all structures are assigned a fitness value  $F = 1$ . Once the fitness values are assigned, the genetic operations follow.

*Selection.* Three different mechanisms assigning different selection probabilities,  $p$ , to the structures are implemented:

roulette wheel	$p_i = \frac{F_i}{\sum_{n=1}^N F_i}$
reversed roulette wheel	$p_i = \frac{F_{N+1-i}}{\sum_{n=1}^N F_i}$
uniform	$p_i = \frac{1}{N}$

Based on the probability values, two distinct structures are selected and are referred to as 'parents'.

*Crossing-over.* The goal of the crossing-over is to exchange the information encoded by the parents. The crossing-over procedure is two-step: (i) the lists of values of the DOFs are combined (Figure 6) and (ii) the newly created lists of values are used to generate new structures ('children').

The structures generated by the crossing-over must be sensible. Otherwise the crossing-over is repeated until sensible geometries are generated or a maximum number of attempts is reached. If no sensible geometries can be generated, the children are exact copies of the parents.

*Mutation.* A mutation is an operation that alters selected DOF values and results in changes in the 3D structure. The mutations are performed indepen-

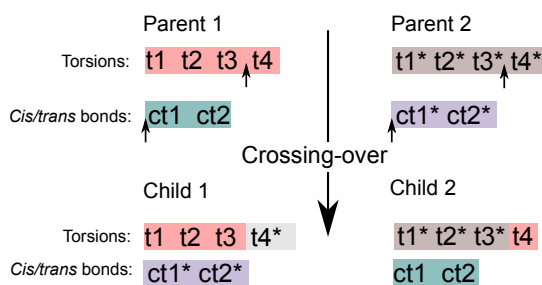


Figure 6: Crossing-over procedure.

dently for each DOF type. Both children undergo mutations. The probability for a mutation of a particular DOF type is controlled with a dedicated parameter. If a DOF type is selected for the mutation, the number of values to be changed needs to be assigned. The maximum number of allowed alterations for a specific DOF type can be controlled with a parameter. The location and actual number of alterations is then decided randomly. Mutation is only successful if the 3D structure created according to the alterations is sensible and unique. Otherwise the mutation is repeated until a sensible and unique structure is generated or a maximum number of attempts is exceeded.

*Local optimization and update.* The structures, that are created by the aforementioned genetic operations, are passed to the external software for local optimization. Once the optimization is completed, the values of the DOFs are updated and the structures are added to the blacklist. After adding the newly optimized structures to the population, the two individuals with the highest energy are eliminated.

**TERMINATION** After a minimum number of iterations, the convergence of the algorithm is evaluated. The algorithm terminates if at least one of the following criteria is met:

- the lowest energy has not changed by more than a defined threshold during a defined number of iterations
- the lowest energy reached a defined value
- the maximum number of iteration is exceeded

**PARAMETERS** The Fafoom parameters can be assigned to three groups: (i) molecular settings, (ii) run settings and (iii) GA settings. The list of all parameters (together with descriptions and defaults) can be found in the Appendix A.1.

### 3.3 FAFOOM: BENCHMARK CALCULATIONS

GA runs for two data sets: seven amino acid dipeptides and eight drug-like molecules were performed in order to evaluate the performance of Fafoom. The evaluation focuses on the following aspects: (i) the quality of the sampling in terms of the generation of diverse low-energy conformers (ii) the ability of the approach to reproduce reference data and (iii) the impact of the utilized parameters.

The DFT calculations were performed with the PBE functional. To correctly describe the energetics and structures of the investigated molecules, the dispersion corrections were always included.

### 3.3.1 Amino acid dipeptides

**REFERENCE DATA** The chemical structures of the selected amino acid dipeptides are shown in Figure 7.

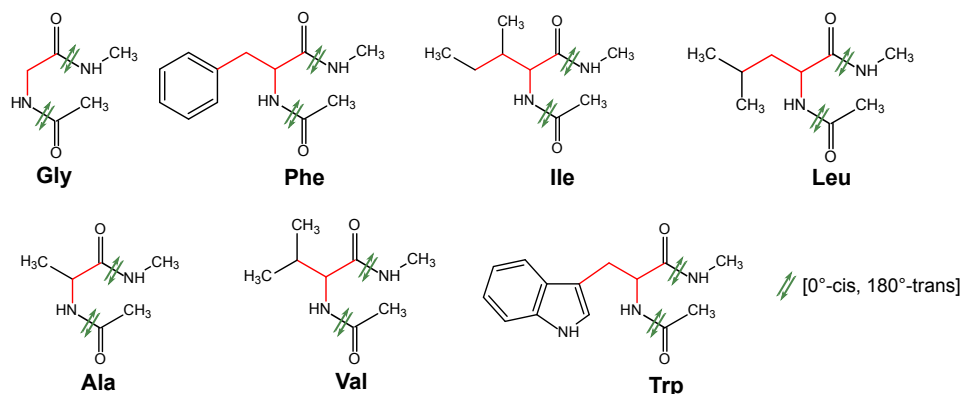


Figure 7: Chemical structures of the selected amino acid dipeptides. Rotatable bond are marked in red and the *cis/trans* bonds are marked in green.

The reference data consists of energy hierarchies of the conformers, i.e. 3D geometries together with the energy values. The reference data was extracted from a large database for amino acid dipeptide structures generated in a combined basin-hopping/multi-tempering based search [164]. The reference conformers were re-optimized at the PBE+vdW<sup>TS</sup> level with *light* computational settings, and the *tier 1* basis set [80]. For benchmarking the performance of the GA for conformer prediction, all structures with a relative energy up to 0.4 eV are considered. The number of considered conformers is summarized in Table 2.

Table 2: Reference data set: seven amino acid dipeptides.

Amino acid dipeptide	Abbr.	No. of atoms	No. of rotatable bonds + No. of <i>cis/trans</i> bonds	No. of conformers (below 0.4 eV)
Glycine	Gly	19	2+2	15 (15)
Alanine	Ala	22	2+2	28 (17)
Phenylalanine	Phe	32	4+2	64 (37)
Valine	Val	28	3+2	60 (40)
Tryptophan	Trp	36	4+2	141 (77)
Leucine	Leu	31	4+2	183 (103)
Isoleucine	Ile	31	4+2	176 (107)

**RESULTS** For each of the amino acid dipeptides 50 GA runs were performed. The GA runs are independent and different from each other, as each of them starts with an initial population of randomly created individuals. We

selected the size of the initial population equal to 5 and set the number of iterations to 10, i.e. two new children are created 10 times. This resulted in a total of  $50 \cdot (5 + 10 \cdot 2) = 1250$  geometry optimization per investigated molecule. The utilized settings are provided in Appendix A.2.

In Table 3 the probabilities for finding the global minimum, known from the reference data, in one single GA run are shown.

Table 3: Average (from 50 GA runs) probability for finding the energy global minimum in one GA run.

Molecule	Gly	Ala	Phe	Val	Trp	Leu	Ile
TDOFs	4	4	6	5	6	6	6
Probability for global minimum (/1 run)	0.82	0.79	0.53	0.60	0.22	0.20	0.10

As expected, the probability for finding the global minimum in a single GA run drops with the increasing number of TDOFs and the related increasing number of stable low-energy conformers. The probability for finding the global minimum of the glycine dipeptide is high mainly due to the low number (15) of conformers. In contrast, the larger and more flexible isoleucine dipeptide can adopt more than 100 low-energy conformers. This decreases the chance for finding the corresponding global minimum in a single GA run with only 10 iterations.

The total number of conformers identified by the GA is summarized in Figure 8. Almost all of the reference conformers are found by the GA and none of the missed structures has a relative energy lower than 0.2 eV. In addition, several new structures are predicted.

Given the results, the following question emerges: how many GA runs are enough to reach a satisfactory coverage of the conformational space, i.e. to identify 80% of reference conformers? In order to provide an answer to this question, the results of randomly chosen 5, 10, 15, 20, 25 GA runs were merged and the number of found reference structures was determined. This procedure was repeated 10,000 times and the result is shown in Figure 9.

For dipeptides with a small number of reference minima (glycine and alanine), already 5 GA runs yield a very good result. The number of GA runs needed to obtain a satisfactory coverage of the conformational space increases with the number of TDOFs, e.g. 20 GA runs are needed in the case of isoleucine.

**PARAMETER SENSITIVITY** Additional GA runs testing the impact of varying settings were performed for the isoleucine dipeptide. The effect of changing the following parameters was tested : (i) the selection mechanism (roulette wheel-default, reverse roulette wheel, random), (ii) the cut-off for blacklisting (0.2 Å-default, 0.05 Å), and (iii) the maximum number of iterations (10-default, 15, 20 and 25). For cases (i) and (ii) 100 GA runs were performed for each of the settings. In order to assess the effect of the number of iterations, 100 runs with a maximal number of iterations equal to 25 were performed and subsequently only considered up to a maximum of 15, 20, 25 maximum iterations. Additionally, 50 GA runs with a maximal number of iterations equal to 100

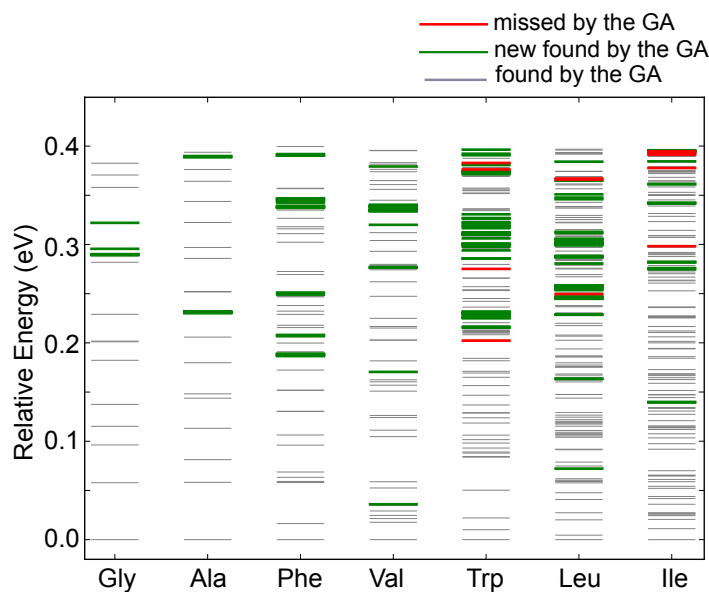


Figure 8: Difference hierarchies for the amino acid dipeptides. Red lines depict structures from the reference data set that have not been found by the GA. Green lines depict structures found by the GA that were absent in the reference data set. Gray lines depict structures from the reference data set that were found by the GA. The results from all 50 GA runs for each dipeptide were taken into account.

*FHI-aims:  
PBE+vdW<sup>TS</sup>,  
light species  
defaults*

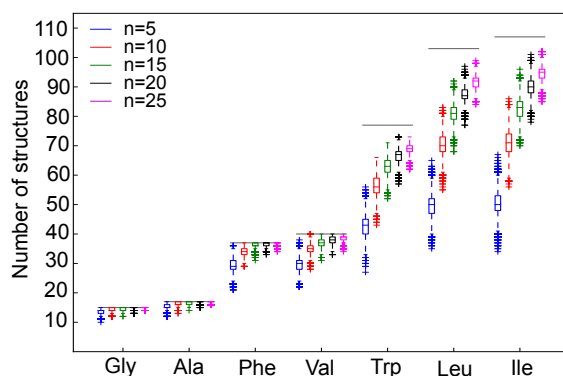


Figure 9: The number of minima found by the GA for seven amino acid dipeptides. The horizontal lines depict the total number of minima for the given molecule as predicted by Ropo *et al.* [164]. From a total of 50 GA runs, 5, 10, 15, 20, 25 GA runs were randomly selected and the discovered structures were counted. This procedure was repeated 10,000 times and the resulting distributions are summarized in box plots. The line inside the box is the median, the bottom and the top of the box are given by the lower ( $Q_{0.25}$ ) and upper ( $Q_{0.75}$ ) quartile. The length of the whisker is given by  $1.5 \cdot (Q_{0.75} - Q_{0.25})$ . Outliers (any data not included between the whiskers) are plotted as crosses.

were performed. In all mentioned cases convergence criteria were evaluated only starting from 10th iteration.

Figure 10 presents the conformational coverage for isoleucine and Table 4 summarizes the probability to find the global minimum in one run.

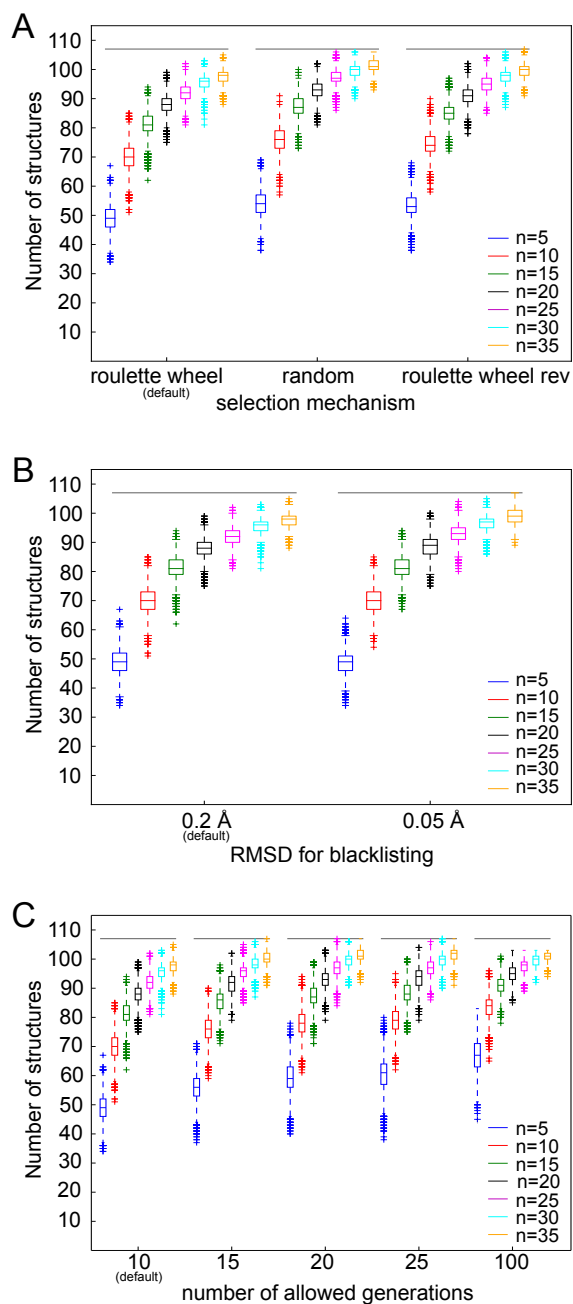


Figure 10: Conformational coverage of the GA search with different settings for Ile. The horizontal lines depict the number of reference minima [164] (107). From a total of 100 GA runs, 5, 10, 15, 20, 25, 30, 35 GA runs were randomly selected and the identified structures were counted. This procedure was repeated 10,000 times and the resulting distributions are summarized in box plots. The line inside the box is the median, the bottom and the top of the box are given by the lower ( $Q_{0.25}$ ) and upper ( $Q_{0.75}$ ) quartile. The length of the whisker is given by  $1.5 \cdot (Q_{0.75} - Q_{0.25})$ . Any data not included between the whiskers is plotted as an outlier with a cross. Conformational coverage hardly changes by using different selection mechanisms (A) or changing the blacklisting cut-off (B). (C) Increasing the number of GA iterations improves conformational coverage.

Table 4: Probability of finding the global minimum of isoleucine in one run for different setups. The default settings include roulette wheel selection mechanism, 0.2 Å cut-off for the blacklisting and maximal number of iteration equal to 10. The numbers in brackets denote the average number of iterations needed for convergence.

Setup		Probability of finding the global minimum (per run)
default		0.17
Selection mechanism	rev. roulette wheel	0.18
	random	0.13
Max. number of iterations	15 (13)	0.20
	20 (15)	0.25
	25 (16)	0.25
	100 (22)	0.46
Cut-off for blacklisting	0.05 Å	0.14

We find that none of the three selection mechanisms has a distinct impact on the quality of the conformational coverage. Similarly, no substantial change was observed for the probability of finding the global minimum. The number of found reference minima increases with an increased number of iterations. This is due to the increased number of samples from the conformational space.

### 3.3.2 Drug-like molecules

**REFERENCE DATA** We utilize the Astex Diverse Set [165], a collection of structures obtained from X-ray crystal structures from the Protein Data Bank (PDB), to construct a small set of relevant flexible organic molecules. Eight ligands were selected (Figure 11) that differ in composition, with 15 to 32 non-hydrogen atoms and 6 to 13 rotatable bonds. All X-X-O-H torsion angles we classified as rotatable bonds.

**RESULTS** For each of the selected ligands three independent GA runs with max. 30 iterations each and a population size of 10 were performed with Fafoom. The full list of utilized settings can be found in the Appendix A.2. The summary of the results is presented in Table 5. The number of found conformers is obtained after removing duplicates among all conformers obtained across all GA runs. For each of the molecules, we calculate the RMSD between the non-hydrogen atoms of each of the obtained structures and the reference ligand (Figure 12A) (hydrogens in the Astex Diverse Set set are the result of modeling and not part of the experimental result). With this, we identify the *best match*, i.e. the conformer which is most similar to the reference ligand. Furthermore, the reference ligand structures were optimized with DFT and are added to the respective plots for completeness. Figure 12B shows the overlay between the reference ligand (before the DFT optimization) and the *best match* for all molecules.

A large number of conformers spread over a wide energy window was obtained for all drug-like ligands. This satisfies our primary goal of obtaining

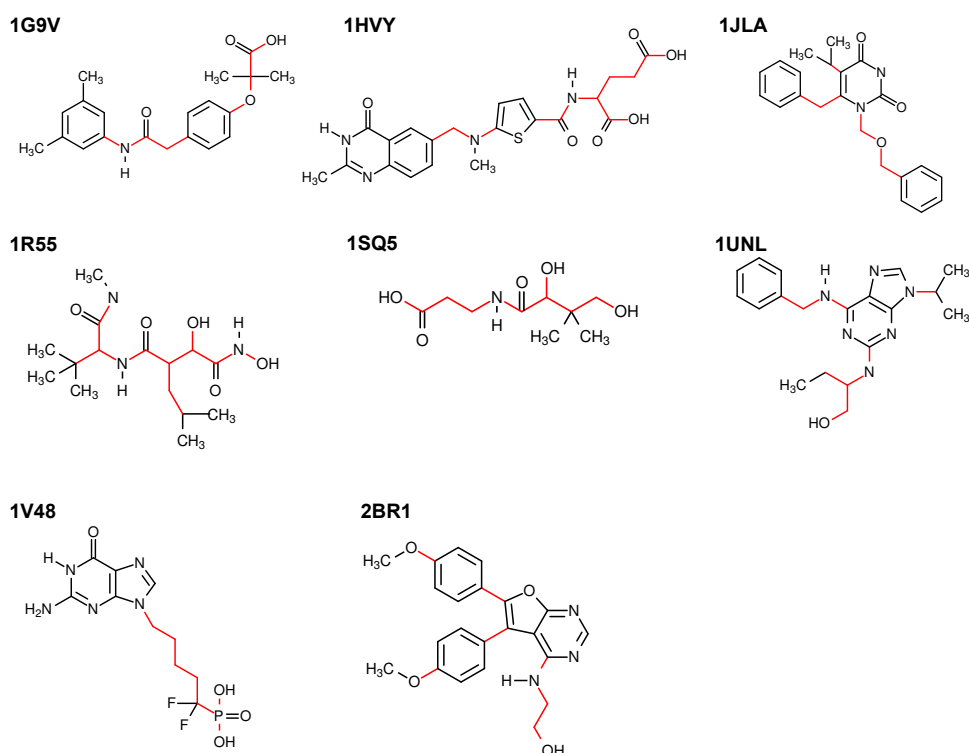


Figure 11: Chemical structures of eight selected drug-like ligands together with the PDB-IDs of the respective X-ray structures of the target proteins. Rotatable bonds are marked in red.

Table 5: Selected ligands from the Astex Diverse Set. The number of found conformers is obtained after removing duplicates among all obtained conformers. The *best match* is the conformer which is most similar to the geometry of the reference ligand.

Target protein	No. of non-hydrogen atoms	No. of rotatable bonds	No. of found conformers	RMSD ( $\text{\AA}$ ) between the ligand and the		$\Delta E$ (eV) between the GA minimum and the	
				<i>best match</i>	GA minimum	<i>best match</i>	optimized ligand
1G9V	25	8	70	1.43	1.66	0.536	0.035
1HVY	32	10	176	1.2	2.73	0.707	0.505
1JLA	27	7	41	0.56	1.44	0.339	0.268
1R55	23	13	116	0.88	1.59	0.315	0.326
1SQ5	15	10	152	0.77	2.14	0.661	0.555
1UNL	26	9	166	0.65	2.23	0.076	0.026
1V48	22	6	118	0.72	2.23	0.696	0.722
2BR1	29	8	73	0.28	0.63	0.005	0.002

a diverse set of conformers with a reliable energy hierarchy in a straightforward fashion. Moreover, in most of the cases, the RMSD between the *best match* and reference ligand is satisfactory (i.e. below  $1.0 \text{ \AA}$ ). Here we would



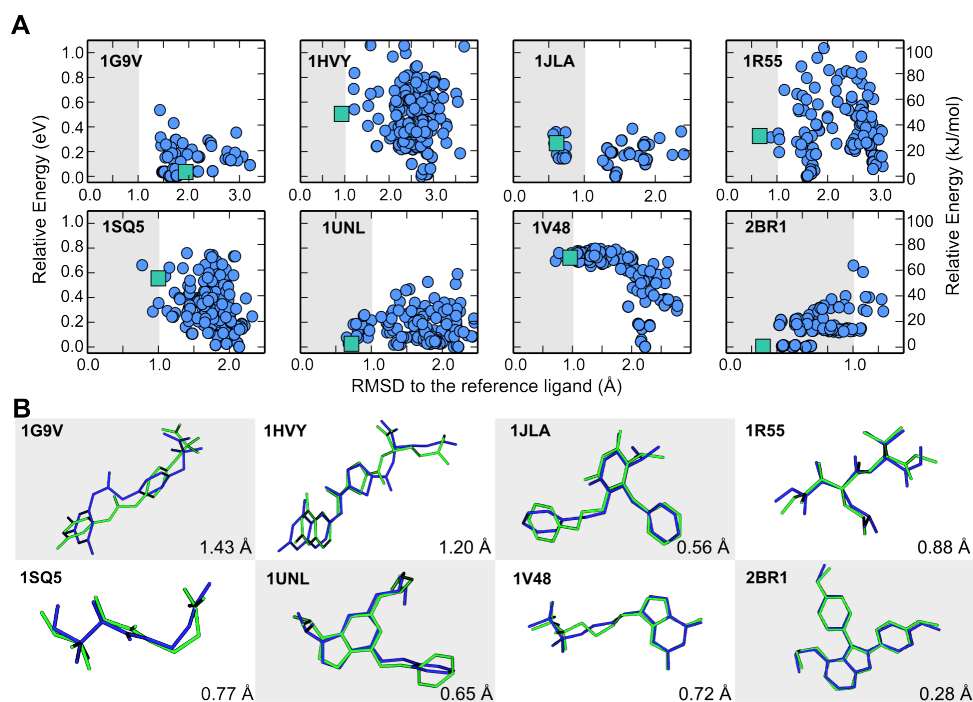


Figure 12: Evaluation of the results for a subset of the Astex Diverse Set. (A) Relative energy of all found conformers as a function of the RMSD to the reference ligand (blue circles). The green squares depict the reference ligand structures after DFT optimization. (B) An overlay between the original structure of the reference ligand (green) and the *best match* (blue) is presented together with the corresponding RMSD value.

*FHI-aims:*  
*PBE+MBD,*  
*light species*  
*defaults*

like to note that our energy evaluations are performed in the gas phase while the reference ligand is obtained from a X-ray crystal structure. The energy of the *best match* is significantly higher than the energy of the GA minimum in most of the cases. This finding supports the need for providing a broad range of conformers instead of only focusing on the global minimum of the particular energy function.

A few cases require further analysis. For two ligands, with the targets 1G9V and 1HVY, the RMSD values between the *best match* and the reference ligand structure exceed the threshold value (1.0 Å). One possible reason is that the reference ligand is not a minimum on the PES sampled by the GA. Another trivial cause might be the insufficient number of the performed GA runs. Further, we note that the optimization of the orientation of the hydroxy groups is required for obtaining a meaningful conformational ensemble.

**PARAMETER SENSITIVITY** In addition to the short exploratory GA runs, one long GA run has been performed for each of the selected drug-like ligand for comparison. In order to obtain comparable results by means of the number of performed DFT optimization, the following parameters have been adjusted: `max_iter=80`, `iter_limit_conv=70`. We compare the results in terms of: (i) energy of the most stable structure, (ii) similarity to reference ligand structure and (iii) number of found conformers. Detailed data about the results of the single long runs are given in Table 6.

Table 6: Comparison of the results obtained from one long GA run (max. 80 iterations) and three short GA runs (each max. 30 iterations).  $\Delta E$  is the difference between the most stable structures found in the compared setups.

Target protein	Number of found conformers		RMSD (Å) between the ligand and the <i>best match</i>		$\Delta E$ (eV)
	1 x max. 80 iterations	3 x max. 30 iterations	1 x max. 80 iterations	3 x max. 30 iterations	
1G9V	45	70	0.57	1.43	0.0
1HVY	146	176	1.07	1.2	0.211
1JLA	40	41	0.62	0.56	-0.005
1R55	85	116	0.69	0.88	0.081
1SQ5	99	152	0.64	0.77	0.031
1UNL	93	166	0.8	0.65	0.003
1V48	115	118	0.81	0.72	0.054
2BR1	47	73	0.28	0.28	-0.005

In terms of finding the *best match*, the single long GA run performs better than the three short GA runs together for some of the molecules. On the other hand, the number of found structures is significantly higher if three short GA runs are performed instead of a single long run for most of the molecules.

The results of the exploratory structure searches for the eight drug-like ligands suggest that performing one long GA run instead of three short GA runs may increase the chance for finding the global minimum and simultaneously decrease the number of identified unique conformers.

### 3.4 FAFOOM: AND BEYOND

Initially, Fafoom was intended to be an implementation of a search with genetic algorithm for organic molecules to be used together with FHI-aims. Over time, Fafoom has developed into a flexible library for a multitude of tasks. Below, several aspects of the flexibility and applicability of Fafoom are outlined, accompanied by examples.

#### 3.4.1 Applications

Fafoom is a library and can be imported and used in Python scripts and applications. A very basic example for application of Fafoom in the daily work of computational chemists is, e.g. assigning values to the torsions of the molecule and checking if clashes in the 3D geometry occur. Fafoom can also be used to perform searches in the conformational space of the molecule with more techniques than just the genetic algorithm. Here we present a short study on the performance of three different search methods, among which two are performed with Fafoom and the third is performed with a different software and is supported by Fafoom.

Mycophenolic acid (target protein: 1MEH) is a very flexible molecule with 43 atoms, 8 rotatable bonds and 1 *cis/trans* bond (Figure 13).

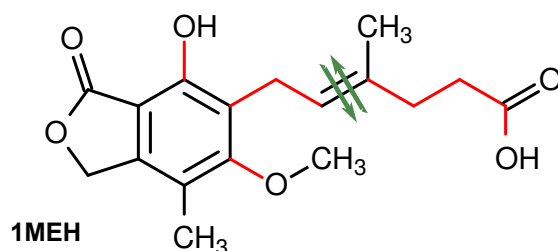


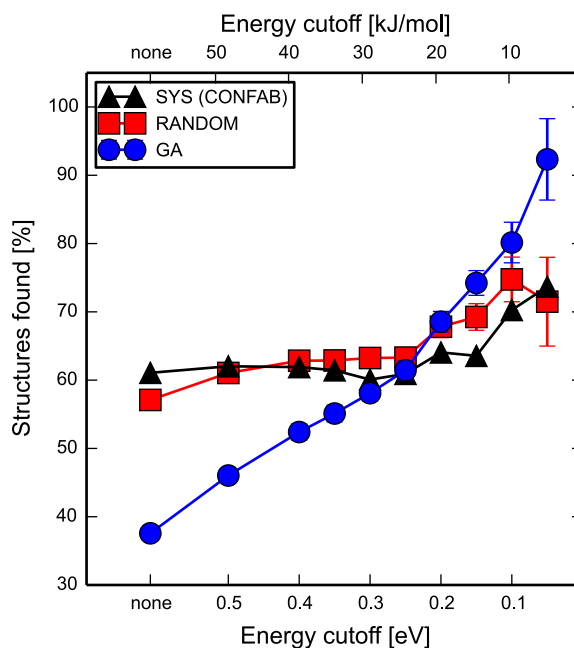
Figure 13: The chemical structure of the mycophenolic acid. Rotatable bonds are marked in red and the *cis/trans* bond is marked with double arrows.

The applicability and performance of three search techniques (1-3 below) in combination with first-principles methods were tested on the example of mycophenolic acid:

1. Genetic algorithm - 50 independent GA runs with max. 30 iterations each and a population size of 10 were performed with Fafoom. A total of 3208 structures were generated.
2. Random search - 3200 random and clash-free structures were generated with Fafoom and further relaxed with DFT.
3. Systematic search with Confab - first, 293 conformers were generated with Confab [97] (assessed via Open Babel, used settings: **RMSD cutoff** = 0.65 and **Energy cutoff** = 15 kcal/mol). In order to account for two different values for the *cis/trans* bond and the X-X-O-H torsions ( $0^\circ$  and  $180^\circ$ ), eight starting structures for each of the Confab conformers were generated with Fafoom. This procedure yields 2344 structures overall. After removing geometries with clashes, **2094** structures were subjected to DFT optimization.

Finally, all DFT optimized structures were merged into a common pool and the duplicates were removed (criteria:  $\Delta E < 10\text{meV}$  and  $\text{tRMSD} < 0.1\pi$ ). In total, 1436 unique structures were found. For the purposes of the following analysis, we refer to these 1436 unique structures as 'all', i.e. 100%. However, it should be noted that there are probably more stable structures in the conformational space of the mycophenolic acid. For the analysis, the three following subsets are defined: (i) 'GA' is a random selection of 25 GA runs (approx. 1600 structures); (ii) 'SYS (CONFAB)' is the set of all 2094 structures generated in the systematic search; and (iii) 'RANDOM' is a random selection of 1600 structures generated in the random search. For the performance evaluation we count the number of reference structures found by the respective search technique. This procedure was repeated 1000 times for each of the energy cutoffs. The results are shown in Figure 14.

All of the search techniques found the global minimum several times. When no energy cutoff is applied, none of the methods was able to find all local minima in the conformational space, i.e. more calculations would be needed. With a decreasing energy cutoff, an improved coverage of the conformational space can be observed. The fact that the GA is a global optimization techniques is clearly visible, as it performs better in the low-energy ( $< 0.2\text{ eV}$ ) region, whereas the random and systematic search perform uniformly, but not perfectly, independent of the energy cutoff used for the evaluation.



*FHI-aims:  
PBE+MBD,  
light species  
defaults*

Figure 14: Share of the number of reference structures found by three search techniques: GA (blue circles), random search (red squares) and systematic search with Confab (black triangles) as a function of the applied energy cutoff.

### 3.4.2 Free choice of the external simulation package

Fafoom can be used with external software for the local optimization. The communication between Fafoom and the external software is limited to: (i) passing the 3D structure to the external software along with the software-specific parameters and (ii) receiving the optimized 3D structure together with the energy value. Given this, all settings for the local optimization can be chosen by the user. On the other hand, Fafoom cannot control the accuracy nor the meaningfulness of the obtained result.

Fafoom can be extended to work with other external software by writing a dedicated wrapper (guidelines are provided with the Fafoom manual). Fafoom is currently interfaced to four molecular simulations packages: FHI-aims [80], RDKit [37, 166], NWChem [167] and ORCA [81]. The wrapper for ORCA was developed together with Philipp Traber (FSU Jena).

### 3.4.3 Extensibility of the kind of optimized degrees of freedom

One of few fixed assumptions in Fafoom is that the objective function is the total energy of the conformer. In contrast, the number and type of the considered degrees of freedom is user-defined. The DOFs currently supported by Fafoom are depicted in Figure 15. As already discussed in the opening: in the case of organic molecules, dihedral angles are descriptive enough and are nearly always a good choice for the DOFs to be optimized. By default, all integers in the range  $-179^\circ$  to  $180^\circ$  can be adopted by the dihedral angles. However, if some of the dihedral angles are known to have clearly defined preferences, e.g. the peptide bond, occurring in either the *cis* or *trans* isomer,

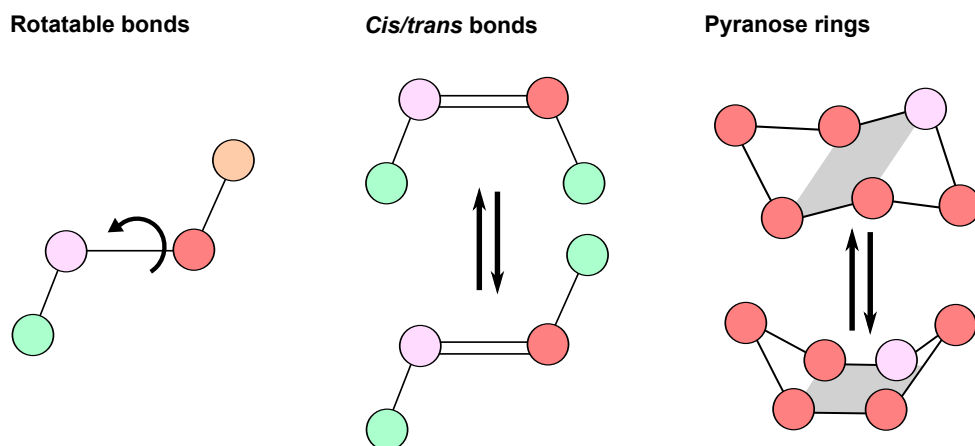


Figure 15: Types of degrees of freedom (DOFs) supported by Fafoom. The depiction of the 'pyranose ring' DOF is simplified: a total of 38 ring conformations can be adopted.

restrictions can be made. By restricting the search space, redundant sampling of highly unlikely conformations can be prevented. Fafoom allows for handling selected dihedral angles in the *cis/trans* mode, i.e. the adopted value needs to be either  $0^\circ$  or  $180^\circ$ . These values can be easily changed and more values can be added if needed. This introduction of additional information from chemical knowledge can significantly increase the performance of the algorithm.

A further complication in the world of flexible organic molecule are cyclic groups. The classification of the cyclic groups depends on the number and type of the elements in the ring and on the saturation status. Many of the unsaturated groups, e.g. the phenyl group, pyridine or thiophene, are planar and have only one favorable conformation. In contrast, the saturated groups have considerably more conformational freedom, e.g. pyranose can adopt 38 distinct conformations [162, 163] with the 'chair' conformation being the most stable one. Although the 'chair' conformation has the highest chance to be adopted, the relative energies of the competitive conformations are still low enough to occur. This presents a challenge for the sampling strategy and the software. Fafoom can generate different conformations for the pyranose rings and thus optimize the pyranose ring DOFs. The method generating different pyranose ring conformations was developed together with Mateusz Marianski (FHI Berlin).

The ability of Fafoom to sample different ring conformations is illustrated by the example of  $\alpha$ -D-glucopyranose known as  $\alpha$ -glucose. A search composed of 18 independent GA runs yielded 429 distinct  $\alpha$ -glucose structures at the PBE+MBD level. The type of the pucker was determined for each of the structures before and after the DFT relaxation. The following pucker types can be adopted: two chairs ('C'), six boats ('B'), six skew-boats ('S'), 12 half-chairs ('H') and 12 envelopes ('E'). The characterization of the different pucker types can be found in the literature [162, 163]. Figure 16 depicts the results of the assignment.

Fafoom samples the different pucker types well (Figure 16, left): 37 out of 38 different puckers were generated at least once. In the pool of structures after the DFT relaxation, only 13 different types of pucker occur: chairs, boats

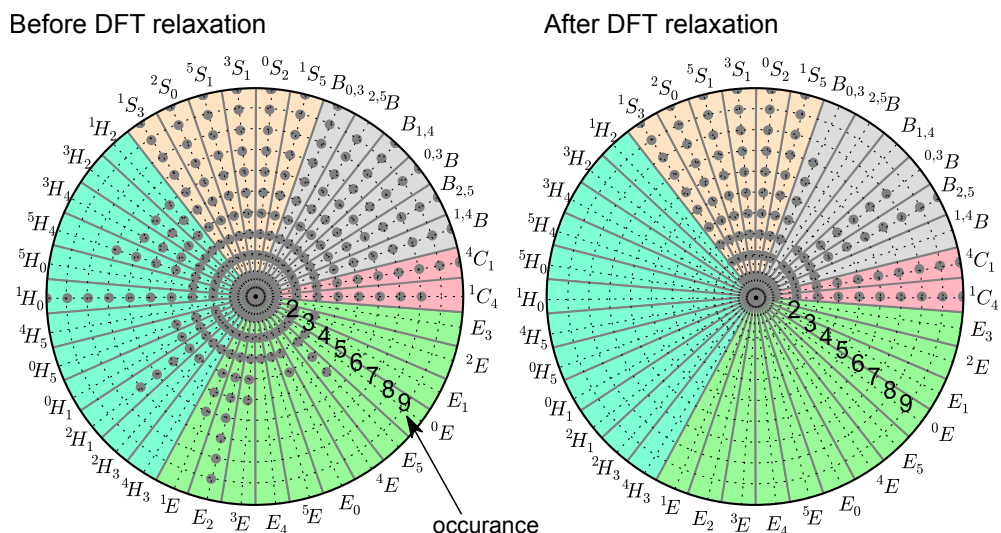


Figure 16: Sampling of the conformations space of  $\alpha$ -glucose. There are 38 pucker types: two chairs (C), six boats (B) six skew-boats (S), 12 half-chairs (H) and 12 envelopes (E). The occurrence of the different pucker types is depicted: before (left) and after DFT relaxation (right). The maximum plotted occurrence is 10 and should be read as  $\geq 10$ .

and skew-boats. There are no envelopes and half-chairs among the locally optimized  $\alpha$ -glucose structures. This is consistent with the results of a study of Mayes and colleagues [168], who performed a systematic search and constructed a library of low-energy minima and transition states for  $\alpha$ -glucose and four further pyranose sugars.

The implementation of the degree of freedom handling the pyranose rings can be easily adapted to handle similar flexible ring with six components. Further types of degrees of freedom can be added to Fafoom, either by modifying and adapting the existing ones or adding new types (guidelines are provided in the Fafoom manual).

#### 3.4.4 Parallelization

There are two ways to utilize parallel computational resources when performing searches with Fafoom. One option is that multiple runs can be started in parallel and the blacklist can be shared between different and subsequent runs. The second option is to run calculations across distributed nodes if the molecular simulation package is efficiently parallelized (e.g. in FHI-aims [150]). Both options can also be used simultaneously.



## REDUCED MOLECULAR POTENTIAL-ENERGY LANDSCAPES

---

This chapter proposes a multi-step framework that allows for constructing a reduced potential-energy surface consisting of minima and selected transitions that can be visualized as an energy barrier tree. In Section 4.1 the motivation for the work is presented and the theoretical foundations of the framework are reviewed. Section 4.2 gives the technical details of the procedure. A proof of concept at a simplified level of theory is introduced and examined in section 4.3. Finally, section 4.4 presents and discusses the results of the framework for a real system at the first-principles level.

The contents of the chapter are published in:

- Supady A. and Baldauf C. Reduced potential-energy landscapes. 2016, in preparation.

### 4.1 MOTIVATION

The preceding chapter focused on the relevance of sampling the PES for minima and introduced Fafoom, a tool facilitating implementation of user-tailored sampling algorithms. Although valuable insights into the PES can be gained if the relevant PES minima are known, it is undoubtedly attractive to also know the relevant transition states on the PES. Unfortunately, finding transition states is not as straightforward as sampling for PES minima.

The approaches for identifying transition states can be loosely divided into two types: (i) dynamics-based and (ii) energy-landscape based [12]. The first approach utilizes data from MD simulations by collecting information about transition states from several MD trajectories. This approach also offers the additional possibility to construct Markov State Models (MSM). The construction of a MSM involves: (i) a discretization of the conformational space and (ii) determination of the transition probabilities between these discrete states [144]. The MSMs can give insight into the systems kinetics and simplify the interpretation of raw time-series data [12, 169, 170]. One drawback of the MD-based approach is that high-energy transition states are unlikely to be observed with unbiased MD techniques. For this reason, enhanced-sampling methods, e.g. metadynamics [148, 171], are utilized to overcome high barriers and to reconstruct the underlying free-energy surface. Metadynamics adds a bias potential that acts on selected degrees of freedom, that are referred to as collective variables (CVs). Despite the wide application of the metadynamics approach, it should be noted that a nonoptimal choice of CVs can lead to a prohibitively large computational time needed for convergence [143].

In the energy-landscape approach, the transition states are obtained from saddle points along the minimum energy paths connecting the minima [12]. This is considerably more systematic than the MD-based approach as allows exploring high-energy transition states. On the other hand, the number of potential saddle points depends quadratically on the number of minima, and

computing saddle points for larger molecules is computationally demanding. Therefore, the high computational cost limits the applicability of this method to small systems, i.e. systems with less than few dozen atoms and not more than a few degrees of freedom.

In this chapter, we propose how the energy-landscape approach can be adapted to be used for more complex systems. Our main idea is to limit the number of minima pairs for which the saddle points are computed. Given a reasonable choice of the subset of minima pairs, a number of relevant transition states can be identified.

Selecting the right subset of minima pairs is critical for the success of the approach. The following considerations are taken as a basis for our strategy. First, it should be noted that only a subset of the minima pairs are connected via *direct transitions*, i.e. with only a single saddle point between them (Figure 17, left panel). The majority of the minima pairs are not connected via *direct transitions* but via *combined transition paths*. *Combined transition paths* are constructed from  $k$  ( $k \geq 2$ ) *direct transitions* (Figure 17, middle and right panels).

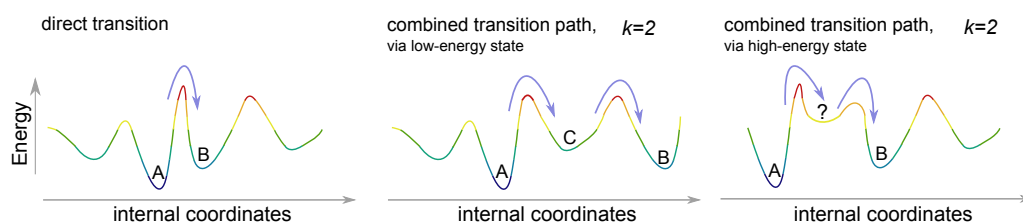


Figure 17: Different transition types between two low-energy conformers A and B. The left panel depicts the *direct transition* with one transition state between. The middle and the right panels depict *combined transition paths* constructed from two *direct transitions* ( $k = 2$ ) via a low/high-energy state.

Based on the fact that *combined transition paths* can be split into *direct transitions*, it is sufficient to compute saddle points for minima pairs connected via *direct transitions*. Unfortunately, it is not known beforehand which of the minima pairs are connected via *direct transitions* and a guess needs to be made. Prior to this, the following should be taken into consideration: (i) conformers of flexible organic molecules interconvert by rotations around single bonds and (ii) such rotations alter the torsion angles of the conformers. Thus, it is advisable to describe the conformers utilizing the torsional degrees of freedom. Besides this, two conformers are likely to interconvert if the needed rotations: (i) involve only few bonds and (ii) the corresponding rotation angles are small. Based on the presented considerations, we formulate the following hypothesis for the selection of the minima pairs: if two minima are similar in terms of torsional degrees of freedom, there is a high chance for a *direct transition*.

Once the subset of the minima pairs has been selected, the corresponding saddle points and the associated energy barriers can be identified (see Section 2.3.1.3).

Finally, the knowledge of the PES minima can be combined with the identified energy barriers and presented in the form of an energy barrier tree (Figure 18). A barrier tree is a convenient representation of the energy sur-



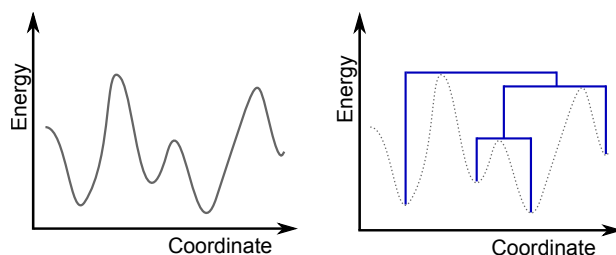


Figure 18: Representations of a model energy surface. The energy profile (left) and the corresponding barrier tree (right).

face [14, 172–174]. An alternative representation of the high-dimensional landscapes are disconnectivity graphs [13, 175].

#### 4.2 DESCRIPTION OF THE METHOD

Figure 19 gives an overview of the method that we developed for constructing energy barrier trees for flexible organic molecules. Further relevant implementation details are given in the text.

For a better understanding, clarifying figures are added to the text. The figures refer to a common abstract example and are not based on any specific data.

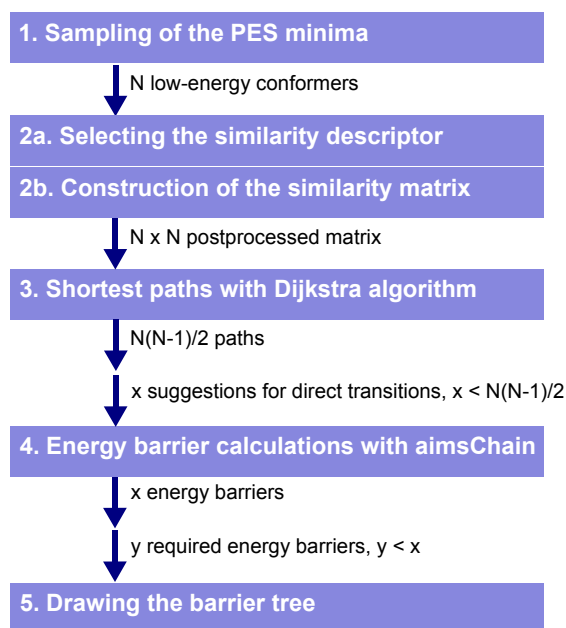
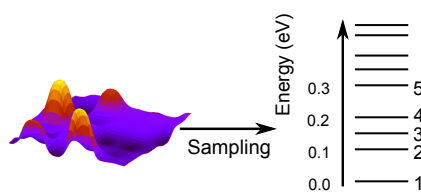


Figure 19: Scheme representation of the workflow

#### 1. SAMPLING OF THE PES MINIMA

At first, the sampling of the PES minima needs to be performed in order to obtain a representative set of conformers containing  $N$  low-energy conformers left after sorting out high-energy conformers.



**2A. SELECTING THE SIMILARITY DESCRIPTOR** The most critical step of the framework is the decision about the subset of minima pairs for which the transition state optimization will be performed. As already introduced, the molecular similarity is an indicator for a high chance of a direct transition between two minima. Here we use the tRMSD (see Section 2.1.2) to determine the molecular similarity.

**2B. CONSTRUCTION OF THE SIMILARITY MATRIX** The similarity matrix  $S$  is a symmetric  $N \times N$  matrix. For  $i, j \in \mathbb{N}, i \leq N, j \leq N$ , the corresponding matrix element is calculated as follows:

$$A_{i,j} = A_{j,i} = \text{tRMSD} = \sqrt{\frac{1}{K} \sum_{k=1}^K d_k^2},$$

	1	2	3	4	5
1	0.0	0.19	0.68	0.09	0.22
2		0.0	0.33	0.25	0.75
3			0.0	0.61	0.27
4				0.0	0.11
5					0.0

(18)

where:

$d_k$  is the difference in the torsion value, and

$K$  is the total number of torsions in the molecule.

All diagonal elements of the similarity matrix  $S$  are equal zero. Values exceeding a certain similarity threshold are penalized in the preprocessing step prior to the determination of the shortest paths between the conformers. The unit of the tRMSD is  $\pi$ . To give an example, a tRMSD value equal  $0.1\pi$  corresponds to: (i) a change of  $18^\circ$  per each of the  $K$  torsions or (ii) a  $18^\circ \cdot \sqrt{K}$  change of a single dihedral angle.

**3. SHORTEST PATHS** In order to ensure that all selected minima are connected, the connections between all  $N(N-1)/2$  conformer pairs need to be predicted. The problem to solve is to find suggestions for the shortest connections between all minima pairs based on the modified similarity matrix. This can be formulated as a special case of the shortest path problem, known from graph theory.

The shortest path problem is to find a path between two nodes in a graph so that the sum of the weights of the connecting edges is as small as possible. A widely used algorithm for solving the shortest path problem is the Dijkstra algorithm [176]. For each of the minima pairs the shortest connection is determined with the Dijkstra algorithm resulting in a number of

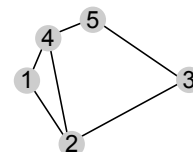
10 resulting shortest paths:

$1 \rightarrow 2$   
 $1 \rightarrow 4 \rightarrow 5 \rightarrow 3$   
 $1 \rightarrow 4$   
 $1 \rightarrow 4 \rightarrow 5$   
 $2 \rightarrow 3$   
 $2 \rightarrow 4$   
 $2 \rightarrow 4 \rightarrow 5$   
 $3 \rightarrow 5 \rightarrow 4$   
 $3 \rightarrow 5$   
 $4 \rightarrow 5$

decompose → remove duplicates

Suggested direct transitions:

$1 \rightarrow 2$   
 $1 \rightarrow 4$   
 $2 \rightarrow 3$   
 $2 \rightarrow 4$   
 $3 \rightarrow 5$   
 $4 \rightarrow 5$



*direct* and *combined transition paths*. The *combined transition paths* are decomposed into lists of *direct transitions*. Finally, all *direct transitions* are combined into a common list while the duplicates are being removed.

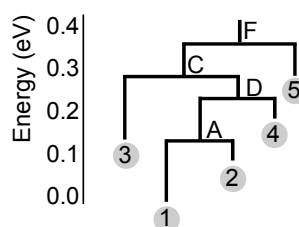
4. ENERGY BARRIER CALCULATIONS For each suggested direct transition, the transition state and the associated height of the energy barrier is determined. The calculations are performed with the aimsChain routine. In order to improve the convergence of the calculations, an initial guess for path connecting the minima is made (see Section 2.3.1.3).

#### 5. DRAWING THE BARRIER TREE

Once the energy barriers are identified, the barrier tree can be constructed. To this end, the energy barriers are sorted in ascending order. Starting from the lowest barrier, the corresponding minima are successively connected. This procedure is continued until all minima of interest are connected. Once this is achieved, the set of required barriers is created. The barriers that are not required to connect the minima are discarded. Prior to the plotting, the heights of the barriers are made relative to the energy of the global minimum. Finally, the barrier is plotted starting from the lowest barrier.

		height	height rel. to global min
barrier A:	1 → 2	0.15 eV	0.15 eV
barrier B:	1 → 4	0.22 eV	0.22 eV
barrier C:	2 → 3	0.20 eV	0.30 eV
barrier D:	2 → 4	0.16 eV	0.26 eV
barrier E:	3 → 5	0.25 eV	0.40 eV
barrier F:	4 → 5	0.18 eV	0.38 eV

Barriers sorted by height: A, D, F, C, B, E  
 Required barriers: A, D, F, C  
 Redundant barriers: B, E  
 Required barriers sorted by rel. height: A, D, C, F



### 4.3 THE EXAMPLE OF ACALA<sub>3</sub>NME

In this subsection we present a 'proof of concept' of the introduced methodology. The system studied here is the alanine tetrapeptide AcAla<sub>3</sub>NMe (Figure 20).

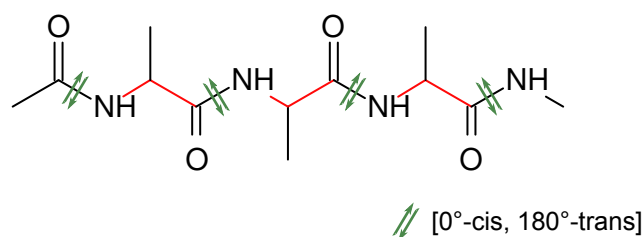


Figure 20: Structure of AcAla<sub>3</sub>NMe. The rotatable bonds are marked in red and the *cis/trans* bonds are marked in green.

The AcAla<sub>3</sub>NMe molecule was selected for the following features: (i) it is easy to compute as it has only 20 heavy atoms; (ii) it has several relevant DOFs: six rotatable bonds and four *cis/trans* peptide bonds and (iii) due its flexibility, it is likely that there will be a number of diverse low-energy conformers. The aim of the study is to compare the information content of a barrier tree of the PES with a free-energy MSM.

## 4.3.1 Reference data for comparison

The reference data for the dynamics of AcAla<sub>3</sub>NMe in form of a free-energy Markov State Model (MSM) were provided by Mateusz Marianski. The model was created with the following framework:

1. 100 molecular dynamics (MD) trajectories of the length 100 ns each at 300 K were computed with Gromacs 4.6.3 [177] utilizing the Amber99sb force field [157].
2. The resulting structures were clustered based on the tRMSD similarity measure.
3. The obtained representative 597 structures were divided into 8 free-energy basins with msmbuilder 2.8.3 [178].
4. After final reduction, the Markov State Model containing 5 macrostates was constructed (Figure 21).

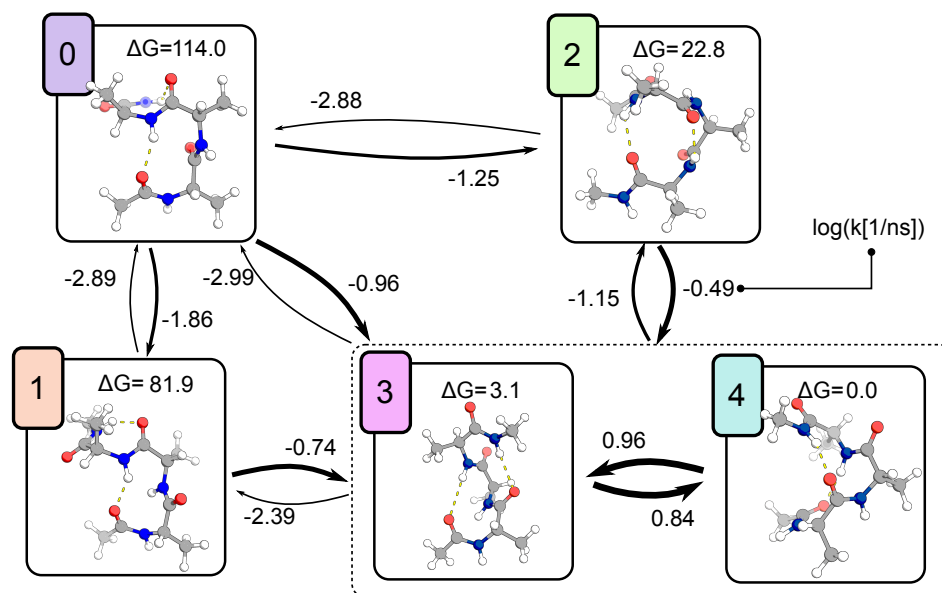


Figure 21: The Markov State Model for the AcAla<sub>3</sub>NMe provided by Mateusz Marianski (FHI Berlin). The free energy unit is meV. The numbers next to the arrows denote the logarithm of the corresponding transition rate  $k$  (1/ns). The free energy was computed at 300 K utilizing the Amber99sb force field.

The presented MSM model gives the information about the most significant features of the FES of AcAla<sub>3</sub>NMe.

## 4.3.2 Sampling of the potential-energy surface

In order to allow for a comparison between the free-energy MSM and a barrier tree derived exclusively from the PES, the sampling of the PES needs to be performed first. According to our definition, the studied system has 10 relevant DOFs. The peptide bond has a strong preference to adopt the *trans* configuration. Thus, we restricted the freedom of the peptide bonds and

performed exclusively constrained searches that do not permit the peptide bond to adopt the *cis* configuration. All energy evaluations were performed with the Amber99sb force field [157] as provided with the Tinker distribution. Two techniques were employed to sample the conformational space of AcAla<sub>3</sub>NMe:

- a GA based search with Fafoom interfaced to Tinker (30 independent runs)
- a basin hopping search with the Tinker SCAN method [113]

The GA search resulted in 750 structures (with duplicates among them). The Tinker search resulted in 112 distinct structures. These two pools of structures were merged to a common pool of structures for clustering. A combined geometry/energy criterion was applied for the clustering. Based on the Cartesian RMSD values, a hierarchical clustering with a threshold of 0.1 was performed that resulted in 125 clusters. For each resulting cluster, the energy range of the structures within the cluster was calculated. None of the energy ranges exceeded a threshold of 10 meV. 81 from the 125 clusters contain structures from both searches. Further, 30 clusters contain exclusively structures from the Tinker search and 14 contain structures found only by the GA search. The energetically most favorable structure from each cluster was then stored and the remaining structures were discarded. With this, 125 unique conformers mark the starting points at the PES for the further investigation.

#### 4.3.3 Towards the network of states

As motivated in 4.1 and described in 4.2, we utilize the tRMSD as the descriptor for the chance of a direct transition between two conformers. The tRMSD was calculated for all pairs of structures and all 10 DOFs, the six rotatable bonds and the four peptide bonds, were considered. The peptide bonds were considered in order to account for the small deviations from the targeted value 180° resulting from our restriction that the peptide bonds can only adopt the *trans* configuration. With this, the resulting similarity matrix (125 × 125) was constructed. Values exceeding tRMSD = 0.25 π were penalized to effectively prevent transitions. We selected the 20 structures with the lowest energies from the introduced set of 125 structures for constructing the network. Then, based on the similarity matrix, suggestions for  $\frac{20 \cdot 19}{2} = 190$  paths were generated. The paths were generated independently of each other. Each of the paths was split into a number of single connections between pairs of structures. The single connections resulting from splitting all paths were combined to a common list. After removing the duplicates, 81 connections remained in the list. Thus, a total of 81 connections is used to connect the 20 lowest-energy structures and can be visualized in a form of a network (Figure 22).

From the network, one can read, which paths were suggested in order to connect the 20 lowest-energy structures (nodes 1-20). Several of the paths are *direct transitions* (e.g. 1 ⇒ 2 or 11 ⇒ 13). Further, *combined transition paths* via one low-energy state (e.g. 12 ⇒ 9 ⇒ 17) or via one high-energy state (e.g. 13 ⇒ 58 ⇒ 20) are suggested. However, most of the suggested paths

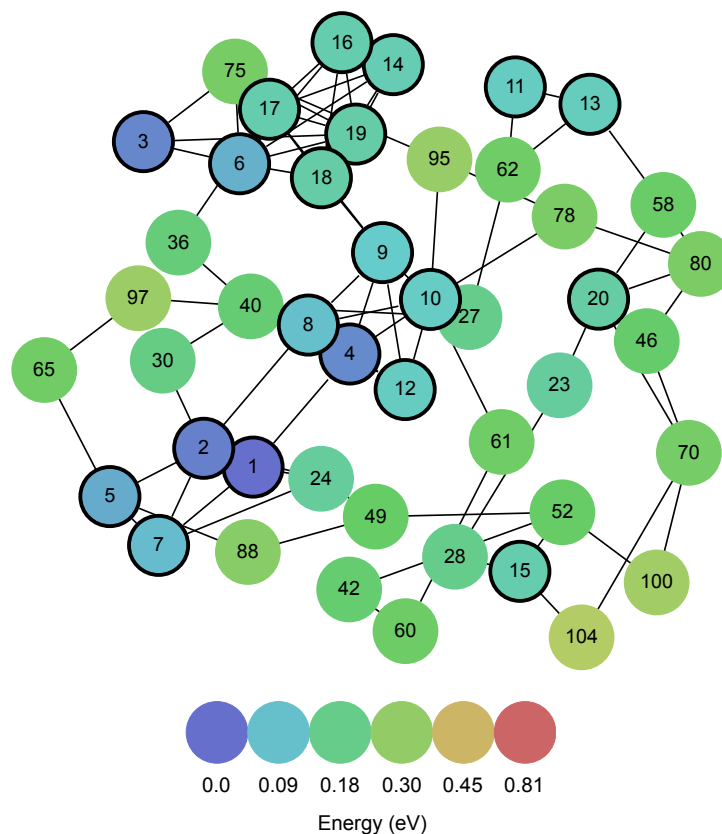


Figure 22: The geometrical-similarity network connecting the 20 lowest energy structures of AcAla<sub>3</sub>NMe. The nodes correspond to the structures. The numbering corresponds to the energy values in ascending order, i.e. the global minimum is labeled with '1'. The distance between two nodes corresponds to the similarity. Only the pairs of nodes connected with an edge are selected for the barrier calculations. The visualization is an approximation of the N-dimensional graph projected onto a 2D space. The graph was generated with Graphviz [179] utilizing the *neato* utility.

are *combined transition paths* via a number of different states (e.g.  $4 \Rightarrow 1 \Rightarrow 24 \Rightarrow 49 \Rightarrow 52 \Rightarrow 15$ ). Due to the fact that the suggestions for the paths were generated independently from each other, there are more connections in the graph than potentially needed. This is a natural consequence of the independent optimization of the paths and this occurs with intent. The here suggested network of connections contains information about the structural similarity of the conformers. Once the values of the barriers are known, the paths can be re-evaluated and a network containing information about the energy of the transitions can be established.

#### 4.3.4 Calculations of the energy barriers

For each of the selected 81 connections an energy barrier was identified. First, a clash-free initial guess for a path with 20 images connecting the initial and the final structure was generated for each pair. The paths were optimized

with aimsChain interfaced to Tinker, utilizing the string method and the TR optimizer. The convergence criterion was set to 5 meV.<sup>1</sup>

The resulting barriers were sorted in ascending order. Then, starting from the pair of structures connected via the lowest barrier, the states were connected to construct the network. This procedure was continued until all 20 lowest energy structures were connected. This was achieved after connecting the states associated with the first 45 barriers. The resulting network contains a number of 'dead ends', i.e. single-connected states. As we are only interested in the connections between the 20 energy-lowest structures, all 'dead ends' that do not connect to any interesting structures are deleted. The final network consists of 31 states and 30 connections (Figure 23).

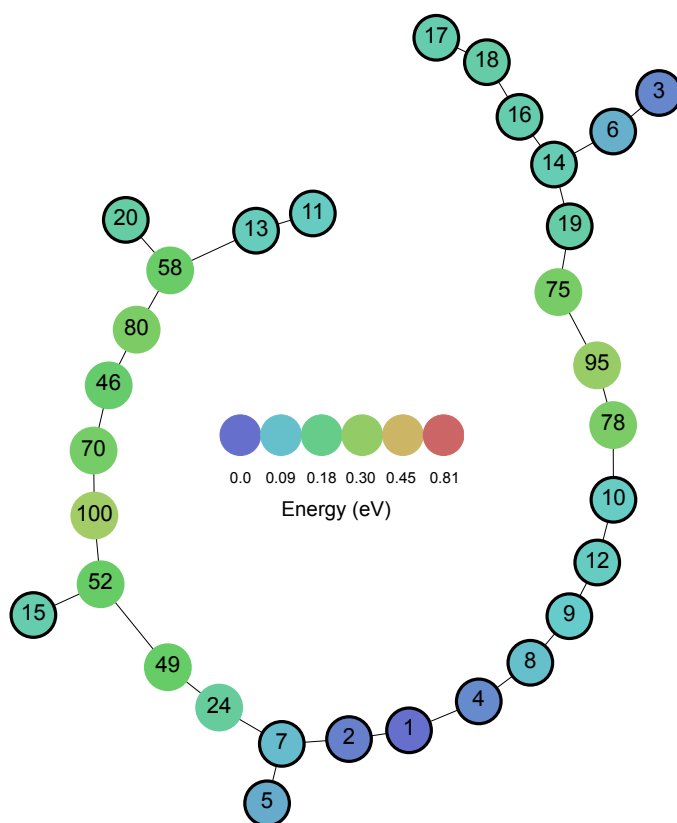


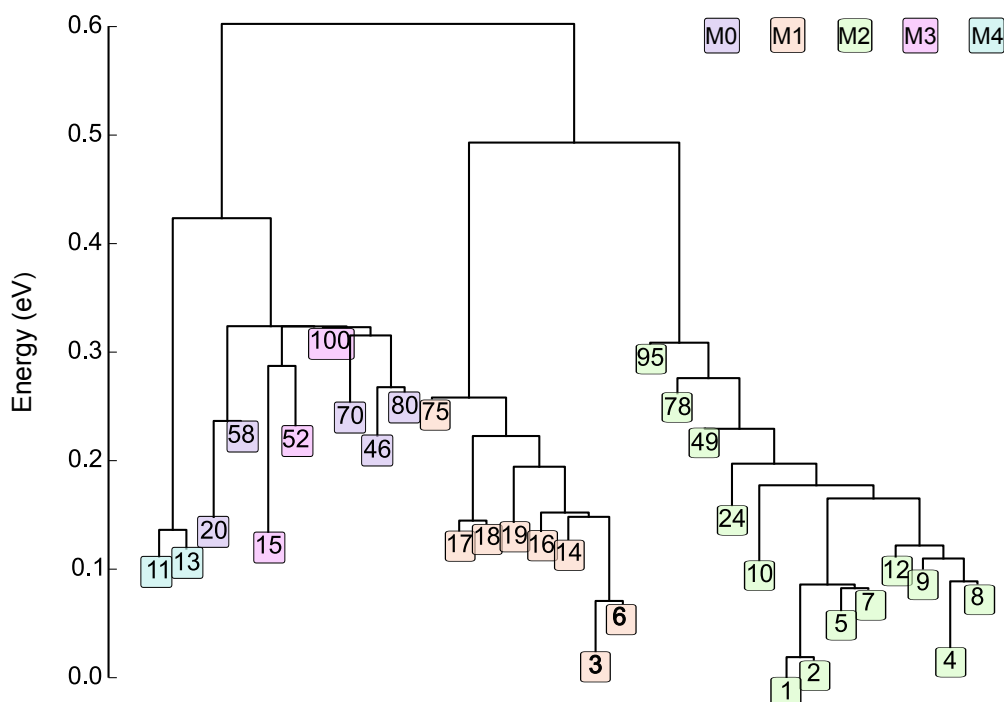
Figure 23: The energy barrier network connecting the 20 lowest energy structures of AcAla<sub>3</sub>NMe. The nodes correspond to the structures. The numbering corresponds to the energy values in ascending order, i.e. the global minimum is labeled with '1'. The distance between two nodes connected with an edge corresponds to the height of the energy barrier. The distances between nodes not connected with an edge are arbitrary.

*Tinker:  
Amber99sb force  
field*

The length of the edges corresponds to the height of energy barrier between the connected nodes. For each pair of connected structures, the height of the barrier is relative to the energy of the energetically more stable structure in the pair. The resulting network has a shape of a long chain with a few side chains. Although a lot of information can be recovered from such depiction, it would be favorable to connect the information about the energy of the states and the energy of the barriers. Barrier trees offer an opportunity to visualize

<sup>1</sup> In some cases, if redundant oscillations occurred, the calculation was declared converged at a slightly higher threshold, i.e. ca. 10 meV.

the energy relation in a more convenient way. In order to construct the barrier tree, energy barriers are made relative to the globally most stable structure. Figure 24 depicts the resulting barrier tree.



*Tinker:  
Amber99sb force  
field*

Figure 24: The barrier tree for the 20 lowest-energy structures of AcAla<sub>3</sub>NMe calculated at the Amber99sb force field level. All energy values are relative to the global minimum, i.e. structure labeled '1'. Each of the structures was assigned to one of the reference MSM macrostates and colored accordingly.

For each of the structures plotted in the barrier tree, the most similar structure among the reference FES structures used to construct the MSM model was identified. With this, each of the structures can be assigned to one of the reference macrostates.

There is a good agreement between the division of structures into the macrostates and the shape of the barrier tree. Structures belonging to the macrostates M<sub>1</sub>, M<sub>2</sub> and M<sub>4</sub> are clearly separated (Figure 21).

The results presented here are based exclusively on a selection of PES structures. Valuable insight into the system and the relation to the reference FES could have been obtained without performing free-energy MD simulations.

It should be emphasized that the shape of the barrier tree is strongly dependent on the level of theory utilized for the computation of the PES. Already for small systems, the PES undergoes rearrangements depending on the utilized energy function. This was demonstrated e.g. for alanine dipeptide (Figure 3.1 in [30]). Given the fact that alanine tetrapeptide is more flexible than alanine dipeptide, even more severe PES rearrangements should be expected. Though, the presented alanine tetrapeptide study was performed to merely prove the concept that the PES-based barrier tree can give some insights into dynamical properties of the systems. Encouraged by the results



of this 'proof of concept', we present in the next section a study where the energy was calculated at the first principles level.

#### 4.4 REDUCED PES OF A SYNTHETIC PEPTIDE FROM FIRST PRINCIPLES

This section is dedicated to the exploration of the PES of a well-characterized synthetic peptide,  $\gamma\alpha$  (Figure 25). The  $\gamma\alpha$  peptide was studied experimentally in the gas-phase using conformation-selective spectroscopy [180]. In addition to the experimental findings, the study of Kusaka, reported computational results of a conformational search for minima on the gas-phase PES.

The goal of the study presented here is to: (i) verify how well our computational search strategy samples the PES of  $\gamma\alpha$  in comparison to the published computational and experimental results; and (ii) calculate a barrier tree for a selection of low-energy conformers.

The  $\gamma\alpha$  peptide has 8 rotatable bonds and 3 *cis/trans* peptide bonds suggesting great flexibility and a number of low-energy conformers. All energy evaluations for this molecule were performed with the PBE functional and the MBD dispersion correction.

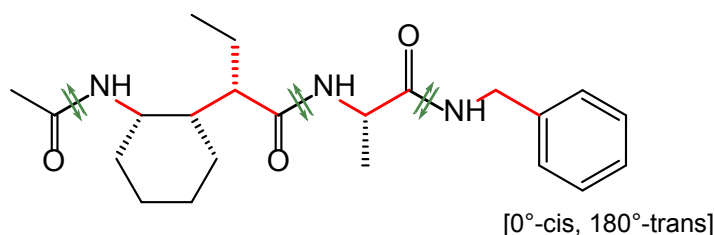


Figure 25: Structure of the  $\gamma\alpha$  peptide. The rotatable bonds are marked in red and the *cis/trans* bonds are marked in green.

##### 4.4.1 Sampling of the potential-energy surface

###### 4.4.1.1 GA searches

The PES of the synthetic  $\gamma\alpha$  peptide was sampled with a GA search performed with Fafoom. 30 independent GA runs were performed yielding 750 structures optimized at the PBE+MBD level. The duplicate removal resulted in 471 structures, with 71 structures having a relative energy below 0.3 eV.

###### 4.4.1.2 Reference data

In the computational part of the study of Kusaka, 18 distinct conformers of the  $\gamma\alpha$  peptide were reported. These structures were optimized with the highly parameterized hybrid meta exchange-correlation functional M05-2X functional [181]. Prior to the comparison with the structures obtained from our GA searches, these reference structures were reoptimized at the PBE+MBD level. From the set of the optimized structures, duplicates were removed so that the total number of structures was reduced to 16. All 16 structures have a relative energy lower than 0.3 eV.

#### 4.4.1.3 Comparison of the reference and GA data

The 16 reference structures were compared to the 471 structures obtained with the GA based on the geometrical similarity. As a result, 14 out of 16 reference structures also were identified by the GA search. Further, two structures from the reference set were not found by our GA: one with a relative energy of ca. 90 meV, and another one with a relative energy of ca. 230 meV. Moreover, the GA search identified 57 structures that were not present in the reference set but are also low-energy (i.e. with a relative energy lower than 0.3 eV). The energy hierarchy of the  $\gamma\alpha$  peptide is depicted in Figure 26.

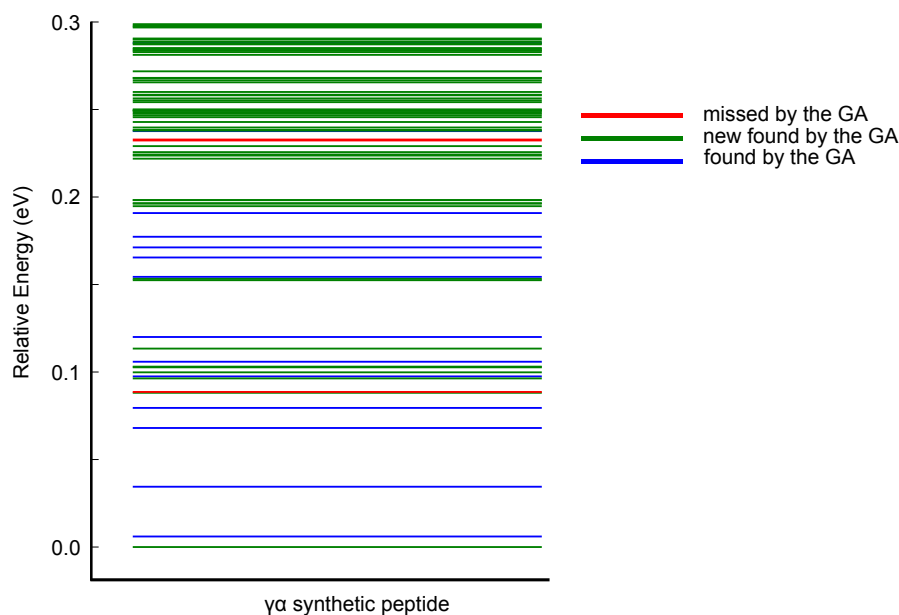


Figure 26: Low energy structures of the  $\gamma\alpha$  peptide.

*FHI-aims:  
PBE+MBD,  
light species  
defaults*

#### 4.4.1.4 Final set of structures

We aimed at constructing a representative set of low-energy ( $< 0.3\text{eV}$ ) structures. To this end, we combined: the 14 structures known from the reference set and identified by the GA, the 57 structures identified only by the GA and 2 structures missed by the GA but known from the reference set. The final set consists of 73 structures.

#### 4.4.2 Towards a network of states

Similarly as in the AcAla<sub>3</sub>NMe-study, tRMSD was selected for the calculation of the ( $73 \times 73$ ) similarity matrix. All 11 DOFs, i.e. 8 rotatable bonds and 3 *cis/trans* bonds, were considered for the tRMSD calculation. Values exceeding  $\text{tRMSD} = 0.3 \pi$  were removed from the matrix.

The 10 lowest-energy structures were selected for the construction of the network and independent suggestions for 45 paths were generated. The paths were split into single connections and merged to one list with 55 unique

connections. The network constructed from this connection is presented in Figure 27.

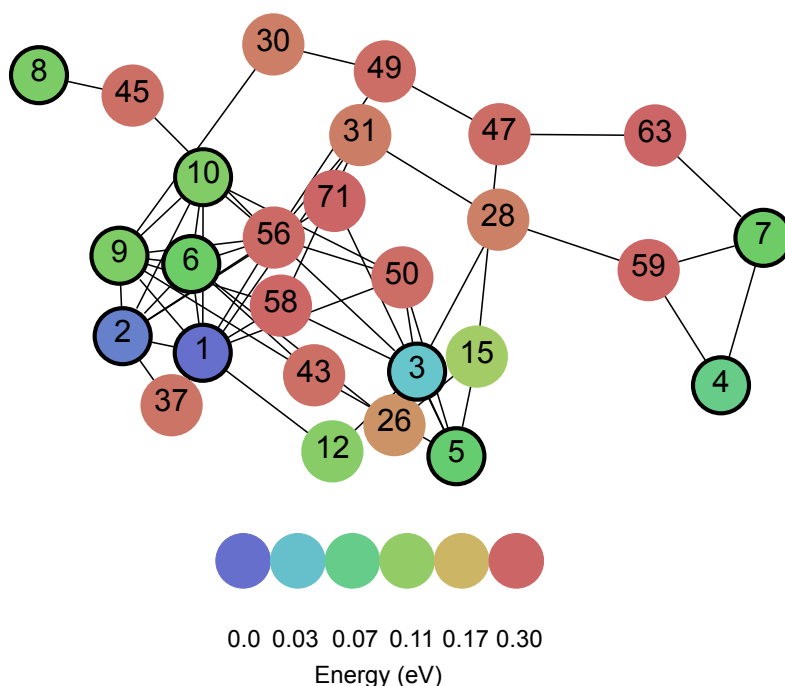


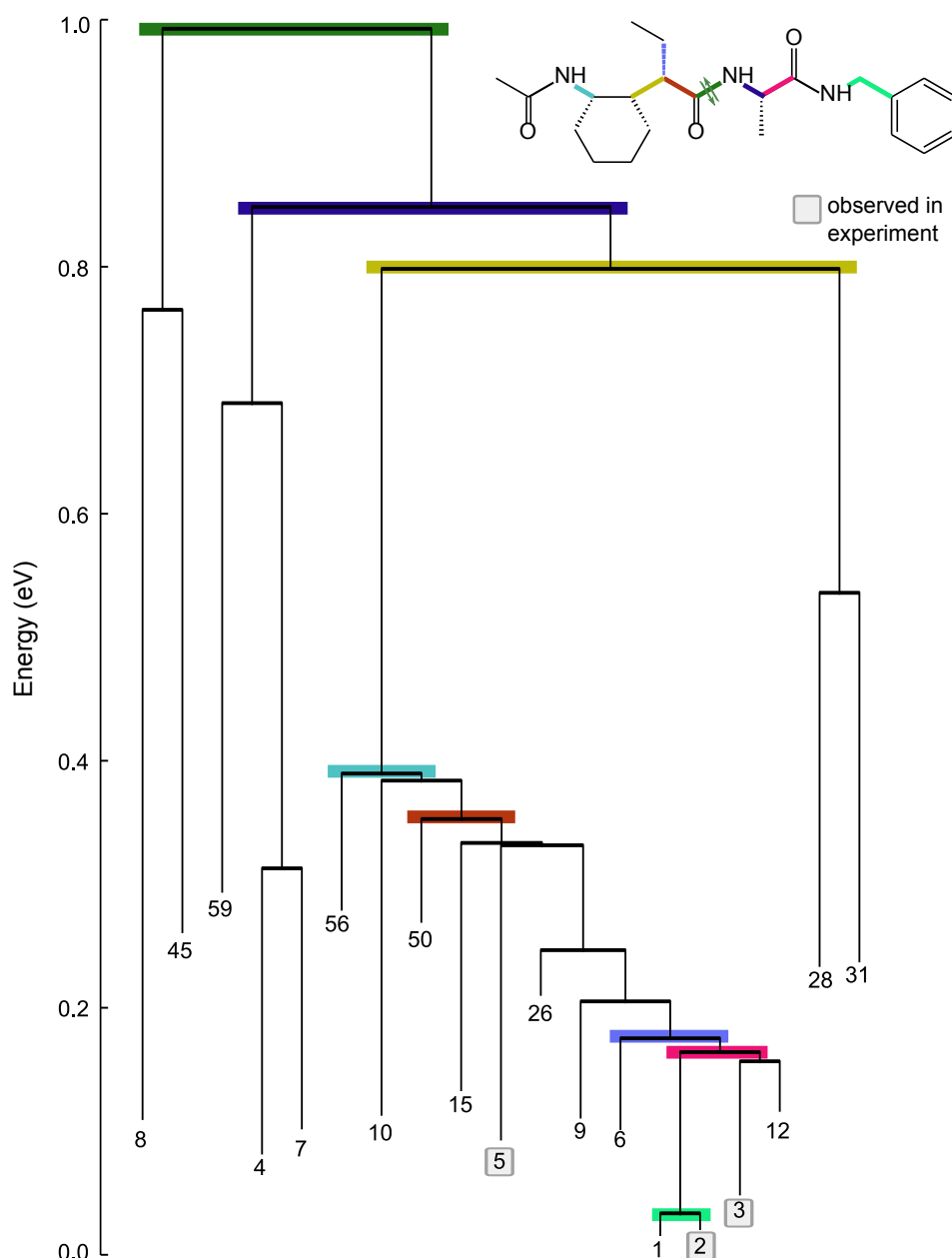
Figure 27: The geometrical-similarity network connecting the 10 lowest energy structures of the  $\gamma\alpha$  peptide. The nodes corresponds to the structures. The numbering corresponds to the energy values in ascending order, i.e. the global minimum is labeled with '1'. The distance between two nodes corresponds to the similarity, also if no edge is plotted. Only the pairs of nodes connected with an edge are selected for the barrier calculations. The visualization is an approximation of the N-dimensional graph projected onto a 2D space. The graph was generated with Graphviz [179] utilizing the *neato* utility.

The network shows the suggested paths connecting the 10 low-energy states of the  $\gamma\alpha$  peptide. Only 12 of the 45 path suggestions are *direct transitions* (e.g.  $2 \Rightarrow 6$ ) while more than a half of the path suggestions are *combined transition paths* via at least two different states. Barriers are calculated for all single connections depicted in the graph.

#### 4.4.3 Calculations of the energy barriers

For each of the selected 55 connections an energy barrier was identified. First, a clash-free initial guess for a path with 10 images connecting the initial and the final structure was generated for each pair. The path was optimized with aimsChain interfaced to FHI-aims, utilizing the string method and the TR optimizer. The convergence criterion was set to 5 meV.<sup>2</sup> The energy and force evaluations were evaluated at the PBE+MBD level. The minimum number of barriers needed to connect the 10 lowest-energy states is 18 and is the resulting barrier tree is visualized in Figure 28.

<sup>2</sup> In some cases, if redundant oscillations occurred, the calculation was declared converged at a slightly higher threshold, i.e. ca. 10 meV.



*FHI-aims:  
PBE+MBD,  
light species  
defaults*

Figure 28: The barrier tree for the 10 lowest-energy structure of the  $\gamma\alpha$  peptide calculated at the PBE+MBD level. All energy values are relative to the global minimum, i.e. structure labeled '1'.

In order to investigate the relationship between the structural change and the energy barriers, a further analysis was performed. For each of the plotted barriers, the geometries of the two respective structures were compared. To this end, the values of the respective DOFs were compared and the DOF that deviated most was identified. This identified DOF was then assigned to the barrier and is referred to as its character. Among barriers of the same character, the barrier with the lowest energy is selected and marked with a dedicated color. (Figure 28). Note that the character assignment is approximate, i.e. for most of the barriers several DOFs change.

The highest energy barrier in the barrier tree corresponds to the isomerisation of the central *cis/trans* bond. This is due to the peptide bond isomerisation being energetically more demanding than the rotation around the single bonds. Moreover, there is a general trend that the heights of the barriers decrease with the increasing distance between the barrier character and the center of the molecule.

In addition, the experimentally identified structures are also marked in the barrier tree. It is worth noting that there are no high energy barriers between these structures. The global minimum ('1') was not observed explicitly in the experiment. This is likely due to the fact that the second best structure (observed in experiment) is structurally similar and very close in energy. The energy barrier between these two structures is predicted to be very low so that the interconversion rate between them is very high and these two structures can be effectively regarded as one minimum.

There are several low-energy structures (e.g. '4') that were not observed in the experiment. From the barrier tree it is clear that there is a high barrier separating the groups of minima. Thus, it is likely that some low-energy structures are not observed during the experiment because of the kinetic trapping.



## GLOBAL STRUCTURE SEARCH FROM FIRST PRINCIPLES: THIOUREA CATALYSTS

---

This chapter describes how quantum mechanical investigations can advance the understanding of structure-dependent activity of (organo)catalysts. Section 5.1 motivates our interest in studying conformational preferences of catalysts. Further, it gives a brief introduction into the field of organocatalysis with a focus on thiourea catalysts. Section 5.2 presents the studied systems. Section 5.3 discuss our findings on isolated molecules. The molecular complexes of the catalyst and a model substrate are investigated in Section 5.4. The contents of the chapter will be published:

- Supady A., Ingram T., Hecht S., and Baldauf C. Catalytic activity of thiourea derivatives requires structural rearrangements via intermolecular interactions. 2016, in preparation.

### 5.1 MOTIVATION

The impressive advances in chemical synthesis allows nowadays for an automatized synthesis of peptides, nucleic acids and oligosaccharides [182]. Recently, an approach was proposed that allows for a more generalized automation of the synthesis of a number of distinct classes of small-molecule [183]. It is currently less an issue to synthesize a given molecule than it is to design a molecule carrying a desired property. Besides the property of interest, several other properties, e.g. stability, need to be satisfied in order to make the molecule functional. A further objective in compound design may be the reduction of the generation of hazardous substances in their synthesis or application. Computational techniques can support solving such multi-objective optimization problems [184]. Though this is only feasible if the desired property can be defined as a computable observable.

#### 5.1.1 *Structure-dependent catalytic activity*

Catalysts lower the energy barriers of chemical reactions. Compared to uncatalyzed reactions, catalysts open alternative reaction routes via competing transition states. Typically, catalysis involves the following steps: (i) the educt(s) form a complex with the catalyst; (ii) a transition state is formed; (iii) a product-catalyst complex is formed, and finally (iv) the product and the catalyst separate and the catalyst is released. The chemical equilibrium between the educts and the product is not influenced by the use of a catalyst. The efficiency of such catalytic process, i.e. the reaction rate increase, depends on many different factors. The most prominent ones are environmental conditions, the availability of the substrate and the catalytic activity of the catalyst. The catalyst's activity depends on: (i) its chemical composition, (ii) adopted 3D structure and (iii) environment-dependent interactions with the substrate.

Further, to design a functional catalyst, additional aspects like toxicity, product inhibition, and production applicability need to be addressed.

The 3D structure of a molecule determines its physical and chemical properties (e.g. catalytic activity). This concept is often referred to as structure-property-relationship. Flexible organic molecules can adopt different 3D structures exhibiting different properties. In consequence, 3D structures of the same molecule with different properties can co-exist.

Here we study the structural preferences of potential organic catalysts, with the focus on the relation between the adopted 3D structure and catalytic activity. Several drivers may increase the probability for adopting an active conformation by the molecule. Four of them are introduced here and will be referred to later in the chapter. The active conformation can be:

1. Energetically the most favorable one with the remaining conformations being significantly less stable
2. Induced only or enhanced by the presence of the substrate
3. Induced by self-association of the catalyst leading to a relative destabilization of the catalytically inactive conformations
4. In a dynamic equilibrium with inactive conformations but have the lowest energy of the corresponding transition state on the way to the product ('Curtin-Hammett principle' [185])

Curtin-Hammett principle (Figure 29) states that the product ratio ( $P_1 : P_2$ ) does not exclusively depend on the energy difference between the particular conformers ( $\Delta G_0$ ) but essentially on the energy difference of the corresponding transition states ( $\Delta G_{TS}$ ). The Curtin-Hammett principle is valid only if the interconversion rate between  $A_1$  and  $A_3$  is significantly higher than the reaction rates  $k_1$  ( $A_1 \xrightarrow{k_1} P_1$ ) and  $k_2$  ( $A_3 \xrightarrow{k_2} P_2$ ).

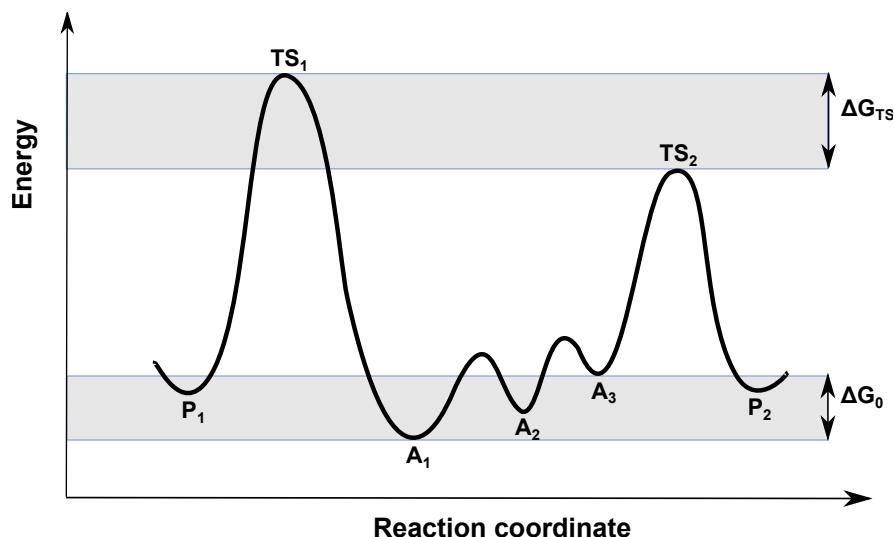


Figure 29: Energy profile for the illustration of the Curtin-Hammett principle.  $A_1$ ,  $A_2$  and  $A_3$  are conformational isomers of the substrate while  $TS_1$  and  $TS_2$  are the transition states going to products  $P_1$  and  $P_2$ . In the presented example, product  $P_2$  will be preferred because  $TS_2 < TS_1$ .



Investigation of the factors facilitating the presence of active conformation is further complicated by the fact that several of the listed drivers can occur simultaneously.

### 5.1.2 Thioureas acting as organocatalysts

The type of catalysis whereby the reaction is accelerated by an organic catalyst is referred to as organocatalysis. Organocatalysts are an attractive alternative to metal(ion)-based catalysts as they are environment-friendly and commonly do not require strong enthalpic binding. Making use of weaker interactions can result in diverse benefits, e.g. the potential loss of purity of the product can be prevented. Significant advances have been achieved in this field in the last 20 years [186–193].

A number of diverse organocatalysts have been developed acting via different catalytic mechanisms. Here the focus will be given to these organocatalysts that trigger a reaction by the means of nucleophilic/electrophilic properties of the catalyst. In such case, the transition state is stabilized by the hydrogen bonding with the catalyst acting like a Lewis acid. This mechanism is employed by, e.g. the urea- and thiourea-based catalysts. Figure 30 illustrates, how the thiourea subunit of a catalyst coordinates an exemplary substrate. It has been shown that a broad range of substituted thioureas is able to catalyse Diels-Alder reactions [186] and that the increase in the reaction rate depends more on the kind of the substituents than on the solvent.

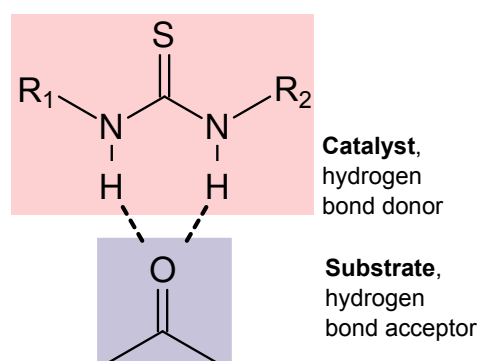


Figure 30: Simplified scheme of the interactions between the thiourea-based catalyst and exemplary substrate.

Organocatalysts are commonly flexible molecules and can adopt a variety of conformations with different properties such as catalytic activity. As depicted in Figure 30, the catalytic activity of the thiourea moieties can be defined as the capability of the molecule to participate in a double hydrogen bond. In consequence, only the conformations of the catalyst are active where both hydrogens of the thiourea group are directed to the opposite direction in the sulfur atom.

## 5.1.3 Studied problem

A number of theoretical studies [187, 194, 195] have been published explaining the mechanistic details of diverse organocatalytic reactions. Nevertheless, only a few studies [188, 189, 196] explored the structural freedom of the catalysts and its consequences.

Here we study the relation between the structural preferences and the catalytic activity on the example of thioureas. In the following we utilize diverse computational techniques to reveal which of the four proposed drivers may play the leading role in adopting the catalytically active conformation by the thiourea catalysts.

## 5.2 MOLECULAR SYSTEMS

It was observed that subtle differences in the composition of the substituents may result in distinct activity levels of the thioureas [186]. To better understand this finding, a set of molecules sharing a similar scaffold with different substituents was created. The set includes eight molecules, four ureas and four thioureas (Figure 31). The substituents range from methyl groups to substituted phenyl/cyclohexane rings.

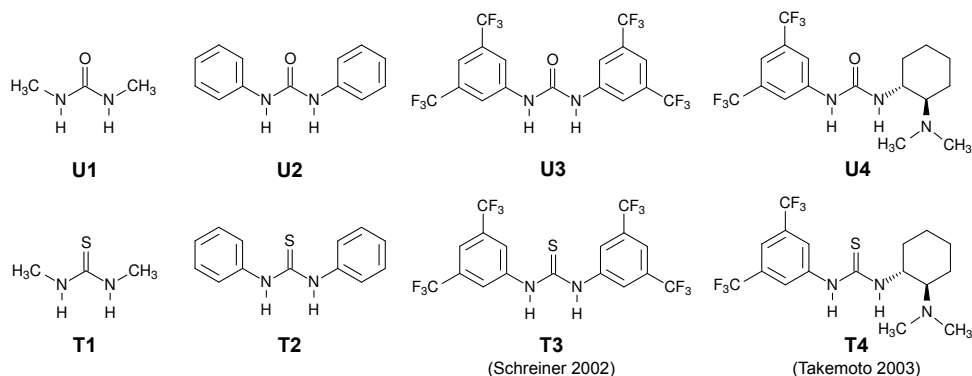


Figure 31: Schematic representation of investigated urea (**U1-U4**) and thiourea (**T1-T4**) derivatives. **T3** is known as the *Schreiner catalyst* and **T4** as the *Takemoto catalyst*.

Two out of the eight presented molecules are known to exhibit a distinctive catalytic activity: **T3** and **T4**.

**T3**, the *Schreiner catalyst* (*N,N'*-bis[[3,5-bis(trifluoromethyl)phenyl]thiourea) is a very popular and effective catalyst [186, 191]. It is easy to synthesise, soluble, electron poor and does not cause product inhibition.

**T4**, the *Takemoto catalyst* is a bifunctional chiral thiourea derivative, catalysing e.g. the Michael reaction of malonates to various nitroolefins with high enantioselectivity [192].

The relative positions of the substituents to the oxygen/sulfur center determine the orientation type of the structure. Four isomer types can be defined: *syn-syn*, *anti-anti*, *anti-syn* and *syn-anti* (Figure 32). For molecules carrying two identical substituents (**U1-U3** and **T1-T3**) the *anti-syn* and *syn-anti* conformers are equivalent, i.e. the number of orientation types to be considered can be reduced to three. Bearing in mind the interactions initiating the catalytic re-

action (Figure 30) and the corresponding structural requirement, exclusively the *anti-anti* conformer can be referred to as 'active' conformation.

all:

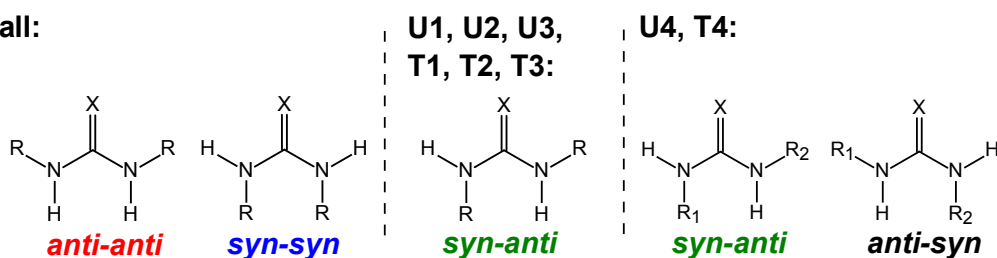


Figure 32: Types of isomers.

### 5.3 ISOLATED MOLECULES

In this section, the energetics of the presented set of molecules at different levels of theory is investigated. Further, the accuracy of the applied levels of theory is evaluated. The workflow of this multi-step theoretical investigation is depicted in Figure 33 and commented in detail further in the text. At last, energy barriers between different conformations on the example of two thiourea molecules are presented.

#### 5.3.1 Energetics

**1. SAMPLING OF THE PES** The initial conformers were generated: (i) manually (**U1** and **T1**) or (ii) with a genetic algorithm based search performed with Fafoom [95](all remaining molecules). The DFT optimization was performed with the FHI-aims code with the PBE functional, pairwise vdW correction and *light* species defaults. The number of resulting conformers after duplicate removal is shown in Table 7.

Table 7: Number of conformers generated in a genetic algorithm based search performed with Fafoom at the PBE+vdW level. The duplicates were removed.

molecule	No. of conformers	molecule	No. of conformers
<b>U1</b>	3	<b>T1</b>	3
<b>U2</b>	3	<b>T2</b>	3
<b>U3</b>	10	<b>T3</b>	8
<b>U4</b>	70	<b>T4</b>	63

**2. FHI-AIMS REFINEMENT** The structures resulting from the sampling of the PES were further refined with FHI-aims [80], utilizing the following setup: PBE0, MBD, and tight species defaults. The number of structures resulting after this refinement and after removing duplicates is summarized in Table 8.

**3. ORCA REFINEMENT** The resulting conformers were utilized as starting points for the refinement with the ORCA software that supports wave-based

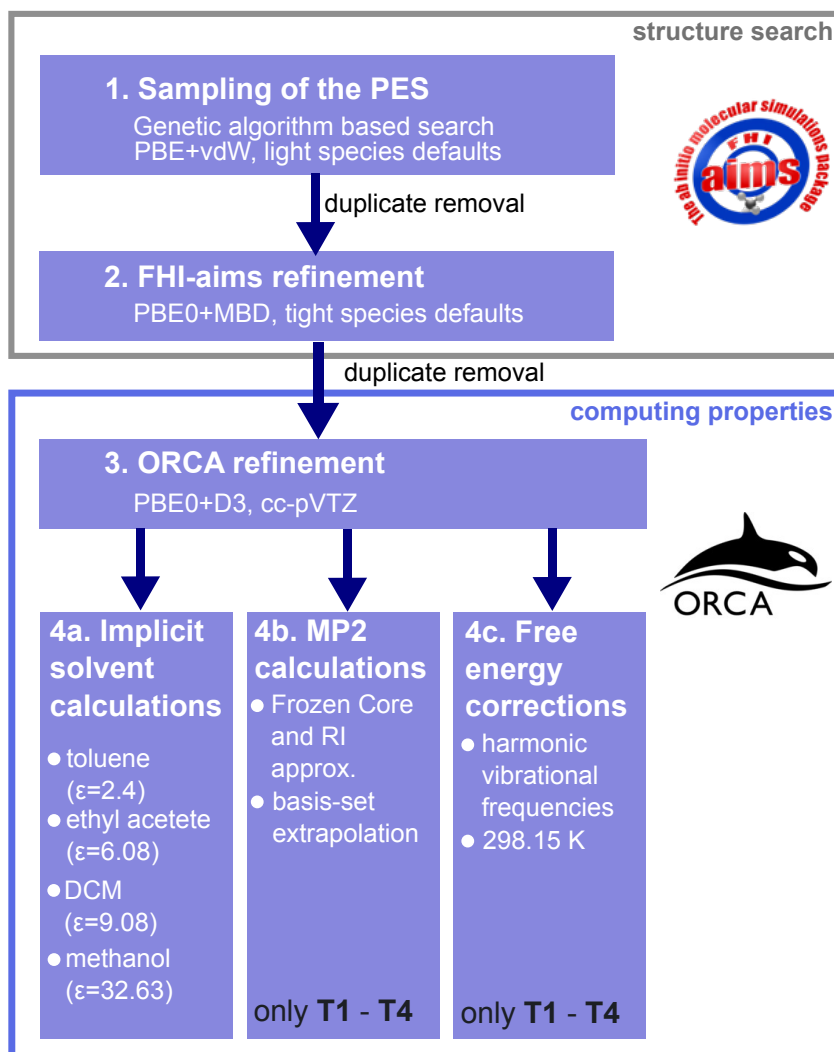


Figure 33: Scheme representation of the workflow.

Table 8: Number of conformers after the refinement at the PBE0+MBD level after the duplicate removal.

molecule	No. of conformers	molecule	No. of conformers
<b>U<sub>1</sub></b>	3	<b>T<sub>1</sub></b>	3
<b>U<sub>2</sub></b>	3	<b>T<sub>2</sub></b>	3
<b>U<sub>3</sub></b>	5	<b>T<sub>3</sub></b>	3
<b>U<sub>4</sub></b>	50	<b>T<sub>4</sub></b>	5 <sup>1</sup>

energy functions and implicit solvent modeling. The structures were reoptimized with the PBE0 functional but with an alternative dispersion scheme and another basis set. Here, the D3 dispersion correction and cc-pVTZ basis set were utilized in contrast to the MBD scheme and tight species defaults used in the FHI-aims refinement. The resulting energy hierarchies are depicted in the first panel of Figure 34. No significant change in the energetics between the PBE0+MBD scheme with tight species defaults and PBE0+D3 with cc-pVTZ was observed (see Appendix A.3).

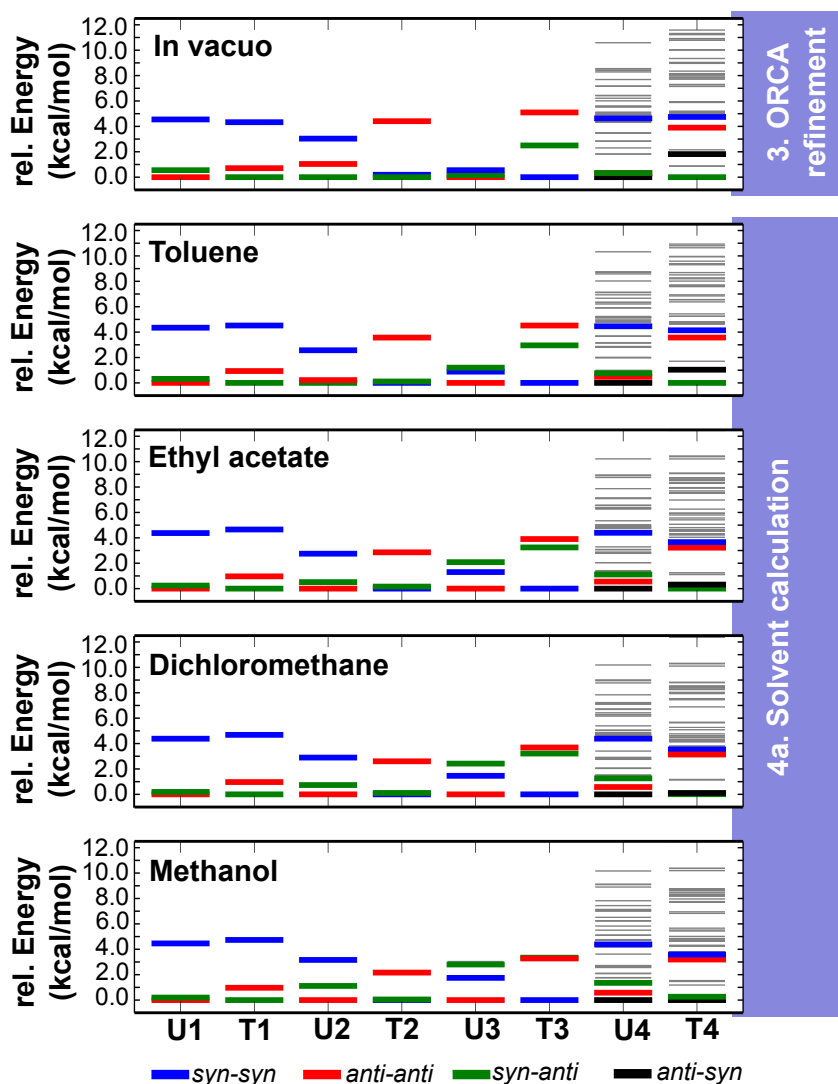


Figure 34: Relative energy of investigated molecules: in vacuo, in toluene, in ethyl acetate, in dichloromethane, and in methanol. Only the most stable structure of each of the isomer types ("representative structure") is depicted in color: anti-anti in red, syn-syn in blue, syn-anti in green and anti-syn (only for U<sub>4</sub> and T<sub>4</sub>) in black. Structures, that are less stable than the corresponding "representative structure" are marked in gray

ORCA:  
PBE0+D<sub>3</sub>,  
cc-pVTZ basis  
set

**4A. IMPLICIT SOLVENT CALCULATIONS** For the conformers refined with ORCA, calculations with the COSMO solvent model were performed (Figure 34). The effect of the following solvents was investigated: toluene ( $\epsilon = 2.4$ ), ethyl acetate ( $\epsilon = 6.08$ ), dichloromethane ( $\epsilon = 9.08$ ), and methanol ( $\epsilon = 32.63$ ).

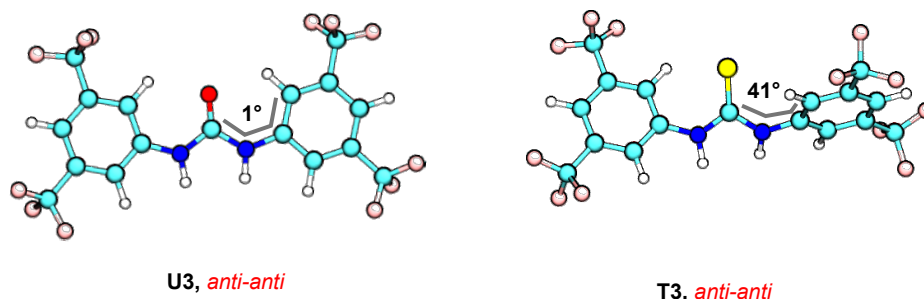
Consistently with previous studies [197–199], the *anti-anti* and the *syn-anti* conformers of U<sub>1</sub> and T<sub>1</sub> are close in energy ( $< 1$  kcal/mol), while the *syn-syn* conformer is higher in energy by more than 4 kcal/mol. Further, no significant solvent influence is observed, only the *syn-anti* conformer of U<sub>1</sub> gets slightly stabilized.

The most stable form of U<sub>2</sub> in the gas phase is the *syn-anti* conformer, that is in agreement with previous studies [200, 201]. Comparing with U<sub>1</sub>, the U<sub>2</sub>

*syn-syn* conformer is energetically more favorable. This can be assigned to the  $\pi - \pi$  stacking effect. This effect becomes even more evident in the case of **T2** due to enhanced polarizability of the sulfur atom. In solution, the *anti-anti* conformers get stabilized, both in **U2** and **T2**.

Kirsten *et al.* [188] analyzed the conformational preferences of **T3** and found the *syn-anti* conformer to be the most stable conformer in the gas-phase. In solution, the *anti-anti* conformer was preferred, followed by the *syn-syn* conformer. Lippert *et al.* [195] reported similar results, but in this study the energetics of the *syn-syn* conformer was not evaluated. Our results clearly disagree with the previous studies: the most stable conformer of **T3** features the *syn-syn* orientation while the *anti-anti* conformer gets stabilized in the solvent. In the case of **U3** the energy differences between the conformers are small.

It should be noted, that the geometries of the *anti-anti* conformers of **U3** and **T3** have significant differences (Figure 35). While **U3** is nearly planar, a  $41^\circ$  angle between the phenyl ring plane and the thiourea-group plane is observed in **T3**. The larger angle is due to a stronger repulsion between the sulfur center atom and the ortho-hydrogens of the phenyl ring compared to the repulsion in the case of an oxygen center atom [202].



ORCA:  
PBE0+D3,  
cc-pVTZ basis  
set

Figure 35: The *anti-anti* conformers of **U3** and **T3**. Different atom types are colored as follows: sulfur (yellow), oxygen (red), carbon (light blue), nitrogen (dark blue), fluor (rosa), and hydrogen (white).

It is known from the crystal structure of the Takemoto catalyst [192] (**T4**) that the NH protons of the thiourea group and the tertiary amino group are oriented in the same direction. Further investigations revealed that the different conformers of the free catalyst are in a dynamic equilibrium, in gas-phase and in solution, and that the catalyst cannot be regarded as rigid [189]. We find that the catalyst preferentially adopts catalytically inactive conformations (*anti-syn* and *syn-anti*) promoted by the intramolecular hydrogen bonding between the basic nitrogen and the thiourea hydrogens (Figure 36). Tarkanyi and colleagues suggested that it is the self-association of the catalyst that promotes the active conformation by destabilising the inactive conformations [196].

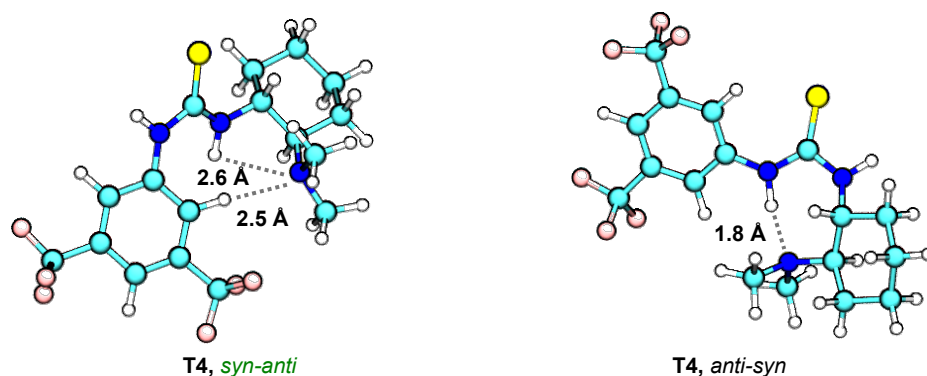


Figure 36: The *syn-anti* and *anti-syn* conformers of **T4**.

ORCA:  
PBE0+D3,  
cc-pVTZ basis  
set

The evaluation of the total energy of the isolated molecules allows us to formulate the following:

- The catalytically active conformers of **T3** and **T4** are not the most stable conformers.
- A thermodynamic equilibrium between the different conformers can be assumed as the energy differences are often small.

Given these hypotheses, it should be noted that the accuracy and level of the energy calculations is critical for the investigation. Diverse shortcomings of popular computation methods were observed. First, the black-box use of popular DFT functionals has several limitations [22, 203]. Second, dispersion interactions cannot be neglected [21, 200]. Further, more subtle accuracy limitations should be considered, e.g. MP2 has the tendency to overestimate the stacking interaction energy [201, 204]. Thus, in the following we critically evaluate the approximations utilized by us. We investigate:

- the influence of the entropic contributions by calculating free-energy corrections
- if the results change if calculations at the MP2 level are performed
- the importance of the dispersion correction

This analysis was performed only for the thiourea molecules. Figure 37 combines the energy hierarchies obtained with different methods.

The reference results obtained after ORCA-refinement are depicted (already shown in Figure 34) in the second column. In the first column the total energies of the exact same conformers as in the second column are depicted with the only difference that the dispersion correction term was ignored. The third column depicts energy hierarchies obtained after optimization with MP2 and the last column depicts the the Gibbs free enthalpies.

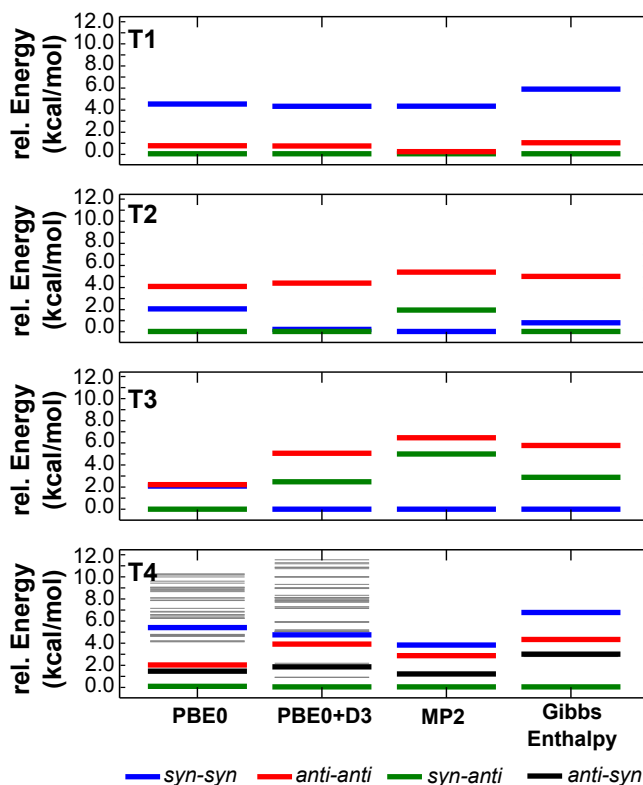


Figure 37: Relative energies of urea and thiourea molecules at different levels of the-ory: PBE0, PBE0+D<sub>3</sub> and MP2. In the last column, the total Gibbs free enthalpy calculated at the PBE0+D<sub>3</sub> level at 298.15 K is depicted.

**4B. MP2 CALCULATIONS** For the ORCA-refined conformers of the thiourea molecules (**T1-T4**) a two-step MP2 optimization with ORCA was performed. First, the geometries were optimized with cc-pVTZ basis set [154]. Next, single point calculations were performed with a 3,4 extrapolation [156] to the CBS limit.

**4C. FREE ENERGY CORRECTIONS** For the ORCA-refined conformers of each of the thiourea molecules harmonic vibrational frequencies were calculated numerically. From the obtained results, the thermochemical properties at 298.15 K can be derived. The calculations were performed in vacuo.

A number of conclusions can be made based on the results depicted in Figure 37. First, the entropic contributions barely influence the relative energies of the conformers. Second, there is a very good agreement between the DFT approximation chosen by us (PBE0+D<sub>3</sub>) and the results obtained at the MP2 level. Only the *syn-syn* conformers of **T2** and **T3** are more stable at the MP2 level. However, this can be assigned to the previously mentioned fact that MP2 tends to overestimate the  $\pi - \pi$  stacking. Further, the close match between the results obtained with PBE0+D<sub>3</sub> and MP2 is clearly facilitated by the incorporation of a dispersion correction. In the case of **T3**, the *syn-syn* conformer becomes the most stable one only if the correction for dispersion is incorporated. On the other hand, the energy difference between the *anti-anti* and *syn-anti* remains unchanged. Summing up, meaningful values can



be obtained at the PBE0 level, but only if the dispersion correction is incorporated. It should be noted that neglecting dispersion will be more significant for larger systems.

### 5.3.2 Energy barriers

The computation of the relative energies of conformers is a mandatory step in studying structure-property relationship in molecules. To obtain a complete picture, finding and characterizing the transition states between the conformers is also required. For example, the accessibility of certain states can be evaluated by means of minimum energy paths and associated energy barriers. As suggested in the previous subsection, the different conformers of some of the investigated molecules may be in a dynamic equilibrium.

Upon further investigation, we selected two molecules known to exhibit significantly different catalytic activity: **T3**, an effective catalyst, along with **T2** that was shown to exhibit poor catalytic activity [186].

**ROTATIONAL BARRIERS** The **T2** and **T3** molecules were selected for the investigation of the energy barriers. Rotational barriers between the following isomer pairs were considered: *anti-syn* & *syn-syn* and *anti-syn* & *anti-anti*. The corresponding minimum energy paths (MEP) were calculated with the string method implemented in aimsChain [142] at the PBE+MBD level with tight species defaults. Two energy barriers were calculated: (i) between the *syn-syn* and *syn-anti* conformers and (ii) between the *syn-anti* and *anti-anti* conformers. The resulting energy profiles are depicted in Figure 38.

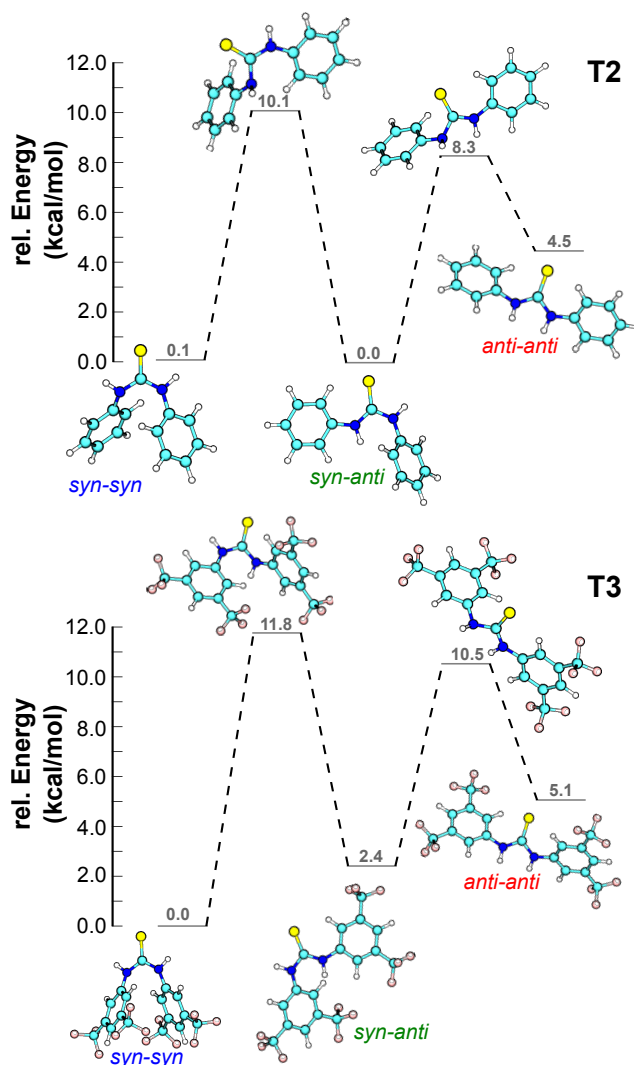
The energy values obtained for the rotation of the C-N bond in **T2** are in close agreement with previously reported values [202]. The energy barriers in **T3** are only slightly higher than in **T2**, suggesting that the CF<sub>3</sub> groups do not hinder the rotation around the C-N bond. Furthermore, the energy barrier heights can be classified as low/middle-high and the corresponding transition states are accessible at room temperature.

The presented results allow us to state that the catalytically active conformation of **T3** is less stable, but remains accessible. The chance that it will be adopted can increase under certain circumstances, e.g. in (polar) solution, the active conformer gets stabilized. Nevertheless, by means of the presented data on the isolated molecules, no clear answer to the question what drives the presence of the active conformation of the **T3** can be given yet. Thus, in the following section, the relation between the **T3** catalyst and a model substrate is studied.

## 5.4 CATALYST-SUBSTRATE COMPLEXES

In the following section we investigate how the relative energy of the different isomers of **T3** (*syn-syn*, *syn-anti* and *anti-anti*) changes in presence of a model substrate.

**STARTING GEOMETRIES** The ORCA refined conformers for the three catalyst isomers are used as starting geometries. We chose formaldehyde to act as a minimal model of a substrate. For each of the three isomers of **T3**, multiple



ORCA:  
PBE0+D<sub>3</sub>,  
cc-pVTZ basis  
set

Figure 38: Energy barriers between the different conformations of T<sub>2</sub> and T<sub>3</sub>. The numbers in denote the relative energy values of the minimum/transition state.

starting geometries together with the model substrate were generated. Upon generation we define the following distances:

- $d_1 = \text{distance}(\text{H}(1)(\text{catalyst}), \text{O}(\text{formaldehyde}))$
- $d_2 = \text{distance}(\text{H}(2)(\text{catalyst}), \text{O}(\text{formaldehyde}))$

For the complexes where T<sub>3</sub> exhibits the *anti-anti* or *syn-syn* orientation 10 random starting geometries were generated. For T<sub>3</sub> in *syn-anti* orientation 30 random starting geometries were generated. The model substrate was placed randomly taking into account the following constraints:

- if T<sub>3</sub> is in *anti-anti* orientation:  $1.5 \text{ \AA} < d_1 < 2.3 \text{ \AA}$  and  $1.5 \text{ \AA} < d_2 < 2.3 \text{ \AA}$
- if T<sub>3</sub> is in *syn-anti* or *syn-syn* orientation:  $1.5 \text{ \AA} < d_1 < 2.3 \text{ \AA}$  or  $1.5 \text{ \AA} < d_2 < 2.3 \text{ \AA}$

**ENERGY OF CATALYST-SUBSTRATE COMPLEXES** The starting structures were first fully optimized at the PBE+D<sub>3</sub> level. Then the duplicates were removed, and a refinement the PBE0+D<sub>3</sub> level was performed. Figure 39 shows the resulting energy hierarchy in comparison to the free T<sub>3</sub>.

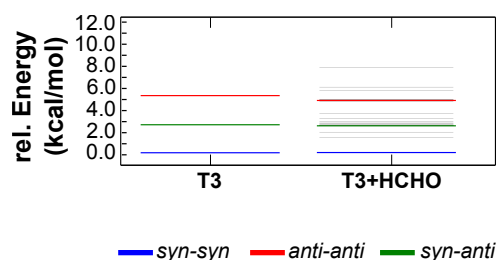


Figure 39: Relative energies of the T<sub>3</sub>-formaldehyde complexes compared to the relative energies of T<sub>3</sub>.

ORCA:  
PBE0+D<sub>3</sub>,  
cc-pVTZ basis  
set

The relative energetics of the T<sub>3</sub>-complexes remains similar to the energetics of isolated T<sub>3</sub> (for comparison see Figure 34). This implies that the active form of the catalyst (*anti-anti*) is also not the most favored one in the presence of the formaldehyde substrate. Figure 40 depicts the energetically most stable complexes of T<sub>3</sub> and the model substrate. The following characteristics were measured in order to evaluate the character of the hydrogen bonds between the formaldehyde and T<sub>3</sub>: (i) the distance between the H(catalyst) and O(formaldehyde); (ii) the angle O(formaldehyde)-H(catalyst)-N(catalyst); and (iii) the angle C(formaldehyde)-O(formaldehyde)-H(catalyst). In the case of the inactive, i.e. *syn-syn* and *syn-anti* conformers of T<sub>3</sub>, the single hydrogen bonds are shorter and of more linear character than the hydrogen bonds built by the active, i.e. *anti-anti* form of T<sub>3</sub>.

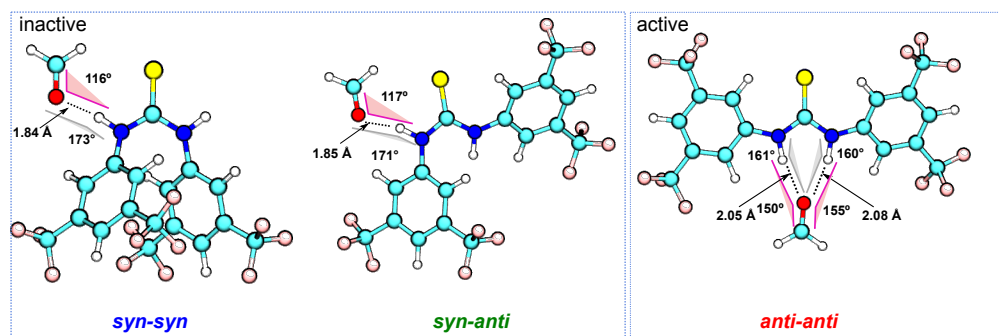


Figure 40: Energetically favorable complexes of T<sub>3</sub> conformations *syn-syn*, *syn-anti*, and *anti-anti* with formaldehyde. The distance between H(catalyst) and O(formaldehyde) is marked with an arrow. The angle O(formaldehyde)-H(catalyst)-N(catalyst) is marked in gray and the angle C(formaldehyde)-O(formaldehyde)-H(catalyst) is marked in pink.

ORCA:  
PBE0+D<sub>3</sub>,  
cc-pVTZ basis  
set

**REACTIVITY** The reactivity of the substrate is crucial for the catalytic reaction. For the three depicted complexes (Figure 40) an investigation was performed in order to quantify the differences in the reactivity of the substrate

in different complexes and alone (for comparison). The data is summarized in Table 9.

Table 9: The total charge, distribution of the partial charges, LUMO energy, and the dipole magnitude of formaldehyde. The parameters were calculated in four setups: for the formaldehyde alone and in complexes with three different T<sub>3</sub> isomers

	Partial charge of HCHO atoms				Total charge of HCHO	Dipole magnitude (debye)	LUMO (eV)
	C	O	H	H			
HCHO	0.150	-0.232	0.041	0.041	0.0	2.272	-1.15
T <sub>3</sub> ( <i>syn-syn</i> ) & HCHO	0.114	-0.256	0.101	0.087	0.046	3.473	-2.078
T <sub>3</sub> ( <i>anti-syn</i> ) & HCHO	0.115	-0.259	0.090	0.102	0.048	3.885	-2.017
T <sub>3</sub> ( <i>anti-anti</i> ) & HCHO	0.119	-0.267	0.111	0.105	0.067	8.618	-2.857

A significant difference in both the total charge and the charge distribution in the formaldehyde molecule can be observed between the complexes where the catalyst is inactive (*syn-syn* or *anti-syn*) and the complex with the active catalyst form (*anti-anti*). In the complex with the active form, the HCHO molecule carries more positive charge and the distribution of the charge is more polar. That is consistent with the magnitude of the corresponding dipole moment.

Electrophiles are attracted to free electrons and can react with nucleophiles forming a bond. The reactivity between a nucleophile and an electrophile can be evaluated with the help of Fukui's frontier molecular orbital model [205] and the Klopman-Salem equation [206–208]. With the Klopman-Salem equation the energy gained or lost during the overlap of two reacting orbitals can be calculated [209]. In summary, three major forces play a role: (i) the occupied orbitals of the reactants repel each other; (ii) if the reactants carry opposite charges they will attract each other; and (iii) the occupied orbitals interact with the unoccupied orbitals, while the HOMO-LUMO interaction contributes most. In a nucleophile-electrophile reaction the important orbitals are the HOMO of the nucleophile and the LUMO of the electrophile [209]. The smaller the difference between the HOMO of the nucleophile and the LUMO of the electrophile, the higher the approximated reactivity will be.

The formaldehyde can act as an electrophile and react with a potential nucleophile. The LUMO of formaldehyde decreases in the presence of T<sub>3</sub> (Table 9). This effect is most evident in the T<sub>3</sub>(*anti-anti*) & HCHO complex, i.e. in the complex with the active form of the catalyst. This finding indicates that formaldehyde will be most reactive when complexed with the active form of the catalyst.

## CONCLUSIONS AND OUTLOOK

---

In this thesis I present my contributions facilitating the exploration of molecular energy surfaces: a tool for sampling and a framework for a reduced representation of energy surfaces. I present the applicability of the developed methods for a range of small organic molecules. In the following, the results are briefly reviewed and possible directions for future development are outlined.

### 6.1 SUMMARY

#### 6.1.1 *Fafoom - Flexible algorithm for optimization of molecules*

In Chapter 3, I present Fafoom, an easy-to-use tool that I developed for efficient sampling of molecular energy surfaces. The performance of the genetic algorithm (GA) based search implemented in Fafoom was evaluated for a set of seven amino acid dipeptides and eight drug-like ligands [95]. For all investigated systems the GA search was successful at generating a diverse set of low-energy conformers. In the case of the amino acid dipeptides, the GA search reproduces well the computational reference data [164], and offers an attractive alternative to computationally challenging multi-step search strategies. The results of the investigation of the drug-like ligand structures [165] support the need for a broad PES sampling: conformers closest to the crystallographic receptor-bound structures are often significantly higher in energy than the respective global energy minimum of the isolated molecule. Besides the GA search, Fafoom can be utilized to implement and/or complement alternative, e.g. random or systematic, searches. The performance of different searches that can be performed with Fafoom was evaluated on the example of the mycophenolic acid.

One of the most important features of Fafoom is the possibility to use external molecular simulation packages for the geometry optimization and energy calculation. Currently, four packages, FHI-aims, RDKit, NWChem, and ORCA are supported. Further, Fafoom allows for creating user-specific degrees of freedom. We demonstrated that an extension of Fafoom for handling different puckers of pyranose rings allows for exploring the conformational space of sugars.

#### 6.1.2 *Reduced representation of molecular landscapes*

In Chapter 4, I introduce a framework that I implemented for constructing reduced potential-energy surfaces consisting of the PES minima and selected transitions. In short, the framework comprises the following steps: (i) sampling of the PES; (ii) defining a descriptor for the similarity of the minima; (iii) selecting minima pairs for the calculation of the transition states; (iv) computing transition states and the corresponding energy barriers; (v) drawing

the energy barrier tree. The selection of the similarity descriptor, torsional RMSD, was based the following hypothesis: pairs of conformers that are similar in terms of torsional degrees of freedom have a higher chance to interconvert directly. First, we conducted a proof of concept for alanine tetrapeptide at the force-field level of theory. The resulting potential-energy barrier tree was then compared to reference data obtained from free-energy simulations. Our results, based exclusively on a selection of PES structures, closely match the reference data.

Next, a study at the first-principles level was performed for a synthetic peptide. The resulting energy barrier tree, analyzed together with the molecular structure of the molecule, supported our hypothesis for the minima pairs selection. Furthermore, we showed that there are no high energy barriers between the experimentally observed structures of the synthetic peptide.

### 6.1.3 *Structure-dependent catalytic activity*

The standard view on organocatalysis often neglects the conformational properties of the catalyst itself. In Chapter 5, the relation between the adopted 3D structure and the potential catalytic activity for eight (thio)urea derivatives is presented. In the first part of the study, I investigated conformational preferences of the isolated molecules at different levels of theory. The conformational preferences of the urea and thiourea molecules vary and depend on the substituents and solvent. Furthermore, we showed that the experimentally confirmed catalytically active conformer of **T3**, known as the Schreiner catalyst, is less stable than the inactive structures of **T3**. In the second part of the study, we investigated the relation between the **T3** catalyst and a model substrate, formaldehyde. The conformational preferences of **T3** do not change in the presence of formaldehyde, i.e. the catalytically active form is less stable than the inactive forms. Finally, we investigated the potential reactivity of the model substrate. We found that the LUMO of the model substrate is lowered most in a complex with the active form of the catalyst. This translates to a higher reactivity if exposed to a potential nucleophilic partner, as this would decrease the energy difference between the frontier molecular orbitals of the reaction partners.

## 6.2 CONCLUSIONS

A number of initiatives have been started recently towards collecting and sharing molecular and materials simulation data [e.g. 210–212]. Moreover, the focus of the NOMAD Center of Excellence [213] is to provide a set of analytic tools along with the data and, in the long-term, to establish a Materials Encyclopedia. The collected and analyzed data is typically generated in automated high-throughput studies. However, the quality of the final data remains dependent on the accuracy of the underlying model.

Based on the results of this thesis, I formulate the following conclusions that address the challenges that are crucial for obtaining accurate results in modeling flexible organic molecules:

- A broad sampling of the energy surface of flexible organic molecules is mandatory. Focusing only on the global minimum might be misleading

because: (i) a number of favorable conformers with a competing energies can exist; (ii) identifying the global minimum is very sensitive to the applied level of theory; (iii) the global minimum can be useless for meaningful predictions of properties.

- The selection and curation of the degrees of freedom significantly impacts the success of the search. In the case of flexible organic molecules, it is often sufficient to consider the torsional degrees of freedom for the global sampling. However, it can be beneficial to: (i) constrain this selection based on chemical knowledge, and allow only a subset of values to be adopted; (ii) expand the selection by, e.g. considering X-X-O-H torsions as rotatable in systems where hydrogen bonding matters; or (iii) defining additional degrees of freedom, e.g. ring puckers.

### 6.3 OUTLOOK

The modular design of Fafoom facilitates the addition of new functionalities. The development of Fafoom currently advances in two directions: (i) wrappers for further simulation packages, e.g. DFTB+ [214], are being developed and (ii) new degrees of freedom, e.g. orientation of the molecule, are being implemented [215]. The targeted ability of Fafoom to handle orientation of a molecule can in the future be applied to searches for systems with more than one molecule, e.g. a ligand in a docking site [5], dimers, or catalyst-substrate complexes.

In the current version of Fafoom, the objective function is the potential energy of the molecule. In order to increase the flexibility of Fafoom, the objective function could also be user-defined. This would be helpful in searches for conformations where a specific property, e.g. the LUMO energy, needs to be optimized. In addition, a flexible definition of the objective function would allow filtering out structures that do not satisfy constraints, e.g. that do not match experimentally determined collisional cross sections (CCS).

The long-term development plan for Fafoom aims at enabling searches across chemical space. This could be initiated by, e.g. considering different structural isomers of the input molecule. The main challenge in this approach is definition of the objective function, as the relative energy cannot be used for comparison of different molecules.

The next step in the development of the framework for reducing the dimensionality of the energy surfaces would be to construct a model of the dynamics of the system. The required conformational transition rates can be calculated from the values of the energy barriers between the states. The parametrized model can then be used to perform stochastic dynamics simulations capturing the time evolution of the system. Such simulations could be used to, e.g. estimate how long certain conformations can be kinetically trapped.

A natural continuation of the study of structure-dependent catalytic activity would be to simulate the catalytic reaction. This would involve the following reaction intermediates: the catalyst with substrates, transition state, and the catalyst with product together with the corresponding interconversions. With this, the energy barriers could be obtained. One of the main challenges in the simulation of such a reaction is to obtain a stable initial state, i.e. a



complex of the catalyst together with two substrates. The search for such an energetically favorable complex could be facilitated by the ability of Fafoom to optimize complexes of molecules. An ultimate goal of the project is to use the information obtained from the PES investigations to define sensible starting points and collective variables for biased MD simulations. With this, a more realistic simulation of a system where dynamics and temperature play a role could be approached.



## APPENDIX

## A.1 FAFOOM PARAMETER

## MOLECULE SETTINGS

<b>SMILES</b>	simplified one-line notation of the molecule
<b>optimize_torsion</b> (default = True)	if set to True, rotatable bonds will be among the optimized degrees of freedom
<b>optimize_cistrans</b> (default = False)	if set to True, <i>cis/trans</i> bonds will be among the optimized degrees of freedom
<b>optimize_pyranosering</b> (default = False)	if set to True, the pyranose ring will be among the optimized degrees of freedom
<b>smart_torsion</b> , default="[*]~[\$(##*)&!D1]-&!@[!\$(##*)&!D1]~[*]"	pattern for searching for rotatable bonds
<b>smart_cistrans</b>	pattern for searching for <i>cis/trans</i> bonds
<b>filter_smart_torsion</b>	pattern for sorting out rotatable bonds, that should be ignored
<b>list_of_torsion</b> , <b>list_of_cistrans</b> , <b>list_of_pyranosering</b>	the location of degrees of freedom can be passed directly as lists of tuples of atom indices
<b>distance_cutoff_1</b> (default = 1.3 Å)	if two non-bonded atoms are closer to each other than <b>distance_cutoff_1</b> (Å) the structure will not pass the geometry check.
<b>distance_cutoff_2</b> (default = 2.15 Å)	if two bonded atoms are further from other than <b>distance_cutoff_2</b> (Å) the structure will not pass the geometry check.
<b>rmsd_type</b> (default = "cartesian")	option for similarity evaluation
<b>rmsd_cutoff_uniq</b> (default = 0.2 Å)	threshold for the evaluation of the similarity of the structures
<b>chiral</b> (default = True)	if set to False, mirrors of the 3D structures will be evaluated
<b>weights_torsion</b> , <b>weights_cistrans</b> , <b>weights_pyranosering</b>	weights can be assigned to the value options of the degrees of freedom. E.g.: For the <i>cis/trans</i> bonds, if one wants to generate and optimize conformations only with <i>trans</i> bond (i.e. equal to 180°), it is possible with: <b>weights_cistrans</b> = [0., 1.]

## GA SETTINGS

<b>popsize</b> (default = 10)	number of structures in the population
<b>energy_var</b> (default = 0.001 eV)	if the difference between the highest and lowest energy within the population is lower than the <b>energy_var</b> , all the individuals will be assigned the same fitness of 1.0.
<b>selection</b> (default="roulette_wheel")	options for the selection mechanisms of the individuals. Another options are random and roulette_wheel_reverse.
<b>fitness_sum_limit</b> (default = 1.2)	if the sum of the fitness values for all individuals is lower than this threshold the selection will be conducted independently from the chosen mechanism. The best and a random individual will be selected.
<b>prob_for_crossing</b> (default = 1.0)	probability for the crossing over.
<b>prob_for_mut_torsion,</b> <b>prob_for_mut_cistrans,</b> <b>prob_for_mut_pyranosering</b>	probability for a mutation in torsions/ <i>cis/trans</i> bonds/ pyranose rings
<b>max_mutations_torsions,</b> <b>max_mutations_cistrans,</b> <b>max_mutations_pyranosering</b>	Maximal number of mutations for torsions/ <i>cis/trans</i> bonds/ pyranose rings.

## RUN SETTINGS

<b>energy_function</b>	Name of the software/method to be used for the energy evaluations. Currently supported options are: FHI-aims, NWChem, RDKit and ORCA.
<b>max_iter</b> (default = 30)	Number of iterations that will be performed after the initialization is finished.
<b>iter_limit_conv</b> (default = 20)	Minimal number of iterations to be performed before any convergence criteria are checked.
<b>energy_diff_conv</b> (default = 0.001 eV)	Parameter for checking the convergence. If the lowest energy hasn't change by more than <b>energy_diff_conv</b> (eV) after <b>iter_limit_conv</b> iterations, the GA-run is considered to be converged. Attention: convergence doesn't necessarily mean that the global minimum was found.
<b>energy_wanted</b>	If the energy of the global minimum is known it can also be used for checking if the convergence is achieved.
<b>sourcedir</b>	Name of your directory with control.in file.
<b>aims_call</b>	String for execution of FHI-aims.

## A.2 FAFOOM BENCHMARK: PARAMETER LISTS

Table 10: GA parameters for isoleucine dipeptide

	Parameter	Value
Molecule	SMILES	<chem>CC(=O)N[C@H](C(=O)NC)[C@H](CC)C</chem>
	distance_cutoff_1	1.2 Å
	distance_cutoff_2	2.0 Å
	rmsd_cutoff_uniq	0.2 Å
	chiral	True
Run settings	max_iter	10
	iter_limit_conv	10
	energy_diff_conv	0.001 eV
GA settings	popsize	5
	energy_var	0.001 eV
	selection	roulette wheel
	fitness_sum_limit	1.2
	prob_for_crossing	0.95
	cross_trial	20
	prob_for_mut_cistrans	0.5
	prob_for_mut_rot	0.5
	max_mutations_cistrans	1
	max_mutations_torsions	2
mut_trial	100	

Table 11: GA parameters for the drug-like ligands

	Parameter	Value
Molecule	distance_cutoff_1	1.2 Å
	distance_cutoff_2	2.15 Å
	rmsd_cutoff_uniq	0.2 Å
	chiral	True
Run settings	max_iter	30
	iter_limit_conv	20
	energy_diff_conv	0.001 eV
GA settings	popsize	10
	energy_var	0.001 eV
	selection	roulette wheel
	fitness_sum_limit	1.2
	prob_for_crossing	0.95
	cross_trial	100
	prob_for_mut_cistrans	1.0
	prob_for_mut_rot	0.8
	max_mutations_cistrans	1
	max_mutations_torsions	3
mut_trial	100	

## A.3 (THIO)UREA MOLECULES: IMPACT OF DIFFERENT DISPERSION CORRECTIONS

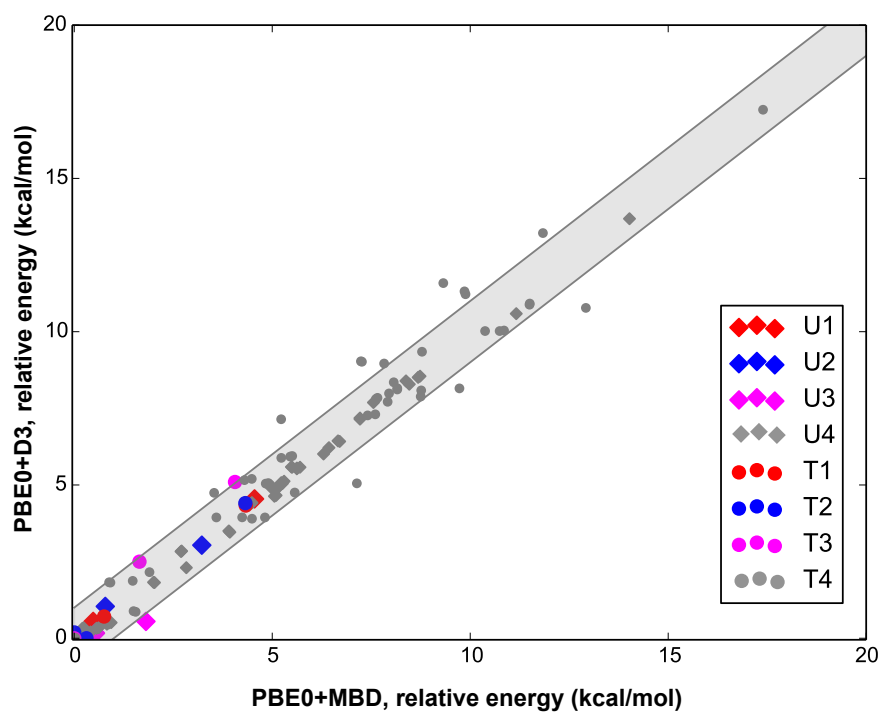


Figure 41: The relative energy of the (thio)urea molecules calculated with PBE0+MBD (x-axis) versus PBE0+D3 (y-axis). The points within the shaded area correspond to energy differences lower than 1 kcal/mol.

## SELBSTÄNDIGKEITSERKLÄRUNG

---

Hiermit versichere ich, alle Hilfsmittel und Hilfen angegeben zu haben und auf dieser Grundlage die Arbeit selbstständig verfasst zu haben. Die Arbeit wurde nicht schon einmal in einem früheren Promotionsverfahren angenommen oder als ungenügend beurteilt.

*Berlin, den 30. Juni 2016*

---

Adriana Supady



## CURRICULUM VITAE

---

For reasons of data protection, the curriculum vitae is not included in the online version.





## PUBLICATIONS

---

Publication related to this thesis:

- Supady A., Blum V. and Baldauf C. First-principles molecular structure search with a genetic algorithm. *J. Chem. Inf. Model.*, **2015**, 55, 2338-2348.

Publications concerning other topics:

- Supady A., Klipp E. and Barberis M. A variable fork rate affects timing of origin firing and S phase dynamics in *Saccharomyces cerevisiae*. *J. Biotechnol.*, **2013**, 168(2), 174-84.
- Hoffman-Sommer M., Supady A. and Klipp E. Cell-to-cell communication circuits: Quantitative analysis of synthetic logic gates. *Front. Physio.* **2012**, 3, 287.



## ACKNOWLEDGMENTS

---

First of all I thank my supervisor, Carsten Baldauf, for his guidance, ideas and invaluable support. Thank you for your patience, feedback and suggestions.

I thank Matthias Scheffler for giving me the opportunity to be a PhD student at the Theory Department of the Fritz Haber Institute and for refereeing this thesis. I also thank Roland Netz for being the second referee.

I thank Volker Blum, Luca Ghiringhelli and Stefan Hecht for sharing instructive insights, many inspiring discussions, and fruitful collaboration.

I am very grateful to my office mates: Franziska, for her patience in answering my questions and the invaluable support since my first days at the FHI; and to Arvid, for explaining the mysteries of scientific computing and for many valuable hints. Big thanks to Markus, Mateusz, and Teresa for the effective teamwork and for making the Biogroup such a stimulating environment. Further, I thank Norina, Lydia, Tanja, Björn, Johannes, Franz, Wael, and all my present and former colleagues from the Theory Department for the motivational discussions and all the attended/co-organized conferences, workshops and events.

A special thanks goes to Birgit, Julia, Hanna, and Steffen for their help and encouraging and friendly chats.

Finally I thank my friends and my family, especially my parents, Marianna and Grzegorz, and my sister, Dominika, for their support and encouragement. I thank Kurt for always being there for me and for his unconditional support.



## BIBLIOGRAPHY

---

- [1] Alan R. Katritzky, Minati Kuanar, Svetoslav Slavov, C. Dennis Hall, Mati Karelson, Iris Kahn, and Dimitar A. Dobchev. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chem. Rev.*, 110(10):5714–5789, 2010. doi: 10.1021/cr900238d.
- [2] Martin Karplus. Development of multiscale models for complex chemical systems: from H+H<sub>2</sub> to biomolecules (Nobel Lecture). *Angew. Chem. Int. Ed.*, 53(38):9992–10005, 2014. doi: 10.1002/anie.201403924.
- [3] Gareth Jones, Peter Willett, Robert C. Glen, Andrew R. Leach, and Robin Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267(3):727–748, 1997. doi: 10.1006/jmbi.1996.0897.
- [4] Garrett M. Morris, David S. Goodsell, Robert S. Halliday, Ruth Huey, William E. Hart, Richard K. Belew, and Arthur J. Olson. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, 19(14):1639–1662, 1998. doi: 10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B.
- [5] René Meier, Martin Pippel, Frank Brandt, Wolfgang Sippl, and Carsten Baldauf. ParaDockS: a framework for molecular docking with population-based metaheuristics. *J. Chem. Inf. Model.*, 50(5):879–889, 2010. doi: 10.1021/ci900467x.
- [6] Carsten Baldauf and Mariana Rossi. Going clean: structure and dynamics of peptides in the gas phase and paths to solvation. *J. Phys. Condens. Matter*, 27(49):493002, 2015. doi: 10.1088/0953-8984/27/49/493002.
- [7] Scott M. Woodley and Richard Catlow. Crystal structure prediction from first principles. *Nature Mater.*, 7(12):937–946, 2008. doi: 10.1038/nmat2321.
- [8] Xunhua Zhao, Xiang Shao, Yuichi Fujimori, Saswata Bhattacharya, Luca M. Ghiringhelli, Hans-Joachim Freund, Martin Sterrer, Niklas Nilius, and Sergey V. Levchenko. Formation of Water Chains on CaO(001): What Drives the 1D Growth? *J. Phys. Chem. Lett.*, 6(7):1204–1208, 2015. doi: 10.1021/acs.jpcclett.5b00223.
- [9] Elizabeth C. Beret, Luca M. Ghiringhelli, and Matthias Scheffler. Free gold clusters: beyond the static, monostructure description. *Farad. Discuss.*, 152:153, 2011. doi: 10.1039/c1fd00027f.
- [10] World Community Grid. URL <http://www.worldcommunitygrid.org/>.
- [11] David J. Wales. Perspective: Insight into reaction coordinates and dynamics from the potential energy landscape. *J. Chem. Phys.*, 142(13):130901, 2015. doi: 10.1063/1.4916307.

- [12] Frank Noé and Stefan Fischer. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.*, 18(2):154–162, 2008. doi: 10.1016/j.sbi.2008.01.008.
- [13] Oren M. Becker and Martin Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.*, 106(4):1495, 1997. doi: 10.1063/1.473299.
- [14] Christoph Flamm, Ivo L. Hofacker, Peter F. Stadler, and Michael T. Wolfinger. Barrier Trees of Degenerate Landscapes. *Z. Phys. Chem.*, 216(2/2002):155, 2002. doi: 10.1524/zpch.2002.216.2.155.
- [15] David J. Wales. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. Cambridge University Press, Cambridge, 2003. ISBN 0521814154.
- [16] Carsten Baldauf, Kevin Pagel, Stephan Warnke, Gert von Helden, Beate Koksich, Volker Blum, and Matthias Scheffler. How cations change peptide structure. *Chem.-Eur. J.*, 19(34):11224–11234, 2013. doi: 10.1002/chem.201204554.
- [17] Mariana Rossi, Sucismita Chutia, Matthias Scheffler, and Volker Blum. Validation Challenge of Density-Functional Theory for Peptides-Example of Ac-Phe-Ala<sub>5</sub>-LysH<sup>+</sup>. *J. Phys. Chem. A*, 118(35):7349–7359, 2014. doi: 10.1021/jp412055r.
- [18] Robert Sedlak, Tomasz Janowski, Michal Pitoňák, Jan Řezáč, Peter Pulay, and Pavel Hobza. The accuracy of quantum chemical methods for large noncovalent complexes. *J. Chem. Theory Comput.*, 9(8):3364–3374, 2013. doi: 10.1021/ct400036b.
- [19] Qin Wu and Weitao Yang. Empirical correction to density functional theory for van der Waals interactions. *J. Chem. Phys.*, 116(2):515, 2002. doi: 10.1063/1.1424928.
- [20] Alexandre Tkatchenko, Mariana Rossi, Volker Blum, Joel Ireta, and Matthias Scheffler. Unraveling the Stability of Polypeptide Helices: Critical Role of van der Waals Interactions. *Phys. Rev. Lett.*, 106(11):118102, 2011. doi: 10.1103/PhysRevLett.106.118102.
- [21] J. Philipp Wagner and Peter R. Schreiner. London Dispersion in Molecular Chemistry-Reconsidering Steric Effects. *Angew. Chem. Int. Ed.*, 54(42):12274–12296, 2015. doi: 10.1002/anie.201503476.
- [22] Holger Kruse, Lars Goerigk, and Stefan Grimme. Why the standard B<sub>3</sub>LYP/6-31G\* model chemistry should not be used in DFT calculations of molecular thermochemistry: understanding and correcting the problem. *J. Org. Chem.*, 77(23):10824–10834, 2012. doi: 10.1021/jo302156p.
- [23] Franziska Schubert, Mariana Rossi, Carsten Baldauf, Kevin Pagel, Stephan Warnke, Gert von Helden, Frank Filsinger, Peter Kupser, Gerard Meijer, Mario Salwiczek, Beate Koksich, Matthias Scheffler, and Volker Blum. Exploring the conformational preferences of 20-residue peptides in isolation: Ac-Ala<sub>19</sub>-Lys+H<sup>+</sup> vs. Ac-Lys-Ala<sub>19</sub>+H<sup>+</sup> and the

- current reach of DFT. *Phys. Chem. Chem. Phys.*, 2015. doi: 10.1039/C4CP05541A.
- [24] Johannes Kirchmair, Christian Laggner, Gerhard Wolber, and Thierry Langer. Comparative analysis of protein-bound ligand conformations with respect to catalyst's conformational space subsampling algorithms. *J. Chem. Inf. Model.*, 45(2):422–430, 2005. doi: 10.1021/ci049753l.
- [25] Dimitris K. Agrafiotis, Alan C. Gibbs, Fangqiang Zhu, Sergei Izrailev, and Eric Martin. Conformational sampling of bioactive molecules: a comparative study. *J. Chem. Inf. Model.*, 47(3):1067–1086, 2007. doi: 10.1021/ci6005454.
- [26] Fafoom - flexible algorithm for optimization of molecules, . URL <https://github.com/adrianasupady/fafoom>.
- [27] Wolfram Koch and Max C. Holthausen. *A Chemist's Guide to Density Functional Theory*. Wiley-VCH, Weinheim, 2001. ISBN 978-3-527-30372-4.
- [28] Bernd Hartke, lecture scripts. URL <https://ravel.pctc.uni-kiel.de/teaching/>.
- [29] Mariana Rossi. *Ab initio study of alanine-based polypeptides secondary-structure motifs in the gas phase*. PhD thesis, TU Berlin and Fritz-Haber-Institut der Max-Planck-Gesellschaft, 2011.
- [30] Franziska Schubert. *Conformational equilibria and spectroscopy of gas-phase homologous peptides from first principles*. PhD thesis, FU Berlin and Fritz-Haber-Institut der Max-Planck-Gesellschaft, 2014.
- [31] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005.
- [32] Noel M. O'Boyle. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *J. Cheminform.*, 4(1):22, 2012. doi: 10.1186/1758-2946-4-22.
- [33] Stephen R. Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.*, 7(1):23, 2015. doi: 10.1186/s13321-015-0068-4.
- [34] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. Sect. A*, 32(5):922–923, 1976. doi: 10.1107/S0567739476001873.
- [35] Daylight tutorials: SMILES, SMARTS and SMIRKS. URL [http://www.daylight.com/dayhtml\\_tutorials/index.html](http://www.daylight.com/dayhtml_tutorials/index.html).
- [36] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open Babel: An open chemical toolbox. *J. Cheminform.*, 3(1):33, 2011. doi: 10.1186/1758-2946-3-33.

- [37] RDKit: Cheminformatics and Machine Learning Software. URL <http://www.rdkit.org/>.
- [38] Marcus D. Hanwell, Donald E. Curtis, David C. Lonie, Tim Vandermeersch, Eva Zurek, and Geoffrey R. Hutchison. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminform.*, 4(1):17, 2012. doi: 10.1186/1758-2946-4-17.
- [39] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual molecular dynamics. *J. Mol. Graphics*, 14(1):33–38, 1996. doi: 10.1016/0263-7855(96)00018-5.
- [40] Jmol: an open-source Java viewer for chemical structures in 3D. URL <http://www.jmol.org/>.
- [41] Petr Jurecka, Jirí Sponer, Jirí Cerný, and Pavel Hobza. Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys. Chem. Chem. Phys.*, 8(17):1985–1993, 2006. doi: 10.1039/b600027d.
- [42] Thomas A. Halgren. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.*, 17(5–6):490–519, 1996.
- [43] William L. Jorgensen, David S. Maxwell, and Julian Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, 1996. doi: 10.1021/ja9621760.
- [44] Scott J. Weiner, Peter A. Kollman, David A. Case, U. Chandra Singh, Caterina Ghio, Giuliano Alagona, Salvatore Profeta, and Paul Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106(3):765–784, 1984.
- [45] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197, 1995. doi: 10.1021/ja00124a002.
- [46] Karl N. Kirschner, Austin B. Yongye, Sarah M. Tschampel, Jorge González-Outeiriño, Charlisa R. Daniels, B. Lachele Foley, and Robert J. Woods. GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem.*, 29(4):622–655, 2008. doi: 10.1002/jcc.20820.
- [47] Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4(2):187–217, 1983. doi: 10.1002/jcc.540040211.
- [48] Chris Oostenbrink, Alessandra Villa, Alan E. Mark, and Wilfred F. van Gunsteren. A biomolecular force field based on the free enthalpy of



- hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.*, 25(13):1656–1676, 2004. doi: 10.1002/jcc.20090.
- [49] Jay W. Ponder, Chuanjie Wu, Pengyu Ren, Vijay S. Pande, John D. Chodera, Michael J. Schnieders, Imran Haque, David L. Mobley, Daniel S. Lambrecht, Robert A. DiStasio, Martin Head-Gordon, Gary N. I. Clark, Margaret E. Johnson, and Teresa Head-Gordon. Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B*, 114(8):2549–2564, 2010. doi: 10.1021/jp910674d.
- [50] Hannah H. Aogy-David and Hanoch Senderowitz. Toward Focusing Conformational Ensembles on Bioactive Conformations: A Molecular Mechanics/Quantum Mechanics Study. *J. Chem. Inf. Model.*, 55(10):2154–2167, 2015. doi: 10.1021/acs.jcim.5b00259.
- [51] Francesca Vitalini, Antonia S. J. S. Mey, Frank Noé, and Bettina G. Keller. Dynamic properties of force fields. *J. Chem. Phys.*, 142(8):084101, 2015. doi: 10.1063/1.4909549.
- [52] Max Born and Robert Oppenheimer. Zur Quantentheorie der Molekeln. *Ann. Phys. (Berlin)*, 389(20):457–484, 1927. doi: 10.1002/andp.19273892002.
- [53] Douglas R. Hartree. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part II. Some Results and Discussion. *Math. Proc. Cambridge Phil. Soc.*, 24(1):111–132, 1928.
- [54] John C. Slater. Note on Hartree’s Method. *Phys. Rev.*, 35(2):210–211, 1930.
- [55] Vladimir Fock. Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Z. Phys.*, 61(1-2):126–148, 1930. doi: 10.1007/BF01340294.
- [56] Christian Møller and Milton S. Plesset. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.*, 46(7):618–622, 1934. doi: 10.1103/PhysRev.46.618.
- [57] Jiří Čížek. On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods. *J. Chem. Phys.*, 45(11):4256, 1966. doi: 10.1063/1.1727484.
- [58] Rodney J. Bartlett and Monika Musiał. Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys.*, 79(1):291–352, 2007. doi: 10.1103/RevModPhys.79.291.
- [59] Varun Rishi, Ajith Perera, and Rodney J. Bartlett. Assessing the distinguishable cluster approximation based on the triple bond-breaking in the nitrogen molecule. *J. Chem. Phys.*, 144(12):124117, 2016. ISSN 0021-9606. doi: 10.1063/1.4944087.

- [60] Kevin E. Riley, Michal Pitonák, Petr Jurecka, and Pavel Hobza. Stabilization and structure calculations for noncovalent interactions in extended molecular systems based on wave function and density functional theories. *Chem. Rev.*, 110(9):5023–5063, 2010. doi: 10.1021/cr1000173.
- [61] Christoph Riplinger and Frank Neese. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.*, 138(3):034106, 2013. doi: 10.1063/1.4773581.
- [62] Dimitrios G. Liakos, Manuel Sparta, Manoj K. Kesharwani, Jan M. L. Martin, and Frank Neese. Exploring the Accuracy Limits of Local Pair Natural Orbital Coupled-Cluster Theory. *J. Chem. Theory Comput.*, 11(4):1525–1539, 2015. doi: 10.1021/ct501129s.
- [63] Pierre Hohenberg and Walter Kohn. Inhomogeneous Electron Gas. *Phys. Rev.*, 136(3B):B864–B871, 1964. doi: 10.1103/PhysRev.136.B864.
- [64] Walter Kohn and Lu J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.*, 140(4A):A1133–A1138, 1965. doi: 10.1103/PhysRev.140.A1133.
- [65] John P. Perdew and Karla Schmidt. Jacob’s ladder of density functional approximations for the exchange-correlation energy. *AIP Conf. Proc.*, 577(1):1–20, 2001. doi: 10.1063/1.1390175.
- [66] John P. Perdew and Adrienn Ruzsinszky. Fourteen easy lessons in density functional theory. *Int. J. Quantum Chem.*, 110(May):2801–2807, 2010. doi: 10.1002/qua.22829.
- [67] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.*, 77(18):3865–3868, 1996. doi: 10.1103/PhysRevLett.77.3865.
- [68] Carlo Adamo and Vincenzo Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.*, 110(13):6158, 1999. doi: 10.1063/1.478522.
- [69] David Bohm and David Pines. A Collective Description of Electron Interactions: III. Coulomb Interactions in a Degenerate Electron Gas. *Phys. Rev.*, 92(3):609–625, 1953. doi: 10.1103/PhysRev.92.609.
- [70] Xinguo Ren, Patrick Rinke, Christian Joas, and Matthias Scheffler. Random-phase approximation and its applications in computational chemistry and materials science. *J. Mater. Sci.*, 47(21):7447–7471, 2012. doi: 10.1007/s10853-012-6570-4.
- [71] Anthony M. Reilly and Alexandre Tkatchenko. Van der Waals dispersion interactions in molecular materials: beyond pairwise additivity. *Chem. Sci.*, 6(6):3289–3301, 2015. doi: 10.1039/C5SC00410A.
- [72] Maxime Dion, Henrik Rydberg, Elsebeth Schröder, David C. Langreth, and Bengt I. Lundqvist. Van der Waals density functional for general geometries. *Phys. Rev. Lett.*, 92(24):246401, 2004. doi: 10.1103/PhysRevLett.92.246401.

- [73] Yan Zhao and Donald G. Truhlar. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.*, 125(19):194101, 2006. doi: 10.1063/1.2370993.
- [74] Jiří Klimeš and Angelos Michaelides. Perspective: Advances and challenges in treating van der Waals dispersion forces in density functional theory. *J. Chem. Phys.*, 137(12):120901, 2012. doi: 10.1063/1.4754130.
- [75] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.*, 132(15):154104, 2010. doi: 10.1063/1.3382344.
- [76] Alexandre Tkatchenko and Matthias Scheffler. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.*, 102(7):073005, 2009. doi: 10.1103/PhysRevLett.102.073005.
- [77] Noa Marom, Alexandre Tkatchenko, Mariana Rossi, Vivekanand V. Gore, Oded Hod, Matthias Scheffler, and Leeor Kronik. Dispersion Interactions with Density-Functional Theory: Benchmarking Semiempirical and Interatomic Pairwise Corrected Density Functionals. *J. Chem. Theory Comput.*, 7(12):3944–3951, 2011. doi: 10.1021/ct2005616.
- [78] Alexandre Tkatchenko, Robert A. DiStasio, Roberto Car, and Matthias Scheffler. Accurate and Efficient Method for Many-Body van der Waals Interactions. *Phys. Rev. Lett.*, 108(23):236402, 2012. doi: 10.1103/PhysRevLett.108.236402.
- [79] Alberto Ambrosetti, Anthony M. Reilly, Robert A. DiStasio, and Alexandre Tkatchenko. Long-range correlation energy calculated from coupled atomic response functions. *J. Chem. Phys.*, 140(18):18A508, 2014. doi: 10.1063/1.4865104.
- [80] Volker Blum, Ralf Gehrke, Felix Hanke, Paula Havu, Ville Havu, Xinguo Ren, Karsten Reuter, and Matthias Scheffler. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.*, 180(11):2175–2196, 2009. doi: 10.1016/j.cpc.2009.06.022.
- [81] Frank Neese. The ORCA program system. *WIREs: Comput. Mol. Sci.*, 2(1):73–78, 2012. doi: 10.1002/wcms.81.
- [82] Jacopo Tomasi, Benedetta Mennucci, and Roberto Cammi. Quantum mechanical continuum solvation models. *Chem. Rev.*, 105(8):2999–3093, 2005. doi: 10.1021/cr9904009.
- [83] Andreas Klamt and Gerrit Schüürmann. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2*, (5):799, 1993. doi: 10.1039/p29930000799.
- [84] H. Bernhard Schlegel. Geometry optimization. *WIREs: Comput. Mol. Sci.*, 1(5):790–809, 2011. doi: 10.1002/wcms.34.

- [85] John A. Nelder and Roger Mead. A Simplex Method for Function Minimization. *Comput. J.*, 7(4):308–313, 1965. doi: 10.1093/comjnl/7.4.308.
- [86] Charles G. Broyden. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA J. Appl. Math.*, 6(1):76–90, 1970. doi: 10.1093/imamat/6.1.76.
- [87] Roger Fletcher. New approach to variable metric algorithms. 13(3): 317–322, 1970.
- [88] Donald Goldfarb. A family of variable metric methods derived by variational means. 24:23–26, 1970.
- [89] David F. Shanno. Conditioning of quasi-Newton methods for function minimization. 24:647–656, 1970.
- [90] Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. Springer New York, 2006.
- [91] Roger Fletcher. *Practical Methods of Optimization*. John Wiley & Sons, New York, 1987.
- [92] H. Bernhard Schlegel. Estimating the hessian for gradient-type geometry optimizations. *Theor. Chim. Acta.*, 66(5):333–340, 1984. doi: 10.1007/BF00554788.
- [93] Thomas H. Fischer and Jan Almlöf. General methods for geometry and wave function optimization. *J. Phys. Chem.*, 96(24):9768–9774, 1992.
- [94] Roland Lindh, Anders Bernhardsson, Gunnar Karlström, and Per-Åke Malmqvist. On the use of a Hessian model function in molecular geometry optimizations. *Chem. Phys. Lett.*, 241(4):423–428, 1995. doi: 10.1016/0009-2614(95)00646-L.
- [95] Adriana Supady, Volker Blum, and Carsten Baldauf. First-Principles Molecular Structure Search with a Genetic Algorithm. *J. Chem. Inf. Model.*, 2015. doi: 10.1021/acs.jcim.5b00243.
- [96] Jiabo Li, Tedman Ehlers, Jon Sutter, Shikha Varma-O'Brien, and Johannes Kirchmair. CAESAR: a new conformer generation algorithm based on recursive buildup and local rotational symmetry consideration. *J. Chem. Inf. Model.*, 47(5):1923–1932, 2007. doi: 10.1021/ci700136x.
- [97] Noel M. O'Boyle, Tim Vandermeersch, Christopher J. Flynn, Anita R. Maguire, and Geoffrey R. Hutchison. Confab - Systematic generation of diverse low-energy conformers. *J. Cheminform.*, 3(1):8, 2011. doi: 10.1186/1758-2946-3-8.
- [98] Fariborz Mohamadi, Nigel G. J. Richards, Wayne C. Guida, Rob Liskamp, Mark Lipton, Craig Caufield, George Chang, Thomas Hendrickson, and W. Clark Still. Macromodel - An integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J. Comput. Chem.*, 11(4):440–467, 1990. doi: 10.1002/jcc.540110405.

- [99] MOE (*Molecular Operating Environment*). Chemical Computing Group, Inc.: Montreal, Canada, 2008.
- [100] Javier Klett, Alvaro Cortés-Cabrera, Rubén Gil-Redondo, Federico Gago, and Antonio Morreale. ALFA: Automatic Ligand Flexibility Assignment. *J. Chem. Inf. Model.*, 54(1):314–323, 2014. doi: 10.1021/ci400453n.
- [101] Christin Schärfer, Tanja Schulz-Gasch, Jérôme Hert, Lennart Heinzerling, Benjamin Schulz, Therese Inhester, Martin Stahl, and Matthias Rarey. CONFECT: Conformations from an Expert Collection of Torsion Patterns. *ChemMedChem*, pages 1690–1700, 2013. doi: 10.1002/cmdc.201300242.
- [102] Jens Sadowski, Johann Gasteiger, and Gerhard Klebe. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.*, 34(4):1000–1008, 1994. doi: 10.1021/ci00020a039.
- [103] Steffen Renner, Christof H. Schwab, Johann Gasteiger, and Gisbert Schneider. Impact of conformational flexibility on three-dimensional similarity searching using correlation vectors. *J. Chem. Inf. Model.*, 46(6):2324–2332, 2006. doi: 10.1021/ci050075s.
- [104] Alessio Andronico, Arlo Randall, Ryan W. Benz, and Pierre Baldi. Data-driven high-throughput prediction of the 3-D structure of small molecules: review and progress. *J. Chem. Inf. Model.*, 51(4):760–776, 2011. doi: 10.1021/ci100223t.
- [105] Peter Sadowski and Pierre Baldi. Small-molecule 3D structure prediction using open crystallography data. *J. Chem. Inf. Model.*, 53(12):3127–3130, 2013. doi: 10.1021/ci4005282.
- [106] Paul C. D. Hawkins, A. Geoffrey Skillman, Gregory L. Warren, Benjamin A. Ellingson, and Matthew T. Stahl. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.*, 50(4):572–584, 2010. doi: 10.1021/ci100031x.
- [107] Mikko J. Vainio and Mark S. Johnson. Generating conformer ensembles using a multiobjective genetic algorithm. *J. Chem. Inf. Model.*, 47(6):2462–2474, 2007. doi: 10.1021/ci6005646.
- [108] Xiaofeng Liu, Fang Bai, Sisheng Ouyang, Xicheng Wang, Honglin Li, and Hualiang Jiang. Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation. *BMC Bioinf.*, 10(1):101, 2009. doi: 10.1186/1471-2105-10-101.
- [109] David J. Wales and Jonathan P. K. Doye. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *J. Phys. Chem. A*, 101(28):5111–5116, 1997. doi: 10.1021/jp970984n.

- [110] Stefan Goedecker. Minima hopping: an efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.*, 120(21):9911–9917, 2004. doi: 10.1063/1.1724816.
- [111] Sune R. Bahn and Karsten W. Jacobsen. An object-oriented scripting interface to a legacy electronic structure code. *Comput. Sci. Eng.*, 4(3): 56–66, 2002. doi: 10.1109/5992.998641.
- [112] David J Wales. GMIN: A program for finding global minima and calculating thermodynamic properties from basin-sampling. URL <http://www-wales.ch.cam.ac.uk/GMIN/>.
- [113] Jay W Ponder. Tinker - Software Tools for Molecular Design. URL <http://dasher.wustl.edu/tinker/>.
- [114] John H. Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press, Ann Arbor, MI, 1975. ISBN 0472084607.
- [115] David B. Fogel, editor. *Evolutionary Computation: The Fossil Record*. IEEE Press, Piscataway, NJ, 1998.
- [116] David E. Goldberg. *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley, Reading, MA, 1989.
- [117] Nikhil Nair and Jonathan M. Goodman. Genetic Algorithms in Conformational Analysis. *J. Chem. Inf. Comput. Sci.*, 38(2):317–320, 1998. doi: 10.1021/ci970433u.
- [118] Martin Damsbo, Brian S. Kinnear, Matthew R. Hartings, Peder T. Ruhoff, Martin F. Jarrold, and Mark A. Ratner. Application of evolutionary algorithm methods to polypeptide folding: comparison with experimental results for unsolvated Ac-(Ala-Gly-Gly)<sub>5</sub>-LysH<sup>+</sup>. *Proc. Natl. Acad. Sci. U. S. A.*, 101(19):7215–7222, 2004. doi: 10.1073/pnas.0401659101.
- [119] Niss O. Carstensen, Johannes M. Dieterich, and Bernd Hartke. Design of optimally switchable molecules by genetic algorithms. *Phys. Chem. Chem. Phys.*, 13(7):2903–2910, 2011. doi: 10.1039/c0cp01065k.
- [120] Silvia Carlotto, Laura Orian, and Antonino Polimeno. Heuristic approaches to the optimization of acceptor systems in bulk heterojunction cells: a computational study. *Theor. Chem. Acc.*, 131(3):1191, 2012. doi: 10.1007/s00214-012-1191-1.
- [121] Bernd Hartke. Global geometry optimization of clusters using genetic algorithms. *J. Phys. Chem.*, 97(39):9973–9976, 1993. doi: 10.1021/j100141a013.
- [122] David M. Deaven and Kai-Ming Ho. Molecular geometry optimization with a genetic algorithm. *Phys. Rev. Lett.*, 75(2):288–291, 1995.

- [123] Bernd Hartke. Global cluster geometry optimization by a phenotype algorithm with Niches: Location of elusive minima, and low-order scaling with cluster size. *J. Comput. Chem.*, 20(16):1752–1759, 1999. doi: 10.1002/(SICI)1096-987X(199912)20:16<1752::AID-JCC7>3.0.CO;2-0.
- [124] Roy L. Johnston. Evolving better nanoparticles: Genetic algorithms for optimising cluster geometries. *Dalt. Trans.*, (22):4193–4207, 2003. doi: 10.1039/b305686d.
- [125] Volker Blum, Gus L. W. Hart, Michael J. Walorski, and Alex Zunger. Using genetic algorithms to map first-principles results to model Hamiltonians: Application to the generalized Ising model for alloys. *Phys. Rev. B*, 72(16):165113, 2005. doi: 10.1103/PhysRevB.72.165113.
- [126] Sandro E. Schönborn, Stefan Goedecker, Shantanu Roy, and Artem R. Oganov. The performance of minima hopping and evolutionary algorithms for cluster structure prediction. *J. Chem. Phys.*, 130(14):144108, 2009. doi: 10.1063/1.3097197.
- [127] Marek Sierka. Synergy between theory and experiment in structure resolution of low-dimensional oxides. *Prog. Surf. Sci.*, 85(9-12):398–434, 2010. doi: 10.1016/j.progsurf.2010.07.004.
- [128] Saswata Bhattacharya, Sergey V. Levchenko, Luca M. Ghiringhelli, and Matthias Scheffler. Stability and Metastability of Clusters in a Reactive Atmosphere: Theoretical Evidence for Unexpected Stoichiometries of  $Mg_MO_x$ . *Phys. Rev. Lett.*, 111(13):135501, 2013. doi: 10.1103/PhysRevLett.111.135501.
- [129] Roy L. Johnston, editor. *Applications of Evolutionary Computation in Chemistry*. Structure and Bonding. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-40258-9. doi: 10.1007/b10607.
- [130] Bernd Hartke. Global optimization. *WIREs: Comput. Mol. Sci.*, 1(6): 879–887, 2011. doi: 10.1002/wcms.70.
- [131] Sven Heiles and Roy L. Johnston. Global optimization of clusters using electronic structure methods. *Int. J. Quantum Chem.*, 113(18):2091–2109, 2013. doi: 10.1002/qua.24462.
- [132] Gus L. W. Hart, Volker Blum, Michael J. Walorski, and Alex Zunger. Evolutionary approach for determining first-principles hamiltonians. *Nature Mater.*, 4(5):391–394, 2005. doi: 10.1038/nmat1374.
- [133] Nadine L. Abraham and Matthew I. J. Probert. A periodic genetic algorithm with real-space representation for crystal structure and polymorph prediction. *Phys. Rev. B*, 73(22):224104, 2006. doi: 10.1103/PhysRevB.73.224104.
- [134] Artem R. Oganov and Colin W. Glass. Crystal structure prediction using ab initio evolutionary techniques: principles and applications. *J. Chem. Phys.*, 124(24):244704, 2006. doi: 10.1063/1.2210932.

- [135] William W. Tipton and Richard G. Hennig. A grand canonical genetic algorithm for the prediction of multi-component phase diagrams and testing of empirical potentials. *J. Phys. Condens. Matter*, 25(49):495401, 2013. doi: 10.1088/0953-8984/25/49/495401.
- [136] Christian Neiss and Detlef Schooss. Accelerated cluster structure search using electron diffraction data in a genetic algorithm. *Chem. Phys. Lett.*, 532(null):119–123, 2012. doi: 10.1016/j.cplett.2012.02.062.
- [137] Zoe E. Brain and Matthew A. Addicoat. Optimization of a genetic algorithm for searching molecular conformer space. *J. Chem. Phys.*, 135(17):174106, 2011. doi: 10.1063/1.3656323.
- [138] Graeme Henkelman and Hannes Jónsson. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.*, 113(22):9978, 2000. doi: 10.1063/1.1323224.
- [139] Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. String method for the study of rare events. *Phys. Rev. B*, 66(5):052301, 2002. doi: 10.1103/PhysRevB.66.052301.
- [140] Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *J. Chem. Phys.*, 126(16):164103, 2007. doi: 10.1063/1.2720838.
- [141] Baron Peters, Andreas Heyden, Alexis T. Bell, and Arup Chakraborty. A growing string method for determining transition states: comparison to the nudged elastic band and string methods. *J. Chem. Phys.*, 120(17):7877–7886, 2004. doi: 10.1063/1.1691018.
- [142] Ying Yao, Adriana Supady, Carsten Baldauf, Matthias Scheffler, and Luca. M. Ghiringhelli. Personal communication, manuscript in preparation.
- [143] Omar Valsson, Pratyush Tiwary, and Michele Parrinello. Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint. *Annu. Rev. Phys. Chem.*, 67:159–184, 2016. doi: 10.1146/annurev-physchem-040215-112229.
- [144] John D. Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.*, 25:135–44, 2014. doi: 10.1016/j.sbi.2014.04.002.
- [145] Frank Noé, Illia Horenko, Christof Schütte, and Jeremy C. Smith. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J. Chem. Phys.*, 126(15):155102, 2007. doi: 10.1063/1.2714539.
- [146] Vijay S. Pande, Kyle Beauchamp, and Gregory R. Bowman. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods*, 52(1):99–105, 2010. doi: 10.1016/j.ymeth.2010.06.002.



- [147] Enzo Marinari and Giorgio Parisi. Simulated Tempering: A New Monte Carlo Scheme. *Europhysics Letters (EPL)*, 19(6):451–458, 1992. doi: 10.1209/0295-5075/19/6/002.
- [148] Alessandro Laio and Francesco L. Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.*, 71(12):126601, 2008. doi: 10.1088/0034-4885/71/12/126601.
- [149] Ville Havu, Volker Blum, Paula Havu, and Matthias Scheffler. Efficient integration for all-electron electronic structure calculation using numeric basis functions. *J. Comput. Phys.*, 228(22):8367–8379, 2009. doi: 10.1016/j.jcp.2009.08.008.
- [150] Andreas Marek, Volker Blum, Rainer Johanni, Ville Havu, Bruno Lang, Thomas Auckenthaler, Alexander Heinecke, Hans-Joachim Bungartz, and Hermann Lederer. The ELPA library: scalable parallel eigenvalue solutions for electronic structure theory and computational science. *J. Phys. Condens. Matter*, 26(21):213201, 2014. doi: 10.1088/0953-8984/26/21/213201.
- [151] Kurt Lejaeghere, Veronique Van Speybroeck, Guido Van Oost, and Stefaan Cottenier. Error Estimates for Solid-State Density-Functional Theory Predictions: An Overview by Means of the Ground-State Elemental Crystals. *Crit. Rev. Solid State Mater. Sci.*, 39(1):1–24, 2014. doi: 10.1080/10408436.2013.772503.
- [152] Bernard Delley. An all-electron numerical method for solving the local density functional for polyatomic molecules. *J Chem. Phys.*, 92(1):508, 1990. doi: 10.1063/1.458452.
- [153] FHI-aims: Full-Potential, All-Electron Electronic Structure Theory with Numeric Atom-Centered Basis Functions. URL <https://aimsclub.fhi-berlin.mpg.de/>.
- [154] Thom H. Dunning. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.*, 90(2):1007, 1989. doi: 10.1063/1.456153.
- [155] ORCA - An ab initio, DFT and semiempirical SCF-MO package. URL <https://orcaforum.cec.mpg.de/>.
- [156] Frank Neese and Edward F. Valeev. Revisiting the Atomic Natural Orbital Approach for Basis Sets: Robust Systematic Basis Sets for Explicitly Correlated and Conventional Correlated ab initio Methods? *J. Chem. Theory Comput.*, 7(1):33–43, 2011. doi: 10.1021/ct100396y.
- [157] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, 65(3):712–725, 2006. doi: 10.1002/prot.21123.
- [158] Christof H. Schwab. Conformations and 3D pharmacophore searching. *Drug Discov. Today. Technol.*, 7(4):e245–e253, 2010. doi: 10.1016/j.ddtec.2010.10.003.

- [159] Irwin D. Kuntz, Jeffrey M. Blaney, Stuart J. Oatley, Robert Langridge, and Thomas E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161(2):269–288, 1982.
- [160] Rajendra Kristam, Valerie J. Gillet, Richard A. Lewis, and David Thorner. Comparison of conformational analysis techniques to generate pharmacophore hypotheses using catalyst. *J. Chem. Inf. Model.*, 45(2):461–476, 2005. doi: 10.1021/ci049731z.
- [161] GNU Lesser General Public License. URL <https://www.gnu.org/licenses/lgpl.html>.
- [162] Attila Bérces, Dennis M. Whitfield, and Tomoo Nukada. Quantitative description of six-membered ring conformations following the IUPAC conformational nomenclature. *Tetrahedron*, 57(3):477–491, 2001. doi: 10.1016/S0040-4020(00)01019-X.
- [163] Anthony D. Hill and Peter J. Reilly. Puckering coordinates of monocyclic rings by triangular decomposition. *J. Chem. Inf. Model.*, 47(3):1031–1035, 2007. doi: 10.1021/ci600492e.
- [164] Matti Ropo, Markus Schneider, Carsten Baldauf, and Volker Blum. First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids. *Sci. Data*, 3:160009, 2016. doi: 10.1038/sdata.2016.9.
- [165] Michael J. Hartshorn, Marcel L. Verdonk, Gianni Chessari, Suzanne C. Brewerton, Wijnand T. M. Mooij, Paul N. Mortenson, and Christopher W. Murray. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.*, 50(4):726–741, 2007. doi: 10.1021/jm061277y.
- [166] Paolo Tosco, Nikolaus Stiefl, and Gregory Landrum. Bringing the MMFF force field to the RDKit: implementation and validation. *J. Cheminform.*, 6(1):37, 2014. doi: 10.1186/s13321-014-0037-3.
- [167] Marat Valiev, Eric J. Bylaska, Niranjana Govind, Karol Kowalski, Tjerk P. Straatsma, Hubertus J. J. Van Dam, Dunyou Wang, Jarek Nieplocha, Edoardo Apra, Theresa L. Windus, and Wibe A. de Jong. NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.*, 181(9):1477–1489, 2010. doi: 10.1016/j.cpc.2010.04.018.
- [168] Heather B. Mayes, Linda J. Broadbelt, and Gregg T. Beckham. How sugars pucker: electronic structure calculations map the kinetic landscape of five biologically paramount monosaccharides and their implications for enzymatic catalysis. *J. Am. Chem. Soc.*, 136(3):1008–1022, 2014. doi: 10.1021/ja410264d.
- [169] Frank Noé, Dieter Krachtus, Jeremy C. Smith, and Stefan Fischer. Transition networks for the comprehensive characterization of complex conformational change in proteins. *J. Chem. Theory Comput.*, 2(3):840–857, 2006. doi: 10.1021/ct050162r.

- [170] Maria Cameron and Eric Vanden-Eijnden. Flows in Complex Networks: Theory, Algorithms, and Application to Lennard-Jones Cluster Rearrangement. *J. Stat. Phys.*, 156(3):427–454, 2014. doi: 10.1007/s10955-014-0997-8.
- [171] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. 1(5):826–843, 2011. doi: 10.1002/wcms.31.
- [172] Andreas Richter. Exploration of biopolymer energy landscapes via random sampling. Diplomarbeit, Friedrich-Schiller-Universität Jena, 2009.
- [173] Jacob Midtgaard-Olesen. Barrier trees for continuous landscapes. Master’s thesis, University of Southern Denmark, 2009.
- [174] Martin Kamp Jensen. Energy landscape analysis for peptides using barrier trees. Master’s thesis, University of Southern Denmark, 2011.
- [175] David J Wales. Energy landscapes: some new horizons. *Curr. Opin. Struct. Biol.*, 20(1):3–10, 2010. doi: 10.1016/j.sbi.2009.12.011.
- [176] Edsger W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Math.*, 1(1):269–271, 1959. doi: 10.1007/BF01386390.
- [177] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.*, 4(3):435–447, 2008. doi: 10.1021/ct700301q.
- [178] Kyle A. Beauchamp, Gregory R. Bowman, Thomas J. Lane, Lutz Maibaum, Imran S. Haque, and Vijay S. Pande. MSMBuilder2: Modeling Conformational Dynamics at the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.*, 7(10):3412–3419, 2011. doi: 10.1021/ct200463m.
- [179] Graphviz - graph visualization software. URL <http://graphviz.org/>.
- [180] Ryoji Kusaka, Di Zhang, Patrick S. Walsh, Joseph R. Gord, Brian F. Fisher, Samuel H. Gellman, and Timothy S. Zwier. Role of ring-constrained  $\gamma$ -amino acid residues in  $\alpha/\gamma$ -peptide folding: single-conformation UV and IR spectroscopy. *J. Phys. Chem. A*, 117(42):10847–10862, 2013. doi: 10.1021/jp408258w.
- [181] Yan Zhao, Nathan E. Schultz, , and Donald G. Truhlar. Design of density functionals by combining the method of constraint satisfaction with parametrization for thermochemistry, thermochemical kinetics, and noncovalent interactions. *J. Chem. Theory Comput.*, 2(2):364–382, 2006. doi: 10.1021/ct0502763.
- [182] Peter H. Seeberger. Automated oligosaccharide synthesis. *Chem. Soc. Rev.*, 37(1):19–28, 2008. doi: 10.1039/b511197h.
- [183] Junqi Li, Steven G. Ballmer, Eric P. Gillis, Seiko Fujii, Michael J. Schmidt, Andrea M. E. Palazzolo, Jonathan W. Lehmann, Greg F. Morehouse, and Martin D. Burke. Synthesis of many different types of organic small molecules using one automated process. *Science*, 347(6227):1221–1226, 2015. ISSN 1095-9203. doi: 10.1126/science.aaa5414.

- [184] Kendall N. Houk and Paul H.-Y. Cheong. Computational prediction of small-molecule catalysts. *Nature*, 455(7211):309–313, 2008. ISSN 1476-4687. doi: 10.1038/nature07368.
- [185] Jeffrey I. Seeman. Effect of conformational change on reactivity in organic chemistry. Evaluations, applications, and extensions of Curtin-Hammett Winstein-Holness kinetics. *Chem. Rev.*, 83(2):83–134, 1983. doi: 10.1021/cr00054a001.
- [186] Alexander Wittkopp and Peter R. Schreiner. Metal-Free, Noncovalent Catalysis of Diels-Alder Reactions by Neutral Hydrogen Bond Donors in Organic Solvents and in Water. *Chem.-Eur. J.*, 9(2):407–414, 2003. doi: 10.1002/chem.200390042.
- [187] Peter R. Schreiner. Metal-free organocatalysis through explicit hydrogen bonding interactions. *Chem. Soc. Rev.*, 32(5):289–296, 2003. doi: 10.1039/b107298f.
- [188] Martin Kirsten, Julia Rehbein, Martin Hiersemann, and Thomas Strassner. Organocatalytic claisen rearrangement: theory and experiment. *J. Org. Chem.*, 72(11):4001–4011, 2007. doi: 10.1021/jo062455y.
- [189] Andrea Hamza, Gabor Schubert, Tibor Soós, and Imre Papai. Theoretical studies on the bifunctionality of chiral thiourea-based organocatalysts: competing routes to C-C bond formation. *J. Am. Chem. Soc.*, 128(40):13151–13160, 2006. doi: 10.1021/ja063201x.
- [190] Zhiguo Zhang, Zongbi Bao, and Huabin Xing. N,N'-Bis[3,5-bis(trifluoromethyl)phenyl]thiourea: a privileged motif for catalyst development. *Org. Biomol. Chem.*, 12(20):3151–3162, 2014. doi: 10.1039/c4ob00306c.
- [191] Peter R. Schreiner and Alexander Wittkopp. H-Bonding Additives Act Like Lewis Acid Catalysts. *Org. Lett.*, 4(2):217–220, 2002. doi: 10.1021/ol017117s.
- [192] Tomotaka Okino, Yasutaka Hoashi, and Yoshiji Takemoto. Enantioselective Michael reaction of malonates to nitroolefins catalyzed by bifunctional organocatalysts. *J. Am. Chem. Soc.*, 125(42):12672–12673, 2003. doi: 10.1021/ja036972z.
- [193] Stephan J. Zuend and Eric N. Jacobsen. Mechanism of Amido-Thiourea Catalyzed Enantioselective Imine Hydrocyanation: Transition State Stabilization via Multiple Non-Covalent Interactions. *J. Am. Chem. Soc.*, 131(42):15358–15374, 2009. doi: 10.1021/ja9058958.
- [194] Paul Ha Yeon Cheong, Claude Y. Legault, Joann M. Um, Nihan Çelebi-Ölçüm, and K. N. Houk. Quantum mechanical investigations of organocatalysis: Mechanisms, reactivities, and selectivities. *Chem. Rev.*, 111:5042–5137, 2011. doi: 10.1021/cr100212h.
- [195] Katharina M. Lippert, Kira Hof, Dennis Gerbig, David Ley, Heike Hausmann, Sabine Guenther, and Peter R. Schreiner.

- Hydrogen-Bonding Thiourea Organocatalysts: The Privileged 3,5-Bis(trifluoromethyl)phenyl Group. *Eur. J. Org. Chem.*, 2012(30):5919–5927, 2012. doi: 10.1002/ejoc.201200739.
- [196] Gábor Tárkányi, Péter Király, Tibor Soós, and Szilárd Varga. Active conformation in amine-thiourea bifunctional organocatalysis performed by catalyst aggregation. *Chem.-Eur. J.*, 18(7):1918–22, 2012. doi: 10.1002/chem.201102701.
- [197] Candee C. Chambers, Edet F. Archibong, Ali Jabalameli, Richard H. Sullivan, David J. Giesen, Christopher J. Cramer, and Donald G. Truhlar. Quantum mechanical and  $^{13}\text{C}$  dynamic NMR study of 1,3-dimethylthiourea conformational isomerizations. *J. Mol. Struct.: THEOCHEM*, 425(1-2):61–68, 1998. doi: 10.1016/S0166-1280(97)00137-1.
- [198] Radu Custelcean, Maryna G. Gorbunova, and Peter V. Bonnesen. Steric control over hydrogen bonding in crystalline organic solids: a structural study of  $\text{N,N}'$ -dialkylthioureas. *Chem.-Eur. J.*, 11(5):1459–1466, 2005. doi: 10.1002/chem.200400973.
- [199] John P. Terhorst and William L. Jorgensen. E/Z Energetics for Molecular Modeling and Design. *J. Chem. Theory Comput.*, 6(9):2762–2769, 2010. doi: 10.1021/ct1004017.
- [200] Jonathan Clayden, Ulrich Hennecke, Mark A. Vincent, Ian H. Hillier, and Madeleine Helliwell. The origin of the conformational preference of  $\text{N,N}'$ -diaryl- $\text{N,N}'$ -dimethyl ureas. *Phys. Chem. Chem. Phys.*, 12(45):15056, 2010. doi: 10.1039/c0cp00571a.
- [201] Jhenny F. Galan, Edward Germany, Amanda Pawlowski, Lynette Strickland, and Mary Grace I. Galinato. Theoretical and Spectroscopic Analysis of  $\text{N,N}'$ -Diphenylurea and  $\text{N,N}'$ -Dimethyl- $\text{N,N}'$ -diphenylurea Conformations. *J. Phys. Chem. A*, 118(28):5304–5315, 2014. doi: 10.1021/jp503539m.
- [202] Vyacheslav S. Bryantsev and Benjamin P. Hay. Conformational preferences and internal rotation in alkyl- and phenyl-substituted thiourea derivatives. *J. Phys. Chem. A*, 110(14):4678–4688, 2006. doi: 10.1021/jp056906e.
- [203] Matthew D. Wodrich, Clémence Corminboeuf, Peter R. Schreiner, Andrey A. Fokin, and Paul von Ragué Schleyer. How accurate are DFT treatments of organic energies? *Org. Lett.*, 9(10):1851–1854, 2007. doi: 10.1021/olo70354w.
- [204] Jirí Sponer, Kevin E. Riley, and Pavel Hobza. *Phys. Chem. Chem. Phys.* doi: 10.1039/b719370j.
- [205] Kenichi Fukui, Teijiro Yonezawa, and Haruo Shingu. A Molecular Orbital Theory of Reactivity in Aromatic Hydrocarbons. *J. Chem. Phys.*, 20(4):722, dec 1952. doi: 10.1063/1.1700523.
- [206] Gilles. Klopman. Chemical reactivity and the concept of charge- and frontier-controlled reactions. *J. Am. Chem. Soc.*, 90(2):223–234, 1968.

- [207] Lionel. Salem. Intermolecular orbital theory of the interaction between conjugated systems. II. Thermal and photochemical cycloadditions. *J. Am. Chem. Soc.*, 90(3):553–566, 1968. doi: 10.1021/ja01005a002.
- [208] Lionel. Salem. Intermolecular orbital theory of the interaction between conjugated systems. I. General theory. *J. Am. Chem. Soc.*, 90(3):543–552, 1968. doi: 10.1021/ja01005a001.
- [209] Fleming, Ian. *Frontier Orbitals and Organic Chemical Reactions*. John Wiley & Sons, Chichester, 1976.
- [210] The Materials Project. URL <https://www.materialsproject.org/>.
- [211] Harvard Clean Energy Project. URL <http://cleanenergy.molecularspace.org/>.
- [212] Open PHACTS Foundation. URL <http://www.openphactsfoundation.org/>.
- [213] Novel Materials Discovery (NOMAD) Laboratory. URL <http://nomad-coe.eu/>.
- [214] Bálint Aradi, Ben Hourahine, and Thomas Frauenheim. DFTB+, a sparse matrix-based implementation of the DFTB method. *J. Phys. Chem. A*, 111(26):5678–5684, 2007. doi: 10.1021/jp070186p.
- [215] Fafoom – development branch "orientation\_dof", . URL [https://github.com/adrianasupady/fafoom/tree/orientation\\_dof/](https://github.com/adrianasupady/fafoom/tree/orientation_dof/).