EPFL

Thèse n° 9740

Characterization and prediction of peptide structures on inorganic surfaces

Présentée le 29 juillet 2022

Faculté des sciences et techniques de l'ingénieur
Laboratoire de science computationnelle et modélisation
Programme doctoral en science et génie des matériaux

pour l'obtention du grade de Docteur ès Sciences

par

Dmitrii MAKSIMOV

Acceptée sur proposition du jury

Dr Y. Leterrier, président du jury
Prof. M. Ceriotti, Dr M. Rossi Carvalho, directeurs de thèse
Prof. O. Hofmann, rapporteur
Prof. M. Todorovic, rapporteuse
Prof. N. Marzari, rapporteur

École
polytechnique
fédérale
de Lausanne

2022

To my family, friends and colleagues.

Acknowledgements

First and foremost, I owe a debt of gratitude to Dr. Mariana Rossi, for whom I am eternally grateful. She is an amazing group leader and a humble and kind individual who has been instrumental in guiding me through the often difficult times that come with research and writing a dissertation.
Professor Dr. Michele Ceriotti, who served as my thesis director and allowed me to get practical expertise in machine learning methods deserves a special thank you for his guidance and support.
I would like to express my gratitude to Dr. Carsten Baldauf for his early research oversight.
I would like to express my gratitude to Prof. Dr. Matthias Scheffler for his intriguing study at the Fritz Haber Institute's Theory Department, which he is conducting.
I would like to express my gratitude to the Max Planck-EPFL Center for Molecular Nanoscience and Technology – particularly to Prof. Dr. Klaus Kern and Dr. Klaus Kuhnke – for their support in our collaborative research effort.
Because of the people that worked there, the Fritz Haber Institute's Theory Department has always enjoyed a pleasant working environment. We owe a debt of appreciation to far too many people, including (but not limited to) Karen and Florian; Marcel and Marcin; Xiajuan; Markus; Maria; Christian and Sebastian; Björn and Hanna; Luca and Majid; Sebastian and Henrik and Sebastian; Alan and Alaa; and Yair.
Special thanks to my friends who were with me during my time in Europe: Valeria Volodkina, Evgenii Ikonnikov, Stanislav Podchezerzev and Alexandr Sukhanov.

Hamburg, July 7, 2022 D. M.

Abstract

Interfaces between peptides and metallic surfaces are the subject of great interest for possible use in technological and medicinal applications, mainly since organic systems present an extensive range of functionalities, are abundant, cheap, and exhibit low toxicity. *Exemplary applications* are biosensors that may be sensitive to specific metabolites or harmful compounds. However, these hybrid interfaces pose a challenge to computational modelling, particularly regarding predicting the most relevant configurations at the surface, which determines the electronic properties of the system as a whole. From a theoretical point of view, predicting the most stable interface configuration requires searching through the enormous structure space of flexible biomolecules with respect to the surface for different configurations and performing computational calculations of their properties. However, it is impossible to investigate those parts separately due to complex interactions during adsorption. In order to capture these complex interactions, one has to employ accurate theoretical methods, which are very computationally expensive. In this thesis, we provide a comprehensive description of the complex nature of the interaction of selected amino acids with metallic surfaces using state of the art dimensionality reduction techniques and accurate *ab initio* theoretical methods and creation of tools tailored for the high-throughput investigations of interface systems.

The theoretical methods used in the thesis are described in its first part. The second section looks into the conformational space changes of Arginine (Arg) and its protonated counterpart after adsorption on three noble metallic surfaces. Arg is an excellent testbed because it is tiny enough to be treated using density functional theory, which is considered the best compromise between accuracy and computational efficiency. At the same time, Arg is complex enough due to a highly flexible side-chain that allows for hundreds of different configurations in the gas phase alone. The examination of adsorption behaviour requires creating a database by performing a large number of geometry optimizations of various conformations and orientations. The investigation of that database includes creating a low-dimensional representation of the conformational spaces using recent dimensionality reduction techniques, followed by examining various bonding and charge transfer patterns and how they affect the available conformational spaces.

The third section of the thesis is concerned with developing tools for the automated structure search of interface systems and the modelling of self-assembly patterns formed after adsorption. Different geometry optimization algorithms and a flexible method of preconditioning the quasi-Newton optimization algorithms are implemented in the GenSec package that

was developed. Together, these enable a more straightforward interface with a wide range of quantum chemistry packages for sampling the conformational spaces of flexible molecules in 1D (ions), 2D (surfaces), and 3D (cavities and molecules) systems. Structure search of the conformational space of a flexible molecule using GenSec provided satisfactory results for di-L-alanine adsorbed on Cu(110) surface.

Contents

Abbreviations

AA amino acid. 3–5, 8, 9, 11, 13, 14, 35, 40, 104, 107
AIMD *ab initio* molecular dynamics. 12
Arg Arginine. iii, xi–xiv, xix, 13, 55, 57–79, 81, 83–85
Arg-H$^+$ Arginine-H$^+$. xi–xiv, xix, 13, 55, 57–62, 65–68, 70–79, 81, 83–85
ASE Atomic Simulation Environment. 90–92, 97, 101, 106, 108, 112, 117

BFGS Broyden-Fletcher-Goldfarb-Shanno. 43, 44, 97, 98, 101, 102
BO Born-Oppenheimer. 18, 19

COM center of mass. 91, 92, 94–97

DFA density-functional approximation. 22, 25, 29
DFT density-functional theory. xiii, xiv, xix, 10–13, 17, 20, 22, 24, 26, 27, 30, 35, 40, 42, 55, 73, 74, 79, 83, 84, 116, 117

ES-IBD electrospray ion beam deposition. 9

fcc face-centered cubic. 35, 57, 101
FF force field. 11, 12, 17, 34, 35, 40, 47, 90, 94, 111
FHI-aims Fritz Haber Institute "ab initio molecular simulations". 29–31, 47, 101, 106

GenSec Generation and Search. 89, 90, 94, 97, 100, 101, 103, 107, 108
GGA generalized gradient-approximated. 22, 23

HEG homogeneous electron gas. 22

KS Kohn-Sham. 21, 29

LDA local density approximation. 22, 23
LDOS local density of states. 9
LJ Lennard-Jones. xv, 34, 35, 47, 98–101, 103
LSM line search method. 44, 46
LZK Lifshitz-Zaremba-Kohn. 28

MD molecular dynamics. 12, 13, 40, 43
ML machine learning. 11, 13, 42, 44, 112

NAO numeric atom-centered orbitals. 29

NN neural networks. 44

PBC periodic boundary conditions. 31, 89, 94
PBE Perdew, Burke and Ernzerhof. xiv, xix, 23, 24, 57, 58, 60, 73, 77, 81, 83, 84, 101
PCA principal component analysis. 50
PES potential energy surface. 11, 12, 19, 34, 39–46, 102, 103

QM quantum mechanical. 33, 35

REMatch regularized entropy match kernel. 50
RMSD root mean square displacement. 102

SCF self-consistent field. 31
SOAP smooth overlap of atomic positions. 13, 48, 62, 65
STM scanning tunneling microscopy. 9, 10, 17, 30, 33, 104, 107, 108, 110, 111

TRM trust-region method. 45, 46, 101
TS Tkatchenko and Scheffler. 25–28

vdW van der Waals. 3, 11, 24–29, 34, 41, 57, 73, 94, 98, 100, 104

XC exchange-correlation. 21–23, 25, 27
XPS X-ray photoemission spectroscopy. 9

List of Figures

List of Tables

[50mm] Does anybody really know the secret
Or the combination for this life and where they keep it?
It's kinda sad when you don't know the meanin'
But everything happens for a reason...
"Take a look around", Limp Bizkit

1 Introduction

Because of the fascinating potential applications of hybrid organic-inorganic interfaces, adsorption and self-assembly of organic molecules on surfaces are critical topics in nanoscience and nanotechnology [1]. For example, amino acids that are the building blocks of peptides and their oligomers are particularly intriguing because they are naturally biocompatible and provide a rich functional space already at the amino acid (AA) level. The combinatorial increase in molecular motifs made available by forming peptide bonds can further enlarge this functional space. By immobilizing a bioorganic component on a substrate, an inorganic part acts as a platform to support and capture interactions and reactions, which provide the path for creating different bionanoelectronic devices.

In recent years, a tremendous effort has been expended to identify adsorbates' structure on surfaces and disentangle the processes behind self-assembly that would lead to the rational design of materials and devices with desired properties.

From a theoretical point of view, this poses a challenge to computational modelling, particularly regarding the prediction of stable configurations at the interface at different conditions, which determines the electronic properties of the system as a whole. Even in the gas-phase, single AA have rich conformational spaces, where they can have hundreds of distinct local minima [2], and determination of them requires computationally expensive methods. After adsorption on the surface, the conformational preferences of the AAs can change dramatically due to a combination of factors, such as van der Waals (vdW), electrostatic or ionic interactions, but also due to their reduced flexibility, as well as by intermolecular forces and interactions with the surface itself [3, 4]. The systematic structure search of molecules adsorbed on surfaces and creation of databases including energetic information from the theoretical approaches is of high importance for revealing structure-property relationships of the interface systems, for further developments of the theoretical methods able to describe larger structures, and for disentangling of the mechanisms of self-assembly. However, such studies are challenging as they require (i) accurate energetics for a system containing elements across the periodic table and where considerable charge rearrangement and chemical reactions can occur (ii) sampling and representing a large conformational space, and (iii)

dealing with structure motifs that can only be represented by unit cells containing hundreds of atoms.

The scope of this thesis is the description of the complex nature of the interaction of AAs with metallic surfaces and the creation of tools for high-throughput calculations for investigations of interface systems. An exhaustive structure search for two AAs on three metallic surfaces was performed with the use of *ab initio* methods that are required for analysis of the electronic properties of the interface systems. The database created during the work contains thousands of local minima and is available for further development of the methods that can accelerate the research of self-assembly phenomena. The databases were analyzed with state-of-the-art unsupervised machine learning techniques that help reveal structure-property relationships in that kind of system. Further, we developed a package that automates the structure search of flexible molecules with respect to specified surroundings that connects to most of the electronic structure packages available today, making it freely available and open source. We investigate the adsorption of a di-L-alanine molecule adsorbed on Cu(110) surface using this package.

1.1 Amino acids and peptides

AAs are organic compounds that contain amino ($-NH_2$) and carboxyl ($-COOH$) functional groups, along with a side chain unique to each AA. AAs are known to be the monomer units of peptides and are essential for the existence of life. In the form of proteins, AA residues are the second-largest component of human muscles after water. Analyses of a large number of proteins from nearly every possible source have revealed that all proteins are made up of 20 "standard" AAs. Not all 20 types of AAs are found in every protein, although most proteins contain the majority, if not all, of the 20 types [5]. In addition, AAs and their derivatives are involved in processes as neurotransmitters - chemical messengers for communication between cells. For example, diminished activity of serotonin (tryptophan derivative) pathways plays a causal role in the pathophysiology of depression [6].

The most general formula to represent the common AA which is called α-amino acid, is reported in Fig. 1.1 a: the molecule is distinguished by the presence of a α carbon atom in the center, to which both the amino and carboxyl groups are attached. The rest of the molecule is represented as a side chain (R group), the structure of which uniquely defines all the common AAs. Depending on the molecule's environmental conditions, AAs can exist in three different chemical forms (see Fig. 1.1 a): i) the neutral form is common for isolated molecules; (ii) the zwitterionic form is common for solid AAs crystals and for molecules on poorly reactive surfaces and in solutions. This form appears when a proton is transferred from the carboxylic group to the amino group of the same molecule, which maintains its global neutrality; (iii) the anionic state is typical for AAs that interact strongly with a substrate, resulting in chemical bond breaking/formation and deprotonation of the molecule.

Except for the smallest AA glycine, all other AAs are chiral (Fig. 1.1 b), which implies that they have nonsuperimposable mirror images known as enantiomers of one another. Although

there exist (L) and (D) enantiomers, the (L)-enantiomer is the only one found in living beings; as a result, the vast majority of investigations have been conducted on (L)-type molecules.

As one progresses towards more complicated and "realistic" biomolecules, one comes across peptides, which are polymers of AAs connected by CO-NH peptidic bonds (Fig. 1.1 c). A dipeptide, for example, is formed by the condensation of two AAs, i.e. the reaction between one AA's carboxyl group and the amino group of the second, with the elimination of one water molecule. Peptides are chains of comparable (homopeptides) or different (heteropeptides) AAs. Proteins are the "summum" of a peptide chain, where the sequence of AAs, their location, and their three-dimensional layout regulate the biological activity of the molecule.

AAs exhibit a range of polarity and structural features. AA side chains can be nonpolar (e.g. glycine, alanine, valine, leucine, isoleucine, methionine, proline, phenylalanine, tryptophan), polar (e.g. serine, threonine, asparagine, glutamine, tyrosine, cysteine), or charged (e.g. arginine, lysine, histidine, aspartic acid and glutamic acid). Side chains may be nonpolar or polar (neutral or charged). They may be aliphatic (e.g. alanine) or contain other functional groups such as carboxylic group (e.g. glutamic acid), amino group (e.g. lysine), or sulphur (e.g. cysteine). Additionally, they can be linear (e.g. glutamic acid) or have one heterocycle (e.g. proline) or aromatic (e.g. tyrosine) ring in their side chain. The structures of the twenty most frequent AAs, along with their three-letter notations and side-chain characteristics, are depicted in Fig. 1.2. More comprehensive review considering other properties of AAs and other AAs that are not specified by the “universal” genetic code that is common for almost all life forms can be found in biochemistry textbook [5].

Even with the mentioned AAs, the chemical space of possible configurations is genuinely immense, and peptides that can be formed of different sequences of AAs will vary a lot on their structural configuration and properties, which presents an advantage for the rational design of different nanodevices and functionalization of inorganic surfaces.

1.2 Recent applications of peptide-inorganic surface interface systems

In this section, we would like to show some of the recent applications of peptide-metal interfaces and thus showcase the great potential of such a field of research.

The use of peptides in solar cell applications, inspired by natural photosynthesis processes, is arguably the most straightforward optoelectronic application. Appending a dye to the side chain, or one of the ends, of a peptide, was shown to be effective in extending the absorption spectrum and increasing photocurrent production capacities [7, 8] even when the peptide is physically adsorbed on a gold surface [9]. In the presence of dyes with different excitation wavelengths, the synthesis of mixed monolayers of helical peptides with opposing dipole orientations towards the surface allowed the creation of a molecular photodiode system that can switch photocurrent direction by varying the excitation wavelength [10]. The efficiency

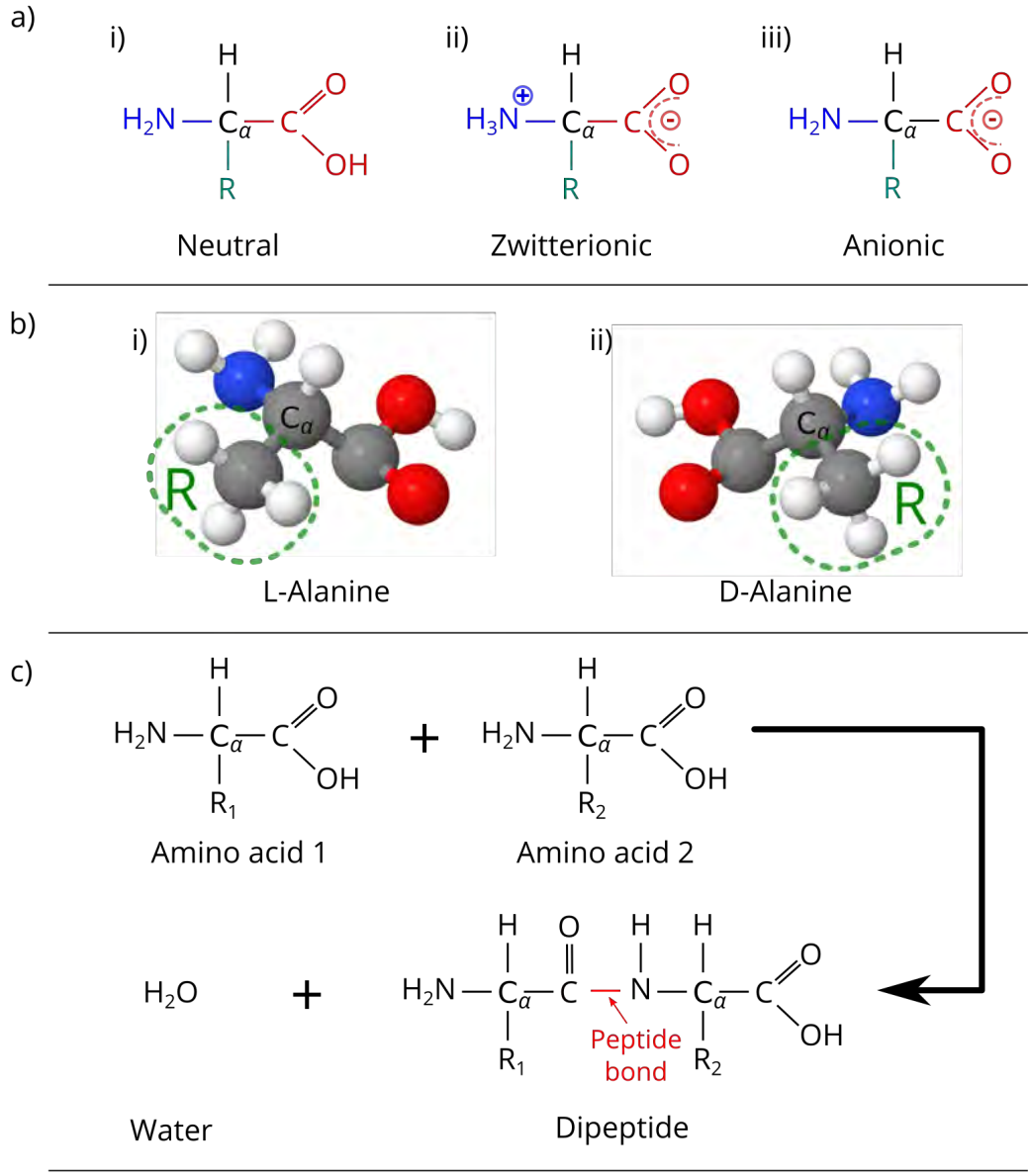


Figure 1.1 – a) The general structure of a α-amino acid in its neutral, zwitterionic, and anionic states. The amino group is highlighted in blue, the carboxylic/carboxylate group is highlighted in red, the α-carbon is highlighted in black, and the side chain is highlighted in green; b) Schematic representation of the Alanine amino acid in its neutral configuration. Red atoms are oxygen; blue atoms are nitrogen; white atoms are hydrogen, and grey atoms are carbon. The R symbol stands for the side-chain (highlighted with green dashes), here represented by the CH_3 group. In (i) L-Alanine, with respect to the central C_α carbon and in (ii) a D-Alanine; c) Schematic representation of the formation of the peptide bond: two amino acids with different side chains R_1 and R_2 react to form a peptide via the production of a water molecule.

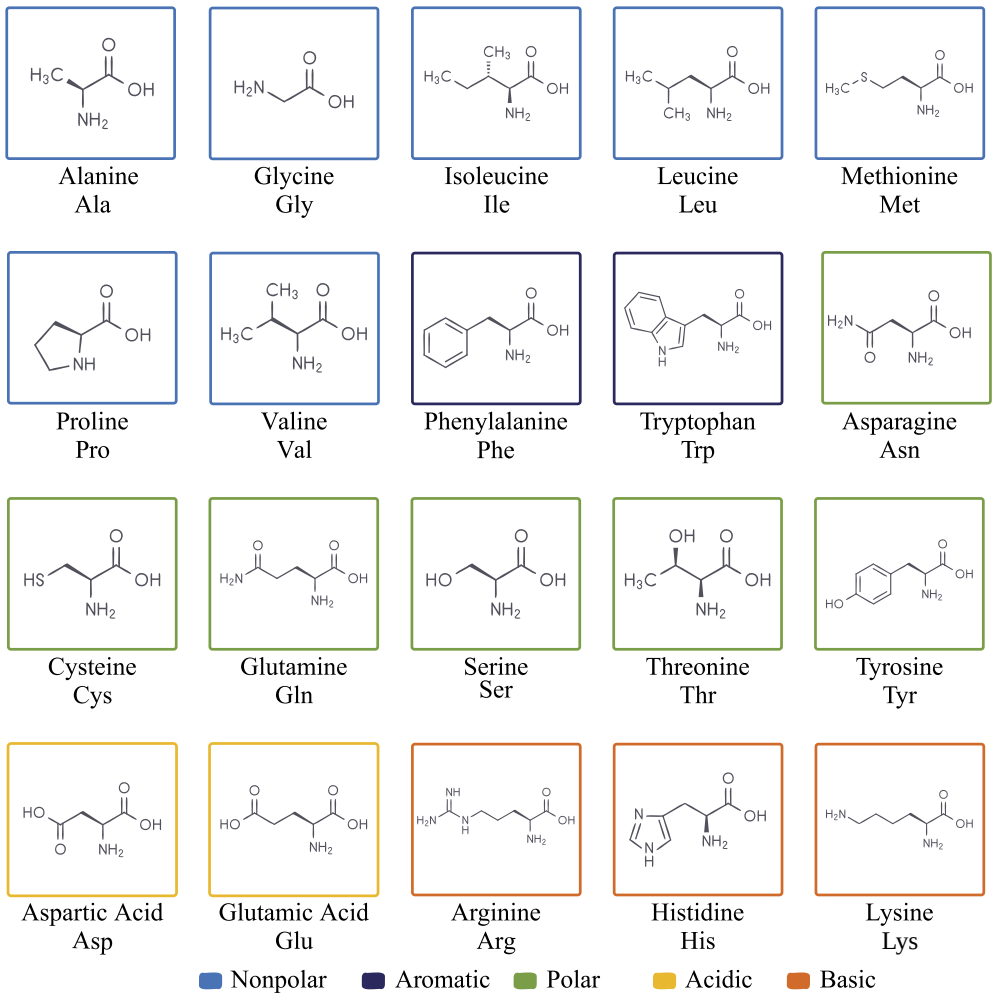


Figure 1.2 – Scheme of the 20 most common α-amino acids present in nature, represented in their neutral form.

of the organic solar cells can also be tuned by interfacial modification with an ultrathin peptide layer that causes changes in the work function of the substrate [11] that is also highly dependent on the peptide sequence and conformation of the backbone [12, 13].

Using peptides as molecular bridges and producing conductive wires is crucial for the next generation of bioelectronic devices. The effectiveness of electronic transport is dependent on the overall charge of protonating side chains which allow controlling I-V characteristics of peptide junctions [14–16]. Self-assemblies on surfaces can provide the unique and flexible way to implement ensembles of low-dimensional quantum confinement geometries [17], for example, of fullerenes that are too mobile on the surface without such a template [18] or for quantitative modulation of the work function of a substrate [19].

Using peptide monolayers as an antifouling coating [20, 21] to inhibit the adherence of

proteins and organisms to surfaces is one of the most potential applications in industry and medicine. Promotion of cell adhesion and proliferation on biomaterials is essential for the successful integration of implants [22]. Cell binding motifs, such as the Arg–Gly–Asp peptide, can be anchored to the surface of a biomaterial to increase its mechanical and biological characteristics [23]. It has been proven that titanium surfaces, a material that is commonly utilized in the implant industry, can be functionalized with cell-binding peptides by employing Cys AAs as the binding factor [24]. Also, the surface reactivity can be altered by using the intrinsic chirality of AAs, which enables chiral separation and enantioselective heterogeneous catalysis [25–27].

Another example is controlling the wettability of graphite surfaces using self-assembled peptides by mixing distinct peptide types (hydrophobic and hydrophilic) in different ratios [28]. The excellent stability of peptide nanostructures, as well as their vast surface area and controlled wettability features, make them an appealing candidate for use as the dielectric layer in supercapacitors [29, 30]. Also, AAs are non-toxic, relatively cheap and easy to produce promising green corrosion inhibitors [31–35].

At the time of writing, the author can not stress enough the need for producing biosensors targeted explicitly for detection of the pathogenic microbes and viruses, where organic molecules provide high biocompatibility and tunable selectivity due to significant variations of accessible chemical configurations [36–42].

Even though there are already many applications and devices, the fundamental mechanisms that govern particular structures adopted on particular surfaces remain unclear. The following section will be devoted to both state-of-the-art experimental and theoretical methods of investigations of organic-inorganic interfaces.

1.3 State of the art

During past decades it has been proven that a large diversity of distinct molecular assemblies may form via adsorption of organic molecules at inorganic surfaces. However, many aspects of the interaction mechanisms of biomolecules and inorganic surfaces are still unclear. Often, the shape of such self-organized structures may be adjusted by carefully controlling the deposition circumstances such as temperature [43–45], coverage [46] or changing of the substrate [47–50]. This section will offer a quick review of the methodologies that are used to investigate the adsorption of AAs on inorganic surfaces.

1.3.1 Experimental techniques

Self-assembly processes between molecules start with the adsorption of individual molecules from the gas phase (or liquid), then diffusion on the surface and further island formation through molecule-molecule interaction. Using a crucible (Knudsen cell) to sublimate AAs from a crystalline form under vacuum conditions is a standard method for generating organic layers on the substrates [51]. Such a technique is limited to relatively small peptides (of up to four AAs) and requires careful adjustment of the sublimation temperature since melting

the powders can damage them. One of the most sophisticated approaches is soft-landing electrospray ion beam deposition (ES-IBD) since the production of intact gas-phase ions by electrospray ionization is not limited by low thermal stability [52–54]. Molecular ions are decelerated before landing, preventing fragmentation and guaranteeing that the molecules remain intact following deposition. The use of mass spectrometry, mass filtering, and soft landing, all of which are essential to the ES-IBD process, ensures the intact and extremely pure deposition of the selected species under ultrahigh vacuum [52, 55, 56].

The most fundamental tool to study the self-assembly patterns of molecules is scanning tunneling microscopy (STM), which is based on the concept of quantum tunnelling. This technique measures the tunnelling current as a function of the sharp conducting tip position, applied voltage, and the local density of states (LDOS) of the sample since electrons can tunnel across the vacuum between tip and sample when the bias voltage is applied [57]. This technique allows one to determine the atomic positions in molecules and the morphology of the substrate. STM allows to obtain a three-dimensional profile of a sample as an image and distinguish different adsorption patterns of a single peptide [58–60] or how the self-assemblies look depending on the different chemical composition of the adsorbates, substrate, and overall deposition conditions [25, 46, 61–65]. However, the interpretation of STM images of molecules adsorbed on surfaces is not straightforward. First of all, STM images are not a topography map but also include electronic information of both the molecule and the underlying surface. In the case of chemisorbed systems, STM images carry information about the chemical bonding that can be extracted only from complementary investigations.

STM is often supplemented with spectroscopic studies that provide chemical state information of the adsorbed molecules and the surroundings of the functional groups. For example, AA adsorption can occur in different protonation states that can be described by proton configuration of carboxyl and amino groups (neutral, anionic or zwitterionic) and by different protonation configuration of Histidine AA. The occurrence of a zwitterionic form can be evidenced by X-ray photoemission spectroscopy (XPS) that allows the investigation of the core levels of the atoms present at the surface. The XPS analysis of core-level shifts will immediately show the presence of a charged NH_3^+ functional group, which causes an upshift on the N 1s photoemission line [25]. It is also possible to estimate the relative co-existing states of the same molecule adsorbed on the surface.

Additionally, a tunable X-ray source allows other types of spectroscopies, like near-edge X-ray absorption fine structure (NEXAFS), where the X-ray adsorption features can be indicated by the photoabsorption cross section for electronic transitions from an atomic core level to final states in the energy range of 50–100 eV above the chosen atomic core level. When employing differently polarized light, the directed electric field vector of the X-rays can only excite those electrons able to move parallel to it, which gives them crucial information on the chemical bonding orientation [49, 66, 67].

Vibrational spectroscopy is another experimental technique that exploits the fact that molecules absorb energy at specific frequencies which resonate with their vibrational modes. Due to interactions with the surface, those specific frequencies are changed relative to gas-phase frequencies but remain characteristic to the adsorption site's chemical groups, configuration, and geometry. On metal surfaces, reflection absorption infrared spectroscopy (RAIRS) [68, 69] or high resolution electron energy loss detection (HEELS) [70, 71] can be employed. However, due to many potential vibrational modes, additional methods are frequently essential for the characterization of the adsorbed system. For more comprehensive experimental techniques, we refer the reader to the reviews [61].

Unfortunately, experimental procedures cannot provide the system with information at the needed level of resolution. The unknown tip geometry and electrical characteristics are usually the most significant uncertainties encountered in detailed STM interpretation. Also, surface diffusion is significant at room temperature, causing tip instability and affecting atomic and electrical characteristics. One of the most significant experimental limitations of spectroscopic approaches is that spectra are obtained by measuring the sample's total yield of electrons or photons. A direct link between the measured spectra and the sample's specific geometry is not guaranteed. Because of the limited resolution, high complexity of the systems, technical difficulties, and cost of the experiments, theoretical approaches become essential for accessing the properties that are not accessible through experiments, resulting in a synergy of theory and experimental data that leads to a deeper understanding of the processes that are taking place on the surface.

1.3.2 Theoretical techniques

In addition to experimental research, model computations are required to bring more insights into the structure and characteristics of the molecule–surface systems. For example, issues that can be addressed theoretically are the nature of the intermolecular interactions, structure of adsorbates, charge configuration of the molecules, their chemical composition, chiral recognition, orientation and preferred adsorption sites. In principle, the theoretical foundation suitable for addressing the problems mentioned above was already fully established with the formulation of quantum mechanics in the first part of the 20th century. However, as Paul Dirac once wrote: "The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble" [72]. We are restricted to different approximations that allow us to model systems of different scales and the available computational power that can treat such calculations.

For modelling systems that consist of hundreds of atoms per unit cell, the most popular theoretical approach nowadays is DFT, which delivers a good compromise between accuracy and computational efficiency. The fundamental theorem behind DFT is that the electronis structure properties of non-degenerate systems are entirely determined by their ground-state electron density, $n(r)$, that alone governs the whole behaviour of the system. The

so-called exchange-correlation functional, which is the $n(r)$-dependent energy contribution caused by quantum-mechanical and many-body deviations from a mean-field description of the electrons, is a fundamental piece of this approach. However, a precise equation for the universal functional still has not been found, giving rise to many suitable approximations for different systems. DFT will be discussed in more detail in the next chapter.

Pioneering works that used DFT were focused on small or rigid AAs, and on a minimal number of trial configurations [73–77] due to the high computational cost of such calculations at the time. Because the first DFT studies of complex systems did not account for vdW interactions, they were affected by a errors in their predictions; however, they are now taken into account in more modern functionals and approaches that result in a significant increase in the quantitative agreement between the predictions and the experimental data [78]. With the use of DFT, one can answer whether a chemical bond is formed between AA and a substrate, what the energy hierarchy of different adsorbed conformational configurations is, as well as determining charge distribution on the adsorbed structures and their height above the surface [61, 79–83].

One of the first studies that were dedicated to larger AAs highlight the challenge of adequately sampling the large structure space of flexible biomolecules [84] that is usually not feasible with the use of DFT due to high computational cost. These studies have clarified that an accurate potential energy surface (PES) is only one of the ingredients needed to correctly predict the structure of peptides at surfaces, with the sampling of structure space being just as important.

DFT calculations not only offer valuable information on their own, but also they can provide the basis to cheaper theoretical approaches and used, for example, as a basis for a classical force field (FF) parameterization [85–87] or for the training of machine learning (ML) models [88–90]. These methods can be several orders of magnitude cheaper to evaluate compared to DFT and, in some cases, FFs specifically developed for modelling simulations between a protein and a surface may be a good approximation. However, to obtain high-quality results, the FF parameters must be derived and calibrated for the systems of interest. Different FFs exist for modelling AAs on metallic surfaces and the most famous ones are GolP-CHARMM FF [85, 91] optimized for Au(111) and Au(100) slabs, AgP-CHARMM FF [86] that is parametrized for simulations on Ag(111) and Ag(100) in aqueous solutions and INTERFACE-FF [87] which includes a broad range of different surfaces available for modelling. The main drawback of using FFs in simulations is their non-transferability to systems other than those to which they were parametrized. Another limitation of these FFs is the inability to model chemical reactions or to capture effects such as charge transfer. While more complex FFs exist, such as bond order-based reactive FF (ReaxFF) [92] that in contrast to the previous FFs allows bond breaking and formation reactions, such FFs require much larger training sets, which can be a limiting factor for using them for various systems. To the best of our knowledge, only one ReaxFF was designed to model adsorption of glycine on Cu(110) [93].

1.3.3 Global structure search

The most challenging part of theoretical modelling is properly sampling the large structure space of flexible biomolecules. Theoretical methods such as DFT and FF allow for the calculation of the forces acting on nuclei based on the input geometry of the structure. It is possible then to determine the nuclei arrangement that results in local or global minima of the system with a given PES.

Finding the global minimum of the system implies sampling the conformational space of complex molecular systems, which frequently arises in the context of molecular dynamics (MD) simulations. With the use of MD methods, Newton's law of motion is solved numerically for the nuclei. It is possible, then, to sample the most likely regions of the PES with an array of different MD flavours, such as Born-Oppenheimer and Car-Parrinello [94]. These are usually denoted as *ab initio* molecular dynamics (AIMD) simulations since the PES is constructed using quantum mechanical approaches. Despite the very limited time scales that can be simulated using AIMD (up to hundreds of ps), studies are employing such methods, for example, to investigate the preferred chemical composition and adsorption sites of glycine and lysine [95, 96], and to study peptide-silica interactions [97] or β-sheet adhesion of gold surfaces [98].

The exploration of PES with the methods described above can be very inefficient since, during such simulations, the system can be trapped in some local minima, which limits the sampling of the conformational space. There are different methods proposed in order to enhance the sampling efficiency of MD simulations and these have been used for investigations of protein-surface interactions [79]. We will discuss them in more detail in Section 3.1.

1.3.4 Analysis of high-dimensional spaces

Analysing complex molecular systems with many degrees of freedom and interpreting of their high-dimensional data is another challenge in understanding the structure-property relationships of flexible molecules adsorbed on inorganic surfaces. There is no analytical method to determine the configurations of the different peptide structures. One of the first representations developed for the analysis of peptide structures was proposed by Ramachandran, which uses dihedral angle rotations around the N-C_α and C_α-C bonds [99] to represent the number of possible conformations for an amino-acid residue in a protein, as well as the distribution of those data points. The Ramachandran approach generally proposes quite a simple metric for qualitative analysis of the secondary structures and distinguishing between amino-acids, but is not suitable for the analysis of the structural changes within one system due to the small number of input parameters, and requires the extraction of specific information such as dihedral angles. The modern approach for visualising the complex conformational space in material science is to use machine learning techniques for dimensionality reduction that rely on introducing suitable molecular descriptors of the whole system and introducing a metric in high-dimensional space. The main properties of such descriptors should be (i) invariance to transformations such as translations, rotations and permutations of atom indexing; (ii) uniqueness that implies that systems different in

structure will be mapped in different representations; (iii) Continuity with respect to changes in atomic coordinates, which is required for stability of ML models and (iv) generality for the ability to describe any system [100].

Different molecular descriptors are used in computational chemistry for representing molecular systems, but most of them do not fulfil all the requirements listed above. For example, descriptors widely used in chemoinformatics such as Simplified molecular-input line-entry system (SMILES) [101], International Chemical Identifier (InChI) [102] that encode in a one-line notation the connectivity, the bond type, and the stereochemical information and fingerprints such as Extended-connectivity fingerprints (ECFPs) [103] violate (ii) and (iii) due to lack of information about the spatial arrangement of atoms. Including the spatial 3D information can be done by using Cartesian coordinates and representation on internal coordinates, but both violate requirement (i). The field of developing molecular descriptors is quite active, with the Coulomb matrix [104], bag of bonds (BoB) [105], many-body tensor representation (MBTR) [106, 107], and bonds angles machine learning (BAML) [108] recently introduced. One of the descriptors that satisfy all the requirements above and can capture local changes of the environment is smooth overlap of atomic positions (SOAP)[109, 110], which is a general representation where the atom-centred local neighbourhood is a sum of Gaussians located at atoms within the local environment. The density is expanded in orthogonal radial, and spherical harmonics basis functions [111]. This descriptor was successfully applied in the visualisation of conformational spaces of biomolecules [109, 110, 112–115]. The overall performance of SOAP descriptors means it appears to be becoming increasingly popular compared to other descriptors [107, 116, 117]. With these descriptors, similarities between atomic configurations can be formulated [107] and dimensionality reduction techniques can be applied [118]. Such techniques were applied for analysis of the MD trajectories [112] and of the AA datasets [110].

1.3.5 Overview of the thesis

In this thesis, we present one of the most extensive and accurate studies of adsorbed AAs (with use of DFT) in the literature up to date. Global structure search of systems with large conformational space is one of the bottlenecks in modern computational studies, and one of the parts of this thesis is explicitly dedicated to this problem.

This thesis is divided into five main chapters. The second chapter is dedicated to the theoretical foundation, mainly to the electronic structure calculation methods used in the thesis. The third chapter is also theoretical and describes the methods for investigating and analyzing conformational spaces of flexible molecules.

The fourth chapter describes the work that was done to investigate the conformational space changes of Arg and its protonated counterpart Arg-H^+ after adsorption on three noble metallic surfaces [83]. Arg was chosen as a good testbed since it is small enough to be treatable using DFT and at the same time challenging enough due to a very flexible side-chain which allows for hundreds of possible configurations in the gas phase alone. Also, Arg is the most

flexible among AAs [2] and least investigated while adsorbed on metallic surfaces [61]. The analysis of the adsorption behaviour required the creation of a database by performing a large number of geometry optimizations of different conformations and orientations. The analysis of that database includes producing a low-dimensional representation of the conformational spaces using modern dimensionality reduction techniques and following analysis of different patterns of bonding and charge transfer and how it can affect the accessible conformational spaces.

The fifth chapter of the thesis deals with developing the tools for the automated investigation of flexible molecules, which also enables the modeling of self-assembly patterns formed after adsorption. Different geometry optimization algorithms are implemented together with a flexible way of preconditioning the quasi-Newton optimization algorithms in the package. Together, these allow a simplified interface with a wide variety of electronic structure packages ready to sample conformational spaces of flexible molecules with respect to 1D (ions), 2D (surfaces), and 3D (cavities and molecules) fixed frames. Also, it shows the application of the package, described in the fourth chapter, where we showcase the structure search algorithm on the di-L-alanine molecule adsorbed on Cu(110) surface and compares our findings with experimental results.

[50mm] Sitting on the shoulders of giants

2 Theoretical methods

The essential ideas, notations, and approximations utilised in this thesis are introduced in this chapter. We will explain and motivate the central approximation in condensed matter physics and quantum chemistry after first explaining the many-body problem, which addresses the electrons as quantum objects. Next, we will present the theoretical technique that will play a major role in this thesis: the density-functional theory (DFT). The fundamentals of DFT will be covered, including a discussion of the most common approximations and modern developments, such as the inclusion of the long-range correlation interactions. This chapter will also discuss the basics of theoretical production of STM images and the calculation of charge transfer effects. Also, a short overview of the FF techniques will be covered at the end.

2.1 The many-body problem

A system composed of nuclei and electrons may be formally characterized in quantum mechanics by solving the time-independent Schrödinger equation. It's non-relativistic form is given by:

$$\hat{\mathrm{H}}\Psi = \mathrm{E}\Psi, \tag{2.1}$$

where $\hat{\mathrm{H}}$ represents the non-relativistic time-independent Hamiltonian operator, E denotes the total energy of the system, and Ψ is the many-body wave function of the system that depends on electronic and nuclear degrees of freedom $\Psi = \Psi(\mathbf{r}_i; \mathbf{R}_I)$, where $\mathbf{r}_i$ and $\mathbf{R}_I$ correspond to the electron and nuclei position vectors. Hamiltonian $\hat{\mathrm{H}}$ in the absence of an external electromagnetic field consists of five terms:

$$\hat{\mathrm{H}} = \hat{T}_{\mathrm{n}} + \hat{T}_{\mathrm{e}} + \hat{V}_{\mathrm{e\text{–}e}} + \hat{V}_{\mathrm{ext}} + \hat{V}_{\mathrm{n\text{–}n}}, \tag{2.2}$$

where $\hat{T}_{\mathrm{n}}$ and $\hat{T}_{\mathrm{e}}$ are the nuclear and electronic kinetic energy operators, $\hat{V}_{\mathrm{e\text{–}e}}$ and $\hat{V}_{\mathrm{n\text{–}n}}$ are the electron–electron and nuclear–nuclear Coulomb repulsion, and $\hat{V}_{\mathrm{ext}}$, is the electron–nuclear Coulomb attraction. For simplicity atomic units are used where the electron mass m_{e}, the elementary charge e, the reduced Planck constant $\hbar$ as well as the vacuum permittivity

factor $4\pi\epsilon_0$ are all set to unity. The Hamiltonian in Eq. 2.2 can be written explicitly as

$$\hat{\mathrm{H}} = -\frac{1}{2}\sum_{I=1}^{M}\frac{\nabla_I^2}{M_I} - \frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} - \sum_{i=1}^{N}\sum_{I=1}^{M}\frac{Z_I}{r_{iI}} + \sum_{I=1}^{M}\sum_{J>I}^{M}\frac{Z_I Z_J}{R_{IJ}}, \tag{2.3}$$

where the indices i, j refer to indexes of N electrons and I, J are indexes of M nuclei so that Z_I denote the nuclear charge, M_I is the nuclear mass, $r_{ij} = \left|\mathbf{r}_i - \mathbf{r}_j\right|$, $r_{iI} = |\mathbf{r}_i - \mathbf{R}_I|$ and $R_{IJ} = \left|\mathbf{R}_I - \mathbf{R}_J\right|$ represent the electron-electron, electron-nucleus and nucleus-nucleus distances respectively. In the above equation, the Laplacian operators ∇_i^2 and ∇_I^2 include differentiation with respect to the ith electron and Ith nucleus coordinates.

Since the nuclei and electrons are not constrained in general, the solution of Eq. 2.1 implies a problem of $3N + 3M$ ($4N$ considering the spin variables) degrees of freedom. Since exact analytical solutions to the Eq. 2.1 are only accessible in a few limited cases, the following sections discusses approximations that allow obtaining a numerical solution for the systems relevant to the scope of this work.

2.2 The Born-Oppenheimer approximation

The Born-Oppenheimer (BO) approximation is a fundamental concept in electronic structure theory that provides a significant simplification of Eq. 2.1 by decoupling the dynamics of electrons and nuclei.

Because nuclei are significantly heavier than electrons for example, for a single proton, the ratio is

$$\frac{m_\mathrm{e}}{M_\mathrm{p}} \approx \frac{1}{1836} \ll 1, \tag{2.4}$$

to a fair approximation, electrons in a molecule can be thought to be travelling in a field of fixed nuclei. Within this approximation, the first term of Eq. 2.3, the nuclei's kinetic energy, may be ignored, and the last component of Eq. 2.3, the nuclei's repulsion, can be assumed to be constant. Any constant introduced to an operator increases the operator's eigenvalues and does not influence the eigenfunctions of the operator. The remaining components in Eq. 2.3 are known as the electronic Hamiltonian $\hat{\mathrm{H}}_e$, which only depends parametrically on the nuclear coordinates $\mathbf{R}$:

$$\hat{\mathrm{H}}_\mathrm{e}(\mathbf{R}) = \hat{T}_\mathrm{e} + \hat{V}_\mathrm{e\text{–}e} + \hat{V}_\mathrm{ext}. \tag{2.5}$$

that describes the motion of N electrons in a field of M point charges. The time-independent Schrödinger equation for electronic part, considering ν electronic eigenfunctions for $\hat{\mathrm{H}}_e$ will be:

$$\hat{\mathrm{H}}_\mathrm{e}\psi_\nu(\mathbf{r};\mathbf{R}) = E_\nu^\mathrm{e}(\mathbf{R})\psi_\nu(\mathbf{r};\mathbf{R}),, \quad \text{with } \nu = 1,\ldots,N \tag{2.6}$$

where E_ν^e is the electronic energy of the electron that moves in the field created by the point charges produced by the given configuration of the nuclei. The total wavefunction Ψ can be

expanded into a nuclear χ and an electronic part ψ as:

$$\Psi(\mathbf{r},\mathbf{R}) = \sum_{\nu} \chi_{\nu}(\mathbf{R})\psi_{\nu}(\mathbf{r};\mathbf{R}), \tag{2.7}$$

where $\chi_{\nu}(\mathbf{R})$ are functions of the nuclear positions and represent the coefficients of such expansion. With the entire Schrödinger Eq. 2.1 and a left-side multiplication by $\langle\psi_{\mu}(\mathbf{r};\mathbf{R})|$ followed by integration over the electronic coordinates and application of chain rules, the equation becomes [119]:

$$\begin{aligned} E\chi_{\mu}(\mathbf{R}) = &\left[\hat{T}_n + \hat{V}_{n-n} + E^{\mathrm{e}}_{\mu}\right]\chi_{\mu}(\mathbf{R}) - \\ &- \sum_{\nu}\sum_{I}\frac{1}{2M_I}\left[2\left\langle\psi_{\mu}(\mathbf{r};\mathbf{R})|\nabla_I|\psi_{\nu}(\mathbf{r};\mathbf{R})\right\rangle\nabla_I + \left\langle\psi_{\mu}(\mathbf{r},\mathbf{R})\middle|\nabla_I^2\middle|\psi_{\nu}(\mathbf{r};\mathbf{R})\right\rangle\right]\chi_{\nu}(\mathbf{R}) \end{aligned} \tag{2.8}$$

where E now is the total energy of the system, where we applied the property

$$\left\langle\psi_{\mu}(\mathbf{r};\mathbf{R})\,|\,\psi_{\nu}(\mathbf{r};\mathbf{R})\right\rangle = \delta_{\mu\nu} \tag{2.9}$$

The off-diagonal elements of the last two terms in the Eq. 2.8 are called non-adiabatic contributions, describing the interaction between different electronic states. Within the BO approximation, these terms are assumed to be zero:

$$\left\langle\psi_{\mu}|\nabla_I|\psi_{\nu}\right\rangle = \left\langle\psi_{\mu}\middle|\nabla_I^2\middle|\psi_{\nu}\right\rangle = 0 \text{ for } \mu \neq \nu, \tag{2.10}$$

which means that the atomic motion does not induce electronic excitations. The elements for $\left\langle\psi_{\mu}\middle|\nabla_I^2\middle|\psi_{\mu}\right\rangle$ can be also neglected in comparison with electronic ones, since electron to proton mass ratio is at least of the order 10^{-4} (Eq. 2.4). With all these assumptions, the BO PES, where the nuclei move, is defined as

$$V_{\mu}^{\mathrm{BO}}(\mathbf{R}) = \hat{V}_{\mathrm{n-n}}(\mathbf{R}) + E^{\mathrm{e}}_{\mu}(\mathbf{R}), \tag{2.11}$$

where $\mu = 0$ is the electronic ground-state. It has to be noted that the BO approximation fails when a transition between electronic states occurs. For example, when examining organic molecules and UV photoabsorption, a conical intersection between the electronic ground and excited states can be observed depending on the geometry of the molecule. In this situation, the excited molecule undergoes an ultrafast non-adiabatic internal conversion, which does not result in the emission of radiation, and violates the condition in Eq. 2.10 [120].

2.3 Density Functional Theory

The Nobel Prize in Chemistry 1998 was divided equally between Walter Kohn "for his development of the density-functional theory" and John A. Pople "for his development of computational methods in quantum chemistry". The initial work on Density Functional Theory (DFT) was reported in two of Kohn's publications with Pierre Hohenberg in 1964 [121]

and with Lu J. Sham in 1965 [122]. The main advantage of the DFT approach is its compromise between accuracy and computational cost, which made it a very popular and common technique for the calculation of the properties of different systems from condensed matter to isolated molecules. DFT is an electronic-structure calculation method that replaces the N-electron wave-function ψ_{e} with the electron density $n(\mathbf{r})$ that depends only on 3 spatial coordinates. From an N-electron wavefunction, the electron density can be obtained by integration:

$$n_0(\mathbf{r}) = N \int \left|\psi_0(\mathbf{r}, \mathbf{r}_2, \ldots, \mathbf{r}_N)\right|^2 d\mathbf{r}_2 \ldots d\mathbf{r}_N, \tag{2.12}$$

where N is the number of electrons in the system and the dependency on the spin is omitted for simplicity.

The foundation of DFT began from the Thomas-Fermi model [123, 124], where the energy of the system was expressed in terms of electron density based on the homogeneous electron gas. Based on this idea, Hohenberg and Kohn developed the mathematical basis of modern DFT that proves that all the ground-state properties of the system can be expressed as functionals of the electronic density [125].

2.3.1 The Hohenberg-Kohn theorems

The electron density contains all necessary information about the system, as was shown by Hohenberg and Kohn in 1964 through two theorems:

1. The external potential $v_{\mathrm{ext}}(\mathbf{r})$ is a unique functional of electron density $n(\mathbf{r})$. This means that the electron density, in fact, uniquely determines the Hamiltonian and thus all electronic properties of the system, making it possible to describe the properties of the system as a functional of $n(\mathbf{r})$. The total energy of the system has the form

$$E[n(\mathbf{r})] = \int v_{\mathrm{ext}}(\mathbf{r}) n(\mathbf{r}) d\mathbf{r} + F[n(\mathbf{r})] \tag{2.13}$$

The first term depends on the actual system of interest under investigation and includes the electron-nuclei attraction. The second term is universal in the sense that its form does not depend on the number of electrons, nuclei positions and their charges:

$$F[n(\mathbf{r})] = T[n(\mathbf{r})] + E_{e-e}[n(\mathbf{r})], \tag{2.14}$$

where $T[n(\mathbf{r})]$ is the kinetic-energy functional and $E_{e-e}[n(\mathbf{r})]$ is the electron-electron interaction functional.

2. The electron density that minimises the value of the energy functional is the exact ground-state density n_0:

$$E[n_0] \leq E[n(\mathbf{r})] \tag{2.15}$$

The proofs of the two Hohenberg-Kohn theorems are straightforward and can be found

elsewhere [126]. Elimination of the restriction to non-degenerate ground-states was provided by Levy-Lieb [125]. However, these theorems do not give a practical method for solving the equations and obtaining electron densities.

2.3.2 The Kohn-Sham equations

The idea of the Kohn-Sham scheme is to define a non-interacting system of N electrons whose ground-state electron density exactly equals the ground-state density of real, interacting system n_0. The density is then constructed as a sum of single-particle Kohn-Sham (KS) orbitals:

$$n(\mathbf{r}) = \sum_{i}^{N} \phi_i^*(\mathbf{r})\phi_i(\mathbf{r}). \tag{2.16}$$

The KS theorem ensures the existence of an effective external potential such that a system of non-interacting electrons will produce exactly the same ground-state electron density. Then one can rewrite the total energy functional in a way that includes well-defined terms:

$$E[n(\mathbf{r})] = T_S[n(\mathbf{r})] + V_{\mathrm{H}}[n(\mathbf{r})] + E_{\mathrm{xc}}[n(\mathbf{r})] + \int v_{\mathrm{ext}}(\mathbf{r})n(\mathbf{r})d\mathbf{r}, \tag{2.17}$$

where T_S is the kinetic energy operator of non-interacting system and $V_{\mathrm{H}}[n(\mathbf{r})]$ is the Hartree term:

$$\mathrm{T_S}[n(\mathbf{r})] = -\frac{1}{2}\sum_{i}^{\mathrm{N}} \left\langle \phi_{\mathrm{i}}^*(\mathbf{r}) \middle| \nabla^2 \middle| \phi_{\mathrm{i}}(\mathbf{r}) \right\rangle, \tag{2.18}$$

$$V_{\mathrm{H}}[n(\mathbf{r})] = \frac{1}{2}\iint \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}d\mathbf{r}', \tag{2.19}$$

where the factor $1/2$ is present to avoid double counting. The first three terms of Eq. 2.17 are the functional $F[n(\mathbf{r})]$, and the quantum-mechanical many-body complexity is described by $E_{\mathrm{xc}}[n(\mathbf{r})]$, the exchange-correlation (XC) functional that is unknown. $E_{\mathrm{xc}}[n(\mathbf{r})]$ includes the difference between the true kinetic energy $T[n(\mathbf{r})]$ and the kinetic energy of the non-interacting system, as well as all the non-classical electron-electron interactions:

$$E_{\mathrm{xc}}[n(\mathbf{r})] = T[n(\mathbf{r})] - T_S[n(\mathbf{r})] + V_{e-e}[n(\mathbf{r})] - V_{\mathrm{H}}[n(\mathbf{r})]. \tag{2.20}$$

As in the Hartree-Fock method, applying the variational principle and minimizing Eq. 2.17 with respect to the electron density, with the constraint that any electron density must conserve the total number of electrons, yields the set of single-particle KS equations [127]:

$$\hat{h}^{KS}\phi_i(\mathbf{r}) = \epsilon_i\phi_i(\mathbf{r}) \tag{2.21}$$

$$\left(-\frac{1}{2}\nabla^2 + v_{\mathrm{H}}(\mathbf{r}) + v_{\mathrm{xc}}(\mathbf{r}) + v_{\mathrm{ext}}(\mathbf{r})\right)\phi_i(\mathbf{r}) = \epsilon_i\phi_i(\mathbf{r}) \tag{2.22}$$

$$\frac{\delta V_{\mathrm{H}}[n(\mathbf{r})]}{\delta n(\mathbf{r})} = v_{\mathrm{H}}(\mathbf{r}) = \int \frac{n\left(\mathbf{r}_j\right)}{\left|\mathbf{r}-\mathbf{r}_j\right|} d\mathbf{r}_j, \quad \frac{\delta E_{\mathrm{xc}}[n(\mathbf{r})]}{\delta n(\mathbf{r})} = v_{\mathrm{xc}}(\mathbf{r}), \tag{2.23}$$

where v_{H} is called Hartree potential and v_{xc} is XC potential. Usually these three potential are combined in one effective single-particle potential:

$$v_{\mathrm{eff}}(\mathbf{r}) = v_{\mathrm{H}}(\mathbf{r}) + v_{\mathrm{xc}}(\mathbf{r}) + v_{\mathrm{ext}}(\mathbf{r}). \tag{2.24}$$

Starting with a trial electron density and solving the set of single-particle equations from Eq. 2.22 one can obtain a new set of eigenstates from which to obtain a new density, and continuing this procedure minimizes the total energy self-consistently.

2.3.3 Exchange correlation functionals

Until now DFT in itself is a truly *ab initio* method if the exact form of the XC functional could be written down. Since it is not known, approximations to it have to be made, which gives rise to different density-functional approximation (DFA) that can be separated into different types. The simplest is the local density approximation (LDA). The XC energy functional in LDA is written as:

$$E_{\mathrm{xc}}^{\mathrm{LDA}}[n(\mathbf{r})] = \int \epsilon_{\mathrm{xc}}[n(\mathbf{r})] n(\mathbf{r}) d\mathbf{r}, \tag{2.25}$$

where $\epsilon_{\mathrm{xc}}[n(\mathbf{r})]$ is the XC energy per particle of a uniform electron gas of density $n(r)$. This term can be divided into exchange and correlation terms $\epsilon_{\mathrm{xc}}[n(\mathbf{r})] = \epsilon_{\mathrm{x}}[n(\mathbf{r})] + \epsilon_{\mathrm{c}}[n(\mathbf{r})]$ which leads to

$$E_{\mathrm{xc}}[n(\mathbf{r})] = E_{\mathrm{x}}[n(\mathbf{r})] + E_{\mathrm{c}}[n(\mathbf{r})]. \tag{2.26}$$

The exchange energy of the homogeneous electron gas (HEG) has an analytical form:

$$E_x^{\mathrm{LDA}} = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} \int n^{4/3}(\mathbf{r}) d\mathbf{r}. \tag{2.27}$$

The form of the correlation energy is unknown, but accurate approximations to it obtained from Quantum Monte-Carlo calculations exist [128]. For systems such as bulk metals where the electron density varies very slowly, the LDA is quite a good approximation. However, it is known to fail for cases where the electron density cannot be taken as uniformly distributed.

The generalized gradient-approximated (GGA) functionals are the most straightforward extension of LDA to inhomogeneous systems. This class of XC functionals, also known as semi-local functionals, incorporate the gradient of the electron density $\nabla n(\mathbf{r})$ to account for non-locality:

$$E_{\mathrm{xc}}^{\mathrm{GGA}}[n(\mathbf{r})] = \int f(n, \nabla n) d\mathbf{r} = \int \epsilon_{\mathrm{xc}}(n(\mathbf{r})) F_{\mathrm{xc}}(n(\mathbf{r}), \nabla n(\mathbf{r})) n(\mathbf{r}) d\mathbf{r}. \tag{2.28}$$

Numerous efforts have been made in recent years to design and parametrize a variety of GGA functionals. The most popular GGA functional is the PBE functional [129] which is a non-empirical functional, in the sense that all parameters are basic constants, and there is no parametrization dependence on experimental data. GGA functionals outperform the LDA in terms of total energies, atomization energies, energy barriers, and structural energy differences. When used to analyze the structure of molecules, the GGA functionals produce good results, however, they can greatly underestimate the binding energies of weakly bound systems [130].

Another type of functionals consist of mixing of Hartree-Fock-exchange energy with the exchange and correlation of the semi-local functional proposed by Becke [131]:

$$E_{\mathrm{xc}}^{\mathrm{hybrid}}[n(\mathbf{r})] = \alpha E_{\mathrm{X}}^{\mathrm{HF}}[n(\mathbf{r})] + (1-\alpha)E_{\mathrm{x}}^{\mathrm{GGA}}[n(\mathbf{r})] + E_{\mathrm{c}}^{\mathrm{GGA}}[n(\mathbf{r})], \tag{2.29}$$

where the parameter α regulates the mixing. The exact exchange is taken from Hartree–Fock theory [132]:

$$E_{\mathrm{X}}^{\mathrm{HF}} = -\frac{1}{2}\sum_{i,j}\iint \phi_i^*(\mathbf{r}_1)\phi_j^*(\mathbf{r}_2)\frac{1}{r_{12}}\phi_j(\mathbf{r}_1)\phi_i(\mathbf{r}_2)\,d\mathbf{r}_1 d\mathbf{r}_2 \tag{2.30}$$

There are hundreds of different functionals nowadays [133] and an informal classification, where XC functionals of similar capabilities are placed at the rungs of the "Jacob's ladder" was proposed by Perdew [134]. Comprehensive information about different types of functionals can be found in the literature [135]. The functional that will be mostly used in this thesis is PBE and in some cases PBE0 [136, 137]. For PBE, the XC functional is expressed as

$$E_{\mathrm{xc}}^{\mathrm{PBE}}[n(\mathbf{r})] = E_{\mathrm{x}}^{\mathrm{PBE}}[n(\mathbf{r})] + E_{\mathrm{c}}^{\mathrm{PBE}}[n(\mathbf{r})], \tag{2.31}$$

where the exchange functional $E_{\mathrm{x}}^{\mathrm{PBE}}[n(\mathbf{r})]$ is

$$E_{\mathrm{x}}^{\mathrm{PBE}}[n(\mathbf{r})] = \int n(\mathbf{r})\epsilon_{\mathrm{x}}^{\mathrm{LDA}}[n(\mathbf{r})]F_{\mathrm{x}}(s)\mathrm{d}\mathbf{r}, \tag{2.32}$$

where

$$\epsilon_{\mathrm{x}}[n(\mathbf{r})] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} n(\mathbf{r})^{1/3} \tag{2.33}$$

is the exchange energy density in the uniform electron gas (see Eq. 2.27) with

$$n(\mathbf{r}) = \frac{3}{4\pi}\frac{1}{r_s^3}, \tag{2.34}$$

where r_s denotes the radius of a sphere that contains one electron on average. $F_{\mathrm{x}}(s)$ denotes the GGA enhancement factor depending on a dimensionless density gradient s which is defined as $s = |\nabla n(\mathbf{r})|/(2k_F n(\mathbf{r}))$, where $k_F = \left(3\pi^2 n(\mathbf{r})\right)^{1/3}$ is the Fermi wave vector. The

enhancement factor $F_{\mathrm{X}}(s)$ has to satisfy a formal conditions [129] and is expressed as

$$F_{\mathrm{x}}(s) = 1 + \kappa - \frac{\kappa}{1 + \mu s^2/\kappa}, \tag{2.35}$$

with $\mu = \beta\left(\pi^2/3\right)$, $\beta = 0.066725$, and $\kappa = 0.804$.

The correlation energy in PBE is expressed as the local correlation plus a correction term $H(r_s, \zeta, t)$ [129] and has the following form

$$E_{\mathrm{c}}^{\mathrm{PBE}}[n(\mathbf{r})] = \int d\boldsymbol{r}\, n(\mathbf{r})\left[\varepsilon_{\mathrm{c}}^{\mathrm{LDA}}(r_s, \zeta) + H(r_s, \zeta, t)\right] \tag{2.36}$$

where $\varepsilon_{\mathrm{c}}^{\mathrm{LDA}}$ is the correlation energy density in PW-LDA approximation [138], ζ is the magnetization density and t is dimensionless gradient (See details in Ref. [129, 139]).

The functional PBE0 mixes $a_0 = 0.25$ of exact exchange (E EX) to the PBE functional, having the form:

$$E_{xc}^{PBE0} = a_0 E_X^{HF} + (1 - a_0) E_x^{PBE} + E_c^{PBE}, \tag{2.37}$$

where the value $a_0 = 0.25$ was chosen based on considerations from fourth order many-body perturbation theory [140].

2.4 Long-range van der Waals interactions

Even though the exact DFT would include all correlation effects, the approximations representing the state-of-the-art density functionals are typically unable to describe dispersion and non-local correlation effects by construction [141]. However, the accurate incorporation of weak vdW interactions are especially crucial for calculation of the properties of such systems as biomolecules [142–144], molecular crystals [145, 146] and interface systems [130, 147–154] due to their collective nature. Even if after adsorption the molecule covalently binds to the surface, the accurate description of the vdW interactions are crucial for such kind of systems that makes it possible to obtain deviations in theoretical adsorption heights within 0.1-0.2 Å within experimental values [78]. A theoretically accurate method for a description of the vdW interactions was recently developed and takes into account electronic screening and the many-body nature of the dispersion term [155].

There are many groups working on inclusion of the vdW corrections and introduction to different approaches that also can be classified in the similar way as well-known “Jacob’s ladder” of functionals introduced by Perdew [134] can be found in the literature [156]. One of the most wide-spread way to account for vdW interactions nowadays are so-called pairwise-additive dispersion correction schemes, where vdW energies are calculated analytically after the convergence of the electronic self-consistency cycle [157–164]. The total energy in this case will be:

$$E_{\mathrm{tot}} = E_{\mathrm{DFA}} + E_{\mathrm{vdW}}, \tag{2.38}$$

where E_{DFA} is the total energy of the system obtained with particular DFA. The dispersion contribution E_{vdW} is defined as the interaction between mutually induced charge fluctuations arising from the instantaneous quantum mechanical excitations of electrons. At large distances, the dispersion interaction can be expressed via a multipolar expansion of the Coulomb potential, as a series in inverse powers of R and, by taking the first term $1/R^6$ that corresponds to the instantaneous induced dipole–induced dipole interaction that is the main contribution, we get:

$$E_{\mathrm{vdW}} = -\frac{1}{2}\sum_{A\neq B}\frac{C_{6,AB}}{R_{AB}^6}, \tag{2.39}$$

where the indices A and B refer to two different atoms, and the sum runs over all possible combinations of atoms in the system, $C_{6,AB}$ is the dispersion coefficient of the two atoms and R_{AB} is the interatomic distance between them. One drawback of using formula 2.39 is the fact that for small interatomic distances it is clearly diverges, and so the damping function $f_{\mathrm{damp}}\,(R_{AB})$ is needed to remove this divergence and also to minimize the overlap between the short-range contributions of the XC functional and of the vdW correction. In this case the formula for dispersion correction looks like:

$$E_{\mathrm{vdW}} = -\frac{1}{2}\sum_{A\neq B}\frac{C_{6,AB}}{R_{AB}^6} f_{\mathrm{damp}}\,(R_{AB}). \tag{2.40}$$

In the simplest approach the $C_{6,AB}$ coefficients are constant and isotropic. Such methods do not include many-body dispersion effects such as screening in metals [165] and keeping of the C_6 coefficients constant neglect the environmental contributions. Obtaining the C_6 coefficients could involve experimental ionization potentials and polarizabilities [166], however, this imposes a constraint on the list of components that may be handled to those found in organic compounds.

The next step to increase the accuracy of the dispersion correction is to include environment-dependent C_6 corrections where the dispersion coefficient of an atom in a molecule depends on the effective volume of the atom. The most popular schemes developed in this direction are DFT-D3 by Grimme [159], Becke-Johnson model [167] and the method of Tkatchenko and Scheffler (TS) [163]. Grimme's model employs the concept of fractional coordination numbers where the function calculating the number of neighbors continuously interpolates between the tabulated reference values. Becke-Johnson model exploits the fact that around an electron there will be a XC hole that produces non-zero dipole and higher-order electrostatic moments causing polarization in other atoms leading to an attractive dipole-induced dipole interaction.

The way of fitting of the damping function is crucial since it defines the shape of the binding curve that tat has to be compatible to XC functional of choice and to definition of vdW radii of atoms [156] and giving rise to broader family of the different approaches [158, 168–170].

The TS approach is much more cost effective compared to the Becke-Johnson model and uses precalculated C_6 coefficients instead of hole dipole moments. The extension of the vdW-TS method tailored to model interface systems was used in this work and its scope will be described in more details in the further section.

2.5 Tkatchenko-Scheffler vdW method

The energy in the TS method is computed using the formula in Eq. 2.39, which is a sum over pairwise interatomic C_6/R_6 terms. The expression for the isotropic C_6 coefficients that describe the vdW interactions between two well-separated fragments is derived from Casimir-Polder formula [171]:

$$C_{6,AB} = \frac{3}{\pi}\int_0^\infty \alpha_A(i\omega)\alpha_B(i\omega)d\omega, \tag{2.41}$$

where $\alpha_A(i\omega)$ is the average dynamic polarizability for atom A and ω is the excitation frequency. Retaining only the leading term of the Padé [172] series, the polarizability of spherical free atoms can be approximated and gives:

$$\alpha_A^1(\omega) = \frac{\alpha_A^0}{1-(\omega/\omega_A)^2}, \tag{2.42}$$

where α_A^0 is the static polarizability of atom A and ω_A is the effective excitation frequency. After substitution into Eq. 2.41 with the static polarizabilities the integral can be solved analytically and the C_6 coefficient can be written as:

$$C_{6,AB} = \frac{3}{2}\alpha_A^0\alpha_B^0\frac{\omega_A\omega_B}{(\omega_A+\omega_B)}. \tag{2.43}$$

For the homonuclear $C_{6,AA}$ coefficient, the effective excitation frequency of atom A can be expressed in terms of the static polarizability:

$$\omega_A = \frac{4}{3}\frac{C_{6,AA}}{\left(\alpha_A^0\right)^2}. \tag{2.44}$$

After that the expression for $C_{6,AB}$ can be obtained by substitution of effective excitation frequencies in Equation 2.43:

$$C_{6,AB} = \frac{2C_{6,AA}C_{6,BB}}{\left(\frac{\alpha_B^0}{\alpha_A^0}C_{6,AA}+\frac{\alpha_A^0}{\alpha_B^0}C_{6,BB}\right)}. \tag{2.45}$$

Then, the C_6 coefficients can be accurately computed using the free-atom parameters α_A^0 and $C_{6,AA}$ obtained from from high-level self-interaction corrected time-dependent DFT reference data [173].

from high-level self-interaction corrected TDDFT reference data For the atoms inside a molecule or solid the proceeding formulation can be adapted to make the TS scheme environment-dependent by introducing the proportional coefficient k, which comes from assuming that the polarizability depends linearly on volume [174]: $k_A^{\text{free}}\,\alpha_A^{\text{free}} = V_A^{\text{free}}$, where "free" refers to free atoms. By obtaining the effective volume of the atom inside a molecule or solid the parameter k can be computed as ratio between effective volume and its free value in order to rescale all the quantities introduced earlier. In the TS scheme the effective volume is obtained from the electron density of the system and the Hirshfeld partitioning of the density (via Hirshfeld weight $w_A(\mathbf{r})$) [175]:

$$\frac{k_A^{\text{eff}}\alpha_A^{\text{eff}}}{k_A^{\text{free}}\alpha_A^{\text{free}}} = \frac{V_A[n(\mathbf{r})]}{V_A^{\text{free}}} = \frac{\int r^3 w_A(\mathbf{r}) n(\mathbf{r}) d\mathbf{r}}{\int r^3 n_A^{\text{free}}(\mathbf{r}) d\mathbf{r}} = \gamma_A[n(\mathbf{r})], \tag{2.46}$$

where the electron density $n(\mathbf{r})$ is taken from DFT calculations, $n_A^{\text{free}}(\mathbf{r})$ is the free atom spherically averaged reference density and $r = |\mathbf{r} - \mathbf{R}_A|$ is the distance between the nucleus of atom A and the point $\mathbf{r}$. The effective quantities are then determined from the free ones as:

$$\alpha_A^{0,\text{eff}} = \gamma_A[n(\mathbf{r})]\alpha_A^{0,\text{ free}}, \tag{2.47}$$

$$C_{6,AA}^{\text{eff}} = \left(\gamma_A[n(\mathbf{r})]\right)^2 C_{6,AA}^{\text{free}}, \tag{2.48}$$

$$R_A^{0,\text{eff}} = \left(\gamma_A[n(\mathbf{r})]\right)^{1/3} R_A^{0,\text{ free}}, \tag{2.49}$$

where the R is vdW radius. The TS scheme was tested on a database of 1225 intermolecular C_6 pairs and showed a mean absolute error of 5.5% compared to experimental results irrespective of the employed XC functional [163].

As was mentioned above, the sum of pairwise $C_{6,AB}/R_{AB}^6$ terms diverges for small interatomic distances and the damping function has to be introduced (Eq. 2.40). The damping function in the case of the TS method is a Fermi-type function:

$$f_{damp}^{AB}\left(R_{AB}, R_{AB}^0[n(\mathbf{r})]\right) = \frac{1}{1+\exp\left[-d\left(\frac{R_{AB}}{s_R R_{AB}^{0,\text{ eff}}[n(\mathbf{r})]}-1\right)\right]}, \tag{2.50}$$

where R_{AB} is the interatomic distance, $R_{AB}^{0,\text{ eff}} = R_A^{0,\text{ eff}} + R_B^{0,\text{ eff}}$ is the sum of the vdW radii associated with atoms A and B that depend on the electron density through the effective volume (Eq. 2.49) and parameters d and s_R are empirical values that need to be determined for a given XC functional. The parameters d, that affects the steepness of the damping, and the parameter s_R, that scales the vdW radii and regulates the extent of the vdW correction for a given XC functional, were fitted for different functionals with use of S22 database [176].

2.6 Tkatchenko-Scheffler vdW^{surf} method

In order to include the non-local collective response of the substrate surface in the vdW energy the extension of the TS-vdW scheme (vdW^{surf} [177]) for modelling of interfaces relies on Lifshitz-Zaremba-Kohn (LZK) theory [178, 179] for the vdW interaction between an atom and a solid surface. This leads to a set of C_6 coefficients that incorporate dielectric screening of the bulk, and in the case of solids the reference vdW parameters have to be determined taking into account atom-in-a-solid environmental effects [180]. In LZK theory the atom-surface dispersion interaction beyond the distance of the orbital overlap is given by [179, 181]:

$$E_{\mathrm{vdW}} \simeq -\frac{C_3^{aS}}{(H-H_0)^3}, \tag{2.51}$$

where H is the distance between an adsorbate atom a and the topmost layer of the surface S. The reference plane H_0 can be obtained from the jellium model yeilding $H_0 = h/2$, where h is the interlayer distance of the solid. The term C_3^{aS} describes the dielectric response of the bulk solid to the instantaneous dipole moment of particles and depends on the dipole polarizability $\alpha(i\omega)$ of the adsorbate and dielectric function $\epsilon_S(i\omega)$ of the solid:

$$C_3^{aS} = \frac{1}{4\pi}\int_0^{+\infty} \alpha(i\omega)\left[\frac{\epsilon_S(i\omega)-1}{\epsilon_S(i\omega)+1}\right] d\omega. \tag{2.52}$$

The screening effects inside the bulk are incorporated in the Eq. 2.52 by dependence on the dielectric function $\epsilon_S(i\omega)$. Next step is determination of the vdW interaction between an adsorbate atom a with a solid S by a summation of the pair potentials $-C_6/R^6$ between an atom a and atoms s in the infinite half-space infinite of the solid S. After that, the connection to LZK expression can be achieved by the relation:

$$C_{3,aS} = n_S\left(\frac{\pi}{6}\right)\mathrm{C}_{6,as}, \tag{2.53}$$

where n_S is the number of atoms per unit volume in the bulk of the substrate, and

$$\mathrm{C}_{6,as} = \frac{2\mathrm{C}_{6,aa}\mathrm{C}_{6,ss}}{\frac{\alpha_s^0}{\alpha_a^0}\mathrm{C}_{6,aa} + \frac{\alpha_a^0}{\alpha_s^0}\mathrm{C}_{6,ss}}, \tag{2.54}$$

where the $\mathrm{C}_{6,as}$, α_s^0 and R_s^0 are the new set of parameters that depend on dielectric function $\epsilon_S(i\omega)$ and thus inherit the many-body collective response (screening) of the solid. The only difference from the TS method is that the effective quantities, that were including the effects of polarization with use of the Hirshfeld weight, are now obtained from the LZK parameters and not from the free atom reference. The dielectric function can be computed from first-principles and was shown to reasonably agree with the results obtained from reflection electron energy-loss spectroscopy (REELS) experiments [182]. In case of transition

metals, the inclusion of collective response of the solid leads to reducing the C_6 coefficients by up to a factor of ten compared to reference free atom values [130].

Investigating interface systems, the inclusion of the vdW parameters should only be applied when appropriate: for example inside metal surfaces there are already good approximations from DFA functionals and inclusion of the vdW interactions, even if the results are improved compared to experimental, can be considered as effect of cancellation of the errors [150].

2.7 Basis sets

In order to solve the set of single-particle KS eigenvalue equations (Eq. 2.22) it is a common technique to use basis functions to expand the single-particle orbitals:

$$\phi_\nu(\mathbf{r}) = \sum_n c_{n\nu}\xi_n(\mathbf{r}) \tag{2.55}$$

A basis set allows us to write the Schrödinger equation as a generalized eigenvalue problem:

$$\sum_n h_{mn}c_{n\nu} = \epsilon_\nu \sum_n s_{mn}c_{n\nu}, \tag{2.56}$$

where $h_{mn} = \left\langle \xi_m \middle| \hat{h}^{KS} \middle| \xi_n \right\rangle$ is the matrix element of the Hamiltonian, and $s_{mn} = \langle \xi_m | \xi_n \rangle$ is the overlap matrix element. A suitable choice of basis functions depends on the system under investigation. For this thesis we use the all-electron/full potential Fritz Haber Institute "ab initio molecular simulations" (FHI-aims) code [183, 184], which adopts the tabulated numeric atom-centered orbitals (NAO) basis functions of the form:

$$\xi_i(\mathbf{r}) = \frac{u_i(r)}{r} Y_{lm}(\Omega) \tag{2.57}$$

where the function $u_i(r)$ has radial symmetry and is numerically tabulated and $Y_{lm}(\Omega)$ are the spherical harmonics. The particular form of the NAOs allows to include the radial functions of free-atom orbitals and can be constructed using a Schrödinger-like radial equation:

$$\left[-\frac{1}{2}\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2} + v_i(r) + v_{\text{cut}}(r)\right] u_i(r) = \epsilon_i u_i(r) \tag{2.58}$$

where l is the angular quantum number. The potential $v_i(r)$ defines the shape of $u_i(r)$ and the term $v_{\text{cut}}(r)$ is the confining potential, which ensures a decay to zero of the radial functions. Minimal basis consists of the core and valence functions of spherically symmetric free atoms by setting $v_i(r)$ to the self-consistent free-atom radial potential $v_{\text{at}}^{\text{free}}$. The construction of the accurate and transferable basis sets that allow up meV-level total energy convergence relies on addition of the candidate functions from a large pool of different radial functions (e.g. hydrogen-like, cation-like or atom-like) with different confinement potential to minimal basis set until no further significant improvement on total energy results [185].

The analytical form of the confining potential is not unique and along with a smooth decay, it must ensure that the function and its derivatives do not have any discontinuities. The confining potential in FHI-aims is provided by:

$$v_{\text{cut}}(r) = \begin{cases} 0 & r \leq r_{\text{onset}} \\ \frac{s}{(r-r_{\text{cut}})^2} \exp\left(\frac{w}{r-r_{\text{onset}}}\right) & r_{\text{onset}} < r < r_{\text{cut}} \\ +\infty & r \geq r_{\text{cut}}\,, \end{cases} \tag{2.59}$$

where s is a global scaling parameter and $w = r_{\text{cut}} - r_{\text{onset}}$ sets the width of the region, where potential is defined. The selection of the parameters r_{cut} and r_{onset}, is essential for both the accuracy of the results and numerical efficiency. For example, a large value of r_{cut} would result in extended radial functions, increasing the computational cost of the calculation. Setting r_{onset} to a very small value will result in unphysical results since radial functions will be limited in a very narrow region surrounding the atom.

In the case of periodic systems, the Kohn-Sham Eqs. 2.56 are **k**-space dependent. This leads to separate matrices $h_{mn}(\mathbf{k})$, $s_{mn}(\mathbf{k})$ and solutions $\phi_{\nu,\mathbf{k}}(\mathbf{r})$ that have to be obtained for different **k**-points in the first Brillouin zone. For that Bloch-like generalized basis functions $\varphi_i(\mathbf{r})$ that are centered in unit cells shifted by translation vectors $\boldsymbol{T}(\boldsymbol{N})[\boldsymbol{N} = (N_1, N_2, N_3)]$ are introduced in the code:

$$\chi_{i,\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\boldsymbol{N}} \exp[i\boldsymbol{k}\cdot\boldsymbol{T}(\boldsymbol{N})]\cdot\varphi_i[\boldsymbol{r} - R_A + \boldsymbol{T}(\boldsymbol{N})]\,. \tag{2.60}$$

Such definition brings **k**-dependent matrix elements

$$\begin{aligned} h_{ij}(\boldsymbol{k}) &= \left\langle \chi_{i,\boldsymbol{k}} \left| \hat{h}^{\text{KS}} \right| \chi_{j,\boldsymbol{k}} \right\rangle \\ &= \sum_{\boldsymbol{M},\boldsymbol{N}} \exp\{i\boldsymbol{k}\cdot[\boldsymbol{T}(\boldsymbol{N}) - \boldsymbol{T}(\boldsymbol{M})]\} \left\langle \varphi_{i,\boldsymbol{M}} \left| \hat{h}^{\text{KS}} \right| \varphi_{j,\boldsymbol{N}} \right\rangle \end{aligned} \tag{2.61}$$

with the real-space basis functions $\phi_i(\mathbf{r},\mathbf{M})$ and $\phi_i(\mathbf{r},\mathbf{N})$ that are centered in different unit cells **M** and **N**. In practice, all integration points and pieces of are mapped back to the original unit cell in order to avoid breaking down lattice sums in Eq. 2.61 due to periodicity since the integration volumes could extend over several unit cells in the integrals $\left\langle \varphi_{i,\boldsymbol{M}} \left| \hat{h}^{\text{KS}} \right| \varphi_{j,\boldsymbol{N}} \right\rangle$. Since all basis functions are bounded by the confinement potential, only a finite number of inequivalent real-space matrix elements are non-zero.

2.8 Charge transfer and binding energy calculations

The interaction of individual molecules with metallic surfaces constitutes one of the central topics of surface science partially because experimental techniques such as the STM could be easily operated on conductive substrates. The final electronic structure of interface system can be calculated with accurate computational methods such as DFT. Understanding of the mechanisms that leads to the particular adsorption pattern of the molecule and identifying the molecular donor/acceptor parts can give more insights towards rational design and

self-assembly processes of the interfaces. In this thesis we are interested in both neutral and positively charged molecules adsorbing on metallic surfaces and in this section we would like to address the procedure that we use to investigate adsorption.

While modelling the interface structures we must use periodic boundary conditions (PBC). Organic-inorganic interface systems could incur a dipole moment in the direction perpendicular to the surface due to charge rearrangements at the interface or due to polar adsorbates which leads to appearing of the electric field that generates a potential gradient in the unit cell compensating the potential shift induced by the system's dipole moment. The interaction of the interface dipole with this electric field also leads to charge rearrangements between ends of the entire slab that in turn affects the total energy of the system. The most common way to deal with the spurious polarization is to introduce discontinuity in the electrostatic potential within the vacuum region and referred in the literature as "dipole correction" [186] and to use large vacuum regions since the magnitude of this spurious electric field depends inversely on the thickness of the vacuum region. In FHI-aims, the magnitude of dipole correction is obtained from the gradient of the long-range Hartree potential term of the Ewald sum (which is evaluated in reciprocal space). The surface plane is placed parallel to $x\,y$ plane in the deep vacuum region that is further than 6 Å away from the nearest atom.

Simulation of charged unit cells is required for several physical problems such as dealing with charged defects [187, 188] or when the electron transfer from the adsorbed molecules is quenched and they can exhibit metastable charge-states [189, 190]. This brings the problem that the repeated slab approach imply that all unit cells in the system carry a charge and such a periodic arrangement of charges results in a diverging energy that prevents convergence of the self-consistent field (SCF) algorithm. Basically, there is a Coulomb interaction between the delocalized homogenous background charge and the excess charge that is localized in the slab that significantly contributes to the total energy of the system. The spurious energy contribution originates from the spurious net dipole of the unit cell and, hence, scales linearly with the thickness of the vacuum region. Two types of approaches were developed to deal with such cases. The first class neutralizes the interaction between charged cells perpendicular to the substrate *via a posteriori* correction based on the dielectric profile of the interface [191] or by interfering Poisson equation that describes electrostatic potential [192, 193]. The second class intentionally adds spatially localized countercharges into the system ensuring charge neutrality of the such that leads to the absence of compensating background charge. The virtual crystal approximation [187] provides a fixed number of free charge carriers per volume, the Charge Reservoir Electrostatic Sheet Technique [194] models the countercharges as a charged sheet, which is placed below the substrate and the generalized dipole correction approach [195] introduces a monopole sheet as a "computational electrode" and a dipole layer in the vacuum region.

In our case for adsorption of the both neutral and positively charged species the unit cell is set to have neutral charge. After adsorption of the charged molecules on the surface the charge transfer will occur from surface since it has infinite pool of electrons that will neutralize the

unit cell. This comes from the fact that energy of lowest unoccupied orbital of the positively charged molecules are way below the Fermi energy of the metallic surfaces. Having that, the binding energies for neutral molecule adsorbed on the surface were calculated as

$$E_{\mathrm{b}} = E_{\mathrm{mol@surf}} - E_{\mathrm{surf}} - E_{\mathrm{mol}}, \tag{2.62}$$

where $E_{\mathrm{mol@surf}}$ corresponds to the total energy of the interface, E_{surf} is the total energy of the pristine metallic slab and E_{mol} the total energy of the lowest energy gas-phase conformer.

For charged molecules, we considered the binding energy of a two-step reaction. *First*, the interface is formed between the charged molecule and the clean surface:

$$E_{\mathrm{b1}} = E_{\mathrm{mol^+@surf}} - E_{\mathrm{surf}} - E_{\mathrm{mol^+}}, \tag{2.63}$$

where $E_{\mathrm{mol^+}}$ is the total energy of the most stable gas-phase conformer of the isolated charged molecule. *Second*, an electron from the metal neutralizes the unit cell where the adsorbed molecule is located, yielding

$$E_{\mathrm{b2}} = E_{\mathrm{mol@surf}} - E_{\mathrm{mol^+@surf}} - E_f, \tag{2.64}$$

where E_f corresponds to the Fermi energy of the metallic surface. The final binding energy is thus considered to be

$$E_{\mathrm{b}}^{+} = E_{\mathrm{b1}} + E_{\mathrm{b2}} = E_{\mathrm{mol@surf}} - E_{\mathrm{surf}} - E_f - E_{\mathrm{mol^+}}. \tag{2.65}$$

To address charge rearrangements after adsorption on the surface, we compute the electron density differences for selected with

$$\Delta\rho = \rho_{\mathrm{mol@surf}} - \rho_{\mathrm{surf}} - \rho_{\mathrm{mol}}, \tag{2.66}$$

and in the case of neutral molecule and

$$\Delta\rho^{(+)} = \rho_{\mathrm{mol@surf}} - \rho_{\mathrm{surf}} - \rho_{\mathrm{mol^{(+)}}}, \tag{2.67}$$

in the case of charged molecule. In these expressions, $\rho_{\mathrm{mol@surf}}$ is the total electron density of the interface, ρ_{surf} is the electron density of the slab without molecule, and ρ_{mol} and $\rho_{\mathrm{mol^+}}$ are electron densities of neutral and charged molecules with the same geometries as in interface. The + sign denotes that the final density difference integrates to +1 electron in the case of charged molecule. These densities allow us to identify charge build up on particular functional group, as well as charge transfer to the surface.

2.9 Modelling of the STM images

One of the ways to validate theoretical investigations is to directly compare experimental measurements with theoretically modelled properties of the system. In that respect modelling of STM images can be a very useful tool to identify the system geometry. Using Bardeen's expression [196] one can write the current flowing from a metallic tip to the sample as

$$I_{t\to s} = \frac{2\pi e}{\hbar}\int |M_{ts}|^2 N_t(E-eV)N_s(E)f_t(E-eV)\big[1-f_s(E)\big]\,dE, \tag{2.68}$$

where V is the applied voltage, $N_t(E)$ and $N_s(E)$ are the density of states of the tip and the sample respectively, $f(E)$ is their Fermi-Dirac distribution. The effective matrix element M_{ts} couples a tip wave function, Ψ_t, to a substrate wave function, Ψ_s, by the expression

$$M_{ts} = \frac{\hbar^2}{2m}\int \left(\Psi_t^*\nabla\Psi_s - \Psi_s\nabla\Psi_t^*\right) d\boldsymbol{S}, \tag{2.69}$$

where the integral is taken over a surface separating the tip and sample.

For modelling STM images one of the most widely used approaches is the scheme proposed by Tersoff and Hamman [197]. One of the main assumptions made within this model is that complex electronic structure of the tip is assumed to be simple atomic s-wave-function since only the orbitals that localized at the outermost tip atom are important for tunneling process taking into account that this wave-function decays exponentially into the vacuum. The total current flowing from the tip to the sample within the zero temperature approximation and low bias voltage is:

$$I = \frac{2\pi e^2}{\hbar}V\sum_N |M_{ts}|^2\,\delta(E_s-E_F)\,\delta(E_t-E_F), \tag{2.70}$$

where V is the voltage applied and the energy conservation is ensured by δ-functions.

The advantage of the Tersoff-Hamann theory is that the tip Ψ_t wavefunction can be modelled as a solution in a locally spherical potential with curvature R about its center r and asymptotically the is chosen to have the form of an s-wave. So the matrix element M_{ts} is proportional to the sample wavefunction evaluated at the tip center of curvature ($M_{ts}\propto\Psi_s(r_0)$) leading to:

$$I \propto V N_t(E_F)\sum_s |\Psi_s(r_0)|^2\,\delta(E_s-E_F), \tag{2.71}$$

where the sum represents the local density of states of the sample (LDOS) around the Fermi level evaluated at the tip center.

2.10 Force field methods

In previous sections we addressed the methods for simulations of the interface systems at quantum mechanical (QM) levels of theory of high computational costs, applicable only to systems of few hundreds of atoms. In this section we briefly describe the applications and

limitations of the FF methods that are less accurate but orders of magnitude cheaper to perform and thus enable the simulation of the systems that consist of millions of atoms.

Most commonly within classical FFs PES functions are expressed as a sum of bonded and nonbonded interaction terms. Hence, the description of a of FF is given by its potential energy $E_{\text{pot}}^{\text{FF}}\left(\boldsymbol{R}^N\right)$ that is given as a function of positions $\boldsymbol{R}_1,\ldots,\boldsymbol{R}_N$ the N nuclei of the system is given by

$$E_{\text{pot}}^{\text{FF}}\left(\boldsymbol{R}^N\right) = E_{\text{bonded}}\left(\boldsymbol{R}^N\right) + E_{\text{nonbonded}}\left(\boldsymbol{R}^N\right). \tag{2.72}$$

For example in CHARMM22 [198], one of the popular FF for simulation of biomolecules, the "bonded" terms are of the following form:

$$E_{\text{bonds}}\left(\boldsymbol{R}^N\right) = \sum_{\text{bonds}} \frac{k_r}{2}(R-R_0)^2 \tag{2.73}$$

$$E_{\text{angles}}\left(\boldsymbol{R}^N\right) = \sum_{\text{angles}} \frac{k_\theta}{2}(\theta-\theta_0)^2 \tag{2.74}$$

$$E_{\text{torsions}}\left(\boldsymbol{R}^N\right) = \sum_{\text{torsions}} \frac{k_\tau}{2}(1+\cos(n\tau-\delta)) \tag{2.75}$$

$$E_{\text{impropers}}\left(\boldsymbol{R}^N\right) = \sum_{\text{impropers}} k_\omega(\omega-\omega_0)^2 \tag{2.76}$$

$$E_{\text{Urey-Bradley}}\left(\boldsymbol{R}^N\right) = \sum_{\text{Urey-Bradley}} k_u(u-u_0)^2 \tag{2.77}$$

where n is the multiplicity of the function, δ is the phase shift, k_R, k_θ, k_τ, k_ω, and k_u are the bond, angle, dihedral angle, improper dihedral angle and Urey–Bradley force constants, respectively; R, θ, τ, ω, and u are the bond length, bond angle, dihedral angle, improper torsion angle and Urey–Bradley 1,3-distance, respectively, with the subscript zero representing the equilibrium values for the individual terms. "Nonbonded" interaction terms are included for all atoms separated by three or more covalent bonds and include electrostatic interactions

$$E_{\text{Coulomb}}\left(\boldsymbol{R}^N\right) = \sum_{\text{A B}} \frac{q_A q_B}{R^{AB}}, \tag{2.78}$$

and vdW intra- and inter-molecular interactions

$$E_{\text{vdW}}\left(\boldsymbol{R}^N\right) = \sum_{\text{A B}} \varepsilon_{AB}\left[\left(\frac{R_0^{\text{vdW}}}{R^{AB}}\right)^{12} - 2\left(\frac{R_0^{\text{vdW}}}{R^{AB}}\right)^{6}\right], \tag{2.79}$$

where q_A is the charge of the atom A, R^{AB} is the distance between atoms A and B and ε_{AB} is the energy required to separate the atoms. In the Lennard-Jones (LJ) potential above, the R_0^{vdW} term is not the minimum of the potential, but rather where the LJ potential is zero.

In recent years a lot of effort has been invested to adapt biomolecular FFs to the simulation of interfaces between biomolecules and inorganic materials which resulted in creation of a class of general bio-inorganic FFs. One of these FFs is called INTERFACE FF, in which LJ parameters for eight neutral face-centered cubic (fcc) metals (Ag, Al, Au, Cu, Ni, Pb and Pd) based on experimentally determined densities and surface tensions under ambient conditions have been added to CHARMM22 for the simulation of metal surfaces in contact with biomolecules [87, 199, 200].

One of the greatest limitations of most common FFs is the inability to simulate chemical reactions that involve the formation or dissociation of chemical bonds. Modelling of chemisorption processes require QM or employing of reactive FFs that have to be used with great caution [92, 93]. Second disadvantage is the sensitivity of FF parameters to deviations from the reference state, for which they were derived that imply nontransferability of the parameters to different systems.

For much broader overview of different FFs designed for modelling of the protein-inorganic surface interaction can be found in the review [79]. For modelling of AA adsorption on metallic surfaces using INTERFACE FF we use NAMD package [201] and compare results obtained with DFT in the Section 4.0.5.

[45mm] And to the man he said, “Since you listened to your wife and ate from the tree whose fruit I commanded you not to eat, the ground is cursed because of you. All your life you will struggle to scratch a living from it.” Genesis 3.17 about curse of dimensionality (author’s note)

3 Structure search and analysis of conformational spaces

A major challenge in computational chemistry is the search of low-energy conformers for a given flexible molecule. Organic molecules that are flexible can adopt a number of energetically favourable conformations with varying chemical and physical characteristics (Fig. 3.1 a). As a result, examining the attributes of a single randomly created conformer may result in incorrect results. The environment, as well as interactions with other molecules and surfaces, can all impact the likelihood of a given shape being adopted (Fig. 3.1 b). Further, it has been shown that the bioactive conformation of drug-like molecules can be higher in energy than the respective global minimum [202]. Structures that can be trapped in metastable local minima during growth process, can be accessed at finite temperatures or under pressure. As a result, we aim at not just finding the conformer expressing the PES's global minimum, but at covering relevant portions of the available conformational space.

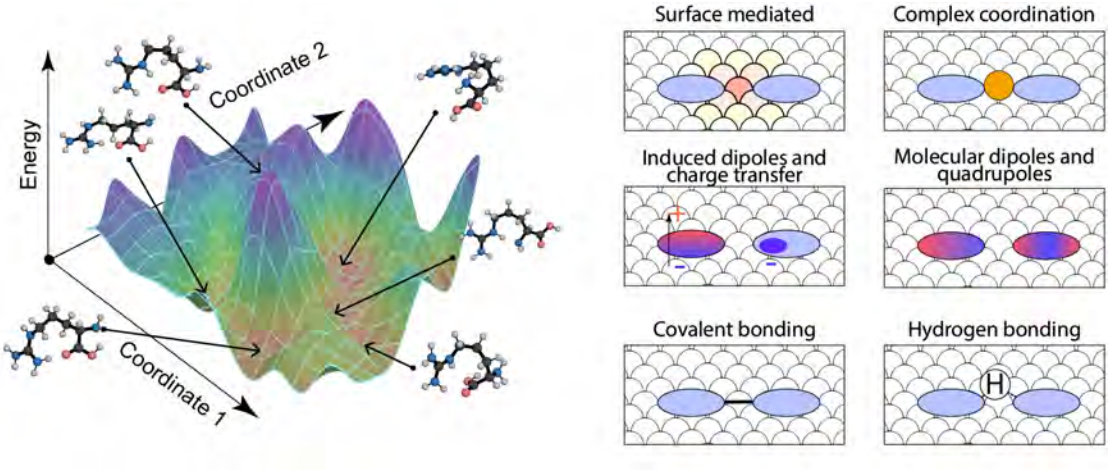


Figure 3.1 – a) Pictorial representation of the multiple local minima of PES of a flexible molecule with respect to arbitrary coordinates. b) Examples of complex interactions that appear during self-assembly processes on the surfaces

3.1 Global structure search techniques

Finding the most stable configurations of assembly of atoms is a challenging problem due to the fact that the number of stationary points in the particular PES can grow exponentially with the number of atoms in the system. Finding the global minimum of the system, in general, requires searching through many local minima which is effectively prohibitive for large systems due to finite computational resources available. There is a broad field of computational search techniques. Below we will briefly mention techniques based on MD and then present a more in-depth characterization of other techniques more directly relevant for this thesis. But first of all, it should be mentioned, that performance of all algorithms that search for the global minimum of an energy function is the same when averaged over all possible energy functions - this is known in literature as "no free lunch theorem" [203]. This implies, that there is no possible way to find algorithm that would perform better than another in all scenarios.

MD-based techniques

Replica exchange molecular dynamics (REMD) simulations combine MD simulation with the Monte Carlo algorithm and are used to sample the configurational space of a system, e.g. at different temperatures or with a different Hamiltonian [204]. Structures locked in local energy minima can traverse the energy landscape by exchanging the replicas, improving Boltzmann-weighted sampling. Unlike a standard MD simulation, REMD allows sampling various configurations in different potential wells separated by huge energy barriers.

Umbrella sampling is another standard method for enhancing the sampling of configurational space of the system [205]. This technique defines a reaction coordinate as a link between two thermodynamic states. The reaction coordinate is usually determined based on a distance or an angle. The reaction coordinate is then divided into windows, each exposed to a bias (umbrella) potential. Each window has its simulation to sample the area around the associated coordinate point. The simulations are then reweighted to account for the biased ensembles using, for example, the weighted histogram analysis method (WHAM) [206], or its generalization [207]. Umbrella sampling method, for example, was used for simulations of adsorption of AA side chain analogues on the TiO_2(100) surface [208] using FF models. There are subtleties in determining the best computationally efficient approach to apply the umbrella sampling method, as outlined in the book [209].

MD simulations can be also evaluated using classical FFs followed by static DFT refinement of the obtained data [91, 210, 211]. The main disadvantage of such approaches is that the PES obtained from a FF or DFT can be very different, which will result in the need to reevaluate all the sampled geometries using more accurate methods in order to obtain correct hierarchy [212, 213].

Other techniques

One of the simplest methods to explore PES that allows to effectively overcome potential energy barriers is a random search. Random search implies that next trial structure is

not dependent on the information that was already accumulated during the search. Of course, simply creation of assembly of atoms and calculation of their energies would be far from effective. Concerning investigation of adsorbates on surfaces, to make such a strategy efficient, one has to impose limits on the generated structures, by creating only "sensible" structures. Structures that have some of the atoms that are very close to each other cannot be in the local minimum. Those structures should not be investigated and this already significantly decreases number of candidates that have to be calculated. Having that one can apply criteria for non-bonded atoms, for example, their vdW radii should not overlap which prevents modelling of unwanted chemical reactions. Concerning molecular systems and surfaces, one already has *a priori* information on bonds in the system, most of which should not change. After that different structures can be generated and followed with geometry optimization (See Sec. 3.2) to find local minima. Application of the random structure search to investigate organic-inorganic materials, in particular, the procedure of generating different molecular conformers with respect to specified surrounding, will be discussed in Chapter 5. Random structure search has been effectively utilized in the field of materials research, demonstrating that even random sampling has a decent probability of identifying low-energy basins[214–216]. The advantages of such strategy are the small amount of the parameters that have to be set for the investigation of PES and covering broader volume of PES without biasing of the search itself. Many other methods to some extent depend on routines for producing random structures. More sophisticated techniques that were developed in recent years by introducing bias for the structure search that aims to find global minima faster [217].

For example, the heuristic technique ranks candidates in a search at each branching step depending on the information provided to determine which branch to take next. One of the most famous representatives among the class of heuristic methods are genetic algorithms (GA) [218]. Based on the principle of evolution, new candidates have to exhibit high fitness with respect to some function in order to survive in the natural selection. As a result, they must devise strategies to maintain the diversity of genes within their populations, two of which are well-known: chromosomal crossover during mating and mutation in-place. Individuals with poor fitness will be removed after many generations of natural selection, and the general fitness of the entire population will improve as a result of the selection process. In principle, the "genetic code" is chosen as an array of relevant parameters for the system at hand. Initially, a random set of candidate structures is generated then a fraction of the population is selected with bias towards fittest, then those structures that were selected are paired up for "recombination" and mutation step may be performed. New candidate structures are added to the pool and the whole process is repeated. GA were applied to investigations of molecular structures, clusters and crystals [219–221].

Another popular family of global geometry optimization techniques include Monte Carlo based approaches, like basin-hopping [222] and minima hopping [223]. The basic strategy in these algorithms is to find one of the local minimum of the system and then with the trial moves escape from the basin and reach another local minimum. Then, with some

probability that depends on a specified effective temperature use this new minimum as new starting point. An example of such algorithm that works on internal coordinates plus local translation and rotation of independent geometrical subunits, was demonstrated for molecules adsorbed on surfaces and interfaces [224].

Also, ML algorithms can be used to approximate the PES [225, 226]. In the active learning family, Bayesian optimization search techniques are used to fit a surrogate PES to the data points acquired from DFT calculations, and then improve this potential by acquiring new data points at places where the exploratory lower confidence bound acquisition function is minimized [227–230].

3.2 Geometry optimizations on the Born-Oppenheimer Potential Energy Surface

In the algorithms discussed in the last part of the previous section, the exploration of the conformational space relies on the creation of sample points, followed by local geometry optimizations. Finding local minima requires the computation of the derivatives of the energy with respect to atomic positions (forces) and for more sophisticated and efficient methods estimation of the second derivatives – the Hessian matrix. Here the basic concepts of local structure optimization will be described, introducing the trust region and line search methods, following the textbook by Nocedal and Wright [231]. The starting point to perform local geometry optimization is obtaining the atomic forces, that are defined by $-dE/d\mathbf{R}$. Within density-functional-theory the energy derivative with respect to atom γ is

$$\frac{dE}{dR_\gamma} = \frac{\partial}{\partial R_\gamma}\left[E_v[n(\mathbf{r})] + \frac{1}{2}\sum_{\alpha,\beta}^{N_a}\frac{Z_\alpha Z_\beta}{|R_\alpha - R_\beta|}\right] + \int \frac{\delta E_v[n(\mathbf{r})]}{\delta n(\mathbf{r})}\frac{\partial n(\mathbf{r})}{\partial R_\gamma}d^3r, \tag{3.1}$$

where the implicit R_γ-dependence of the electron density is taken into account in the second term and the only term that explicitly depends on R_γ in $E_v[n(\mathbf{r})$ is the external potential. Computations of the forces that arise by embedding each nucleus into the electrostatic fields of the electron density and all other nuclei, corresponding to the first term of Eq. 3.1, are performed using the Hellmann–Feynman expression [183]:

$$f^\gamma_{HF} = \sum_{\alpha,\alpha\neq\gamma}^{N_A} Z_\gamma Z_\alpha \frac{R_\gamma - R_\alpha}{|R_\gamma - R_\alpha|^3} - \int n(\mathbf{r})\frac{Z_\gamma(R_\gamma - \mathbf{r})}{|R_\gamma - \mathbf{r}|^3}d^3r. \tag{3.2}$$

In codes such as FHI-aims where an atom centered basis set is employed, the basis functions φ_i "move" with R_γ, which leads to additional force contributions (Pulay forces) that arise from the second term in Equation 3.1

$$f^\gamma_{\text{Pulay}} = -\int \frac{\partial E_v[n(\mathbf{r})]}{\partial n(\mathbf{r})}\frac{\partial n(\mathbf{r})}{\partial R_\gamma}d^3r = -2\,\text{Re}\sum_i f_i\left\langle \frac{\partial \varphi_i}{\partial R_\gamma}\middle| -\frac{1}{2}\nabla^2 + v_{\text{KS}} - \epsilon_i\middle|\varphi_i\right\rangle, \tag{3.3}$$

where f_i are the occupation numbers, ϵ_i are the KS eigenvalues. For the details how to take into account additional contributions such as grid effects and multipole correction to the Hartree potential we refer to the original paper of FHI-aims [183].

3.2.1 Local minima finding

Apart from MD, any structure search technique heavily relies on geometry optimization routines that, use energies and forces of the system to find the nearest local minimum of the initial input geometry. The iterative nature of the geometry optimization procedure is denoted with use of a label k, so that for a system with N particles let $x_k \in \mathbb{R}^{3N}$ denote the configuration at the kth optimization step, and the corresponding forces on the system at step k is $f_k = f(x_k)$. The necessary condition for a point on the smooth PES to be a local minimum is the requirement that forces vanish:

$$f(x_k) = 0. \tag{3.4}$$

and that the Hessian matrix $H_k = \partial^2 E / \partial x_k^2$ at this point x_k is positive semidefinite. The standard optimization schemes iteratively search for structures that minimize the energy of a system until Eq. 3.4 is satisfied with desired accuracy usually this threshold is less than 10^{-2}eV Å^{-1}. The simplest methods such as steepest descent and conjugate gradient simply follow calculated gradients and move the atoms in the direction of calculated forces. These are guaranteed to converge, but are among of the most inefficient optimization techniques since they tend to primarily follow the degrees of freedom for which the small displacements lead to large energy changes which results in very poor convergence near the local minimum. The most popular optimization technique is the quasi-Newton scheme that uses the information about the second derivatives of the PES to search for the optimization direction more efficiently [232, 233]. The basic idea is to approximate the PES by an harmonic model with respect to x_k:

$$M_n(x_k + s_k) := x_k - f_k^T s_k + \frac{1}{2} s_k^T H_k s_k, \tag{3.5}$$

where s_k is displacement. The calculation of the exact Hessian requires substantial computational effort, but for the optimization techniques described below this is not necessary and instead an approximation to the Hessian is used, that is updated during the geometry optimization process. The most widely used scheme for updating the Hessian matrix is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) formula [232, 233],

$$H_{k+1} = H_k - \frac{H_k \Delta x_k \Delta x_k^T H_k}{\Delta x_k^T H_k \Delta x_k} - \frac{\Delta f_k \Delta f_k^T}{\Delta f_k^T \Delta x_k}, \tag{3.6}$$

where $\Delta x_k = x_{k+1} - x_k$ and $\Delta f_k = f_{k+1} - f_k$. In this method the initial guess H_0 is important and can dramatically improve the efficiency of finding of the local minima and, in some cases, even lead to different results when different initial guesses are used [234]. A naive choice of the initial guess is to take the scaled identity matrix $H_0 = \beta \cdot I$ where $\beta > 0$. This scheme is very efficient if the PES is truly harmonic and the Hessian is explicitly known. The

first assumption is valid when the structure is already near a local minimum. Usually when dealing with flexible molecules it is impossible to generate the initial guess geometries to be near local minima. Regarding the second term, the first guess for the Hessian matrix must be chosen carefully since it can influence even the qualitative outcome of the optimization [234]. Different preconditioning schemes perform quite differently for different materials systems [235–237] and these will be discussed further in Sec. 5.6.

In recent years, applying a ML model in geometry optimization became a significant field of research, but it is still in the very early stages of adoption. For example, a neural networks (NN) was used to accelerate the saddle-point search by construction of an approximate energy surface [238], and Gaussian Process Regression (GPR) can help to predict derivatives and smoothness of energy function together with their uncertainties during the geometry optimization [239, 240]. The area of active-learning application in geometry optimization looks quite promising [241–245], however, the high flexibility adds a computational cost since a large number of training points (on the order of tens of thousands) are required to ensure that the NN PES has the proper form.

3.2.2 Line search method

Within an optimization algorithm, one has to define a search direction to displace the atoms. One of the approaches for prediction of the search direction for optimization step is the line search method (LSM). Starting from the quadratic model of the PES (Eq. 3.5), one needs to obtain a search direction p_n along which the optimization step s is obtained according to a step length α_n

$$s_n = \alpha_n p_n. \tag{3.7}$$

Then the new configuration is obtained with $x_{k+1} = x_k + s_k$. The search direction p_k for which the energy decreases is the descent direction: $f_k^T p_k > 0$. From the harmonic model the search direction, which is also called quasi Newton step for an approximate Hessian, that minimizes the energy is

$$p_k = H_k^{-1} f_k. \tag{3.8}$$

After finding the search direction one has to determine the step length α_k. Estimation of the step length is done by imposing Wolfe conditions on it [246, 247]:

$$E\left(x_k + \alpha_k p_k\right) \leq E\left(x_k\right) - c_1 \alpha_k f_k^T p_k, \quad c_1 \in (0,1), \tag{3.9}$$

$$f\left(x_k + \alpha_k p_k\right)^T p_k \leq c_2 f_k^T p_k, \quad c_2 \in (c_1,1). \tag{3.10}$$

The last one is also called the Armijo condition [248] and assures a sufficient decrease in the objective function along the search direction. The line search method with a BFGS update for the approximate Hessian is summarized in Algorithm 1:

Algorithm 1 BFGS line search

Require: $x_0, H_0, \epsilon > 0$
  $k \leftarrow 0$
  while $\left\|f_k\right\|_\infty > \epsilon$ **do**
    Get p_n from Equation 3.8.
    Get α_n ensuring Wolfe conditions (Eqs. 3.9, 3.10) are satisfied.
    $x_{k+1} = x_k + s_k = x_k + \alpha_k p_k$
    Update approximate Hessian $\Leftarrow$ using Eq. 3.6
    $k \leftarrow k+1$
    end

3.2.3 Trust-region method

Another approach that is widely used is the trust-region method (TRM) that assumes that the harmonic model of the PES is correct within trust-region radius Δ_k near x_k. The trial step is then obtained by minimizing the quadratic model function:

$$s_k = \underset{s_k^* \in T_k}{\arg\min} M_k\left(x_k + s_k^*\right), \tag{3.11}$$

where $T_k := \left\{s_k^* : \left\|s_k^*\right\|_2 \leq \Delta_k\right\}$. Then the quality of the harmonic model is calculated as the ratio between the actual reduction of the total energy E when the trial step s_k is taken and the reduction that is predicted by the model function M_k:

$$\rho_k = \frac{E(x_k) - E(x_k + s_k)}{M_k(x_k) - M_k(x_k + s_k)} \tag{3.12}$$

If ρ_k is negative, the energy increases with the taken step. For negative and small values of ρ_k, the step is rejected, and the trust-radius is reduced. If ρ_k is close to one, the PES around x_k is in agreement with the harmonic model, and thus the trust-radius can be increased, and the step is accepted. The criteria for adjustment of the trust-radius can be summarized in the following:

$$\Delta_{k+1} = \begin{cases} \frac{1}{4}\Delta_k & \text{if } \rho_k < \frac{1}{4} \\ \min\{2\Delta_k, \Delta_{\max}\} & \text{if } \rho_k > \frac{3}{4} \wedge \|s_n\|_2 = \Delta_k \\ \Delta_k & \text{else ,} \end{cases} \tag{3.13}$$

where $\Delta_{\max}$ is the maximum allowed displacement length that is defined for the geometry optimization of the system. Iteration continues until the trust-radius is adjusted so that the step is accepted. The iteration then continues until the convergence condition for the forces $\left\|f_k\right\|_\infty < \epsilon$ is met. The TRM method is summarized in the Algorithm 2:

Algorithm 2 BFGS TRM

Require: $x_0, \Delta_0 \in (0, \Delta_{\max}), \epsilon > 0$
$k \leftarrow 0$
while $\left\| f_k \right\|_\infty > \epsilon$ **do**
    Get s_k from Equation 3.11.
    Get ρ_k from Equation 3.12.
    Update trust-region radius $\Leftarrow$ using Eqs. 3.13
    if $\rho_k > \frac{1}{4}$ **then**
        $x_{k+1} = x_k + s_k$
        Update approximate Hessian using Eq. 3.6
    else
        $x_{k+1} = x_k$
    end
    $k \leftarrow k + 1$
end

The minimization in Eq. 3.11 can be solved approximately; for further details, we refer the reader to textbook [85].

In conclusion, the LSM and the TRM optimization techniques are both reasonably simple and robust. It is possible to classify both techniques as modified quasi-Newton approaches since they are based on a quadratic model PES and do not require knowledge of the actual Hessian. They are looking for stationary places at which the force disappears, and as a result, they rely on the assumption of a smooth PES. Even though this assumption appears fair for physical systems, it may not necessarily hold in all cases, especially if the system is far away from the local minimum and the electronic structure changes dramatically with respect to structural changes. However, it should be noted that both techniques are only capable of locating local energy minima; the global energy minimum, on the other hand, requires, in addition, sampling of the PES.

For the LSM, the step length is often determined iteratively until the Wolfe criteria are met. For *ab initio* approaches, this can result in an unacceptably large number of energy and force evaluations and may be unstable due to the numerical inaccuracies of the forces. Because the TRM does not require any extra energy calculations to calculate the trust radius, it is more suitable for *ab initio* structure optimization than the LSM. Thus TRM is the method used in this thesis and default search strategy implemented in the global structure search package discussed in Section 5.

3.2.4 Preconditioning schemes for geometry optimizations

The challenge for quasi-Newton optimization methods is that the Hessian is unknown and has to be approximated for the guess geometry since the calculation of the exact Hessian requires enormous computational effort - it requires $6N$ force evaluations, where N is the number of atoms in the system. Another way to calculate the Hessian matrix is to employ density functional perturbation theory, which is also computationally inefficient

[234]. Different ways to construct the initial guess of the Hessian matrix are proposed in the literature. This is referred to as preconditioning, and it may be thought of as a coordinate transformation to a new coordinate system with a better-conditioned optimization problem; as a result, algorithms converge faster and are more robust. Different preconditioning schemes perform with different efficiency for different systems: for example, for covalently bonded periodic systems, the Exponential (Laplacian) preconditioning scheme was found to be simple and effective [237]:

$$H^{Exp}_{(3A+i),(3B+j)} = \begin{cases} -\mu \exp\left(-\alpha\left(\frac{R^{AB}}{R_{nn}} - 1\right)\right), & R^{AB} < R_{\text{cut}} \text{ and } i = j \\ 0, & R^{AB} \geq R_{\text{cut}} \text{ or } i \neq j \end{cases} \tag{3.14}$$

where i, j are Cartesian coordinates and R_{nn} is the maximal nearest-neighbour distance:

$$R_{nn} = \max_{A}(\min_{B} R^{AB}) \tag{3.15}$$

α is chosen arbitrarily to provide damping of atomic interactions, R_{cut} can be reasonably taken as $2R_{nn}$, and the scaling parameter μ can be automatically identified from test displacements of the atoms [249]. By setting $\alpha = 0$ and $\mu = 1$ the Hessian reduces to the Laplacian matrix, a generalization of which is used to represent undirected graphs.

For systems such as molecules in a gas phase or molecular crystals, the Exponential preconditioner scheme does not perform as well as for bulk systems due to the wide range of different interactions. For molecular systems, the use of internal coordinates [250, 251] and FF like preconditioner techniques are much more efficient. For example, the FF model Hessian in Lindh preconditioning scheme is described in the original paper [236] and introduces the analytic form of the energy function that consists of quadratic terms for all distances, angles, and dihedrals in the molecule. The positive-definite requirement for such a preconditioning scheme is fulfilled by assuming that the current geometry is its local minimum. This approach will also be used in the derivation of the Section 5.6, where we derive the preconditioning LJ scheme. Using a simple 15-parameter function of the nuclear positions, the model Hessian can be constructed for any molecule with atoms from the first three rows of the periodic table. This approach yields great performance and is implemented in many electronic structure packages, including FHI-aims[252]. Other FF based initial Hessian matrices take into account the many-body terms such as bond stretch, angles and dihedrals that are specifically parametrized for a system under investigation and also be used in combination with other preconditioning schemes tailored to systems like molecular crystals [235].

3.3 Comparing molecules across structural space

The large quantities of high dimensional data obtained from structure searches and molecular dynamics simulations require automated tools to produce representations, analyses and classifications. The strategy for representing the high dimensional spaces in a human-readable low-dimensional format usually consists of several steps: a) choosing a represen-

tation for the molecules; b) calculating the dissimilarity covariance matrix between these representations; c) performing a dimensionality reduction.

SOAP [111] is an elegant representation that is invariant to rotations, translations, and permutations of equivalent atoms. The main idea of SOAP is to expand the molecular structure into a set of local atomic environments $\mathscr{X}$ and then use their combinations to measure a global similarity between structures. The local environment density around the central atom is approximated as a sum of Gaussian functions with variance σ^2 centred at atom positions $\mathbf{x}_i$ within the environment $\mathscr{X}$:

$$\rho_{\mathscr{X}}(\mathbf{r}) = \sum_{i\in\mathscr{X}} \exp\left(-\frac{(\mathbf{x}_i-\mathbf{r})^2}{2\sigma^2}\right) \tag{3.16}$$

The similarity kernel between two local environments $\mathscr{X}$ and $\mathscr{X}'$ is defined as

$$\tilde{k}\left(\mathscr{X},\mathscr{X}'\right) = \int \mathrm{d}\hat{R} \left| \int \rho_{\mathscr{X}}(\mathbf{r})\rho_{\mathscr{X}'}(\hat{R}\mathbf{r})\mathrm{d}\mathbf{r}\right|^2, \tag{3.17}$$

which is the overlap of the two local atomic environment densities integrated over all three-dimensional rotations $\hat{R}$. The self-similarity of any kernel should be unity, so the final normalized kernel has a form

$$k\left(\mathscr{X},\mathscr{X}'\right) = \tilde{k}\left(\mathscr{X},\mathscr{X}'\right)/\sqrt{\tilde{k}(\mathscr{X},\mathscr{X})\tilde{k}(\mathscr{X}',\mathscr{X}')}. \tag{3.18}$$

The integration over all rotations can done analytically if the atomic neighbourhood densities are expanded in a basis composed of orthogonal radial basis functions g_n and (angular) spherical harmonics $Y_{l,m}$:

$$\rho_{\mathscr{X}}(\mathbf{r}) = \sum_{n,l,m} c_{n,l,m} g_n(|\mathbf{r}|)\mathrm{Y}_{lm}(\mathbf{r}), \tag{3.19}$$

where $c_{n,l,m}$ are expansion coefficients. From these coefficients, rotationally invariant quantities can be constructed, such as the *power spectrum* that is given by

$$p(\mathscr{X})_{n,n',\ell} = \sum_m c_{n,\ell,m}\, c^*_{n',\ell,m} \tag{3.20}$$

The elements of the power spectrum are then collected into a unit-length vector $\hat{\mathbf{p}}(\mathscr{X})$, so that the SOAP kernel is given as [111]

$$k\left(\mathscr{X},\mathscr{X}'\right) = \hat{\mathbf{p}}(\mathscr{X})\cdot\hat{\mathbf{p}}\left(\mathscr{X}'\right). \tag{3.21}$$

The numerical hyper parameters that have to be tuned are the maximal number of radial and angular basis functions, the broadening width, and the cut-off radius. For the details of the derivation of the SOAP kernels for multi-species environments we refer the reader to the

detailed explanation in Ref. [115].

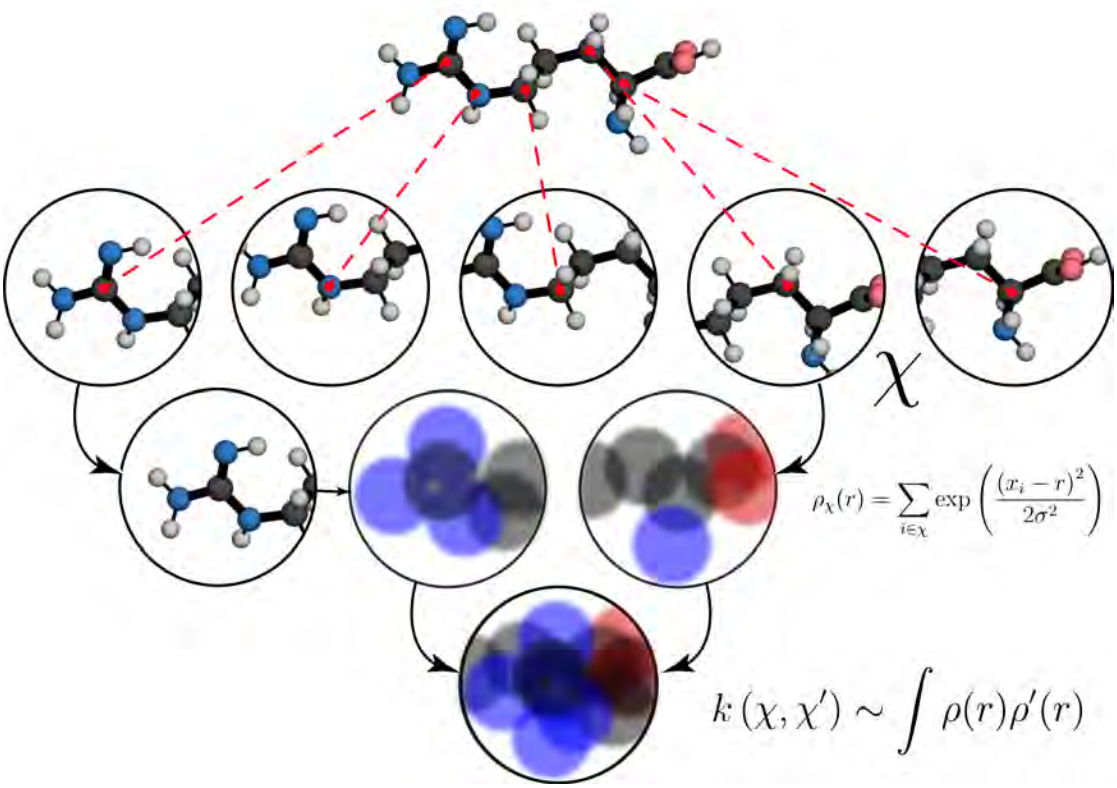


Figure 3.2 – Atom-density-based structural representations, in which the structure is mapped onto a smooth atom density constructed as a superposition of smooth atom-centered functions that also reflect the chemical composition information.

After the mathematical formulation to compare two local environments is established, the next step is to introduce the global kernel to compare two structures. For two structures with the same number of atoms N, one can compute an environment covariance matrix that contains all the possible pairings of environments

$$C_{ij}(A,B) = k\left(\mathscr{X}_i^A, \mathscr{X}_j^B\right), \tag{3.22}$$

where indices i, j run through all of the atoms contained in structures A and B. The simplest way to introduce a global metric is to use the average kernel

$$\bar{K}(A,B) = \frac{1}{N^2}\sum_{ij} C_{ij}(A,B) = \left[\frac{1}{N}\sum_i \mathbf{p}\left(\mathscr{X}_i^A\right)\right] \cdot \left[\frac{1}{N}\sum_j \mathbf{p}\left(\mathscr{X}_j^B\right)\right] \tag{3.23}$$

The main drawback of this approach is that two very different structures can appear to be very similar if their environments give the same fingerprints upon averaging.

Another possibility is to find the best matching between the environments of the two struc-

tures

$$\hat{K}(A,B) = \max_{\mathbf{P}\in\mathscr{U}(N,N)} \sum_{ij} C_{ij}(A,B) P_{ij}, \tag{3.24}$$

by finding the permutation matrix P_{ij} that maximizes the value of $\hat{K}(A,B)$. Here $\mathscr{U}(N,N)$ is the set of $N \times N$ scaled doubly stochastic matrices whose rows and columns sum to $1/N$, i.e. $\sum_i P_{ij} = \sum_j P_{ij} = 1/N$. This is a very computationally expensive procedure that can be computed in polynomial time using the Hungarian Method [253]. This method has discontinuous derivatives whenever the matching of environments change. This problem can be solved by introducing the regularized entropy match kernel (REMatch) that combines the features of *average* and the *best-match kernel* and smoothly interpolates between them. It relies on ideas from optimal transport theory [254] that regularize this problem by adding a penalty that aims to maximize the information entropy for the matrix P_{ij}:

$$\hat{K}^{\gamma}(A,B) = \mathrm{Tr}\,\mathbf{P}^{\gamma}\mathbf{C}(A,B) \tag{3.25}$$

$$\mathbf{P}^{\gamma} = \underset{\mathbf{P}\in\mathscr{U}(N,N)}{\mathrm{argmin}} \sum_{ij} P_{ij}\left(1 - C_{ij}(A,B) + \gamma \ln P_{ij}\right), \tag{3.26}$$

where the entropy term $E(\mathbf{P}) = -\sum_{ij} P_{ij} \ln P_{ij}$ introduces the regularization. This allows the computation P_{ij} with $\mathscr{O}\left(N^2\right)$ effort using the Sinkhorn algorithm [254]. For small values of γ this penalty becomes negligible and we obtain the best-match kernel. For the large values of γ the permutation matrix with the least informational content must be selected $P_{ij} = 1/N^2$, which reduces Eq. 3.25 to the average kernel limit. The definition of the distance would be

$$D(A,B) = \sqrt{2 - 2K(A,B)}, \tag{3.27}$$

where $K(A,B)$ is the global similarity kernel.

After introducing the kernel-induced metric, one can calculate the dissimilarity matrix of a set of structures and employ one of the dimensionality reduction schemes to obtain two dimensional map that represents proximity relations between structures. The simplest method among all the schemes is principal component analysis (PCA) which constructs a linear combination of variables extracting the maximum variance from the input features. PCA and its variances are widely applied in material science for analysing different systems [255–261]. The interested reader can find more details on the dimensionality reduction techniques such as ISOMAP [262, 263], t-SNE [264] applied to analyse biomolecular systems in nice reviews [265–267].

For the dimensionality-reduced representation, we here chose to use the metric multi-dimensional scaling (MDS) algorithm as implemented in the `scikit-learn` package[268]. This algorithm is similar to the Sketch-map algorithm previously employed in Ref. [110], but

we found it to be more suitable for the data at hand, which is composed of decorrelated local stationary-points, instead of structures generated from molecular dynamics trajectories. The low-dimensional map is obtained through an iterative minimization of the stress function:

$$\delta = \sum_{A \neq B} (D(A,B) - d(A,B))^2 , \tag{3.28}$$

where $D(A,B)$ is the distance between structure A and B in high-dimensional space and $d(A,B)$ is the Euclidean distance in the low-dimensional space. The result of the procedure will be set of two dimensional coordinates y_N reflecting the mutual distances between structures. For tracking the changes of the conformational spaces one can use one of the two dimensional points as reference and project other structures with use of out-of-sample embedding technique. Finding the low-dimensional coordinates x for structure with high-dimensional representation $\mathscr{X}$ is done through minimization of the stress function δ_P considering the known low-dimensional coordinates for N structures y_N and their high-dimensional representations $\mathscr{X}_N$

$$\delta_P = \sum_{n=1}^{N} (D(\mathscr{X}, \mathscr{X}_N) - d(x, x_N))^2 , \tag{3.29}$$

where the sum runs over all structures in the reference dataset.

[55mm] I am a dwarf and I'm digging a hole
Diggy, diggy hole! Diggy, diggy hole! a song of Simon Lane (Honeydew)

4 The conformational space of a flexible amino acid at metallic surfaces

This chapter is dedicated to the description of single molecule adsorption on metallic surfaces. Amino acids are the building blocks of proteins when connected in a sequence via peptide bonds $(\text{N-C}_\alpha\text{-C(O)})_n$, and can be great test systems for methodological developments since they are small enough to be computationally feasible for modern accurate theoretical methods and flexible enough to provide a challenge for their structure search.

In this chapter the adsorption preferences of the most flexible amino acid Arg and its charged counterpart Arg-H^+ were investigated using an exhaustive conformational search. This case is further complicated by the fact that after adsorption the neutral Arg and positively charged Arg-H^+ undergo complex charge rearrangement (see Fig. 4.1). The adsorption was modeled on three noble metal surfaces Cu(111), Ag(111) and Au(111), to study the adsorption behaviour depending of the reactivity of the model surfaces. A depiction of the Arg molecule including the labeling of the different chemical groups and specific atoms we will refer to in the thesis is shown in Fig. 4.2(a). In this context we use the term *protonation state* to distinguish between Arg and its singly-protonated form Arg-H^+. We use the word *protomers* to distinguish between different arrangements of protons within molecules of the same sum formula, for example the protomers **P1** to **P5** of Arg or the protomers **P6** and **P7** of Arg-H^+, shown in Fig. 4.2(b) and (c).

Another important aspect to address is the chemical composition of Arg after adsorption. In general, amino acids tend to adsorb in their zwitterionic form, when the molecule has termination groups COO^- and NH_3^+ [61]. However, deprotonation is also possible, with the anionic (COO^- and NH_2) and an extra hydrogen atom being adsorbed on the surface [269].

In order to establish the conformational preferences of adsorbed Arg and Arg-H^+, the relative energies of these conformers must be calculated. This can be done using DFT, which can also describe any charge rearrangements that occur following adsorption. In addition, DFT provides insights on the modification of molecular energy levels when forming an interface [73, 211, 270] that are crucial to understanding transport phenomena in molecular

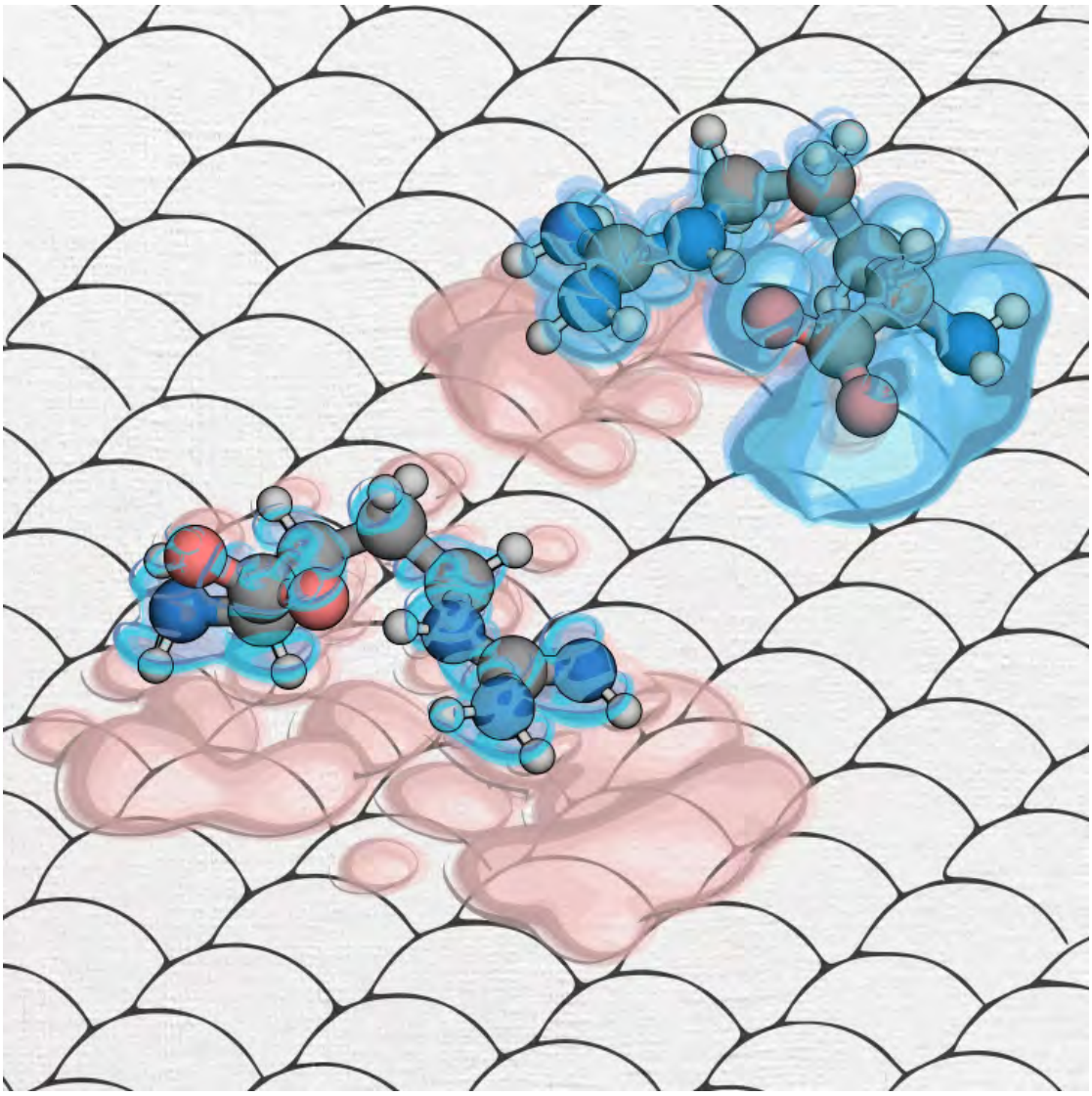


Figure 4.1 – The picture shows a sketch of the electronic density rearrangement that happens when arginine and protonated arginine adsorb on Cu(111) surface. The electron accumulation is depicted in red and electron depletion depicted in blue.

electronic devices. Information on the particular preference of adsorption sites and binding energy strengths that depend on the interacting groups are important in understanding self-assembly patterns that are formed on surfaces [271, 272].

The starting point for this investigation was the creation of a database with thousands of stationary states of different conformers on metal surfaces. The procedure of this database generation with a description of the computational setup and convergence tests is described in the next section. A shortened version of this chapter was published in International Journal of Quantum Chemistry [83].

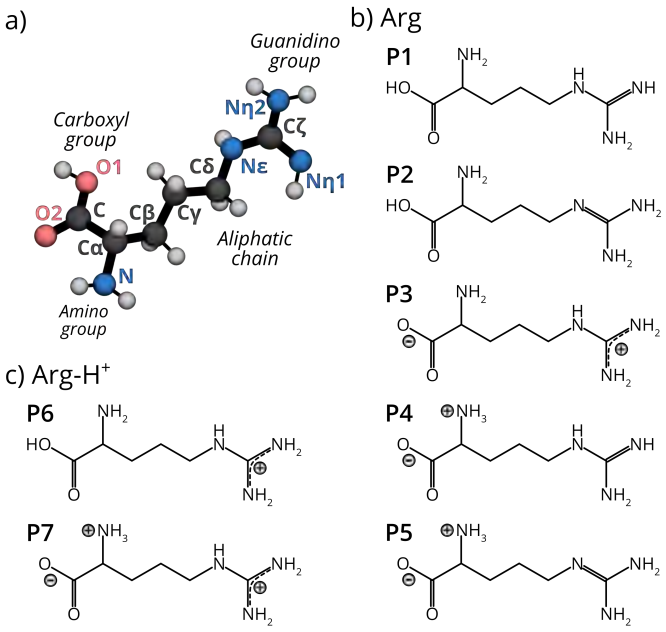


Figure 4.2 – a) Pictorial representation of the arginine amino acid, including labels of chemical groups and atoms. b) Protomers of Arg that are addressed in this work. c) Protomers of Arg-H^+ that are addressed in this work.

4.0.1 Computational setup

For modeling the adsorbed molecules, we first had to create model slabs on which to perform an exhaustive structure search. The bulk lattice constants for Cu, Ag and Au were determined by optimizing the fcc unit cell with convergence criteria set to 0.001 eV/Å for the final forces, 10^{-4} e/Bohr^3 for the charge density, and 10^{-5} eV for the total energy of the system, and a 30×30×30 k-grid mesh was used for the sampling of the Brillouin zone. The lattice constants, obtained with the PBE functional[273] are shown in Table 4.1. We also compare the PBE lattice constants with those obtained including pairwise vdW dispersion from the original Tkatchenko-Scheffler scheme (+vdW)[163] and with the one that includes an effective electronic screening optimized for metallic surfaces (+vdW^{surf})[130].

Table 4.1 – Lattice constants (in Å) of bulk metals determined with the PBE, PBE+vdW and PBE+vdW^{surf} functionals (*light* settings).

Method	Cu	Ag	Au
PBE	3.633	4.156	4.157
PBE+vdW	3.545	4.077	4.114
PBE+vdW^{surf}	3.604	4.022	4.173
Exp [274]	3.598	4.079	4.064

Since the PBE lattice constants for Cu, Ag, and Au are already in good agreement with experimental data [274] (Table 4.1) and with previous works [150, 275], and given the absence of a systematic improvement by the inclusion of these types of vdW interactions [130] in

metals, we chose to use the simplest setup and proceed with PBE lattice constants for generating the metal slabs.

For simulations of Arg adsorbed onto surfaces, a 5×6 surface unit cell with $4 \times 4 \times 1$ k-point sampling was employed. The slab contains 4 layers, and we added a 50 Å vacuum in the z direction in order to separate periodic images of the system. Convergence plots in Fig. 4.3 show that this is sufficient to obtain the correct energy hierarchy for different conformers. However, a surface unit cell of this size does not completely isolate neighboring molecules on the surface plane. In order to estimate the magnitude of this spurious interaction, we

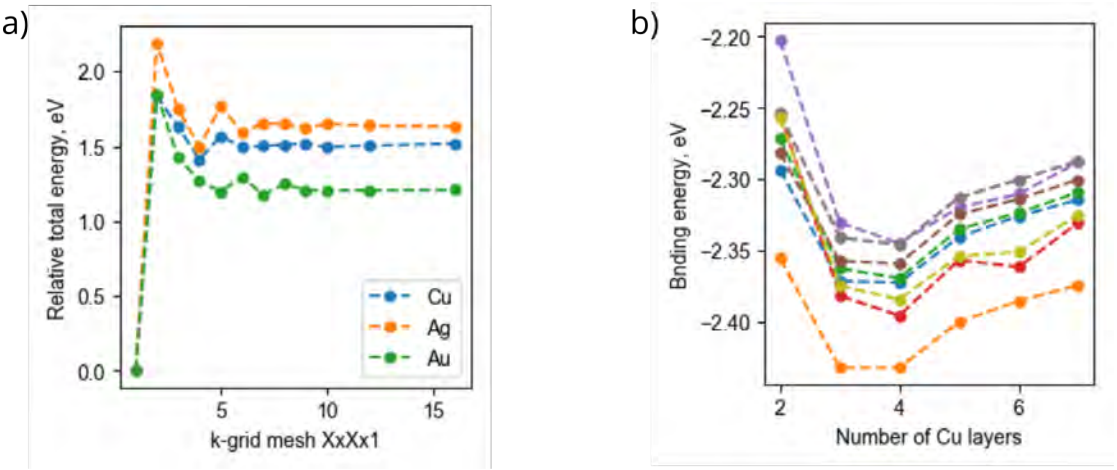


Figure 4.3 – a) Relative total energy convergence of with respect to k-grid mesh for different 5×6 slabs. b) Binding energy hierarchy calculated for different structures on Cu(111) surface with different amount of layers.

calculated binding energies for three Arg and three Arg-H^+ structures adsorbed on Cu(111) using different surface unit cell sizes. These structures are shown in Fig. 4.4. As shown in Table 4.2, the relative binding energies change by no more than 50 meV when reaching a 10×12 cell. Furthermore, the energetic hierarchy of the structures does not change with increasing the unit cell size and to save computational resources we proceed with a 5×6 unit cell size.

All the electronic structure calculations were carried out using the numeric atom-centered basis set of the all-electron code FHI-aims [183, 184]. We used the standard *light* settings of FHI-aims for all species with use of PBE+vdW^{surf} functional, except when stated otherwise. Relativistic effects were considered by the zeroth order regular approximation (ZORA) [276, 277]. To prevent an artificial relaxation of the metal surfaces, we did not use vdW interactions between metal atoms since we created slabs with PBE lattice constants. We also fixed the two bottom layers of the slabs in all optimizations. A dipole correction was applied in the z direction to compensate for the dipole formed by the asymmetric surface configurations. With this setup, we placed different conformations of Arg and Arg-H^+ in different orientations with respect to the slab and performed a geometry optimization with the BFGS algorithm

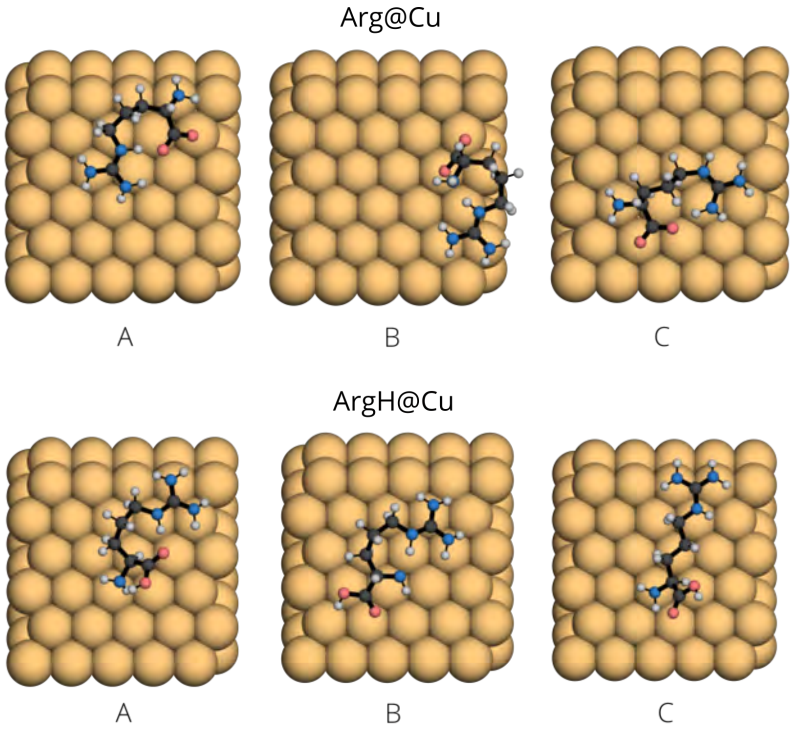


Figure 4.4 – Structures that were used for the surface unit cell size convergence test of Arg@Cu (first row) and ArgH@Cu (second row). Image unit cell size is 5×6.

using the trust region method, until all forces in the system were below 0.01 eV/Å . Database generation is described in the next section.

For reference, we report the values we used for E_f at each surface in Table 4.3.

4.0.2 Database Generation

The sampling of the structure space of Arginine in two protonation states on metallic surfaces was performed by starting from a previously published dataset comprising the stationary points of isolated amino acids and dipeptides [2, 278]. For Arg, 1206 structures are present in the database. In order to reduce the number of possibilities, but keeping a representative share of the structures, we considered the 300 lowest energy conformers, the 27 highest energy conformers, and 125 conformers uniformly spanning the energy range in between. For the Arg-H^+ amino acid, all 215 structures present in the gas-phase data set were used in this study.

We distinguish *upstanding* positions of the molecules where the largest eigenvector of the rigid-body moment of the inertia tensor is approximately perpendicular to the surface plane, from *flat lying* positions with an arrangement parallel to the surface. For Arg, 3 flat lying configurations per structure were generated by randomly placing the molecule flat on the Cu(111) surface and then rotating it by 120° around the principal axis. Two upstanding configurations were generated for the 25 of gas-phase structures by first placing the molecule

Table 4.2 – Relative binding energies (in eV) of relaxed Arg@Cu and ArgH@Cu for different surface unit cell sizes with a 8×8×1 k-grid for the cell sizes less than 10×12 and 4×4×1 for the 10×12 unit cell. All numbers are reported with respect to the binding energy for the structure A modelled with a 5×6 surface unit cell.

	Arg@Cu			ArgH@Cu		
slab size	A	B	C	A	B	C
5×6	0.000	0.011	0.216	0.000	0.080	0.035
6×6	-0.011	-0.013	0.190	-0.050	0.041	-0.017
6×7	-0.021	-0.030	0.174	-0.055	0.029	-0.033
10×12	-0.048	-0.053	0.151	-0.044	-0.007	-0.057

Table 4.3 – Fermi energies calculated with the PBE functional for the 4-layer slabs with (111) surface orientation used in our calculations of the binding energies of charged molecules to the different surfaces. All values in eV.

	Cu	Ag	Au
Slab E_f	-4.73	-4.30	-5.02

in a random upright orientation, and then flipping it. For Arg-H^+ a similar procedure was adopted: flat lying positions were created by 90° rotations around the principal axis and upstanding configurations were created for 27 structures. In summary, we considered a total of 1156 conformers of Arg@Cu(111) and 914 conformers of Arg-H^+@Cu(111).

Every optimized structure that fell within a range of 0.5 eV from the global minimum on Cu(111) were transferred to Ag(111) and Au(111) and further optimized. In addition, we randomly picked 105 Arg-H^+ structures representing the higher energy range on Cu(111) to be further optimized on Ag(111) and Au(111). Moreover, for Arg 180 randomly picked structures representing the higher energy range were considered on Ag(111) and 61 on Au(111). The total amount of calculated structures for each case is summarized in Table 4.4.

We checked that this strategy ensured a sufficient sampling of the low-energy range of both Arg and Arg-H^+ on Ag(111) and Au(111) by analyzing the alterations in relative energy hierarchies on the different surfaces. In Fig. 4.5, each dot corresponds to a conformer that was optimized first on the Cu(111) surface and then post-relaxed on Ag(111) or Au(111). Within the lowest 0.5 eV range, we do not observe any significant rearrangement of the

Table 4.4 – Number of calculated Arg and Arg-H structures in isolation and adsorbed on Cu(111), Ag(111) and Au(111).

	Gas phase	Cu(111)	Ag(111)	Au(111)
Arg	1206	1156	327	209
Arg-H^+	215	914	718	721

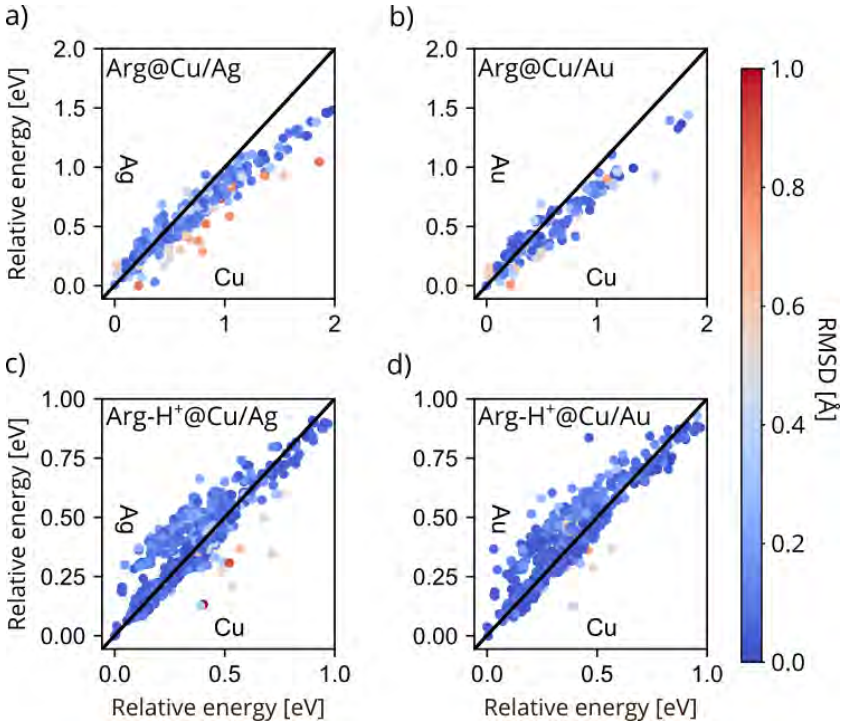


Figure 4.5 – (a-d) Correlation plots of relative energies of Arg or Arg-H^+ conformers on Cu, Ag, and Au (111) surfaces. Each dot corresponds to the same conformer optimized on the two surfaces addressed in each panel, color coded with respect to the RMSD (heavy atoms only) between the superimposed optimized structures without taking surface atoms into consideration.

energy hierarchy with respect to the Cu(111) surface. The energy hierarchies of both Arg and Arg-H^+ on the Ag(111) and Au(111) surfaces are almost identical. The most pronounced outliers in all plots correlate with a higher root mean square displacement (RMSD) of the molecular atoms (i.e. disregarding the surface-adsorption site), thus pointing to a structural rearrangement of the molecule.

4.0.3 Structure space representation

As was mentioned in the introduction, the simplest and one of the oldest representations developed for analysis of peptide structures was the Ramachandran plot, which can be seen in Fig. 4.6. As one can see the dihedral angles of the Arg and Arg-H^+ conformers are distributed in 8 clusters, but this information is not enough to draw conclusions about structure-property relationships, since Arg has 4 rotatable dihedral angles. Therefore, we proceed to analyse the database of isolated molecules and introduce further notation for later color coding of the results.

We analyse the structure space of all systems considered by employing a dimensionality reduction procedure that makes it more intuitive to understand the high-dimensional space. Following Ref. [110], we represent the local atom-centered environments of the structures

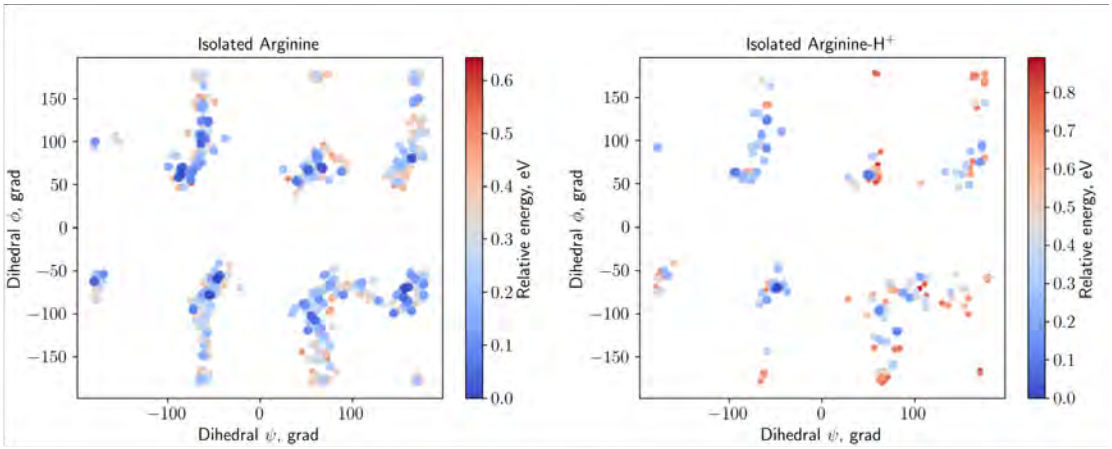


Figure 4.6 – Ramachandran plots for Arg (left) and Arg-H^+ (right) in isolation.

through SOAP[109] descriptors. We then obtain the similarity matrix between different conformers with the REMatch algorithm [115]. We used SOAP descriptors with a cutoff of 5.0 Å, a Gaussian broadening of $\sigma = 0.5$ Å and an intermediate regularization parameter γ=0.01 defined in Sec. 3.3. SOAP kernels were calculated only considering the heavy atoms in the molecule (disregarding metal and hydrogen atoms) and were obtained using the GLOSIM package [115, 279].

For the dimensionality-reduced representation, we here chose to use the metric multi-dimensional scaling (MDS) algorithm as implemented in the `scikit-learn` package[268]. This algorithm is similar to the Sketch-map algorithm previously employed in Ref. [110], but we found it more suitable for the data at hand, which is composed of decorrelated local stationary-points, instead of structures generated from molecular dynamics trajectories. In short, the low-dimensional map was obtained considering all calculated structures of Arg in the gas-phase and through an iterative minimization of the stress function, according to the procedure described in Section 3.3. We then projected structures in different environments onto the pre-computed map of gas-phase Arg by fixing the parameters of the map and finding the low-dimensional coordinates of the adsorbed molecules. The coordinates obtained as a result of the iterative metric MDS are not explicitly shown as axes on the plots since they are correlated to the descriptors used for the structural representation, which does not allow for a direct physical interpretation. These scatter plots just offer a visualization of the similarity matrix in lower dimensions. In order to classify structural patterns, we employ the following notations: We represent the protomers by the labels shown in Fig. 4.2(b) and (c). We identify the presence of strong intramolecular hydrogen bonds (H-bonds) whenever the distances between the hydrogen connecting donor and acceptor are below 2.5 Å. We label the H-bond pattern between two atoms in the molecules according to the nomenclature shown in Fig. 4.7. We further classify the structures according to the longest distance between two heavy atoms in the molecule. After describing of the results obtained for isolated Arg and Arg-H^+ molecules we will proceed to the description of adsorbed structures on surfaces.

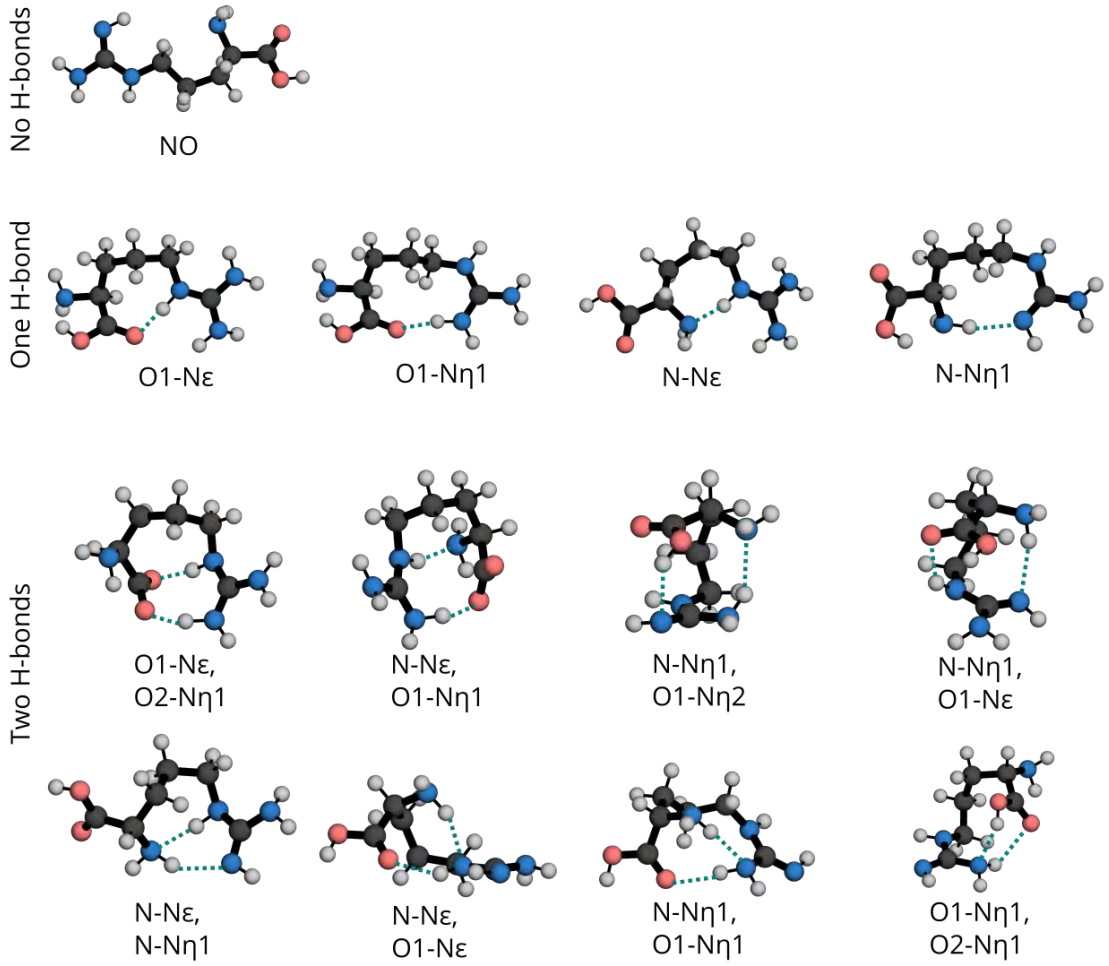


Figure 4.7 – Labeling of all H-bond patterns considered in this thesis.

The unconstrained structure space: Arg in isolation

We start by analysing the unconstrained conformational space of Arg in isolation, which is formed by more than 1200 local stationary states [2, 278]. In order to rationalize the different structural arrangements in this space, we utilize the dimensionality-reduction MDS algorithm and build a two-dimensional map. On this map, shown in Figure 4.8, each dot represents one structure. A close proximity between dots implies similarities between the heavy-atom arrangement between the conformations. This is the low-dimensional map that is taken as a reference for comparison throughout this manuscript.

We proceed to color-code the dots on the map according to different properties. In Fig. 4.8(a) we show the map colored by the relative energy ΔE_{rel} of each structure with respect to the global minimum. We only color structures with $\Delta E_{\text{rel}} < 0.5$ eV. The region with $\Delta E_{\text{rel}} < 0.1$ eV is colored red and is represented by 32 different structures that occupy different parts of the map. The dominant protomer among these conformers (29 out of 32, >90%) is the one labeled **P1** in Fig. 4.2, i.e. non-zwitterionic. However, the lowest energy structure, labeled *a*

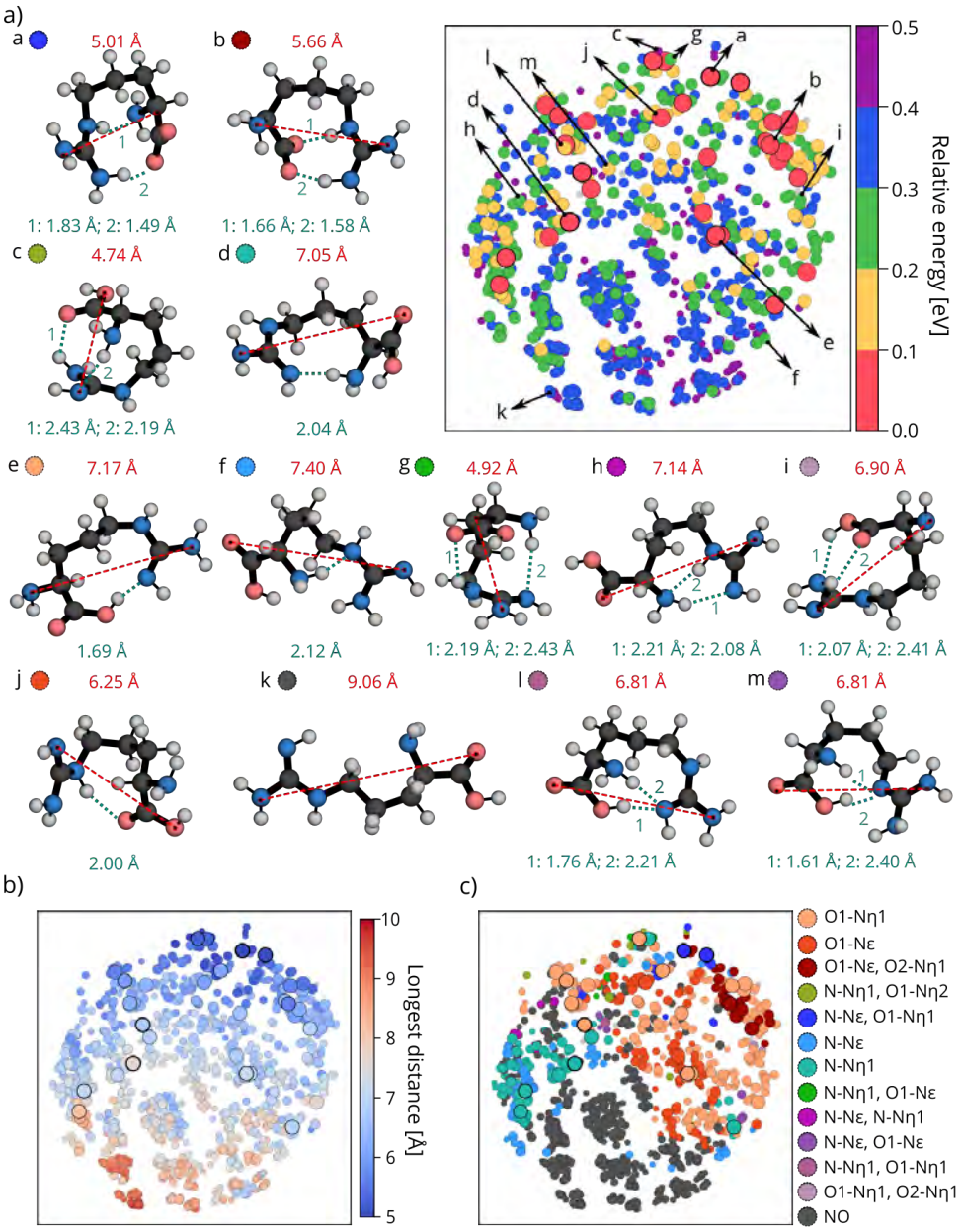


Figure 4.8 – Low-dimensional map of Arg stationary points on the PES. Only points linked to structures with a relative energy of 0.5 eV or lower are colored. Representative structures of all conformer families are visualized as well as their H-bond distances (in turquoise) and longest distance between two heavy atoms (in red) of the molecule. The maps are colored with respect to a) relative energy, b) longest distance, and c) H-bond pattern. The size of the dots also reflect their relative energy, with larger dots corresponding to lower energy structures.

in panel (a) of Fig. 4.8, is protomer **P3**, with a shared proton between the carboxylic and the guanidino group. This structure is compact, with the longest distance within the molecule of only 5.01 Å and presenting two strong intramolecular H-bonds. Zwitterionic protomers, denoted as **P4** and **P5** in Fig. 4.2, do not appear in the gas-phase.

Inspecting the map in Fig. 4.8(a), it is clear that low-energy conformers are almost exclusively present in the upper hemisphere of the plot. This can be rationalized in terms of the structural motifs that occupy these two halves of conformational space: In Fig. 4.8(b), we color-code the dots in terms of the longest extension of the conformers. While the upper hemisphere features compact structures, the lower hemisphere of the map is populated by extended conformers (with longest extensions between 7.5 Å and 10.0 Å). Many of them do not contain any H-bonds, or contain only one H-bond between the carboxyl and amino group. Extended conformers of Arg are energetically unfavoured in the gas-phase as the formation of strong H-bonds is crucial for the stabilization of Arg in isolation. Comparing the different plots in Figure 4.8, we see that the low-energy structures with $\Delta E_{\text{rel}} < 0.1$ eV are indeed compact with one or two H-bonds.

In Fig. 4.8(c), we identify in total 13 different configurational families with respect to the number and character of H-bonds in the molecule, with $\Delta E_{\text{rel}} < 0.5$ eV. Representative structures of all families are shown in panel (a). This family classification helps us understand why in Fig. 4.8(a) there are structures of higher energies in similar regions as structures with lower energies. Even though these structures are typically in the same protomeric state and have a similar arrangement of heavy atoms, the carboxyl group can rotate, giving rise to different H-bond patterns. These different patterns can give rise to energy differences of up to 0.2 eV, as exemplified in Fig. 4.9. Including hydrogens in the SOAP descriptors used to build the 2D map could provide a better energy separation, but would prevent us from comparing different protonation states, as shown in the next section.

Adding a proton: Arg-H$^+$ in isolation

Arg-H$^+$ is the most abundant form of Arginine under physiological pH conditions [280], and we thus investigate changes of the conformational space introduced by the addition of a proton to the Arg amino-acid. To that end, we plot a projection of all stationary points of the Arg-H$^+$ PES with $\Delta E_{\text{rel}} < 0.5$ eV (referenced to its own global minimum) onto the map that was previously created for Arg. In Fig. 4.10(a), we color the dots in the map according to ΔE_{rel}, in Fig. 4.10(b) according to the longest distance between heavy atoms in the molecule, and in Fig. 4.10(c) according to the H-bond pattern. The grey dots in the maps represent all points in the Arg map of Figure 4.8 and are shown for ease of comparison.

The unique conformation types of Arg-H$^+$ can be grouped into 8 different families in this energy range, which are represented in Fig. 4.10(a). Most families only have one H-bond and there are no zwitterionic protomers. This means that in isolation only the protomer **P6** is populated. It is worth noting that under physiological conditions (in solution), the zwitterionic protomer **P7** is preferred.

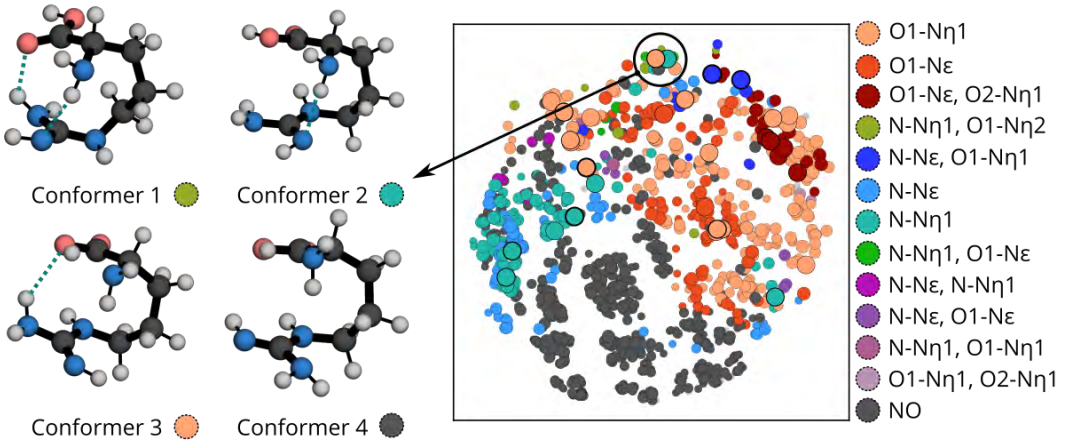


Figure 4.9 – Representative conformers with similar backbone structure but different H-bonds within the molecule. The different H-bond pattern can cause energy differences of up to 0.2 eV for similar structures, as discussed in the main text.

There are only two (very similar) structures with $\Delta E_{\text{rel}} < 0.1$ eV in this case. The global minimum, labeled *a* in Fig. 4.10(a), contains two H-bonds within the molecule, between atoms N-Nε and O1-Nη (see Fig. 4.2). This particular structure resembles the lowest-energy structure of Arg with a proton added to the carboxyl group. This protonation results in an extension of the molecule by around 1 Å. That correlates with the location of the lowest-energy structure being slightly shifted on the map towards the region containing more extended structures.

The structure space of Arg-H$^+$ is contained within the conformational space of Arg and also drastically reduced in number when if compared to Arg: There are only 108 structures with $\Delta E_{\text{rel}} < 0.5$ eV, compared to 1179 structures in the Arg case. In this energy range, regions of the map with very compact and very extended structures are not populated in this protonation state. This can be traced to the constraint imposed by the addition of the proton, that make extended structures less stable due to the strong driving force to neutralize the charge imbalance created by the proton on the guanidino group. To rationalize why the most compact conformers are also less populated, we show in Fig. 4.11 the electron-density differences between the lowest energy Arg-H$^+$ conformer and an Arg conformer created by fixing the same Arg-H$^+$ structure, but neutralizing the charge and removing the hydrogen connected to the carboxyl group. This modification yields the same covalent connectivity observed in the global minimum of Arg. We show isosurfaces corresponding to electron accumulation in Arg-H$^+$ in red and electron depletion in Arg-H$^+$ (accumulation in Arg) in blue. We observe a density surplus between the O1 and Nη atoms in Arg, favoring the formation of a stronger H-bond leading to a more compact structure.

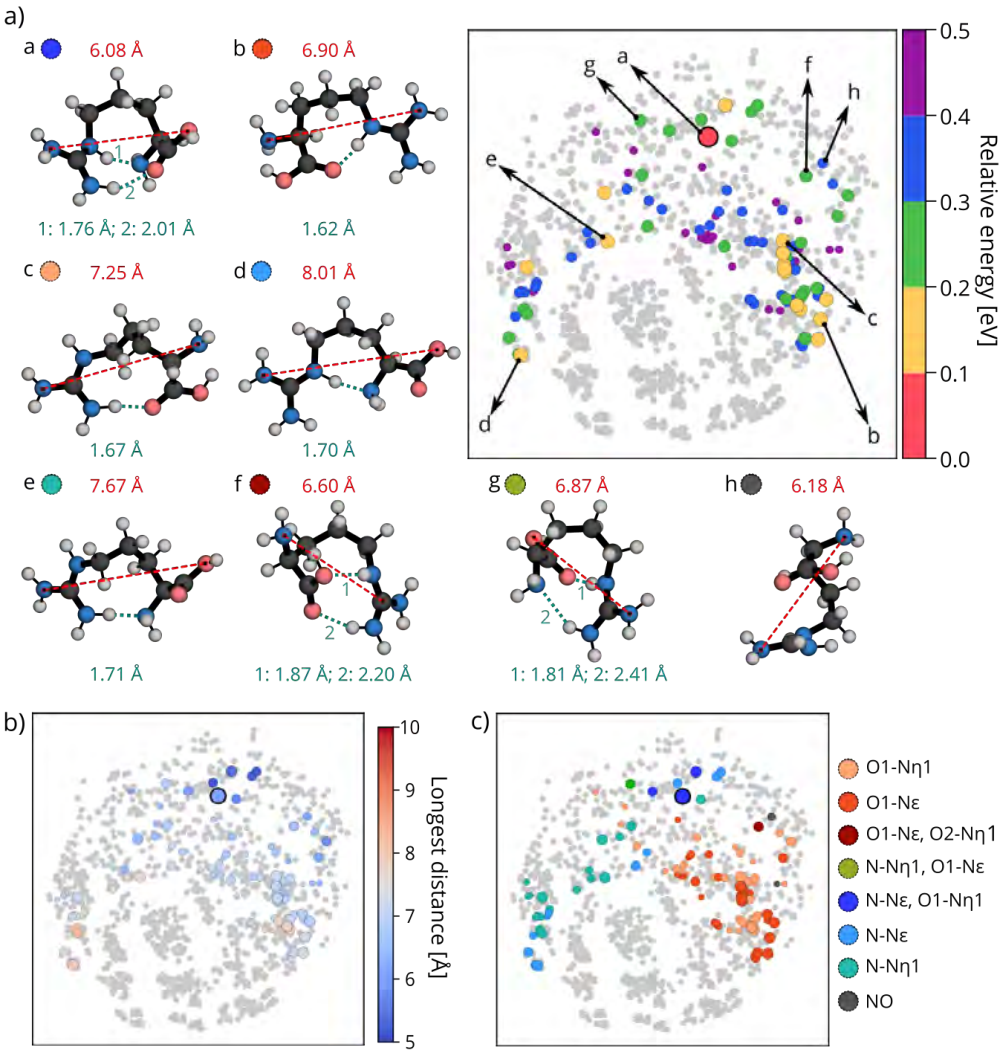


Figure 4.10 – Representative conformers of the populated structure families within 0.5 eV of the global minimum of isolated Arg-H^+ and low-dimensional projections of all populated conformers onto the Arg map. Grey dots represent all structures from the original map of isolated Arg in Fig. 4.10, and serve as a guide to the eye. The maps are colored with respect to a) relative energy, b) longest distance within the molecule, and c) H-bond pattern.

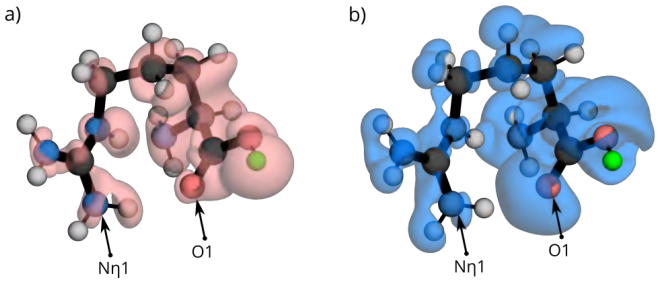


Figure 4.11 – Electron density difference between Arg-H$^+$ and Arg calculated by neutralizing the charge and removing the hydrogen connected to the carboxyl group (marked in green) from the lowest energy structure of Arg-H$^+$. The isosurfaces of electron density with value ±0.005 e/Bohr^3 corresponding to the a) regions of electron accumulation on Arg-H$^+$ and b) where the electron depletion on Arg-H$^+$, both compared to Arg.

Adsorption of Arg on Cu, Ag, Au (111) surfaces

We now turn to the analysis of the conformational space of Arg when in contact with metal surfaces, namely Cu(111), Ag(111), and Au(111). In Figure 4.12, we show map-projections of the stationary points with $\Delta E_{\text{rel}} < 0.5$ eV (referenced to the respective global minimum) of Arg adsorbed on the three surfaces. The conformational space of Arg upon adsorption is reduced and the adsorbed conformers occupy similar regions of the map as the conformers of Arg-H$^+$. We will learn in the following that this is mainly due to the formation of strong bonds with the surface that result in steric constraints of the space, and also partially due to electron donation from the molecule to the metallic surfaces.

The lowest energy structure lies on the same part of the map on all surfaces, which is different from the area where the gas-phase global minimum of Arg was located. These conformers, labeled *a* in Figure 4.12(a), (b) and (c), form a strong H-bond between atoms O1 and Nϵ. The longest distance within the molecule lies between 7.20-7.35 Å in all cases. This structure binds strongly to all three surfaces through both its amino and carboxyl groups.

Other low-energy structures on all surfaces form strong bonds to the surfaces only through the carboxyl group, as exemplified by the structure labeled *b* in all panels of Fig. 4.12. These bonds are formed most favorably on *top* positions, i.e. vertically on top of a surface metal atom. In particular for Cu(111), the atomic spacing of the Cu atoms on the surface favors both oxygens to bind on *top* positions simultaneously. The favorable formation of these bonds is connected with the fact that all conformers with $\Delta E_{\text{rel}} < 0.2$ eV are in the protomeric state **P3**, in which the carboxyl group is deprotonated. The bonds to the surface and a favorable vdW attraction effectively flatten the molecular conformation, thus energetically favoring more elongated structures. Protomers of type **P1**, which were dominant in the gas-phase, only appear with $\Delta E_{\text{rel}} > 0.3$ eV on Cu and Ag, and with $\Delta E_{\text{rel}} > 0.2$ eV on Au. Zwitterionic protomers **P4** and **P5** are again not observed. Regarding the intramolecular H-bond patterns, within 0.5 eV from the global minimum we can identify 7 different families on Cu(111), and 6 families on both Ag(111) and Au(111). These families contain H-bonds where the carboxyl

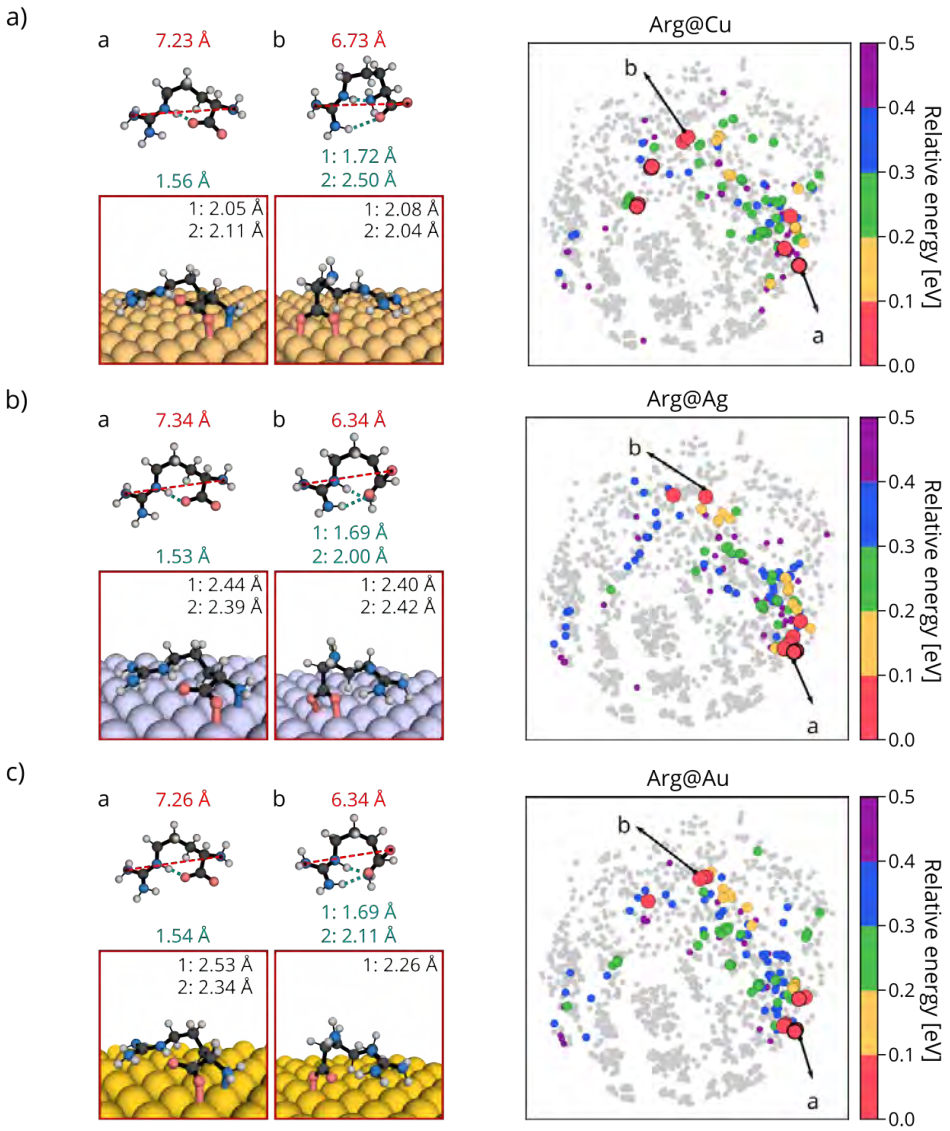


Figure 4.12 – Low-dimensional projections of conformers of Arg adsorbed on a) Cu(111), b) Ag(111), and c) Au(111), onto the gas-phase Arg map of Fig. 4.8. Only conformers within 0.5 eV of their respective global minimum are colored. Grey dots represent all structures from the original map of gas-phase Arg, and serve as a guide to the eye. In each panel, representative structures are shown from two perspectives: a side view where molecule and surface are shown (bottom), and the corresponding top view (top) where only the molecule is shown. The longest distance within each visualized conformer is reported in red and H-bond lengths are reported in turquoise.

group predominantly participates. All families are represented in Fig. 4.18.

Adsorption of Arg-H$^+$ on Cu, Ag, Au (111) surfaces

Finally, we characterize the conformational-space changes arising from the simultaneous addition of a proton and the adsorption onto metallic surfaces. In Figure 4.13, we show the projection of the low-dimensional representations of Arg-H$^+$ conformers adsorbed on Cu, Ag, and Au(111) onto the map of isolated Arg conformers. These projections, in particular the comparison of the plots in Figs. 4.12 and 4.13, reveal that the conformational space of adsorbed Arg-H$^+$ is larger than the one of adsorbed Arg. While Arg-H$^+$ features more than 500 conformers within $\Delta E_{\text{rel}} < 0.5$ eV, Arg only counts about 150 conformers in the same energy range. Interestingly, the adsorption of Arg-H$^+$ to a metal surface also results in an increase of the occupied structure space in comparison to isolated Arg-H$^+$ (108 structures with $\Delta E_{\text{rel}} < 0.5$ eV), shown in Fig. 4.10. In fact, the structures occupy similar regions of the map as the ones occupied by Arg-H$^+$, with the addition of extended structures that are located in the bottom of the map.

We identify 4 different families on Cu(111) and 3 on Ag(111) and Au(111) with $\Delta E_{\text{rel}} < 0.1$ eV. Representative conformers of these families are shown in Fig. 4.13. The lowest energy conformer, labeled *a* in Fig. 4.13(a)-(c), appears on all surfaces at the same region of the map as for adsorbed Arg. The largest distance within the molecule lies around 7 Å and it also has a strong H-bond linking the carboxyl-O and the Nε atoms. The structure, however, does not present the same orientation to the surface as compared to the lowest energy conformer of Arg, and does not form strong bonds with the surface. With the exception of the extended structure on Cu(111), labeled *d* in Fig. 4.13(a), all conformers with $\Delta E_{\text{rel}} < 0.1$ eV on all surfaces contain one intramolecular H-bond involving either carboxyl-O and Nε atom (labeled a), backbone N and Nε atoms (labeled b) or carboxyl O and a Nη atom (labeled c). Compared to adsorbed Arg, adsorbed Arg-H$^+$ structures become on average 1.0 Å more extended as shown in Fig. 4.14. The protomer **P6**, the only one present in the gas-phase, is dominantly populated also on the surfaces. However, we do observe a few conformers in the zwitterionic **P7** state. These structures are at least 0.2 eV higher in energy than than the global minimum.

With respect to the number of bonds that Arg-H$^+$ forms with the surface, the picture is very different from adsorbed Arg. Within the lower 0.15 eV, we do not observe short (strong) bonds of O or N atoms to the surfaces. This lack of constraint by the surface contributes to the increased structure space of adsorbed Arg-H$^+$ in comparison to Arg. In addition, the molecule accepts electrons from the surface, becoming less positively charged, as we discuss in detail in the next section. We conclude that Arg-H$^+$ interacts with the metallic surfaces mostly through van der Waals and electrostatic interactions.

4.0.4 Electronic structure and trends across surfaces

In the previous section we focused on structural aspects of the adsorbed molecules and the most prominent bonds the molecules make with the metallic surfaces. In the following, we

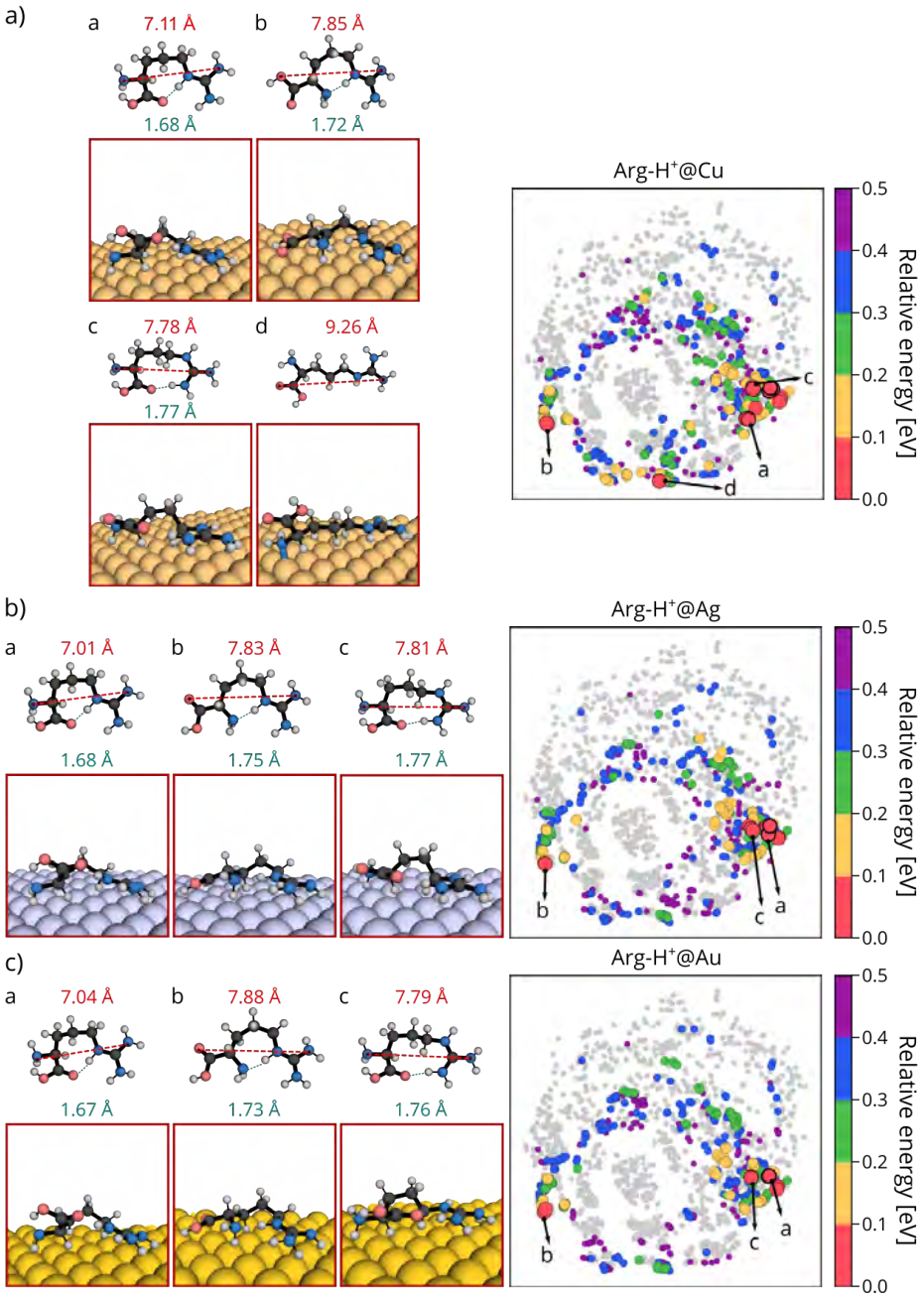


Figure 4.13 – Low-dimensional projections of conformers of Arg-H^+ adsorbed on a) Cu(111), b) Ag(111), and c) Au(111), plotted on the gas-phase Arg map of Fig. 4.8. Only conformers within 0.5 eV of their respective global minimum are colored. Grey dots represent all structures from the original map of gas-phase Arg, and serve as a guide to the eye. In each panel, representative structures are shown from two perspectives: a side view where molecule and surface are shown (bottom), and the corresponding top view (top) where only the molecule is shown. The longest distance within each visualized conformer is reported in red and H-bond lengths are reported in turquoise.

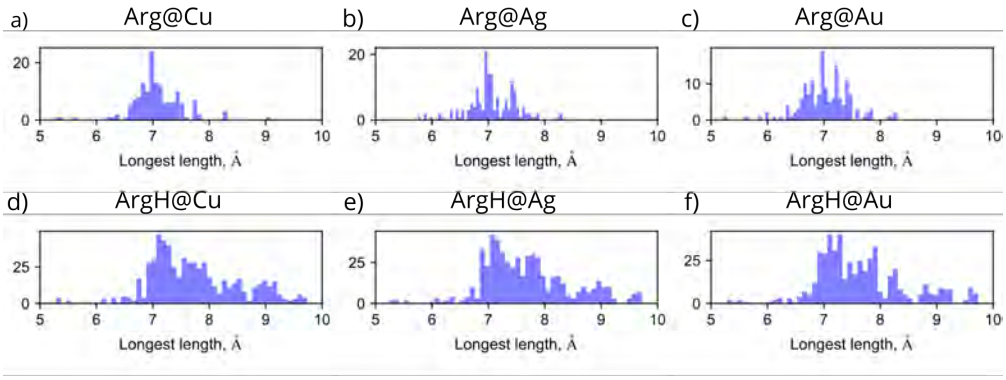


Figure 4.14 – Histogram of the longest distances of adsorbed molecules on different surfaces

will discuss different aspects of the molecule-surface interactions, with the goal of identifying trends across these systems.

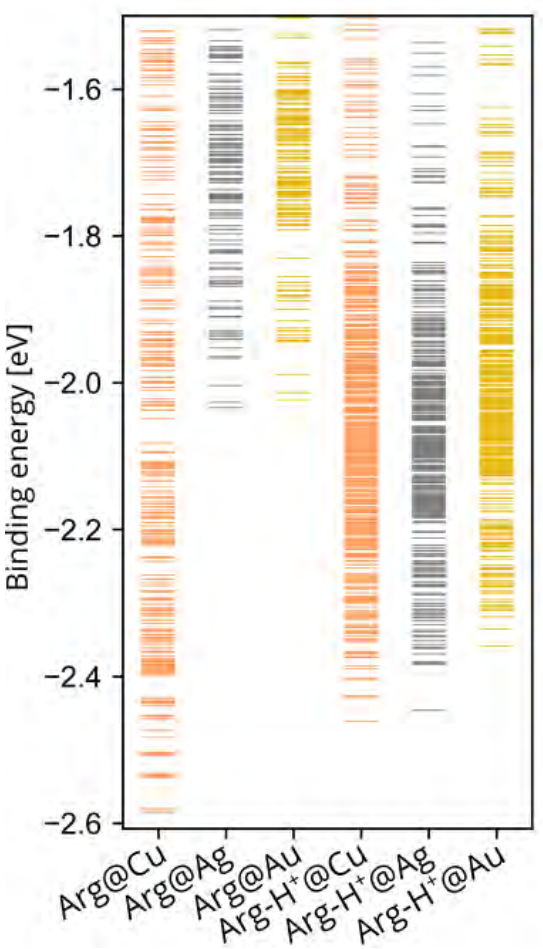


Figure 4.15 – Binding energies of Arg and Arg-H^+ on Cu(111), Ag(111) and Au(111) surfaces.

We begin by analysing the binding energies between the molecules and surface, which are

shown in Fig. 4.15. The binding energies for all surfaces were calculated as discussed in Section 2.8. The larger negative values in Fig. 4.15 correspond to stronger binding of the molecule to the surface. In the case of adsorbed Arg, many conformers bind to Cu more strongly than to Ag and Au, with the binding of the deprotonated carboxyl group of Arg to the Cu(111) surface geometrically favored as discussed above. In the case of adsorbed Arg-H^+, there is no pronounced difference in binding strengths to the different surfaces, and the values are comparable to the binding energies obtained for Arg adsorbed on Cu(111). This correlates with the observation that the interaction of Arg-H^+ with the surfaces happens mostly through dispersion and electrostatic interactions. Despite the strong binding to the surface, it is also visible from comparing Figs. 4.12 and 4.13 that the interaction of Arg-H^+ with the surface does not strongly template the conformations of this molecule, implying a low corrugation (i.e. homogeneity) of the molecule-surface interaction and allowing for a larger variety of conformers with similiar energies. This is in contrast to the molecule-surface interation of Arg, that is more inhomogeneous due to the formation of bonds through specific chemical groups. In realistic applications, the thermal energy will result in vibrational contributions to the stability of a conformer, potentially changing the energy hierarchy. In order to address the question about thermal stability of adsorbed structure, the free energies at finite temperatures within the harmonic approximation [281, 282] can be calculated:

$$F_{\text{harm}}(T) = E_{\text{PES}} + F_{\text{vib}}(\text{T}), \tag{4.1}$$

where E_{PES} is the total energy obtained from DFT (PBE+vdW$^{\text{surf}}$ functional), and we have used textbook expressions for the harmonic vibrational Helmholtz free energy $F_{\text{vib}}(T)$:

$$F_{\text{vib}}(\text{T}) = \sum_{i}^{3N-6} \left[\frac{\hbar\omega_i}{2} + k_{\text{B}} T \ln\left(1 - e^{-\beta\hbar\omega_i}\right) \right],$$

where N is the total number of atoms in the molecule (metal atoms were not displaced and were taken into account in external field), k_{B} is Boltzmann constant, T is the temperature, ω_i are vibrational frequencies obtained by diagonalization of Hessian matrix with use of developing version of phonopy-FHI-aims [283, 284]. For the adsorbed conformers, rotational contributions are completely neglected since rotation around all principal axes of the molecule become internal vibrational modes of the system.

We have estimated harmonic vibrational free energies for representative conformers with $\Delta E_{\text{rel}} < 0.1$ eV in each surface. In contrast to what has been reported for longer helical peptides [285, 286], the global minimum remains the same in all cases, as reported in Fig. 4.16. For Arg-H^+ we observe relative energy rearrangements of up to 50 meV at 300 K, which changes the relative energy hierarchy of conformers less stable than the global minimum. Therefore, vibrational effects must be considered in order to obtain an accurate energy hierarchy at a given temperature.

We then focus on the distance between the molecule and the surfaces. We define this quantity by measuring the distance of the center of mass (COM) of the molecule with respect to the

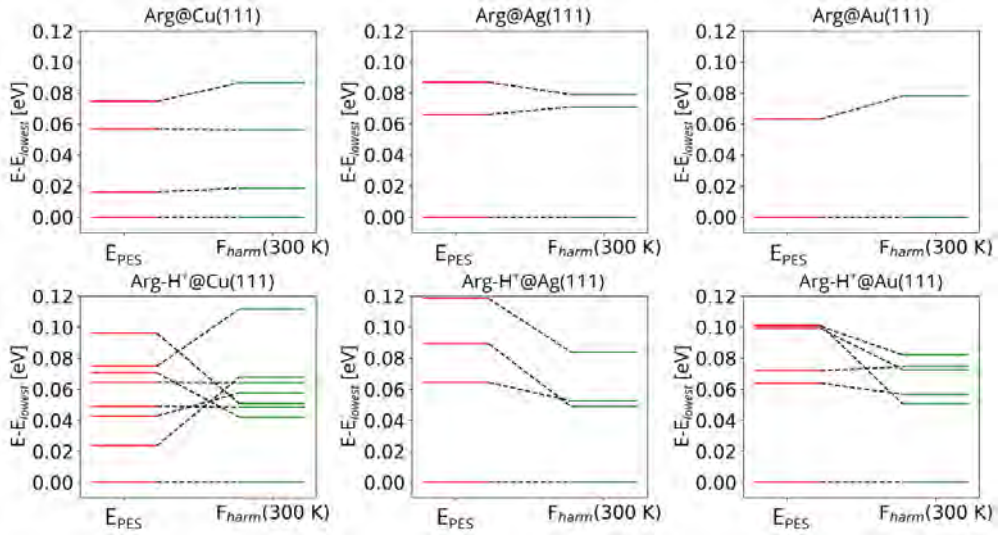


Figure 4.16 – Harmonic free energies calculated for adsorbed structures within the lowest 0.1 eV total-energy range. $E_{\mathbf{PES}}$ corresponds to the total energy of the system obtained at DFT level and $F_{\mathbf{harm}}$ corresponds to the free energy of the system at 300 K calculated as described above.

surface plane defined by the top layer of surface atoms. These distances are collected in Fig. 4.17. The COM is closer to Cu(111) than to Ag(111) and Au(111) for both Arg and Arg-H^+, because of higher reactivity of Cu. In addition, in all surfaces, Arg lies closer than Arg-H^+, in agreement with the observation that Arg forms covalent bonds to the surface. The extended structures of Arg-H^+, at the bottom of the maps, tend to be closer to the surface than those that have H-bonds within the molecule, likely due to the stronger vdW attraction to the surface by extended conformations.

The difference in COM distances to the surfaces between Arg and Arg-H^+ is apparently related to the preferred orientations of the chiral center of the molecule to the surface. The chiral C_α carbon can point its bonded hydrogen towards the surface (labeled *down* in the following), or towards the vacuum region (labeled *up* in the following). Examples of these different molecular orientation are shown in Fig. 4.19(a).

The dominant orientation with respect to the surface is different in the cases of Arg and Arg-H^+, as evidenced by the numbers presented in Fig. 4.19(b). The lower energy structures are mostly in the *up* orientation for Arg and mostly in the *down* orientation for Arg-H^+ (see also map in Fig. 4.20), consistent with the typically smaller distance to the surface for adsorbed Arg. However, despite the different orientations of their $C_\alpha H$ groups, the lowest energy structures for both molecules adsorbed on each surface have very similar conformations. Since the addition or removal of a proton can apparently alter the preference of the chiral-center orientation, we propose that it could template different chiralities of self-assembled super-structures on the surface [27].

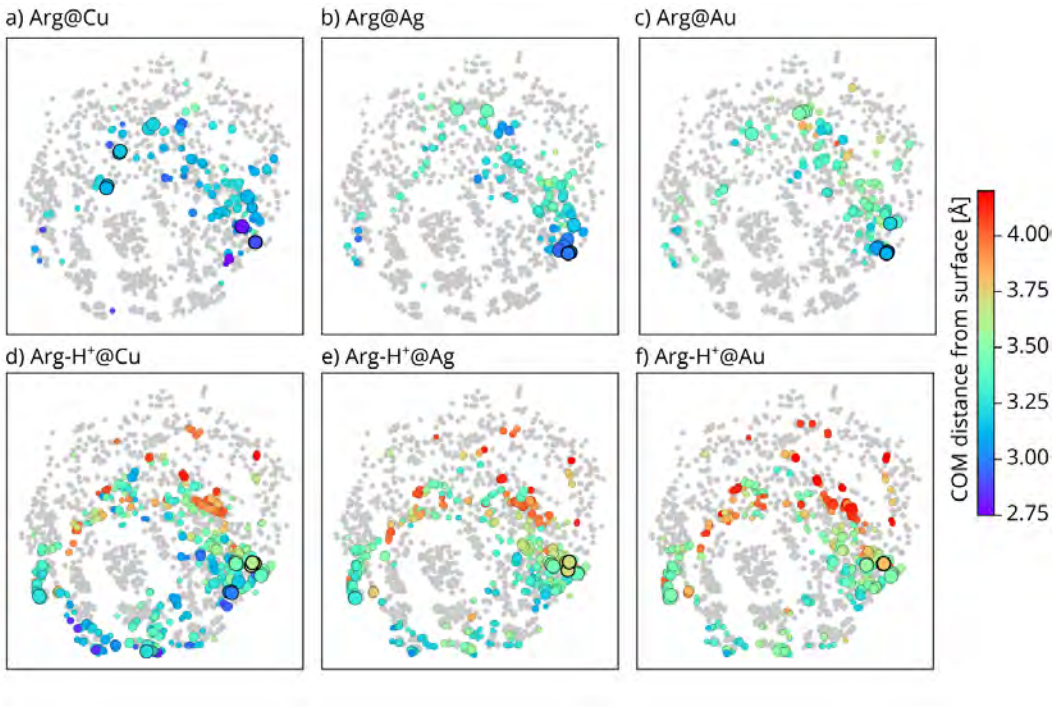


Figure 4.17 – Low dimensional projections of adsorbed Arg and Arg-H^+ on Cu(111), Ag(111) and Au(111) color-coded with respect to the distance of the center of mass of the molecule with respect to the surface. Grey dots represent all structures from the original map of isolated Arg where the projection was made, and serve as a guide to the eye.

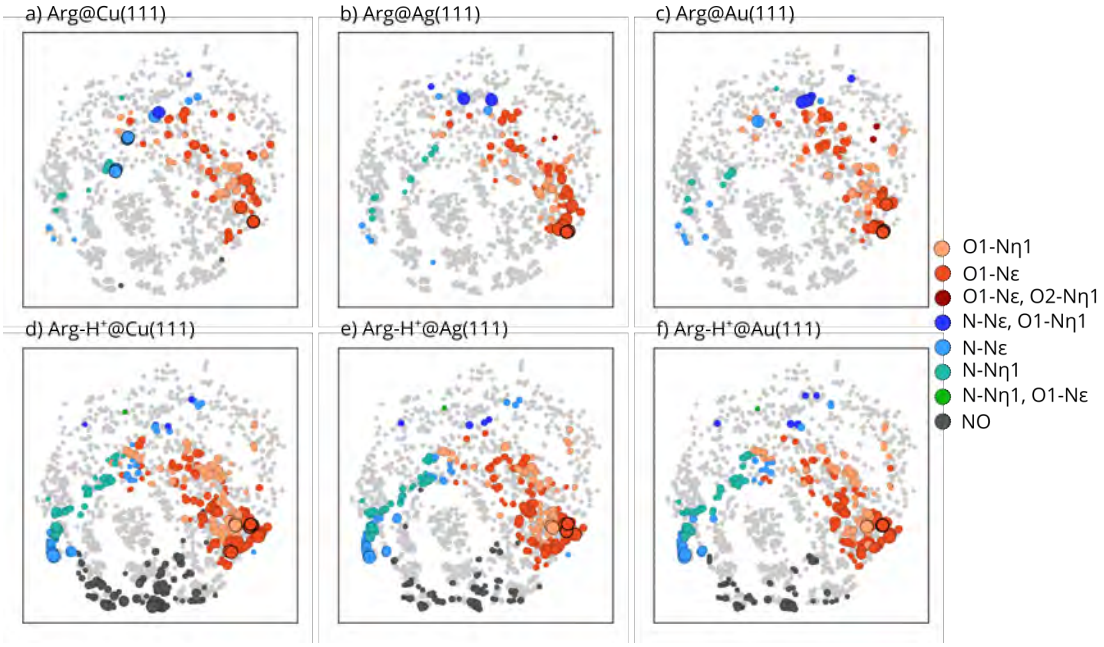
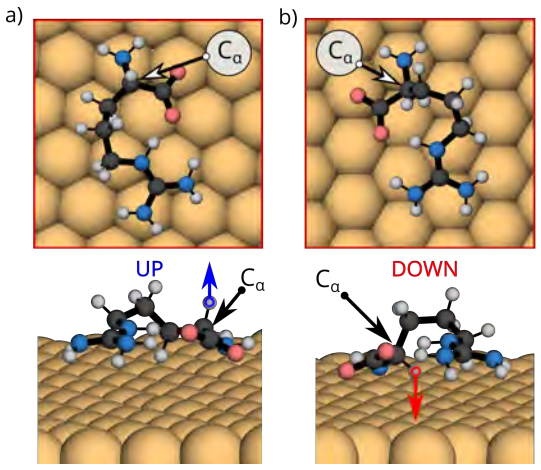


Figure 4.18 – Projection of Arg and Arg-H^+ conformers adsorbed on the different metalic surfaces on the low-dimensional map of gas-phase Arg, colored according to the H-bond pattern.



	Arg		Arg-H^+	
	up	*down*	*up*	*down*
Cu	12/100	0/48	8/255	13/355
Ag	10/110	0/50	3/255	10/292
Au	13/104	1/53	2/219	12/284

Figure 4.19 – Orientation of the $C_\alpha H$ group in a) *up* orientation (hydrogen pointing towards vacuum) and b) *down* orientation (hydrogen pointing towards the surfaces). c) The amount of structures with *up* and *down* orientation within 0.1/0.5 eV from the global minimum of each surface.

We then investigated the rearrangement of the electronic density upon binding of the molecules to the different surfaces. In Fig. 4.21 we show the electronic density rearrangement created by the lowest energy conformer at each surface, integrated over the axis parallel to the surface, overlaid on the side-view of the 3D density rearrangement. In addition, we show

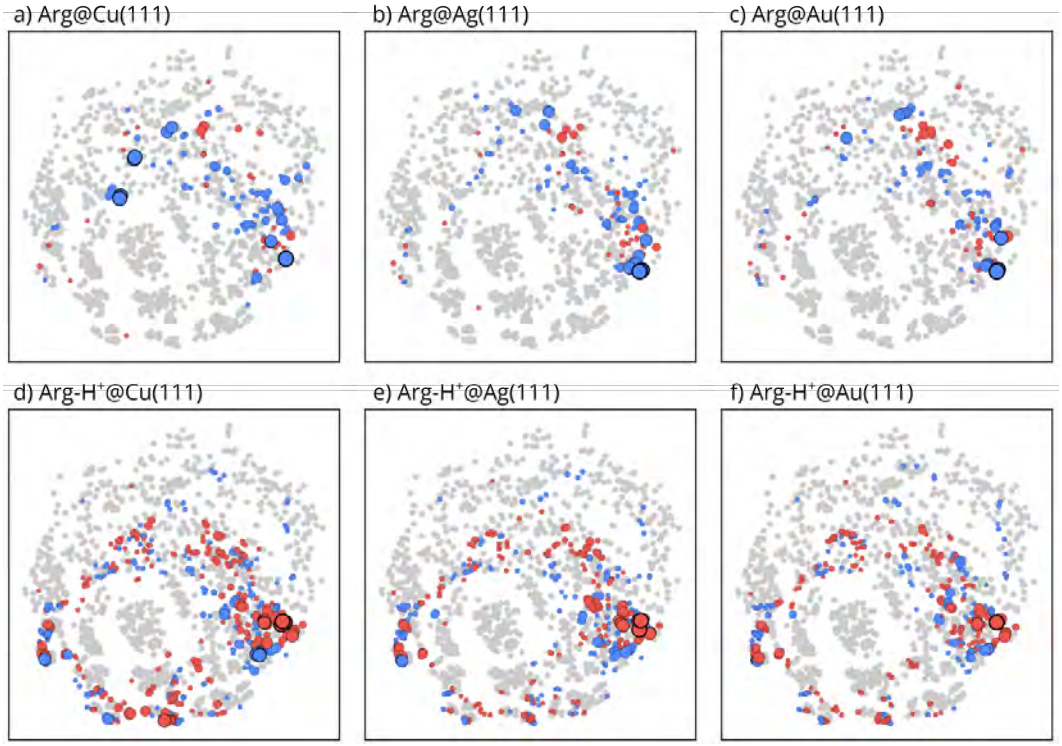


Figure 4.20 – Low dimensional maps of Arg and Arg-H^+ adsorbed on Cu(111), Ag(111) and Au(111) color-coded with respect to the orientation of the $C_\alpha H$ group. Blue correspond to *up* orientation and red correspond to *down* orientation of the $C_\alpha H$ group.

a top view of the density rearrangement in each case. Examples of further conformers are summarized in the Appendix. The data shows that Arg donates electrons to the surface, while Arg-H^+ accepts electrons from the surface. We have checked this propensity for selected conformers by integration of the electronic density rearrangement around the molecule and by calculating the Hirshfeld charge remaining on the molecule for the full database (see Table 4.5). When comparing Hirshfeld charges on the molecule and those obtained from the electronic density rearrangement, we observe that Hirshfeld charges are always 0.3-0.5 e underestimated, making them an unreliable method to analyse charge trasfer.

In addition, we observe that the depletion and accumulation of charge is not uniform through the lateral extension of the molecule. This behavior is consistent with the level alignment predicted by the PBE Kohn-Sham energy levels, as shown in Fig. 4.22. However, we note that quantitative values of charge transfer are often inaccurate at this level of theory, as characterized in Refs. [287, 288]. Optimally tuned range-separated hybrid functionals would yield more accurate values, but their computational cost is prohibitive for the use in this whole database. Nevertheless, hybrid-functional calculations (PBE0) of selected conformers (Fig. 4.23) confirm the qualitative trend. Therefore, we conclude that the protonation state again critically impacts these systems, in this case by qualitatively changing the redistribution of electronic charge.

It was observed experimentally that amino acids can undergo deprotonation on reactive surfaces [289–294]. Here we also investigated whether deprotonation of Arg and Arg-H^+ was favorable on any of the surfaces studied here. In Arg, we found it most favorable to detach the proton from the guanidino group, while for Arg-H^+, it was most favorable to detach the proton from the carboxyl group. We chose three representative conformers at each surface: the lowest energy structure and two others with different H-bonds within the molecule. We placed the detached proton at a distance of at least 2.5 Å from the molecule and fully optimized the dissociated structures. Comparing the energy difference between the final

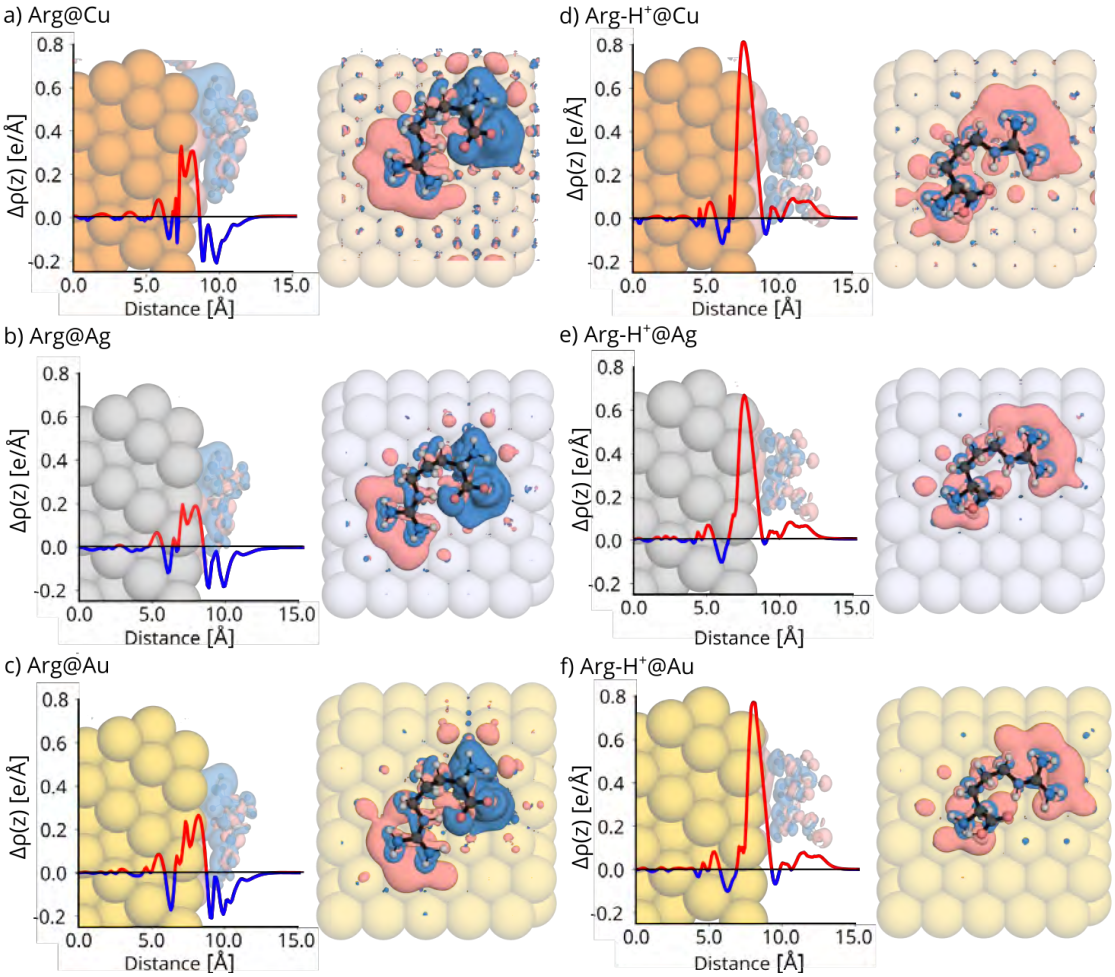


Figure 4.21 – Electronic-density difference averaged over the directions parallel to the surface for the lowest energy conformers of Arg adsorbed on Cu(111) (a), Ag(111) (b), and Au(111) (c), as well as of Arg-H^+ adsorbed on Cu(111) (d), Ag(111) (e), and Au(111) (f). Positive values (red) correspond to electron density accumulation and negative values (blue) correspond to electron density depletion. In each panel, we also show a side and top view of the 3D electronic density rearrangement. Blue isosurfaces correspond to an electron density of +0.05 e/Bohr^3 and red isosurfaces to -0.05 e/Bohr^3.

Table 4.5 – Calculated charge on the molecule with use of Hirshfeld partial charge analysis and by integration of the electron density difference in the molecular region. Values are in electrons.

Conformer	Hirshfeld	Integral	Conformer	Hirshfeld	Integral
Arg@Cu			Arg-H^+@Cu		
a	0.11	0.19	a	0.29	0.85
b	0.03	0.30	b	0.30	0.85
c	0.04	0.31	c	0.31	0.84
d	0.08	0.26	d	0.43	0.88
e	0.01	0.24	e	0.46	0.85
f	0.11	0.30	f	0.38	0.82
Arg@Ag			Arg-H^+@Ag		
a	0.04	0.15	a	0.28	0.83
b	-0.08	0.23	b	0.30	0.83
c	-0.03	0.24	c	0.31	0.82
d	-0.06	0.21	d	0.43	0.86
e	-0.13	0.16	e	0.46	0.85
f	0.05	0.14	f	0.36	0.86
Arg@Au			Arg-H^+@Au		
a	0.06	0.05	a	0.32	0.86
b	-0.01	0.29	b	0.29	0.86
c	0.00	0.30	c	0.34	0.85
d	-0.10	0.25	d	0.48	0.91
e	0.01	0.23	e	0.49	0.90
f	0.06	0.31	f	0.43	0.92

and initial states gives a lower limit for the dissociation barrier:

$$\Delta E = E_{\text{dissociated}} - E_{\text{lowest}}. \tag{4.2}$$

The results are summarized in Figs. 4.24 and 4.25. They show that, however, only the deprotonation of Arg-H^+ is favorable on Cu(111), such that Arg-H^+ would be predominantely deprotonated. However, we have not observed any spontaneous dissociation upon optimization of Arg-H^+ on Cu(111), leading us to conclude that, although favorable, this dissociation of H does not occur without a barrier. In all other surfaces, the barrier for dissociation would be rather high for both molecules.

4.0.5 Comparison of DFT with INTERFACE FF

Comparing DFT results with existing FFs is usually beneficial since it helps develop less expensive and more accurate potentials. All the local minima obtained at DFT level of theory were optimized with the INTERFACE-FF [213] using the NAMD package [201]. Calculations were performed with periodic boundary conditions with the same cell size and number of Cu atoms as in the DFT calculations. We obtained parameters for certain protonation

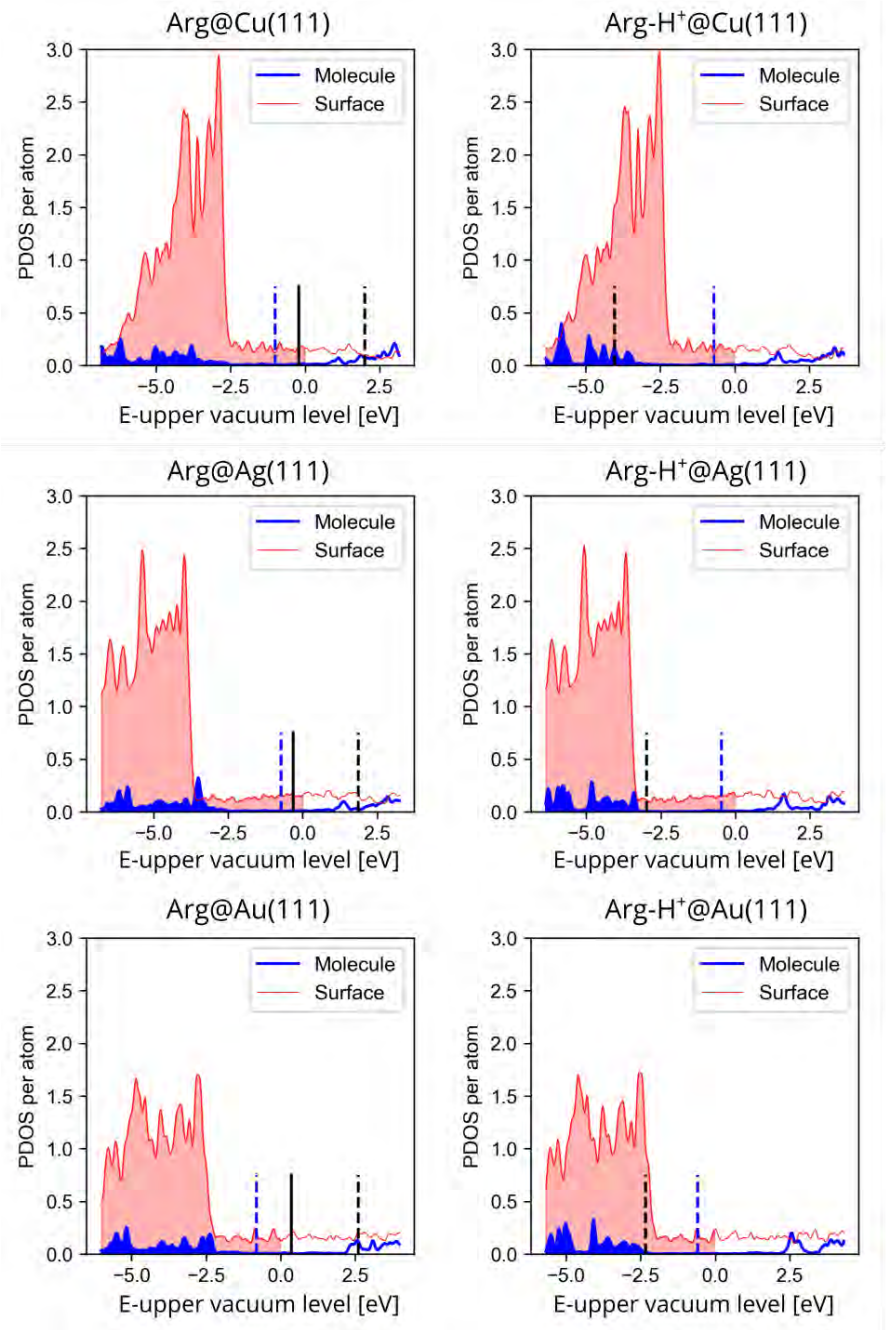


Figure 4.22 – Projected densities of states of the lowest energy structures on each surface. Filled area corresponds to the occupied states below highest occupied state (VBM) of the whole system. HOMO (black solid line) and LUMO (black dashed line) are the states of the corresponding gas-phase molecular conformer calculated with the same geometry as it adopts when adsorbed. The Fermi energy of the pristine slab is depicted with blue dashed line.

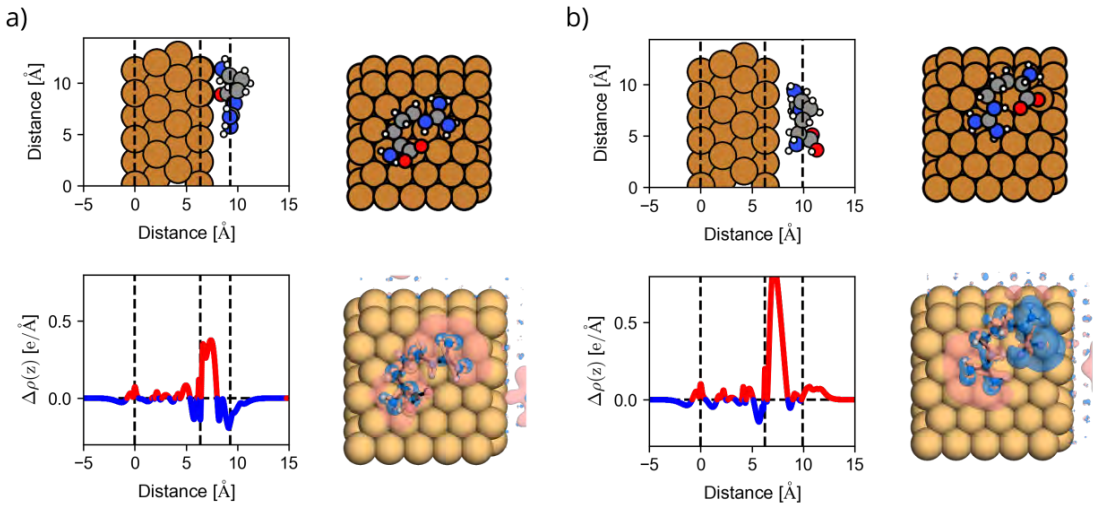


Figure 4.23 – Side and top views of the adsorbed structures of a) Arg on Cu(111) and b) Arg-H$^+$ on Cu(111). Dashed black lines correspond to: average z position of the atoms in the lowest layer of the surface (left), average z position of atoms in the highest layer of the surface (middle), centre of the mass of the molecule (right). Red/blue solid lines (and also red/blue regions) correspond to the electron density accumulation/depletion with PBE0 functional.

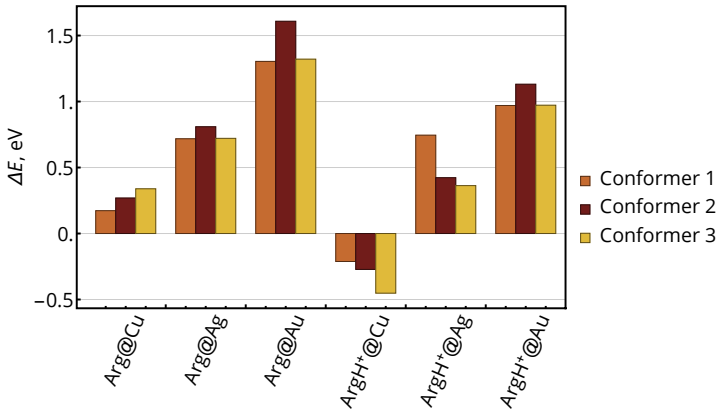


Figure 4.24 – Energy differences upon hydrogen dissociation for selected conformers of Arg and Arg-H$^+$ on all metallic surfaces. $\Delta E = E_{\mathbf{dep}} - E$, where $E_{\mathbf{dep}}$ is the total energy of the dissociated structure after optimization (including the adsorbed hydrogen) and E the energy of the optimized intact structure. A negative ΔE indicates that deprotonation is favored.

states from existing parametrization of Arg and Arg-H$^+$ available from CHARMM FF. For the calculation of Arg, two protomers **P1** and **P3** had to be prepared.

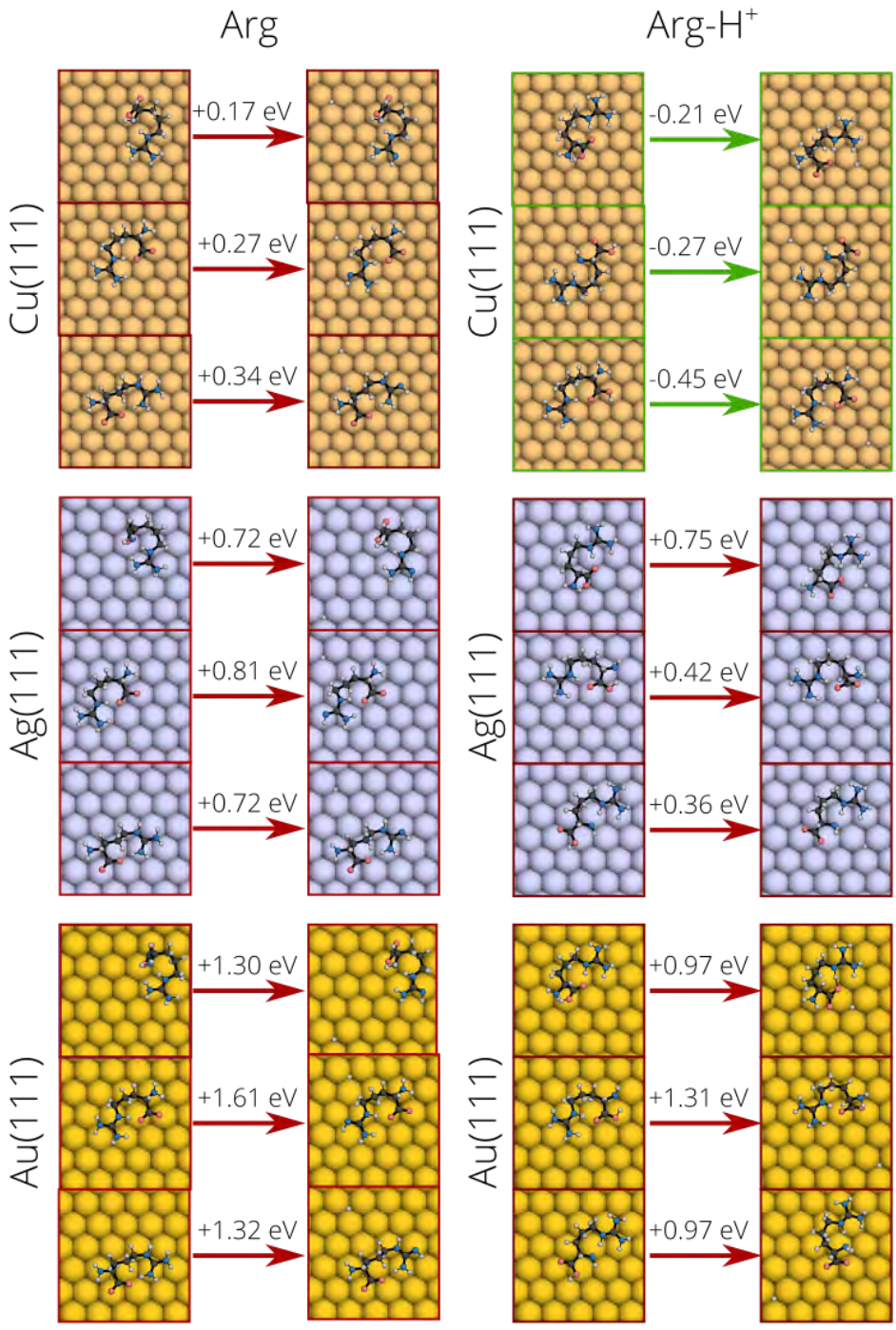


Figure 4.25 – All structures that were analyzed for the calculation of the deprotonation energies. ΔE is also reported in each panel.

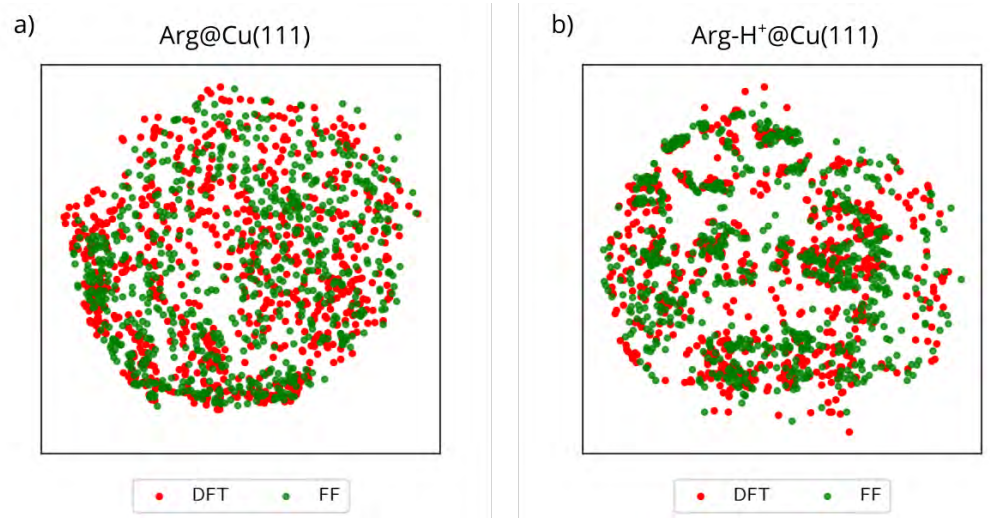


Figure 4.26 – Low-dimensional map of the conformational space of the Arg and Arg-H^+ molecules adsorbed on the Cu(111) surface. The map was optimized considering all DFT and INTERFACE-FF structures. Green dots represent conformations obtained at DFT level of theory and red dots represent conformations obtained after geometry optimization with INTERFACE-FF. Close proximity of the dots reflects their structural similarity.

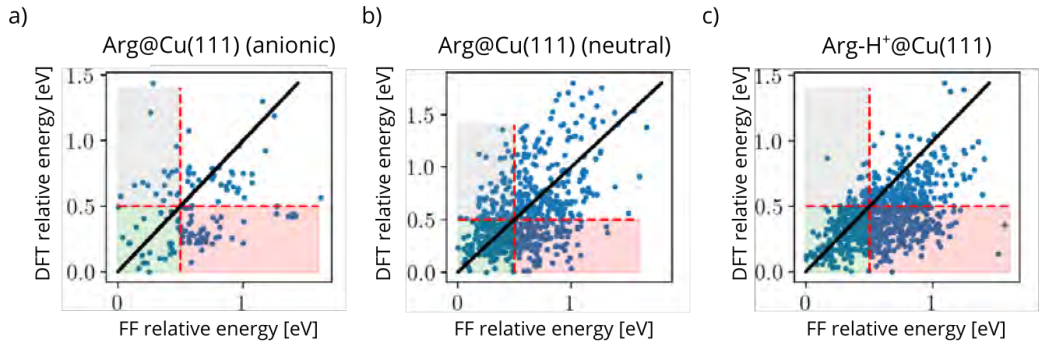


Figure 4.27 – Comparison of the relative energies obtained from DFT optimized structures and the same structures after post-relaxation in with the INTERFACE force field.

From Fig. 4.26, we conclude that both levels of theory represent a similar conformational space. However, Fig. 4.27 shows the comparison of the relative energies obtained from DFT optimized structures and the same structures after post-relaxation in with the INTERFACE-FF. Dots on the diagonal line represent an optimal correlation. The red area marks structures that lie in the lower 0.5 eV energy range in DFT but above the 0.5 eV energy range in INTERFACE-FF. The green area marks the structures that are in the lower 0.5 eV energy range regardless of the level of theory. The grey area marks the structures that are above the 0.5 eV energy range in DFT but below the 0.5 eV energy range in INTERFACE-FF. From this, we conclude that DFT (PBE+vdW^{surf}) and the INTERFACE-FF yield very different energy hierarchies. Furthermore, Table 4.6 shows that DFT and the FF yield different adsorption site preferences for the amino and carboxyl groups. In particular, DFT predicts that O will adsorb almost exclusively on top

sites, consistent with the accepted adsorption site preference of CO groups on the pristine Cu(111) surface. The FF predicts a larger population of other adsorption sites, in particular hollow sites, compared to DFT.

Table 4.6 – Surface site adsorption preferences of chosen chemical groups in Arg and Arg-H^+. All numbers are reported as a percentage of the total number of conformers optimized with DFT (PBE+vdW^{surf}) and the INTERFACE-FF.

	Arg@Cu(111)				Arg-H^+@Cu(111)			
	Amino		Carboxyl		Amino		Carboxyl	
Adsorption site	DFT	FF	DFT	FF	DFT	FF	DFT	FF
Top	80	53	76	48	59	50	70	45
Bridge	9	18	14	18	18	20	15	22
Hollow-FCC	5	13	4	17	13	15	7	16
Hollow-HCP	6	16	5	17	10	15	9	18

INTERFACE-FF is not reliable for estimation of the energy hierarchies of the molecules, even though the conformational spaces of DFT and FF are very similar. To go beyond single molecules we still need better FFs or ML potentials.

4.0.6 Conclusions

One of the results of this chapter is the creation of the database of Arg and Arg-H^+ adsorbed on three metal surfaces (Cu(111), Ag(111) and Au(111)) containing thousands of structures optimized using DFT. This database is publicly available to download via NOMAD repository [295]. In order to accelerate the development of parametrization of FFs and the training of ML potentials, it is necessary to share these databases to overcome the bottleneck of computationally expensive DFT geometry optimizations, which are required for obtaining relevant information about structure-property relations of interface systems.This is required to achieve the synergy between theory and experiment, in which computational findings may shed light on characteristics of systems that are not accessible via experiment.

Then, using a state-of-the-art dimensionality reduction method, we investigated the conformational spaces of Arg and Arg-H^+ in isolation and after adsorption on metal surfaces. The unsupervised dimensionality reduction technique appeared to be a very powerful tool for the rapid analysis of systems with a large number of degrees of freedom. We managed to easily conclude that all structural motifs of all adsorbed systems are already represented in the conformational space of Arg. In comparison to isolated Arg-H^+, the number of accessible conformations substantially increased after adsorption. Another intriguing discovery that might be easily overlooked without conformational analysis is that the lowest energy structures of adsorbed Arg and Arg-H^+ have remarkably similar conformations since they occur in the same regions of the low-dimensional maps. A closer examination of these lowest energy structures reveals that the dominating orientation of the $C_\alpha H$ group relative to the surface varies between Arg and Arg-H^+. This feature should be studied further for other systems since it may govern the templating of various chiralities of self-assembled structures

on the surface. Additionally, a visual depiction of the accessible regions of a conformational space can be provided. For example, spiral-like conformations that lack H-bonds are unfavourable for both Arg and Arg-H^+, while extended structures are favourable for just Arg-H^+. After that, we have specifically investigated why different parts of the conformational space become accessible or are excluded depending on the protonation state and the environment, demonstrating the importance of bond formation and charge rearrangement in these systems.

Arg adsorption occurs through the formation of strong bonds with the surface, with carboxyl and amino groups playing major roles. The surface bindings limit the conformations of this molecule, reducing the number of possible configurations with respect to the numbers observed in the gas-phase. In contrast, Arg-H^+ receives electrons from the surface and becomes less positively charged, which leads to the number of allowed conformations to increase compared to isolated Arg-H^+, which is due to the weakening of intramolecular H-bonds.

After adsorption on Cu, Ag, and Au surfaces, we analyzed the patterns observed for Arg and Arg-H^+. When the substrate is changed, the relative energy order of conformers is mainly conserved, which is a pretty counterintuitive observation. The average adsorption height of the molecules is following the trend: Cu(111) < Ag(111) < Au(111), and Arg is always closer to the same respective surface than Arg-H^+. Most adsorbed Arg conformers bind to Cu(111) surface more strongly than to Ag(111) or Au(111). However, adsorbed Arg-H^+ has similar binding strengths to all surfaces as Arg adsorbed on Cu(111). The computation of dissociation energies leads us to the conclusion that deprotonation of Arg-H^+ is only energetically favourable on Cu(111).

Finally, we show that while INTERFACE-FF may sample the relevant conformational space of these adsorbed molecules, it cannot capture consistent energy hierarchies. Databases like the ones we established will be a valuable source of data for future parameterization and development of cheaper potentials.

In general, there is no accessible collection of isolated local minima conformers to start a structure search from, for any random system of interest. Few suitable packages exist for such tasks, and all methods for creating starting structures with different molecular orientations with respect to the surface must be established manually. In the following chapter, we will present a package that will assist in carrying out these calculations, paving the way for the acceleration of database development for interface systems.

[50mm] It turns out that any repetitive endeavour – whatever the industry – can be automated within the context of rising digitisation. “Fully Automated Luxury Communism: A Manifesto”, Aaron Bastani

5 Generation and search of the flexible molecules with respect to fixed surroundings

The previous chapter was devoted to the study of single molecule adsorption on various surfaces, and it required a significant degree of human engagement in terms of data production, data organization, and data interpretation. In order to capture the trends across all the amino acids on different surfaces, such work should be performed for other systems as well. However, it is not common to have an available structure database for gas-phase structures that are useful as beginning structures. Moreover, in cases where the adsorption pattern is composed of repeating templates, it is best to take into account PBC in order to perform the structure search. In the age of high performance computers, the workflow should make use of parallelization in the data acquisition process. Existing software packages that are capable of performing sampling of conformational spaces are typically coupled to a small number of specific electronic structure packages, which limits the usefulness of such packages in practice. Also structure search packages are not tailored to sample flexible adsorbates and their assemblies with respect to specified surroundings e.g surfaces or cavities. In response to these challenges we have developed a program that addresses all of these issues and is meant for sampling the conformational spaces of flexible molecules and their assemblies on surfaces. In this chapter we present an automated workflow that allows us to easily generate and perform geometry optimizations.

5.1 GenSec package for structure search of the interfaces

Random structure search is the basis for more sophisticated methods such as Bayesian optimization [227] and evolutionary algorithms [217, 220], and is the method employed in the Generation and Search (GenSec) package. Random structure search is also used in crystal structure prediction [216, 296] and shows a decent probability of identifying low-energy minima [214, 215]. The efficiency of the random structure search can be increased dramatically first by imposing constraints on the generated structures, avoiding clashes between atoms and keeping the database of previously calculated structures in order to avoid repetitive calculations. Starting from the procedure for generating different conformers of the isolated molecules, we then describe the extension of such procedures to enable simulations of these conformers with respect to fixed surroundings (*fixed frames*) that can be, in general, 1D (e.g.

ions), 2D (e.g. surfaces) or 3D (e.g. solids) static references. In short, GenSec performs a quasi-random global structure search, with the ability to choose different internal degrees of freedom and sample them with respect to specified fixed surroundings. The geometry optimizations are performed by a connection with the Atomic Simulation Environment (ASE) [297] environment, which can be connected to many electronic structure and FF packages and offers the choice of a variety of geometry optimization routines, which we have improved as detailed in Section 5.6. The connection to the ASE database support makes it possible to perform multiple searches in parallel with shared access to the information obtained from all the searches.

GenSec is written using Python 3 and distributed under the GNU Lesser GENERAL Public License and available from:

`https://github.com/sabia-group/gensec`

5.2 Workflow of the GenSec package

The workflow of GenSec consists of the three main steps (Fig. 5.1):

1. Random generation of a candidate structure with specified constraints
2. Comparing the generated structure with the structures already contained in the databases
3. Performing a geometry optimization if the structure is unique, and adding all optimization steps from the geometry relaxation as well as the local minima to the database

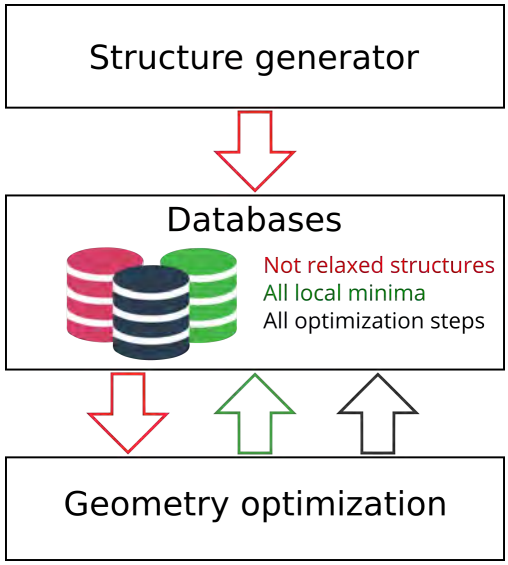


Figure 5.1 – Workflow of the GenSec package.

The search performs a user-specified number of unique relaxations, or the algorithm stops if it cannot find any more unique structures within the user-specified number of trials. The

processes of structure generation and geometry optimization can be parallelized and run independently, and the details of each step are described in the following sections.

5.3 Structure generation

The generation of structures is implemented as a standalone procedure, and can generate structures via multiple independent processes, while creating a central database, where the unique unrelaxed structures are stored. The generation of these structures is based on the internal degrees of freedom of the molecule, such as the dihedral angles, position of center of mass (COM), and orientation of the molecule. Starting from the generation of different conformers of isolated molecules, we then extend the procedure to generate self-assemblies on surfaces.

5.3.1 Internal degrees of freedom: dihedrals

The very first step is to identify the connectivity of the molecule. ASE allows reading the molecule 3D coordinates from a template in multiple chemical formats (Fig. 5.2 a), after which it creates the connectivity matrix based on the covalent radii distances between atoms. If the spheres of two atoms defined by their atomic covalent radii that are tabulated in ASE, overlap, they will be counted as bonded atoms. This connectivity matrix is then represented as a undirected graph that reflects the bonding information between atoms as shown on the Fig. 5.2 b. The dihedral angle for organic chemical systems is defined as the angle between two planes both of which are defined by three atoms that are connected by two bonds and both of the planes have to share the bond that is not the terminal bond of both planes [298]. For producing different conformers with the same chemical bonding we are interested in changing of dihedral angles of those planes, where the shared bond is freely rotatable. The rotatable bonds are identified from the graph in Fig. 5.2 b) with the following rules:

1. First select all the atoms that have two or more bonds - potentially they will be two central atoms forming the dihedral angle if they are not in a cyclic structure

2. Exclude the atoms with exactly 4 bonds, three of which are terminating atoms. Such exclusion removes e.g. CH_3 terminating groups

3. Exclude the atoms with three bonds for which two of the atoms are terminating hydrogens - with that we also exclude groups such as NH_2

4. Finally exclude the atoms that have two bonds, one of which is a terminating hydrogen which appears in carboxyl group

With this procedure the rotatable bonds of the molecules can be automatically identified after construction of the connectivity matrix of the template molecule and in the case of di-L-alanine only four rotatable bonds will be identified and used for creating of the different conformations. It should be mentioned that we pay additional attention to exclusion of the rotatable groups containing light hydrogen atoms. These exclusion can be allowed

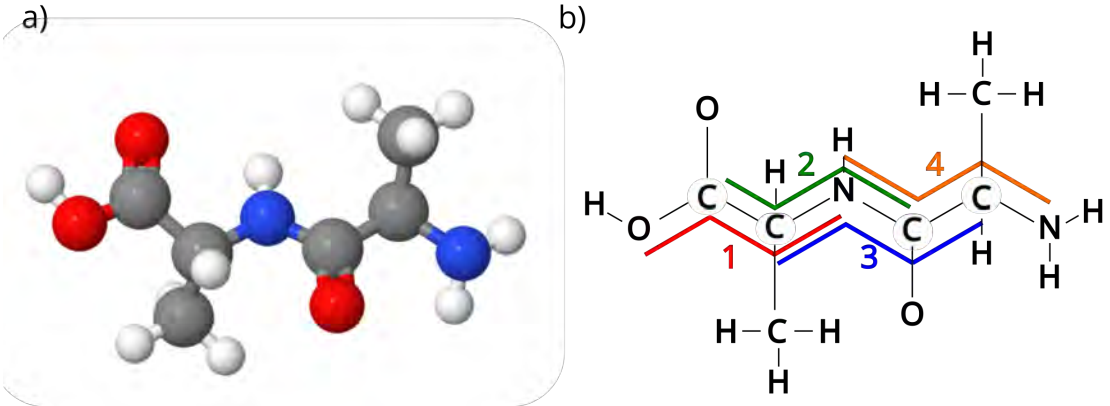


Figure 5.2 – a) 3D representation of a flexible molecule (di-L-Alanine); b) representation of di-L-Alanine as undirected graph together with rotatable bonds automatically identified using GenSec coloured in red, green, blue and orange.

since during geometry optimization light atoms will anyways move if necessary resulting in preferred orientation of the whole chemical group with respect to the rest of the molecule. If particular rotatable bonds are of interest during the search, this information can be anyways manually specified in the parameters file. The only thing left to address is that the rotatable bonds obtained with the algorithm described above can occur in cycles, which creates redundant degrees of freedom. To exclude the rotatable bonds that appear in cycles we use the networkx package [299] that uses Johnson's algorithm to detect cycles in a graph [300]. The rotatable bonds are then excluded by simple filtering that requires at least one of the central atoms not to be in a cycle.

After that, random values of the dihedral angles can be applied to these rotatable bonds through the ASE interface. The resulting molecule is checked for internal clashes by constructing the connectivity matrix again and comparing it with the initial template. The procedure described up to this point enables the generation of random isolated conformers. In order to model adsorbed species, additional degrees of freedom such as orientation and positioning of the molecule with respect to fixed frames had to be implemented.

5.3.2 Generating molecules with respect to fixed frames

In order to sample the configurational space of rigid molecules with respect to fixed frames we added two additional degrees of freedom to a template molecule: the orientation and positioning of the COM of the molecule. The COM of the molecule is a simple translational degree of freedom, which locates the molecule relative to a specific origin in Cartesian coordinates. The COM is defined as $\mathbf{r}^* = \sum_k m_k \mathbf{r}_k / \sum_k m_k$, where m_k and $\mathbf{r}_k$ are the mass and coordinates of the k-th atom in the molecule.

For the orientation of the molecules we must introduce a notation to describe the orientation of the molecule, which is not trivial in the case of the flexible structures. Rotations are performed with using Hamilton's quaternions [301], which are closely related to the

geometrically intuitive angle and axis notation. These are presented as an ordered set of 4 real quantities which we write as

$$\mathbf{q} = [q_0, q_1, q_2, q_3],$$

or as a combination of a scalar and a vector

$$\mathbf{q} = [q_0, \mathbf{v}],$$

where $\mathbf{v}$=$[q_1, q_2, q_3]$. In order to use quaternions for spatial rotations around some unit vector $|\mathbf{v}|$=1 on angle θ, we can use a unit quaternion $\mathbf{q} = [\cos(\theta/2), \mathbf{v}\sin(\theta/2)]$, with rotations implemented as the action of an operator $R_\mathbf{q}$ on a 3-dimensional vector:

$$R_\mathbf{q}(\mathbf{x}) = \mathbf{R}(\mathbf{q}) \cdot \mathbf{x},$$

where $\mathbf{x}$ are Cartesian coordinates of atoms in the system and $\mathbf{R}(\mathbf{q})$ is a matrix which in component form can be written as follows:

$$\mathbf{R}(\mathbf{q}) = \begin{pmatrix} 1-2q_2^2-2q_3^2 & 2q_1q_2-2q_0q_3 & 2q_1q_3+2q_0q_2 \\ 2q_2q_1+2q_0q_3 & 1-2q_3^2-2q_1^2 & 2q_2q_3-2q_0q_1 \\ 2q_3q_1-2q_0q_2 & 2q_3q_2+2q_0q_1 & 1-2q_1^2-2q_2^2 \end{pmatrix}. \tag{5.1}$$

In order to describe a rotation of the molecule such that it can be compared to other rotations, we use the orientation associated with the eigenvectors of the inertial moments of the rigid molecule. The moment of inertia matrix is given by

$$\mathbf{I} = \sum_k m_k((\mathbf{r}_k \cdot \mathbf{r}_k)\mathbf{E} - \mathbf{r}_k \otimes \mathbf{r}_k), \tag{5.2}$$

where m_k and $\mathbf{r}_k = (x_k, y_k, z_k)$ are the masses and coordinates of k-th atom in the molecule, $\mathbf{E}$ is the identity tensor and $\otimes$ is the tensor product. The eigenvector with the lowest corresponding eigenvalue (shortest principal axis) is chosen as the main vector of the molecule. The eigenvector with the corresponding largest eigenvalue (longest principal axis) is chosen as the minor vector of the molecule. The signs of these axes are determined by drawing the vector from the first to the last atom of the molecule and calculation of its dot products with principal axis. The principal axis for which the dot product with this vector is positive are chosen to be main and minor vectors. By default those atoms are literally chosen as first to last heavy atoms provided in template file, but also can be manually defined by user tailored for particular system of interest. The main vector is aligned to the z Cartesian axis and minor vector aligned to the x Cartesian axis - this orientation is considered the “initial” orientation for a particular molecule. All other orientations of the molecule are treated with respect to its “initial” orientation. The representation that is stored as an internal degree of freedom has a human-readable notation similar to quaternions: it is composed of the main vector of the molecule and the angle through which one would have to rotate the molecule around

this axis in order to put the molecule in the "initial" orientation, with the main vector aligned with the z-axis. This also allows for a discretization of the space of orientations. There are three principal axis that are obtained for each molecule and only two of them are needed to identify the "initial" configuration of the molecule.

5.3.3 Self-assembly generation with respect to fixed frames

Fixed frames, with respect to which the sampling of the configurational space is performed, can be of any form i.e., atoms, molecules, 2D periodic structures and 3D cavities. After some unique configuration of the molecule is generated, the distances between all the atoms of the molecule and the fixed surrounding are calculated and, if all of them exceed a certain value (no overlaps found) that can be specified before the search, the structure can proceed to geometry optimization. When dealing with periodic structures with particular PBC one has to take care of potential clashes of the molecule with its periodic images. Using the minimum image convention, all the atoms are mapped inside the unit cell, and checked for clashes, which in the case of a single molecule is also done with the creation of the connectivity matrix that is constructed taking into account PBC.

Having specified the template molecule and the fixed surroundings, one can set the number of molecules that should be produced in the unit cell. GenSec will then produce molecules in an iterative way and assign to them specified values for internal degrees of freedom, that can be the *same* or *different*. For example, one can sample molecules with the same conformations but having different orientations, or with the same overall orientation (for example flat lying) but with different conformations. This allows us to impose some constraints on the generated structures. In the case of generating multiple molecules, the distance between atoms of the molecules can be specified according to the goals of the search.

Examples of self-assembled structures obtained with GenSec for F6-TCNNQ/MoS_2 with 2 molecules in a (4x8) MoS_2 supercell were used for investigation of the temperature-dependent electronic ground-state charge transfer in vdW heterostructures [302] and can be found in Fig. 5.3. GenSec automates routine tasks and does not require using any FFs for the generation of self-assemblies.

5.3.4 Constraints of the search

Without imposing constraints, the number of configurations to sample is too large. For real life applications specific orientations and positions of molecules with respect to specified surroundings have to be targeted. The allowed COM space in GenSec can be specified by a range of points in the x, y, and z directions. For each direction, one can specify the boundaries and number of points that lie within those boundaries. For example, in order to generate all the structures that lie in the same 2D plane, the boundary for the direction perpendicular to this plane must contain only one particular value, which is very useful for modelling planar assemblies on the surfaces.

In the case of orientations, the discretization is performed on the angle of self-rotation. This

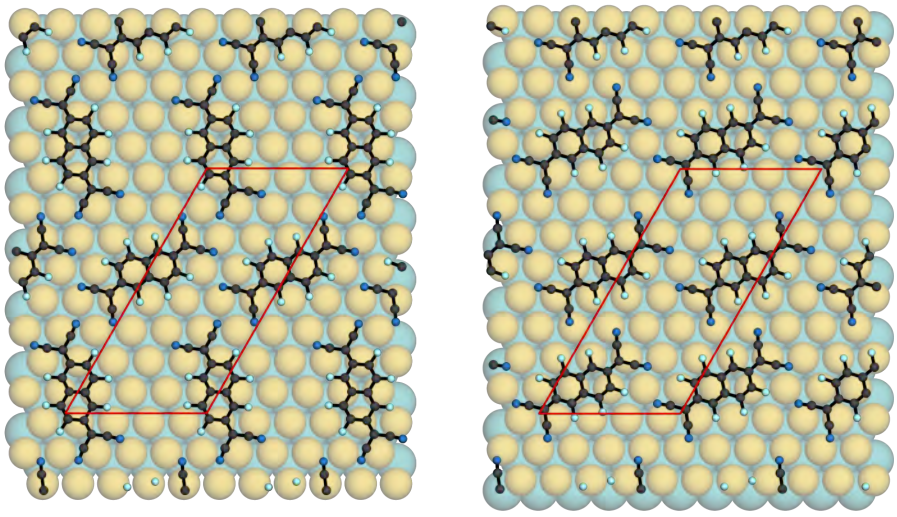


Figure 5.3 – Examples of self-assembled structures obtained with GenSec for F6-TCNNQ/MoS2 with 2 molecules in a (4x8) MoS_2 supercell.

is quantified by specifying the allowed angle of rotation. For example, if the number equals 60, six rotations of the molecule will be generated, and if the number is 360 self-rotations are basically forbidden. In this case the vector, associated with the principal axes corresponding to the lowest eigenvalue of the moment inertia tensor will solely identify the orientations. The main vector of the molecule is sampled from a uniform distribution between specified maximum and minimum values for q_1, q_2 and q_3. An example of different orientations and their notations are reflected in Fig. 5.4.

Having set these routines, one can produce an arbitrary amount of molecules per unit cell with specified orientations and conformations, that will be clash-free structures ready for geometry optimization. However, before geometry optimization, which can be very time consuming, we check the generated configurations against the database, and only if the configuration is unique, is a geometry optimization performed.

5.4 Database creation and filtering of the structures

Here we describe how the uniqueness of a randomly generated structure is checked. The database is created in the SQLite3 format, which is a self-contained, server-less, zero-configuration database. Every row in the database contains atom positions and calculated forces on all atoms together with internal degrees of freedom that represent the system. The internal degrees of freedom are stored in the database separately with the notation “t” for torsion angle numbers which are automatically identified, “q” for orientation, that has 4 values for each molecule and “c” for COM, that has three values that are defined with respect to the Cartesian origin. For a given configuration, one can easily create a query that will extract all the configurations from the database with the same corresponding torsion angles values within a given threshold. If the number of filtered structures is more than one,

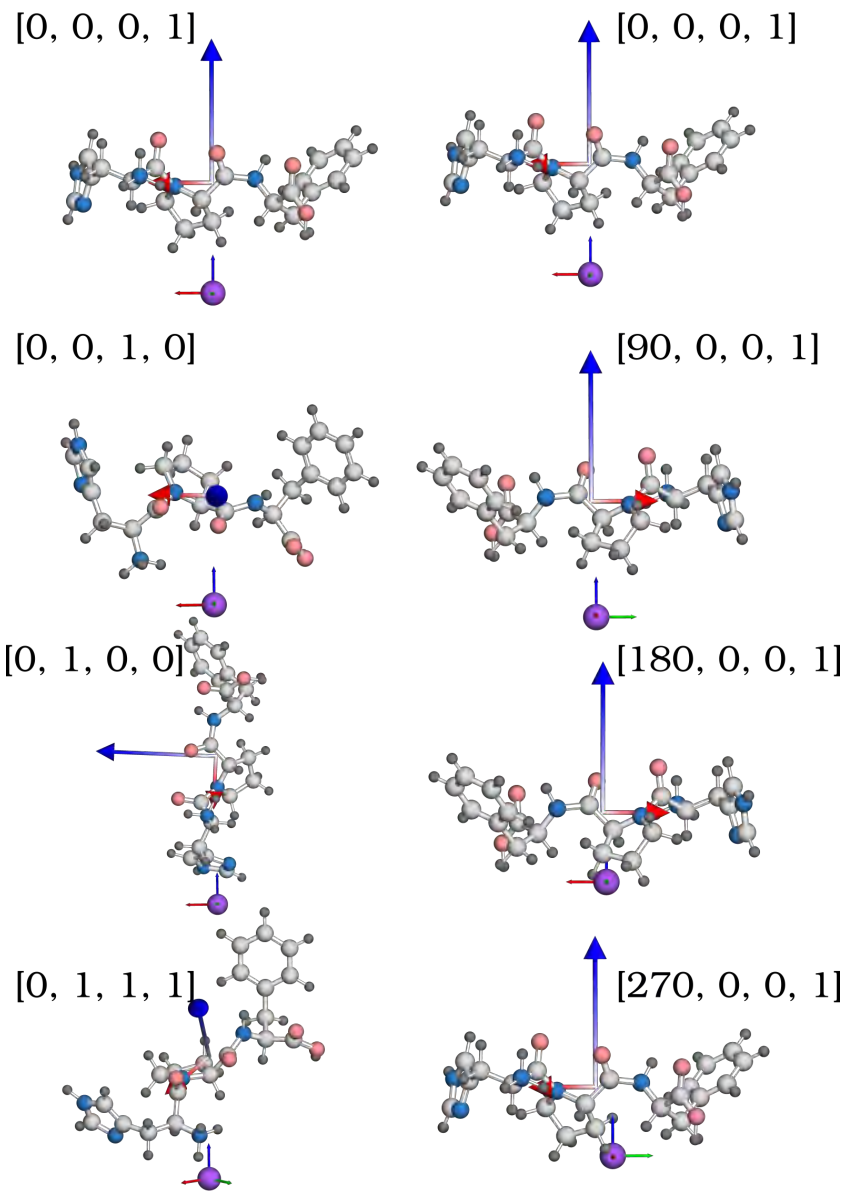


Figure 5.4 – Examples of the orientations for two different conformers. Big blue vector denotes main direction, smaller red vector denotes minor direction. Magenta circle is a Na atom from which one can see three small vectors: red - x-axis, green - y-axis and blue - z-axis. First number in brackets denotes a "self-rotation" around main vector with respect to the "initial" orientation and three other number represent direction of the main vector.

the initially generated structure is not unique and should not be further optimized. This procedure easily extends to multiple molecules. If the number of structures is more than one, and if checks on the orientations and COMs are specified, then each filtered structure will

be compared to the structure under trial. If the distance between COMs of structures within one system is more than specified value (default is 0.5 Å), the structures will be considered as different. For the orientations, the self-rotation and the angle between the main vectors of the molecules are checked separately. If both the difference between self-rotations and the angle between main vectors are greater than specified values (the default values are 30° in both cases), those structures will be considered different. If the generated structure is unique, it will proceed for geometry optimization, deleted from the database of generated structures. After relaxation, the trajectory will be added to the database of trajectories, and the local minimum will be added to the database of local minima.

We also implemented restarting procedure that is very important in the workflow of the GenSec, since it provides seamless way of continuing of the unfinished processes and continuing of the database generation especially when multiple parallel processes utilized for structure search.

5.5 Geometry optimization workflow

One of the strengths of GenSec is that it straightforwardly interfaces with the ASE environment, which allows us to perform energy and force evaluations using the most popular electronic structure packages, as well as empirical potential codes. These packages can be used to obtain energies and forces of the system at each step of geometry optimization to find local minima. The structures from every step of these geometry optimizations are stored in the database that helps to find the new unique trial structure more efficiently and provide more data for training of potentially cheaper potentials. The limitations on the size of the database is limited by capabilities of SQL

The bottleneck of exhaustive searches is *ab initio* geometry optimizations that can be sped up with the use of preconditioning of geometry optimization algorithms. There are some routines already available for geometry optimization in ASE. However, in the following, we describe the preconditioning of the BFGS algorithm that takes into account an approximate Hessian matrix that contains information about connectivity and physical interactions in the system. This allows the algorithm to make a better choice for the next step towards finding the local minima. This implementation is tailored explicitly for interface systems, and its description and performance will be described in the following section.

5.6 Preconditioner for geometry optimization

Having routines for sampling different parts of the conformational space of a system, it is necessary to minimize the system's energy. It was shown that the energy hierarchy of the structures for which only single-point calculations were performed could change dramatically after their geometry optimization [303]. The most popular geometry optimization algorithms are quasi-Newton algorithms that require input information about the energy and forces of a configuration and iteratively find the local minima of the system. Based on the forces and energies of adjacent steps, the algorithm updates the approximate Hessian

matrix. One of the most successful schemes is the BFGS algorithm, which was described in Section 3.2. However, the potential energy surface of the system can be highly anisotropic, which results in a poor performance (slow convergence) of the geometry optimization. In order to make the shape of the potential more isotropic, one can use preconditioners that perform a metric transformation of the coordinate system, thus making the shape of the potential energy surface smoother and improving the efficiency of finding the nearest local minima.

By default, the initial Hessian matrix is a scaled identity matrix, and initializing the Hessian matrix with some information about the system can improve the speed of convergence of the geometry optimization algorithm. A combination of the Hessian matrix with different preconditioning schemes showed a performance gain when applied to molecular crystals [304], for example. For modelling condensed phase systems, the best performance is demonstrated using the Exponential preconditioner [249]. For modelling gas-phase molecular systems, the force-field-like preconditioner proposed by Lindh et al. [236] is widely used due to its simplicity. Specifically for the interfaces, we propose a scheme that allows us to combine these different approximations and apply them to the corresponding parts of the system, i.e. Lindh applied to the molecular part and Exponential to the solid part. We also introduce a vdW part that allows us to calculate a LJ Hessian matrix based on the vdW parameters developed in the TS-vdW method that can be applied to the parts of the Hessian where it can play an important role. The pictorial representation of the proposed scheme can be found in Fig. 5.5. First, we describe the workflow of the LJ preconditioning scheme and then show some results for model systems where the combined preconditioning scheme was applied.

5.6.1 Lennard-Jones-like Hessian matrix

Here we would like to introduce preconditioning scheme that could treat vdW bonded systems. First, we introduce notations used in the scheme:

$$\begin{aligned} &A, B \in \{0, .., N-1\}, A \neq B - \text{interacting atoms,} \\ &i, j \in \{0, 1, 2\} - \text{cartesian axes,} \\ &\delta_{ij} \text{ is Kronecker delta.} \end{aligned} \tag{5.3}$$

We derive the Hessian starting from a Lennard-Jones 12-6 potential:

$$E_{AB}^{\text{LJ}} = 4\varepsilon\left[\left(\frac{\sigma}{R^{AB}}\right)^{12} - \left(\frac{\sigma}{R^{AB}}\right)^{6}\right] = \frac{C_{12}^{AB}}{(R^{AB})^{12}} - \frac{C_{6}^{AB}}{(R^{AB})^{6}}, \tag{5.4}$$

where indices A and B denote different atoms,

$$R^{AB} = ((x_1^B - x_1^A), (x_2^B - x_2^A), (x_3^B - x_3^A)) \tag{5.5}$$

$$R_i^{AB} = x_i^B - x_i^A \tag{5.6}$$

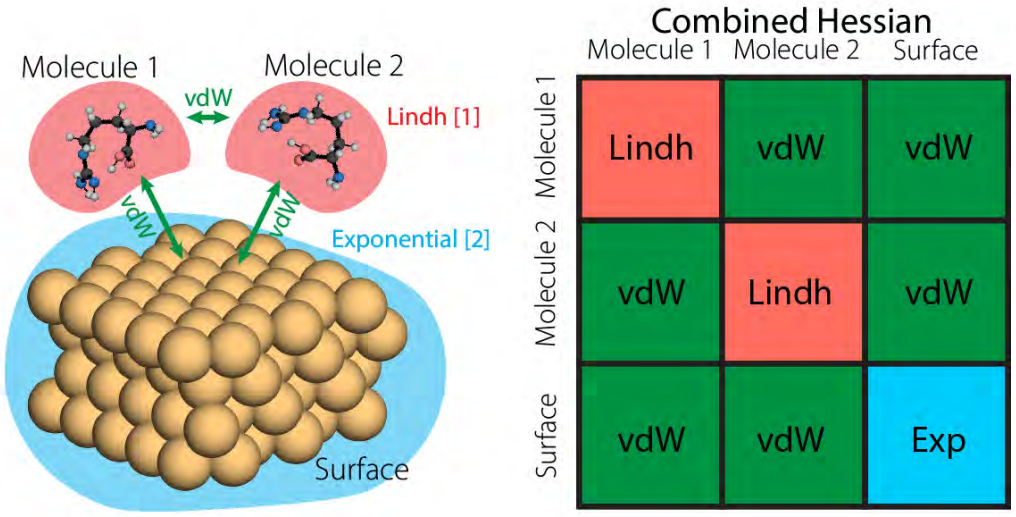


Figure 5.5 – Representation of the construction of the approximated Hessian matrix using different preconditioning schemes a) Representation of the different parts of the system for which different preconditioning schemes can be applied separately; b) the combined approximated Hessian matrix constructed using different preconditioner schemes applied for different parts of the system.

is the distance between atoms A and B and

$$|R^{AB}| = R = \sqrt{(x_1^B - x_1^A)^2 + (x_2^B - x_2^A)^2 + (x_3^B - x_3^A)^2}. \tag{5.7}$$

The C_6 coefficients are taken from [305, 306]. So we proceed to take the first derivative:

$$\frac{dE_{AB}^{\mathrm{LJ}}}{dx^A} = \frac{6C_6}{R^8} R^{AB} - \frac{12C_{12}}{R^{14}} R^{AB}. \tag{5.8}$$

By assuming that LJ potential adopts a minimum at $R_0^{AB} = \frac{R_{\mathrm{vdW}}^A + R_{\mathrm{vdW}}^B}{2}$, one can derive C_{12} from Eq. 5.8 as

$$C_{12}^{AB} = \frac{1}{2} C_6^{AB} * (R_0^{AB})^6, \tag{5.9}$$

as discussed in [163].

After that we take the second derivative and get:

$$\frac{\partial E_{AB}^{\mathrm{LJ}}}{\partial x_i^A \partial x_j^B} = \frac{\partial \left(\frac{6C_6}{R^8}\right)}{\partial x_j^B} R_i^{AB} + \frac{6C_6}{R^8} \frac{\partial R_i^{AB}}{\partial x_j^B} - \frac{\partial \left(\frac{12C_{12}}{R^{14}}\right)}{\partial x_j^B} R_i^{AB} - \frac{12C_{12}}{R^{14}} \frac{\partial R_i^{AB}}{\partial x_j^B} =$$

$$= \frac{48C_6 R_i^{AB} R_j^{AB}}{R^{10}} - \frac{6C_6}{R^8}\delta_{ij} - \frac{168C_{12} R_i^{AB} R_j^{AB}}{R^{16}} + \frac{12C_{12}}{R^{14}}\delta_{ij}. \tag{5.10}$$

After simplification the LJ Hessian will be:

$$H^{\mathrm{LJ}}_{(3A+i),(3B+j)} = \frac{48C_6(x_j^B - x_j^A)}{R^{10}}(x_i^B - x_i^A) - \frac{168C_{12}(x_j^B - x_j^A)}{R^{16}}(x_i^B - x_i^A) - (\frac{6C_6}{R^8} - \frac{12C_{12}}{R^{14}})\delta_{ij}. \tag{5.11}$$

However, in practical simulation, we want to employ preconditioning scheme at situation that may be far from the ideal minimum of such a potential. In that case this constructed Hessian will not be positive definite. To overcome this issue we apply the strategy as in the Lindh approach for constructing the Hessian matrix, where the Hessian is estimated for a particular configuration as it would be if that configuration was a minimum [236]:

$$E(\mathbf{R}) = E\left(R_1^0, \cdots R_N^0\right) + \sum_{i=1}^{N} \frac{\partial E}{\partial R_i}\bigg|_{R_i = R_i^0} \left(R_i - R_i^0\right) + \frac{1}{2}\sum_{i,j=1}^{N}\left(R_i - R_i^0\right)\frac{\partial^2 E}{\partial R_i \partial R_j}\left(R_j - R_j^0\right) + \cdots \tag{5.12}$$

where the second term cancels to zero. Our model Hessian is then

$$H^{\mathrm{LJ}}_{(3A+i),(3B+j)} = \frac{\partial^2 E(R_0^{AB})}{\partial R_i^A \partial R_j^B}|_{R_i^{AB}=R_{0i}^{AB}, R_j^{AB}=R_{0j}^{AB}} = \frac{48C_6 R_{0i}^{AB} R_{0j}^{AB}}{(R_0^{AB})^{10}} - \frac{6C_6}{(R_0^{AB})^8}\delta_{ij} - \frac{168C_{12} R_{0i}^{AB} R_{0j}^{AB}}{(R_0^{AB})^{16}} + \frac{12C_{12}}{(R_0^{AB})^{14}}\delta_{ij} \tag{5.13}$$

Obviously, values R_i^{AB} can be far from equilibrium values and this will lead to Hessian matrix be not positive definite. Instead, the R_i^{AB} is scaled to the length of R_{0i}^{AB} in order to satisfy the assumption that the system is near the local minimum:

$$R_{0i}^{AB} = R_i^{AB} * \frac{|R_0^{AB}|}{R} \tag{5.14}$$

With use of prefactor coefficient ρ_{AB} for the whole Hessian matrix we set the vdW interaction at the distances larger than $2 \times R_0^{AB}$ to be negligible, basically setting preconditioning only for nearest neighbour atoms:

$$\rho_{\mathrm{AB}} = \exp\left[\alpha_{\mathrm{AB}}\left((R_0^{\mathrm{AB}})^2 - R^2\right)\right], \tag{5.15}$$

by fitting of the the parameters α_{AB} for each pair of R_0^{AB}. Finally we get

$$H^{\mathrm{LJ}}_{(3A+i),(3B+j)} = \rho_{\mathrm{AB}} \frac{\partial^2 E(R_0^{AB})}{\partial R_i^A \partial R_j^B}|_{R_i^{AB}=R_{0i}^{AB}, R_j^{AB}=R_{0j}^{AB}} \tag{5.16}$$

This scheme is implemented in GenSec and available with use of the flag “vdW” for preconditioning of the geometry optimization. The scheme was tested on model LJ Ar_n clusters,

where n reflects the number of atoms in the cluster, the local minima of which were taken from the database [307]. For all the minima random displacements of 0.01 Å were applied for each atom. The BFGS TRM method was used for geometry optimization. The geometry optimizations were carried out with the vdW preconditioning scheme and with the scaled identity matrix using 70 as the scaling factor (default in ASE) as initial Hessian (which is also will be noted as unpreconditioned case). The performance gain is calculated as the number of steps required to reach the local minima for unpreconditioned case divided by the number of steps required to reach the same local minima with use of initial preconditioned vdW Hessian matrix, and shown as a function of cluster size in Fig. 5.6.

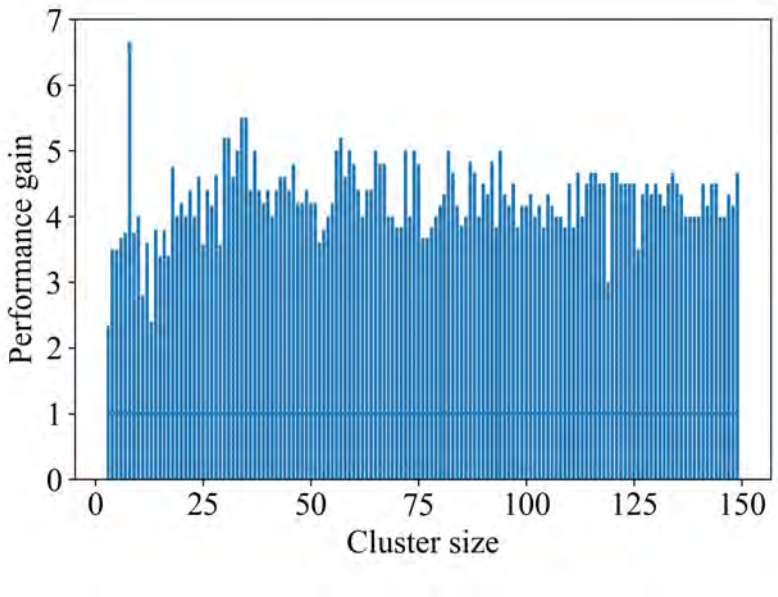


Figure 5.6 – Performance gain for the geometry optimization of LJ clusters of different sizes using vdW preconditioning scheme, compared to the unpreconditioned case.

Identical structures are obtained with and without application of the preconditioner, and our preconditioning scheme shows significant performance gains for these systems, where the only force acting on the atoms is LJ force. Now we proceed to the combination of the different Hessian schemes, and apply the combined Hessians to model interface systems.

Next we adapted the Exponential and Lindh preconditioning schemes described in Sec. 3.2.4 into the workflow of GenSec. To test the performance of the Exponential preconditioner, we optimized bulk $N \times N \times N$ fcc Cu unit cells were optimized using Effective Medium Theory (EMT) potential implemented in ASE [308], and to test the Lindh preconditioning scheme we used PBE with light settings implemented in FHI-aims to relax different Alanine dipeptide conformers obtained with GenSec. The results are shown in Fig. 5.7 - in both cases a significant performance gain is observed.

Randomly generated geometries can be far away from any local minima. Especially for

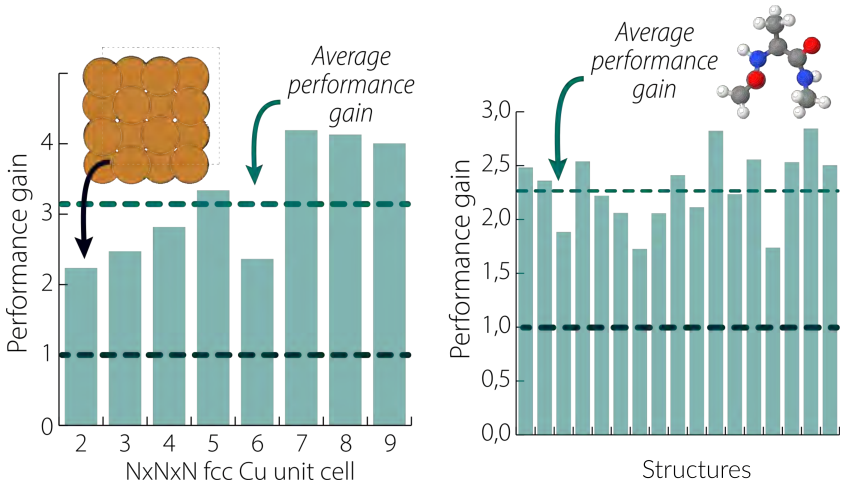


Figure 5.7 – Performance gain for geometry optimization with Exponential preconditioning scheme applied to Cu bulk systems (left) and performance gain of the Lindh preconditioning scheme applied to geometry optimization of different conformers of Alanine dipeptide structures (right).

flexible molecular systems, the local environments can change dramatically during geometry optimization due to torsional rotations. In this case, the local PES cannot be approximated quadratically. To overcome this issue, one of the approaches could be to restart the BFGS procedure and reinitialize the Hessian matrix during the geometry optimization. One way to do this is to "reset" Hessian matrix after some fixed number of steps. By contrast, we restart and update the Hessian matrix depending on the change of the root mean square displacement (RMSD) value between snapshots in the geometry optimization trajectory:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \mathrm{d}_i^2}, \tag{5.17}$$

where d_i is the distance between the atomic positions. Randomly created flexible molecules are usually far away from a local minimum which means that harmonic approximation of quasi-Newton procedure that was initially made will not be valid after several optimization steps and reinitialization of the Hessian matrix allows the BFGS algorithm to find local minima faster. For the same set of conformers of Alanine dipeptide presented in Fig. 5.7 we applied this scheme, where the Hessian matrix was reinitialized after the RMSD exceeded the specified value. The definition of the RMSD value is system specific and should be chosen with caution in order to obtain the best performance results - choosing the value to be too small will reinitialize the Hessian update too often, which could lead to decrease of performance of BFGS algorithm. Harmonic approximation could be valid if the atom displacements are within 0.2 Å from their equilibrium positions [234]. The results in Fig. 5.8 show that this strategy can be twice as efficient compared to the case where preconditioning was applied only at the initialization step.

5.6.2 Combining the preconditioners

Having all of the preconditioning schemes implemented in GenSec, we created the model system of one hexane molecule adsorbed on Rh surface to test the performance of the combined preconditioner illustrated in Fig. 5.9. The system can be clearly separated into molecular and surface parts, and the strategy for applying the different preconditioning schemes is the following: the constructed initial Hessian can be obtained for the whole system using Exponential or Lindh. One can apply different preconditioning schemes to different parts, i.e, Exponential for the substrate part and Lindh for the molecular part. For the Hessian matrix elements that correspond to off-block-diagonal elements, that do not correspond solely to molecular or substrate part, one can apply the vdW preconditioning scheme, or simply set those elements to 0.

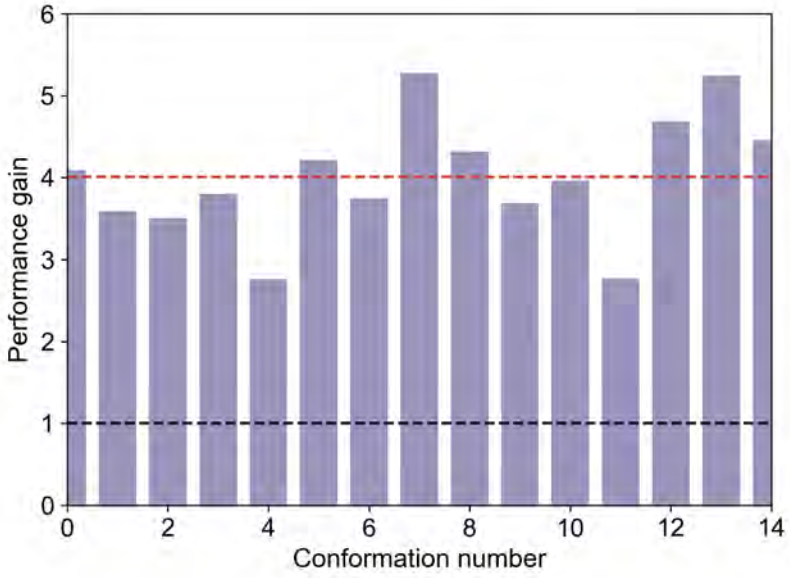


Figure 5.8 – Performance gain for geometry optimization of different randomly generated conformers of Alanine dipeptide with reinitialization of the Hessian after the conformational change exceed 0.1 Å .

$$H_{(3A+i),(3B+j)} = \begin{cases} \text{Lindh term,} & \text{if A, B are in molecule} \\ \text{vdW or 0} & \text{if A, B belong to different parts of the system} \\ \text{Exponential} & \text{if A, B are in surface} \end{cases} \tag{5.18}$$

For the model system the effects of applying the different preconditioning schemes are shown in Fig. 5.9. The PES was constructed using adaptive intermolecular reactive bond order (AIREBO) potentials [309, 310] for carbohydrates, embedded atom model (EAM) interatomic potential for Rh atoms [311, 312] and LJ potential for interactions between molecule and surface. It is clear that applying a combined preconditioner is more efficient than applying a

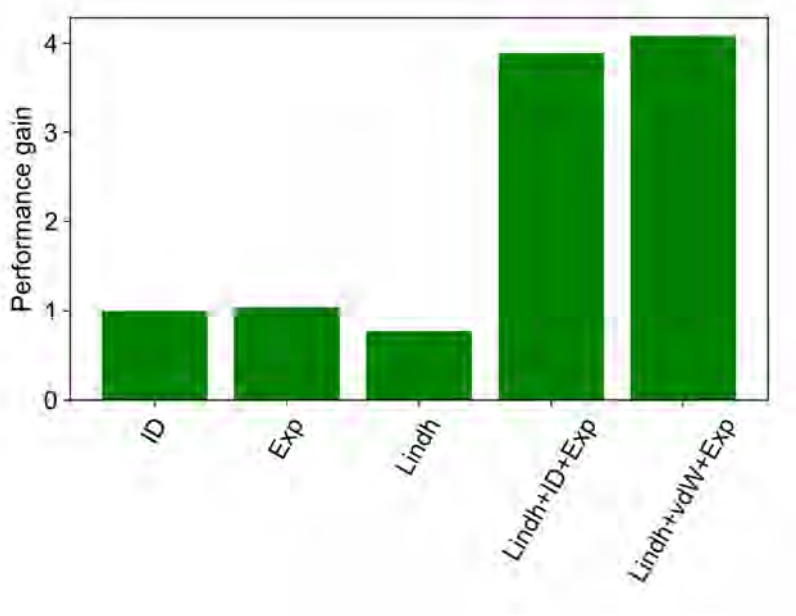


Figure 5.9 – Performance gain for geometry optimization with different preconditioning scheme applied to geometry optimization of hexane on Rh surface.

single preconditioning scheme to the whole system. Inclusion of the vdW preconditioning scheme doesn't give a significant performance gain in this case. This is likely because the vdW forces are never the largest forces in the optimization path. Nevertheless, this strategy can be efficient, and applying it to the broader range of systems with different potentials will be the scope of future investigations.

The package is open-source and ready for usage. Tutorials and documentation can be found at https://github.com/sabia-group/GenSec

5.7 Application to di-L-alanine on Cu(110)

Having presented the GenSec package, we now provide an example of how it can be applied to a system that has been previously investigated experimentally, namely di-L-alanine adsorbed on the Cu(110) surface. STM was utilized to investigate the sub-monolayer formation of this peptide, which is the smallest possible chiral peptide consisting of two AAs (L-alanine), on Cu (110) [313]. At low coverages, these molecules nucleate along the [$\bar{3}$32] direction, forming small, predominantly one-dimensional islands. Coverage increase results in forming elongated, [$\bar{3}$32]-directed islands. At higher coverages, up to one monolayer, the islands merge to form phase barriers across domains with opposite orientations. In Fig. 5.10 and Fig. 5.11, we reproduce the experimental STM images from Ref.[313].

We investigated the adsorption of di-L-alanine on Cu(110) at DFT level of theory. In order to compare experimental and theoretical results we proceeded with comparing of the STM images obtained for the lowest energy structures obtained during structure search. We

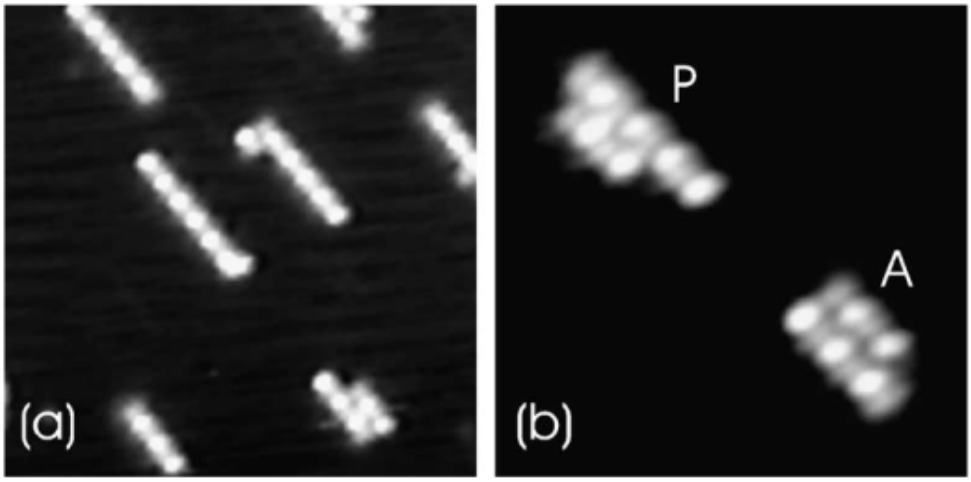


Figure 5.10 – Two STM images of di-L-alanine on Cu(110) at low coverage. The molecules were evaporated at a sample temperature of 248 K and scanning took place at 208 K to freeze out diffuzion: (a) 160 Å × 160 Å , $V_1 = -2.10$ V, $I_1 = -0.34$ nA. (b) Two islands with parallel (P) or anti-parallel (A) di-L-alanine molecules in adjacent rows: 90 Å × 90 Å , $V_1 = -1.68$ V, $I_1 = -0.34$ nA. Reprinted from Surface Science, Volume 545, Issues 1–2, Ivan Stensgaard, Adsorption of di-L-alanine on Cu(110) investigated with scanning tunneling microscopy, Pages L747-L752, Copyright 2003, with permission from Elsevier.

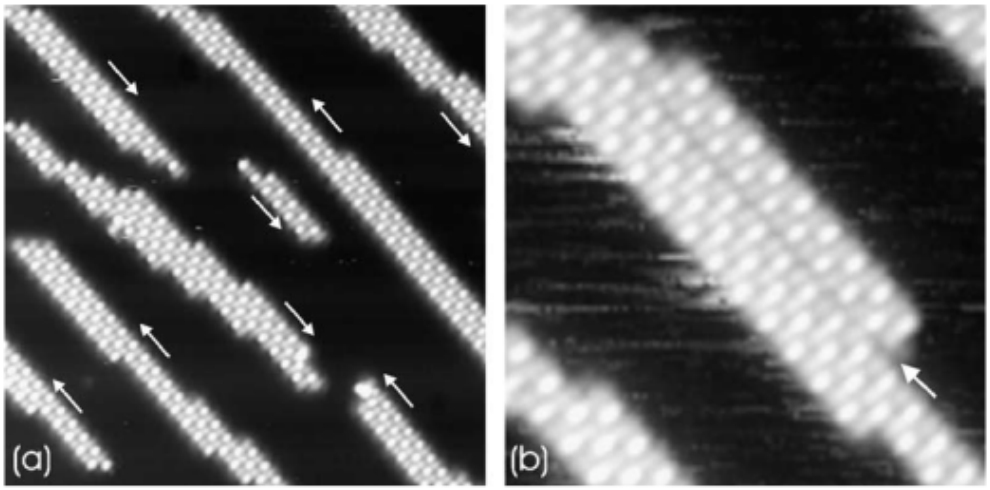


Figure 5.11 – (a) STM image of di-L-alanine on Cu(110). All molecules in an island are oriented parallel of antiparallel to the $[\bar{3}32]$ direction as indicated by the two directions of the arrows. The di-L-alanine was evaporated at a sample temperature of 363 K and imaged at 198 K. Area: 250 Å × 250 Å , $V_1 = -1.25$ V, $I_1 = -0.65$ nA. (b) Formation of a domain boundary (marked with an arrow) between two antiparallel domains. Adsorption temperature: 363 K, imaged at 268 K, 100 Å × 100 Å , $V_1 = -1.68$ V, $I_1 = -1.52$ nA. Reprinted from Surface Science, Volume 545, Issues 1–2, Ivan Stensgaard, Adsorption of di-L-alanine on Cu(110) investigated with scanning tunneling microscopy, Pages L747-L752, Copyright 2003, with permission from Elsevier.

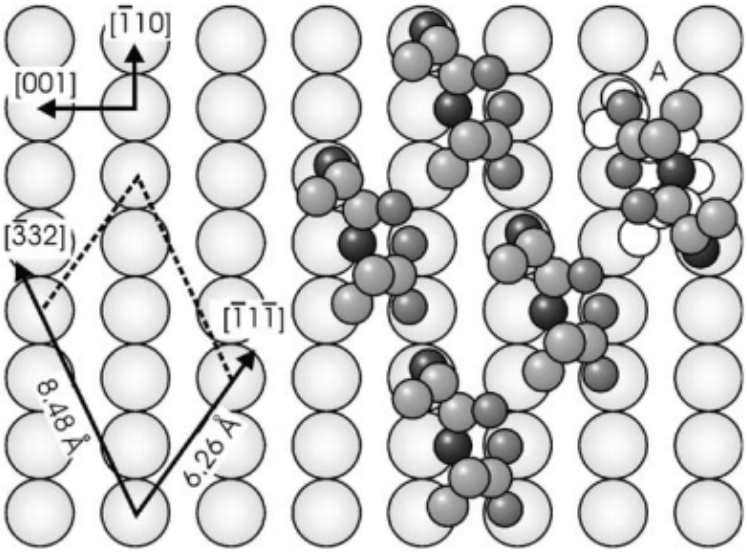


Figure 5.12 – Schematic model of the di-L-alanine surface layer on a Cu(110) substrate. The size and orientation of the unit cell is indicated. The atoms of the molecules are shown in shades of grey going from N (darkest) via O to C (lightest). Hydrogen atoms are left out. The molecule marked A in the upper right corner has been rotated by 180° and shifted slightly to adopt the same local adsorption geometry as the unrotated molecules. The position of the molecule before rotation is shown as an outline. Reprinted from Surface Science, Volume 545, Issues 1–2, Ivan Stensgaard, Adsorption of di-L-alanine on Cu(110) investigated with scanning tunneling microscopy, Pages L747-L752, Copyright 2003, with permission from Elsevier.

analyzed the characteristics of the structures found by the random search, which one seems to be the experimental structure, and how it compares with the structure originally proposed in Ref. [313] and can be found in Fig 5.12(a).

5.7.1 Computational details

The electronic structure calculations were carried out using the numeric atom-centered orbital all-electron code FHI-aims [183, 184]. We used the standard *light* settings of FHI-aims for all species. For modeling the adsorbed molecules, a surface $1\times1\times2$ unit cell with $6\times6\times1$ k-point sampling was employed. The fcc(110) copper slab was produced using ASE package with lattice vectors directions [$\bar{3}$32], [$\bar{1}$1$\bar{1}$] and [110] that resulted in 4 layers in the slab with parameters a = 8.52 Å and b = 6.29 Å compared to experimental 8.48 Å and 6.29 Å lattice vectors lengths in [$\bar{3}$32], [$\bar{1}$1$\bar{1}$] respectively. The lattice parameter employed was 3.63 Å as in our previous works [83]. In order to isolate periodic images we added a 100 Å vacuum in the z direction and also employed the dipole correction. We employed the PBE+vdW^{surf} functional [130] which contains an effective screening of the vdW interactions optimized for metallic surfaces. The two bottom layers of the surface was constrained and a geometry optimization was performed until all forces in the system were below 0.01 eV/Å .

STM images were produced with Tersoff-Hamman approximation [197] with modelled applied voltage of -2 eV. This voltage was chosen based on the experimental values of the applied voltage for STM picture recording.

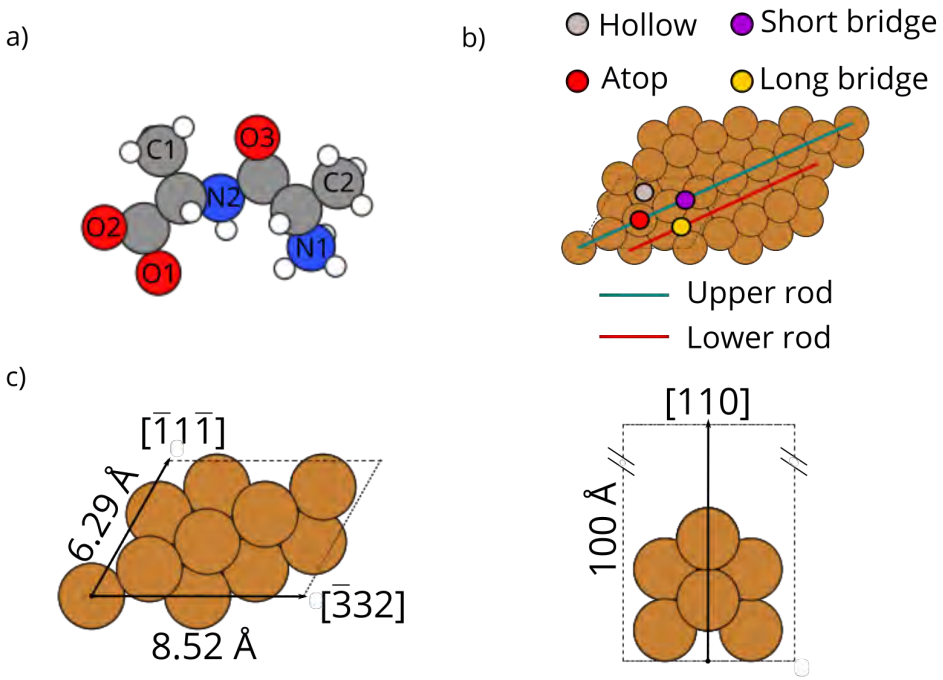


Figure 5.13 – a) Schematic representation of the di-L-alanine amino acid in its zwitterionic configuration. Red atoms are oxygen; blue atoms are nitrogen; white atoms are hydrogen, and grey atoms are carbon. b-d) Schematic representation of Cu(110).

An example of the di-L-alanine molecule and of the Cu(110) unit cell surface that we used for structure search can be found in Fig. 5.13.

5.7.2 Generation of trial structures

Trial structures were randomly produced using GenSec package with one molecule per unit cell. From the experimental study, we learned that we could apply a few constraints in the search. We restrict trial structures to be extended along the [3̄32] direction. The structures were generated in zwitterionic state since Fig. 5.10(a) shows evidence that the molecules within a single-row island are aligned in the same direction at low coverage. This evidence points to a model in which the terminal carboxylic group of one molecule forms a hydrogen bond with the terminal amino group of another molecule. The zwitterionic character of alanine in its solid-state [68], would be a good match for this type of relationship. However, it is impossible to rule out the possibility of deprotonation during the adsorption process, which would result in the formation of an anionic molecule. Investigations of tri-L-alanine for low coverage adsorption on Cu(110) revealed that the AA was bonding in the anionic form [314].

Ten searches were conducted in parallel, sharing the databases that blacklist trial candidates and geometry optimization trajectories obtained from different searches. Machinery implemented in GenSec allowed to perform such a structure search in a high-performance computing infrastructure by utilizing SQLite3 database features in ASE. After sampling of 500 structures we stop the structure search and select all the structures that fall within 1 eV energy range relative to the lowest energy structure and proceed to analysis of the results.

5.7.3 Analysis of the search

The structure that was proposed in Ref. [313] would bind with O1 and O2 oxygen atoms at the atop position to the same upper rod of the Cu(110) surface and atoms C1, N1 and N2 would adsorb also at atop positions on the neighbouring upper rod. Oxygen atom O3 and C2 from methyl group should not be connected to the surface. We manually prepared this structure and performed geometry optimization. This structure is depicted in Fig. 5.12 together with its STM image. As one can see, the patterns on STM images recorded experimentally and theoretically produced do no match: there are no interweaving bright and weak spots and their connectivity between neighbouring strands is absent.

After we performed structure search only 23 unique structures in our database fall within 1 eV from the lowest energy structure. The structures either remain in the zwitterionic state or undergo a deprotonation and adopt an anionic state. For all the 23 structures, we modelled STM images and created repeated images for easier visual comparison. One can clearly see that the patterns can differ considerably from each other. The particular pattern observed in experiment (interweaving of bright and faded spots along a strand, with connections between strands that reminds of a tadpole) is very similar to the ones obtained for structure 7 (Fig. 5.14). All the lowest energy structures together with their STM images can be found in Appendix B.1-B.5 and we proceed to more detailed analysis of the eight lowest energy structures found during the search (Fig. 5.14). The exact structure that was proposed in Ref. [313] was not found during structure search. We prepared this structure manually and performed geometry optimization for it, which results in the structure 8 (Fig. 5.14) but higher by 30 mEv in energy from it due to slightly different adsorption pattern (Fig. 5.15). During geometry optimization C1 atom does not bind to the surface and points towards the vacuum region and thus, we can conclude that structure originally proposed in Ref. [313] is not stable.

The structures denoted 1,2,3 and 17 undergo deprotonation of the molecule adsorbed on the surface during geometry optimization. Most of the structures bind to the surface with at least one oxygen atom from carboxyl group and amino group attached to the different rods. The second oxygen atom from carboxyl group can be attached to the same rod as the first oxygen atom (structures 1, 2, 4, 5, 6, 7, 8, 9, 11, 13, 16, 17, 23), to the same rod as amino group (structures 3, 12, 14, 15, 19) or not attached to the surface. Almost in all cases the interstrand connectivity is done via carboxyl and amino groups, except for the structures 3 and 17. Only in case of structure 22 both methyl groups are parallel to the surface - in this structure amino group is also not attached to the surface, in all other cases at one or both

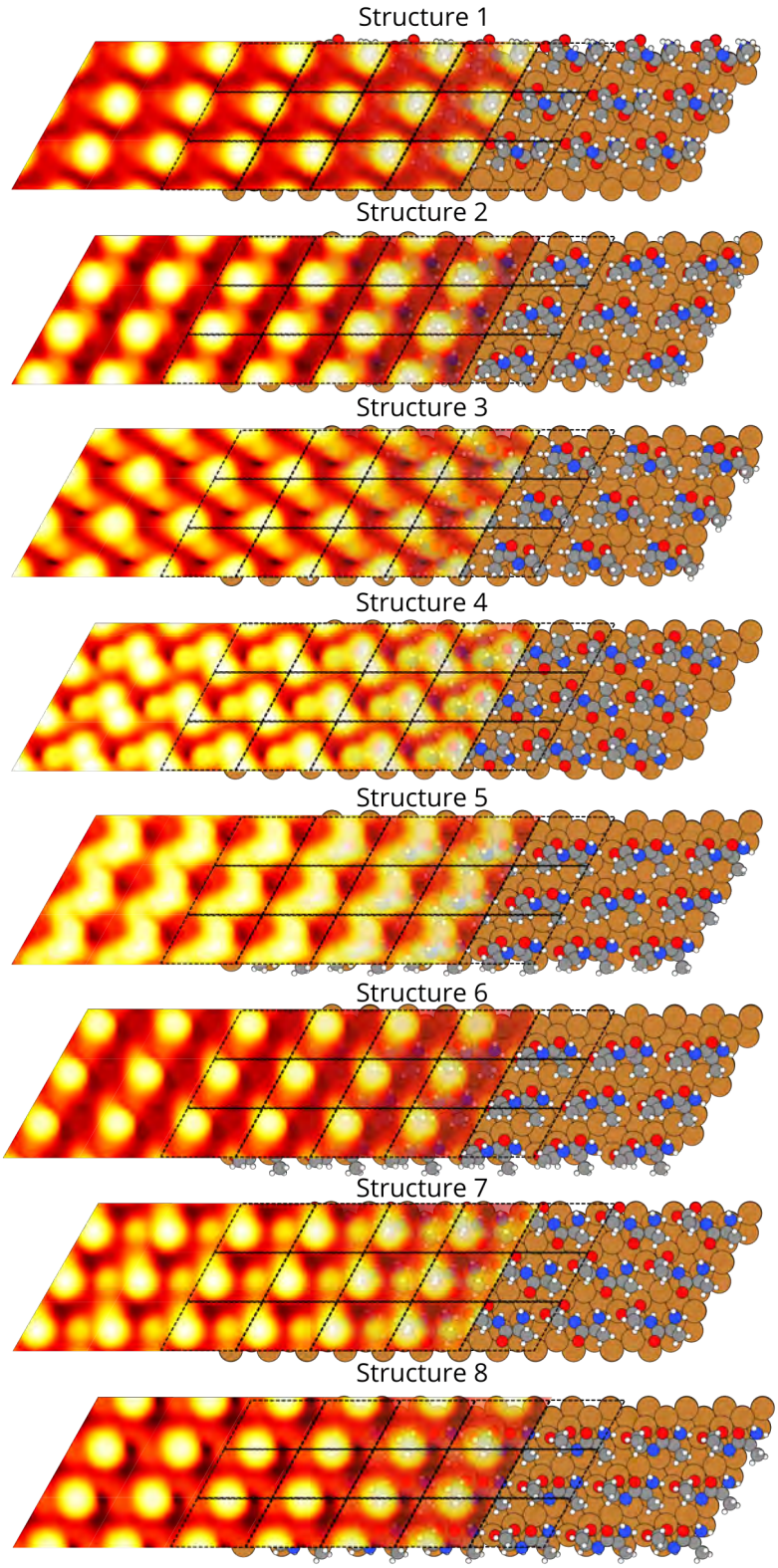


Figure 5.14 – Modelled STM images and structures 1-8 of di-L-alanine molecules adsorbed on Cu(110) surface together with unit cell represented with black dashed lines.

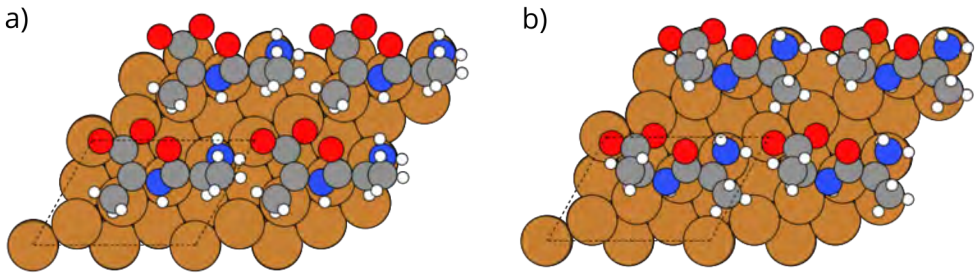


Figure 5.15 – proposed and relaxed structures.

methyl groups pointing towards vacuum. We present in Fig. 5.14 a detailed visualization of structures 1, 2, 3, 7.

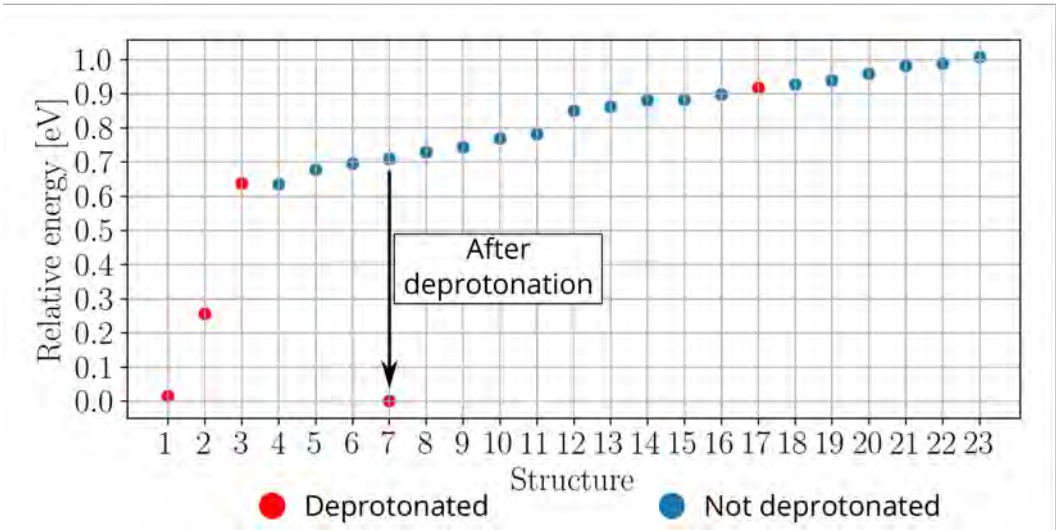


Figure 5.16 – Energy hierarchy of the obtained structures within 1 eV relative energy range.

Structure 7 binds with both O1 and O2 oxygens to the same rod of surface at atop positions and N1 atom binds to another rod which is the same binding proposed in Ref. [313]. This structure differs from the one originally proposed by Stensgaard [313] by orientation of the N2 and O3 atoms and orientation of the C1 and C2 atoms that not interact with the surface in Structure 7. Moreover, amino group interacts not only with carboxyl from the same strand, but also with O3 oxygen atom from another strand.

Since deprotonation is possible and the three lowest energy structures found in this search were deprotonated, we removed the hydrogen atom from the amino group that was pointing towards the surface and performed a new geometry optimization. The resulting structure is 10 mEv lower in energy than the lowest energy structure. The pattern from the modelled STM for this system is more pronounced and seems to agree even better with experimental one, reproduced in Fig. 5.17.

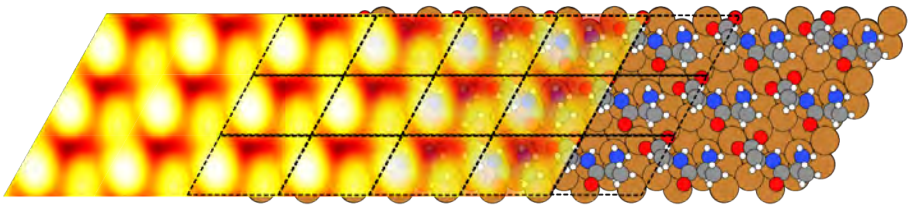


Figure 5.17 – Modelled STM image and structure of structure 7 after deprotonation together with unit cell represented with black dashed lines

Performing an analysis of the different stabilizing interactions of these self-assembled structures, shown in Appendix A.1, we find, interestingly, that the only structures for which the intrastrand and interstrand interactions are almost identical in energy are structures 7 and 11. These structures would likely fall into the exact same minimum if an optimization with a better basis set and accuracy threshold were performed.

The case of di-L-alanine on Cu(110) could be studied much further, but the results presented here show that a random search like the one performed here with GenSec can be a powerful ally of such STM experiments. Because it is based on first principles geometry optimizations, it also automatically identified the propensity for deprotonation of these molecules on such a reactive surface like Cu(110). This is an important point to consider when dealing with other types of theoretical approaches such as FF and schemes that keep molecules "rigid" or "whole".

We conclude that the best candidate theoretically predicted structure neither one that was proposed in the paper [313] nor the lowest energy structure found during the structure search. This supports the idea, that in a particular experiment the global minimum found theoretically may not always the most relevant structure. A random structure search strategy that covers the broadest possible parts of the conformational space (withing few constraints) can be quite effective.

One of the intriguing and yet not completely understood results is that the best candidate structure stands out among other lowest energy structures by having the perfect balance between intrastrand and interstrand interaction energies that could be relevant for understanding of the self assembly processes on surfaces.

The final result of comparing of simulated and experimental STM images can be found in Fig. 5.18.

5.8 Conclusions

GenSec showed satisfactory results for structure search of di-L-alanine adsorbed on Cu(110) surface in both efficient utilization of resources (multiple structure searches were carried out at the same time connected to the shared database) and for unbiased sampling of the

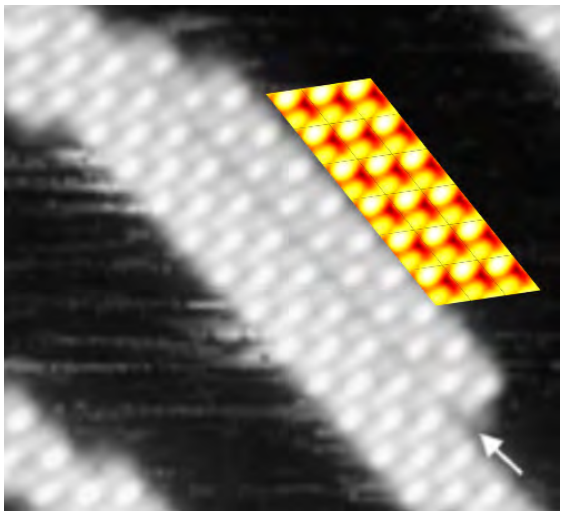


Figure 5.18 – Modelled STM image colored in oranges and experimental STM image colored in grays of di-L-alanine on Cu(110) aligned in direction of strand grow. Reprinted from Surface Science, Volume 545, Issues 1–2, Ivan Stensgaard, Adsorption of di-L-alanine on Cu(110) investigated with scanning tunneling microscopy, Pages L747-L752, Copyright 2003, with permission from Elsevier.

conformational space of flexible molecule. The created infrastructure allows easily specify constraints of the search according to the experimental input and choose from multiple electronic structure codes available to connect through ASE package. Created databases contain not only the lowest energy structure but also all the intermediate steps together with energies and forces in unified format that is convenient to share and reuse.

5.9 Outlook

The workflow of the package already allows an investigation of an arbitrary amount of adsorbates per unit cell with respect to specified surroundings. One of the package's strengths is that it can produce data in a parallel fashion, optimally utilizing available computational resources. The resulting data has general formatting independent of the electronic structure package used for structure search, making it reusable and easy to handle for further processing.

The main directions of further development of the package should be:

- Connection of the package to the ML packages that allow training cheap potentials on the fly for further exhaustive search procedure;
- Connection to the packages that allow to automatically generate low dimension representation of the conformational spaces and visualize them

[50mm] Many people asked me what would I do if I didn't finish the thesis.
We will never know it.

6 Conclusions

In this thesis, we have characterized the conformational space of the arginine amino acid in its neutral and protonated form in different non-biological environments, i.e. in isolation and in contact with metallic surfaces. In particular, we have analyzed how and why different parts of the conformational space become accessible or are excluded, depending on the protonation state and the environment, showing the importance of bond formation and charge rearrangement in these systems.

This study included the construction of a database based on thousands of structures optimized by density-functional theory including dispersion interactions. The construction of this database is a result in itself and we hope that in future systems, that are investigated will be also available for everyone. The analysis of complex systems is still far from being fully automated, and requires a human's creative approach and a tremendous amount of effort and time to be invested in the identification of structure-property relationships. Even the application of modern dimensionality reduction and visualization techniques should be considered only as a first step that can give inspiration for further analysis. Regarding the investigation of interface systems, we found, for example, that it is advantageous to start from a comprehensive sampling of the conformational space of the least-constrained molecular form, which in our case was the neutral Arg amino acid in the gas-phase. This is evidenced by the fact that in our low-dimensional projections, all low-energy conformers we observe on the surfaces for both Arg and Arg-H^+, lie among structural conformations that were already present in the gas-phase sampling of Arg, albeit often with high relative energies. This is not the general case though, and the environment can alter conformational space in an unpredictable manner - this can be seen from the sampling of Arg-H^+ structures, the flexibility of which increased after adsorption. In addition, we find that for Cu, Ag, and Au surfaces, the energy hierarchies of different conformers are largely preserved when changing the substrate.

We illustrate that while INTERFACE-FF can sample relevant areas of conformational space, it is not able to capture consistent energy hierarchies. Additionally, the molecular chemical groups show a preference to adsorb on different surface sites, which could have considerable

impact on self-assembly studies. Databases such as those we created will serve as an important source of data for further parametrization and improvement of these potentials.

Regarding the structural space of Arg and Arg-H^+ adsorbed on (111) surfaces of Cu, Ag and Au, we have learned the following: The adsorption of Arg leads to the formation of strong bonds to the surface that involve mostly the carboxyl and amino groups. This stabilizes the protomer that we label **P3** in this work, where the carboxyl group is deprotonated and the side chain is protonated. This is different to the dominant protomer in the gas phase, with the label **P1**. The bonds to the surface sterically constrain the conformations of this molecule, thus decreasing the number of observed structures with respect to the numbers observed in the gas phase. When adsorbed, Arg donates electrons to the surface, becoming slightly positively charged. We do not observe fully extended structures lying on the surface, and most conformers exhibit intramolecular H-bonds. The majority of conformers of Arg in the low-energy region adsorb with the $C_\alpha H$ chiral center pointing the hydrogen atom away from the surfaces.

Arginine in its protonated form, i.e. Arg-H^+, is the most abundant form of this amino-acid in biological environments, where it typically adopts the zwitterionic protomer **P7**. In the gas-phase, we observe that the non-zwitterionic state **P6** is dominant and that the addition of a proton decreases the number of allowed conformations with respect to isolated Arg due to the added electrostatic interactions, and the neutralization of the carboxyl group that would otherwise be involved in intramolecular H-bonds. Upon adsorption to metallic surfaces, we observe that the protomer **P6** is still dominant and that there are no strong bonds formed to the surface. In addition, this molecule receives electrons from the surface, thus becoming less positively charged. Both effects conspire to yield a homogeneous (flat) molecule-surface interaction, and a relatively high population of different structures in the low-energy range. Contrary to Arg, most low-energy conformers of Arg-H^+ adsorb with the C_α-H chiral center pointing the hydrogen atom towards to the surfaces. Finally, through the calculation of dissociation energies, we also conclude that the deprotonation of Arg-H^+ is energetically favorable only on Cu(111).

Our observations regarding the preferred protomers and deprotonation propensities discussed above are consistent with the observations in the literature that the adsorption of amino acids in their anionic and deprotonated form is common on reactive metals like Cu(111) [61]. One pronounced difference that we find among surfaces is the average adsorption height of the molecules: They follow the trend Cu(111) < Ag(111) < Au(111), and Arg is always closer than Arg-H^+ to the same respective surface.

The set of electronic-structure calculations presented here show that a flexible amino-acid like Arginine presents a rich conformational space involving different protomeric states and molecule orientations with respect to the surface, allied to a complex charge rearrangement. Going forward, it is clear that the likes of this study based solely on DFT cannot become a routine method due to the elevated computational cost. Addressing the whole breadth

of amino acids as well as self assembly of these structures on surfaces will profit from this study as a benchmark and a means to develop models, possibly based on different machine-learning techniques that can bypass the cost of thousands of DFT structure optimizations. Unfortunately, to the best of our knowledge there is still no experimental results available for Arg and Arg-H^+ adsorbed on coinage metals are available.

With the approaching technology of exascale computing, we need to develop software that can efficiently utilize the available computational resources. We developed the GenSec package as a step further in automatising the kind of structure search described above. This can help reduce the effort required to carry out these kinds of investigations, and opens the path for routinely perform high-throughput calculations of interface systems, and also for modelling of self-assemblies formed on inorganic substrates. Many tasks that previously required manually setting parameters are automated in the package, such as the identification of the internal degrees of freedom of the flexible part of the interface. By setting up periodic boundary conditions, the package can produce arbitrary amount of molecules per unit cell that are obtained from the template. The construction of the database in a standardized form will make it possible to efficiently share data between researchers using modern material science repositories, facilitating the general understanding of the processes at the atomic level. Data produced with GenSec is suitable for parametrizing FFs and applying to machine learning methods that would allow the investigation of thermodynamical properties of systems and carry out calculations at longer time scales. The geometry optimization schemes together with their preconditioning schemes increase the efficiency for the most important part in the database generation procedure. The random search strategy implemented in the GenSec can be seen as robust foundation for other global search techniques such as evolutionary algorithms or Bayesian optimization methods, that rely on random generation to some extent. Further development would involve automated schemes for producing low dimensional representations using the procedures used in this thesis. The preconditioning schemes described in the thesis should be tested on the wide range of the different systems, leading to further optimisation of these techniques and increasing the efficiency of databases generation.

As a result of the efficient utilization of resources (multiple structure searches were carried out at the same time connected to the shared database) and the unbiased sampling of the conformational space of a flexible molecule, GenSec provided satisfactory results for the structure search for di-L-alanine adsorbed on Cu(110) surface. As a result of the newly developed infrastructure, it is now possible to establish search constraints based on experimental input and choose from a large number of electronic structure codes that are available to connect through the ASE package. The databases that have been created contain the lowest energy structure and all of the intermediate steps and their energies and forces, all in a single format that is easy to share and reuse.

A Additional information on Arg and Arg-H^+ on metallic surfaces

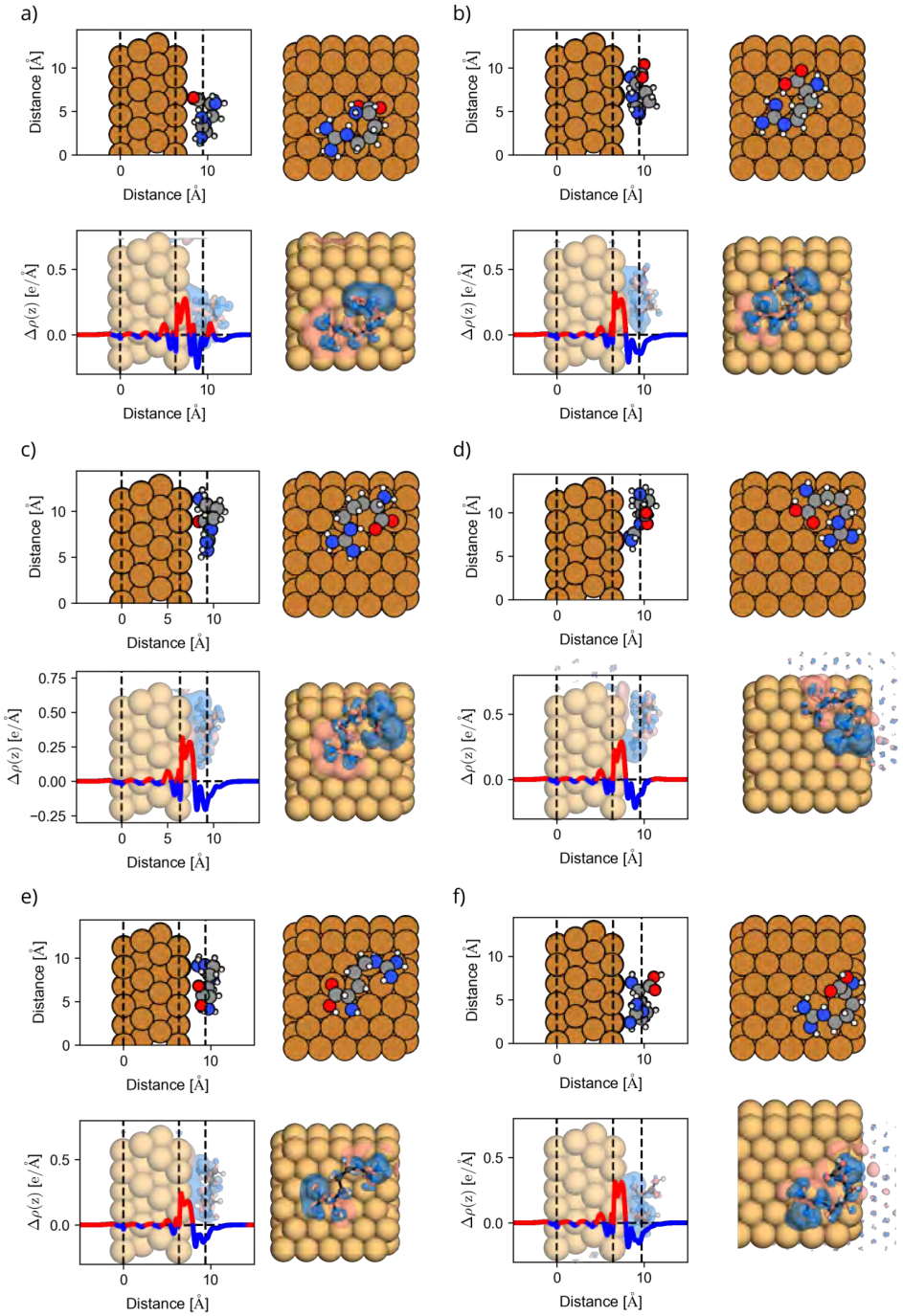


Figure A.1 – Side and top views of the adsorbed structures of Arg on Cu(111). Dashed black lines correspond to: average z position of the atoms in the lowest layer of the surface (left), average z position of atoms in the highest layer of the surface (middle), centre of the mass of the molecule (right). Red/blue solid lines (and also red/blue regions) correspond to the electron density accumulation/depletion.

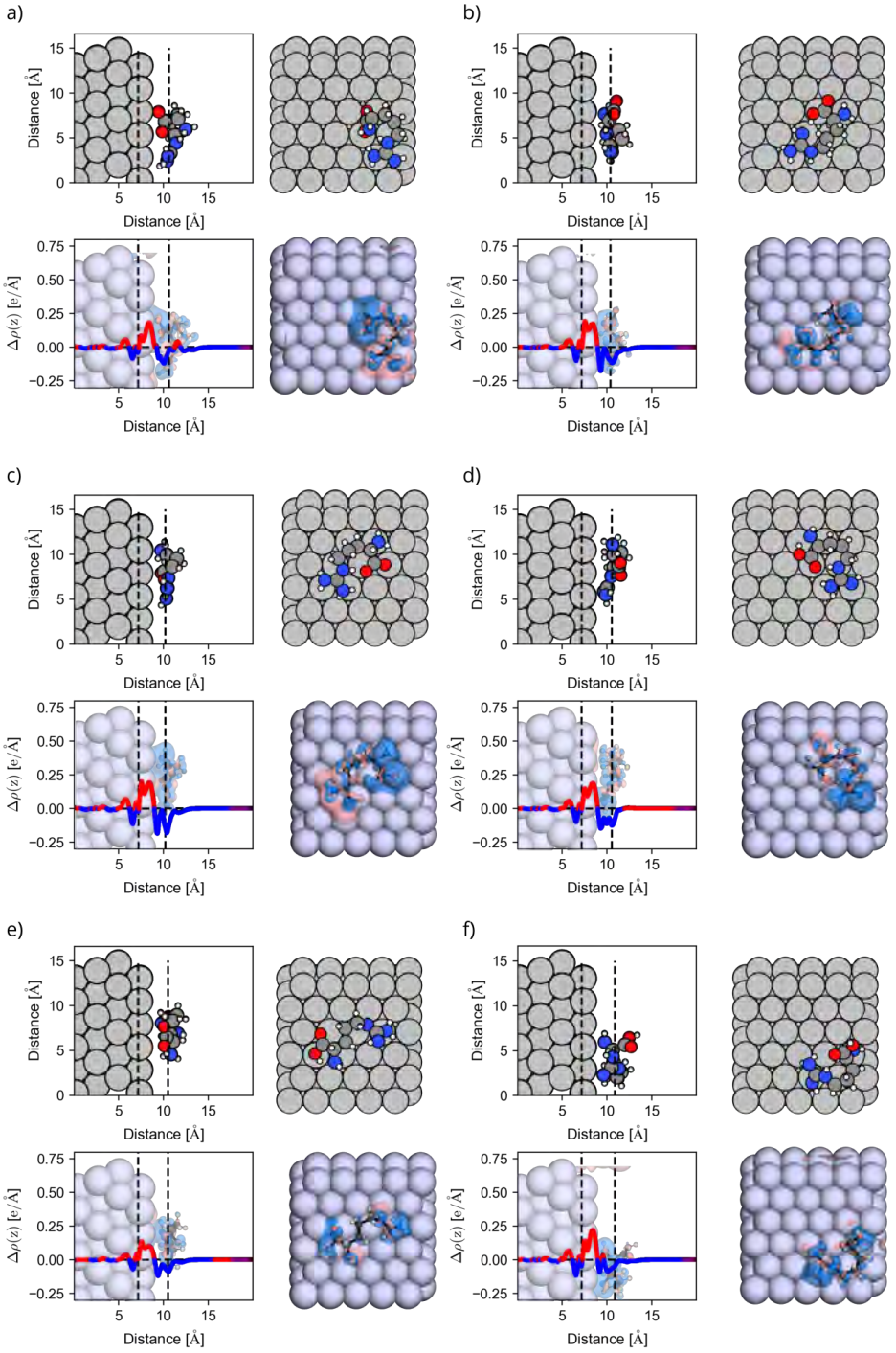


Figure A.2 – Side and top views of the adsorbed structures of Arg on Ag(111). Dashed black lines correspond to: average z position of the atoms in the lowest layer of the surface (left), average z position of atoms in the highest layer of the surface (middle), centre of the mass of the molecule (right). Red/blue solid lines (and also red/blue regions) correspond to the electron density accumulation/depletion.

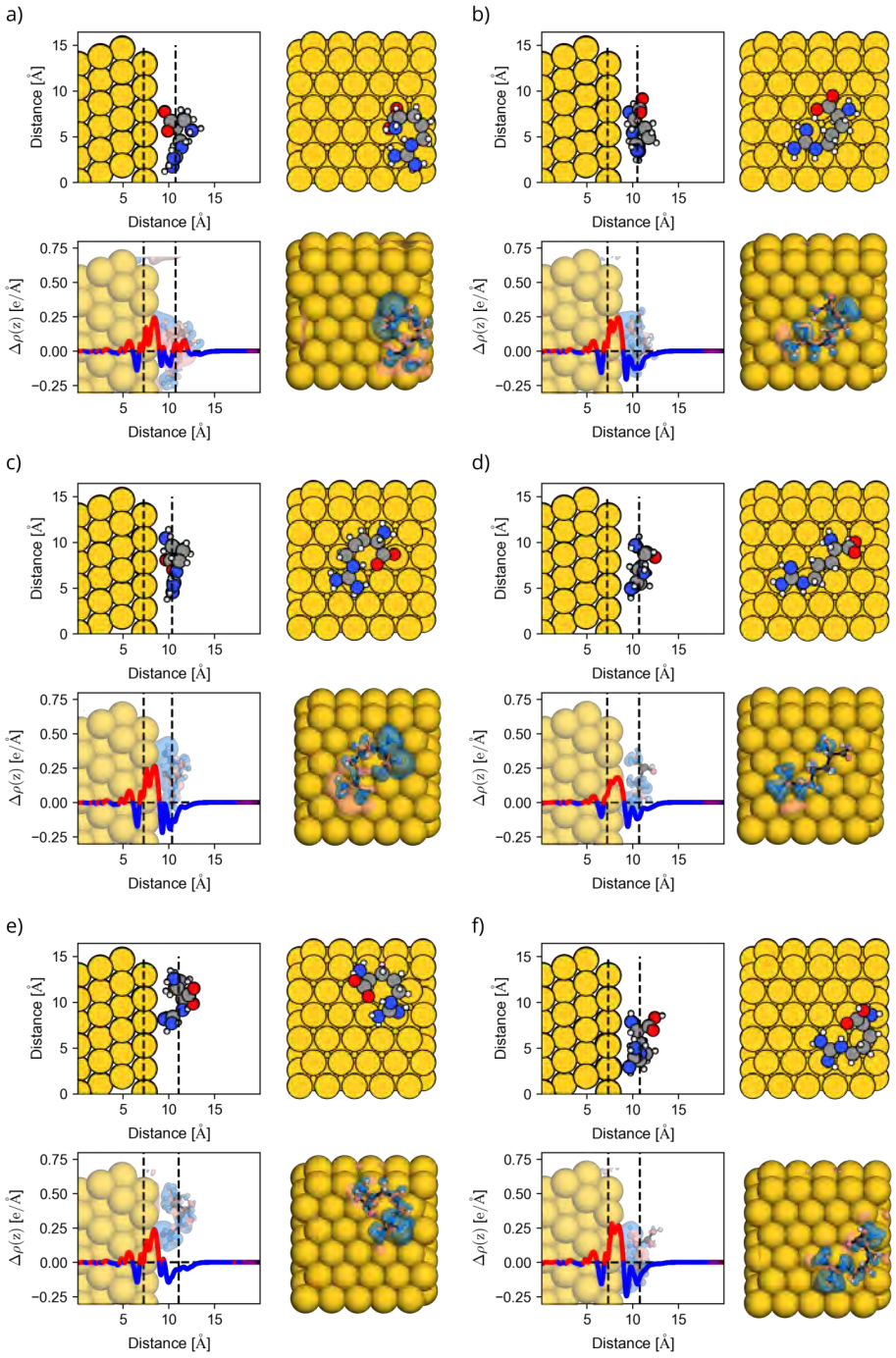


Figure A.3 – Side and top views of the adsorbed structures of Arg on Au(111). Dashed black lines correspond to: average z position of the atoms in the lowest layer of the surface (left), average z position of atoms in the highest layer of the surface (middle), centre of the mass of the molecule (right). Red/blue solid lines (and also red/blue regions) correspond to the electron density accumulation/depletion.

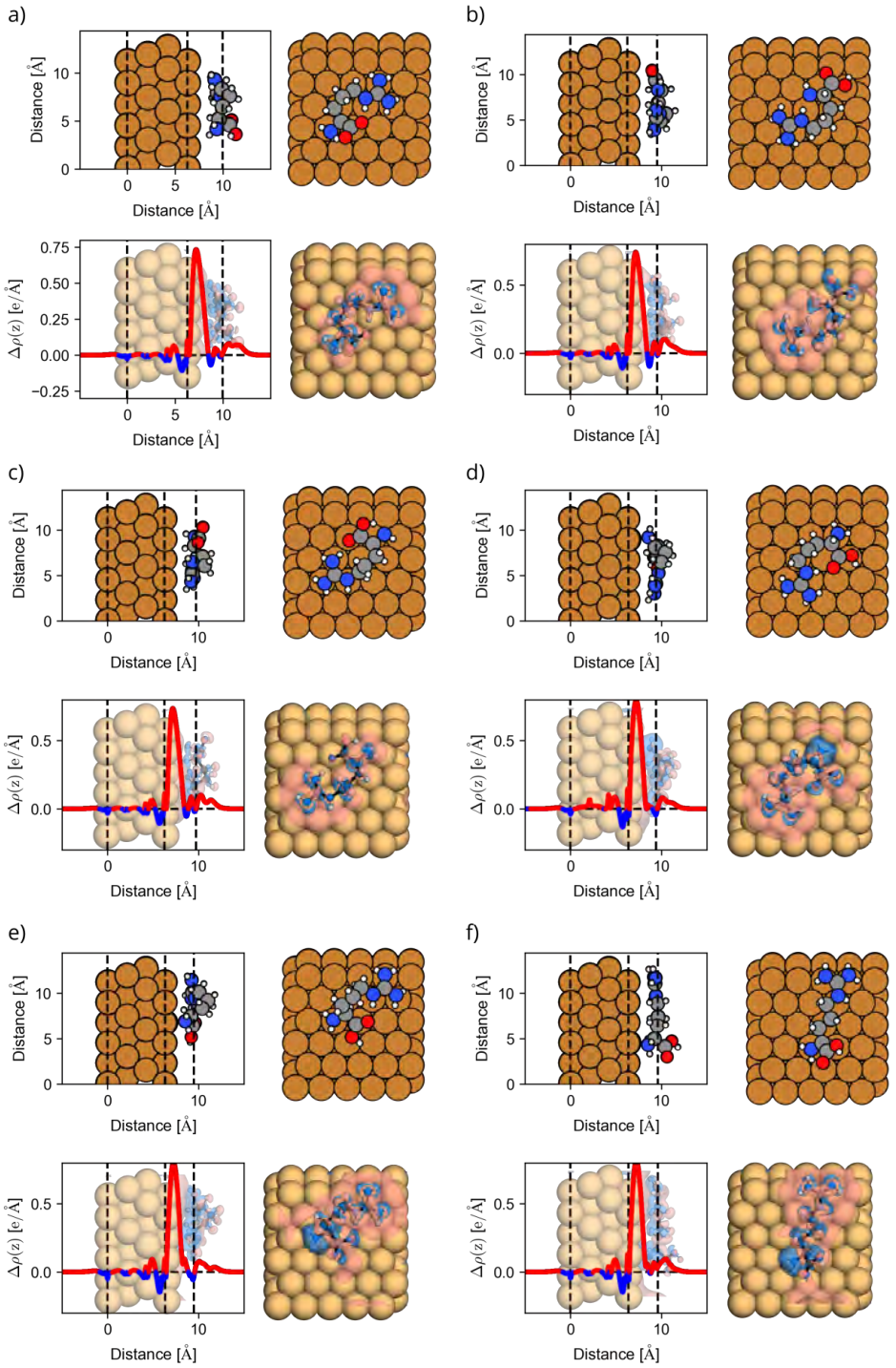


Figure A.4 – Side and top views of the adsorbed structures of Arg-H^+ on Cu(111). Dashed black lines correspond to: average z position of the atoms in the lowest layer of the surface (left), average z position of atoms in the highest layer of the surface (middle), centre of the mass of the molecule (right). Red/blue solid lines (and also red/blue regions) correspond to the electron density accumulation/depletion.

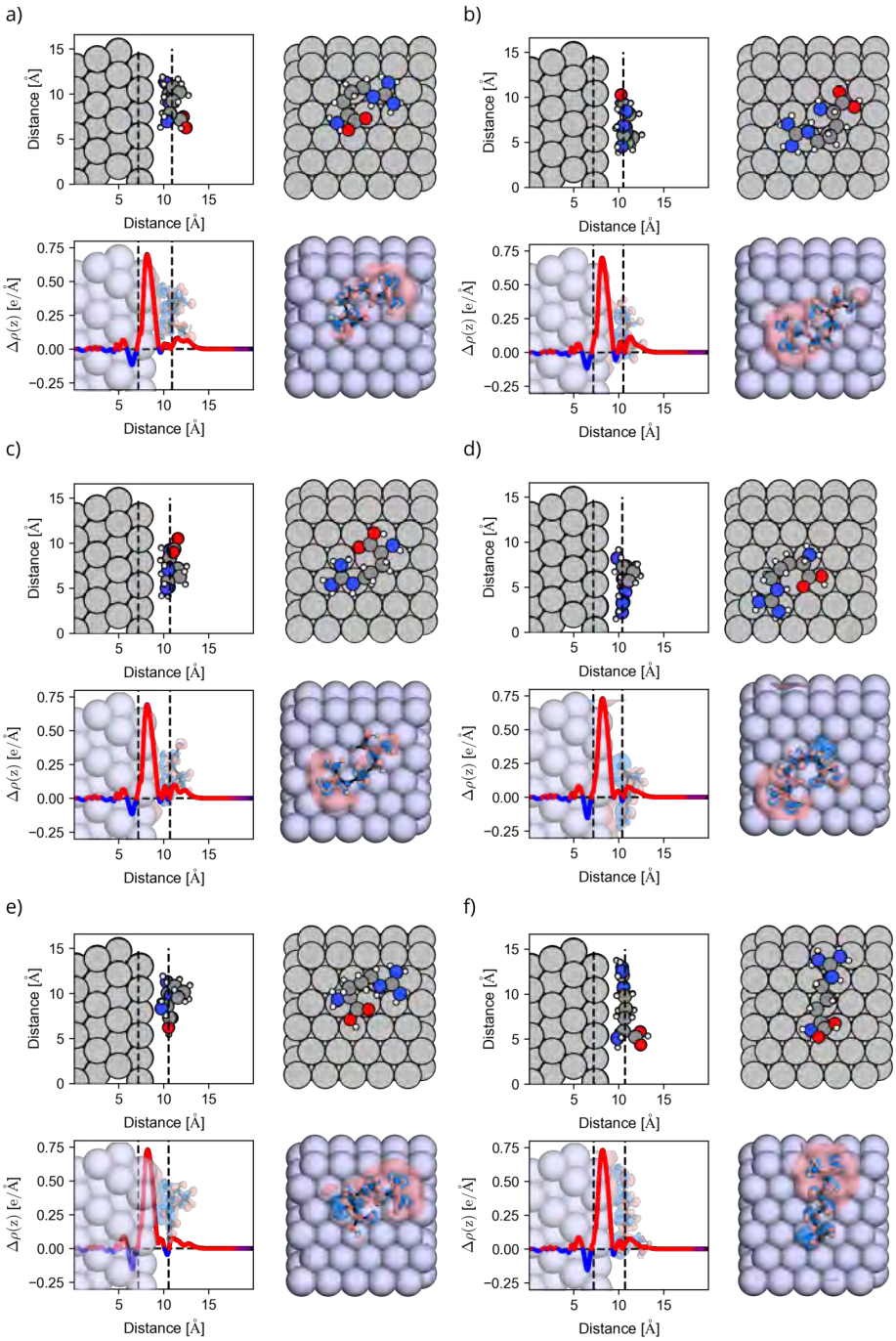


Figure A.5 – Side and top views of the adsorbed structures of Arg-H$^+$ on Ag(111). Dashed black lines correspond to: average z position of the atoms in the lowest layer of the surface (left), average z position of atoms in the highest layer of the surface (middle), centre of the mass of the molecule (right). Red/blue solid lines (and also red/blue regions) correspond to the electron density accumulation/depletion.

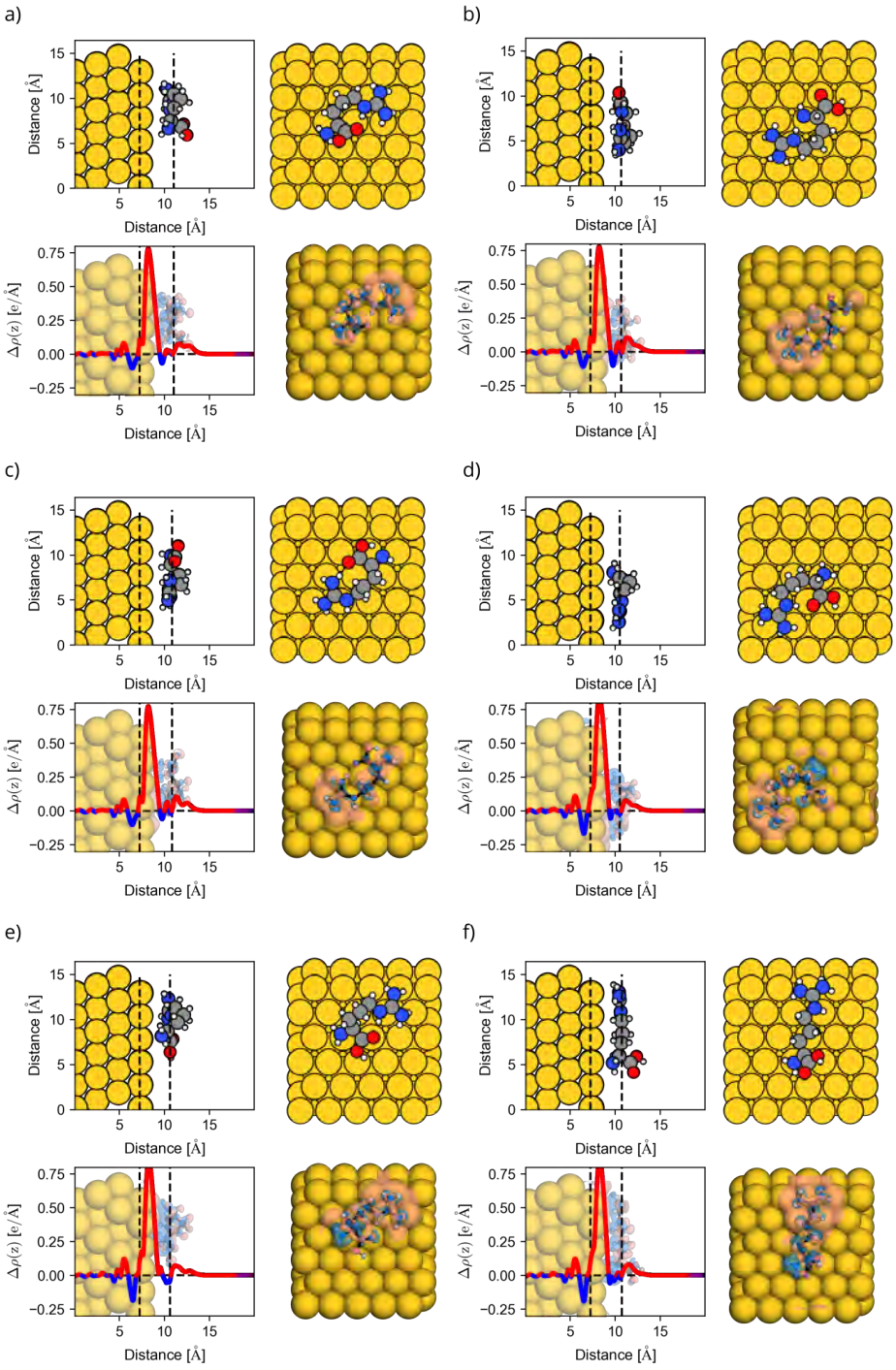


Figure A.6 – Side and top views of the adsorbed structures of Arg-H^+ on Au(111). Dashed black lines correspond to: average z position of the atoms in the lowest layer of the surface (left), average z position of atoms in the highest layer of the surface (middle), centre of the mass of the molecule (right). Red/blue solid lines (and also red/blue regions) correspond to the electron density accumulation/depletion.

B Additional information on di-L-alanine molecule on Cu(110)

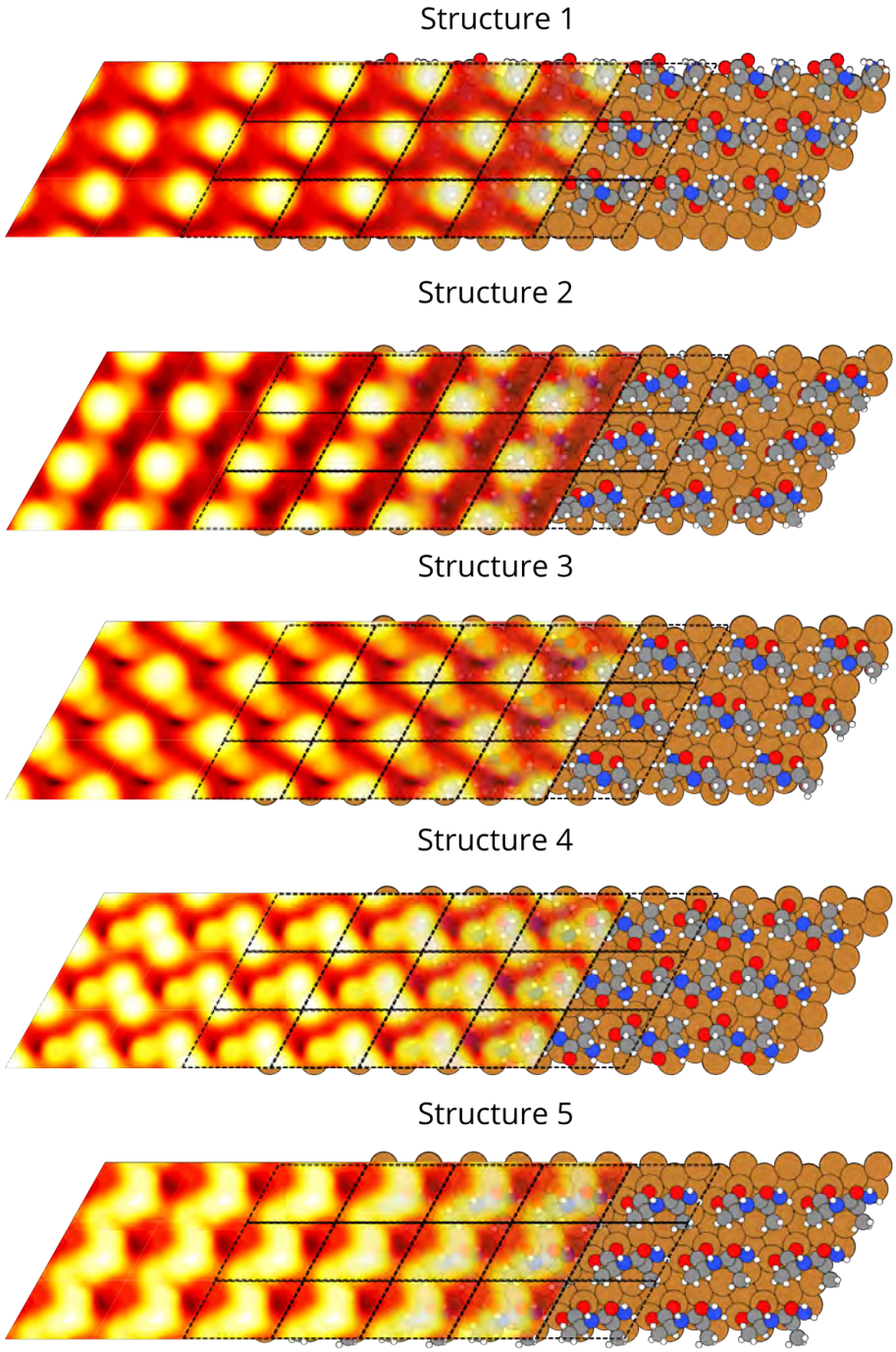


Figure B.1 – Modelled STM images and structures 1-5 of di-L-alanine molecules adsorbed on Cu(110) surface together with unit cell represented with black dashed lines

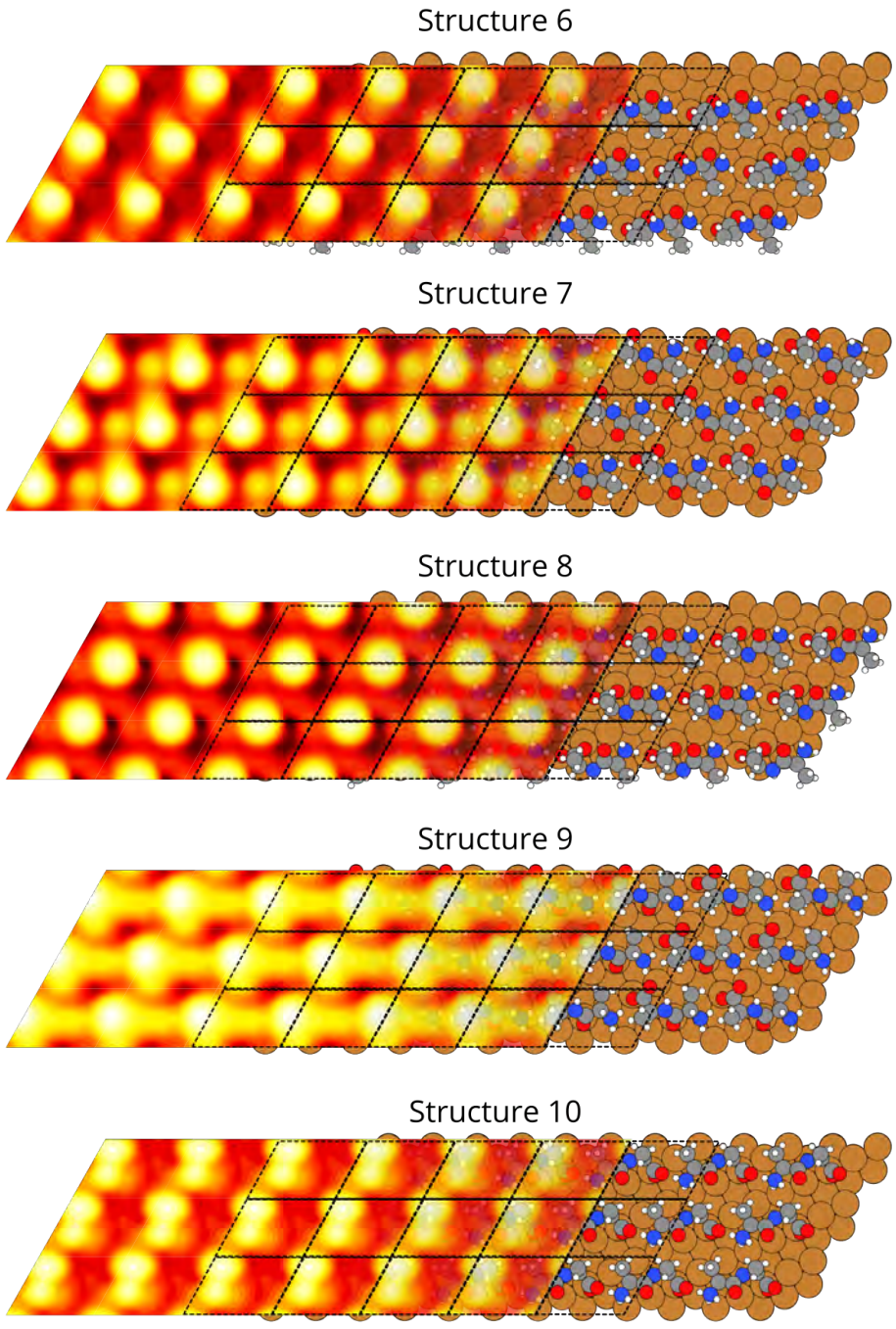


Figure B.2 – Modelled STM images and structures 6-10 of di-L-alanine molecules adsorbed on Cu(110) surface together with unit cell represented with black dashed lines

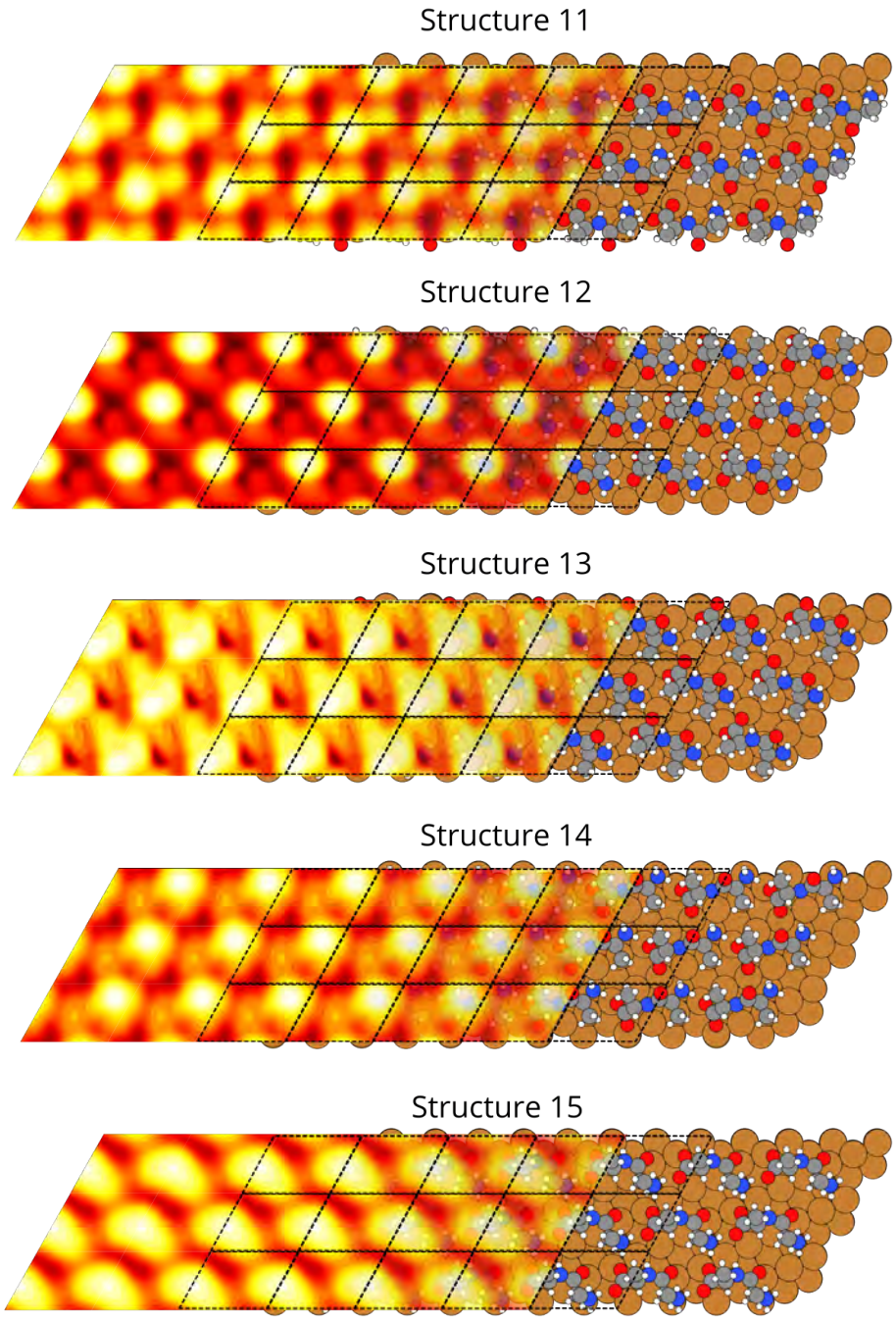


Figure B.3 – Modelled STM images and structures 11-15 of di-L-alanine molecules adsorbed on Cu(110) surface together with unit cell represented with black dashed lines

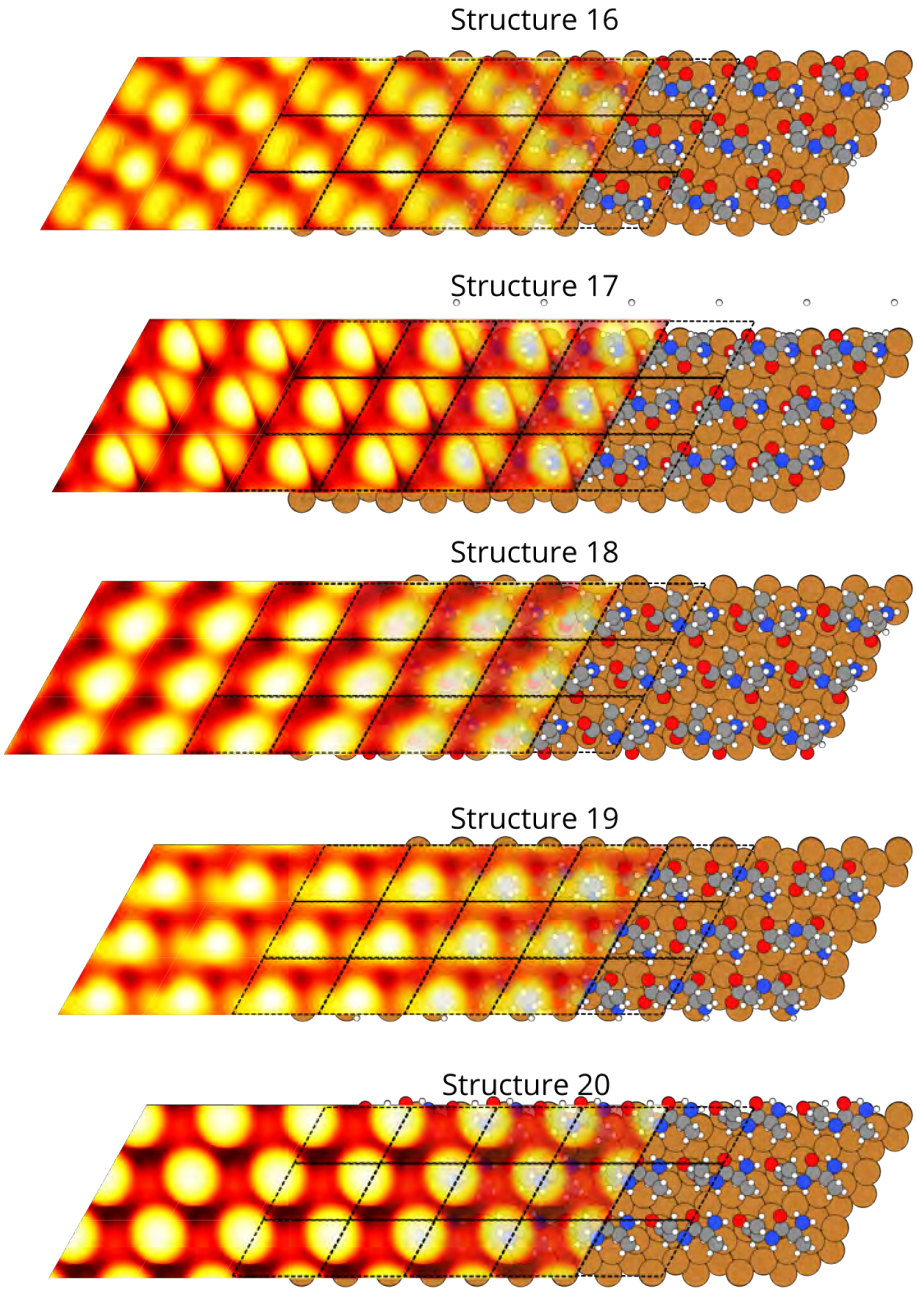


Figure B.4 – Modelled STM images and structures 16-20 of di-L-alanine molecules adsorbed on Cu(110) surface together with unit cell represented with black dashed lines

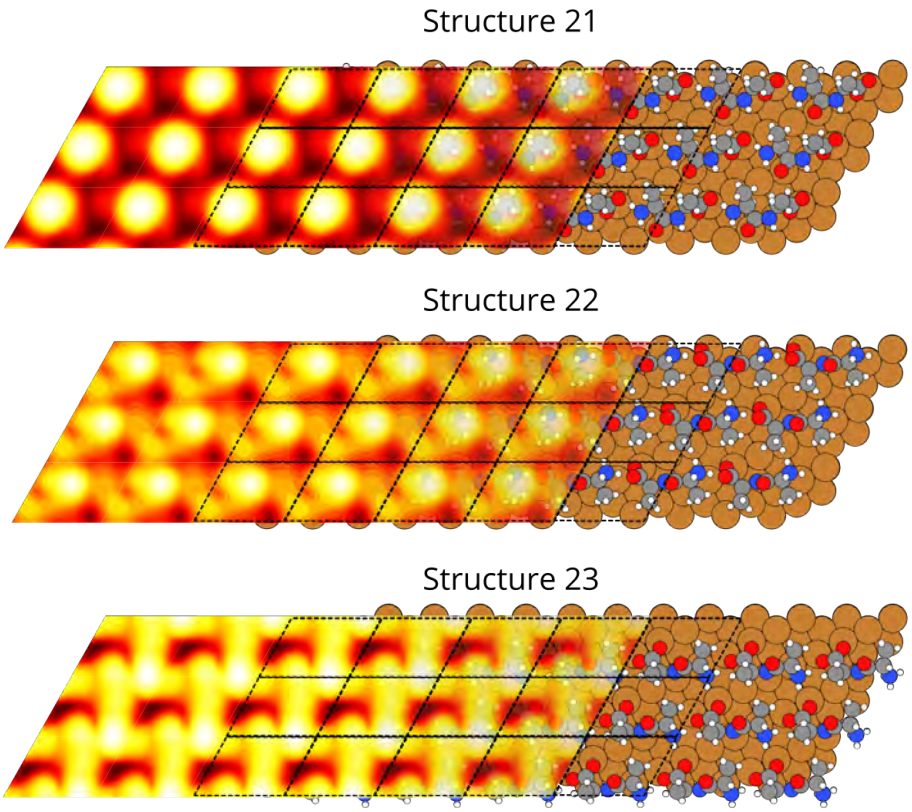


Figure B.5 – Modelled STM images and structures 21-23 of di-L-alanine molecules adsorbed on Cu(110) surface together with unit cell represented with black dashed lines

A Estimation of stabilizing interactions for di-L-alanine on Cu(110)

In order to address the question about molecular inter- and intrastrand interactions we needed to create bigger systems that would isolate only one strand that would allow to compare it to the fully periodic system, where molecule-molecule and molecule-surface interactions could be separated. For that the 6×2 unit cell slab system was prepared on which the different length of one strand will be calculated. For one unit cell sized systems and for bigger systems we applied $10\times10\times1$ and $1\times5\times1$ k-point sampling correspondingly. Now we would like to introduce some notations:

E - Energy
EB - Binding energy
S_1 - fully periodic single unit cell system
S_2^n - isolated system with n molecules on large surface

The binding energy can be calculated as follows:

$$EB_{S_1} = E_{S_1} - E_{\text{mol}} - E_{\text{surf}} \tag{A.1}$$

where E_{mol} is the energy of the isolated molecule in the same configuration and E_{surf} - energy of the small surface taken from S_1. This binding energy EB_{S_1} contains all the contributions between molecules and surfaces, molecular intrastrand interactions (within the strand) and interstrand (between strands).

$$EB_{S_2}^n = E_{S_2}^n - nE_{\text{mol}} - E_{\text{surf}}, \tag{A.2}$$

where $EB_{S_2}^1$ is just molecules-surface interaction that can be subtracted in order to isolate molecule-molecule interactions. Now we create incremental function that calculates increase of molecule-molecule interaction with increasing of amount of molecules:

$$\Delta EB(n) = (E_{S_2}^n - E_{S_2}^{n-1}) - E_{\text{mol}}. \tag{A.3}$$

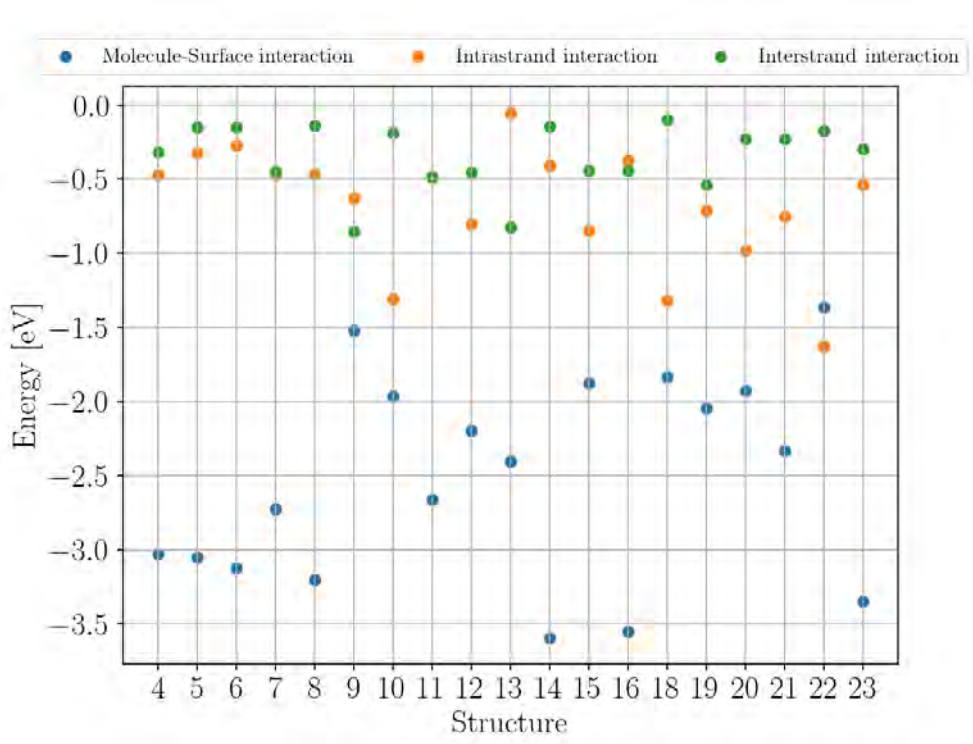


Figure A.1 – Molecule-surface, intrastrand and inerstrand interactions for the lowest energy structures of di-L-alanine adsorbed on Cu(110) surface

After that we need to subtract the molecule-surface interaction and we get the energy gain when we add one molecule to the strand:

$$E_{\text{interstrand}}(n) = \Delta E\,B(n) - E\,B^{1}_{S_2}. \tag{A.4}$$

Finally, the molecule-molecule interaction is also given by:

$$E\,B_{S_1} - E\,B^{1}_{S_2} = E_{\text{mol-mol}} = E_{\text{interstrand}}(\infty) + E_{\text{intrastrand}}(\infty), \tag{A.5}$$

using which one can calculate convergence with large n.

For all the structures that are not deprotonated on surface with the formulas obtained above we decompose the interactions and the results can be found in Fig. A.1. The results should be further processed in order to draw more precise conclusions, but what one can immediately see is that the only structure for which the intrastrand and interstrand interactions are almost identical in energy are structures 7 and 11 that are very similar (probably will fall to the same local minima if we perform geometry optimization with tighter settings).

Bibliography

[1] Tiffany R Walsh and Marc R Knecht. Biointerface Structural Effects on the Properties and Applications of Bioinspired Peptide-Based Nanomaterials. *Chemical Reviews*, 117(20):12641–12704, 2017.

[2] 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. *Scientific Data*, 3:160009, 2016.

[3] Ludwig Bartels. Tailoring molecular layers at metal surfaces, 2010.

[4] Bin Yang, Dave J. Adams, Maria Marlow, and Mischa Zelzer. Surface-mediated supramolecular self-assembly of protein, peptide, and nucleoside derivatives: From surface design to the underlying mechanism and tailored functions. *Langmuir*, 34(50):15109–15125, 2018. PMID: 30032622.

[5] E. J. Wood. Fundamentals of biochemistry: Life at the molecular level (Third Edition) by D. Voet, J. Voet, and C. W. Pratt. *Biochemistry and Molecular Biology Education*, 2008.

[6] Philip J. Cowen and Michael Browning. What has serotonin to do with depression?, 2015.

[7] Emanuela Gatto, Lorenzo Stella, Fernando Formaggio, Claudio Toniolo, Leandro Lorenzelli, and Mariano Venanzi. Electroconductive and photocurrent generation properties of self-assembled monolayers formed by functionalized, conformationally-constrained peptides on gold electrodes. *Journal of Peptide Science*, 2008.

[8] Emanuela Gatto, Alessia Quatela, Mario Caruso, Roberto Tagliaferro, Marta De Zotti, Fernando Formaggio, Claudio Toniolo, Aldo Di Carlo, and Mariano Venanzi. Mimicking nature: A novel peptide-based bio-inspired approach for solar energy conversion. *ChemPhysChem*, 2014.

[9] Emanuela Gatto, Mario Caruso, Alessandro Porchetta, Claudio Toniolo, Fernando Formaggio, Marco Crisma, and Mariano Venanzi. Photocurrent generation through peptide-based self-assembled monolayers on a gold surface: Antenna and junction effects. *Journal of Peptide Science*, 2011.

[10] Shiro Yasutomi, Tomoyuki Morita, Yukio Imanishi, and Shunsaku Kimura. A molecular photodiode system that can switch photocurrent direction. *Science*, 2004.

[11] Riming Nie, Aiyuan Li, and Xianyu Deng. Environmentally friendly biomaterials as an interfacial layer for highly efficient and air-stable inverted organic solar cells. *Journal of Materials Chemistry A*, 2014.

[12] Maayan Matmor and Nurit Ashkenasy. Modulating semiconductor surface electronic properties by inorganic peptide-binders sequence design. *Journal of the American Chemical Society*, 2012.

[13] Maayan Matmor, George A. Lengyel, W. Seth Horne, and Nurit Ashkenasy. Peptide-functionalized semiconductor surfaces: Strong surface electronic effects from minor alterations to backbone composition. *Physical Chemistry Chemical Physics*, 2017.

[14] Xiaoyin Xiao, Bingqian Xu, and Nongjian Tao. Conductance Titration of Single-Peptide Molecules. *Journal of the American Chemical Society*, 2004.

[15] Lior Sepunaru, Sivan Refaely-Abramson, Robert Lovrinčić, Yulian Gavrilov, Piyush Agrawal, Yaakov Levy, Leeor Kronik, Israel Pecht, Mordechai Sheves, and David Cahen. Electronic transport via homopeptides: The role of side chains and secondary structure. *Journal of the American Chemical Society*, 2015.

[16] Cunlan Guo, Xi Yu, Sivan Refaely-Abramson, Lior Sepunaru, Tatyana Bendikov, Israel Pecht, Leeor Kronik, Ayelet Vilan, Mordechai Sheves, and David Cahen. Tuning electronic transport via hepta-alanine peptides junction by tryptophan doping. *Proceedings of the National Academy of Sciences of the United States of America*, 2016.

[17] Y. Pennec, W. Auwärter, A. Schiffrin, A. Weber-Bargioni, A. Riemann, and J. V. Barth. Supramolecular gratings for tuneable confinement of electrons on metal surfaces. *Nature Nanotechnology*, 2007.

[18] Ken Kanazawa, Atsushi Taninaka, Hui Huang, Munenori Nishimura, Shoji Yoshida, Osamu Takeuchi, and Hidemi Shigekawa. Scanning tunneling microscopy/spectroscopy on self-assembly of a glycine/cu(111) nanocavity array. *Chem. Commun.*, 47:11312–11314, 2011.

[19] Clara Shlizerman, Alexander Atanassov, Inbal Berkovich, Gonen Ashkenasy, and Nurit Ashkenasy. De novo designed coiled-coil proteins with variable conformations as components of molecular electronic devices. *Journal of the American Chemical Society*, 2010.

[20] Shengfu Chen, Zhiqiang Cao, and Shaoyi Jiang. Ultra-low fouling peptide surfaces derived from natural amino acids. *Biomaterials*, 2009.

[21] Sivan Nir and Meital Reches. Bio-inspired antifouling approaches: The quest towards non-toxic and non-biocidal materials, 2016.

[22] Mehmet Sarikaya, Candan Tamerler, Alex K.Y. Jen, Klaus Schulten, and François Baneyx. Molecular biomimetics: nanotechnology through biology. *Nature Materials*, 2(9):577–585, 2003.

[23] Carlos Mas-Moruno, Roberta Fraioli, Fernando Albericio, José María Manero, and F. Javier Gil. Novel peptide-based platform for the dual presentation of biologically active peptide motifs on biomaterials. In *ACS Applied Materials and Interfaces*, 2014.

[24] Mareen Pagel, Rayk Hassert, Torsten John, Klaus Braun, Manfred Wießler, Bernd Abel, and Annette G. Beck-Sickinger. Multifunctional Coating Improves Cell Adhesion on Titanium by using Cooperatively Acting Peptides. *Angewandte Chemie - International Edition*, 2016.

[25] S. M. Barlow and R. Raval. Complex organic molecules at metal surfaces: Bonding, organisation and chirality, 2003.

[26] Magalí Lingenfelder, Giulia Tomba, Giovanni Costantini, Lucio Colombi Ciacchi, Alessandro De Vita, and Klaus Kern. Tracking the chiral recognition of adsorbed dipeptides at the single-molecule level. *Angewandte Chemie - International Edition*, 2007.

[27] Magalí Lingenfelder. *Chiral recognition and supramolecular self-assembly of adsorbed amino acids and dipeptides at the submolecular level.* PhD thesis, École Polytechnique Fédèrale de Lausanne, 2008.

[28] Dmitriy Khatayevich, Christopher R. So, Yuhei Hayamizu, Carolyn Gresswell, and Mehmet Sarikaya. Controlling the surface chemistry of graphite by engineered self-assembled peptides. *Langmuir*, 2012.

[29] Lihi Adler-Abramovich, Daniel Aronov, Peter Beker, Maya Yevnin, Shiri Stempler, Ludmila Buzhansky, Gil Rosenman, and Ehud Gazit. Self-assembled arrays of peptide nanotubes by vapour deposition. *Nature Nanotechnology*, 2009.

[30] P. Beker and G. Rosenman. Bioinspired nanostructural peptide materials for supercapacitor electrodes. *Journal of Materials Research*, 2010.

[31] Dharmendr Kumar, Nimesh Jain, Vinay Jain, and Beena Rai. Amino acids as copper corrosion inhibitors: A density functional theory approach. *Applied Surface Science*, 514:145905, 2020.

[32] Anton Kasprzhitskii, Georgy Lazorenko, Tatiana Nazdracheva, Aleksandr Kukharskii, Victor Yavna, and Andrei Kochur. Theoretical evaluation of the corrosion inhibition performance of aliphatic dipeptides. *New Journal of Chemistry*, 2021.

[33] Maryam Dehdab, Mehdi Shahraki, and Sayyed Mostafa Habibi-Khorassani. Theoretical study of inhibition efficiencies of some amino acids on corrosion of carbon steel in acidic media: Green corrosion inhibitors. *Amino Acids*, 2016.

[34] Jia Jun Fu, Su Ning Li, Ying Wang, Lin Hua Cao, and Lu De Lu. Computational and electrochemical studies of some amino acid compounds as corrosion inhibitors for mild steel in hydrochloric acid solution. *Journal of Materials Science*, 2010.

[35] Sanjoy Satpati, Aditya Suhasaria, Subhas Ghosal, Abhijit Saha, Sukalpa Dey, and Dipankar Sukul. Amino acid and cinnamaldehyde conjugated Schiff bases as proficient corrosion inhibitors for mild steel in 1 M HCl at higher temperature and prolonged exposure: Detailed electrochemical, adsorption and theoretical study. *Journal of Molecular Liquids*, 2021.

[36] Manu S. Mannoor, Hu Tao, Jefferson D. Clayton, Amartya Sengupta, David L. Kaplan, Rajesh R. Naik, Naveen Verma, Fiorenzo G. Omenetto, and Michael C. McAlpine. Graphene-based wireless bacteria detection on tooth enamel. *Nature Communications*, 2012.

[37] Hye Jin Hwang, Myung Yi Ryu, Chan Young Park, Junki Ahn, Hyun Gyu Park, Changsun Choi, Sang Do Ha, Tae Jung Park, and Jong Pil Park. High sensitive and selective electrochemical biosensor: Label-free detection of human norovirus using affinity peptide as molecular binder. *Biosensors and Bioelectronics*, 2017.

[38] Ning Xia, Xin Wang, Jie Yu, Yangyang Wu, Shuchao Cheng, Yun Xing, and Lin Liu. Design of electrochemical biosensors with peptide probes as the receptors of targets and the inducers of gold nanoparticles assembly on electrode surface. *Sensors and Actuators, B: Chemical*, 2017.

[39] Giwan Seo, Geonhee Lee, Mi Jeong Kim, Seung Hwa Baek, Minsuk Choi, Keun Bon Ku, Chang Seop Lee, Sangmi Jun, Daeui Park, Hong Gi Kim, Seong Jun Kim, Jeong O. Lee, Bum Tae Kim, Edmond Changkyun Park, and Seung Il Kim. Rapid Detection of COVID-19 Causative Virus (SARS-CoV-2) in Human Nasopharyngeal Swab Specimens Using Field-Effect Transistor-Based Biosensor. *ACS nano*, 2020.

[40] Javad Payandehpeyman, Neda Parvini, Kambiz Moradi, and Nima Hashemian. Detection of sars-cov-2 using antibody–antigen interactions with graphene-based nanomechanical resonator sensors. *ACS Applied Nano Materials*, 0(0):null, 0.

[41] Owen J. Guy, Gregory Burwell, Zari Tehrani, Ambroise Castaing, Kelly Ann Walker, and S.H. Doak. Graphene Nano-Biosensors for Detection of Cancer Risk. *Materials Science Forum*, 711:246–252, 2012.

[42] Dmitriy Khatayevich, Tamon Page, Carolyn Gresswell, Yuhei Hayamizu, William Grady, and Mehmet Sarikaya. Selective detection of target proteins by peptide-enabled graphene biosensor. *Small*, 10(8):1505–1513, 2014.

[43] Vincent Humblot, Christophe Méthivier, Rasmita Raval, and Claire Marie Pradier. Amino acid and peptides on Cu(1 1 0) surfaces: Chemical and structural analyses of l-lysine. *Surface Science*, 2007.

[44] M. Smerieri, L. Vattuone, T. Kravchuk, D. Costa, and L. Savio. (S)-glutamic acid on Ag(100): Self-assembly in the nonzwitterionic form. *Langmuir*, 2011.

[45] S. M. Barlow, S. Louafi, D. Le Roux, J. Williams, C. Muryn, S. Haq, and R. Raval. Polymorphism in supramolecular chiral structures of R- and S-alanine on Cu(1 1 0). *Surface Science*, 2005.

[46] Li Ping Xu, Yibiao Liu, and Xueji Zhang. Interfacial self-assembly of amino acids and peptides: Scanning tunneling microscopy investigation, 2011.

[47] A. Kühnle, L. M. Molina, T. R. Linderoth, B. Hammer, and F. Besenbacher. Growth of unidirectional molecular rows of cysteine on Au(110)-(1 × 2) driven by adsorbate-induced surface rearrangements. *Physical Review Letters*, 2004.

[48] Angelika Kühnle, Trolle R. Linderoth, Michael Schunack, and Flemming Besenbacher. L-cysteine adsorption structures on Au(111) investigated by scanning tunneling microscopy under ultrahigh vacuum conditions. *Langmuir*, 2006.

[49] Sybille Fischer, Anthoula C. Papageorgiou, Matthias Marschall, Joachim Reichert, Katharina Diller, Florian Klappenberger, Francesco Allegretti, Alexei Nefedov, Christof Wöll, and Johannes V. Barth. L -Cysteine on Ag(111): A combined STM and X-ray spectroscopy study of anchorage and deprotonation. *Journal of Physical Chemistry C*, 2012.

[50] Christian Engelbrekt, Renat R. Nazmutdinov, Tamara T. Zinkicheva, Dmitrii V. Glukhov, Jiawei Yan, Bingwei Mao, Jens Ulstrup, and Jingdong Zhang. Chemistry of cysteine assembly on Au(100): Electrochemistry,: in situ STM and molecular modeling. *Nanoscale*, 2019.

[51] Feng Gao, Zhenjun Li, Yilin Wang, Luke Burkholder, and W. T. Tysoe. Chemistry of Alanine on Pd(1 1 1): Temperature-programmed desorption and X-ray photoelectron spectroscopic study. *Surface Science*, 2007.

[52] Julia Laskin, Peng Wang, and Omar Hadjar. Soft-landing of peptide ions onto self-assembled monolayer surfaces: An overview, 2008.

[53] Stephan Rauschenbach, Frank L. Stadler, Eugenio Lunedei, Nicola Malinowski, Sergej Koltsov, Giovanni Costantini, and Klaus Kern. Electrospray ion beam deposition of clusters and biomolecules. *Small*, 2006.

[54] Zoltán Takáts, Justin M. Wiseman, Bogdan Gologan, and R. Graham Cooks. Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science*, 2004.

[55] Stephan Rauschenbach, Ralf Vogelgesang, N. Malinowski, Jürgen W. Gerlach, Mohamed Benyoucef, Giovanni Costantini, Zhitao Deng, Nicha Thontasen, and Klaus Kern. Electrospray ion beam deposition: Soft-landing and fragmentation of functional molecules at solid surfaces. *ACS Nano*, 2009.

[56] Stephan Rauschenbach, Markus Ternes, Ludger Harnau, and Klaus Kern. Mass Spectrometry as a Preparative Tool for the Surface Science of Large Molecules. *Annual Review of Analytical Chemistry*, 9(1):473–498, 2016.

[57] G. Binnig and H. Rohrer. SCANNING TUNNELING MICROSCOPY G. BINNIG and H. ROHRER. *Surface Science*, 1983.

[58] Zhitao Deng, Nicha Thontasen, Nikola Malinowski, Gordon Rinke, Ludger Harnau, Stephan Rauschenbach, and Klaus Kern. A close look at proteins: Submolecular resolution of two- and three-dimensionally folded cytochrome c at surfaces. *Nano Letters*, 12(5):2452–2458, 2012. PMID: 22530980.

[59] Gordon Rinke, Stephan Rauschenbach, Ludger Harnau, Alyazan Albarghash, Matthias Pauly, and Klaus Kern. Active conformation control of unfolded proteins by hyperthermal collision with a metal surface. *Nano Letters*, 14(10):5609–5615, 2014. PMID: 25198655.

[60] Stephan Rauschenbach, Gordon Rinke, Rico Gutzler, Sabine Abb, Alyazan Albarghash, Duy Le, Talat S. Rahman, Michael Duy, Ludger Harnau, and Klaus Kern. Two-Dimensional Folding of Polypeptides into Molecular Nanostructures at Surfaces. *ACS Nano*, 11(3):2420–2427, 2017.

[61] Dominique Costa, Claire-Marie Pradier, Frederik Tielens, and Letizia Savio. Adsorption and self-assembly of bio-organic molecules at model surfaces: A route towards increased complexity. *Surface Science Reports*, 70(4):449–553, 2015.

[62] Marian L. Clegg, Leonardo Morales De La Garza, Sofia Karakatsani, David A. King, and Stephen M. Driver. Chirality in amino acid overlayers on Cu surfaces. *Topics in Catalysis*, 2011.

[63] Yeliang Wang, Magalí Lingenfelder, Stefano Fabris, Guido Fratesi, Riccardo Ferrando, Thomas Classen, Klaus Kern, and Giovanni Costantini. Programming hierarchical supramolecular nanostructures by molecular design. *The Journal of Physical Chemistry C*, 117(7):3440–3445, 2013.

[64] K. E. Wilson, H. A. Früchtl, F. Grillo, and C. J. Baddeley. (s)-lysine adsorption induces the formation of gold nanofingers on au111. *Chem. Commun.*, 47:10365–10367, 2011.

[65] Xiubo Zhao, Fang Pan, and Jian R. Lu. Recent development of peptide self-assembly. *Progress in Natural Science*, 18(6):653–660, 2008.

[66] Joachim Reichert, Agustin Schiffrin, Willi Auwärter, Alexander Weber-Bargioni, Matthias Marschall, Martina Dell'Angela, Dean Cvetko, Gregor Bavdek, Albano Cossaro, Alberto Morgante, and Johannes V. Barth. L-tyrosine on Ag(111): Universality of the amino acid 2D zwitterionic bonding scheme? In *ACS Nano*, 2010.

[67] Vitally Feyer, Oksana Plekan, Tomáš Skála, Vladimír Cháb, Vladimír Matolín, and Kevin C. Prince. The electronic structure and adsorption geometry of L-histidine on Cu(110). *Journal of Physical Chemistry B*, 2008.

[68] J. Williams, S. Haq, and R. Raval. The bonding and orientation of the amino acid L-alanine on Cu {110} determined by RAIRS. *Surface Science*, 1996.

[69] T. E. Jones, C. J. Baddeley, A. Gerbi, L. Savio, M. Rocca, and L. Vattuone. Molecular ordering and adsorbate induced faceting in the Ag{110}-(S)-glutamic acid system. *Langmuir*, 2005.

[70] V. De Renzi, L. Lavagnino, V. Corradini, R. Biagi, M. Canepa, and U. Del Pennino. Very low energy vibrational modes as a fingerprint of H-bond network formation: L-cysteine on Au(111). *Journal of Physical Chemistry C*, 2008.

[71] M. Smerieri, L. Vattuone, M. Rocca, and L. Savio. Spectroscopic evidence for neutral and anionic adsorption of (S)-glutamic acid on Ag(111). *Langmuir*, 2013.

[72] Paul Adrien Maurice Dirac and Ralph Howard Fowler. Quantum mechanics of many-electron systems. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 123(792):714–733, 1929.

[73] Rosa Di Felice, Annabella Selloni, and Elisa Molinari. DFT Study of Cysteine Adsorption on Au(111). *The Journal of Physical Chemistry B*, 107(5):1151–1156, 2003.

[74] Rosa Di Felice and Annabella Selloni. Adsorption modes of cysteine on Au(111): Thiolate, amino-thiolate, disulfide. *The Journal of Chemical Physics*, 120(10):4906–4914, 2004.

[75] Luca M Ghiringhelli, Pim Schravendijk, and Luigi Delle Site. Adsorption of alanine on a Ni(111) surface: A multiscale modeling oriented density functional study. *Phys. Rev. B*, 74(3):35437, 2006.

[76] Corinne Arrouvel, Boubakar Diawara, Dominique Costa, and Philippe Marcus. DFT Periodic Study of the Adsorption of Glycine on the Anhydrous and Hydroxylated (0001) Surfaces of α-Alumina. *The Journal of Physical Chemistry C*, 111(49):18164–18173, 2007.

[77] Francesco Iori, Stefano Corni, and Rosa Di Felice. Unraveling the Interaction between Histidine Side Chain and the Au(111) Surface: A DFT Study. *The Journal of Physical Chemistry C*, 112(35):13540–13545, 2008.

[78] Wei Liu, Alexandre Tkatchenko, and Matthias Scheffler. Modeling adsorption and reactions of organic molecules at metal surfaces. *Accounts of Chemical Research*, 2014.

[79] Musa Ozboyaci, Daria B. Kokh, Stefano Corni, and Rebecca C. Wade. Modeling and simulation of protein-surface interactions: Achievements and challenges, 2016.

[80] Sung Sik Lee, Bongsoo Kim, and Sungyul Lee. Structures and bonding properties of Gold-Arg-Cys complexes: DFT study of simple peptide-coated metal. *Journal of Physical Chemistry C*, 2014.

[81] R. R. Nazmutdinov, I. R. Manyurov, T. T. Zinkicheva, J. Jang, and J. Ulstrup. Cysteine adsorption on the Au(111) surface and the electron transfer in configuration of a scanning tunneling microscope: A quantum-chemical approach. *Russian Journal of Electrochemistry*, 2007.

[82] José L.C. Fajín, José R.B. Gomes, and M. Natália D.S. Cordeiro. DFT study of the adsorption of d-(l-)cysteine on flat and chiral stepped gold surfaces. *Langmuir*, 2013.

[83] Dmitrii Maksimov, Carsten Baldauf, and Mariana Rossi. The conformational space of a flexible amino acid at metallic surfaces. *International Journal of Quantum Chemistry*, 121(3):e26369, 2021.

[84] Gongyi Hong, Hendrik Heinz, Rajesh R Naik, Barry L Farmer, and Ruth Pachter. Toward Understanding Amino Acid Adsorption at Metallic Interfaces: A Density Functional Theory Study. *ACS Applied Materials & Interfaces*, 1(2):388–392, 2009.

[85] Louise B. Wright, P. Mark Rodger, Stefano Corni, and Tiffany R. Walsh. GolP-CHARMM: First-principles based force fields for the interaction of proteins with Au(111) and Au(100). *Journal of Chemical Theory and Computation*, 2013.

[86] Zak E. Hughes, Louise B. Wright, and Tiffany R. Walsh. Biomolecular adsorption at aqueous silver interfaces: First-principles calculations, polarizable force-field simulations, and comparisons with gold. *Langmuir*, 2013.

[87] Hendrik Heinz, R. A. Vaia, B. L. Farmer, and R. R. Naik. Accurate simulation of surfaces and interfaces of face-centered cubic metals using 12-6 and 9-6 lennard-jones potentials. *Journal of Physical Chemistry C*, 112(44):17281–17290, 2008.

[88] Andrea Grisafi and Michele Ceriotti. Incorporating long-range physics in atomic-scale machine learning. *Journal of Chemical Physics*, 2019.

[89] Alexandre Tkatchenko. Machine learning for chemical discovery, 2020.

[90] Jörg Behler. Four Generations of High-Dimensional Neural Network Potentials, 2021.

[91] Zdenek Futera and Jochen Blumberger. Adsorption of Amino Acids on Gold: Assessing the Accuracy of the GolP-CHARMM Force Field and Parametrization of Au-S Bonds. *Journal of Chemical Theory and Computation*, 2019.

[92] Thomas P. Senftle, Sungwook Hong, Md Mahbubul Islam, Sudhir B. Kylasa, Yuanxia Zheng, Yun Kyung Shin, Chad Junkermeier, Roman Engel-Herbert, Michael J. Janik, Hasan Metin Aktulga, Toon Verstraelen, Ananth Grama, and Adri C.T. Van Duin. The ReaxFF reactive force-field: Development, applications and future directions, 2016.

[93] Susanna Monti, Cui Li, and Vincenzo Carravetta. Reactive dynamics simulation of monolayer and multilayer adsorption of glycine on Cu(110). *Journal of Physical Chemistry C*, 2013.

[94] Dominik Marx and Jürg Hutter. Ab initio molecular dynamics: Theory and implementation. *Modern methods and algorithms of quantum chemistry*, 2000.

[95] Alessandro Motta, Marie Pierre Gaigeot, and Dominique Costa. AIMD evidence of inner sphere adsorption of glycine on a stepped (101) boehmite AlOOH surface. *Journal of Physical Chemistry C*, 2012.

[96] Frederik Tielens, Vincent Humblot, and Claire Marie Pradier. Elucidation of the low coverage chiral adsorption assembly of l-lysine on Cu(1 1 0) surface: A theoretical study. *Surface Science*, 2008.

[97] Julian Schneider and Lucio Colombi Ciacchi. Specific material recognition by small peptides mediated by the interfacial solvent structure. *Journal of the American Chemical Society*, 2012.

[98] Arrigo Calzolari, Giancarlo Cicero, Carlo Cavazzoni, Rosa Di Felice, Alessandra Catellani, and Stefano Corni. Hydroxyl-rich β-sheet adhesion to the gold surface in water by first-principle simulations. *Journal of the American Chemical Society*, 2010.

[99] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95–99, 1963.

[100] O. Anatole von Lilienfeld, Raghunathan Ramakrishnan, Matthias Rupp, and Aaron Knoll. Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *International Journal of Quantum Chemistry*, 115(16):1084–1093, 2015.

[101] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[102] Stephen R. Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics*, 2015.

[103] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. PMID: 20426451.

[104] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108:058301, Jan 2012.

[105] Katja Hansen, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O. Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The Journal of Physical Chemistry Letters*, 6(12):2326–2331, 2015. PMID: 26113956.

[106] Haoyan Huo and Matthias Rupp. Unified representation of molecules and crystals for machine learning, 2017.

[107] Marcel F. Langer, Alex Goeßmann, and Matthias Rupp. Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning, 2020.

[108] Bing Huang and O. Anatole von Lilienfeld. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *The Journal of Chemical Physics*, 145(16):161102, 2016.

[109] Albert P. Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R. Kermode, Gábor Csányi, and Michele Ceriotti. Machine learning unifies the modeling of materials and molecules. *Science Advances*, 3(12), 2017.

[110] Sandip De, Felix Musil, Teresa Ingram, Carsten Baldauf, and Michele Ceriotti. Mapping and classifying molecules from a high-throughput structural database. *Journal of Cheminformatics*, 9(1):6, 2017.

[111] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87:184115, May 2013.

[112] M. Ceriotti, G. A. Tribello, and M. Parrinello. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences*, 108(32):13023–13028, 2011.

[113] Gareth A Tribello, Michele Ceriotti, and Michele Parrinello. Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 109(14):5196–5201, 2012.

[114] Michele Ceriotti, Gareth A Tribello, and Michele Parrinello. Demonstrating the Transferability and the Descriptive Power of Sketch-Map. *Journal of Chemical Theory and Computation*, 9(3):1521–1532, 2013.

[115] Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.*, 18:13754–13769, 2016.

[116] Lauri Himanen, Marc O.J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, and Adam S. Foster. Dscribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020.

[117] Felix Mayr and Alessio Gagliardi. Global property prediction: A benchmark study on open-source, perovskite-like datasets. *ACS Omega*, 6(19):12722–12732, 2021.

[118] Michele Ceriotti. Unsupervised machine learning in atomistic simulations, between predictions and understanding, 2019.

[119] Jerzy Leszczynski, Anna Kaczmarek-Kedziera, Tomasz Puzyn, Manthos G. Papadopoulos, Heribert Reis, and Manoj K. Shukla. *Handbook of computational chemistry*. 2017.

[120] Mario Barbatti, Matthias Ruckenbauer, Jaroslaw J. Szymczak, Adélia J.A. Aquino, and Hans Lischka. Nonadiabatic excited-state dynamics of polar π-systems and related model compounds of biological relevance. *Physical Chemistry Chemical Physics*, 2008.

[121] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

[122] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965.

[123] L. H. Thomas. The calculation of atomic fields. *Mathematical Proceedings of the Cambridge Philosophical Society*, 1927.

[124] Enrico Fermi. Statistical method to determine some properties of atoms. *Rend. Accad. Naz. Lincei*, 1927.

[125] M. Levy. Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem. *Proceedings of the National Academy of Sciences of the United States of America*, 1979.

[126] Matt Probert. Electronic Structure: Basic Theory and Practical Methods, by Richard M. Martin. *Contemporary Physics*, 2011.

[127] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A), 1965.

[128] D. M. Ceperley and B. J. Alder. Ground state of the electron gas by a stochastic method. *Physical Review Letters*, 1980.

[129] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 1996.

[130] Victor G. Ruiz, Wei Liu, and Alexandre Tkatchenko. Density-functional theory with screened van der Waals interactions applied to atomic and molecular adsorbates on close-packed and non-close-packed surfaces. *Physical Review B*, 93(3):035118, 2016.

[131] Axel D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 1993.

[132] Attila Szabo and Neil Ostlund. Szabo and Ostlund - Modern Quantum Chemistry, 1996.

[133] Susi Lehtola, Conrad Steigemann, Micael J.T. Oliveira, and Miguel A.L. Marques. Recent developments in LIBXC — A comprehensive library of functionals for density functional theory. *SoftwareX*, 2018.

[134] John P. Perdew. Jacob's ladder of density functional approximations for the exchange-correlation energy. 2003.

[135] Eberhard Engel and Reiner M. Dreizler. Density Functional Theory: An Advanced Course. *Theoretical and Mathematical Physics*, 2011.

[136] Matthias Ernzerhof and Gustavo E. Scuseria. Assessment of the Perdew-Burke-Ernzerhof exchange-correlation functional. *Journal of Chemical Physics*, 1999.

[137] Carlo Adamo and Vincenzo Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *Journal of Chemical Physics*, 1999.

[138] John P. Perdew and Yue Wang. Accurate and simple analytic representation of the electron-gas correlation energy. *Physical Review B*, 1992.

[139] Jorge Kohanoff. *Electronic structure calculations for solids and molecules: Theory and computational methods*. 2006.

[140] John P. Perdew, Matthias Ernzerhof, and Kieron Burke. Rationale for mixing exact exchange with density functional approximations. *Journal of Chemical Physics*, 1996.

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

[142] Robert L. Baldwin. Energetics of protein folding. *Journal of Molecular Biology*, 371(2):283–301, 2007.

[143] Robert A. DiStasio, O. Anatole von Lilienfeld, and Alexandre Tkatchenko. Collective many-body van der waals interactions in molecular systems. *Proceedings of the National Academy of Sciences*, 109(37):14791–14795, 2012.

[144] Mariana Rossi, Wei Fang, and Angelos Michaelides. Stability of complex biomolecular structures: van der waals, hydrogen bond cooperativity, and nuclear quantum effects. *The Journal of Physical Chemistry Letters*, 6(21):4233–4238, 2015. PMID: 26722963.

[145] Anthony M. Reilly, Richard I. Cooper, Claire S. Adjiman, Saswata Bhattacharya, A. Daniel Boese, Jan Gerit Brandenburg, Peter J. Bygrave, Rita Bylsma, Josh E. Campbell, Roberto Car, David H. Case, Renu Chadha, Jason C. Cole, Katherine Cosburn, Herma M. Cuppen, Farren Curtis, Graeme M. Day, Robert A. DiStasio Jr, Alexander Dzyabchenko, Bouke P. van Eijck, Dennis M. Elking, Joost A. van den Ende, Julio C. Facelli, Marta B. Ferraro, Laszlo Fusti-Molnar, Christina-Anna Gatsiou, Thomas S.

Gee, René de Gelder, Luca M. Ghiringhelli, Hitoshi Goto, Stefan Grimme, Rui Guo, Detlef W. M. Hofmann, Johannes Hoja, Rebecca K. Hylton, Luca Iuzzolino, Wojciech Jankiewicz, Daniël T. de Jong, John Kendrick, Niek J. J. de Klerk, Hsin-Yu Ko, Liudmila N. Kuleshova, Xiayue Li, Sanjaya Lohani, Frank J. J. Leusen, Albert M. Lund, Jian Lv, Yanming Ma, Noa Marom, Artëm E. Masunov, Patrick McCabe, David P. McMahon, Hugo Meekes, Michael P. Metz, Alston J. Misquitta, Sharmarke Mohamed, Bartomeu Monserrat, Richard J. Needs, Marcus A. Neumann, Jonas Nyman, Shigeaki Obata, Harald Oberhofer, Artem R. Oganov, Anita M. Orendt, Gabriel I. Pagola, Constantinos C. Pantelides, Chris J. Pickard, Rafal Podeszwa, Louise S. Price, Sarah L. Price, Angeles Pulido, Murray G. Read, Karsten Reuter, Elia Schneider, Christoph Schober, Gregory P. Shields, Pawanpreet Singh, Isaac J. Sugden, Krzysztof Szalewicz, Christopher R. Taylor, Alexandre Tkatchenko, Mark E. Tuckerman, Francesca Vacarro, Manolis Vasileiadis, Alvaro Vazquez-Mayagoitia, Leslie Vogt, Yanchao Wang, Rona E. Watson, Gilles A. de Wijs, Jack Yang, Qiang Zhu, and Colin R. Groom. Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallographica Section B*, 72(4):439–459, Aug 2016.

[146] Noa Marom, Robert A. DiStasio Jr., Viktor Atalla, Sergey Levchenko, Anthony M. Reilly, James R. Chelikowsky, Leslie Leiserowitz, and Alexandre Tkatchenko. Many-body dispersion interactions in molecular crystal polymorphism. *Angewandte Chemie International Edition*, 52(26):6629–6632, 2013.

[147] Javier Carrasco, Wei Liu, Angelos Michaelides, and Alexandre Tkatchenko. Insight into the description of van der waals forces for benzene adsorption on transition metal (111) surfaces. *The Journal of Chemical Physics*, 140(8):084704, 2014.

[148] Wei Liu, Friedrich Maaß, Martin Willenbockel, Christopher Bronner, Michael Schulze, Serguei Soubatch, F. Stefan Tautz, Petra Tegeder, and Alexandre Tkatchenko. Quantitative prediction of molecular adsorption: Structure and binding of benzene on coinage metals. *Phys. Rev. Lett.*, 115:036104, Jul 2015.

[149] Reinhard J. Maurer, Victor G. Ruiz, Javier Camarillo-Cisneros, Wei Liu, Nicola Ferri, Karsten Reuter, and Alexandre Tkatchenko. Adsorption structures and energetics of molecules on metal surfaces: Bridging experiment and theory. *Progress in Surface Science*, 91(2):72–100, 2016.

[150] Wei Liu, Victor G. Ruiz, Guo Xu Zhang, Biswajit Santra, Xinguo Ren, Matthias Scheffler, and Alexandre Tkatchenko. Structure and energetics of benzene adsorbed on transition-metal surfaces: Density-functional theory with van der Waals interactions including collective substrate response. *New Journal of Physics*, 2013.

[151] Wei Liu, Javier Carrasco, Biswajit Santra, Angelos Michaelides, Matthias Scheffler, and Alexandre Tkatchenko. Benzene adsorbed on metals: Concerted effect of covalency and van der Waals bonding. *Phys. Rev. B*, 86(24):245405, 2012.

[152] W A Al-Saidi, Haijun Feng, and Kristen A Fichthorn. Adsorption of Polyvinylpyrrolidone on Ag Surfaces: Insight into a Structure-Directing Agent. *Nano Letters*, 12(2):997–1001, 2012.

[153] Jan van Ruitenbeek. Dispersion forces unveiled. *Nature Materials*, 11:834–835, 2012.

[154] C. Wagner, N. Fournier, F. S. Tautz, and R. Temirov. Measurement of the binding energies of the organic-metal perylene-teracarboxylic-dianhydride/au(111) bonds by molecular manipulation using an atomic force microscope. *Phys. Rev. Lett.*, 109:076102, 2012.

[155] Jan Hermann and Alexandre Tkatchenko. Density-functional model for van der Waals interactions: Unifying atomic approaches with nonlocal functionals. *arXiv e-prints*, 2019.

[156] Jiří Klimeš and Angelos Michaelides. Perspective: Advances and challenges in treating van der waals dispersion forces in density functional theory. *The Journal of Chemical Physics*, 137(12):120901, 2012.

[157] Stefan Grimme. Accurate description of van der Waals complexes by density functional theory including empirical corrections. *Journal of Computational Chemistry*, 2004.

[158] Stefan Grimme. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of Computational Chemistry*, 2006.

[159] 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. *Journal of Chemical Physics*, 2010.

[160] Axel D. Becke and Erin R. Johnson. Exchange-hole dipole moment and the dispersion interaction. *Journal of Chemical Physics*, 2005.

[161] Axel D. Becke and Erin R. Johnson. A density-functional model of the dispersion interaction. *Journal of Chemical Physics*, 2005.

[162] Erin R. Johnson and Axel D. Becke. A post-Hartree-Fock model of intermolecular interactions. *Journal of Chemical Physics*, 2005.

[163] Alexandre Tkatchenko and Matthias Scheffler. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Physical Review Letters*, 2009.

[164] Takeshi Sato and Hiromi Nakai. Density functional method including weak interactions: Dispersion coefficients based on the local response approximation. *Journal of Chemical Physics*, 2009.

[165] John F. Dobson, Angela White, and Angel Rubio. Asymptotics of the dispersion interaction: Analytic benchmarks for van der Waals energy functionals. *Physical Review Letters*, 2006.

[166] Marcus Elstner, Pavel Hobza, Thomas Frauenheim, Sándor Suhai, and Efthimios Kaxiras. Hydrogen bonding and stacking interactions of nucleic acid base pairs: A density functional-theory based treatment. *Journal of Chemical Physics*, 2001.

[167] Axel D. Becke and Erin R. Johnson. Exchange-hole dipole moment and the dispersion interaction revisited. *Journal of Chemical Physics*, 2007.

[168] Petr Jurečka, Jiří Černý, Pavel Hobza, and Dennis R. Salahub. Density functional theory augmented with an empirical dispersion term. Interaction energies and geometries of 80 noncovalent complexes compared with ab initio quantum mechanics calculations. *Journal of Computational Chemistry*, 2007.

[169] Urs Zimmerli, Michele Parrinello, and Petros Koumoutsakos. Dispersion corrections to density functionals for water aromatic interactions. *Journal of Chemical Physics*, 2004.

[170] Marcus A. Neumann and Marc Antoine Perrin. Energy ranking of molecular crystals using density functional theory calculations and an empirical van der waals correction. *Journal of Physical Chemistry B*, 2005.

[171] H. B. G. Casimir and D. Polder. The influence of retardation on the london-van der waals forces. *Phys. Rev.*, 73:360–372, Feb 1948.

[172] K. T. Tang and M. Karplus. Padé-approximant calculation of the nonretarded van der waals coefficients for two and three helium atoms. *Phys. Rev.*, 171:70–74, Jul 1968.

[173] X. Chu and A. Dalgarno. Linear response time-dependent density functional theory for van der waals coefficients. *The Journal of Chemical Physics*, 121(9):4083–4088, 2004.

[174] Tore Brinck, Jane S. Murray, and Peter Politzer. Polarizability and volume. *The Journal of Chemical Physics*, 98(5):4305–4306, 1993.

[175] F. L. Hirshfeld. Bonded-atom fragments for describing molecular charge densities. *Theoretica Chimica Acta*, 1977.

[176] Petr Jurečka, Jiří Šponer, Jiří Černý, 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. *Physical Chemistry Chemical Physics*, 2006.

[177] Victor G. Ruiz, Wei Liu, Egbert Zojer, Matthias Scheffler, and Alexandre Tkatchenko. Density-functional theory with screened van der Waals interactions for the modeling of hybrid inorganic-organic systems. *Physical Review Letters*, 2012.

[178] E M Lifshitz. The theory of molecular attractive forces between solids. *Journal of Experimental and Theoretical Physics*, 1956.

[179] E. Zaremba and W. Kohn. Van der Waals interaction between an atom and a solid surface. *Physical Review B*, 1976.

[180] Guo Xu Zhang, Alexandre Tkatchenko, Joachim Paier, Heiko Appel, and Matthias Scheffler. Van der Waals interactions in ionic and semiconductor solids. *Physical Review Letters*, 2011.

[181] Todd Raeker. Physical Adsorption: Forces and Phenomena (Bruch, L.W.; Cole, Milton W.; Zaremba, Eugene). *Journal of Chemical Education*, 1998.

[182] Wolfgang S.M. Werner, Kathrin Glantschnig, and Claudia Ambrosch-Draxl. Optical constants and inelastic electron-scattering data for 17 elemental metals. *Journal of Physical and Chemical Reference Data*, 2009.

[183] 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. *Computer Physics Communications*, 180:2175–2196, 2009.

[184] V. Havu, V. Blum, P. Havu, and M. Scheffler. Efficient o(n) integration for all-electron electronic structure calculation using numeric basis functions. *Journal of Computational Physics*, 228(22):8367–8379, 2009.

[185] Igor Ying Zhang, Xinguo Ren, Patrick Rinke, Volker Blum, and Matthias Scheffler. Numeric atom-centered-orbital basis sets with valence-correlation consistency from H to Ar. *New Journal of Physics*, 2013.

[186] Jörg Neugebauer and Matthias Scheffler. Adsorbate-substrate and adsorbate-adsorbate interactions of Na and K adlayers on Al(111). *Physical Review B*, 1992.

[187] Norina A. Richter, Sabrina Sicolo, Sergey V. Levchenko, Joachim Sauer, and Matthias Scheffler. Concentration of vacancies at metal-oxide surfaces: Case study of MgO(100). *Physical Review Letters*, 2013.

[188] Daniel Berger, Andrew J. Logsdail, Harald Oberhofer, Matthew R. Farrow, C. Richard A. Catlow, Paul Sherwood, Alexey A. Sokol, Volker Blum, and Karsten Reuter. Embedded-cluster calculations in a numeric atomic orbital density-functional theory framework. *Journal of Chemical Physics*, 2014.

[189] Wolfram Steurer, Shadi Fatayer, Leo Gross, and Gerhard Meyer. Probe-based measurement of lateral single-electron transfer between individual molecules. *Nature Communications*, 2015.

[190] Daniel Hernangómez-Pérez, Jakob Schlör, David A. Egger, Laerte L. Patera, Jascha Repp, and Ferdinand Evers. Reorganization energy and polaronic effects of pentacene on NaCl films. *Physical Review B*, 2020.

[191] Hannu Pekka Komsa and Alfredo Pasquarello. Finite-size supercell correction for charged defects at surfaces and interfaces. *Physical Review Letters*, 2013.

[192] Ismaila Dabo, Boris Kozinsky, Nicholas E. Singh-Miller, and Nicola Marzari. Electrostatics in periodic boundary conditions and real-space corrections. *Physical Review B - Condensed Matter and Materials Physics*, 2008.

[193] M. Otani and O. Sugino. First-principles calculations of charged surfaces and interfaces: A plane-wave nonrepeated slab approach. *Physical Review B - Condensed Matter and Materials Physics*, 2006.

[194] Ofer Sinai, Oliver T. Hofmann, Patrick Rinke, Matthias Scheffler, Georg Heimel, and Leeor Kronik. Multiscale approach to the electronic structure of doped semiconductor surfaces. *Physical Review B - Condensed Matter and Materials Physics*, 2015.

[195] Christoph Freysoldt, Arpit Mishra, Michael Ashton, and Jörg Neugebauer. Generalized dipole correction for charged surfaces in the repeated-slab approach. *Physical Review B*, 2020.

[196] J. Bardeen. Tunnelling from a many-particle point of view. *Physical Review Letters*, 1961.

[197] J. Tersoff and D. R. Hamann. Theory and application for the scanning tunneling microscope. *Phys. Rev. Lett.*, 50:1998–2001, Jun 1983.

[198] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B*, 1998.

[199] Hendrik Heinz, Barry L. Farmer, Ras B. Pandey, Joseph M. Slocik, Soumya S. Patnaik, Ruth Pachter, and Rajesh R. Naik. Nature of molecular interactions of peptides with gold, palladium, and Pd-Au bimetal surfaces in aqueous solution. *Journal of the American Chemical Society*, 2009.

[200] Jie Feng, Ras B. Pandey, Rajiv J. Berry, Barry L. Farmer, Rajesh R. Naik, and Hendrik Heinz. Adsorption mechanism of single amino acid and surfactant molecules to Au {111} surfaces in aqueous solution: Design rules for metal-binding molecules. *Soft Matter*, 2011.

[201] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with namd. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[202] 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. *Journal of Chemical Information and Modeling*, 2005.

[203] David H. Wolpert and William G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1997.

[204] Ruxi Qi, Guanghong Wei, Buyong Ma, and Ruth Nussinov. Replica exchange molecular dynamics: A practical application protocol with solutions to common problems and a peptide aggregation and self-assembly example. In *Methods in Molecular Biology*. 2018.

[205] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 1977.

[206] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. THE weighted histogram analysis method for free[U+2010]energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 1992.

[207] Christian Bartels. Analyzing biased Monte Carlo and molecular dynamics simulations. *Chemical Physics Letters*, 2000.

[208] Erik G. Brandt and Alexander P. Lyubartsev. Molecular Dynamics Simulations of Adsorption of Amino Acid Side Chain Analogues and a Titanium Binding Peptide on the TiO2 (100) Surface. *Journal of Physical Chemistry C*, 2015.

[209] Daan Frenkel, Berend Smit, Jan Tobochnik, Susan R. McKay, and Wolfgang Christian. Understanding Molecular Simulation. *Computers in Physics*, 1997.

[210] Martin Hoefling, Francesco Iori, Stefano Corni, and Kay Eberhard Gottschalk. The conformations of amino acids on a gold(111) surface. *ChemPhysChem*, 2010.

[211] Zdenek Futera. Amino-acid interactions with the Au(111) surface: Adsorption, band alignment, and interfacial electronic coupling. *Physical Chemistry Chemical Physics*, 2021.

[212] Carsten Baldauf and Mariana Rossi. Going clean: Structure and dynamics of peptides in the gas phase and paths to solvation, 2015.

[213] Hendrik Heinz and Hadi Ramezani-Dakhel. Simulations of inorganic–bioorganic interfaces to discover new materials: insights, comparisons to experiment, challenges, and opportunities. *Chem. Soc. Rev.*, 45(2):412–448, 2016.

[214] Chris J. Pickard and R. J. Needs. High-pressure phases of silane. *Physical Review Letters*, 2006.

[215] Georg Schusteritsch and Chris J. Pickard. Predicting interface structures: From SrTiO3 to graphene. *Physical Review B - Condensed Matter and Materials Physics*, 2014.

[216] Miri Zilka, Dmytro V. Dudenko, Colan E. Hughes, P. Andrew Williams, Simone Sturniolo, W. Trent Franks, Chris J. Pickard, Jonathan R. Yates, Kenneth D.M. Harris, and Steven P. Brown. Ab initio random structure searching of organic molecular solids: Assessment and validation against experimental data. *Physical Chemistry Chemical Physics*, 2017.

[217] Sandro E Schönborn, Stefan Goedecker, Shantanu Roy, and Artem R Oganov. The performance of minima hopping and evolutionary algorithms for cluster structure prediction. *The Journal of Chemical Physics*, 130(14):144108, 2009.

[218] Thomas Bäck. *Evolutionary Algorithms in Theory and Practice*. 1996.

[219] Bernd Hartke. Efficient Global Geometry Optimization of Atomic and Molecular Clusters. In *Global Optimization*. 2006.

[220] Adriana Supady, Volker Blum, and Carsten Baldauf. First-Principles Molecular Structure Search with a Genetic Algorithm. *Journal of Chemical Information and Modeling*, 55(11):2338–2348, 2015.

[221] Farren Curtis, Xiayue Li, Timothy Rose, Álvaro Vázquez-Mayagoitia, Saswata Bhattacharya, Luca M. Ghiringhelli, and Noa Marom. GAtor: A First-Principles Genetic Algorithm for Molecular Crystal Structure Prediction. *Journal of Chemical Theory and Computation*, 2018.

[222] 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. *Journal of Physical Chemistry A*, 1997.

[223] Stefan Goedecker. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *Journal of Chemical Physics*, 2004.

[224] Konstantin Krautgasser, Chiara Panosetti, Dennis Palagin, Karsten Reuter, and Reinhard J. Maurer. Global structure search for molecules on surfaces: Efficient sampling with curvilinear coordinates. *Journal of Chemical Physics*, 2016.

[225] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 2010.

[226] J. Behler. Representing potential energy surfaces by high-dimensional neural network potentials, 2014.

[227] Milica Todorović, Michael U. Gutmann, Jukka Corander, and Patrick Rinke. Bayesian inference of atomistic structure in functional materials. *npj Computational Materials*, 2019.

[228] Jari Järvi, Patrick Rinke, and Milica Todorović. Detecting stable adsorbates of (1 S)-camphor on Cu(111) with Bayesian optimization . *Beilstein Journal of Nanotechnology*, 2020.

[229] Jari Järvi, Benjamin Alldritt, Ondřej Krejčí, Milica Todorović, Peter Liljeroth, and Patrick Rinke. Integrating Bayesian Inference with Scanning Probe Experiments for Robust Identification of Surface Adsorbate Configurations. *Advanced Functional Materials*, 2021.

[230] Lincan Fang, Esko Makkonen, Milica Todorović, Patrick Rinke, and Xi Chen. Efficient Amino Acid Conformer Search with Bayesian Optimization. *Journal of Chemical Theory and Computation*, 2021.

[231] Jorge Nocedal and Stephen J. Wright. *Numerical optimization 2nd edition*. 2000.

[232] C. G. Broyden. Quasi-Newton methods and their application to function minimisation. *Mathematics of Computation*, 1967.

[233] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 1970.

[234] Oliver T. Hofmann, Egbert Zojer, Lukas Hörmann, Andreas Jeindl, and Reinhard J. Maurer. First-principles calculations of hybrid inorganic-organic interfaces: From state-of-the-art to best practice, 2021.

[235] Letif Mones, Christoph Ortner, and Gábor Csányi. Preconditioners for the geometry optimisation and saddle point search of molecular systems. *Scientific Reports*, 2018.

[236] Roland Lindh, Anders Bernhardsson, Gunnar Karlström, and Per Åke Malmqvist. On the use of a hessian model function in molecular geometry optimizations. *Chemical Physics Letters*, 241(4):423 – 428, 1995.

[237] David Packwood, James Kermode, Letif Mones, Noam Bernstein, John Woolley, Nicholas Gould, Christoph Ortner, and Gábor Csányi. A universal preconditioner for simulating condensed phase materials. *Journal of Chemical Physics*, 2016.

[238] Andrew A. Peterson. Acceleration of saddle-point searches with machine learning. *Journal of Chemical Physics*, 2016.

[239] Olli Pekka Koistinen, Freyja B. Dagbjartsdóttir, Vilhjálmur Ásgeirsson, Aki Vehtari, and Hannes Jónsson. Nudged elastic band calculations accelerated with Gaussian process regression. *Journal of Chemical Physics*, 2017.

[240] José A. Garrido Torres, Paul C. Jennings, Martin H. Hansen, Jacob R. Boes, and Thomas Bligaard. Low-Scaling Algorithm for Nudged Elastic Band Calculations Using a Surrogate Machine Learning Model. *Physical Review Letters*, 2019.

[241] Nongnuch Artrith and Jörg Behler. High-dimensional neural network potentials for metal surfaces: A prototype study for copper. *Physical Review B - Condensed Matter and Materials Physics*, 2012.

[242] Estefanía Garijo Del Río, Jens Jørgen Mortensen, and Karsten Wedel Jacobsen. Local Bayesian optimizer for atomic structures. *Physical Review B*, 2019.

[243] Estefanía Garijo Del Río, Sami Kaappa, José A. Garrido Torres, Thomas Bligaard, and Karsten Wedel Jacobsen. Machine learning with bond information for local structure optimizations in surface science. *The Journal of chemical physics*, 2020.

[244] Muhammed Shuaibi, Saurabh Sivakumar, Rui Qi Chen, and Zachary W Ulissi. Enabling robust offline active learning for machine learning potentials using simple physics-based priors. *Machine Learning: Science and Technology*, 2021.

[245] Ryosuke Jinnouchi, Kazutoshi Miwa, Ferenc Karsai, Georg Kresse, and Ryoji Asahi. On-the-Fly Active Learning of Interatomic Potentials for Large-Scale Atomistic Simulations, 2020.

[246] WOLFE P. CONVERGENCE CONDITIONS FOR ASCENT METHODS. *SIAM Review*, 1969.

[247] Philip Wolfe. Convergence Conditions for Ascent Methods. II: Some Corrections. *SIAM Review*, 1971.

[248] Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 1966.

[249] David Packwood, James Kermode, Letif Mones, Noam Bernstein, John Woolley, Nicholas Gould, Christoph Ortner, and Gábor Csányi. A universal preconditioner for simulating condensed phase materials. *The Journal of Chemical Physics*, 144(16):164109, 2016.

[250] Thomas H. Fischer and Jan Almlöf. General methods for geometry and wave function optimization. *Journal of Physical Chemistry*, 1992.

[251] Geza Fogarasi, Xuefeng Zhou, Patterson W. Taylor, and Peter Pulay. The Calculation of ab Initio Molecular Geometries: Efficient Optimization by Natural Internal Coordinates and Empirical Correction by Offset Forces. *Journal of the American Chemical Society*, 1992.

[252] Letif Mones, Christoph Ortner, and Gábor Csányi. Preconditioners for the geometry optimisation and saddle point search of molecular systems. *Scientific Reports*, 2018.

[253] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 1955.

[254] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[255] P. W.A. Howe. Principal components analysis of protein structure ensembles calculated using NMR data. *Journal of Biomolecular NMR*, 2001.

[256] Lee Wei Yang, Eran Eyal, Ivet Bahar, and Akio Kitao. Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): Insights into functional dynamics. *Bioinformatics*, 2009.

[257] Katrijn Van Deun, Elise A.V. Crompvoets, and Eva Ceulemans. Obtaining insights from high-dimensional data: Sparse principal covariates regression. *BMC Bioinformatics*, 2018.

[258] Andrea Anelli, Edgar A. Engel, Chris J. Pickard, and Michele Ceriotti. Generalized convex hull construction for materials discovery. *Physical Review Materials*, 2018.

[259] Mazen Ahmad, Volkhard Helms, Olga V. Kalinina, and Thomas Lengauer. Relative Principal Components Analysis: Application to Analyzing Biomolecular Conformational Changes. *Journal of Chemical Theory and Computation*, 2019.

[260] Benjamin A Helfrecht, Rose K Cersonsky, Guillaume Fraux, and Michele Ceriotti. Structure-property maps with Kernel principal covariates regression. *Machine Learning: Science and Technology*, 2020.

[261] Piero Gasparotto, Maria Fischer, Daniele Scopece, Maciej O. Liedke, Maik Butterling, Andreas Wagner, Oguz Yildirim, Mathis Trant, Daniele Passerone, Hans J. Hug, and Carlo A. Pignedoli. Mapping the Structure of Oxygen-Doped Wurtzite Aluminum Nitride Coatings from Ab Initio Random Structure Search and Experiments. *ACS Applied Materials and Interfaces*, 2021.

[262] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000.

[263] Vin De Silva and Joshua B Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, 2004.

[264] Laurens Van Der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008.

[265] Gareth A. Tribello and Piero Gasparotto. Using Data-Reduction Techniques to Analyze Biomolecular Trajectories. In *Methods in Molecular Biology*. 2019.

[266] Lauri Himanen, Amber Geurts, Adam Stuart Foster, and Patrick Rinke. Data-Driven Materials Science: Status, Challenges, and Perspectives, 2019.

[267] Teng Zhou, Zhen Song, and Kai Sundmacher. Big Data Creates New Opportunities for Materials Research: A Review on Methods and Applications of Machine Learning for Materials Design, 2019.

[268] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[269] E. Mateo Marti, Ch Methivier, P. Dubot, and C. M. Pradier. Adsorption of (S)-histidine on Cu(110) and oxygen-covered Cu(110), a combined Fourier transform reflection absorption infrared spectroscopy and force field calculation study. *Journal of Physical Chemistry B*, 2003.

[270] Luiza Buimaga-Iarinca, Calin G. Floare, Adrian Calborean, and Ioan Turcu. DFT study on cysteine adsorption mechanism on Au(111) and Au(110). In *AIP Conference Proceedings*, 2013.

[271] S. Blankenburg and W. G. Schmidt. Glutamic acid adsorbed on Ag(110): Direct and indirect molecular interactions. *Journal of Physics Condensed Matter*, 2009.

[272] Tanja Deckert-Gaudig, Eva Rauls, and Volker Deckert. Aromatic amino acid monolayers sandwiched between gold and silver: A combined tip-enhanced Raman and theoretical approach. *Journal of Physical Chemistry C*, 2010.

[273] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Physical Review Letters*, 78(7):1396–1396, 1997.

[274] Wheeler P. Davey. Precision measurements of the lattice constants of twelve common metals. *Physical Review*, 1925.

[275] Philipp Haas, Fabien Tran, and Peter Blaha. Calculation of the lattice constant of solids with semilocal functionals. *Phys. Rev. B*, 79(8):85104, 2009.

[276] J.H. van Lenthe, S. Faas, and J.G. Snijders. Gradients in the ab initio scalar zeroth-order regular approximation (zora) approach. *Chemical Physics Letters*, 328(1):107–112, 2000.

[277] Christoph van Wüllen. Molecular density functional calculations in the regular relativistic approximation: Method, application to coinage metal diatomics, hydrides, fluorides and chlorides, and comparison with first-order relativistic calculations. *The Journal of Chemical Physics*, 109(2):392–399, 1998.

[278] Matti Ropo, Volker Blum, and Carsten Baldauf. Trends for isolated amino acids and dipeptides: Conformation, divalent ion binding, and remarkable similarity of binding to calcium and lead. *Scientific Reports*, 6:35772, 2016.

[279] Michele Ceriotti, Sandip De, and Felix Musil. Glosim package. Code assessed in 2020-01-01.

[280] Michael B. Bolger. Chapter 9–computational techniques in macromolecular structural analysis. In Jay A. Glasel, Murray P. Deutscher, and Murray P. Deutscher, editors, *Introduction to Biophysical Methods for Protein and Nucleic Acid Research*, pages 433–490. Academic Press, San Diego, 1995.

[281] Donald A. McQuarrie. *Statistical Mechanics*. University Science Books, 2000.

[282] Brent Fultz. Vibrational thermodynamics of materials. *Progress in Materials Science*, 55(4):247–352, 2010.

[283] A Togo and I Tanaka. First principles phonon calculations in materials science. *Scr. Mater.*, 108:1–5, 2015.

[284] Karen Fidanyan. Development version of phonopy. Accessed: 2019-03-01.

[285] Mariana Rossi, Matthias Scheffler, and Volker Blum. Impact of Vibrational Entropy on the Stability of Unsolvated Peptide Helices with Increasing Length. *Journal Of Physical Chemistry B*, 117(18):5574–5584, 2013.

[286] Franziska Schubert, Mariana Rossi, Carsten Baldauf, Kevin Pagel, Stephan Warnke, Gert von Helden, Frank Filsinger, Peter Kupser, Gerard Meijer, Mario Salwiczek, Beate Koksch, Matthias Scheffler, and Volker Blum. Exploring the conformational preferences of 20-residue peptides in isolation: Ac-Ala 19 -Lys + H + vs. Ac-Lys-Ala 19 + H + and the current reach of DFT . *Physical Chemistry Chemical Physics*, 17(11):7373–7385, 2015.

[287] David A Egger, Zhen-Fei Liu, Jeffrey B Neaton, and Leeor Kronik. Reliable Energy Level Alignment at Physisorbed Molecule-Metal Interfaces from Density Functional Theory. *Nano Letters*, 15:2448–2455, 2015.

[288] Zhen-Fei Liu, David A Egger, Sivan Refaely-Abramson, Leeor Kronik, and Jeffrey B Neaton. Energy level alignment at molecule-metal interfaces from an optimally tuned range-separated hybrid functional. *The Journal of Chemical Physics*, 146(9):092326, February 2017.

[289] S.M Barlow, K.J Kitching, S Haq, and N.V Richardson. A study of glycine adsorption on a cu110 surface using reflection absorption infrared spectroscopy. *Surface Science*, 401(3):322–335, 1998.

[290] Susan M. Barlow, Souheila Louafi, Delphine Le Roux, Jamie Williams, Christopher Muryn, Sam Haq, and Rasmita Raval. Supramolecular assembly of strongly chemisorbed size- and shape-defined chiral clusters: S- and r-alanine on cu(110). *Langmuir*, 20(17):7171–7176, 2004.

[291] Christophe Méthivier, Vincent Humblot, and Claire-Marie Pradier. l-methionine adsorption on cu(110), binding and geometry of the amino acid as a function of coverage. *Surface Science*, 632:88–92, 2015.

[292] E Mateo Marti, S.M Barlow, S Haq, and R Raval. Bonding and assembly of the chiral amino acid s-proline on cu(110): the influence of structural rigidity. *Surface Science*, 501(3):191–202, 2002.

[293] E. Mateo Marti, Alida Quash, Ch. Methivier, P. Dubot, and C.M. Pradier. Interaction of s-histidine, an amino acid, with copper and gold surfaces, a comparison based on rairs analyses. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 249(1):85–89, 2004.

[294] Tugce Eralp, Andrey Shavorskiy, Zhasmina V. Zheleva, Georg Held, Nataliya Kalashnyk, Yanxiao Ning, and Trolle R. Linderoth. Global and local expression of chirality in serine on the cu110 surface. *Langmuir*, 26(24):18841–18851, 2010.

[295] Dmitrii Maksimov, Carsten Baldauf, and Mariana Rossi. Database of arg and arg-h^+ adsorbed on cu(111), ag(111) and au(111) in the NOMAD repository.

[296] Ziheng Lu, Bonan Zhu, Benjamin W.B. Shires, David O. Scanlon, and Chris J. Pickard. Ab initio random structure searching for battery cathode materials. *Journal of Chemical Physics*, 2021.

[297] Ask Hjorth Larsen, Jens JØrgen Mortensen, Jakob Blomqvist, Ivano E. Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N. Groves, BjØrk Hammer, Cory Hargus, Eric D. Hermes, Paul C. Jennings, Peter Bjerre Jensen, James Kermode, John R. Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob SchiØtz, Ole Schütt, Mikkel Strange, Kristian S. Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W. Jacobsen. The atomic simulation environment - A Python library for working with atoms, 2017.

[298] Encyclopedia of physical science and technology. *Choice Reviews Online*, 2002.

[299] Vincent Frappier, Madeleine Duran, and Amy E. Keating. Erratum to: PixelDB: Protein-peptide complexes annotated with structural conservation of the peptide binding mode: PixelDB: Protein—Peptide Complexes (Protein Science, (2018), 27, 1, (276-285), 10.1002/pro.3320), 2018.

[300] Donald B. Johnson. Finding All the Elementary Circuits of a Directed Graph. *SIAM Journal on Computing*, 1975.

[301] Charles F.F. Karney. Quaternions in molecular modeling. *Journal of Molecular Graphics and Modelling*, 25(5):595 – 604, 2007.

[302] Soohyung Park, Haiyuan Wang, Thorsten Schultz, Dongguen Shin, Ruslan Ovsyannikov, Marios Zacharias, Dmitrii Maksimov, Matthias Meissner, Yuri Hasegawa, Takuma Yamaguchi, Satoshi Kera, Areej Aljarb, Mariam Hakami, Lain Jong Li, Vincent Tung, Patrick Amsalem, Mariana Rossi, and Norbert Koch. Temperature-Dependent Electronic Ground-State Charge Transfer in van der Waals Heterostructures. *Advanced Materials*, 2021.

[303] Lincan Fang, Esko Makkonen, Milica Todorovic, Patrick Rinke, and Xi Chen. Efficient cysteine conformer search with bayesian optimization, 2020.

[304] Letif Mones, Christoph Ortner, and Gábor Csányi. Preconditioners for the geometry optimisation and saddle point search of molecular systems. *Scientific Reports*, 2018.

[305] X. Chu and A. Dalgarno. Linear response time-dependent density functional theory for van der waals coefficients. *The Journal of Chemical Physics*, 121(9):4083–4088, 2004.

[306] J Mitroy, M S Safronova, and Charles W Clark. Theory and applications of atomic and ionic polarizabilities. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 43(20):202001, oct 2010.

[307] Database of lennard-jones clusters. http://doye.chem.ox.ac.uk/jon/structures/LJ.html.

[308] K. W. Jacobsen, P. Stoltze, and J. K. Nørskov. A semi-empirical effective medium theory for metals and alloys. *Surface Science*, 1996.

[309] Steven J. Stuart, Alan B. Tutein, and Judith A. Harrison. A reactive potential for hydrocarbons with intermolecular interactions. *Journal of Chemical Physics*, 2000.

[310] Donald W. Brenner, Olga A. Shenderova, Judith A. Harrison, Steven J. Stuart, Boris Ni, and Susan B. Sinnott. A second-generation reactive empirical bond order (REBO) potential energy expression for hydrocarbons. *Journal of Physics Condensed Matter*, 2002.

[311] Murray S. Daw and M. I. Baskes. Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals. *Physical Review B*, 1984.

[312] Murray S. Daw, Stephen M. Foiles, and Michael I. Baskes. The embedded-atom method: a review of theory and applications, 1993.

[313] I. Stensgaard. Adsorption of di-L-alanine on Cu(1 1 0) investigated with scanning tunneling microscopy. *Surface Science*, 2003.

[314] S. M. Barlow, S. Haq, and R. Raval. Bonding, organization, and dynamical growth behavior of tripeptides on a defined metal surface: Tri-L-alanine and Tri-L-leucine on Cu{100}. *Langmuir*, 2001.

Dmitrii Maksimov

Github: maksimovdmitrii
LinkedIn: dmitrii-maksimov-7ab0a887
Google Scholar: Dmitrii Maksimov
ResearchGate: Dmitrii_Maksimov
ORCID: 0000-0003-4448-8848
ScopusID: 56850160300
ResearcherID: M-7271-2013

Education

- **ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE (EPFL)** — Lausanne, Switzerland
  PhD Materials Science and Engineering — *June 2018 – July 2022*
- **St. Petersburg State University** — St. Petersburg, Russia
  Master of Physics - Biophysics — *September 2013 – July 2015*
- **Ural Federal University** — Yekaterinburg, Russia
  Bachelor of Physics - Theoretical Physics — *September 2009 – July 2013*

Working experience

- **Max Planck Institute for the Structure and Dynamics of Matter** — Hamburg, Germany
  Visiting researcher — *February 2020 – July 2022*
- **Fritz Haber Institute of the Max Planck Society** — Berlin, Germany
  Doctoral student — *February 2017 – July 2022*
- **St. Petersburg State University** — St. Petersburg, Russia
  Researcher — *May 2014 – January 2017*

Projects

- **Generation and Search (GenSec) package**: Created from scratch open-source structure search code for organic/inorganic hybrid materials. The package is designed to unify many electronic structure programs and produce material databases in a unified format suitable for further machine learning projects and training of the empirical potentials. One of the package's strengths is the parallel utilizing of resources on high-performance computational platforms. Tech: Python, ASE (July 2021 - Present time). Publication is in preparation.
- **Structure search of organic molecules adsorbed on vacancies of MoS_2 surface**: Preparation of candidate structures with GenSec followed by *ab initio* molecular dynamics simulation. Expected theoretical results should correspond to experimental scanning tunelling microscope (STM) images (Jan 2022 - Present time). Publication is in preparation.
- **Structure search of triglycine molecule with different protonation states**: Different conformers with different protonation states of triglycine molecule in isolation were investigated using *ab initio* structure search. Modelled theoretical properties for low-energy conformers are investigated corresponding to experimental near edge X-ray absorption fine structure (NEXAFS) spectra (September 2021 - Present time). Publication is in preparation.
- **Temperature-Dependent Electronic Ground-State Charge Transfer in van der Waals Heterostructures**: Investigation of self-assembly candidates of F4-TCNQ and F6-TCNQ molecules adsorbed on MoS_2 surface (July 2020 – July 2021).
- **The conformational space of a flexible amino acid at metallic surfaces**: Generation of the *ab initio* (PBE+vdW^{surf}) database of the Arg and Arg-H^+ molecules adsorbed on different metallic surfaces and analysis of the structure-property relationships employing unsupervised dimensionality reduction techniques (multidimensional scaling) using smooth overlap of atomic positions kernel molecular descriptors (March 2018 – June 2020).
- **Investigation of stacking geometries of nucleobases and their complexes with metal clusters**: Analysis of available databases and investigation of excited-state properties of nucleobases complexes with metal clusters (March 2015 – December 2016).

Publications

Temperature-Dependent Electronic Ground-State Charge Transfer in van der Waals Heterostructures, S. Park et al., Advanced Materials 33 (29), 2008677, (2021)
The conformational space of a flexible amino acid at metallic surfaces, D. Maksimov, C. Baldauf, M. Rossi, International Journal of Quantum Chemistry 121 (3), e26284, (2021)
Excitation spectra of Ag3–DNA bases complexes: a benchmark study, D. Maksimov, V. Pomogaev, A. Kononov, Chemical Physics Letters 673, 11-18, (2017)
Ag–DNA emitter: metal nanorod or supramolecular complex?, R. Ramazanov et al., The Journal of Physical Chemistry Letters 7 (18), 3560-3566, (2016)
Noncanonical stacking geometries of nucleobases as a preferred target for solar radiation, R. Ramazanov, D. Maksimov, A. Kononov, Journal of the American Chemical Society 137 (36), 11656-11665, (2015)

Skills Summary

- **Languages**: Python, Bash
- **Frameworks**: Scikit, Matplotlib, Pandas, Plotly
- **Tools**: GIT, ASE, PyMol, Blender, Jmol
- **Electronic structure packages**: FHI-aims, ORCA, Turbomole, GAMESS, Firefly

Teaching

- **Summer schools tutoring** 2017 – 2019
  Tutor at workshops
  - **Hands-on Workshop Density-Functional Theory and Beyond: Accuracy, Efficiency and Reproducibility in Computational Materials Science**: Humboldt University, Berlin, Germany, July 31 to August 11, 2017
  - **Hands-on DFT and Beyond: Frontiers of advanced electronic structure and Molecular Dynamics Methods**: Peking University, Beijing, China, July 30th to August 10th, 2018
  - **Hands-on DFT and Beyond: High-Throughput Screening and Big-Data Analytics, Towards Exascale Computational Materials Science**: University of Barcelona, Barcelona, Spain, August 26th to September 6th, 2019

Characterization and prediction of peptide structures on inorganic surfaces

THIS IS A TEMPORARY TITLE PAGE
It will be replaced for the final print by a version
provided by the registrar's office.

Thèse n. 1234 2022
présentée le 23 août 2022
à la Faculté des sciences de base
laboratoire SuperScience
programme doctoral en SuperScience
École polytechnique fédérale de Lausanne

pour l'obtention du grade de Docteur ès Sciences
par

Dmitrii Maksimov

acceptée sur proposition du jury :

Prof Name Surname, président du jury
Prof Name Surname, directeur de thèse
Prof Name Surname, rapporteur
Prof Name Surname, rapporteur
Prof Name Surname, rapporteur

Lausanne, EPFL, 2022

To my family, friends and colleagues.

Acknowledgements

First and foremost, I owe a debt of gratitude to Dr. Mariana Rossi, for whom I am eternally grateful. She is an amazing group leader and a humble and kind individual who has been instrumental in guiding me through the often difficult times that come with research and writing a dissertation.
Professor Dr. Michele Ceriotti, who served as my thesis director and allowed me to get practical expertise in machine learning methods deserves a special thank you for his guidance and support.
I would like to express my gratitude to Dr. Carsten Baldauf for his early research oversight.
I would like to express my gratitude to Prof. Dr. Matthias Scheffler for his intriguing study at the Fritz Haber Institute's Theory Department, which he is conducting.
I would like to express my gratitude to the Max Planck-EPFL Center for Molecular Nanoscience and Technology – particularly to Prof. Dr. Klaus Kern and Dr. Klaus Kuhnke – for their support in our collaborative research effort.
Because of the people that worked there, the Fritz Haber Institute's Theory Department has always enjoyed a pleasant working environment. We owe a debt of appreciation to far too many people, including (but not limited to) Karen and Florian; Marcel and Marcin; Xiajuan; Markus; Maria; Christian and Sebastian; Björn and Hanna; Luca and Majid; Sebastian and Henrik and Sebastian; Alan and Alaa; and Yair.
Special thanks to my friends who were with me during my time in Europe: Valeria Volodkina, Evgenii Ikonnikov, Stanislav Podchezerzev and Alexandr Sukhanov.

Hamburg, August 23, 2022 D. M.

Abstract

Interfaces between peptides and metallic surfaces are the subject of great interest for possible use in technological and medicinal applications, mainly since organic systems present an extensive range of functionalities, are abundant, cheap, and exhibit low toxicity. *Exemplary applications* are biosensors that may be sensitive to specific metabolites or harmful compounds. However, these hybrid interfaces pose a challenge to computational modelling, particularly regarding predicting the most relevant configurations at the surface, which determines the electronic properties of the system as a whole. From a theoretical point of view, predicting the most stable interface configuration requires searching through the enormous structure space of flexible biomolecules with respect to the surface for different configurations and performing computational calculations of their properties. However, it is impossible to investigate those parts separately due to complex interactions during adsorption. In order to capture these complex interactions, one has to employ accurate theoretical methods, which are very computationally expensive. In this thesis, we provide a comprehensive description of the complex nature of the interaction of selected amino acids with metallic surfaces using state of the art dimensionality reduction techniques and accurate *ab initio* theoretical methods and creation of tools tailored for the high-throughput investigations of interface systems.

The theoretical methods used in the thesis are described in its first part. The second section looks into the conformational space changes of Arginine (Arg) and its protonated counterpart after adsorption on three noble metallic surfaces. Arg is an excellent testbed because it is tiny enough to be treated using density functional theory, which is considered the best compromise between accuracy and computational efficiency. At the same time, Arg is complex enough due to a highly flexible side-chain that allows for hundreds of different configurations in the gas phase alone. The examination of adsorption behaviour requires creating a database by performing a large number of geometry optimizations of various conformations and orientations. The investigation of that database includes creating a low-dimensional representation of the conformational spaces using recent dimensionality reduction techniques, followed by examining various bonding and charge transfer patterns and how they affect the available conformational spaces.

The third section of the thesis is concerned with developing tools for the automated structure search of interface systems and the modelling of self-assembly patterns formed after adsorption. Different geometry optimization algorithms and a flexible method of preconditioning the quasi-Newton optimization algorithms are implemented in the GenSec package that was developed. Together, these enable a more straightforward interface with a wide range of

quantum chemistry packages for sampling the conformational spaces of flexible molecules in 1D (ions), 2D (surfaces), and 3D (cavities and molecules) systems. Structure search of the conformational space of a flexible molecule using GenSec provided satisfactory results for di-L-alanine adsorbed on Cu(110) surface.

Contents

Abbreviations

AA amino acid. 1–3, 6, 7, 9, 11, 12, 31, 33, 93, 97
AIMD *ab initio* molecular dynamics. 10
Arg Arginine. i, iv, viii–xi, xv, 12, 45–69, 71, 73–75
Arg-H$^+$ Arginine-H$^+$. iv, viii–xi, xv, 12, 45–53, 55–58, 60–69, 71, 73–75
ASE Atomic Simulation Environment. 78–80, 85, 89, 90, 93, 97, 101, 106

BFGS Broyden-Fletcher-Goldfarb-Shanno. 37, 38, 85, 86, 89–91
BO Born-Oppenheimer. 14, 15

COM center of mass. 79, 81, 83, 85

DFA density-functional approximation. 18, 21, 25
DFT density-functional theory. x, xi, xv, 9–13, 16, 18, 20, 23, 27, 31, 34, 35, 46, 63, 64, 73, 74, 105

ES-IBD electrospray ion beam deposition. 7

fcc face-centered cubic. 31, 47, 90
FF force field. 9, 10, 13, 30, 31, 33, 34, 40, 41, 78, 82, 100
FHI-aims Fritz Haber Institute “ab initio molecular simulations”. 25–27, 41, 90, 93

GenSec Generation and Search. 78, 82, 83, 85, 89–91, 96, 97
GGA generalized gradient-approximated. 18–20

HEG homogeneous electron gas. 18

KS Kohn-Sham. 17, 25

LDA local density approximation. 18, 19
LDOS local density of states. 7
LJ Lennard-Jones. xii, 31, 40, 86, 88, 89, 92
LSM line search method. 37, 39, 40
LZK Lifshitz-Zaremba-Kohn. 24, 25

MD molecular dynamics. 10, 11, 32–34, 36
ML machine learning. 9, 11, 35, 37, 102

NAO numeric atom-centered orbitals. 25

NN neural networks. 37

PBC periodic boundary conditions. 27, 77, 82
PBE Perdew, Burke and Ernzerhof. xi, xv, 19, 20, 47–50, 63, 68, 71, 74, 90
PCA principal component analysis. 44
PES potential energy surface. 9, 10, 15, 30, 32, 34–39, 90, 92

QM quantum mechanical. 30, 31

REMatch regularized entropy match kernel. 43
RMSD root mean square displacement. 90, 91

SCF self-consistent field. 27
SOAP smooth overlap of atomic positions. 11, 41, 42, 52, 55
STM scanning tunneling microscopy. 7, 8, 13, 26, 29, 92, 93, 96, 97, 100, 101

TRM trust-region method. 38–40, 89
TS Tkatchenko and Scheffler. 21–24

vdW van der Waals. 1, 9, 20–25, 30, 34, 47, 63, 82, 86, 88, 92

XC exchange-correlation. 17–19, 21, 23, 24
XPS X-ray photoemission spectroscopy. 7

List of Figures

List of Tables

Does anybody really know the secret
Or the combination for this life and where they keep it?
It's kinda sad when you don't know the meanin'
But everything happens for a reason...

"Take a look around", Limp Bizkit

1 Introduction

Because of the fascinating potential applications of hybrid organic-inorganic interfaces, adsorption and self-assembly of organic molecules on surfaces are critical topics in nanoscience and nanotechnology [1]. For example, amino acids that are the building blocks of peptides and their oligomers are particularly intriguing because they are naturally biocompatible and provide a rich functional space already at the amino acid (AA) level. The combinatorial increase in molecular motifs made available by forming peptide bonds can further enlarge this functional space. By immobilizing a bioorganic component on a substrate, an inorganic part acts as a platform to support and capture interactions and reactions, which provide the path for creating different bionanoelectronic devices.

In recent years, a tremendous effort has been expended to identify adsorbates' structure on surfaces and disentangle the processes behind self-assembly that would lead to the rational design of materials and devices with desired properties.

From a theoretical point of view, this poses a challenge to computational modelling, particularly regarding the prediction of stable configurations at the interface at different conditions, which determines the electronic properties of the system as a whole. Even in the gas-phase, single AA have rich conformational spaces, where they can have hundreds of distinct local minima [2], and determination of them requires computationally expensive methods. After adsorption on the surface, the conformational preferences of the AAs can change dramatically due to a combination of factors, such as van der Waals (vdW), electrostatic or ionic interactions, but also due to their reduced flexibility, as well as by intermolecular forces and interactions with the surface itself [3, 4]. The systematic structure search of molecules

adsorbed on surfaces and creation of databases including energetic information from the theoretical approaches is of high importance for revealing structure-property relationships of the interface systems, for further developments of the theoretical methods able to describe larger structures, and for disentangling of the mechanisms of self-assembly. However, such studies are challenging as they require (i) accurate energetics for a system containing elements across the periodic table and where considerable charge rearrangement and chemical reactions can occur (ii) sampling and representing a large conformational space, and (iii) dealing with structure motifs that can only be represented by unit cells containing hundreds of atoms.

The scope of this thesis is the description of the complex nature of the interaction of AAs with metallic surfaces and the creation of tools for high-throughput calculations for investigations of interface systems. An exhaustive structure search for two AAs on three metallic surfaces was performed with the use of *ab initio* methods that are required for analysis of the electronic properties of the interface systems. The database created during the work contains thousands of local minima and is available for further development of the methods that can accelerate the research of self-assembly phenomena. The databases were analyzed with state-of-the-art unsupervised machine learning techniques that help reveal structure-property relationships in that kind of system. Further, we developed a package that automates the structure search of flexible molecules with respect to specified surroundings that connects to most of the electronic structure packages available today, making it freely available and open source. We investigate the adsorption of a di-L-alanine molecule adsorbed on Cu(110) surface using this package.

1.1 Amino acids and peptides

AAs are organic compounds that contain amino ($-NH_2$) and carboxyl ($-COOH$) functional groups, along with a side chain unique to each AA. AAs are known to be the monomer units of peptides and are essential for the existence of life. In the form of proteins, AA residues are the second-largest component of human muscles after water. Analyses of a large number of proteins from nearly every possible source have revealed that all proteins are made up of 20 “standard” AAs. Not all 20 types of AAs are found in every protein, although most proteins contain the majority, if not all, of the 20 types [5]. In addition, AAs and their derivatives are involved in processes as neurotransmitters - chemical messengers for communication between cells. For example, diminished activity of serotonin (tryptophan derivative) pathways plays a causal role in the pathophysiology of depression [6].

The most general formula to represent the common AA which is called α-amino acid, is reported in Fig. 1.1 a: the molecule is distinguished by the presence of a α carbon atom in the center, to which both the amino and carboxyl groups are attached. The rest of the molecule is represented as a side chain (R group), the structure of which uniquely defines all the common AAs. Depending on the molecule’s environmental conditions, AAs can exist in three different chemical forms (see Fig. 1.1 a): i) the neutral form is common for isolated

molecules; (ii) the zwitterionic form is common for solid AAs crystals and for molecules on poorly reactive surfaces and in solutions. This form appears when a proton is transferred from the carboxylic group to the amino group of the same molecule, which maintains its global neutrality; (iii) the anionic state is typical for AAs that interact strongly with a substrate, resulting in chemical bond breaking/formation and deprotonation of the molecule.

Except for the smallest AA glycine, all other AAs are chiral (Fig. 1.1 b), which implies that they have nonsuperimposable mirror images known as enantiomers of one another. Although there exist (L) and (D) enantiomers, the (L)-enantiomer is the only one found in living beings; as a result, the vast majority of investigations have been conducted on (L)-type molecules.

As one progresses towards more complicated and "realistic" biomolecules, one comes across peptides, which are polymers of AAs connected by CO-NH peptidic bonds (Fig. 1.1 c). A dipeptide, for example, is formed by the condensation of two AAs, i.e. the reaction between one AA's carboxyl group and the amino group of the second, with the elimination of one water molecule. Peptides are chains of comparable (homopeptides) or different (heteropeptides) AAs. Proteins are the "summum" of a peptide chain, where the sequence of AAs, their location, and their three-dimensional layout regulate the biological activity of the molecule.

AAs exhibit a range of polarity and structural features. AA side chains can be nonpolar (e.g. glycine, alanine, valine, leucine, isoleucine, methionine, proline, phenylalanine, tryptophan), polar (e.g. serine, threonine, asparagine, glutamine, tyrosine, cysteine), or charged (e.g. arginine, lysine, histidine, aspartic acid and glutamic acid). Side chains may be nonpolar or polar (neutral or charged). They may be aliphatic (e.g. alanine) or contain other functional groups such as carboxylic group (e.g. glutamic acid), amino group (e.g. lysine), or sulphur (e.g. cysteine). Additionally, they can be linear (e.g. glutamic acid) or have one heterocycle (e.g. proline) or aromatic (e.g. tyrosine) ring in their side chain. The structures of the twenty most frequent AAs, along with their three-letter notations and side-chain characteristics, are depicted in Fig. 1.2. More comprehensive review considering other properties of AAs and other AAs that are not specified by the "universal" genetic code that is common for almost all life forms can be found in biochemistry textbook [5].

Even with the mentioned AAs, the chemical space of possible configurations is genuinely immense, and peptides that can be formed of different sequences of AAs will vary a lot on their structural configuration and properties, which presents an advantage for the rational design of different nanodevices and functionalization of inorganic surfaces.

1.2 Recent applications of peptide-inorganic surface interface systems

In this section, we would like to show some of the recent applications of peptide-metal interfaces and thus showcase the great potential of such a field of research.

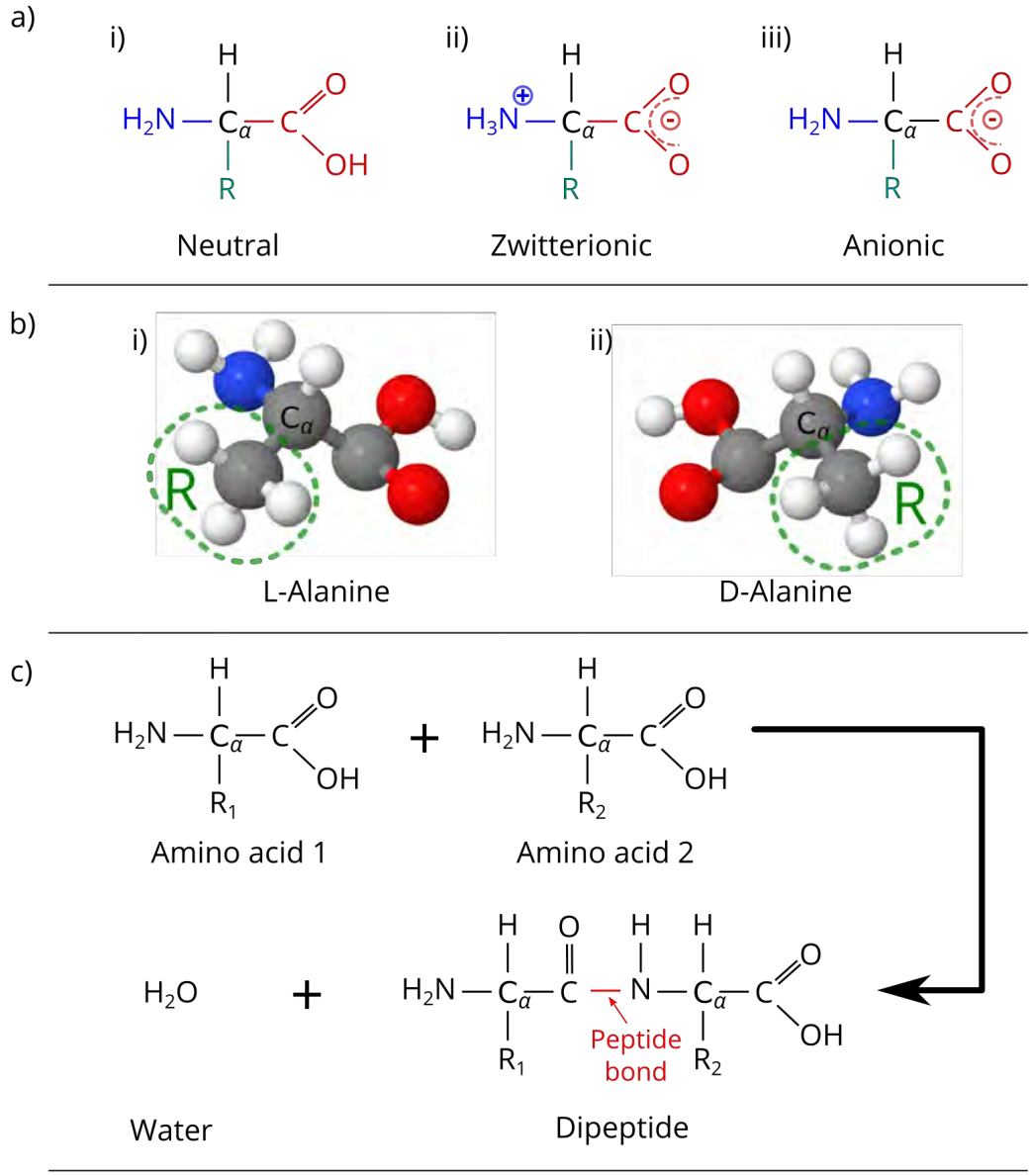


Figure 1.1 – a) The general structure of a α-amino acid in its neutral, zwitterionic, and anionic states. The amino group is highlighted in blue, the carboxylic/carboxylate group is highlighted in red, the α-carbon is highlighted in black, and the side chain is highlighted in green; b) Schematic representation of the Alanine amino acid in its neutral configuration. Red atoms are oxygen; blue atoms are nitrogen; white atoms are hydrogen, and grey atoms are carbon. The R symbol stands for the side-chain (highlighted with green dashes), here represented by the CH_3 group. In (i) L-Alanine, with respect to the central C_α carbon and in (ii) a D-Alanine; c) Schematic representation of the formation of the peptide bond: two amino acids with different side chains R_1 and R_2 react to form a peptide via the production of a water molecule.

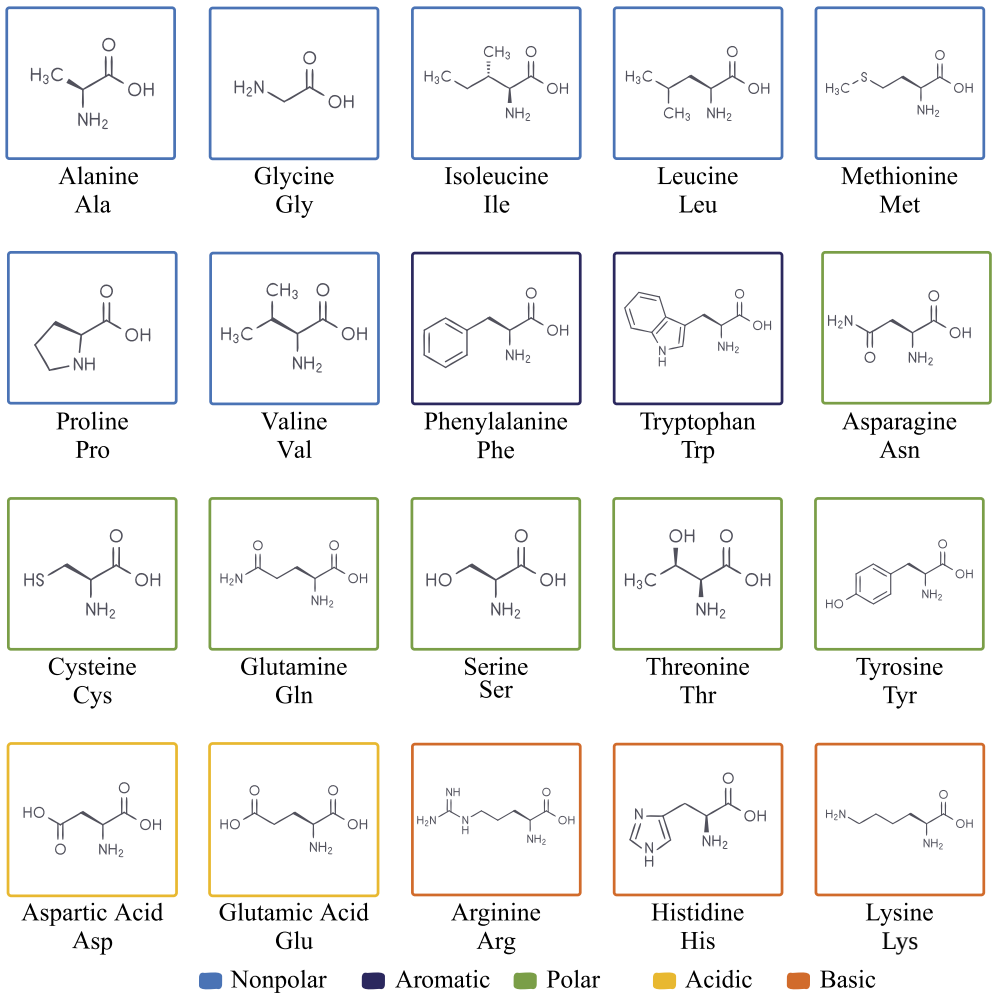


Figure 1.2 – Scheme of the 20 most common α-amino acids present in nature, represented in their neutral form.

The use of peptides in solar cell applications, inspired by natural photosynthesis processes, is arguably the most straightforward optoelectronic application. Appending a dye to the side chain, or one of the ends, of a peptide, was shown to be effective in extending the absorption spectrum and increasing photocurrent production capacities [7, 8] even when the peptide is physically adsorbed on a gold surface [9]. In the presence of dyes with different excitation wavelengths, the synthesis of mixed monolayers of helical peptides with opposing dipole orientations towards the surface allowed the creation of a molecular photodiode system that can switch photocurrent direction by varying the excitation wavelength [10]. The efficiency of the organic solar cells can also be tuned by interfacial modification with an ultrathin peptide layer that causes changes in the work function of the substrate [11] that is also highly dependent on the peptide sequence and conformation of the backbone [12, 13].

Using peptides as molecular bridges and producing conductive wires is crucial for the next

generation of bioelectronic devices. The effectiveness of electronic transport is dependent on the overall charge of protonating side chains which allow controlling I-V characteristics of peptide junctions [14–16]. Self-assemblies on surfaces can provide the unique and flexible way to implement ensembles of low-dimensional quantum confinement geometries [17], for example, of fullerenes that are too mobile on the surface without such a template [18] or for quantitative modulation of the work function of a substrate [19].

Using peptide monolayers as an antifouling coating [20, 21] to inhibit the adherence of proteins and organisms to surfaces is one of the most potential applications in industry and medicine. Promotion of cell adhesion and proliferation on biomaterials is essential for the successful integration of implants [22]. Cell binding motifs, such as the Arg–Gly–Asp peptide, can be anchored to the surface of a biomaterial to increase its mechanical and biological characteristics [23]. It has been proven that titanium surfaces, a material that is commonly utilized in the implant industry, can be functionalized with cell-binding peptides by employing Cys AAs as the binding factor [24]. Also, the surface reactivity can be altered by using the intrinsic chirality of AAs, which enables chiral separation and enantioselective heterogeneous catalysis [25–27].

Another example is controlling the wettability of graphite surfaces using self-assembled peptides by mixing distinct peptide types (hydrophobic and hydrophilic) in different ratios [28]. The excellent stability of peptide nanostructures, as well as their vast surface area and controlled wettability features, make them an appealing candidate for use as the dielectric layer in supercapacitors [29, 30]. Also, AAs are non-toxic, relatively cheap and easy to produce promising green corrosion inhibitors [31–35].

At the time of writing, the author can not stress enough the need for producing biosensors targeted explicitly for detection of the pathogenic microbes and viruses, where organic molecules provide high biocompatibility and tunable selectivity due to significant variations of accessible chemical configurations [36–42].

Even though there are already many applications and devices, the fundamental mechanisms that govern particular structures adopted on particular surfaces remain unclear. The following section will be devoted to both state-of-the-art experimental and theoretical methods of investigations of organic-inorganic interfaces.

1.3 State of the art

During past decades it has been proven that a large diversity of distinct molecular assemblies may form via adsorption of organic molecules at inorganic surfaces. However, many aspects of the interaction mechanisms of biomolecules and inorganic surfaces are still unclear. Often, the shape of such self-organized structures may be adjusted by carefully controlling the deposition circumstances such as temperature [43–45], coverage [46] or changing of the substrate [47–50]. This section will offer a quick review of the methodologies that are used to investigate the adsorption of AAs on inorganic surfaces.

1.3.1 Experimental techniques

Self-assembly processes between molecules start with the adsorption of individual molecules from the gas phase (or liquid), then diffusion on the surface and further island formation through molecule-molecule interaction. Using a crucible (Knudsen cell) to sublimate AAs from a crystalline form under vacuum conditions is a standard method for generating organic layers on the substrates [51]. Such a technique is limited to relatively small peptides (of up to four AAs) and requires careful adjustment of the sublimation temperature since melting the powders can damage them. One of the most sophisticated approaches is soft-landing electrospray ion beam deposition (ES-IBD) since the production of intact gas-phase ions by electrospray ionization is not limited by low thermal stability [52–54]. Molecular ions are decelerated before landing, preventing fragmentation and guaranteeing that the molecules remain intact following deposition. The use of mass spectrometry, mass filtering, and soft landing, all of which are essential to the ES-IBD process, ensures the intact and extremely pure deposition of the selected species under ultrahigh vacuum [52, 55, 56].

The most fundamental tool to study the self-assembly patterns of molecules is scanning tunneling microscopy (STM), which is based on the concept of quantum tunnelling. This technique measures the tunnelling current as a function of the sharp conducting tip position, applied voltage, and the local density of states (LDOS) of the sample since electrons can tunnel across the vacuum between tip and sample when the bias voltage is applied [57]. This technique allows one to determine the atomic positions in molecules and the morphology of the substrate. STM allows to obtain a three-dimensional profile of a sample as an image and distinguish different adsorption patterns of a single peptide [58–60] or how the self-assemblies look depending on the different chemical composition of the adsorbates, substrate, and overall deposition conditions [25, 46, 61–65]. However, the interpretation of STM images of molecules adsorbed on surfaces is not straightforward. First of all, STM images are not a topography map but also include electronic information of both the molecule and the underlying surface. In the case of chemisorbed systems, STM images carry information about the chemical bonding that can be extracted only from complementary investigations.

STM is often supplemented with spectroscopic studies that provide chemical state information of the adsorbed molecules and the surroundings of the functional groups. For example, AA adsorption can occur in different protonation states that can be described by proton configuration of carboxyl and amino groups (neutral, anionic or zwitterionic) and by different protonation configuration of Histidine AA. The occurrence of a zwitterionic form can be evidenced by X-ray photoemission spectroscopy (XPS) that allows the investigation of the core levels of the atoms present at the surface. The XPS analysis of core-level shifts will immediately show the presence of a charged NH_3^+ functional group, which causes an upshift on the N 1s photoemission line [25]. It is also possible to estimate the relative co-existing states of the same molecule adsorbed on the surface.

Additionally, a tunable X-ray source allows other types of spectroscopies, like near-edge X-ray

absorption fine structure (NEXAFS), where the X-ray adsorption features can be indicated by the photoabsorption cross section for electronic transitions from an atomic core level to final states in the energy range of 50–100 eV above the chosen atomic core level. When employing differently polarized light, the directed electric field vector of the X-rays can only excite those electrons able to move parallel to it, which gives them crucial information on the chemical bonding orientation [49, 66, 67].

Vibrational spectroscopy is another experimental technique that exploits the fact that molecules absorb energy at specific frequencies which resonate with their vibrational modes. Due to interactions with the surface, those specific frequencies are changed relative to gas-phase frequencies but remain characteristic to the adsorption site's chemical groups, configuration, and geometry. On metal surfaces, reflection absorption infrared spectroscopy (RAIRS) [68, 69] or high resolution electron energy loss detection (HEELS) [70, 71] can be employed. However, due to many potential vibrational modes, additional methods are frequently essential for the characterization of the adsorbed system. For more comprehensive experimental techniques, we refer the reader to the reviews [61].

Unfortunately, experimental procedures cannot provide the system with information at the needed level of resolution. The unknown tip geometry and electrical characteristics are usually the most significant uncertainties encountered in detailed STM interpretation. Also, surface diffusion is significant at room temperature, causing tip instability and affecting atomic and electrical characteristics. One of the most significant experimental limitations of spectroscopic approaches is that spectra are obtained by measuring the sample's total yield of electrons or photons. A direct link between the measured spectra and the sample's specific geometry is not guaranteed. Because of the limited resolution, high complexity of the systems, technical difficulties, and cost of the experiments, theoretical approaches become essential for accessing the properties that are not accessible through experiments, resulting in a synergy of theory and experimental data that leads to a deeper understanding of the processes that are taking place on the surface.

1.3.2 Theoretical techniques

In addition to experimental research, model computations are required to bring more insights into the structure and characteristics of the molecule–surface systems. For example, issues that can be addressed theoretically are the nature of the intermolecular interactions, structure of adsorbates, charge configuration of the molecules, their chemical composition, chiral recognition, orientation and preferred adsorption sites. In principle, the theoretical foundation suitable for addressing the problems mentioned above was already fully established with the formulation of quantum mechanics in the first part of the 20th century. However, as Paul Dirac once wrote: "The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble" [72]. We are restricted to different approximations that allow us to model systems of different scales and the available computational power that

can treat such calculations.

For modelling systems that consist of hundreds of atoms per unit cell, the most popular theoretical approach nowadays is DFT, which delivers a good compromise between accuracy and computational efficiency. The fundamental theorem behind DFT is that the electronis structure properties of non-degenerate systems are entirely determined by their ground-state electron density, $n(r)$, that alone governs the whole behaviour of the system. The so-called exchange-correlation functional, which is the $n(r)$-dependent energy contribution caused by quantum-mechanical and many-body deviations from a mean-field description of the electrons, is a fundamental piece of this approach. However, a precise equation for the universal functional still has not been found, giving rise to many suitable approximations for different systems. DFT will be discussed in more detail in the next chapter.

Pioneering works that used DFT were focused on small or rigid AAs, and on a minimal number of trial configurations [73–77] due to the high computational cost of such calculations at the time. Because the first DFT studies of complex systems did not account for vdW interactions, they were affected by a errors in their predictions; however, they are now taken into account in more modern functionals and approaches that result in a significant increase in the quantitative agreement between the predictions and the experimental data [78]. With the use of DFT, one can answer whether a chemical bond is formed between AA and a substrate, what the energy hierarchy of different adsorbed conformational configurations is, as well as determining charge distribution on the adsorbed structures and their height above the surface [61, 79–83].

One of the first studies that were dedicated to larger AAs highlight the challenge of adequately sampling the large structure space of flexible biomolecules [84] that is usually not feasible with the use of DFT due to high computational cost. These studies have clarified that an accurate potential energy surface (PES) is only one of the ingredients needed to correctly predict the structure of peptides at surfaces, with the sampling of structure space being just as important.

DFT calculations not only offer valuable information on their own, but also they can provide the basis to cheaper theoretical approaches and used, for example, as a basis for a classical force field (FF) parameterization [85–87] or for the training of machine learning (ML) models [88–90]. These methods can be several orders of magnitude cheaper to evaluate compared to DFT and, in some cases, FFs specifically developed for modelling simulations between a protein and a surface may be a good approximation. However, to obtain high-quality results, the FF parameters must be derived and calibrated for the systems of interest. Different FFs exist for modelling AAs on metallic surfaces and the most famous ones are GolP-CHARMM FF [85, 91] optimized for Au(111) and Au(100) slabs, AgP-CHARMM FF [86] that is parametrized for simulations on Ag(111) and Ag(100) in aqueous solutions and INTERFACE-FF [87] which includes a broad range of different surfaces available for modelling. The main drawback of using FFs in simulations is their non-transferability to systems other than those to which

they were parametrized. Another limitation of these FFs is the inability to model chemical reactions or to capture effects such as charge transfer. While more complex FFs exist, such as bond order-based reactive FF (ReaxFF) [92] that in contrast to the previous FFs allows bond breaking and formation reactions, such FFs require much larger training sets, which can be a limiting factor for using them for various systems. To the best of our knowledge, only one ReaxFF was designed to model adsorption of glycine on Cu(110) [93].

1.3.3 Global structure search

The most challenging part of theoretical modelling is properly sampling the large structure space of flexible biomolecules. Theoretical methods such as DFT and FF allow for the calculation of the forces acting on nuclei based on the input geometry of the structure. It is possible then to determine the nuclei arrangement that results in local or global minima of the system with a given PES.

Finding the global minimum of the system implies sampling the conformational space of complex molecular systems, which frequently arises in the context of molecular dynamics (MD) simulations. With the use of MD methods, Newton's law of motion is solved numerically for the nuclei. It is possible, then, to sample the most likely regions of the PES with an array of different MD flavours, such as Born-Oppenheimer and Car-Parrinello [94]. These are usually denoted as *ab initio* molecular dynamics (AIMD) simulations since the PES is constructed using quantum mechanical approaches. Despite the very limited time scales that can be simulated using AIMD (up to hundreds of ps), studies are employing such methods, for example, to investigate the preferred chemical composition and adsorption sites of glycine and lysine [95, 96], and to study peptide-silica interactions [97] or β-sheet adhesion of gold surfaces [98].

The exploration of PES with the methods described above can be very inefficient since, during such simulations, the system can be trapped in some local minima, which limits the sampling of the conformational space. There are different methods proposed in order to enhance the sampling efficiency of MD simulations and these have been used for investigations of protein-surface interactions [79]. We will discuss them in more detail in Section 3.1.

1.3.4 Analysis of high-dimensional spaces

Analysing complex molecular systems with many degrees of freedom and interpreting of their high-dimensional data is another challenge in understanding the structure-property relationships of flexible molecules adsorbed on inorganic surfaces. There is no analytical method to determine the configurations of the different peptide structures. One of the first representations developed for the analysis of peptide structures was proposed by Ramachandran, which uses dihedral angle rotations around the N-C_α and C_α-C bonds [99] to represent the number of possible conformations for an amino-acid residue in a protein, as well as the distribution of those data points. The Ramachandran approach generally proposes quite a simple metric for qualitative analysis of the secondary structures and distinguishing between amino-acids, but is not suitable for the analysis of the structural changes within

one system due to the small number of input parameters, and requires the extraction of specific information such as dihedral angles. The modern approach for visualising the complex conformational space in material science is to use machine learning techniques for dimensionality reduction that rely on introducing suitable molecular descriptors of the whole system and introducing a metric in high-dimensional space. The main properties of such descriptors should be (i) invariance to transformations such as translations, rotations and permutations of atom indexing; (ii) uniqueness that implies that systems different in structure will be mapped in different representations; (iii) Continuity with respect to changes in atomic coordinates, which is required for stability of ML models and (iv) generality for the ability to describe any system [100].

Different molecular descriptors are used in computational chemistry for representing molecular systems, but most of them do not fulfil all the requirements listed above. For example, descriptors widely used in chemoinformatics such as Simplified molecular-input line-entry system (SMILES) [101], International Chemical Identifier (InChI) [102] that encode in a one-line notation the connectivity, the bond type, and the stereochemical information and fingerprints such as Extended-connectivity fingerprints (ECFPs) [103] violate (ii) and (iii) due to lack of information about the spatial arrangement of atoms. Including the spatial 3D information can be done by using Cartesian coordinates and representation on internal coordinates, but both violate requirement (i). The field of developing molecular descriptors is quite active, with the Coulomb matrix [104], bag of bonds (BoB) [105], many-body tensor representation (MBTR) [106, 107], and bonds angles machine learning (BAML) [108] recently introduced. One of the descriptors that satisfy all the requirements above and can capture local changes of the environment is smooth overlap of atomic positions (SOAP)[109, 110], which is a general representation where the atom-centred local neighbourhood is a sum of Gaussians located at atoms within the local environment. The density is expanded in orthogonal radial, and spherical harmonics basis functions [111]. This descriptor was successfully applied in the visualisation of conformational spaces of biomolecules [109, 110, 112–115]. The overall performance of SOAP descriptors means it appears to be becoming increasingly popular compared to other descriptors [107, 116, 117]. With these descriptors, similarities between atomic configurations can be formulated [107] and dimensionality reduction techniques can be applied [118]. Such techniques were applied for analysis of the MD trajectories [112] and of the AA datasets [110].

1.3.5 Overview of the thesis

In this thesis, we present one of the most extensive and accurate studies of adsorbed AAs (with use of DFT) in the literature up to date. Global structure search of systems with large conformational space is one of the bottlenecks in modern computational studies, and one of the parts of this thesis is explicitly dedicated to this problem.

This thesis is divided into five main chapters. The second chapter is dedicated to the theoretical foundation, mainly to the electronic structure calculation methods used in the thesis. The third chapter is also theoretical and describes the methods for investigating and analyzing

conformational spaces of flexible molecules.

The fourth chapter describes the work that was done to investigate the conformational space changes of Arg and its protonated counterpart Arg-H^+ after adsorption on three noble metallic surfaces [83]. Arg was chosen as a good testbed since it is small enough to be treatable using DFT and at the same time challenging enough due to a very flexible side-chain which allows for hundreds of possible configurations in the gas phase alone. Also, Arg is the most flexible among AAs [2] and least investigated while adsorbed on metallic surfaces [61]. The analysis of the adsorption behaviour required the creation of a database by performing a large number of geometry optimizations of different conformations and orientations. The analysis of that database includes producing a low-dimensional representation of the conformational spaces using modern dimensionality reduction techniques and following analysis of different patterns of bonding and charge transfer and how it can affect the accessible conformational spaces.

The fifth chapter of the thesis deals with developing the tools for the automated investigation of flexible molecules, which also enables the modeling of self-assembly patterns formed after adsorption. Different geometry optimization algorithms are implemented together with a flexible way of preconditioning the quasi-Newton optimization algorithms in the package. Together, these allow a simplified interface with a wide variety of electronic structure packages ready to sample conformational spaces of flexible molecules with respect to 1D (ions), 2D (surfaces), and 3D (cavities and molecules) fixed frames. Also, it shows the application of the package, described in the fourth chapter, where we showcase the structure search algorithm on the di-L-alanine molecule adsorbed on Cu(110) surface and compares our findings with experimental results.

2

Theoretical methods

The essential ideas, notations, and approximations utilised in this thesis are introduced in this chapter. We will explain and motivate the central approximation in condensed matter physics and quantum chemistry after first explaining the many-body problem, which addresses the electrons as quantum objects. Next, we will present the theoretical technique that will play a major role in this thesis: the density-functional theory (DFT). The fundamentals of DFT will be covered, including a discussion of the most common approximations and modern developments, such as the inclusion of the long-range correlation interactions. This chapter will also discuss the basics of theoretical production of STM images and the calculation of charge transfer effects. Also, a short overview of the FF techniques will be covered at the end.

2.1 The many-body problem

A system composed of nuclei and electrons may be formally characterized in quantum mechanics by solving the time-independent Schrödinger equation. It's non-relativistic form is given by:

$$\hat{H}\Psi = E\Psi, \tag{2.1}$$

where $\hat{H}$ represents the non-relativistic time-independent Hamiltonian operator, E denotes the total energy of the system, and Ψ is the many-body wave function of the system that depends on electronic and nuclear degrees of freedom $\Psi = \Psi(\mathbf{r}_i; \mathbf{R}_I)$, where $\mathbf{r}_i$ and $\mathbf{R}_I$ correspond to the electron and nuclei position vectors. Hamiltonian $\hat{H}$ in the absence of an

external electromagnetic field consists of five terms:

$$\hat{\mathrm{H}} = \hat{T}_{\mathrm{n}} + \hat{T}_{\mathrm{e}} + \hat{V}_{\mathrm{e-e}} + \hat{V}_{\mathrm{ext}} + \hat{V}_{\mathrm{n-n}}, \tag{2.2}$$

where $\hat{T}_{\mathrm{n}}$ and $\hat{T}_{\mathrm{e}}$ are the nuclear and electronic kinetic energy operators, $\hat{V}_{\mathrm{e-e}}$ and $\hat{V}_{\mathrm{n-n}}$ are the electron–electron and nuclear–nuclear Coulomb repulsion, and $\hat{V}_{\mathrm{ext}}$, is the electron–nuclear Coulomb attraction. For simplicity atomic units are used where the electron mass m_{e}, the elementary charge e, the reduced Planck constant $\hbar$ as well as the vacuum permittivity factor $4\pi\epsilon_0$ are all set to unity. The Hamiltonian in Eq. 2.2 can be written explicitly as

$$\hat{\mathrm{H}} = -\frac{1}{2}\sum_{I=1}^{M}\frac{\nabla_I^2}{M_I} - \frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} - \sum_{i=1}^{N}\sum_{I=1}^{M}\frac{Z_I}{r_{iI}} + \sum_{I=1}^{M}\sum_{J>I}^{M}\frac{Z_I Z_J}{R_{IJ}}, \tag{2.3}$$

where the indices i, j refer to indexes of N electrons and I, J are indexes of M nuclei so that Z_I denote the nuclear charge, M_I is the nuclear mass, $r_{ij} = \left|\mathbf{r}_i - \mathbf{r}_j\right|$, $r_{iI} = |\mathbf{r}_i - \mathbf{R}_I|$ and $R_{IJ} = \left|\mathbf{R}_I - \mathbf{R}_J\right|$ represent the electron-electron, electron-nucleus and nucleus-nucleus distances respectively. In the above equation, the Laplacian operators ∇_i^2 and ∇_I^2 include differentiation with respect to the ith electron and Ith nucleus coordinates.

Since the nuclei and electrons are not constrained in general, the solution of Eq. 2.1 implies a problem of $3N + 3M$ ($4N$ considering the spin variables) degrees of freedom. Since exact analytical solutions to the Eq. 2.1 are only accessible in a few limited cases, the following sections discusses approximations that allow obtaining a numerical solution for the systems relevant to the scope of this work.

2.2 The Born-Oppenheimer approximation

The Born-Oppenheimer (BO) approximation is a fundamental concept in electronic structure theory that provides a significant simplification of Eq. 2.1 by decoupling the dynamics of electrons and nuclei.

Because nuclei are significantly heavier than electrons for example, for a single proton, the ratio is

$$\frac{m_{\mathrm{e}}}{M_{\mathrm{p}}} \approx \frac{1}{1836} \ll 1, \tag{2.4}$$

to a fair approximation, electrons in a molecule can be thought to be travelling in a field of fixed nuclei. Within this approximation, the first term of Eq. 2.3, the nuclei's kinetic energy, may be ignored, and the last component of Eq. 2.3, the nuclei's repulsion, can be assumed to be constant. Any constant introduced to an operator increases the operator's eigenvalues and does not influence the eigenfunctions of the operator. The remaining components in Eq. 2.3 are known as the electronic Hamiltonian $\hat{\mathrm{H}}_e$, which only depends parametrically on the nuclear coordinates $\mathbf{R}$:

$$\hat{\mathrm{H}}_{\mathrm{e}}(\mathbf{R}) = \hat{T}_{\mathrm{e}} + \hat{V}_{\mathrm{e-e}} + \hat{V}_{\mathrm{ext}}. \tag{2.5}$$

that describes the motion of N electrons in a field of M point charges. The time-independent

Schrödinger equation for electronic part, considering ν electronic eigenfunctions for $\hat{\mathrm{H}}_e$ will be:

$$\hat{\mathrm{H}}_{\mathrm{e}}\psi_\nu(\mathbf{r};\mathbf{R}) = E_\nu^{\mathrm{e}}(\mathbf{R})\psi_\nu(\mathbf{r};\mathbf{R}),, \quad \text{with } \nu = 1,\ldots,N \tag{2.6}$$

where E_ν^{e} is the electronic energy of the electron that moves in the field created by the point charges produced by the given configuration of the nuclei. The total wavefunction Ψ can be expanded into a nuclear χ and an electronic part ψ as:

$$\Psi(\mathbf{r},\mathbf{R}) = \sum_\nu \chi_\nu(\mathbf{R})\psi_\nu(\mathbf{r};\mathbf{R}), \tag{2.7}$$

where $\chi_\nu(\mathbf{R})$ are functions of the nuclear positions and represent the coefficients of such expansion. With the entire Schrödinger Eq. 2.1 and a left-side multiplication by $\langle\psi_\mu(\mathbf{r};\mathbf{R})|$ followed by integration over the electronic coordinates and application of chain rules, the equation becomes [119]:

$$\begin{aligned} E\chi_\mu(\mathbf{R}) &= \left[\hat{T}_n + \hat{V}_{n-n} + E_\mu^{\mathrm{e}}\right]\chi_\mu(\mathbf{R}) - \\ &- \sum_\nu \sum_I \frac{1}{2M_I}\left[2\left\langle\psi_\mu(\mathbf{r};\mathbf{R})|\nabla_I|\psi_\nu(\mathbf{r};\mathbf{R})\right\rangle\nabla_I + \left\langle\psi_\mu(\mathbf{r},\mathbf{R})\left|\nabla_I^2\right|\psi_\nu(\mathbf{r};\mathbf{R})\right\rangle\right]\chi_\nu(\mathbf{R}) \end{aligned} \tag{2.8}$$

where E now is the total energy of the system, where we applied the property

$$\left\langle\psi_\mu(\mathbf{r};\mathbf{R})\,|\,\psi_\nu(\mathbf{r};\mathbf{R})\right\rangle = \delta_{\mu\nu} \tag{2.9}$$

The off-diagonal elements of the last two terms in the Eq. 2.8 are called non-adiabatic contributions, describing the interaction between different electronic states. Within the BO approximation, these terms are assumed to be zero:

$$\left\langle\psi_\mu|\nabla_I|\psi_\nu\right\rangle = \left\langle\psi_\mu\left|\nabla_I^2\right|\psi_\nu\right\rangle = 0 \text{ for } \mu \neq \nu, \tag{2.10}$$

which means that the atomic motion does not induce electronic excitations. The elements for $\left\langle\psi_\mu\left|\nabla_I^2\right|\psi_\mu\right\rangle$ can be also neglected in comparison with electronic ones, since electron to proton mass ratio is at least of the order 10^{-4} (Eq. 2.4). With all these assumptions, the BO PES, where the nuclei move, is defined as

$$V_\mu^{\mathrm{BO}}(\mathbf{R}) = \hat{V}_{\mathrm{n-n}}(\mathbf{R}) + E_\mu^{\mathrm{e}}(\mathbf{R}), \tag{2.11}$$

where $\mu = 0$ is the electronic ground-state. It has to be noted that the BO approximation fails when a transition between electronic states occurs. For example, when examining organic molecules and UV photoabsorption, a conical intersection between the electronic ground and excited states can be observed depending on the geometry of the molecule. In this situation, the excited molecule undergoes an ultrafast non-adiabatic internal conversion, which does not result in the emission of radiation, and violates the condition in Eq. 2.10 [120].

2.3 Density Functional Theory

The Nobel Prize in Chemistry 1998 was divided equally between Walter Kohn "for his development of the density-functional theory" and John A. Pople "for his development of computational methods in quantum chemistry". The initial work on Density Functional Theory (DFT) was reported in two of Kohn's publications with Pierre Hohenberg in 1964 [121] and with Lu J. Sham in 1965 [122]. The main advantage of the DFT approach is its compromise between accuracy and computational cost, which made it a very popular and common technique for the calculation of the properties of different systems from condensed matter to isolated molecules. DFT is an electronic-structure calculation method that replaces the N-electron wave-function ψ_e with the electron density $n(\mathbf{r})$ that depends only on 3 spatial coordinates. From an N-electron wavefunction, the electron density can be obtained by integration:

$$n_0(\mathbf{r}) = N \int \left|\psi_0(\mathbf{r}, \mathbf{r}_2, \ldots, \mathbf{r}_N)\right|^2 d\mathbf{r}_2 \ldots d\mathbf{r}_N, \tag{2.12}$$

where N is the number of electrons in the system and the dependency on the spin is omitted for simplicity.

The foundation of DFT began from the Thomas-Fermi model [123, 124], where the energy of the system was expressed in terms of electron density based on the homogeneous electron gas. Based on this idea, Hohenberg and Kohn developed the mathematical basis of modern DFT that proves that all the ground-state properties of the system can be expressed as functionals of the electronic density [125].

2.3.1 The Hohenberg-Kohn theorems

The electron density contains all necessary information about the system, as was shown by Hohenberg and Kohn in 1964 through two theorems:

1. The external potential $v_{\text{ext}}(\mathbf{r})$ is a unique functional of electron density $n(\mathbf{r})$. This means that the electron density, in fact, uniquely determines the Hamiltonian and thus all electronic properties of the system, making it possible to describe the properties of the system as a functional of $n(\mathbf{r})$. The total energy of the system has the form

$$E[n(\mathbf{r})] = \int v_{\text{ext}}(\mathbf{r}) n(\mathbf{r}) d\mathbf{r} + F[n(\mathbf{r})] \tag{2.13}$$

The first term depends on the actual system of interest under investigation and includes the electron-nuclei attraction. The second term is universal in the sense that its form does not depend on the number of electrons, nuclei positions and their charges:

$$F[n(\mathbf{r})] = T[n(\mathbf{r})] + E_{e-e}[n(\mathbf{r})], \tag{2.14}$$

where $T[n(\mathbf{r})]$ is the kinetic-energy functional and $E_{e-e}[n(\mathbf{r})]$ is the electron-electron interaction functional.

2. The electron density that minimises the value of the energy functional is the exact ground-state density n_0:

$$E[n_0] \leq E[n(\mathbf{r})] \tag{2.15}$$

The proofs of the two Hohenberg-Kohn theorems are straightforward and can be found elsewhere [126]. Elimination of the restriction to non-degenerate ground-states was provided by Levy-Lieb [125]. However, these theorems do not give a practical method for solving the equations and obtaining electron densities.

2.3.2 The Kohn-Sham equations

The idea of the Kohn-Sham scheme is to define a non-interacting system of N electrons whose ground-state electron density exactly equals the ground-state density of real, interacting system n_0. The density is then constructed as a sum of single-particle Kohn-Sham (KS) orbitals:

$$n(\mathbf{r}) = \sum_{i}^{N} \phi_i^*(\mathbf{r})\phi_i(\mathbf{r}). \tag{2.16}$$

The KS theorem ensures the existence of an effective external potential such that a system of non-interacting electrons will produce exactly the same ground-state electron density. Then one can rewrite the total energy functional in a way that includes well-defined terms:

$$E[n(\mathbf{r})] = T_S[n(\mathbf{r})] + V_{\mathrm{H}}[n(\mathbf{r})] + E_{\mathrm{xc}}[n(\mathbf{r})] + \int v_{\mathrm{ext}}(\mathbf{r})n(\mathbf{r})d\mathbf{r}, \tag{2.17}$$

where T_S is the kinetic energy operator of non-interacting system and $V_{\mathrm{H}}[n(\mathbf{r})]$ is the Hartree term:

$$\mathrm{T_S}[n(\mathbf{r})] = -\frac{1}{2}\sum_{i}^{\mathrm{N}} \left\langle \phi_{\mathrm{i}}^*(\mathbf{r}) \middle| \nabla^2 \middle| \phi_{\mathrm{i}}(\mathbf{r}) \right\rangle, \tag{2.18}$$

$$V_{\mathrm{H}}[n(\mathbf{r})] = \frac{1}{2}\iint \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}d\mathbf{r}', \tag{2.19}$$

where the factor $1/2$ is present to avoid double counting. The first three terms of Eq. 2.17 are the functional $F[n(\mathbf{r})]$, and the quantum-mechanical many-body complexity is described by $E_{\mathrm{xc}}[n(\mathbf{r})]$, the exchange-correlation (XC) functional that is unknown. $E_{\mathrm{xc}}[n(\mathbf{r})]$ includes the difference between the true kinetic energy $T[n(\mathbf{r})]$ and the kinetic energy of the non-interacting system, as well as all the non-classical electron-electron interactions:

$$E_{\mathrm{xc}}[n(\mathbf{r})] = T[n(\mathbf{r})] - T_S[n(\mathbf{r})] + V_{e-e}[n(\mathbf{r})] - V_{\mathrm{H}}[n(\mathbf{r})]. \tag{2.20}$$

As in the Hartree-Fock method, applying the variational principle and minimizing Eq. 2.17 with respect to the electron density, with the constraint that any electron density must conserve the total number of electrons, yields the set of single-particle KS equations [127]:

$$\hat{h}^{KS}\phi_i(\mathbf{r}) = \epsilon_i\phi_i(\mathbf{r}) \tag{2.21}$$

$$\left(-\frac{1}{2}\nabla^2 + v_{\mathrm{H}}(\mathbf{r}) + v_{\mathrm{xc}}(\mathbf{r}) + v_{\mathrm{ext}}(\mathbf{r})\right)\phi_i(\mathbf{r}) = \epsilon_i\phi_i(\mathbf{r}) \tag{2.22}$$

$$\frac{\delta V_{\mathrm{H}}[n(\mathbf{r})]}{\delta n(\mathbf{r})} = v_{\mathrm{H}}(\mathbf{r}) = \int \frac{n\left(\mathbf{r}_j\right)}{\left|\mathbf{r}-\mathbf{r}_j\right|} d\mathbf{r}_j, \quad \frac{\delta E_{\mathrm{xc}}[n(\mathbf{r})]}{\delta n(\mathbf{r})} = v_{\mathrm{xc}}(\mathbf{r}), \tag{2.23}$$

where v_{H} is called Hartree potential and v_{xc} is XC potential. Usually these three potential are combined in one effective single-particle potential:

$$v_{\mathrm{eff}}(\mathbf{r}) = v_{\mathrm{H}}(\mathbf{r}) + v_{\mathrm{xc}}(\mathbf{r}) + v_{\mathrm{ext}}(\mathbf{r}). \tag{2.24}$$

Starting with a trial electron density and solving the set of single-particle equations from Eq. 2.22 one can obtain a new set of eigenstates from which to obtain a new density, and continuing this procedure minimizes the total energy self-consistently.

2.3.3 Exchange correlation functionals

Until now DFT in itself is a truly *ab initio* method if the exact form of the XC functional could be written down. Since it is not known, approximations to it have to be made, which gives rise to different density-functional approximation (DFA) that can be separated into different types. The simplest is the local density approximation (LDA). The XC energy functional in LDA is written as:

$$E_{\mathrm{xc}}^{\mathrm{LDA}}[n(\mathbf{r})] = \int \epsilon_{\mathrm{xc}}[n(\mathbf{r})]n(\mathbf{r})d\mathbf{r}, \tag{2.25}$$

where $\epsilon_{\mathrm{xc}}[n(\mathbf{r})]$ is the XC energy per particle of a uniform electron gas of density $n(r)$. This term can be divided into exchange and correlation terms $\epsilon_{\mathrm{xc}}[n(\mathbf{r})] = \epsilon_{\mathrm{x}}[n(\mathbf{r})] + \epsilon_{\mathrm{c}}[n(\mathbf{r})]$ which leads to

$$E_{\mathrm{xc}}[n(\mathbf{r})] = E_{\mathrm{x}}[n(\mathbf{r})] + E_{\mathrm{c}}[n(\mathbf{r})]. \tag{2.26}$$

The exchange energy of the homogeneous electron gas (HEG) has an analytical form:

$$E_x^{\mathrm{LDA}} = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} \int n^{4/3}(\mathbf{r})d\mathbf{r}. \tag{2.27}$$

The form of the correlation energy is unknown, but accurate approximations to it obtained from Quantum Monte-Carlo calculations exist [128]. For systems such as bulk metals where the electron density varies very slowly, the LDA is quite a good approximation. However, it is known to fail for cases where the electron density cannot be taken as uniformly distributed.

The generalized gradient-approximated (GGA) functionals are the most straightforward extension of LDA to inhomogeneous systems. This class of XC functionals, also known as semi-local functionals, incorporate the gradient of the electron density $\nabla n(\mathbf{r})$ to account for

non-locality:

$$E_{\mathrm{xc}}^{\mathrm{GGA}}[n(\mathbf{r})] = \int f(n,\nabla n)d\mathbf{r} = \int \epsilon_{\mathrm{xc}}(n(\mathbf{r}))F_{\mathrm{xc}}(n(\mathbf{r}),\nabla n(\mathbf{r}))n(\mathbf{r})d\mathbf{r}. \tag{2.28}$$

Numerous efforts have been made in recent years to design and parametrize a variety of GGA functionals. The most popular GGA functional is the PBE functional [129] which is a non-empirical functional, in the sense that all parameters are basic constants, and there is no parametrization dependence on experimental data. GGA functionals outperform the LDA in terms of total energies, atomization energies, energy barriers, and structural energy differences. When used to analyze the structure of molecules, the GGA functionals produce good results, however, they can greatly underestimate the binding energies of weakly bound systems [130].

Another type of functionals consist of mixing of Hartree-Fock-exchange energy with the exchange and correlation of the semi-local functional proposed by Becke [131]:

$$E_{\mathrm{xc}}^{\mathrm{hybrid}}[n(\mathbf{r})] = \alpha E_{\mathrm{X}}^{\mathrm{HF}}[n(\mathbf{r})] + (1-\alpha)E_{\mathrm{x}}^{\mathrm{GGA}}[n(\mathbf{r})] + E_{\mathrm{c}}^{\mathrm{GGA}}[n(\mathbf{r})], \tag{2.29}$$

where the parameter α regulates the mixing. The exact exchange is taken from Hartree–Fock theory [132]:

$$E_{\mathrm{X}}^{\mathrm{HF}} = -\frac{1}{2}\sum_{i,j}\iint \phi_i^*(\mathbf{r}_1)\,\phi_j^*(\mathbf{r}_2)\,\frac{1}{r_{12}}\,\phi_j(\mathbf{r}_1)\,\phi_i(\mathbf{r}_2)\,d\mathbf{r}_1 d\mathbf{r}_2 \tag{2.30}$$

There are hundreds of different functionals nowadays [133] and an informal classification, where XC functionals of similar capabilities are placed at the rungs of the "Jacob's ladder" was proposed by Perdew [134]. Comprehensive information about different types of functionals can be found in the literature [135]. The functional that will be mostly used in this thesis is PBE and in some cases PBE0 [136, 137]. For PBE, the XC functional is expressed as

$$E_{\mathrm{xc}}^{\mathrm{PBE}}[n(\mathbf{r})] = E_{\mathrm{x}}^{\mathrm{PBE}}[n(\mathbf{r})] + E_{\mathrm{c}}^{\mathrm{PBE}}[n(\mathbf{r})], \tag{2.31}$$

where the exchange functional $E_{\mathrm{x}}^{\mathrm{PBE}}[n(\mathbf{r})]$ is

$$E_{\mathrm{x}}^{\mathrm{PBE}}[n(\mathbf{r})] = \int n(\mathbf{r})\epsilon_{\mathrm{x}}^{\mathrm{LDA}}[n(\mathbf{r})]F_{\mathrm{x}}(s)\mathrm{d}\mathbf{r}, \tag{2.32}$$

where

$$\epsilon_{\mathrm{x}}[n(\mathbf{r})] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} n(\mathbf{r})^{1/3} \tag{2.33}$$

is the exchange energy density in the uniform electron gas (see Eq. 2.27) with

$$n(\mathbf{r}) = \frac{3}{4\pi}\frac{1}{r_s^3}, \tag{2.34}$$

where r_s denotes the radius of a sphere that contains one electron on average. $F_{\mathrm{X}}(s)$ denotes the GGA enhancement factor depending on a dimensionless density gradient s which is defined as $s = |\nabla n(\mathbf{r})|/(2k_F n(\mathbf{r}))$, where $k_F = \left(3\pi^2 n(\mathbf{r})\right)^{1/3}$ is the Fermi wave vector. The enhancement factor $F_{\mathrm{X}}(s)$ has to satisfy a formal conditions [129] and is expressed as

$$F_{\mathrm{X}}(s) = 1 + \kappa - \frac{\kappa}{1 + \mu s^2/\kappa}, \tag{2.35}$$

with $\mu = \beta\left(\pi^2/3\right)$, $\beta = 0.066725$, and $\kappa = 0.804$.

The correlation energy in PBE is expressed as the local correlation plus a correction term $H(r_s, \zeta, t)$ [129] and has the following form

$$E_{\mathrm{c}}^{\mathrm{PBE}}[n(\mathbf{r})] = \int d\boldsymbol{r}\, n(\mathbf{r})\left[\varepsilon_{\mathrm{c}}^{\mathrm{LDA}}(r_s, \zeta) + H(r_s, \zeta, t)\right] \tag{2.36}$$

where $\varepsilon_{\mathrm{c}}^{\mathrm{LDA}}$ is the correlation energy density in PW-LDA approximation [138], ζ is the magnetization density and t is dimensionless gradient (See details in Ref. [129, 139]).

The functional PBE0 mixes $a_0 = 0.25$ of exact exchange (E EX) to the PBE functional, having the form:

$$E_{xc}^{PBE0} = a_0 E_X^{HF} + (1 - a_0) E_x^{PBE} + E_c^{PBE}, \tag{2.37}$$

where the value $a_0 = 0.25$ was chosen based on considerations from fourth order many-body perturbation theory [140].

2.4 Long-range van der Waals interactions

Even though the exact DFT would include all correlation effects, the approximations representing the state-of-the-art density functionals are typically unable to describe dispersion and non-local correlation effects by construction [141]. However, the accurate incorporation of weak vdW interactions are especially crucial for calculation of the properties of such systems as biomolecules [142–144], molecular crystals [145, 146] and interface systems [130, 147–154] due to their collective nature. Even if after adsorption the molecule covalently binds to the surface, the accurate description of the vdW interactions are crucial for such kind of systems that makes it possible to obtain deviations in theoretical adsorption heights within 0.1-0.2 Å within experimental values [78]. A theoretically accurate method for a description of the vdW interactions was recently developed and takes into account electronic screening and the many-body nature of the dispersion term [155].

There are many groups working on inclusion of the vdW corrections and introduction to different approaches that also can be classified in the similar way as well-known “Jacob’s ladder” of functionals introduced by Perdew [134] can be found in the literature [156]. One of the most wide-spread way to account for vdW interactions nowadays are so-called pairwise-additive dispersion correction schemes, where vdW energies are calculated analytically after the convergence of the electronic self-consistency cycle [157–164]. The total energy in this

case will be:

$$E_{\text{tot}} = E_{\text{DFA}} + E_{\text{vdW}}, \tag{2.38}$$

where E_{DFA} is the total energy of the system obtained with particular DFA. The dispersion contribution E_{vdW} is defined as the interaction between mutually induced charge fluctuations arising from the instantaneous quantum mechanical excitations of electrons. At large distances, the dispersion interaction can be expressed via a multipolar expansion of the Coulomb potential, as a series in inverse powers of R and, by taking the first term $1/R^6$ that corresponds to the instantaneous induced dipole–induced dipole interaction that is the main contribution, we get:

$$E_{\text{vdW}} = -\frac{1}{2}\sum_{A\neq B}\frac{C_{6,AB}}{R_{AB}^6}, \tag{2.39}$$

where the indices A and B refer to two different atoms, and the sum runs over all possible combinations of atoms in the system, $C_{6,AB}$ is the dispersion coefficient of the two atoms and R_{AB} is the interatomic distance between them. One drawback of using formula 2.39 is the fact that for small interatomic distances it is clearly diverges, and so the damping function $f_{\text{damp}}(R_{AB})$ is needed to remove this divergence and also to minimize the overlap between the short-range contributions of the XC functional and of the vdW correction. In this case the formula for dispersion correction looks like:

$$E_{\text{vdW}} = -\frac{1}{2}\sum_{A\neq B}\frac{C_{6,AB}}{R_{AB}^6} f_{\text{damp}}(R_{AB}). \tag{2.40}$$

In the simplest approach the $C_{6,AB}$ coefficients are constant and isotropic. Such methods do not include many-body dispersion effects such as screening in metals [165] and keeping of the C_6 coefficients constant neglect the environmental contributions. Obtaining the C_6 coefficients could involve experimental ionization potentials and polarizabilities [166], however, this imposes a constraint on the list of components that may be handled to those found in organic compounds.

The next step to increase the accuracy of the dispersion correction is to include environment-dependent C_6 corrections where the dispersion coefficient of an atom in a molecule depends on the effective volume of the atom. The most popular schemes developed in this direction are DFT-D3 by Grimme [159], Becke-Johnson model [167] and the method of Tkatchenko and Scheffler (TS) [163]. Grimme's model employs the concept of fractional coordination numbers where the function calculating the number of neighbors continuously interpolates between the tabulated reference values. Becke-Johnson model exploits the fact that around an electron there will be a XC hole that produces non-zero dipole and higher-order electrostatic moments causing polarization in other atoms leading to an attractive dipole-induced dipole interaction.

The way of fitting of the damping function is crucial since it defines the shape of the binding curve that tat has to be compatible to XC functional of choice and to definition of vdW

radii of atoms [156] and giving rise to broader family of the different approaches [158, 168–170].

The TS approach is much more cost effective compared to the Becke-Johnson model and uses precalculated C_6 coefficients instead of hole dipole moments. The extension of the vdW-TS method tailored to model interface systems was used in this work and its scope will be described in more details in the further section.

2.5 Tkatchenko-Scheffler vdW method

The energy in the TS method is computed using the formula in Eq. 2.39, which is a sum over pairwise interatomic C_6/R_6 terms. The expression for the isotropic C_6 coefficients that describe the vdW interactions between two well-separated fragments is derived from Casimir-Polder formula [171]:

$$C_{6,AB} = \frac{3}{\pi}\int_0^\infty \alpha_A(i\omega)\alpha_B(i\omega)d\omega, \tag{2.41}$$

where $\alpha_A(i\omega)$ is the average dynamic polarizability for atom A and ω is the excitation frequency. Retaining only the leading term of the Padé [172] series, the polarizability of spherical free atoms can be approximated and gives:

$$\alpha_A^1(\omega) = \frac{\alpha_A^0}{1-(\omega/\omega_A)^2}, \tag{2.42}$$

where α_A^0 is the static polarizability of atom A and ω_A is the effective excitation frequency. After substitution into Eq. 2.41 with the static polarizabilities the integral can be solved analytically and the C_6 coefficient can be written as:

$$C_{6,AB} = \frac{3}{2}\alpha_A^0\alpha_B^0\frac{\omega_A\omega_B}{(\omega_A+\omega_B)}. \tag{2.43}$$

For the homonuclear $C_{6,AA}$ coefficient, the effective excitation frequency of atom A can be expressed in terms of the static polarizability:

$$\omega_A = \frac{4}{3}\frac{C_{6,AA}}{\left(\alpha_A^0\right)^2}. \tag{2.44}$$

After that the expression for $C_{6,AB}$ can be obtained by substitution of effective excitation frequencies in Equation 2.43:

$$C_{6,AB} = \frac{2C_{6,AA}C_{6,BB}}{\left(\frac{\alpha_B^0}{\alpha_A^0}C_{6,AA}+\frac{\alpha_A^0}{\alpha_B^0}C_{6,BB}\right)}. \tag{2.45}$$

Then, the C_6 coefficients can be accurately computed using the free-atom parameters α_A^0 and $C_{6,AA}$ obtained from from high-level self-interaction corrected time-dependent DFT reference data [173].

from high-level self-interaction corrected TDDFT reference data For the atoms inside a molecule or solid the proceeding formulation can be adapted to make the TS scheme environment-dependent by introducing the proportional coefficient k, which comes from assuming that the polarizability depends linearly on volume [174]: $k_A^{\text{free}}\,\alpha_A^{\text{free}} = V_A^{\text{free}}$, where "free" refers to free atoms. By obtaining the effective volume of the atom inside a molecule or solid the parameter k can be computed as ratio between effective volume and its free value in order to rescale all the quantities introduced earlier. In the TS scheme the effective volume is obtained from the electron density of the system and the Hirshfeld partitioning of the density (via Hirshfeld weight $w_A(\mathbf{r})$) [175]:

$$\frac{k_A^{\text{eff}}\alpha_A^{\text{eff}}}{k_A^{\text{free}}\alpha_A^{\text{free}}} = \frac{V_A[n(\mathbf{r})]}{V_A^{\text{free}}} = \frac{\int r^3 w_A(\mathbf{r}) n(\mathbf{r}) d\mathbf{r}}{\int r^3 n_A^{\text{free}}(\mathbf{r}) d\mathbf{r}} = \gamma_A[n(\mathbf{r})], \tag{2.46}$$

where the electron density $n(\mathbf{r})$ is taken from DFT calculations, $n_A^{\text{free}}(\mathbf{r})$ is the free atom spherically averaged reference density and $r = |\mathbf{r} - \mathbf{R}_A|$ is the distance between the nucleus of atom A and the point $\mathbf{r}$. The effective quantities are then determined from the free ones as:

$$\alpha_A^{0,\text{eff}} = \gamma_A[n(\mathbf{r})]\alpha_A^{0,\,\text{free}}, \tag{2.47}$$

$$C_{6,AA}^{\text{eff}} = \left(\gamma_A[n(\mathbf{r})]\right)^2 C_{6,AA}^{\text{free}}, \tag{2.48}$$

$$R_A^{0,\text{eff}} = \left(\gamma_A[n(\mathbf{r})]\right)^{1/3} R_A^{0,\,\text{free}}, \tag{2.49}$$

where the R is vdW radius. The TS scheme was tested on a database of 1225 intermolecular C_6 pairs and showed a mean absolute error of 5.5% compared to experimental results irrespective of the employed XC functional [163].

As was mentioned above, the sum of pairwise $C_{6,AB}/R_{AB}^6$ terms diverges for small interatomic distances and the damping function has to be introduced (Eq. 2.40). The damping function in the case of the TS method is a Fermi-type function:

$$f_{damp}^{AB}\left(R_{AB}, R_{AB}^0[n(\mathbf{r})]\right) = \frac{1}{1+\exp\left[-d\left(\frac{R_{AB}}{s_R R_{AB}^{0,\,\text{eff}}[n(\mathbf{r})]}-1\right)\right]}, \tag{2.50}$$

where R_{AB} is the interatomic distance, $R_{AB}^{0,\,\text{eff}} = R_A^{0,\,\text{eff}} + R_B^{0,\,\text{eff}}$ is the sum of the vdW radii associated with atoms A and B that depend on the electron density through the effective volume (Eq. 2.49) and parameters d and s_R are empirical values that need to be determined for a given XC functional. The parameters d, that affects the steepness of the damping, and

the parameter s_R, that scales the vdW radii and regulates the extent of the vdW correction for a given XC functional, were fitted for different functionals with use of S22 database [176].

2.6 Tkatchenko-Scheffler vdW^{surf} method

In order to include the non-local collective response of the substrate surface in the vdW energy the extension of the TS-vdW scheme (vdW^{surf} [177]) for modelling of interfaces relies on Lifshitz-Zaremba-Kohn (LZK) theory [178, 179] for the vdW interaction between an atom and a solid surface. This leads to a set of C_6 coefficients that incorporate dielectric screening of the bulk, and in the case of solids the reference vdW parameters have to be determined taking into account atom-in-a-solid environmental effects [180]. In LZK theory the atom-surface dispersion interaction beyond the distance of the orbital overlap is given by [179, 181]:

$$E_{\mathrm{vdW}} \simeq -\frac{C_3^{aS}}{(H-H_0)^3}, \tag{2.51}$$

where H is the distance between an adsorbate atom a and the topmost layer of the surface S. The reference plane H_0 can be obtained from the jellium model yeilding $H_0 = h/2$, where h is the interlayer distance of the solid. The term C_3^{aS} describes the dielectric response of the bulk solid to the instantaneous dipole moment of particles and depends on the dipole polarizability $\alpha(i\omega)$ of the adsorbate and dielectric function $\epsilon_S(i\omega)$ of the solid:

$$C_3^{aS} = \frac{1}{4\pi}\int_0^{+\infty} \alpha(i\omega)\left[\frac{\epsilon_S(i\omega)-1}{\epsilon_S(i\omega)+1}\right] d\omega. \tag{2.52}$$

The screening effects inside the bulk are incorporated in the Eq. 2.52 by dependence on the dielectric function $\epsilon_S(i\omega)$. Next step is determination of the vdW interaction between an adsorbate atom a with a solid S by a summation of the pair potentials $-C_6/R^6$ between an atom a and atoms s in the infinite half-space infinite of the solid S. After that, the connection to LZK expression can be achieved by the relation:

$$C_{3,aS} = n_S\left(\frac{\pi}{6}\right)\mathrm{C}_{6,as}, \tag{2.53}$$

where n_S is the number of atoms per unit volume in the bulk of the substrate, and

$$\mathrm{C}_{6,as} = \frac{2\mathrm{C}_{6,aa}\mathrm{C}_{6,ss}}{\frac{\alpha_s^0}{\alpha_a^0}\mathrm{C}_{6,aa} + \frac{\alpha_a^0}{\alpha_s^0}\mathrm{C}_{6,ss}}, \tag{2.54}$$

where the $\mathrm{C}_{6,as}$, α_s^0 and R_s^0 are the new set of parameters that depend on dielectric function $\epsilon_S(i\omega)$ and thus inherit the many-body collective response (screening) of the solid. The only difference from the TS method is that the effective quantities, that were including the

effects of polarization with use of the Hirshfeld weight, are now obtained from the LZK parameters and not from the free atom reference. The dielectric function can be computed from first-principles and was shown to reasonably agree with the results obtained from reflection electron energy-loss spectroscopy (REELS) experiments [182]. In case of transition metals, the inclusion of collective response of the solid leads to reducing the C_6 coefficients by up to a factor of ten compared to reference free atom values [130].

Investigating interface systems, the inclusion of the vdW parameters should only be applied when appropriate: for example inside metal surfaces there are already good approximations from DFA functionals and inclusion of the vdW interactions, even if the results are improved compared to experimental, can be considered as effect of cancellation of the errors [150].

2.7 Basis sets

In order to solve the set of single-particle KS eigenvalue equations (Eq. 2.22) it is a common technique to use basis functions to expand the single-particle orbitals:

$$\phi_\nu(\mathbf{r}) = \sum_n c_{n\nu}\xi_n(\mathbf{r}) \tag{2.55}$$

A basis set allows us to write the Schrödinger equation as a generalized eigenvalue problem:

$$\sum_n h_{mn}c_{n\nu} = \epsilon_\nu \sum_n s_{mn}c_{n\nu}, \tag{2.56}$$

where $h_{mn} = \left\langle \xi_m \middle| \hat{h}^{KS} \middle| \xi_n \right\rangle$ is the matrix element of the Hamiltonian, and $s_{mn} = \langle \xi_m | \xi_n \rangle$ is the overlap matrix element. A suitable choice of basis functions depends on the system under investigation. For this thesis we use the all-electron/full potential Fritz Haber Institute "ab initio molecular simulations" (FHI-aims) code [183, 184], which adopts the tabulated numeric atom-centered orbitals (NAO) basis functions of the form:

$$\xi_i(\mathbf{r}) = \frac{u_i(r)}{r} Y_{lm}(\Omega) \tag{2.57}$$

where the function $u_i(r)$ has radial symmetry and is numerically tabulated and $Y_{lm}(\Omega)$ are the spherical harmonics. The particular form of the NAOs allows to include the radial functions of free-atom orbitals and can be constructed using a Schrödinger-like radial equation:

$$\left[-\frac{1}{2}\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2} + v_i(r) + v_{\text{cut}}(r)\right] u_i(r) = \epsilon_i u_i(r) \tag{2.58}$$

where l is the angular quantum number. The potential $v_i(r)$ defines the shape of $u_i(r)$ and the term $v_{\text{cut}}(r)$ is the confining potential, which ensures a decay to zero of the radial functions. Minimal basis consists of the core and valence functions of spherically symmetric free atoms by setting $v_i(r)$ to the self-consistent free-atom radial potential $v_{\text{at}}^{\text{free}}$. The construction of the accurate and transferable basis sets that allow up meV-level total energy convergence relies

on addition of the candidate functions from a large pool of different radial functions (e.g. hydrogen-like, cation-like or atom-like) with different confinement potential to minimal basis set until no further significant improvement on total energy results [185].

The analytical form of the confining potential is not unique and along with a smooth decay, it must ensure that the function and its derivatives do not have any discontinuities. The confining potential in FHI-aims is provided by:

$$v_{\text{cut}}(r) = \begin{cases} 0 & r \leq r_{\text{onset}} \\ \frac{s}{(r-r_{\text{cut}})^2} \exp\left(\frac{w}{r-r_{\text{onset}}}\right) & r_{\text{onset}} < r < r_{\text{cut}} \\ +\infty & r \geq r_{\text{cut}}\,, \end{cases} \tag{2.59}$$

where s is a global scaling parameter and $w = r_{\text{cut}} - r_{\text{onset}}$ sets the width of the region, where potential is defined. The selection of the parameters r_{cut} and r_{onset}, is essential for both the accuracy of the results and numerical efficiency. For example, a large value of r_{cut} would result in extended radial functions, increasing the computational cost of the calculation. Setting r_{onset} to a very small value will result in unphysical results since radial functions will be limited in a very narrow region surrounding the atom.

In the case of periodic systems, the Kohn-Sham Eqs. 2.56 are **k**-space dependent. This leads to separate matrices $h_{mn}(\mathbf{k})$, $s_{mn}(\mathbf{k})$ and solutions $\phi_{\nu,\mathbf{k}}(\mathbf{r})$ that have to be obtained for different **k**-points in the first Brillouin zone. For that Bloch-like generalized basis functions $\varphi_i(\mathbf{r})$ that are centered in unit cells shifted by translation vectors $\boldsymbol{T}(\boldsymbol{N})[\boldsymbol{N} = (N_1, N_2, N_3)]$ are introduced in the code:

$$\chi_{i,\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\boldsymbol{N}} \exp[i\boldsymbol{k}\cdot\boldsymbol{T}(\boldsymbol{N})]\cdot\varphi_i[\boldsymbol{r} - R_A + \boldsymbol{T}(\boldsymbol{N})]\,. \tag{2.60}$$

Such definition brings **k**-dependent matrix elements

$$\begin{aligned} h_{ij}(\boldsymbol{k}) &= \left\langle \chi_{i,\boldsymbol{k}} \left| \hat{h}^{\text{KS}} \right| \chi_{j,\boldsymbol{k}} \right\rangle \\ &= \sum_{\boldsymbol{M},\boldsymbol{N}} \exp\{i\boldsymbol{k}\cdot[\boldsymbol{T}(\boldsymbol{N}) - \boldsymbol{T}(\boldsymbol{M})]\} \left\langle \varphi_{i,\boldsymbol{M}} \left| \hat{h}^{\text{KS}} \right| \varphi_{j,\boldsymbol{N}} \right\rangle \end{aligned} \tag{2.61}$$

with the real-space basis functions $\phi_i(\mathbf{r},\mathbf{M})$ and $\phi_i(\mathbf{r},\mathbf{N})$ that are centered in different unit cells **M** and **N**. In practice, all integration points and pieces of are mapped back to the original unit cell in order to avoid breaking down lattice sums in Eq. 2.61 due to periodicity since the integration volumes could extend over several unit cells in the integrals $\left\langle \varphi_{i,\boldsymbol{M}} \left| \hat{h}^{\text{KS}} \right| \varphi_{j,\boldsymbol{N}} \right\rangle$. Since all basis functions are bounded by the confinement potential, only a finite number of inequivalent real-space matrix elements are non-zero.

2.8 Charge transfer and binding energy calculations

The interaction of individual molecules with metallic surfaces constitutes one of the central topics of surface science partially because experimental techniques such as the STM could be

easily operated on conductive substrates. The final electronic structure of interface system can be calculated with accurate computational methods such as DFT. Understanding of the mechanisms that leads to the particular adsorption pattern of the molecule and identifying the molecular donor/acceptor parts can give more insights towards rational design and self-assembly processes of the interfaces. In this thesis we are interested in both neutral and positively charged molecules adsorbing on metallic surfaces and in this section we would like to address the procedure that we use to investigate adsorption.

While modelling the interface structures we must use periodic boundary conditions (PBC). Organic-inorganic interface systems could incur a dipole moment in the direction perpendicular to the surface due to charge rearrangements at the interface or due to polar adsorbates which leads to appearing of the electric field that generates a potential gradient in the unit cell compensating the potential shift induced by the system's dipole moment. The interaction of the interface dipole with this electric field also leads to charge rearrangements between ends of the entire slab that in turn affects the total energy of the system. The most common way to deal with the spurious polarization is to introduce discontinuity in the electrostatic potential within the vacuum region and referred in the literature as "dipole correction" [186] and to use large vacuum regions since the magnitude of this spurious electric field depends inversely on the thickness of the vacuum region. In FHI-aims, the magnitude of dipole correction is obtained from the gradient of the long-range Hartree potential term of the Ewald sum (which is evaluated in reciprocal space). The surface plane is placed parallel to $x\,y$ plane in the deep vacuum region that is further than 6 Å away from the nearest atom.

Simulation of charged unit cells is required for several physical problems such as dealing with charged defects [187, 188] or when the electron transfer from the adsorbed molecules is quenched and they can exhibit metastable charge-states [189, 190]. This brings the problem that the repeated slab approach imply that all unit cells in the system carry a charge and such a periodic arrangement of charges results in a diverging energy that prevents convergence of the self-consistent field (SCF) algorithm. Basically, there is a Coulomb interaction between the delocalized homogenous background charge and the excess charge that is localized in the slab that significantly contributes to the total energy of the system. The spurious energy contribution originates from the spurious net dipole of the unit cell and, hence, scales linearly with the thickness of the vacuum region. Two types of approaches were developed to deal with such cases. The first class neutralizes the interaction between charged cells perpendicular to the substrate *via a posteriori* correction based on the dielectric profile of the interface [191] or by interfering Poisson equation that describes electrostatic potential [192, 193]. The second class intentionally adds spatially localized countercharges into the system ensuring charge neutrality of the such that leads to the absence of compensating background charge. The virtual crystal approximation [187] provides a fixed number of free charge carriers per volume, the Charge Reservoir Electrostatic Sheet Technique [194] models the countercharges as a charged sheet, which is placed below the substrate and the generalized dipole correction approach [195] introduces a monopole sheet as a "computational electrode" and a dipole layer in the vacuum region.

In our case for adsorption of the both neutral and positively charged species the unit cell is set to have neutral charge. After adsorption of the charged molecules on the surface the charge transfer will occur from surface since it has infinite pool of electrons that will neutralize the unit cell. This comes from the fact that energy of lowest unoccupied orbital of the positively charged molecules are way below the Fermi energy of the metallic surfaces. Having that, the binding energies for neutral molecule adsorbed on the surface were calculated as

$$E_{\mathrm{b}} = E_{\mathrm{mol@surf}} - E_{\mathrm{surf}} - E_{\mathrm{mol}}, \tag{2.62}$$

where $E_{\mathrm{mol@surf}}$ corresponds to the total energy of the interface, E_{surf} is the total energy of the pristine metallic slab and E_{mol} the total energy of the lowest energy gas-phase conformer.

For charged molecules, we considered the binding energy of a two-step reaction. *First*, the interface is formed between the charged molecule and the clean surface:

$$E_{\mathrm{b1}} = E_{\mathrm{mol^+@surf}} - E_{\mathrm{surf}} - E_{\mathrm{mol^+}}, \tag{2.63}$$

where $E_{\mathrm{mol^+}}$ is the total energy of the most stable gas-phase conformer of the isolated charged molecule. *Second*, an electron from the metal neutralizes the unit cell where the adsorbed molecule is located, yielding

$$E_{\mathrm{b2}} = E_{\mathrm{mol@surf}} - E_{\mathrm{mol^+@surf}} - E_f, \tag{2.64}$$

where E_f corresponds to the Fermi energy of the metallic surface. The final binding energy is thus considered to be

$$E_{\mathrm{b}}^{+} = E_{\mathrm{b1}} + E_{\mathrm{b2}} = E_{\mathrm{mol@surf}} - E_{\mathrm{surf}} - E_f - E_{\mathrm{mol^+}}. \tag{2.65}$$

To address charge rearrangements after adsorption on the surface, we compute the electron density differences for selected with

$$\Delta\rho = \rho_{\mathrm{mol@surf}} - \rho_{\mathrm{surf}} - \rho_{\mathrm{mol}}, \tag{2.66}$$

and in the case of neutral molecule and

$$\Delta\rho^{(+)} = \rho_{\mathrm{mol@surf}} - \rho_{\mathrm{surf}} - \rho_{\mathrm{mol^{(+)}}}, \tag{2.67}$$

in the case of charged molecule. In these expressions, $\rho_{\mathrm{mol@surf}}$ is the total electron density of the interface, ρ_{surf} is the electron density of the slab without molecule, and ρ_{mol} and $\rho_{\mathrm{mol^+}}$ are electron densities of neutral and charged molecules with the same geometries as in interface. The + sign denotes that the final density difference integrates to +1 electron in the case of charged molecule. These densities allow us to identify charge build up on

particular functional group, as well as charge transfer to the surface.

2.9 Modelling of the STM images

One of the ways to validate theoretical investigations is to directly compare experimental measurements with theoretically modelled properties of the system. In that respect modelling of STM images can be a very useful tool to identify the system geometry. Using Bardeen's expression [196] one can write the current flowing from a metallic tip to the sample as

$$I_{t\to s} = \frac{2\pi e}{\hbar}\int |M_{ts}|^2 N_t(E-eV)N_s(E)f_t(E-eV)\big[1-f_s(E)\big]\,dE, \tag{2.68}$$

where V is the applied voltage, $N_t(E)$ and $N_s(E)$ are the density of states of the tip and the sample respectively, $f(E)$ is their Fermi-Dirac distribution. The effective matrix element M_{ts} couples a tip wave function, Ψ_t, to a substrate wave function, Ψ_s, by the expression

$$M_{ts} = \frac{\hbar^2}{2m}\int \left(\Psi_t^*\nabla\Psi_s - \Psi_s\nabla\Psi_t^*\right) d\boldsymbol{S}, \tag{2.69}$$

where the integral is taken over a surface separating the tip and sample.

For modelling STM images one of the most widely used approaches is the scheme proposed by Tersoff and Hamman [197]. One of the main assumptions made within this model is that complex electronic structure of the tip is assumed to be simple atomic s-wave-function since only the orbitals that localized at the outermost tip atom are important for tunneling process taking into account that this wave-function decays exponentially into the vacuum. The total current flowing from the tip to the sample within the zero temperature approximation and low bias voltage is:

$$I = \frac{2\pi e^2}{\hbar} V \sum_N |M_{ts}|^2\,\delta(E_s - E_F)\,\delta(E_t - E_F), \tag{2.70}$$

where V is the voltage applied and the energy conservation is ensured by δ-functions.

The advantage of the Tersoff-Hamann theory is that the tip Ψ_t wavefunction can be modelled as a solution in a locally spherical potential with curvature R about its center r and asymptotically the is chosen to have the form of an s-wave. So the matrix element M_{ts} is proportional to the sample wavefunction evaluated at the tip center of curvature ($M_{ts} \propto \Psi_s(r_0)$) leading to:

$$I \propto V N_t(E_F)\sum_s |\Psi_s(r_0)|^2\,\delta(E_s - E_F), \tag{2.71}$$

where the sum represents the local density of states of the sample (LDOS) around the Fermi level evaluated at the tip center.

2.10 Force field methods

In previous sections we addressed the methods for simulations of the interface systems at quantum mechanical (QM) levels of theory of high computational costs, applicable only to systems of few hundreds of atoms. In this section we briefly describe the applications and limitations of the FF methods that are less accurate but orders of magnitude cheaper to perform and thus enable the simulation of the systems that consist of millions of atoms.

Most commonly within classical FFs PES functions are expressed as a sum of bonded and nonbonded interaction terms. Hence, the description of a of FF is given by its potential energy $E_{\text{pot}}^{\text{FF}}\left(\boldsymbol{R}^N\right)$ that is given as a function of positions $\boldsymbol{R}_1,\ldots,\boldsymbol{R}_N$ the N nuclei of the system is given by

$$E_{\text{pot}}^{\text{FF}}\left(\boldsymbol{R}^N\right)=E_{\text{bonded}}\left(\boldsymbol{R}^N\right)+E_{\text{nonbonded}}\left(\boldsymbol{R}^N\right). \tag{2.72}$$

For example in CHARMM22 [198], one of the popular FF for simulation of biomolecules, the "bonded" terms are of the following form:

$$E_{\text{bonds}}\left(\boldsymbol{R}^N\right)=\sum_{\text{bonds}}\frac{k_r}{2}(R-R_0)^2 \tag{2.73}$$

$$E_{\text{angles}}\left(\boldsymbol{R}^N\right)=\sum_{\text{angles}}\frac{k_\theta}{2}(\theta-\theta_0)^2 \tag{2.74}$$

$$E_{\text{torsions}}\left(\boldsymbol{R}^N\right)=\sum_{\text{torsions}}\frac{k_\tau}{2}(1+\cos(n\tau-\delta)) \tag{2.75}$$

$$E_{\text{impropers}}\left(\boldsymbol{R}^N\right)=\sum_{\text{impropers}}k_\omega(\omega-\omega_0)^2 \tag{2.76}$$

$$E_{\text{Urey-Bradley}}\left(\boldsymbol{R}^N\right)=\sum_{\text{Urey-Bradley}}k_u(u-u_0)^2 \tag{2.77}$$

where n is the multiplicity of the function, δ is the phase shift, k_R, k_θ, k_τ, k_ω, and k_u are the bond, angle, dihedral angle, improper dihedral angle and Urey–Bradley force constants, respectively; R, θ, τ, ω, and u are the bond length, bond angle, dihedral angle, improper torsion angle and Urey–Bradley 1,3-distance, respectively, with the subscript zero representing the equilibrium values for the individual terms. "Nonbonded" interaction terms are included for all atoms separated by three or more covalent bonds and include electrostatic interactions

$$E_{\text{Coulomb}}\left(\boldsymbol{R}^N\right)=\sum_{\text{A B}}\frac{q_A q_B}{R^{AB}}, \tag{2.78}$$

and vdW intra- and inter-molecular interactions

$$E_{\text{vdW}}\left(\boldsymbol{R}^N\right)=\sum_{\text{A B}}\varepsilon_{AB}\left[\left(\frac{R_0^{\text{vdW}}}{R^{AB}}\right)^{12}-2\left(\frac{R_0^{\text{vdW}}}{R^{AB}}\right)^{6}\right], \tag{2.79}$$

where q_A is the charge of the atom A, R^{AB} is the distance between atoms A and B and ε_{AB} is the energy required to separate the atoms. In the Lennard-Jones (LJ) potential above, the R_0^{vdW} term is not the minimum of the potential, but rather where the LJ potential is zero.

In recent years a lot of effort has been invested to adapt biomolecular FFs to the simulation of interfaces between biomolecules and inorganic materials which resulted in creation of a class of general bio-inorganic FFs. One of these FFs is called INTERFACE FF, in which LJ parameters for eight neutral face-centered cubic (fcc) metals (Ag, Al, Au, Cu, Ni, Pb and Pd) based on experimentally determined densities and surface tensions under ambient conditions have been added to CHARMM22 for the simulation of metal surfaces in contact with biomolecules [87, 199, 200].

One of the greatest limitations of most common FFs is the inability to simulate chemical reactions that involve the formation or dissociation of chemical bonds. Modelling of chemisorption processes require QM or employing of reactive FFs that have to be used with great caution [92, 93]. Second disadvantage is the sensitivity of FF parameters to deviations from the reference state, for which they were derived that imply nontransferability of the parameters to different systems.

For much broader overview of different FFs designed for modelling of the protein-inorganic surface interaction can be found in the review [79]. For modelling of AA adsorption on metallic surfaces using INTERFACE FF we use NAMD package [201] and compare results obtained with DFT in the Section 4.5.

And to the man he said, "Since you listened to your wife and ate from the tree whose fruit I commanded you not to eat, the ground is cursed because of you. All your life you will struggle to scratch a living from it."

Genesis 3.17 about THE curse of dimensionality (author's note)

3

Structure search and analysis of conformational spaces

A major challenge in computational chemistry is the search of low-energy conformers for a given flexible molecule. Organic molecules that are flexible can adopt a number of energetically favourable conformations with varying chemical and physical characteristics (Fig. 3.1 a). As a result, examining the attributes of a single randomly created conformer may result in incorrect results. The environment, as well as interactions with other molecules and surfaces, can all impact the likelihood of a given shape being adopted (Fig. 3.1 b). Further, it has been shown that the bioactive conformation of drug-like molecules can be higher in energy than the respective global minimum [202]. Structures that can be trapped in metastable local minima during growth process, can be accessed at finite temperatures or under pressure. As a result, we aim at not just finding the conformer expressing the PES's global minimum, but at covering relevant portions of the available conformational space.

3.1 Global structure search techniques

Finding the most stable configurations of assembly of atoms is a challenging problem due to the fact that the number of stationary points in the particular PES can grow exponentially with the number of atoms in the system. Finding the global minimum of the system, in general, requires searching through many local minima which is effectively prohibitive for large systems due to finite computational resources available. There is a broad field of computational search techniques. Below we will briefly mention techniques based on MD and then present a more in-depth characterization of other techniques more directly relevant

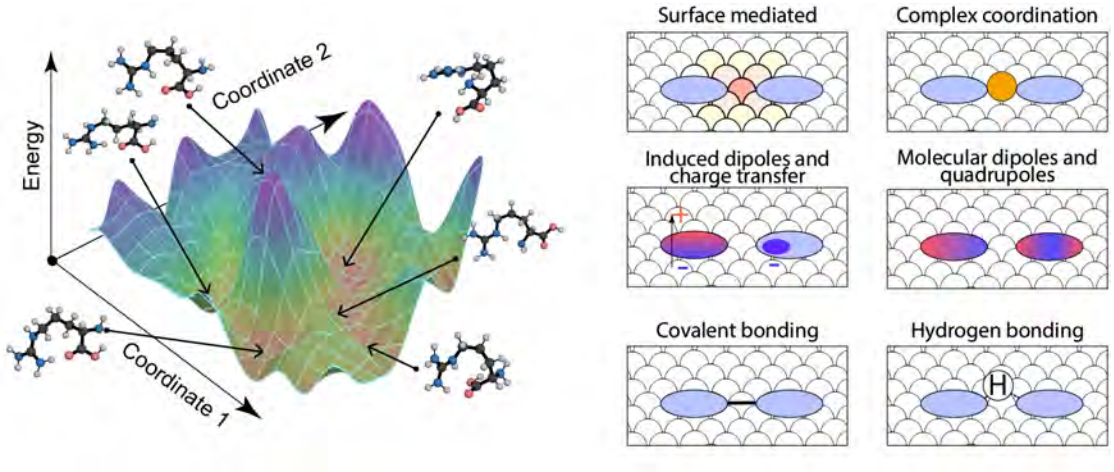


Figure 3.1 – a) Pictorial representation of the multiple local minima of PES of a flexible molecule with respect to arbitrary coordinates. b) Examples of complex interactions that appear during self-assembly processes on the surfaces.

for this thesis. But first of all, it should be mentioned, that performance of all algorithms that search for the global minimum of an energy function is the same when averaged over all possible energy functions - this is known in literature as “no free lunch theorem” [203]. This implies, that there is no possible way to find algorithm that would perform better than another in all scenarios.

3.1.1 MD-based techniques

Replica exchange molecular dynamics (REMD) simulations combine MD simulation with the Monte Carlo algorithm and are used to sample the configurational space of a system, e.g. at different temperatures or with a different Hamiltonian [204]. Structures locked in local energy minima can traverse the energy landscape by exchanging the replicas, improving Boltzmann-weighted sampling. Unlike a standard MD simulation, REMD allows sampling various configurations in different potential wells separated by huge energy barriers.

Umbrella sampling is another standard method for enhancing the sampling of configurational space of the system [205]. This technique defines a reaction coordinate as a link between two thermodynamic states. The reaction coordinate is usually determined based on a distance or an angle. The reaction coordinate is then divided into windows, each exposed to a bias (umbrella) potential. Each window has its simulation to sample the area around the associated coordinate point. The simulations are then reweighted to account for the biased ensembles using, for example, the weighted histogram analysis method (WHAM) [206], or its generalization [207]. Umbrella sampling method, for example, was used for simulations of adsorption of AA side chain analogues on the TiO_2(100) surface [208] using FF models. There are subtleties in determining the best computationally efficient approach to apply the umbrella sampling method, as outlined in the book [209].

MD simulations can be also evaluated using classical FFs followed by static DFT refinement of the obtained data [91, 210, 211]. The main disadvantage of such approaches is that the PES obtained from a FF or DFT can be very different, which will result in the need to reevaluate all the sampled geometries using more accurate methods in order to obtain correct hierarchy [212, 213].

3.1.2 Other techniques

One of the simplest methods to explore PES that allows to effectively overcome potential energy barriers is a random search. Random search implies that next trial structure is not dependent on the information that was already accumulated during the search. Of course, simply creation of assembly of atoms and calculation of their energies would be far from effective. Concerning investigation of adsorbates on surfaces, to make such a strategy efficient, one has to impose limits on the generated structures, by creating only "sensible" structures. Structures that have some of the atoms that are very close to each other cannot be in the local minimum. Those structures should not be investigated and this already significantly decreases number of candidates that have to be calculated. Having that one can apply criteria for non-bonded atoms, for example, their vdW radii should not overlap which prevents modelling of unwanted chemical reactions. Concerning molecular systems and surfaces, one already has *a priori* information on bonds in the system, most of which should not change. After that different structures can be generated and followed with geometry optimization (See Sec. 3.2) to find local minima. Application of the random structure search to investigate organic-inorganic materials, in particular, the procedure of generating different molecular conformers with respect to specified surrounding, will be discussed in Chapter 5. Random structure search has been effectively utilized in the field of materials research, demonstrating that even random sampling has a decent probability of identifying low-energy basins[214–216]. The advantages of such strategy are the small amount of the parameters that have to be set for the investigation of PES and covering broader volume of PES without biasing of the search itself. Many other methods to some extent depend on routines for producing random structures. More sophisticated techniques that were developed in recent years by introducing bias for the structure search that aims to find global minima faster [217].

For example, the heuristic technique ranks candidates in a search at each branching step depending on the information provided to determine which branch to take next. One of the most famous representatives among the class of heuristic methods are genetic algorithms (GA) [218]. Based on the principle of evolution, new candidates have to exhibit high fitness with respect to some function in order to survive in the natural selection. As a result, they must devise strategies to maintain the diversity of genes within their populations, two of which are well-known: chromosomal crossover during mating and mutation in-place. Individuals with poor fitness will be removed after many generations of natural selection, and the general fitness of the entire population will improve as a result of the selection process. In principle, the "genetic code" is chosen as an array of relevant parameters for the

system at hand. Initially, a random set of candidate structures is generated then a fraction of the population is selected with bias towards fittest, then those structures that were selected are paired up for "recombination" and mutation step may be performed. New candidate structures are added to the pool and the whole process is repeated. GA were applied to investigations of molecular structures, clusters and crystals [219–221].

Another popular family of global geometry optimization techniques include Monte Carlo based approaches, like basin-hopping [222] and minima hopping [223]. The basic strategy in these algorithms is to find one of the local minimum of the system and then with the trial moves escape from the basin and reach another local minimum. Then, with some probability that depends on a specified effective temperature use this new minimum as new starting point. An example of such algorithm that works on internal coordinates plus local translation and rotation of independent geometrical subunits, was demonstrated for molecules adsorbed on surfaces and interfaces [224].

Also, ML algorithms can be used to approximate the PES [225, 226]. In the active learning family, Bayesian optimization search techniques are used to fit a surrogate PES to the data points acquired from DFT calculations, and then improve this potential by acquiring new data points at places where the exploratory lower confidence bound acquisition function is minimized [227–230].

3.2 Geometry optimizations on the Born-Oppenheimer Potential Energy Surface

In the algorithms discussed in the last part of the previous section, the exploration of the conformational space relies on the creation of sample points, followed by local geometry optimizations. Finding local minima requires the computation of the derivatives of the energy with respect to atomic positions (forces) and for more sophisticated and efficient methods estimation of the second derivatives – the Hessian matrix. Here the basic concepts of local structure optimization will be described, introducing the trust region and line search methods, following the textbook by Nocedal and Wright [231]. The starting point to perform local geometry optimization is obtaining the atomic forces, that are defined by $-dE/d\mathbf{R}$. Within density-functional-theory the energy derivative with respect to atom γ is

$$\frac{dE}{dR_\gamma} = \frac{\partial}{\partial R_\gamma}\left[E_v[n(\mathbf{r})] + \frac{1}{2}\sum_{\alpha,\beta}^{N_a}\frac{Z_\alpha Z_\beta}{|R_\alpha - R_\beta|}\right] + \int \frac{\delta E_v[n(\mathbf{r})]}{\delta n(\mathbf{r})}\frac{\partial n(\mathbf{r})}{\partial R_\gamma}d^3r, \tag{3.1}$$

where the implicit R_γ-dependence of the electron density is taken into account in the second term and the only term that explicitly depends on R_γ in $E_v[n(\mathbf{r})$ is the external potential. Computations of the forces that arise by embedding each nucleus into the electrostatic fields of the electron density and all other nuclei, corresponding to the first term of Eq. 3.1, are

performed using the Hellmann–Feynman expression [183]:

$$f_{HF}^{\gamma} = \sum_{\alpha,\alpha\neq\gamma}^{N_A} Z_\gamma Z_\alpha \frac{R_\gamma - R_\alpha}{\left|R_\gamma - R_\alpha\right|^3} - \int n(\mathbf{r}) \frac{Z_\gamma \left(R_\gamma - \mathbf{r}\right)}{\left|R_\gamma - \mathbf{r}\right|^3} d^3 r. \tag{3.2}$$

In codes such as FHI-aims where an atom centered basis set is employed, the basis functions φ_i "move" with R_γ, which leads to additional force contributions (Pulay forces) that arise from the second term in Equation 3.1

$$f_{\text{Pulay}}^{\gamma} = -\int \frac{\partial E_v[n(\mathbf{r})]}{\partial n(\mathbf{r})} \frac{\partial n(\mathbf{r})}{\partial R_\gamma} d^3 r = -2\,\text{Re} \sum_i f_i \left\langle \frac{\partial \varphi_i}{\partial R_\gamma} \middle| -\frac{1}{2}\nabla^2 + v_{\text{KS}} - \epsilon_i \middle| \varphi_i \right\rangle, \tag{3.3}$$

where f_i are the occupation numbers, ϵ_i are the KS eigenvalues. For the details how to take into account additional contributions such as grid effects and multipole correction to the Hartree potential we refer to the original paper of FHI-aims [183].

3.2.1 Local minima finding

Apart from MD, any structure search technique heavily relies on geometry optimization routines that, use energies and forces of the system to find the nearest local minimum of the initial input geometry. The iterative nature of the geometry optimization procedure is denoted with use of a label k, so that for a system with N particles let $x_k \in \mathbb{R}^{3N}$ denote the configuration at the kth optimization step, and the corresponding forces on the system at step k is $f_k = f(x_k)$. The necessary condition for a point on the smooth PES to be a local minimum is the requirement that forces vanish:

$$f(x_k) = 0. \tag{3.4}$$

and that the Hessian matrix $H_k = \partial^2 E / \partial x_k^2$ at this point x_k is positive semidefinite. The standard optimization schemes iteratively search for structures that minimize the energy of a system until Eq. 3.4 is satisfied with desired accuracy usually this threshold is less than 10^{-2}eV Å^{-1}. The simplest methods such as steepest descent and conjugate gradient simply follow calculated gradients and move the atoms in the direction of calculated forces. These are guaranteed to converge, but are among of the most inefficient optimization techniques since they tend to primarily follow the degrees of freedom for which the small displacements lead to large energy changes which results in very poor convergence near the local minimum. The most popular optimization technique is the quasi-Newton scheme that uses the information about the second derivatives of the PES to search for the optimization direction more efficiently [232, 233]. The basic idea is to approximate the PES by an harmonic model with respect to x_k:

$$M_n(x_k + s_k) := x_k - f_k^T s_k + \frac{1}{2} s_k^T H_k s_k, \tag{3.5}$$

where s_k is displacement. The calculation of the exact Hessian requires substantial computational effort, but for the optimization techniques described below this is not necessary

and instead an approximation to the Hessian is used, that is updated during the geometry optimization process. The most widely used scheme for updating the Hessian matrix is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) formula [232, 233],

$$H_{k+1} = H_k - \frac{H_k \Delta x_k \Delta x_k^T H_k}{\Delta x_k^T H_k \Delta x_k} - \frac{\Delta f_k \Delta f_k^T}{\Delta f_k^T \Delta x_k}, \tag{3.6}$$

where $\Delta x_k = x_{k+1} - x_k$ and $\Delta f_k = f_{k+1} - f_k$. In this method the initial guess H_0 is important and can dramatically improve the efficiency of finding of the local minima and, in some cases, even lead to different results when different initial guesses are used [234]. A naive choice of the initial guess is to take the scaled identity matrix $H_0 = \beta \cdot I$ where $\beta > 0$. This scheme is very efficient if the PES is truly harmonic and the Hessian is explicitly known. The first assumption is valid when the structure is already near a local minimum. Usually when dealing with flexible molecules it is impossible to generate the initial guess geometries to be near local minima. Regarding the second term, the first guess for the Hessian matrix must be chosen carefully since it can influence even the qualitative outcome of the optimization [234]. Different preconditioning schemes perform quite differently for different materials systems [235–237] and these will be discussed further in Sec. 5.6.

In recent years, applying a ML model in geometry optimization became a significant field of research, but it is still in the very early stages of adoption. For example, a neural networks (NN) was used to accelerate the saddle-point search by construction of an approximate energy surface [238], and Gaussian Process Regression (GPR) can help to predict derivatives and smoothness of energy function together with their uncertainties during the geometry optimization [239, 240]. The area of active-learning application in geometry optimization looks quite promising [241–245], however, the high flexibility adds a computational cost since a large number of training points (on the order of tens of thousands) are required to ensure that the NN PES has the proper form.

3.2.2 Line search method

Within an optimization algorithm, one has to define a search direction to displace the atoms. One of the approaches for prediction of the search direction for optimization step is the line search method (LSM). Starting from the quadratic model of the PES (Eq. 3.5), one needs to obtain a search direction p_n along which the optimization step s is obtained according to a step length α_n

$$s_n = \alpha_n p_n. \tag{3.7}$$

Then the new configuration is obtained with $x_{k+1} = x_k + s_k$. The search direction p_k for which the energy decreases is the descent direction: $f_k^T p_k > 0$. From the harmonic model the search direction, which is also called quasi Newton step for an approximate Hessian, that minimizes the energy is

$$p_k = H_k^{-1} f_k. \tag{3.8}$$

After finding the search direction one has to determine the step length α_k. Estimation of the step length is done by imposing Wolfe conditions on it [246, 247]:

$$E\left(x_k+\alpha_k p_k\right) \leq E\left(x_k\right)-c_1\alpha_k f_k^T p_k, \quad c_1 \in (0,1), \tag{3.9}$$

$$f\left(x_k+\alpha_k p_k\right)^T p_k \leq c_2 f_k^T p_k, \quad c_2 \in (c_1,1). \tag{3.10}$$

The last one is also called the Armijo condition [248] and assures a sufficient decrease in the objective function along the search direction. The line search method with a BFGS update for the approximate Hessian is summarized in Algorithm 1:

Algorithm 1 BFGS line search

Require: $x_0, H_0, \epsilon > 0$
  $k \leftarrow 0$
  while $\left\|f_k\right\|_\infty > \epsilon$ **do**
    Get p_n from Equation 3.8.
    Get α_n ensuring Wolfe conditions (Eqs. 3.9, 3.10) are satisfied.
    $x_{k+1} = x_k + s_k = x_k + \alpha_k p_k$
    Update approximate Hessian $\Leftarrow$ using Eq. 3.6
    $k \leftarrow k+1$
    end

3.2.3 Trust-region method

Another approach that is widely used is the trust-region method (TRM) that assumes that the harmonic model of the PES is correct within trust-region radius Δ_k near x_k. The trial step is then obtained by minimizing the quadratic model function:

$$s_k = \underset{s_k^* \in T_k}{\arg\min} M_k\left(x_k+s_k^*\right), \tag{3.11}$$

where $T_k := \left\{s_k^* : \left\|s_k^*\right\|_2 \leq \Delta_k\right\}$. Then the quality of the harmonic model is calculated as the ratio between the actual reduction of the total energy E when the trial step s_k is taken and the reduction that is predicted by the model function M_k:

$$\rho_k = \frac{E\left(x_k\right)-E\left(x_k+s_k\right)}{M_k\left(x_k\right)-M_k\left(x_k+s_k\right)} \tag{3.12}$$

If ρ_k is negative, the energy increases with the taken step. For negative and small values of ρ_k, the step is rejected, and the trust-radius is reduced. If ρ_k is close to one, the PES around x_k is in agreement with the harmonic model, and thus the trust-radius can be increased, and the step is accepted. The criteria for adjustment of the trust-radius can be summarized

in the following:

$$\Delta_{k+1} = \begin{cases} \frac{1}{4}\Delta_k & \text{if } \rho_k < \frac{1}{4} \\ \min\{2\Delta_k, \Delta_{\max}\} & \text{if } \rho_k > \frac{3}{4} \wedge \|s_n\|_2 = \Delta_k \\ \Delta_k & \text{else ,} \end{cases} \tag{3.13}$$

where $\Delta_{\max}$ is the maximum allowed displacement length that is defined for the geometry optimization of the system. Iteration continues until the trust-radius is adjusted so that the step is accepted. The iteration then continues until the convergence condition for the forces $\left\|f_k\right\|_\infty < \epsilon$ is met. The TRM method is summarized in the Algorithm 2:

Algorithm 2 BFGS TRM

Require: $x_0, \Delta_0 \in (0, \Delta_{\max}), \epsilon > 0$
$k \leftarrow 0$
while $\left\|f_k\right\|_\infty > \epsilon$ **do**
    Get s_k from Equation 3.11.
    Get ρ_k from Equation 3.12.
    Update trust-region radius $\Leftarrow$ using Eqs. 3.13
    if $\rho_k > \frac{1}{4}$ **then**
        $x_{k+1} = x_k + s_k$
        Update approximate Hessian using Eq. 3.6
    else
        $x_{k+1} = x_k$
    end
    $k \leftarrow k+1$
end

The minimization in Eq. 3.11 can be solved approximately; for further details, we refer the reader to textbook [85].

In conclusion, the LSM and the TRM optimization techniques are both reasonably simple and robust. It is possible to classify both techniques as modified quasi-Newton approaches since they are based on a quadratic model PES and do not require knowledge of the actual Hessian. They are looking for stationary places at which the force disappears, and as a result, they rely on the assumption of a smooth PES. Even though this assumption appears fair for physical systems, it may not necessarily hold in all cases, especially if the system is far away from the local minimum and the electronic structure changes dramatically with respect to structural changes. However, it should be noted that both techniques are only capable of locating local energy minima; the global energy minimum, on the other hand, requires, in addition, sampling of the PES.

For the LSM, the step length is often determined iteratively until the Wolfe criteria are met. For *ab initio* approaches, this can result in an unacceptably large number of energy and force evaluations and may be unstable due to the numerical inaccuracies of the forces. Because

the TRM does not require any extra energy calculations to calculate the trust radius, it is more suitable for *ab initio* structure optimization than the LSM. Thus TRM is the method used in this thesis and default search strategy implemented in the global structure search package discussed in Section 5.

3.2.4 Preconditioning schemes for geometry optimizations

The challenge for quasi-Newton optimization methods is that the Hessian is unknown and has to be approximated for the guess geometry since the calculation of the exact Hessian requires enormous computational effort - it requires $6N$ force evaluations, where N is the number of atoms in the system. Another way to calculate the Hessian matrix is to employ density functional perturbation theory, which is also computationally inefficient [234]. Different ways to construct the initial guess of the Hessian matrix are proposed in the literature. This is referred to as preconditioning, and it may be thought of as a coordinate transformation to a new coordinate system with a better-conditioned optimization problem; as a result, algorithms converge faster and are more robust. Different preconditioning schemes perform with different efficiency for different systems: for example, for covalently bonded periodic systems, the Exponential (Laplacian) preconditioning scheme was found to be simple and effective [237]:

$$H^{Exp}_{(3A+i),(3B+j)} = \begin{cases} -\mu \exp\left(-\alpha\left(\frac{R^{AB}}{R_{nn}} - 1\right)\right), & R^{AB} < R_{\text{cut}} \text{ and } i = j \\ 0, & R^{AB} \geq R_{\text{cut}} \text{ or } i \neq j \end{cases} \tag{3.14}$$

where i, j are Cartesian coordinates and R_{nn} is the maximal nearest-neighbour distance:

$$R_{nn} = \max_{A}(\min_{B} R^{AB}) \tag{3.15}$$

α is chosen arbitrarily to provide damping of atomic interactions, R_{cut} can be reasonably taken as $2R_{nn}$, and the scaling parameter μ can be automatically identified from test displacements of the atoms [249]. By setting $\alpha = 0$ and $\mu = 1$ the Hessian reduces to the Laplacian matrix, a generalization of which is used to represent undirected graphs.

For systems such as molecules in a gas phase or molecular crystals, the Exponential preconditioner scheme does not perform as well as for bulk systems due to the wide range of different interactions. For molecular systems, the use of internal coordinates [250, 251] and FF like preconditioner techniques are much more efficient. For example, the FF model Hessian in Lindh preconditioning scheme is described in the original paper [236] and introduces the analytic form of the energy function that consists of quadratic terms for all distances, angles, and dihedrals in the molecule. The positive-definite requirement for such a preconditioning scheme is fulfilled by assuming that the current geometry is its local minimum. This approach will also be used in the derivation of the Section 5.6, where we derive the preconditioning LJ scheme. Using a simple 15-parameter function of the nuclear positions, the model Hessian can be constructed for any molecule with atoms from the first three rows of the periodic table. This approach yields great performance and is implemented

in many electronic structure packages, including FHI-aims[252]. Other FF based initial Hessian matrices take into account the many-body terms such as bond stretch, angles and dihedrals that are specifically parametrized for a system under investigation and also be used in combination with other preconditioning schemes tailored to systems like molecular crystals [235].

3.3 Comparing molecules across structural space

The large quantities of high dimensional data obtained from structure searches and molecular dynamics simulations require automated tools to produce representations, analyses and classifications. The strategy for representing the high dimensional spaces in a human-readable low-dimensional format usually consists of several steps: a) choosing a representation for the molecules; b) calculating the dissimilarity covariance matrix between these representations; c) performing a dimensionality reduction.

SOAP [111] is an elegant representation that is invariant to rotations, translations, and permutations of equivalent atoms. The main idea of SOAP is to expand the molecular structure into a set of local atomic environments $\mathscr{X}$ and then use their combinations to measure a global similarity between structures. The local environment density around the central atom is approximated as a sum of Gaussian functions with variance σ^2 centred at atom positions $\mathbf{x}_i$ within the environment $\mathscr{X}$:

$$\rho_{\mathscr{X}}(\mathbf{r}) = \sum_{i \in \mathscr{X}} \exp\left(-\frac{(\mathbf{x}_i - \mathbf{r})^2}{2\sigma^2}\right) \tag{3.16}$$

The similarity kernel between two local environments $\mathscr{X}$ and $\mathscr{X}'$ is defined as

$$\tilde{k}\left(\mathscr{X}, \mathscr{X}'\right) = \int \mathrm{d}\hat{R} \left| \int \rho_{\mathscr{X}}(\mathbf{r}) \rho_{\mathscr{X}'}(\hat{R}\mathbf{r}) \mathrm{d}\mathbf{r} \right|^2, \tag{3.17}$$

which is the overlap of the two local atomic environment densities integrated over all three-dimensional rotations $\hat{R}$. The self-similarity of any kernel should be unity, so the final normalized kernel has a form

$$k\left(\mathscr{X}, \mathscr{X}'\right) = \tilde{k}\left(\mathscr{X}, \mathscr{X}'\right) / \sqrt{\tilde{k}(\mathscr{X}, \mathscr{X}) \tilde{k}(\mathscr{X}', \mathscr{X}')}. \tag{3.18}$$

The integration over all rotations can done analytically if the atomic neighbourhood densities are expanded in a basis composed of orthogonal radial basis functions g_n and (angular) spherical harmonics $Y_{l,m}$:

$$\rho_{\mathscr{X}}(\mathbf{r}) = \sum_{n,l,m} c_{n,l,m} g_n(|\mathbf{r}|) \mathrm{Y}_{lm}(\mathbf{r}), \tag{3.19}$$

where $c_{n,l,m}$ are expansion coefficients. From these coefficients, rotationally invariant quan-

tities can be constructed, such as the *power spectrum* that is given by

$$p(\mathscr{X})_{n,n',\ell} = \sum_m c_{n,\ell,m} c^*_{n',\ell,m} \tag{3.20}$$

The elements of the power spectrum are then collected into a unit-length vector $\hat{\mathbf{p}}(\mathscr{X})$, so that the SOAP kernel is given as [111]

$$k\left(\mathscr{X},\mathscr{X}'\right) = \hat{\mathbf{p}}(\mathscr{X}) \cdot \hat{\mathbf{p}}\left(\mathscr{X}'\right). \tag{3.21}$$

The numerical hyper parameters that have to be tuned are the maximal number of radial and angular basis functions, the broadening width, and the cut-off radius. For the details of the derivation of the SOAP kernels for multi-species environments we refer the reader to the detailed explanation in Ref. [115].

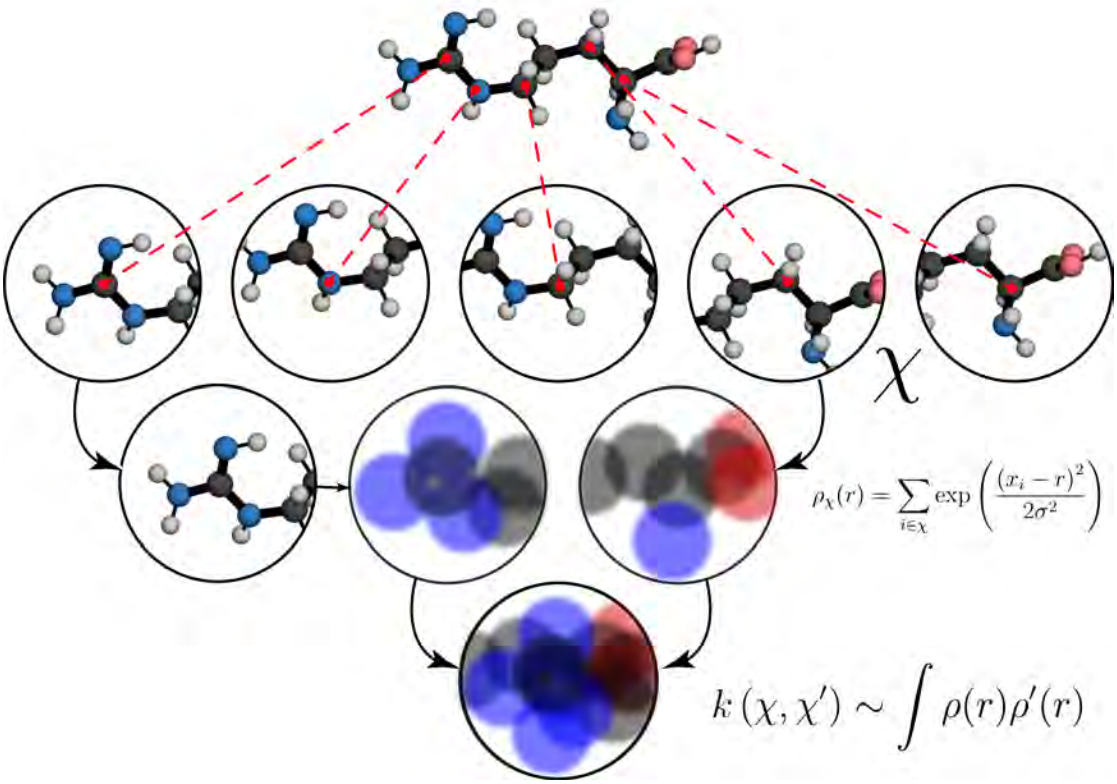


Figure 3.2 – Atom-density-based structural representations, in which the structure is mapped onto a smooth atom density constructed as a superposition of smooth atom-centered functions that also reflect the chemical composition information.

After the mathematical formulation to compare two local environments is established, the next step is to introduce the global kernel to compare two structures. For two structures with the same number of atoms N, one can compute an environment covariance matrix

that contains all the possible pairings of environments

$$C_{ij}(A,B) = k\left(\mathscr{X}_i^A, \mathscr{X}_j^B\right), \tag{3.22}$$

where indices i, j run through all of the atoms contained in structures A and B. The simplest way to introduce a global metric is to use the average kernel

$$\bar{K}(A,B) = \frac{1}{N^2}\sum_{ij} C_{ij}(A,B) = \left[\frac{1}{N}\sum_i \mathbf{p}\left(\mathscr{X}_i^A\right)\right]\cdot\left[\frac{1}{N}\sum_j \mathbf{p}\left(\mathscr{X}_j^B\right)\right] \tag{3.23}$$

The main drawback of this approach is that two very different structures can appear to be very similar if their environments give the same fingerprints upon averaging.

Another possibility is to find the best matching between the environments of the two structures

$$\hat{K}(A,B) = \max_{\mathbf{P}\in\mathscr{U}(N,N)} \sum_{ij} C_{ij}(A,B)P_{ij}, \tag{3.24}$$

by finding the permutation matrix P_{ij} that maximizes the value of $\hat{K}(A,B)$. Here $\mathscr{U}(N,N)$ is the set of $N\times N$ scaled doubly stochastic matrices whose rows and columns sum to $1/N$, i.e. $\sum_i P_{ij} = \sum_j P_{ij} = 1/N$. This is a very computationally expensive procedure that can be computed in polynomial time using the Hungarian Method [253]. This method has discontinuous derivatives whenever the matching of environments change. This problem can be solved by introducing the regularized entropy match kernel (REMatch) that combines the features of *average* and the *best-match kernel* and smoothly interpolates between them. It relies on ideas from optimal transport theory [254] that regularize this problem by adding a penalty that aims to maximize the information entropy for the matrix P_{ij}:

$$\hat{K}^\gamma(A,B) = \mathrm{Tr}\,\mathbf{P}^\gamma\mathbf{C}(A,B) \tag{3.25}$$

$$\mathbf{P}^\gamma = \underset{\mathbf{P}\in\mathscr{U}(N,N)}{\mathrm{argmin}} \sum_{ij} P_{ij}\left(1 - C_{ij}(A,B) + \gamma\ln P_{ij}\right), \tag{3.26}$$

where the entropy term $E(\mathbf{P}) = -\sum_{ij} P_{ij}\ln P_{ij}$ introduces the regularization. This allows the computation P_{ij} with $\mathscr{O}\left(N^2\right)$ effort using the Sinkhorn algorithm [254]. For small values of γ this penalty becomes negligible and we obtain the best-match kernel. For the large values of γ the permutation matrix with the least informational content must be selected $P_{ij} = 1/N^2$, which reduces Eq. 3.25 to the average kernel limit. The definition of the distance would be

$$D(A,B) = \sqrt{2 - 2K(A,B)}, \tag{3.27}$$

where $K(A,B)$ is the global similarity kernel.

After introducing the kernel-induced metric, one can calculate the dissimilarity matrix of a set of structures and employ one of the dimensionality reduction schemes to obtain two dimensional map that represents proximity relations between structures. The simplest method among all the schemes is principal component analysis (PCA) which constructs a linear combination of variables extracting the maximum variance from the input features. PCA and its variances are widely applied in material science for analysing different systems [255–261]. The interested reader can find more details on the dimensionality reduction techniques such as ISOMAP [262, 263], t-SNE [264] applied to analyse biomolecular systems in nice reviews [265–267].

For the dimensionality-reduced representation, we here chose to use the metric multi-dimensional scaling (MDS) algorithm as implemented in the `scikit-learn` package[268]. This algorithm is similar to the Sketch-map algorithm previously employed in Ref. [110], but we found it to be more suitable for the data at hand, which is composed of decorrelated local stationary-points, instead of structures generated from molecular dynamics trajectories. The low-dimensional map is obtained through an iterative minimization of the stress function:

$$\delta = \sum_{A \neq B} (D(A,B) - d(A,B))^2 , \tag{3.28}$$

where $D(A,B)$ is the distance between structure A and B in high-dimensional space and $d(A,B)$ is the Euclidean distance in the low-dimensional space. The result of the procedure will be set of two dimensional coordinates y_N reflecting the mutual distances between structures. For tracking the changes of the conformational spaces one can use one of the two dimensional points as reference and project other structures with use of out-of-sample embedding technique. Finding the low-dimensional coordinates x for structure with high-dimensional representation $\mathscr{X}$ is done through minimization of the stress function δ_P considering the known low-dimensional coordinates for N structures y_N and their high-dimensional representations $\mathscr{X}_N$

$$\delta_P = \sum_{n=1}^{N} (D(\mathscr{X}, \mathscr{X}_N) - d(x, x_N))^2 , \tag{3.29}$$

where the sum runs over all structures in the reference dataset.

I am a dwarf and I'm digging a hole
Diggy, diggy hole! Diggy, diggy hole!
A song of Simon Lane (Honeydew)

4

The conformational space of a flexible amino acid at metallic surfaces

This chapter is dedicated to the description of single molecule adsorption on metallic surfaces. Amino acids are the building blocks of proteins when connected in a sequence via peptide bonds $(\text{N-C}_\alpha\text{-C(O)})_n$, and can be great test systems for methodological developments since they are small enough to be computationally feasible for modern accurate theoretical methods and flexible enough to provide a challenge for their structure search.

In this chapter the adsorption preferences of the most flexible amino acid Arg and its charged counterpart Arg-H^+ were investigated using an exhaustive conformational search. This case is further complicated by the fact that after adsorption the neutral Arg and positively charged Arg-H^+ undergo complex charge rearrangement (see Fig. 4.1). The adsorption was modeled on three noble metal surfaces Cu(111), Ag(111) and Au(111), to study the adsorption behaviour depending of the reactivity of the model surfaces. A depiction of the Arg molecule including the labeling of the different chemical groups and specific atoms we will refer to in the thesis is shown in Fig. 4.2(a). In this context we use the term *protonation state* to distinguish between Arg and its singly-protonated form Arg-H^+. We use the word *protomers* to distinguish between different arrangements of protons within molecules of the same sum formula, for example the protomers **P1** to **P5** of Arg or the protomers **P6** and **P7** of Arg-H^+, shown in Fig. 4.2(b) and (c).

Another important aspect to address is the chemical composition of Arg after adsorption. In general, amino acids tend to adsorb in their zwitterionic form, when the molecule has

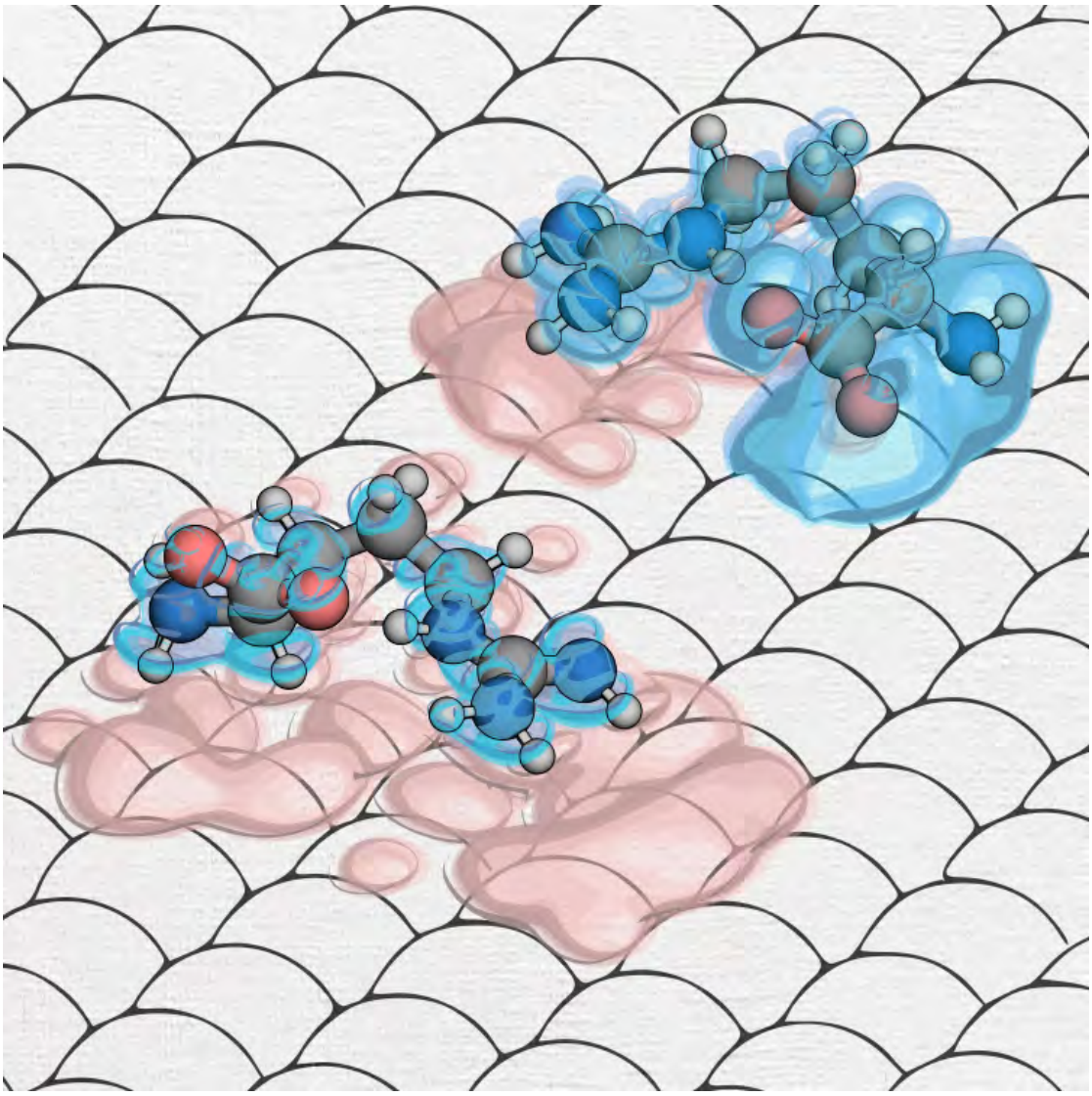


Figure 4.1 – A sketch of the electronic density rearrangement that happens when arginine and protonated arginine adsorb on Cu(111) surface. The electron accumulation is depicted in red and electron depletion is depicted in blue.

termination groups COO^- and NH_3^+ [61]. However, deprotonation is also possible, with the anionic (COO^- and NH_2) and an extra hydrogen atom being adsorbed on the surface [269].

In order to establish the conformational preferences of adsorbed Arg and Arg-H^+, the relative energies of these conformers must be calculated. This can be done using DFT, which can also describe any charge rearrangements that occur following adsorption. In addition, DFT provides insights on the modification of molecular energy levels when forming an interface [73, 211, 270] that are crucial to understanding transport phenomena in molecular electronic devices. Information on the particular preference of adsorption sites and binding

energy strengths that depend on the interacting groups are important in understanding self-assembly patterns that are formed on surfaces [271, 272].

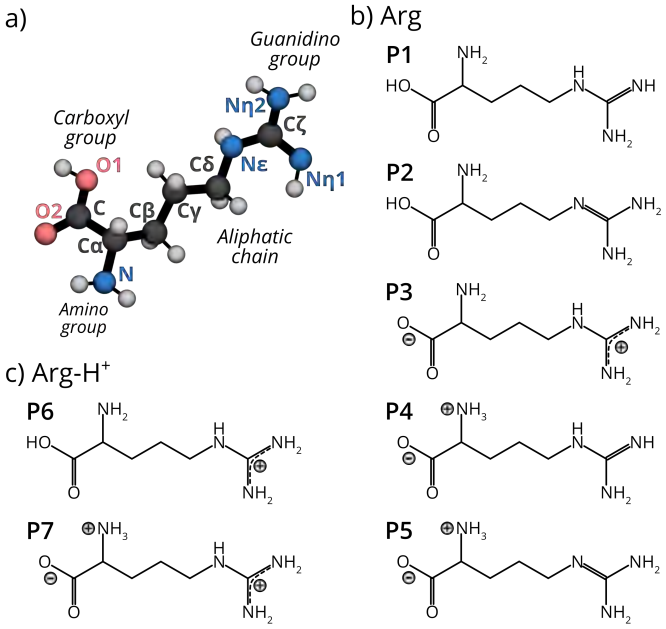


Figure 4.2 – a) Pictorial representation of the arginine amino acid, including labels of chemical groups and atoms. b) Protomers of Arg that are addressed in this work. c) Protomers of Arg-H^+ that are addressed in this work.

The starting point for this investigation was the creation of a database with thousands of stationary states of different conformers on metal surfaces. The procedure of this database generation with a description of the computational setup and convergence tests is described in the next section. A shortened version of this chapter was published in International Journal of Quantum Chemistry [83].

4.1 Computational setup

For modeling the adsorbed molecules, we first had to create model slabs on which to perform an exhaustive structure search. The bulk lattice constants for Cu, Ag and Au were determined by optimizing the fcc unit cell with convergence criteria set to 0.001 eV/Å for the final forces, 10^{-4} e/Bohr^3 for the charge density, and 10^{-5} eV for the total energy of the system, and a 30×30×30 k-grid mesh was used for the sampling of the Brillouin zone. The lattice constants, obtained with the PBE functional[273] are shown in Table 4.1. We also compare the PBE lattice constants with those obtained including pairwise vdW dispersion from the original Tkatchenko-Scheffler scheme (+vdW)[163] and with the one that includes an effective electronic screening optimized for metallic surfaces (+vdW^{surf})[130].

Since the PBE lattice constants for Cu, Ag, and Au are already in good agreement with experimental data [274] (Table 4.1) and with previous works [150, 275], and given the absence

Table 4.1 – Lattice constants (in Å) of bulk metals determined with the PBE, PBE+vdW and PBE+vdW^{surf} functionals (*light* settings).

Method	Cu	Ag	Au
PBE	3.633	4.156	4.157
PBE+vdW	3.545	4.077	4.114
PBE+vdW^{surf}	3.604	4.022	4.173
Exp [274]	3.598	4.079	4.064

of a systematic improvement by the inclusion of these types of vdW interactions [130] in metals, we chose to use the simplest setup and proceed with PBE lattice constants for generating the metal slabs.

For simulations of Arg adsorbed onto surfaces, a 5×6 surface unit cell with $4 \times 4 \times 1$ k-point sampling was employed. The slab contains 4 layers, and we added a 50 Å vacuum in the z direction in order to separate periodic images of the system. Convergence plots in Fig. 4.3 show that this is sufficient to obtain the correct energy hierarchy for different conformers. However, a surface unit cell of this size does not completely isolate neighboring molecules on the surface plane. In order to estimate the magnitude of this spurious interaction, we

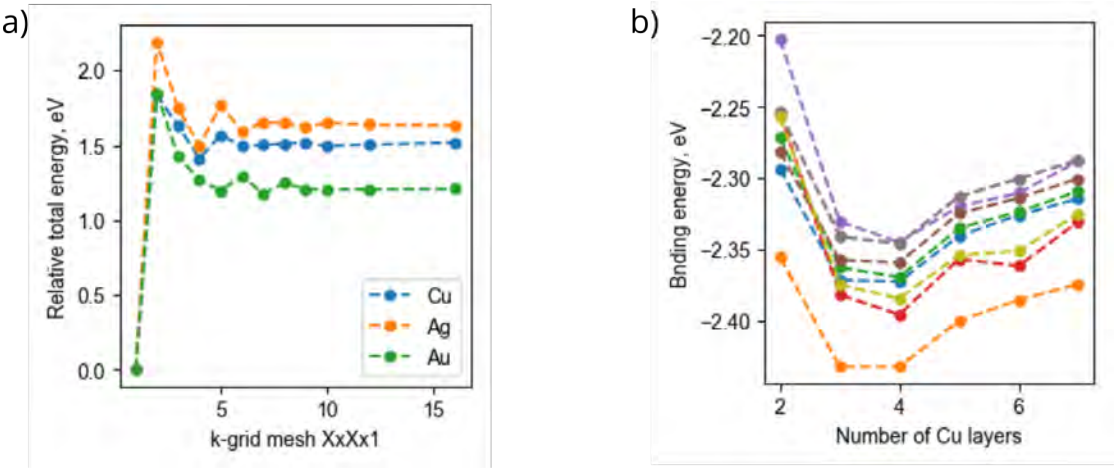


Figure 4.3 – a) Relative total energy convergence of with respect to k-grid mesh for different 5×6 slabs. b) Binding energy hierarchy calculated for different structures on Cu(111) surface with different amount of layers.

calculated binding energies for three Arg and three Arg-H$^+$ structures adsorbed on Cu(111) using different surface unit cell sizes. These structures are shown in Fig. 4.4. As shown in Table 4.2, the relative binding energies change by no more than 50 meV when reaching a 10×12 cell. Furthermore, the energetic hierarchy of the structures does not change with increasing the unit cell size and to save computational resources we proceed with a 5×6 unit cell size.

All the electronic structure calculations were carried out using the numeric atom-centered

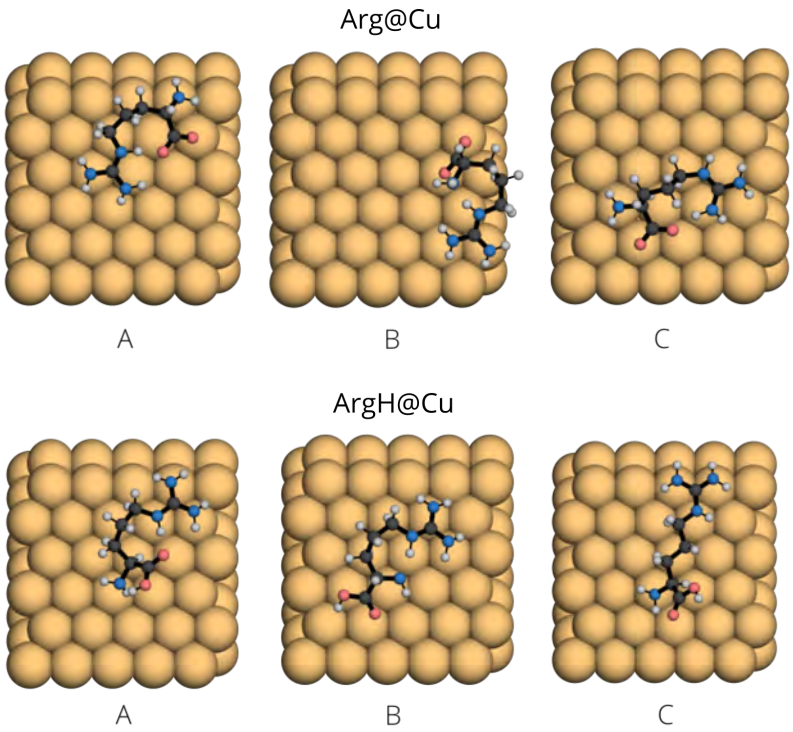


Figure 4.4 – Structures that were used for the surface unit cell size convergence test of Arg@Cu (first row) and ArgH@Cu (second row). Image unit cell size is 5×6.

basis set of the all-electron code FHI-aims [183, 184]. We used the standard *light* settings of FHI-aims for all species with use of PBE+vdW^{surf} functional, except when stated otherwise. Relativistic effects were considered by the zeroth order regular approximation (ZORA) [276, 277]. To prevent an artificial relaxation of the metal surfaces, we did not use vdW interactions between metal atoms since we created slabs with PBE lattice constants. We also fixed the two bottom layers of the slabs in all optimizations. A dipole correction was applied in the z direction to compensate for the dipole formed by the asymmetric surface configurations. With this setup, we placed different conformations of Arg and Arg-H$^+$ in different orientations with respect to the slab and performed a geometry optimization with the BFGS algorithm using the trust region method, until all forces in the system were below 0.01 eV/Å . Database generation is described in the next section.

For reference, we report the values we used for E_f at each surface in Table 4.3.

4.2 Database Generation

The sampling of the structure space of Arginine in two protonation states on metallic surfaces was performed by starting from a previously published dataset comprising the stationary points of isolated amino acids and dipeptides [2, 278]. For Arg, 1206 structures are present in the database. In order to reduce the number of possibilities, but keeping a representative share of the structures, we considered the 300 lowest energy conformers, the 27 highest

Table 4.2 – Relative binding energies (in eV) of relaxed Arg@Cu and ArgH@Cu for different surface unit cell sizes with a 8×8×1 k-grid for the cell sizes less than 10×12 and 4×4×1 for the 10×12 unit cell. All numbers are reported with respect to the binding energy for the structure A modelled with a 5×6 surface unit cell.

	Arg@Cu			ArgH@Cu		
slab size	A	B	C	A	B	C
5×6	0.000	0.011	0.216	0.000	0.080	0.035
6×6	-0.011	-0.013	0.190	-0.050	0.041	-0.017
6×7	-0.021	-0.030	0.174	-0.055	0.029	-0.033
10×12	-0.048	-0.053	0.151	-0.044	-0.007	-0.057

Table 4.3 – Fermi energies calculated with the PBE functional for the 4-layer slabs with (111) surface orientation used in our calculations of the binding energies of charged molecules to the different surfaces. All values in eV.

	Cu	Ag	Au
Slab E_f	-4.73	-4.30	-5.02

energy conformers, and 125 conformers uniformly spanning the energy range in between. For the Arg-H^+ amino acid, all 215 structures present in the gas-phase data set were used in this study.

We distinguish *upstanding* positions of the molecules where the largest eigenvector of the rigid-body moment of the inertia tensor is approximately perpendicular to the surface plane, from *flat lying* positions with an arrangement parallel to the surface. For Arg, 3 flat lying configurations per structure were generated by randomly placing the molecule flat on the Cu(111) surface and then rotating it by 120° around the principal axis. Two upstanding configurations were generated for the 25 of gas-phase structures by first placing the molecule in a random upright orientation, and then flipping it. For Arg-H^+ a similar procedure was adopted: flat lying positions were created by 90° rotations around the principal axis and upstanding configurations were created for 27 structures. In summary, we considered a total of 1156 conformers of Arg@Cu(111) and 914 conformers of Arg-H^+@Cu(111).

Every optimized structure that fell within a range of 0.5 eV from the global minimum on Cu(111) was transferred to Ag(111) and Au(111) and further optimized. In addition, we randomly picked 105 Arg-H^+ structures representing the higher energy range on Cu(111) to be further optimized on Ag(111) and Au(111). Moreover, for Arg 180 randomly picked structures representing the higher energy range were considered on Ag(111) and 61 on Au(111). The total amount of calculated structures for each case is summarized in Table 4.4.

We checked that this strategy ensured a sufficient sampling of the low-energy range of both Arg and Arg-H^+ on Ag(111) and Au(111) by analyzing the alterations in relative energy

Table 4.4 – Number of calculated Arg and Arg-H structures in isolation and adsorbed on Cu(111), Ag(111) and Au(111).

	Gas phase	Cu(111)	Ag(111)	Au(111)
Arg	1206	1156	327	209
Arg-H^+	215	914	718	721

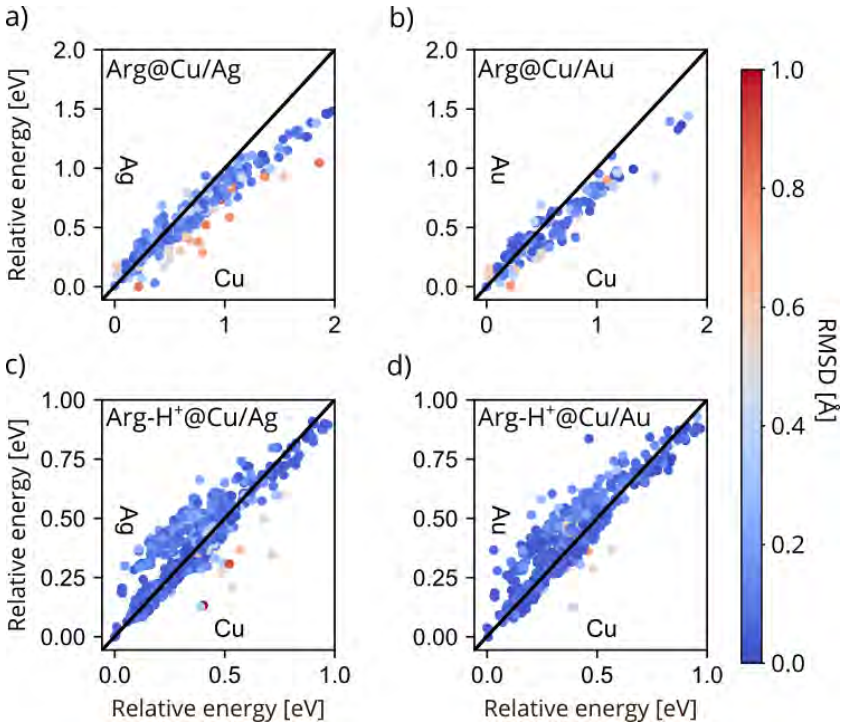


Figure 4.5 – (a-d) Correlation plots of relative energies of Arg or Arg-H^+ conformers on Cu, Ag, and Au (111) surfaces. Each dot corresponds to the same conformer optimized on the two surfaces addressed in each panel, color coded with respect to the RMSD (heavy atoms only) between the superimposed optimized structures without taking surface atoms into consideration.

hierarchies on the different surfaces. In Fig. 4.5, each dot corresponds to a conformer that was optimized first on the Cu(111) surface and then post-relaxed on Ag(111) or Au(111). Within the lowest 0.5 eV range, we do not observe any significant rearrangement of the energy hierarchy with respect to the Cu(111) surface. The energy hierarchies of both Arg and Arg-H^+ on the Ag(111) and Au(111) surfaces are almost identical. The most pronounced outliers in all plots correlate with a higher root mean square displacement (RMSD) of the molecular atoms (i.e. disregarding the surface-adsorption site), thus pointing to a structural rearrangement of the molecule.

4.3 Structure space representation

As was mentioned in the introduction, the simplest and one of the oldest representations developed for the analysis of peptide structures was the Ramachandran plot, which can be seen in Fig. 4.6. As one can see the dihedral angles of the Arg and Arg-H^+ conformers are distributed in 8 clusters, but this information is not enough to draw conclusions about structure-property relationships, since Arg has 4 rotatable dihedral angles. Therefore, we proceed to analyse the database of isolated molecules and introduce further notation for later color coding of the results.

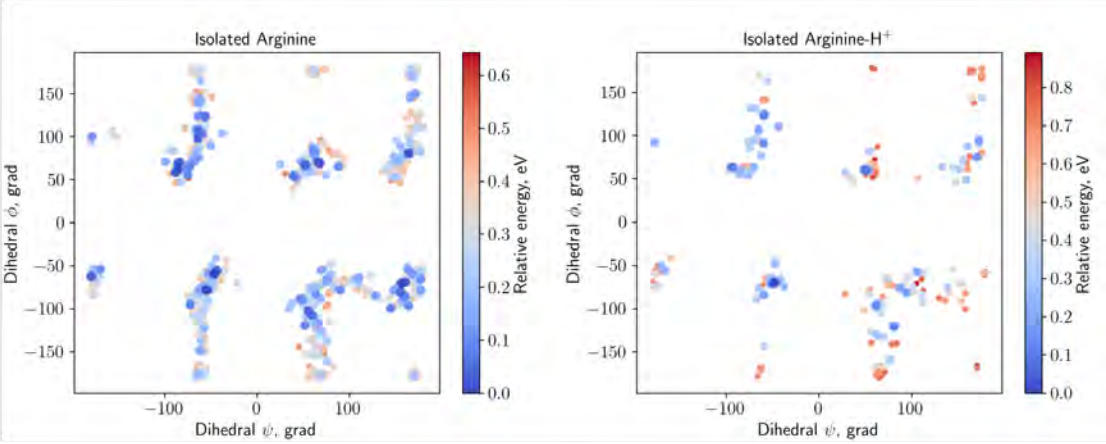


Figure 4.6 – Ramachandran plots for Arg (left) and Arg-H^+ (right) in isolation.

We analyse the structure space of all systems considered by employing a dimensionality reduction procedure that makes it more intuitive to understand the high-dimensional space. Following Ref. [110], we represent the local atom-centered environments of the structures through SOAP[109] descriptors. We then obtain the similarity matrix between different conformers with the REMatch algorithm [115]. We used SOAP descriptors with a cutoff of 5.0 Å, a Gaussian broadening of $\sigma = 0.5$ Å and an intermediate regularization parameter γ=0.01 defined in Sec. 3.3. SOAP kernels were calculated only considering the heavy atoms in the molecule (disregarding metal and hydrogen atoms) and were obtained using the GLOSIM package [115, 279].

For the dimensionality-reduced representation, we here chose to use the metric multi-dimensional scaling (MDS) algorithm as implemented in the `scikit-learn` package[268]. This algorithm is similar to the Sketch-map algorithm previously employed in Ref. [110], but we found it more suitable for the data at hand, which is composed of decorrelated local stationary points, instead of structures generated from molecular dynamics trajectories. In short, the low-dimensional map was obtained considering all calculated structures of Arg in the gas-phase and through iterative minimization of the stress function, according to the procedure described in Section 3.3. We then projected structures in different environments onto the pre-computed map of gas-phase Arg by fixing the parameters of the map and finding the low-dimensional coordinates of the adsorbed molecules. The coordinates obtained as a

result of the iterative metric MDS are not explicitly shown as axes on the plots since they are correlated to the descriptors used for the structural representation, which does not allow for a direct physical interpretation. These scatter plots just offer a visualization of the similarity matrix in lower dimensions. In order to classify structural patterns, we employ the following notations: We represent the protomers by the labels shown in Fig. 4.2(b) and (c). We identify the presence of strong intramolecular hydrogen bonds (H-bonds) whenever the distances between the hydrogen connecting donor and acceptor are below 2.5 Å. We label the H-bond pattern between two atoms in the molecules according to the nomenclature shown in Fig. 4.7. We further classify the structures according to the longest distance between two heavy atoms in the molecule. After describing of the results obtained for isolated Arg and Arg-H^+ molecules we will proceed to the description of adsorbed structures on surfaces.

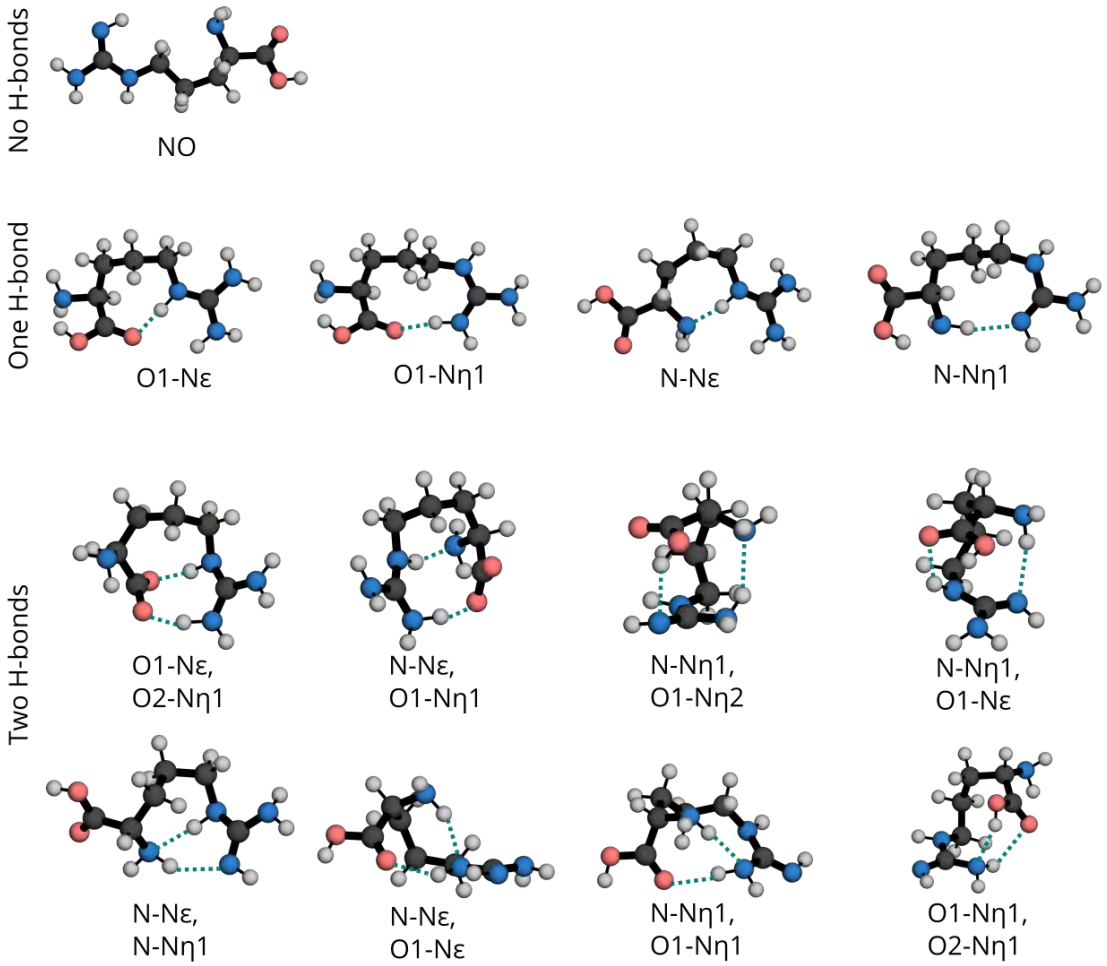


Figure 4.7 – Labeling of all H-bond patterns considered in this thesis.

4.3.1 The unconstrained structure space: Arg in isolation

We start by analysing the unconstrained conformational space of Arg in isolation, which is formed by more than 1200 local stationary states [2, 278]. In order to rationalize the

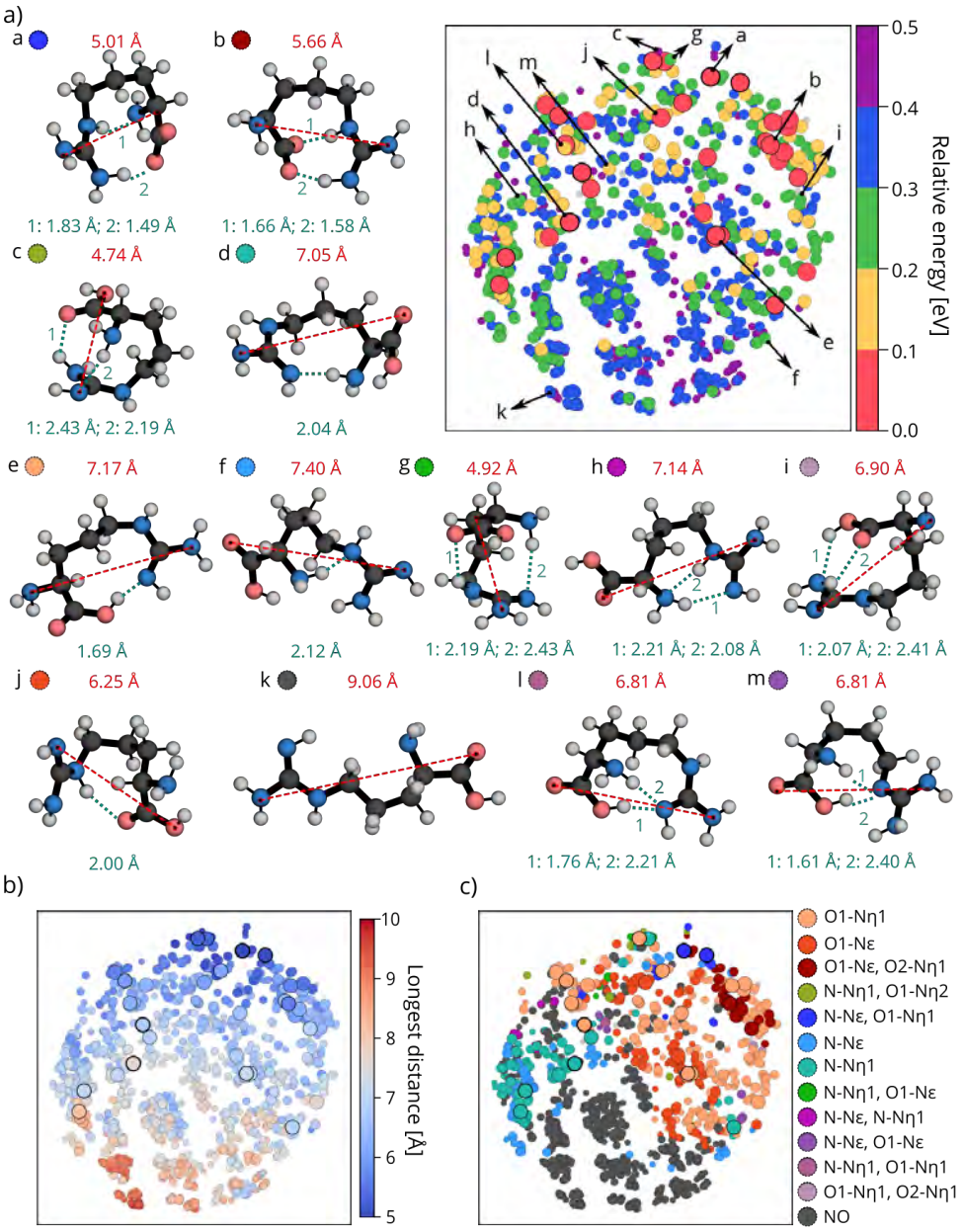


Figure 4.8 – Low-dimensional map of Arg stationary points on the PES. Only points linked to structures with a relative energy of 0.5 eV or lower are colored. Representative structures of all conformer families are visualized as well as their H-bond distances (in turquoise) and longest distance between two heavy atoms (in red) of the molecule. The maps are colored with respect to a) relative energy, b) longest distance, and c) H-bond pattern. The size of the dots also reflect their relative energy, with larger dots corresponding to lower energy structures.

different structural arrangements in this space, we utilize the dimensionality-reduction MDS algorithm and build a two-dimensional map. On this map, shown in Figure 4.8, each dot represents one structure. A close proximity between dots implies similarities between the heavy-atom arrangement between the conformations. This is the low-dimensional map that is taken as a reference for comparison throughout this manuscript.

We proceed to color-code the dots on the map according to different properties. In Fig. 4.8(a) we show the map colored by the relative energy ΔE_{rel} of each structure with respect to the global minimum. We only color structures with $\Delta E_{\text{rel}} < 0.5$ eV. The region with $\Delta E_{\text{rel}} < 0.1$ eV is colored red and is represented by 32 different structures that occupy different parts of the map. The dominant protomer among these conformers (29 out of 32, >90%) is the one labeled **P1** in Fig. 4.2, i.e. non-zwitterionic. However, the lowest energy structure, labeled *a* in panel (a) of Fig. 4.8, is protomer **P3**, with a shared proton between the carboxylic and the guanidino group. This structure is compact, with the longest distance within the molecule of only 5.01 Å and presenting two strong intramolecular H-bonds. Zwitterionic protomers, denoted as **P4** and **P5** in Fig. 4.2, do not appear in the gas-phase.

Inspecting the map in Fig. 4.8(a), it is clear that low-energy conformers are almost exclusively present in the upper hemisphere of the plot. This can be rationalized in terms of the structural motifs that occupy these two halves of conformational space: In Fig. 4.8(b), we color-code the dots in terms of the longest extension of the conformers. While the upper hemisphere features compact structures, the lower hemisphere of the map is populated by extended conformers (with longest extensions between 7.5 Å and 10.0 Å). Many of them do not contain any H-bonds, or contain only one H-bond between the carboxyl and amino group. Extended conformers of Arg are energetically unfavoured in the gas-phase as the formation of strong H-bonds is crucial for the stabilization of Arg in isolation. Comparing the different plots in Figure 4.8, we see that the low-energy structures with $\Delta E_{\text{rel}} < 0.1$ eV are indeed compact with one or two H-bonds.

In Fig. 4.8(c), we identify in total 13 different configurational families with respect to the number and character of H-bonds in the molecule, with $\Delta E_{\text{rel}} < 0.5$ eV. Representative structures of all families are shown in panel (a). This family classification helps us understand why in Fig. 4.8(a) there are structures of higher energies in similar regions as structures with lower energies. Even though these structures are typically in the same protomeric state and have a similar arrangement of heavy atoms, the carboxyl group can rotate, giving rise to different H-bond patterns. These different patterns can give rise to energy differences of up to 0.2 eV, as exemplified in Fig. 4.9. Including hydrogens in the SOAP descriptors used to build the 2D map could provide a better energy separation, but would prevent us from comparing different protonation states, as shown in the next section.

4.3.2 Adding a proton: Arg-H$^+$ in isolation

Arg-H$^+$ is the most abundant form of Arginine under physiological pH conditions [280], and we thus investigate changes of the conformational space introduced by the addition of a

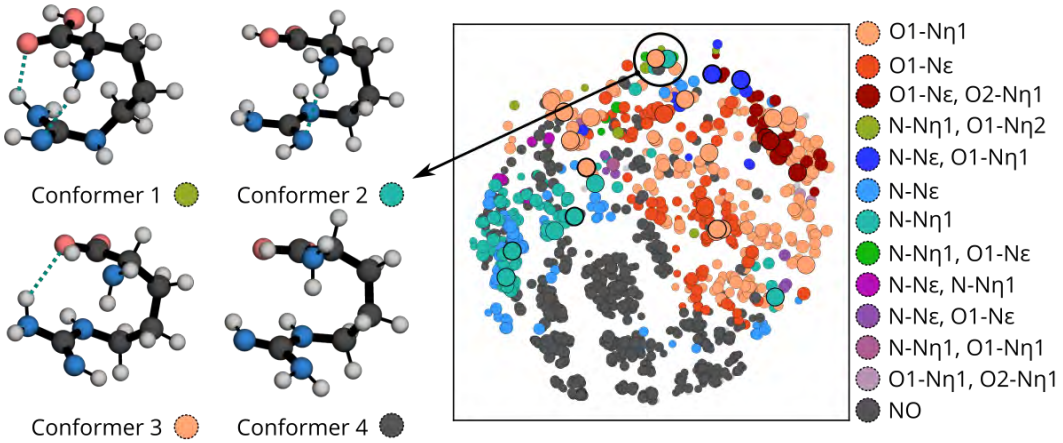


Figure 4.9 – Representative conformers with similar backbone structure but different H-bonds within the molecule. The different H-bond pattern can cause energy differences of up to 0.2 eV for similar structures, as discussed in the main text.

proton to the Arg amino-acid. To that end, we plot a projection of all stationary points of the Arg-H^+ PES with $\Delta E_{\text{rel}} < 0.5$ eV (referenced to its own global minimum) onto the map that was previously created for Arg. In Fig. 4.10(a), we color the dots in the map according to ΔE_{rel}, in Fig. 4.10(b) according to the longest distance between heavy atoms in the molecule, and in Fig. 4.10(c) according to the H-bond pattern. The grey dots in the maps represent all points in the Arg map of Figure 4.8 and are shown for ease of comparison.

The unique conformation types of Arg-H^+ can be grouped into 8 different families in this energy range, which are represented in Fig. 4.10(a). Most families only have one H-bond and there are no zwitterionic protomers. This means that in isolation only the protomer **P6** is populated. It is worth noting that under physiological conditions (in solution), the zwitterionic protomer **P7** is preferred.

There are only two (very similar) structures with $\Delta E_{\text{rel}} < 0.1$ eV in this case. The global minimum, labeled *a* in Fig. 4.10(a), contains two H-bonds within the molecule, between atoms N-Nε and O1-Nη (see Fig. 4.2). This particular structure resembles the lowest-energy structure of Arg with a proton added to the carboxyl group. This protonation results in an extension of the molecule by around 1 Å. That correlates with the location of the lowest-energy structure being slightly shifted on the map towards the region containing more extended structures.

The structure space of Arg-H^+ is contained within the conformational space of Arg and also drastically reduced in number when if compared to Arg: There are only 108 structures with $\Delta E_{\text{rel}} < 0.5$ eV, compared to 1179 structures in the Arg case. In this energy range, regions of the map with very compact and very extended structures are not populated in this protonation state. This can be traced to the constraint imposed by the addition of the proton,

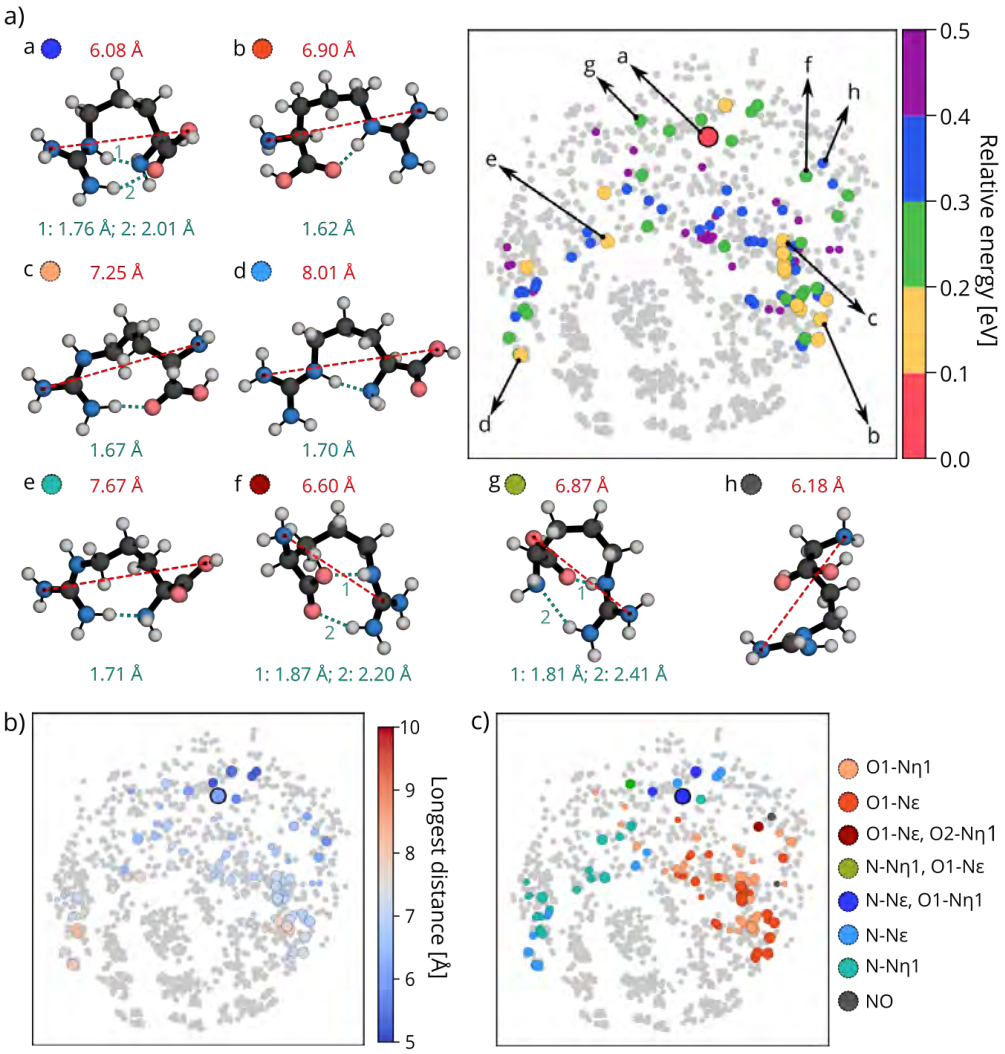


Figure 4.10 – Representative conformers of the populated structure families within 0.5 eV of the global minimum of isolated Arg-H^+ and low-dimensional projections of all populated conformers onto the Arg map. Grey dots represent all structures from the original map of isolated Arg in Fig. 4.10, and serve as a guide to the eye. The maps are colored with respect to a) relative energy, b) longest distance within the molecule, and c) H-bond pattern.

that make extended structures less stable due to the strong driving force to neutralize the charge imbalance created by the proton on the guanidino group. To rationalize why the most compact conformers are also less populated, we show in Fig. 4.11 the electron-density differences between the lowest energy Arg-H^+ conformer and an Arg conformer created by fixing the same Arg-H^+ structure, but neutralizing the charge and removing the hydrogen connected to the carboxyl group. This modification yields the same covalent connectivity observed in the global minimum of Arg. We show isosurfaces corresponding to electron accumulation in Arg-H^+ in red and electron depletion in Arg-H^+ (accumulation in Arg) in blue. We observe a density surplus between the O1 and Nη atoms in Arg, favoring the formation of a stronger H-bond leading to a more compact structure.

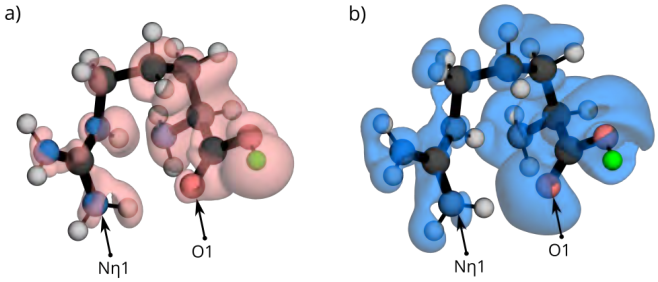


Figure 4.11 – Electron density difference between Arg-H^+ and Arg calculated by neutralizing the charge and removing the hydrogen connected to the carboxyl group (marked in green) from the lowest energy structure of Arg-H^+. The isosurfaces of electron density with value ± 0.005 e/Bohr^3 corresponding to the a) regions of electron accumulation on Arg-H^+ and b) where the electron depletion on Arg-H^+, both compared to Arg.

4.3.3 Adsorption of Arg on Cu, Ag, Au (111) surfaces

We now turn to the analysis of the conformational space of Arg when in contact with metal surfaces, namely Cu(111), Ag(111), and Au(111). In Figure 4.12, we show map-projections of the stationary points with $\Delta E_{\text{rel}} < 0.5$ eV (referenced to the respective global minimum) of Arg adsorbed on the three surfaces. The conformational space of Arg upon adsorption is reduced and the adsorbed conformers occupy similar regions of the map as the conformers of Arg-H^+. We will learn in the following that this is mainly due to the formation of strong bonds with the surface that results in steric constraints of the space, and also partially due to electron donation from the molecule to the metallic surfaces.

The lowest energy structure lies on the same part of the map on all surfaces, which is different from the area where the gas-phase global minimum of Arg was located. These conformers, labeled *a* in Figure 4.12(a), (b) and (c), form a strong H-bond between atoms O1 and Nϵ. The longest distance within the molecule lies between 7.20-7.35 Å in all cases. This structure binds strongly to all three surfaces through both its amino and carboxyl groups.

Other low-energy structures on all surfaces form strong bonds to the surfaces only through the carboxyl group, as exemplified by the structure labeled *b* in all panels of Fig. 4.12. These bonds are formed most favourably on *top* positions, i.e. vertically on top of a surface metal

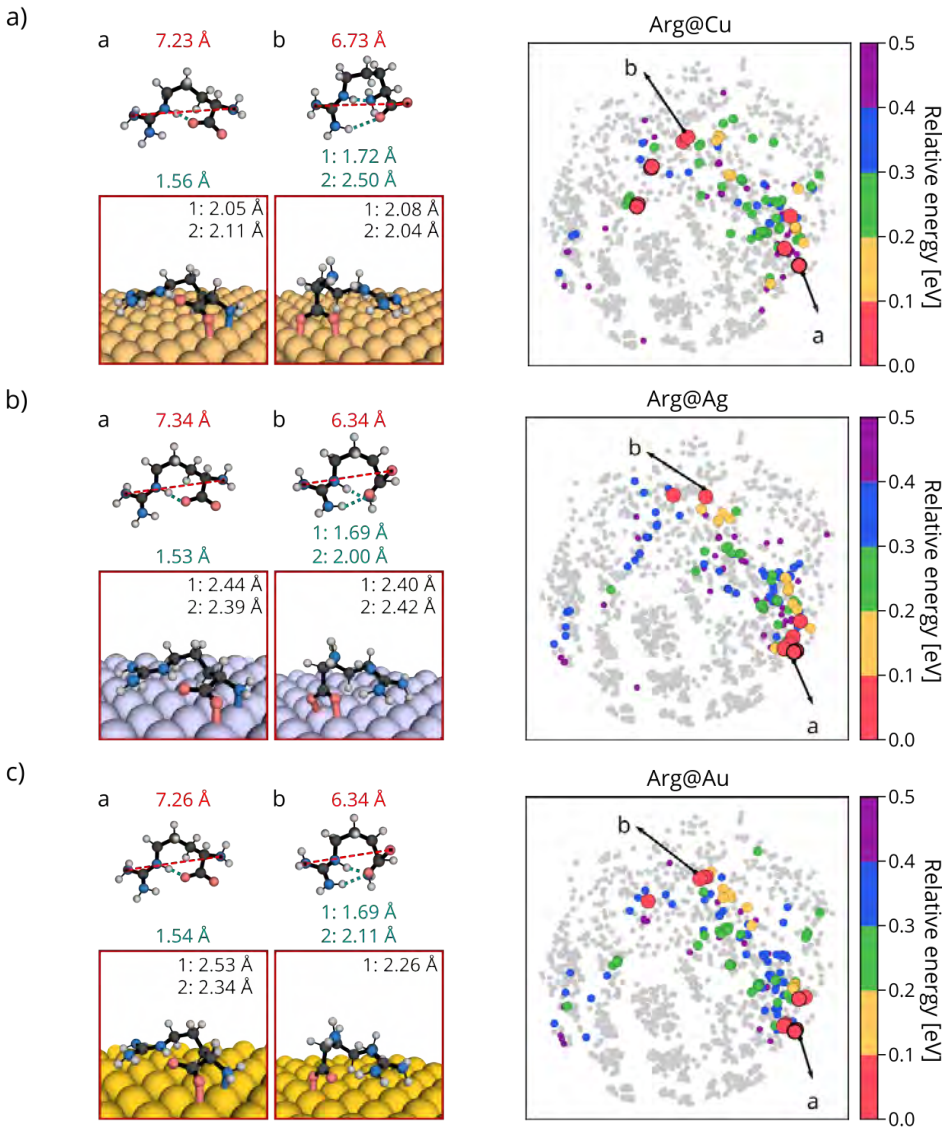


Figure 4.12 – Low-dimensional projections of conformers of Arg adsorbed on a) Cu(111), b) Ag(111), and c) Au(111), onto the gas-phase Arg map of Fig. 4.8. Only conformers within 0.5 eV of their respective global minimum are colored. Grey dots represent all structures from the original map of gas-phase Arg, and serve as a guide to the eye. In each panel, representative structures are shown from two perspectives: a side view where the molecule and surface are shown (bottom), and the corresponding top view (top) where only the molecule is shown. The longest distance within each visualized conformer is reported in red and H-bond lengths are reported in turquoise.

atom. In particular, for Cu(111), the atomic spacing of the Cu atoms on the surface favors both oxygens to bind on *top* positions simultaneously. The favorable formation of these bonds is connected with the fact that all conformers with $\Delta E_{\text{rel}} < 0.2$ eV are in the protomeric state **P3**, in which the carboxyl group is deprotonated. The bonds to the surface and a favorable vdW attraction effectively flatten the molecular conformation, thus energetically favoring more elongated structures. Protomers of type **P1**, which were dominant in the gas-phase, only appear with $\Delta E_{\text{rel}} > 0.3$ eV on Cu and Ag, and with $\Delta E_{\text{rel}} > 0.2$ eV on Au. Zwitterionic protomers **P4** and **P5** are again not observed. Regarding the intramolecular H-bond patterns, within 0.5 eV from the global minimum we can identify 7 different families on Cu(111), and 6 families on both Ag(111) and Au(111). These families contain H-bonds where the carboxyl group predominantly participates. All families are represented in Fig. 4.18.

4.3.4 Adsorption of Arg-H$^+$ on Cu, Ag, Au (111) surfaces

Finally, we characterize the conformational-space changes arising from the simultaneous addition of a proton and the adsorption onto metallic surfaces. In Figure 4.13, we show the projection of the low-dimensional representations of Arg-H$^+$ conformers adsorbed on Cu, Ag, and Au(111) onto the map of isolated Arg conformers. These projections, in particular the comparison of the plots in Figs. 4.12 and 4.13, reveal that the conformational space of adsorbed Arg-H$^+$ is larger than the one of adsorbed Arg. While Arg-H$^+$ features more than 500 conformers within $\Delta E_{\text{rel}} < 0.5$ eV, Arg only counts about 150 conformers in the same energy range. Interestingly, the adsorption of Arg-H$^+$ to a metal surface also results in an increase of the occupied structure space in comparison to isolated Arg-H$^+$ (108 structures with $\Delta E_{\text{rel}} < 0.5$ eV), shown in Fig. 4.10. In fact, the structures occupy similar regions of the map as the ones occupied by Arg-H$^+$, with the addition of extended structures that are located at the bottom of the map.

We identify 4 different families on Cu(111) and 3 on Ag(111) and Au(111) with $\Delta E_{\text{rel}} < 0.1$ eV. Representative conformers of these families are shown in Fig. 4.13. The lowest energy conformer, labeled *a* in Fig. 4.13(a)-(c), appears on all surfaces at the same region of the map as for adsorbed Arg. The largest distance within the molecule lies around 7 Å and it also has a strong H-bond linking the carboxyl-O and the Nε atoms. The structure, however, does not present the same orientation to the surface as compared to the lowest energy conformer of Arg, and does not form strong bonds with the surface. With the exception of the extended structure on Cu(111), labeled *d* in Fig. 4.13(a), all conformers with $\Delta E_{\text{rel}} < 0.1$ eV on all surfaces contain one intramolecular H-bond involving either carboxyl-O and Nε atom (labeled a), backbone N and Nε atoms (labeled b) or carboxyl O and a Nη atom (labeled c). Compared to adsorbed Arg, adsorbed Arg-H$^+$ structures become on average 1.0 Å more extended as shown in Fig. 4.14. The protomer **P6**, the only one present in the gas-phase, is dominantly populated also on the surfaces. However, we do observe a few conformers in the zwitterionic **P7** state. These structures are at least 0.2 eV higher in energy than the global minimum.

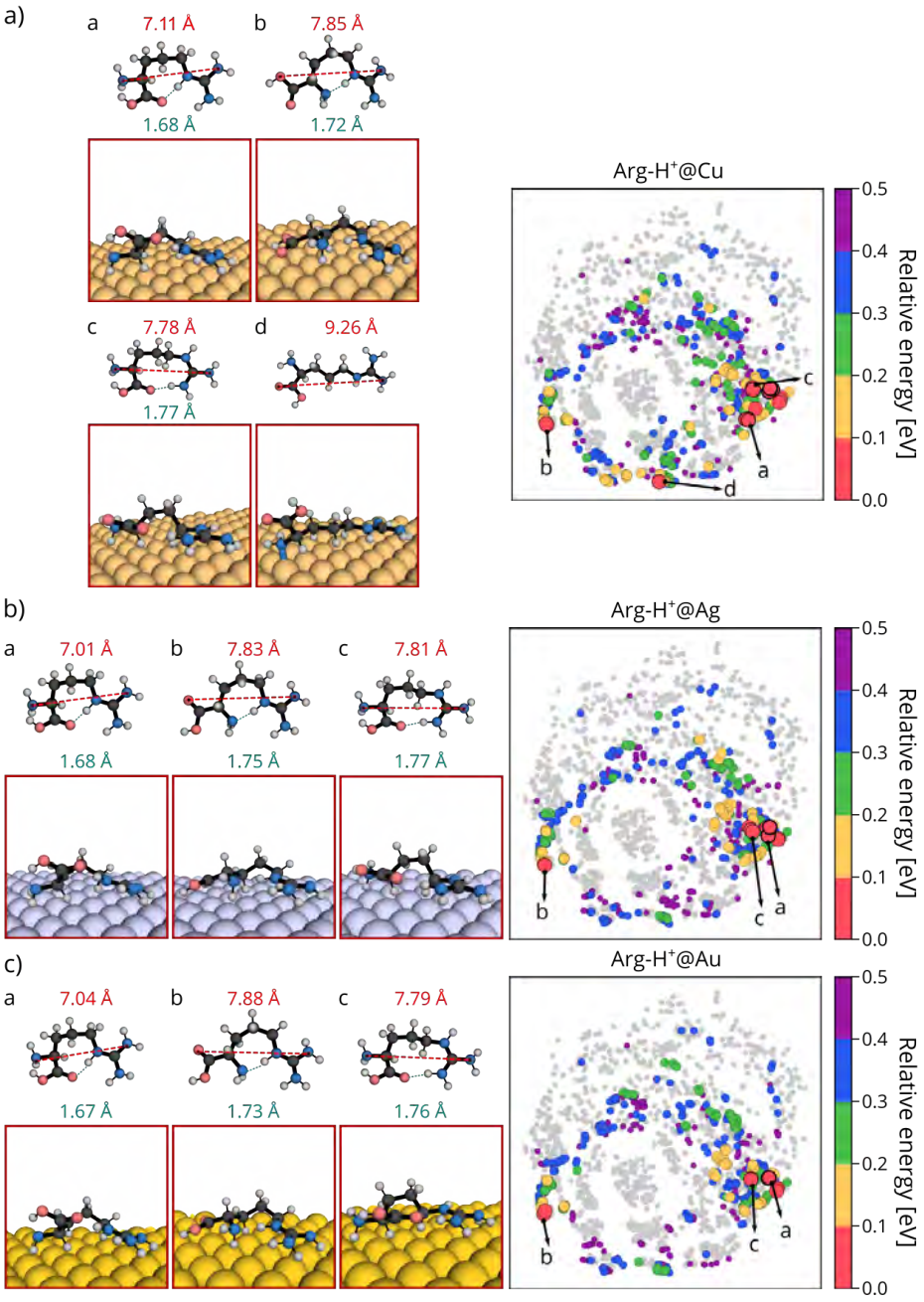


Figure 4.13 – Low-dimensional projections of conformers of Arg-H^+ adsorbed on a) Cu(111), b) Ag(111), and c) Au(111), plotted on the gas-phase Arg map of Fig. 4.8. Only conformers within 0.5 eV of their respective global minimum are colored. Grey dots represent all structures from the original map of gas-phase Arg, and serve as a guide to the eye. In each panel, representative structures are shown from two perspectives: a side view where molecule and surface are shown (bottom), and the corresponding top view (top) where only the molecule is shown. The longest distance within each visualized conformer is reported in red and H-bond lengths are reported in turquoise.

With respect to the number of bonds that Arg-H^+ forms with the surface, the picture is very different from adsorbed Arg. Within the lower 0.15 eV, we do not observe short (strong) bonds of O or N atoms to the surfaces. This lack of constraint by the surface contributes to the increased structure space of adsorbed Arg-H^+ in comparison to Arg. In addition, the molecule accepts electrons from the surface, becoming less positively charged, as we discuss in detail in the next section. We conclude that Arg-H^+ interacts with the metallic surfaces mostly through van der Waals and electrostatic interactions.

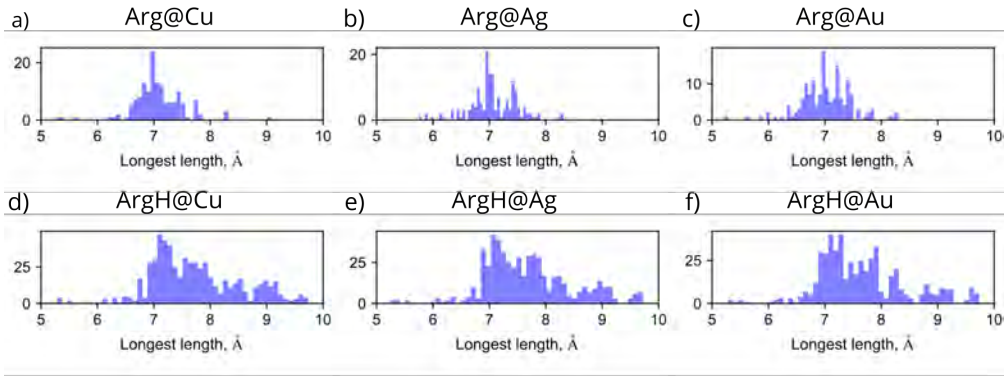


Figure 4.14 – Histogram of the longest distances of adsorbed molecules on different surfaces.

4.4 Electronic structure and trends across surfaces

In the previous section we focused on the structural aspects of the adsorbed molecules and the most prominent bonds the molecules make with the metallic surfaces. In the following, we will discuss different aspects of the molecule-surface interactions, with the goal of identifying trends across these systems.

We begin by analysing the binding energies between the molecules and surface, which are shown in Fig. 4.15. The binding energies for all surfaces were calculated as discussed in Section 2.8. The larger negative values in Fig. 4.15 correspond to the stronger binding of the molecule to the surface. In the case of adsorbed Arg, many conformers bind to Cu more strongly than to Ag and Au, with the binding of the deprotonated carboxyl group of Arg to the Cu(111) surface geometrically favored as discussed above. In the case of adsorbed Arg-H^+, there is no pronounced difference in binding strengths to the different surfaces, and the values are comparable to the binding energies obtained for Arg adsorbed on Cu(111). This correlates with the observation that the interaction of Arg-H^+ with the surfaces happens mostly through dispersion and electrostatic interactions. Despite the strong binding to the surface, it is also visible from comparing Figs. 4.12 and 4.13 that the interaction of Arg-H^+ with the surface does not strongly template the conformations of this molecule, implying a low corrugation (i.e. homogeneity) of the molecule-surface interaction and allowing for a larger variety of conformers with similar energies. This is in contrast to the molecule-surface interaction of Arg, which is more inhomogeneous due to the formation of bonds through specific chemical groups. In realistic applications, the thermal energy will result

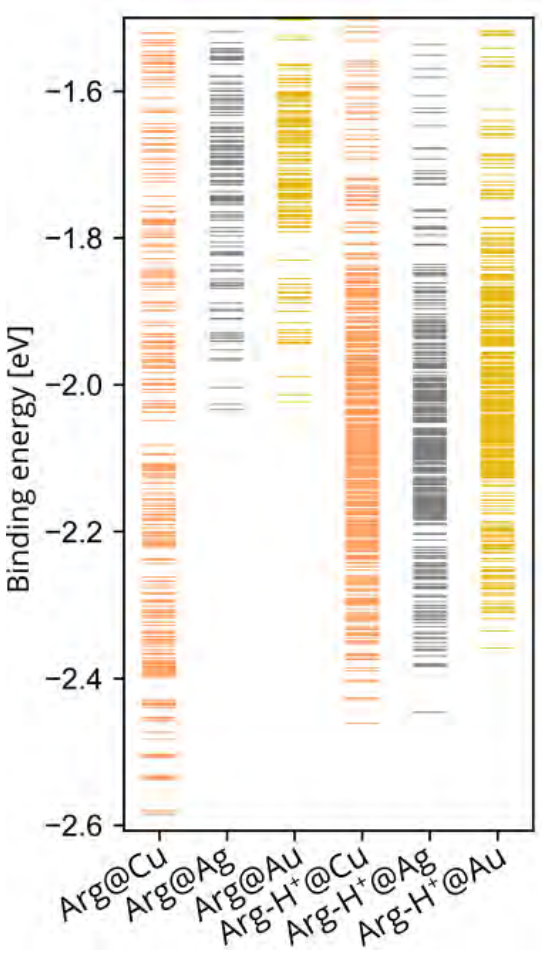


Figure 4.15 – Binding energies of Arg and Arg-H^+ on Cu(111), Ag(111) and Au(111) surfaces.

in vibrational contributions to the stability of a conformer, potentially changing the energy hierarchy. In order to address the question about thermal stability of adsorbed structure, the free energies at finite temperatures within the harmonic approximation [281, 282] can be calculated:

$$F_{\text{harm}}(T) = E_{\text{PES}} + F_{\text{vib}}(\text{T}), \tag{4.1}$$

where E_{PES} is the total energy obtained from DFT (PBE+vdW$^{\text{surf}}$ functional), and we have used textbook expressions for the harmonic vibrational Helmholtz free energy $F_{\text{vib}}(T)$:

$$F_{\text{vib}}(\text{T}) = \sum_{i}^{3N-6} \left[\frac{\hbar\omega_i}{2} + k_{\text{B}} T \ln\left(1 - e^{-\beta\hbar\omega_i}\right) \right],$$

where N is the total number of atoms in the molecule (metal atoms were not displaced and were taken into account in the external field), k_{B} is Boltzmann constant, T is the temperature, ω_i are vibrational frequencies obtained by diagonalization of Hessian matrix with use of developing version of phonopy-FHI-aims [283, 284]. For the adsorbed conformers,

rotational contributions are completely neglected since rotation around all principal axes of the molecule become internal vibrational modes of the system.

We have estimated harmonic vibrational free energies for representative conformers with $\Delta E_{\text{rel}} < 0.1$ eV in each surface. In contrast to what has been reported for longer helical peptides [285, 286], the global minimum remains the same in all cases, as reported in Fig. 4.16. For Arg-H^+ we observe relative energy rearrangements of up to 50 meV at 300 K, which changes the relative energy hierarchy of conformers less stable than the global minimum. Therefore, vibrational effects must be considered in order to obtain an accurate energy hierarchy at a given temperature.

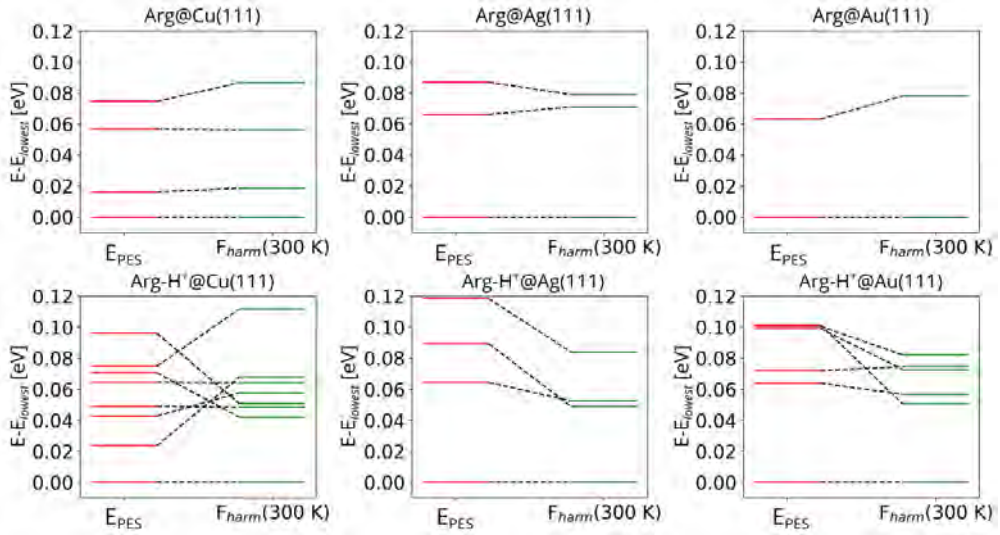


Figure 4.16 – Harmonic free energies calculated for adsorbed structures within the lowest 0.1 eV total-energy range. $E_{\mathbf{PES}}$ corresponds to the total energy of the system obtained at DFT level and $F_{\mathbf{harm}}$ corresponds to the free energy of the system at 300 K calculated as described above.

We then focus on the distance between the molecule and the surfaces. We define this quantity by measuring the distance of the center of mass (COM) of the molecule with respect to the surface plane defined by the top layer of surface atoms. These distances are collected in Fig. 4.17. The COM is closer to Cu(111) than to Ag(111) and Au(111) for both Arg and Arg-H^+, because of higher reactivity of Cu. In addition, in all surfaces, Arg lies closer than Arg-H^+, in agreement with the observation that Arg forms covalent bonds to the surface. The extended structures of Arg-H^+, at the bottom of the maps, tend to be closer to the surface than those that have H-bonds within the molecule, likely due to the stronger vdW attraction to the surface by extended conformations.

The difference in COM distances to the surfaces between Arg and Arg-H^+ is apparently related to the preferred orientations of the chiral center of the molecule to the surface. The chiral C_α carbon can point its bonded hydrogen towards the surface (labeled *down* in the following), or towards the vacuum region (labeled *up* in the following). Examples of these

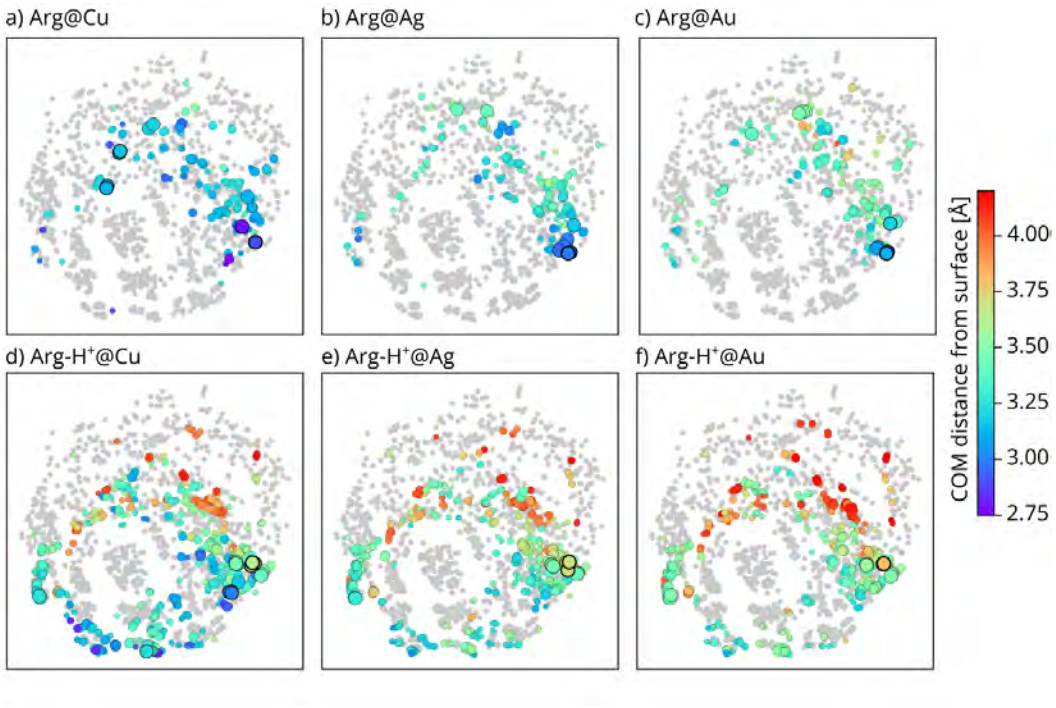


Figure 4.17 – Low dimensional projections of adsorbed Arg and Arg-H^+ on Cu(111), Ag(111) and Au(111) color-coded with respect to the distance of the center of mass of the molecule with respect to the surface. Grey dots represent all structures from the original map of isolated Arg where the projection was made, and serve as a guide to the eye.

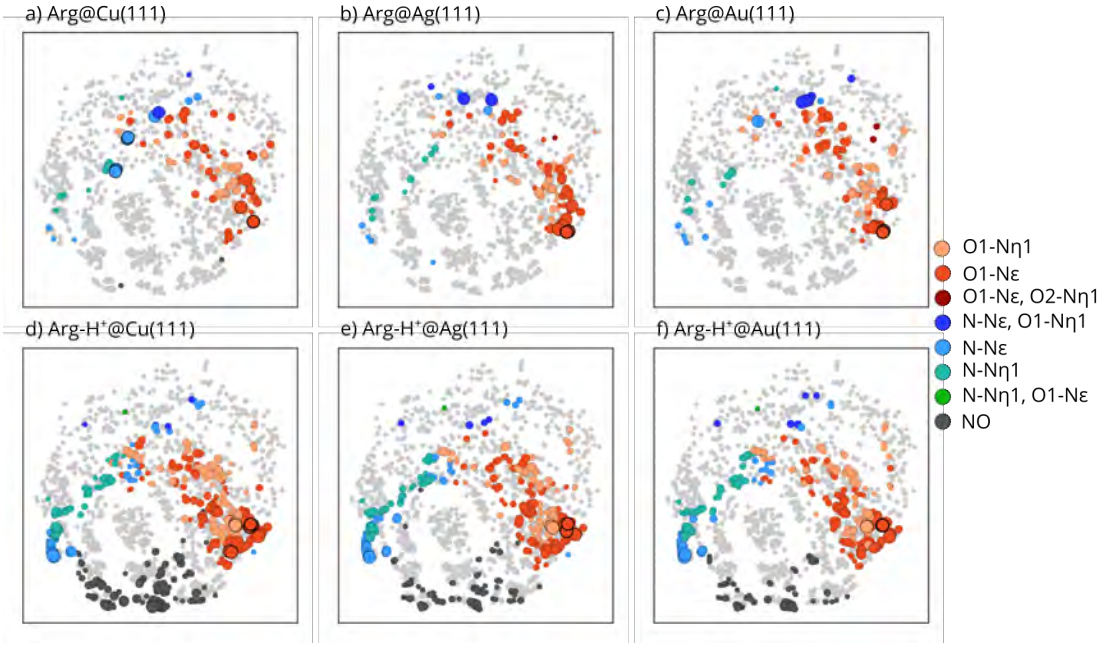
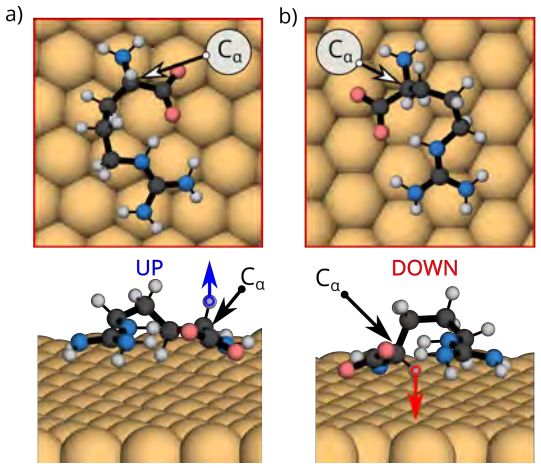


Figure 4.18 – Projection of Arg and Arg-H^+ conformers adsorbed on the different metalic surfaces on the low-dimensional map of gas-phase Arg, colored according to the H-bond pattern.

different molecular orientation are shown in Fig. 4.19(a).



	Arg		Arg-H^+	
	up	*down*	*up*	*down*
Cu	12/100	0/48	8/255	13/355
Ag	10/110	0/50	3/255	10/292
Au	13/104	1/53	2/219	12/284

Figure 4.19 – Orientation of the $C_\alpha H$ group in a) *up* orientation (hydrogen pointing towards vacuum) and b) *down* orientation (hydrogen pointing towards the surfaces). c) The amount of structures with *up* and *down* orientation within 0.1/0.5 eV from the global minimum of each surface.

The dominant orientation with respect to the surface is different in the cases of Arg and Arg-H^+, as evidenced by the numbers presented in Fig. 4.19(b). The lower energy structures are

mostly in the *up* orientation for Arg and mostly in the *down* orientation for Arg-H^+ (see also map in Fig. 4.20), consistent with the typically smaller distance to the surface for adsorbed Arg. However, despite the different orientations of their $C_\alpha H$ groups, the lowest energy

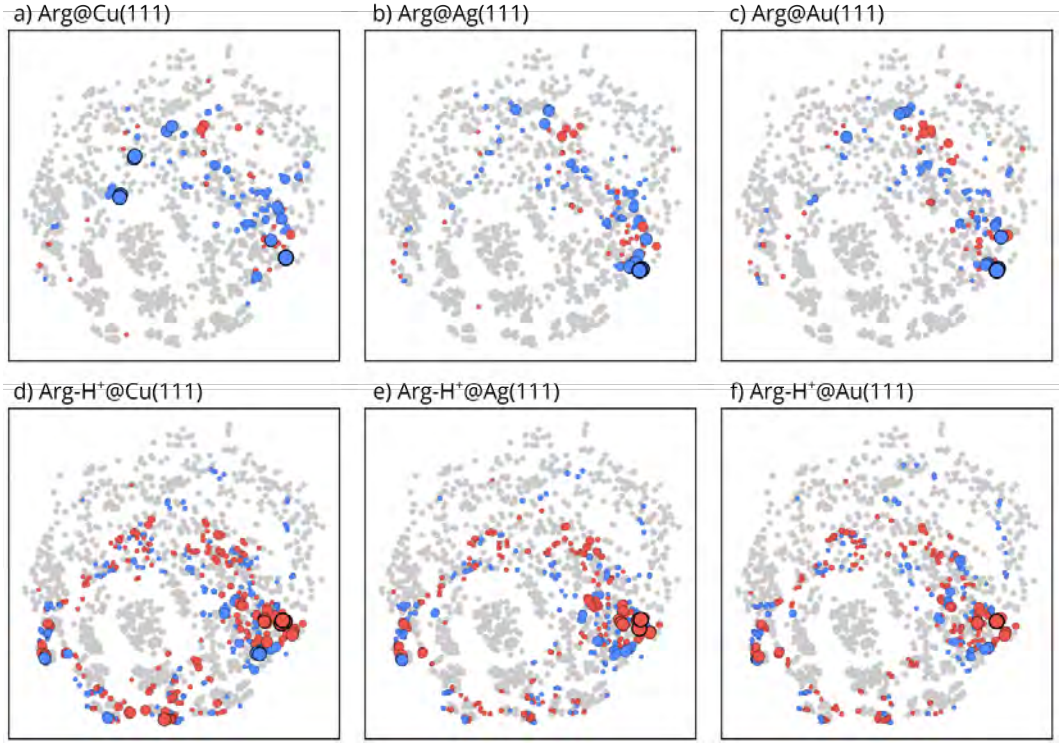


Figure 4.20 – Low dimensional maps of Arg and Arg-H^+ adsorbed on Cu(111), Ag(111) and Au(111) color-coded with respect to the orientation of the $C_\alpha H$ group. Blue correspond to *up* orientation and red correspond to *down* orientation of the $C_\alpha H$ group.

structures for both molecules adsorbed on each surface have very similar conformations. Since the addition or removal of a proton can apparently alter the preference of the chiral-center orientation, we propose that it could template different chiralities of self-assembled super-structures on the surface [27].

We then investigated the rearrangement of the electronic density upon binding of the molecules to the different surfaces. In Fig. 4.21 we show the electronic density rearrangement created by the lowest energy conformer at each surface, integrated over the axis parallel to the surface, overlaid on the side-view of the 3D density rearrangement. In addition, we show a top view of the density rearrangement in each case. Examples of further conformers are summarized in the Appendix. The data shows that Arg donates electrons to the surface, while Arg-H^+ accepts electrons from the surface. We have checked this propensity for selected conformers by integration of the electronic density rearrangement around the molecule and by calculating the Hirshfeld charge remaining on the molecule for the full database (see Table 4.5). When comparing Hirshfeld charges on the molecule and those obtained from the electronic density rearrangement, we observe that Hirshfeld charges are always 0.3-0.5 *e*

underestimated, making them an unreliable method to analyse charge trasfer.

In addition, we observe that the depletion and accumulation of charge is not uniform through the lateral extension of the molecule. This behavior is consistent with the level alignment predicted by the PBE Kohn-Sham energy levels, as shown in Fig. 4.22. However, we note that quantitative values of charge transfer are often inaccurate at this level of theory, as characterized in Refs. [287, 288]. Optimally tuned range-separated hybrid functionals would yield more accurate values, but their computational cost is prohibitive for use in this whole database. Nevertheless, hybrid-functional calculations (PBE0) of selected conformers (Fig.

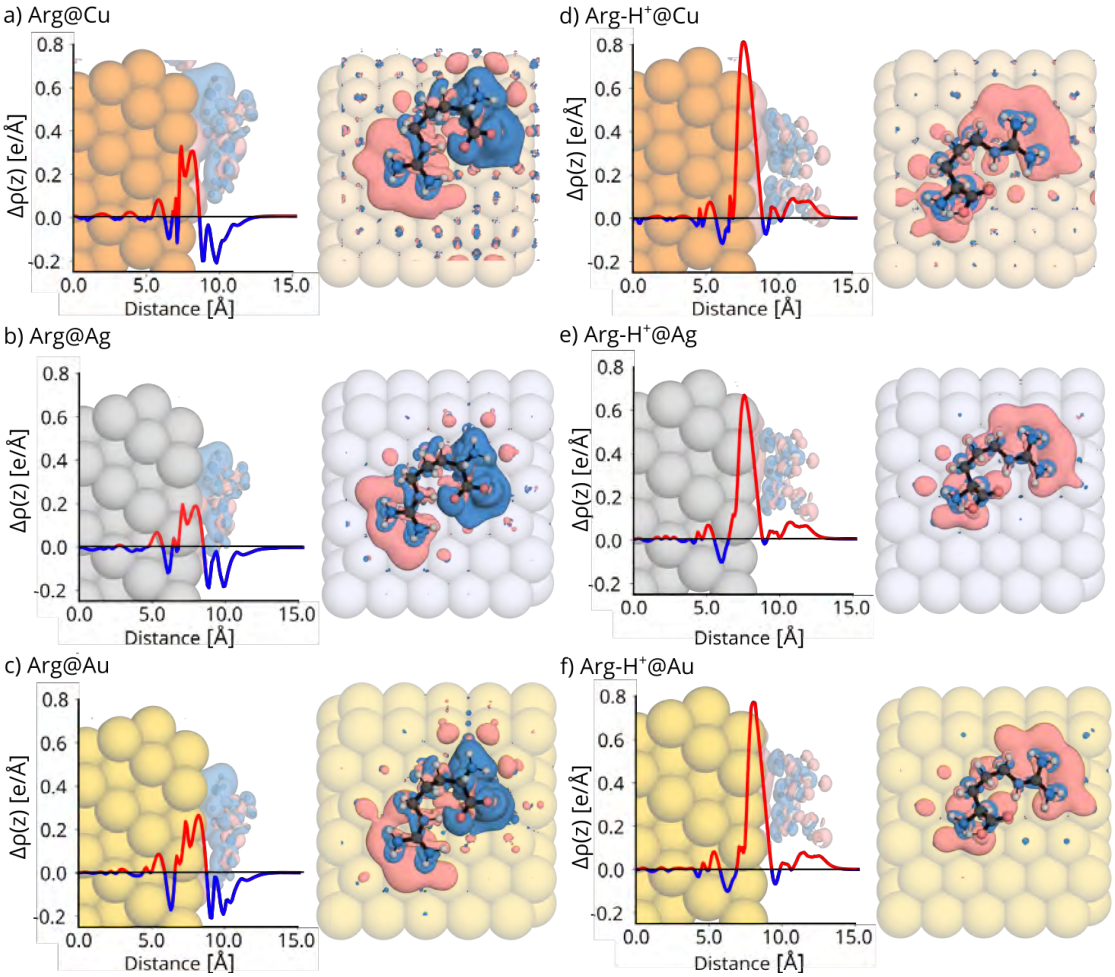


Figure 4.21 – Electronic-density difference averaged over the directions parallel to the surface for the lowest energy conformers of Arg adsorbed on Cu(111) (a), Ag(111) (b), and Au(111) (c), as well as of Arg-H^+ adsorbed on Cu(111) (d), Ag(111) (e), and Au(111) (f). Positive values (red) correspond to electron density accumulation and negative values (blue) correspond to electron density depletion. In each panel, we also show a side and top view of the 3D electronic density rearrangement. Blue isosurfaces correspond to an electron density of +0.05 e/Bohr^3 and red isosurfaces to -0.05 e/Bohr^3.

Table 4.5 – Calculated charge on the molecule with use of Hirshfeld partial charge analysis and by integration of the electron density difference in the molecular region. Values are in electrons.

Conformer	Hirshfeld	Integral	Conformer	Hirshfeld	Integral
Arg@Cu			Arg-H^+@Cu		
a	0.11	0.19	a	0.29	0.85
b	0.03	0.30	b	0.30	0.85
c	0.04	0.31	c	0.31	0.84
d	0.08	0.26	d	0.43	0.88
e	0.01	0.24	e	0.46	0.85
f	0.11	0.30	f	0.38	0.82
Arg@Ag			Arg-H^+@Ag		
a	0.04	0.15	a	0.28	0.83
b	-0.08	0.23	b	0.30	0.83
c	-0.03	0.24	c	0.31	0.82
d	-0.06	0.21	d	0.43	0.86
e	-0.13	0.16	e	0.46	0.85
f	0.05	0.14	f	0.36	0.86
Arg@Au			Arg-H^+@Au		
a	0.06	0.05	a	0.32	0.86
b	-0.01	0.29	b	0.29	0.86
c	0.00	0.30	c	0.34	0.85
d	-0.10	0.25	d	0.48	0.91
e	0.01	0.23	e	0.49	0.90
f	0.06	0.31	f	0.43	0.92

4.23) confirm the qualitative trend. Therefore, we conclude that the protonation state again critically impacts these systems, in this case by qualitatively changing the redistribution of electronic charge.

It was observed experimentally that amino acids can undergo deprotonation on reactive surfaces [289–294]. Here we also investigated whether deprotonation of Arg and Arg-H^+ was favorable on any of the surfaces studied here. In Arg, we found it most favorable to detach the proton from the guanidino group, while for Arg-H^+, it was most favorable to detach the proton from the carboxyl group. We chose three representative conformers at each surface: the lowest energy structure and two others with different H-bonds within the molecule. We placed the detached proton at a distance of at least 2.5 Å from the molecule and fully optimized the dissociated structures. Comparing the energy difference between the final and initial states gives a lower limit for the dissociation barrier:

$$\Delta E = E_{\text{dissociated}} - E_{\text{lowest}}. \tag{4.2}$$

The results are summarized in Figs. 4.24 and 4.25. They show that, however, only the de-

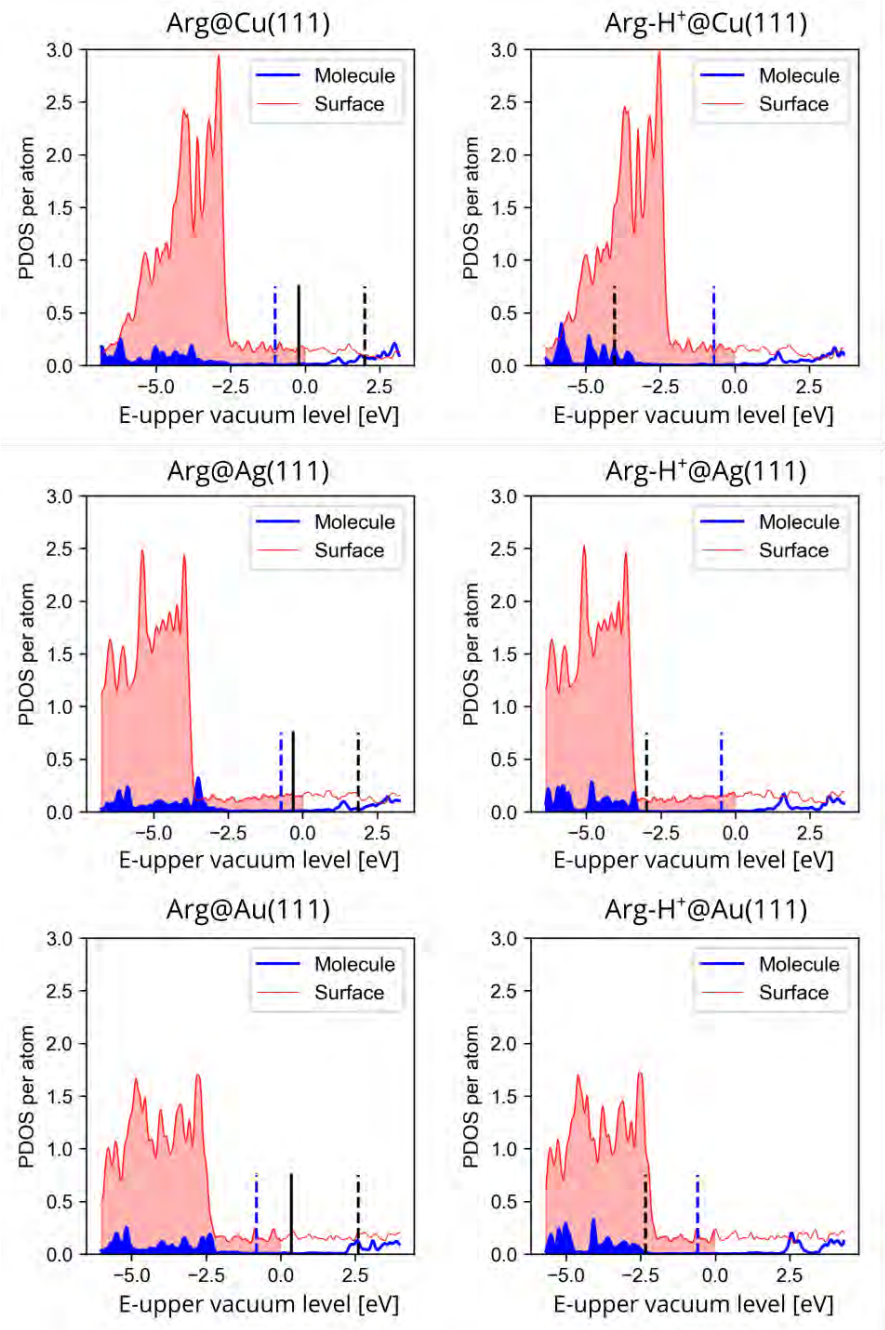


Figure 4.22 – Projected densities of states of the lowest energy structures on each surface. The filled area corresponds to the occupied states below the highest occupied state (VBM) of the whole system. HOMO (black solid line) and LUMO (black dashed line) are the states of the corresponding gas-phase molecular conformer calculated with the same geometry as it adopts when adsorbed. The Fermi energy of the pristine slab is depicted with a blue dashed line.

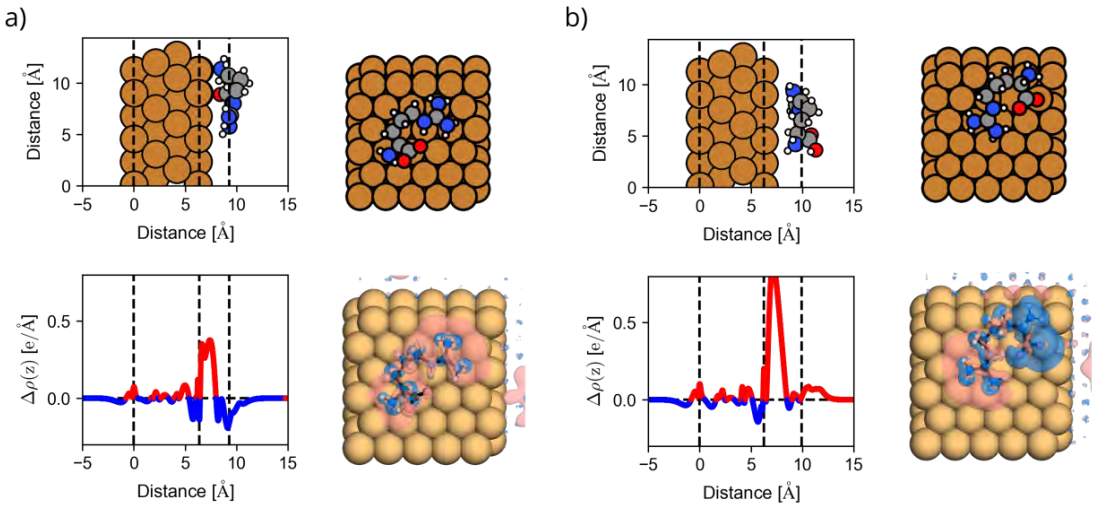


Figure 4.23 – Side and top views of the adsorbed structures of a) Arg on Cu(111) and b) Arg-H$^+$ on Cu(111). Dashed black lines correspond to: the average z position of the atoms in the lowest layer of the surface (left), the average z position of atoms in the highest layer of the surface (middle), the centre of the mass of the molecule (right). Red/blue solid lines (and also red/blue regions) correspond to the electron density accumulation/depletion with PBE0 functional.

protonation of Arg-H$^+$ is favorable on Cu(111), such that Arg-H$^+$ would be predominantely deprotonated. However, we have not observed any spontaneous dissociation upon optimization of Arg-H$^+$ on Cu(111), leading us to conclude that, although favorable, this dissociation of H does not occur without a barrier. On all other surfaces, the barrier for dissociation would be rather high for both molecules.

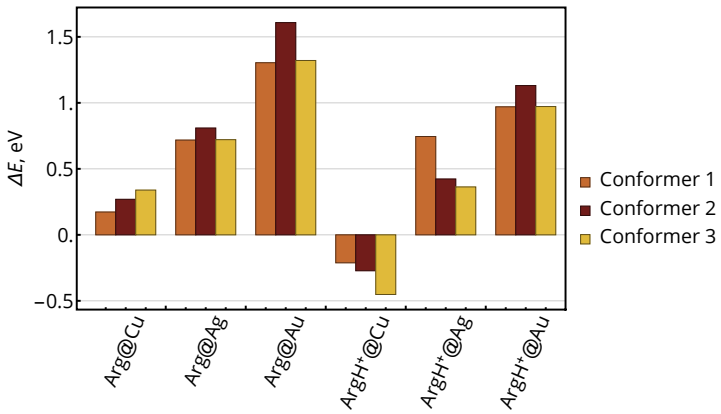


Figure 4.24 – Energy differences upon hydrogen dissociation for selected conformers of Arg and Arg-H$^+$ on all metallic surfaces. $\Delta E = E_{\mathbf{dep}} - E$, where $E_{\mathbf{dep}}$ is the total energy of the dissociated structure after optimization (including the adsorbed hydrogen) and E the energy of the optimized intact structure. A negative ΔE indicates that deprotonation is favored.

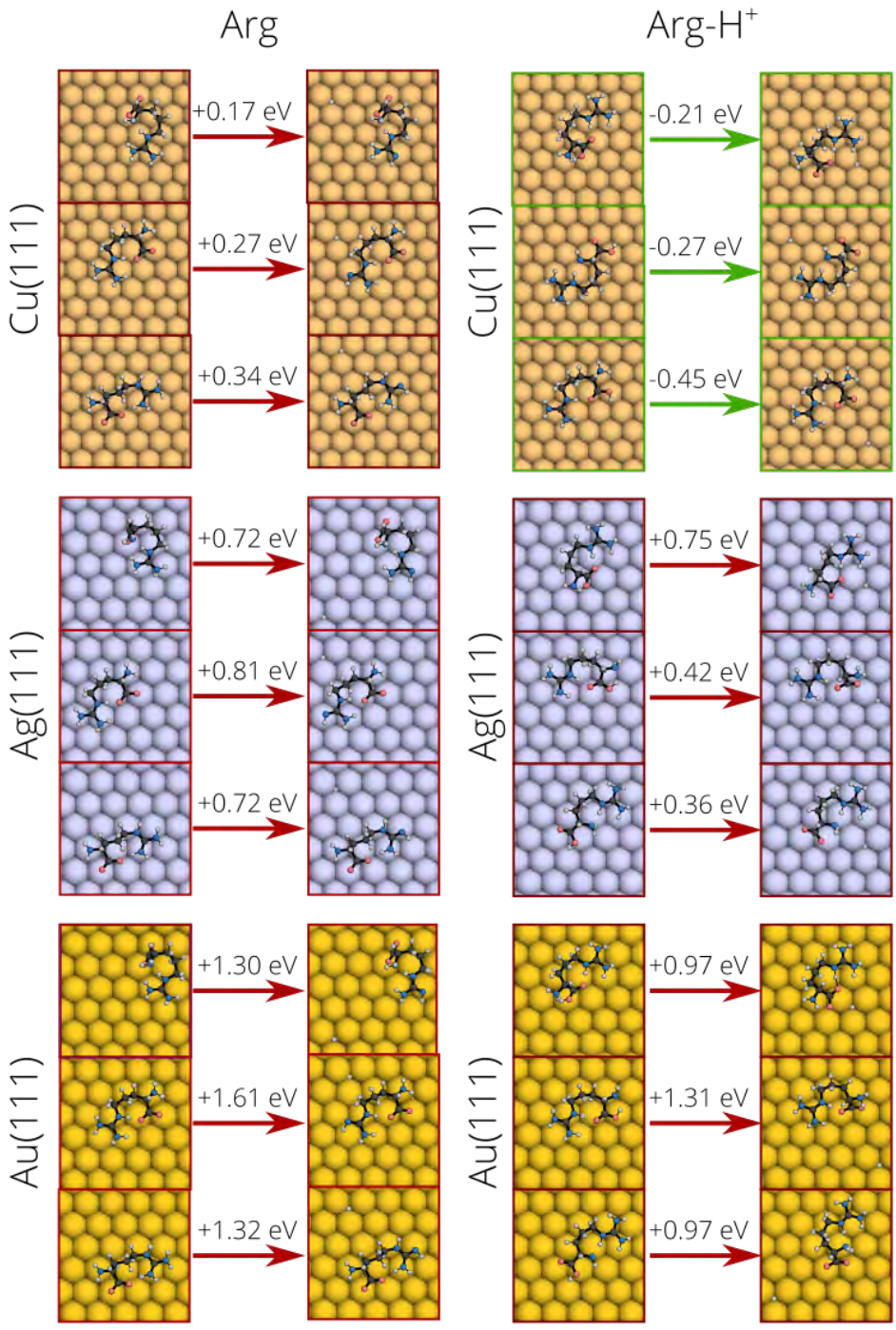


Figure 4.25 – All structures that were analyzed for the calculation of the deprotonation energies. ΔE is also reported in each panel.

4.5 Comparison of DFT with INTERFACE FF

Comparing DFT results with existing FFs is usually beneficial since it helps develop less expensive and more accurate potentials. All the local minima obtained at DFT level of theory were optimized with the INTERFACE-FF [213] using the NAMD package [201]. Calculations were performed with periodic boundary conditions with the same cell size and a number of Cu atoms as in the DFT calculations. We obtained parameters for certain protonation states from existing parametrization of Arg and Arg-H^+ available from CHARMM FF. For the calculation of Arg, two protomers **P1** and **P3** had to be prepared.

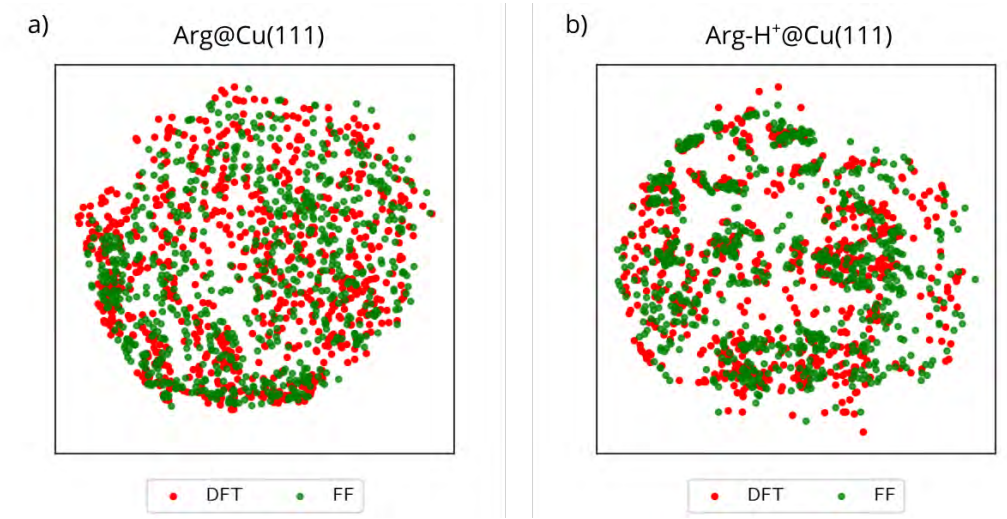


Figure 4.26 – Low-dimensional map of the conformational space of the Arg and Arg-H^+ molecules adsorbed on the Cu(111) surface. The map was optimized considering all DFT and INTERFACE-FF structures. Green dots represent conformations obtained at DFT level of theory and red dots represent conformations obtained after geometry optimization with INTERFACE-FF. Close proximity of the dots reflects their structural similarity.

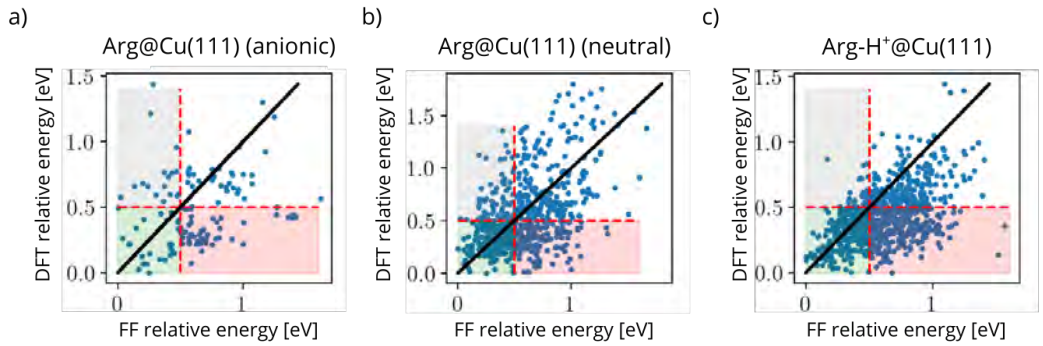


Figure 4.27 – Comparison of the relative energies obtained from DFT optimized structures and the same structures after post-relaxation in with the INTERFACE force field.

From Fig. 4.26, we conclude that both levels of theory represent a similar conformational space. However, Fig. 4.27 shows the comparison of the relative energies obtained from DFT

optimized structures and the same structures after post-relaxation in with the INTERFACE-FF. Dots on the diagonal line represent an optimal correlation. The red area marks structures that lie in the lower 0.5 eV energy range in DFT but above the 0.5 eV energy range in INTERFACE-FF. The green area marks the structures that are in the lower 0.5 eV energy range regardless of the level of theory. The grey area marks the structures that are above the 0.5 eV energy range in DFT but below the 0.5 eV energy range in INTERFACE-FF. From this, we conclude that DFT (PBE+vdW^{surf}) and the INTERFACE-FF yield very different energy hierarchies. Furthermore, Table 4.6 shows that DFT and the FF yield different adsorption site preferences for the amino and carboxyl groups. In particular, DFT predicts that O will adsorb almost exclusively on top sites, consistent with the accepted adsorption site preference of CO groups on the pristine Cu(111) surface. The FF predicts a larger population of other adsorption sites, in particular hollow sites, compared to DFT.

Table 4.6 – Surface site adsorption preferences of chosen chemical groups in Arg and Arg-H$^+$. All numbers are reported as a percentage of the total number of conformers optimized with DFT (PBE+vdW^{surf}) and the INTERFACE-FF.

	Arg@Cu(111)				Arg-H$^+$@Cu(111)			
	Amino		Carboxyl		Amino		Carboxyl	
Adsorption site	DFT	FF	DFT	FF	DFT	FF	DFT	FF
Top	80	53	76	48	59	50	70	45
Bridge	9	18	14	18	18	20	15	22
Hollow-FCC	5	13	4	17	13	15	7	16
Hollow-HCP	6	16	5	17	10	15	9	18

INTERFACE-FF is not reliable for estimation of the energy hierarchies of the molecules, even though the conformational spaces of DFT and FF are very similar. To go beyond single molecules we still need better FFs or ML potentials.

4.6 Conclusions

One of the results of this chapter is the creation of the database of Arg and Arg-H$^+$ adsorbed on three metal surfaces (Cu(111), Ag(111) and Au(111)) containing thousands of structures optimized using DFT. This database is publicly available to download via NOMAD repository [295]. In order to accelerate the development of parametrization of FFs and the training of ML potentials, it is necessary to share these databases to overcome the bottleneck of computationally expensive DFT geometry optimizations, which are required for obtaining relevant information about structure-property relations of interface systems. This is required to achieve the synergy between theory and experiment, in which computational findings may shed light on characteristics of systems that are not accessible via experiment.

Then, using a state-of-the-art dimensionality reduction method, we investigated the conformational spaces of Arg and Arg-H$^+$ in isolation and after adsorption on metal surfaces. The unsupervised dimensionality reduction technique appeared to be a very powerful tool for the rapid analysis of systems with a large number of degrees of freedom. We managed to

easily conclude that all structural motifs of all adsorbed systems are already represented in the conformational space of Arg. In comparison to isolated Arg-H^+, the number of accessible conformations substantially increased after adsorption. Another intriguing discovery that might be easily overlooked without conformational analysis is that the lowest energy structures of adsorbed Arg and Arg-H^+ have remarkably similar conformations since they occur in the same regions of the low-dimensional maps. A closer examination of these lowest energy structures reveals that the dominating orientation of the $C_\alpha H$ group relative to the surface varies between Arg and Arg-H^+. This feature should be studied further for other systems since it may govern the templating of various chiralities of self-assembled structures on the surface. Additionally, a visual depiction of the accessible regions of a conformational space can be provided. For example, spiral-like conformations that lack H-bonds are unfavourable for both Arg and Arg-H^+, while extended structures are favourable for just Arg-H^+. After that, we have specifically investigated why different parts of the conformational space become accessible or are excluded depending on the protonation state and the environment, demonstrating the importance of bond formation and charge rearrangement in these systems.

Arg adsorption occurs through the formation of strong bonds with the surface, with carboxyl and amino groups playing major roles. The surface bindings limit the conformations of this molecule, reducing the number of possible configurations with respect to the numbers observed in the gas-phase. In contrast, Arg-H^+ receives electrons from the surface and becomes less positively charged, which leads to the number of allowed conformations to increase compared to isolated Arg-H^+, which is due to the weakening of intramolecular H-bonds.

After adsorption on Cu, Ag, and Au surfaces, we analyzed the patterns observed for Arg and Arg-H^+. When the substrate is changed, the relative energy order of conformers is mainly conserved, which is a pretty counterintuitive observation. The average adsorption height of the molecules is following the trend: Cu(111) < Ag(111) < Au(111), and Arg is always closer to the same respective surface than Arg-H^+. Most adsorbed Arg conformers bind to Cu(111) surface more strongly than to Ag(111) or Au(111). However, adsorbed Arg-H^+ has similar binding strengths to all surfaces as Arg adsorbed on Cu(111). The computation of dissociation energies leads us to the conclusion that deprotonation of Arg-H^+ is only energetically favourable on Cu(111).

Finally, we show that while INTERFACE-FF may sample the relevant conformational space of these adsorbed molecules, it cannot capture consistent energy hierarchies. Databases like the ones we established will be a valuable source of data for future parameterization and the development of cheaper potentials.

In general, there is no accessible collection of isolated local minima conformers to start a structure search from, for any random system of interest. Few suitable packages exist for such tasks, and all methods for creating starting structures with different molecular

orientations with respect to the surface must be established manually. In the following chapter, we will present a package that will assist in carrying out these calculations, paving the way for the acceleration of database development for interface systems.

It turns out that any repetitive endeavour – whatever the industry – can be automated within the context of rising digitisation.

"Fully Automated Luxury Communism: A Manifesto", Aaron Bastani

5

Generation and search of the flexible molecules with respect to fixed surroundings

The previous chapter was devoted to the study of single molecule adsorption on various surfaces, and it required a significant degree of human engagement in terms of data production, data organization, and data interpretation. In order to capture the trends across all the amino acids on different surfaces, such work should be performed for other systems as well. However, it is not common to have an available structure database for gas-phase structures that are useful as beginning structures. Moreover, in cases where the adsorption pattern is composed of repeating templates, it is best to take into account PBC in order to perform the structure search. In the age of high-performance computers, the workflow should make use of parallelization in the data acquisition process. Existing software packages that are capable of performing sampling of conformational spaces are typically coupled to a small number of specific electronic structure packages, which limits the usefulness of such packages in practice. Also structure search packages are not tailored to sample flexible adsorbates and their assemblies with respect to specified surroundings e.g surfaces or cavities. In response to these challenges we have developed a program that addresses all of these issues and is meant for sampling the conformational spaces of flexible molecules and their assemblies on surfaces. In this chapter we present an automated workflow that allows us to easily generate and perform geometry optimizations.

5.1 GenSec package for structure search of the interfaces

Random structure search is the basis for more sophisticated methods such as Bayesian optimization [227] and evolutionary algorithms [217, 220], and is the method employed in the Generation and Search (GenSec) package. Random structure search is also used in crystal structure prediction [216, 296] and shows a decent probability of identifying low-energy minima [214, 215]. The efficiency of the random structure search can be increased dramatically first by imposing constraints on the generated structures, avoiding clashes between atoms and keeping the database of previously calculated structures in order to avoid repetitive calculations. Starting from the procedure for generating different conformers of the isolated molecules, we then describe the extension of such procedures to enable simulations of these conformers with respect to fixed surroundings (*fixed frames*) that can be, in general, 1D (e.g. ions), 2D (e.g. surfaces) or 3D (e.g. solids) static references. In short, GenSec performs a quasi-random global structure search, with the ability to choose different internal degrees of freedom and sample them with respect to specified fixed surroundings. The geometry optimizations are performed by a connection with the Atomic Simulation Environment (ASE) [297] environment, which can be connected to many electronic structure and FF packages and offers the choice of a variety of geometry optimization routines, which we have improved as detailed in Section 5.6. The connection to the ASE database support makes it possible to perform multiple searches in parallel with shared access to the information obtained from all the searches.

GenSec is written using Python 3 and distributed under the GNU Lesser GENERAL Public License and available from:

`https://github.com/sabia-group/gensec`

5.2 Workflow of the GenSec package

The workflow of GenSec consists of the three main steps (Fig. 5.1):

1. Random generation of a candidate structure with specified constraints
2. Comparing the generated structure with the structures already contained in the databases
3. Performing a geometry optimization if the structure is unique, and adding all optimization steps from the geometry relaxation as well as the local minima to the database

The search performs a user-specified number of unique relaxations, or the algorithm stops if it cannot find any more unique structures within the user-specified number of trials. The processes of structure generation and geometry optimization can be parallelized and run independently, and the details of each step are described in the following sections.

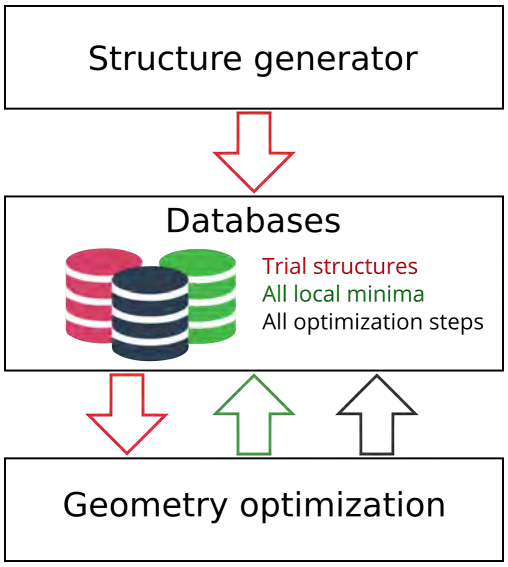


Figure 5.1 – Workflow of the GenSec package.

5.3 Structure generation

The generation of structures is implemented as a standalone procedure, and can generate structures via multiple independent processes, while creating a central database, where the unique unrelaxed structures are stored. The generation of these structures is based on the internal degrees of freedom of the molecule, such as the dihedral angles, position of center of mass (COM), and orientation of the molecule. Starting from the generation of different conformers of isolated molecules, we then extend the procedure to generate self-assemblies on surfaces.

5.3.1 Internal degrees of freedom: dihedrals

The very first step is to identify the connectivity of the molecule. ASE allows reading the molecule 3D coordinates from a template in multiple chemical formats (Fig. 5.2 a), after which it creates the connectivity matrix based on the covalent radii distances between atoms. If the spheres of two atoms defined by their atomic covalent radii that are tabulated in ASE, overlap, they will be counted as bonded atoms. This connectivity matrix is then represented as an undirected graph that reflects the bonding information between atoms as shown in Fig. 5.2 b. The dihedral angle for organic chemical systems is defined as the angle between two planes both of which are defined by three atoms that are connected by two bonds and both of the planes have to share a bond that is not the terminal bond of both planes [298]. For producing different conformers with the same chemical bonding we are interested in changing of dihedral angles of those planes, where the shared bond is freely rotatable. The rotatable bonds are identified from the graph in Fig. 5.2 b) with the following rules:

1. First select all the atoms that have two or more bonds - potentially they will be two central atoms forming the dihedral angle if they are not in a cyclic structure

2. Exclude the atoms with exactly 4 bonds, three of which are terminating atoms. Such exclusion removes e.g. CH_3 terminating groups

3. Exclude the atoms with three bonds for which two of the atoms are terminating hydrogens - with that we also exclude groups such as NH_2

4. Finally exclude the atoms that have two bonds, one of which is terminating hydrogen which appears in the carboxyl group

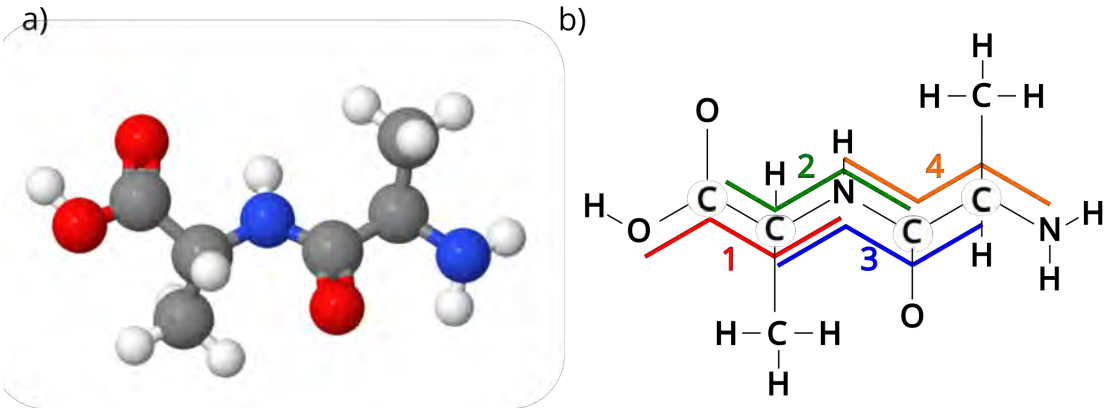


Figure 5.2 – a) 3D representation of a flexible molecule (di-L-Alanine); b) representation of di-L-Alanine as an undirected graph together with rotatable bonds automatically identified using GenSec coloured in red, green, blue and orange.

With this procedure the rotatable bonds of the molecules can be automatically identified after the construction of the connectivity matrix of the template molecule and in the case of di-L-alanine only four rotatable bonds will be identified and used for creating the different conformations. It should be mentioned that we pay additional attention to exclusion of the rotatable groups containing light hydrogen atoms. This exclusion can be allowed since during geometry optimization light atoms will anyways move if necessary resulting in the preferred orientation of the whole chemical group with respect to the rest of the molecule. If particular rotatable bonds are of interest during the search, this information can be anyways manually specified in the parameters file. The only thing left to address is that the rotatable bonds obtained with the algorithm described above can occur in cycles, which creates redundant degrees of freedom. To exclude the rotatable bonds that appear in cycles we use the networkx package [299] that uses Johnson's algorithm to detect cycles in a graph [300]. The rotatable bonds are then excluded by simple filtering that requires at least one of the central atoms not to be in a cycle.

After that, random values of the dihedral angles can be applied to these rotatable bonds through the ASE interface. The resulting molecule is checked for internal clashes by constructing the connectivity matrix again and comparing it with the initial template. The procedure described up to this point enables the generation of random isolated conformers. In order to model adsorbed species, additional degrees of freedom such as orientation and

positioning of the molecule with respect to fixed frames had to be implemented.

5.3.2 Generating molecules with respect to fixed frames

In order to sample the configurational space of rigid molecules with respect to fixed frames we added two additional degrees of freedom to a template molecule: the orientation and positioning of the COM of the molecule. The COM of the molecule is a simple translational degree of freedom, which locates the molecule relative to a specific origin in Cartesian coordinates. The COM is defined as $\mathbf{r^*} = \sum_k m_k \mathbf{r}_k / \sum_k m_k$, where m_k and $\mathbf{r}_k$ are the mass and coordinates of the k-th atom in the molecule.

For the orientation of the molecules we must introduce a notation to describe the orientation of the molecule, which is not trivial in the case of flexible structures. Rotations are performed with using Hamilton's quaternions [301], which are closely related to the geometrically intuitive angle and axis notation. These are presented as an ordered set of 4 real quantities which we write as

$$\mathbf{q} = [q_0, q_1, q_2, q_3],$$

or as a combination of a scalar and a vector

$$\mathbf{q} = [q_0, \mathbf{v}],$$

where $\mathbf{v}$=$[q_1, q_2, q_3]$. In order to use quaternions for spatial rotations around some unit vector $|\mathbf{v}|$=1 on angle θ, we can use a unit quaternion $\mathbf{q} = [\cos(\theta/2), \mathbf{v}\sin(\theta/2)]$, with rotations implemented as the action of an operator $R_\mathbf{q}$ on a 3-dimensional vector:

$$R_\mathbf{q}(\mathbf{x}) = \mathbf{R}(\mathbf{q}) \cdot \mathbf{x},$$

where $\mathbf{x}$ are Cartesian coordinates of atoms in the system and $\mathbf{R}(\mathbf{q})$ is a matrix that in component form can be written as follows:

$$\mathbf{R}(\mathbf{q}) = \begin{pmatrix} 1-2q_2^2-2q_3^2 & 2q_1q_2-2q_0q_3 & 2q_1q_3+2q_0q_2 \\ 2q_2q_1+2q_0q_3 & 1-2q_3^2-2q_1^2 & 2q_2q_3-2q_0q_1 \\ 2q_3q_1-2q_0q_2 & 2q_3q_2+2q_0q_1 & 1-2q_1^2-2q_2^2 \end{pmatrix}. \tag{5.1}$$

In order to describe a rotation of the molecule such that it can be compared to other rotations, we use the orientation associated with the eigenvectors of the inertial moments of the rigid molecule. The moment of inertia matrix is given by

$$\mathbf{I} = \sum_k m_k((\mathbf{r}_k \cdot \mathbf{r}_k)\mathbf{E} - \mathbf{r}_k \otimes \mathbf{r}_k), \tag{5.2}$$

where m_k and $\mathbf{r}_k = (x_k, y_k, z_k)$ are the masses and coordinates of k-th atom in the molecule, $\mathbf{E}$ is the identity tensor and $\otimes$ is the tensor product. The eigenvector with the lowest corresponding eigenvalue (shortest principal axis) is chosen as the main vector of the molecule.

The eigenvector with the corresponding largest eigenvalue (longest principal axis) is chosen as the minor vector of the molecule. The signs of these axes are determined by drawing the vector from the first to the last atom of the molecule and calculating its dot products with the principal axis. The principal axis for which the dot product with this vector is positive is chosen to be the main and minor vectors. By default those atoms are literally chosen as first to last heavy atoms provided in the template file, but also can be manually defined by user tailored for a particular system of interest. The main vector is aligned to the z Cartesian axis and the minor vector is aligned to the x Cartesian axis - this orientation is considered the "initial" orientation for a particular molecule. All other orientations of the molecule are treated with respect to its "initial" orientation. The representation that is stored as an internal degree of freedom has a human-readable notation similar to quaternions: it is composed of the main vector of the molecule and the angle through which one would have to rotate the molecule around this axis in order to put the molecule in the "initial" orientation, with the main vector aligned with the z-axis. This also allows for a discretization of the space of orientations. There are three principal axes that are obtained for each molecule and only two of them are needed to identify the "initial" configuration of the molecule.

5.3.3 Self-assembly generation with respect to fixed frames

Fixed frames, with respect to which the sampling of the configurational space is performed, can be of any form i.e., atoms, molecules, 2D periodic structures and 3D cavities. After some unique configuration of the molecule is generated, the distances between all the atoms of the molecule and the fixed surrounding are calculated and, if all of them exceed a certain value (no overlaps found) that can be specified before the search, the structure can proceed to geometry optimization. When dealing with periodic structures with particular PBC one has to take care of potential clashes of the molecule with its periodic images. Using the minimum image convention, all the atoms are mapped inside the unit cell and checked for clashes, which in the case of a single molecule is also done with the creation of the connectivity matrix that is constructed taking into account PBC.

Having specified the template molecule and the fixed surroundings, one can set the number of molecules that should be produced in the unit cell. GenSec will then produce molecules in an iterative way and assign to them specified values for internal degrees of freedom, which can be the *same* or *different*. For example, one can sample molecules with the same conformations but having different orientations, or with the same overall orientation (for example flat-lying) but with different conformations. This allows us to impose some constraints on the generated structures. In the case of generating multiple molecules, the distance between atoms of the molecules can be specified according to the goals of the search.

Examples of self-assembled structures obtained with GenSec for F6-TCNNQ/MoS_2 with 2 molecules in a (4x8) MoS_2 supercell were used for investigation of the temperature-dependent electronic ground-state charge transfer in vdW heterostructures [302] and can be found in Fig. 5.3. GenSec automates routine tasks and does not require using any FFs for the generation of self-assemblies.

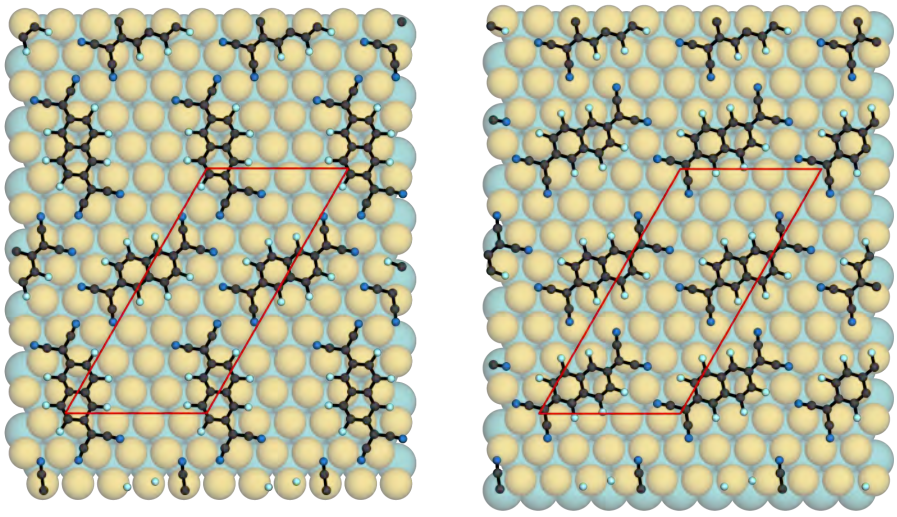


Figure 5.3 – Examples of self-assembled structures obtained with GenSec for F6-TCNNQ/MoS2 with 2 molecules in a (4x8) MoS_2 supercell.

5.3.4 Constraints of the search

Without imposing constraints, the number of configurations to sample is too large. For real-life applications specific orientations and positions of molecules with respect to specified surroundings have to be targeted. The allowed COM space in GenSec can be specified by a range of points in the x, y, and z directions. For each direction, one can specify the boundaries and number of points that lie within those boundaries. For example, in order to generate all the structures that lie in the same 2D plane, the boundary for the direction perpendicular to this plane must contain only one particular value, which is very useful for modelling planar assemblies on the surfaces.

In the case of orientations, the discretization is performed on the angle of self-rotation. This is quantified by specifying the allowed angle of rotation. For example, if the number equals 60, six rotations of the molecule will be generated, and if the number is 360 self-rotations are basically forbidden. In this case the vector, associated with the principal axes corresponding to the lowest eigenvalue of the moment inertia tensor will solely identify the orientations. The main vector of the molecule is sampled from a uniform distribution between specified maximum and minimum values for q_1, q_2 and q_3. An example of different orientations and their notations are reflected in Fig. 5.4.

Having set these routines, one can produce an arbitrary amount of molecules per unit cell with specified orientations and conformations, that will be clash-free structures ready for geometry optimization. However, before geometry optimization, which can be very time-consuming, we check the generated configurations against the database, and only if the configuration is unique, is a geometry optimization performed.

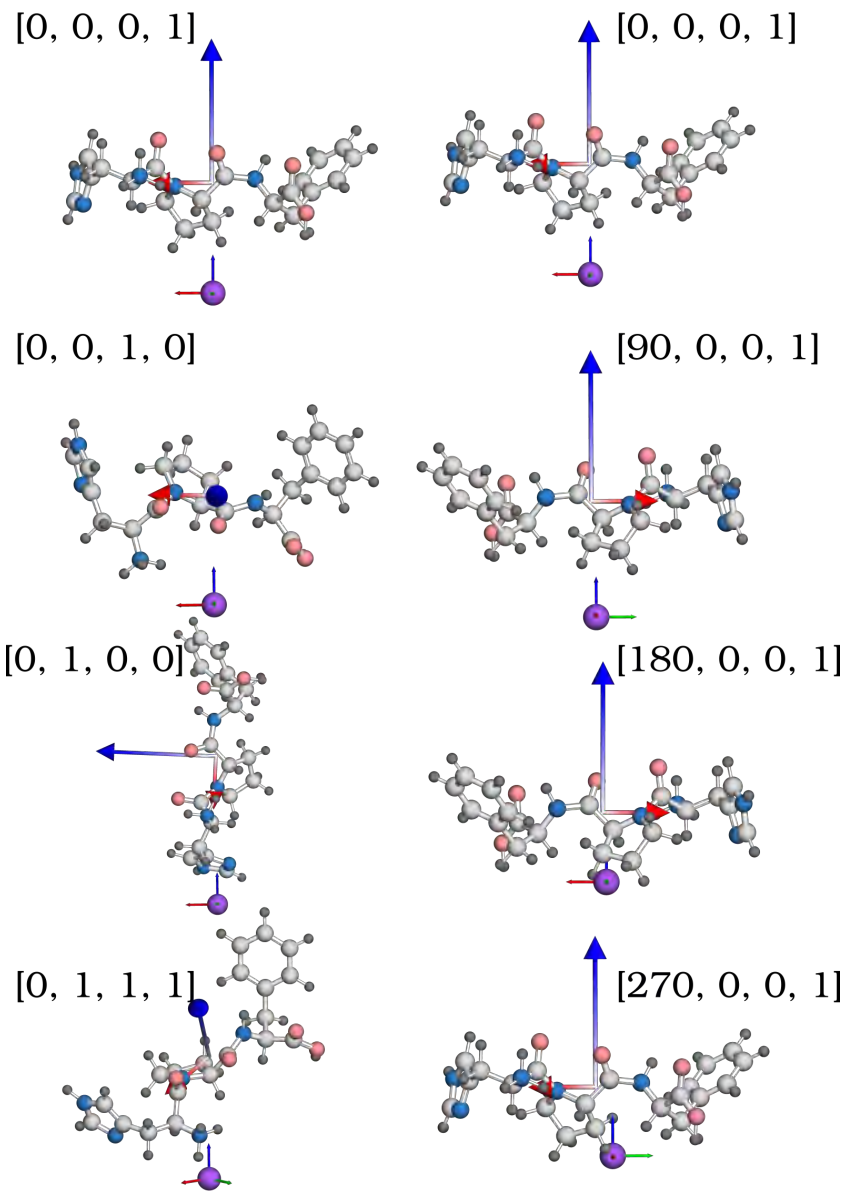


Figure 5.4 – Examples of the orientations for two different conformers. A big blue vector denotes the main direction, smaller red vector denotes the minor direction. The magenta circle is a Na atom from which one can see three small vectors: red - x-axis, green - y-axis and blue - z-axis. The first number in brackets denotes a "self-rotation" around the main vector with respect to the "initial" orientation and three other numbers represent the direction of the main vector.

5.4 Database creation and filtering of the structures

Here we describe how the uniqueness of a randomly generated structure is checked. The database is created in the SQLite3 format, which is a self-contained, server-less, zero-

configuration database. Every row in the database contains atom positions and calculated forces on all atoms together with internal degrees of freedom that represent the system. The internal degrees of freedom are stored in the database separately with the notation “t” for torsion angle numbers which are automatically identified, “q” for orientation, which has 4 values for each molecule and “c” for COM, that has three values that are defined with respect to the Cartesian origin. For a given configuration, one can easily create a query that will extract all the configurations from the database with the same corresponding torsion angles values within a given threshold. If the number of filtered structures is more than one, the initially generated structure is not unique and should not be further optimized. This procedure easily extends to multiple molecules. If the number of structures is more than one, and if checks on the orientations and COMs are specified, then each filtered structure will be compared to the structure under trial. If the distance between COMs of structures within one system is more than the specified value (default is 0.5 Å), the structures will be considered as different. For the orientations, the self-rotation and the angle between the main vectors of the molecules are checked separately. If both the difference between self-rotations and the angle between main vectors are greater than specified values (the default values are 30° in both cases), those structures will be considered different. If the generated structure is unique, it will proceed for geometry optimization, deleted from the database of generated structures. After relaxation, the trajectory will be added to the database of trajectories, and the local minimum will be added to the database of local minima.

We also implemented restarting procedure that is very important in the workflow of the GenSec, since it provides a seamless way of continuing the unfinished processes and continuing the database generation especially when multiple parallel processes are utilized for structure search.

5.5 Geometry optimization workflow

One of the strengths of GenSec is that it straightforwardly interfaces with the ASE environment, which allows us to perform energy and force evaluations using the most popular electronic structure packages, as well as empirical potential codes. These packages can be used to obtain energies and forces of the system at each step of geometry optimization to find local minima. The structures from every step of these geometry optimizations are stored in the database which helps to find the new unique trial structure more efficiently and provides more data for training of potentially cheaper potentials. The limitations on size of the database are limited by the capabilities of SQL

The bottleneck of exhaustive searches is *ab initio* geometry optimizations that can be sped up with the use of preconditioning of geometry optimization algorithms. There are some routines already available for geometry optimization in ASE. However, in the following, we describe the preconditioning of the BFGS algorithm that takes into account an approximate Hessian matrix that contains information about connectivity and physical interactions in the system. This allows the algorithm to make a better choice for the next step toward finding

the local minima. This implementation is tailored explicitly for interface systems, and its description and performance will be described in the following section.

5.6 Preconditioner for geometry optimization

Having routines for sampling different parts of the conformational space of a system is necessary to minimize the system's energy. It was shown that the energy hierarchy of the structures for which only single-point calculations were performed could change dramatically after their geometry optimization [303]. The most popular geometry optimization algorithms are quasi-Newton algorithms that require input information about the energy and forces of a configuration and iteratively find the local minima of the system. Based on the forces and energies of adjacent steps, the algorithm updates the approximate Hessian matrix. One of the most successful schemes is the BFGS algorithm, which was described in Section 3.2. However, the potential energy surface of the system can be highly anisotropic, which results in poor performance (slow convergence) of the geometry optimization. In order to make the shape of the potential more isotropic, one can use preconditioners that perform a metric transformation of the coordinate system, thus making the shape of the potential energy surface smoother and improving the efficiency of finding the nearest local minima.

By default, the initial Hessian matrix is a scaled identity matrix, and initializing the Hessian matrix with some information about the system can improve the speed of convergence of the geometry optimization algorithm. A combination of the Hessian matrix with different preconditioning schemes showed a performance gain when applied to molecular crystals [304], for example. For modelling condensed phase systems, the best performance is demonstrated using the Exponential preconditioner [249]. For modelling gas-phase molecular systems, the force-field-like preconditioner proposed by Lindh et al. [236] is widely used due to its simplicity. Specifically for the interfaces, we propose a scheme that allows us to combine these different approximations and apply them to the corresponding parts of the system, i.e. Lindh applied to the molecular part and Exponential to the solid part. We also introduce a vdW part that allows us to calculate a LJ Hessian matrix based on the vdW parameters developed in the TS-vdW method that can be applied to the parts of the Hessian where it can play an important role. The pictorial representation of the proposed scheme can be found in Fig. 5.5. First, we describe the workflow of the LJ preconditioning scheme and then show some results for model systems where the combined preconditioning scheme was applied.

5.6.1 Lennard-Jones-like Hessian matrix

Here we would like to introduce a preconditioning scheme that could treat vdW bonded systems. First, we introduce the notations used in the scheme:

$$\begin{aligned} &A, B \in \{0, .., N-1\}, A \neq B - \text{interacting atoms,} \\ &i, j \in \{0, 1, 2\} - \text{cartesian axes,} \\ &\delta_{ij} \text{ is Kronecker delta.} \end{aligned} \tag{5.3}$$

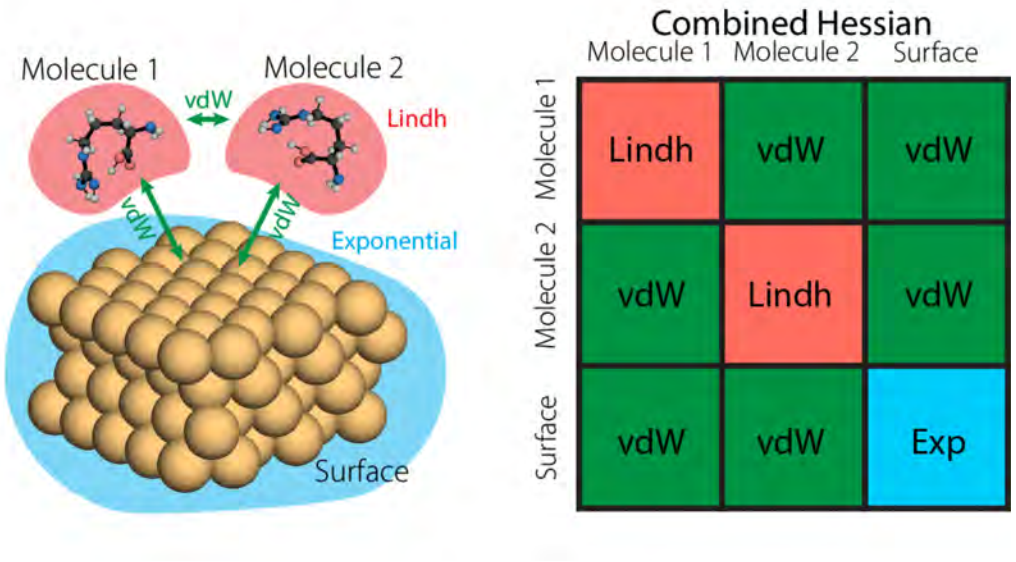


Figure 5.5 – Representation of the construction of the approximated Hessian matrix using different preconditioning schemes a) Representation of the different parts of the system for which different preconditioning schemes can be applied separately; b) the combined approximated Hessian matrix constructed using different preconditioner schemes applied for different parts of the system.

We derive the Hessian starting from a Lennard-Jones 12-6 potential:

$$E_{AB}^{\mathrm{LJ}} = 4\varepsilon\left[\left(\frac{\sigma}{R^{AB}}\right)^{12} - \left(\frac{\sigma}{R^{AB}}\right)^{6}\right] = \frac{C_{12}^{AB}}{(R^{AB})^{12}} - \frac{C_{6}^{AB}}{(R^{AB})^{6}}, \tag{5.4}$$

where indices A and B denote different atoms,

$$R^{AB} = ((x_1^B - x_1^A),(x_2^B - x_2^A),(x_3^B - x_3^A)) \tag{5.5}$$

$$R_i^{AB} = x_i^B - x_i^A \tag{5.6}$$

is the distance between atoms A and B and

$$|R^{AB}| = R = \sqrt{(x_1^B - x_1^A)^2 + (x_2^B - x_2^A)^2 + (x_3^B - x_3^A)^2}. \tag{5.7}$$

The C_6 coefficients are taken from [305, 306]. So we proceed to take the first derivative:

$$\frac{dE_{AB}^{\mathrm{LJ}}}{dx^A} = \frac{6C_6}{R^8}R^{AB} - \frac{12C_{12}}{R^{14}}R^{AB}. \tag{5.8}$$

By assuming that LJ potential adopts a minimum at $R_0^{AB} = \frac{R_{\text{vdW}}^A + R_{\text{vdW}}^B}{2}$, one can derive C_{12} from Eq. 5.8 as

$$C_{12}^{AB} = \frac{1}{2} C_6^{AB} * (R_0^{AB})^6, \tag{5.9}$$

as discussed in [163].

After that we take the second derivative and get:

$$\frac{\partial E_{AB}^{\text{LJ}}}{\partial x_i^A \partial x_j^B} = \frac{\partial \left(\frac{6C_6}{R^8}\right)}{\partial x_j^B} R_i^{AB} + \frac{6C_6}{R^8} \frac{\partial R_i^{AB}}{\partial x_j^B} - \frac{\partial \left(\frac{12C_{12}}{R^{14}}\right)}{\partial x_j^B} R_i^{AB} - \frac{12C_{12}}{R^{14}} \frac{\partial R_i^{AB}}{\partial x_j^B} =$$

$$= \frac{48C_6 R_i^{AB} R_j^{AB}}{R^{10}} - \frac{6C_6}{R^8}\delta_{ij} - \frac{168C_{12} R_i^{AB} R_j^{AB}}{R^{16}} + \frac{12C_{12}}{R^{14}}\delta_{ij}. \tag{5.10}$$

After simplification the LJ Hessian will be:

$$H_{(3A+i),(3B+j)}^{\text{LJ}} = \frac{48C_6(x_j^B - x_j^A)}{R^{10}}(x_i^B - x_i^A) - \frac{168C_{12}(x_j^B - x_j^A)}{R^{16}}(x_i^B - x_i^A) - (\frac{6C_6}{R^8} - \frac{12C_{12}}{R^{14}})\delta_{ij}. \tag{5.11}$$

However, in practical simulation, we want to employ a preconditioning scheme in situations that may be far from the ideal minimum of such potential. In that case this constructed Hessian will not be positive definite. To overcome this issue we apply the strategy as in the Lindh approach for constructing the Hessian matrix, where the Hessian is estimated for a particular configuration as it would be if that configuration was a minimum [236]:

$$E(\mathbf{R}) = E\left(R_1^0, \cdots R_N^0\right) + \sum_{i=1}^{N} \frac{\partial E}{\partial R_i}\bigg|_{R_i = R_i^0} \left(R_i - R_i^0\right) + \frac{1}{2} \sum_{i,j=1}^{N} \left(R_i - R_i^0\right) \frac{\partial^2 E}{\partial R_i \partial R_j} \left(R_j - R_j^0\right) + \cdots \tag{5.12}$$

where the second term cancels to zero. Our model Hessian is then

$$H_{(3A+i),(3B+j)}^{\text{LJ}} = \frac{\partial^2 E(R_0^{AB})}{\partial R_i^A \partial R_j^B}|_{R_i^{AB} = R_{0i}^{AB}, R_j^{AB} = R_{0j}^{AB}} = \frac{48C_6 R_{0i}^{AB} R_{0j}^{AB}}{(R_0^{AB})^{10}} - \frac{6C_6}{(R_0^{AB})^8}\delta_{ij} - \frac{168C_{12} R_{0i}^{AB} R_{0j}^{AB}}{(R_0^{AB})^{16}} + \frac{12C_{12}}{(R_0^{AB})^{14}}\delta_{ij} \tag{5.13}$$

Obviously, values R_i^{AB} can be far from equilibrium values and this will lead to Hessian matrix be not positive definite. Instead, the R_i^{AB} is scaled to the length of R_{0i}^{AB} in order to satisfy the assumption that the system is near the local minimum:

$$R_{0i}^{AB} = R_i^{AB} * \frac{|R_0^{AB}|}{R} \tag{5.14}$$

With use of prefactor coefficient ρ_{AB} for the whole Hessian matrix we set the vdW interaction at the distances larger than $2 \times R_0^{AB}$ to be negligible, basically setting preconditioning only

for nearest neighbour atoms:

$$\rho_{\mathrm{AB}} = \exp\left[\alpha_{\mathrm{AB}}\left((R_0^{\mathrm{AB}})^2 - R^2\right)\right], \tag{5.15}$$

by fitting of the the parameters α_{AB} for each pair of R_0^{AB}. Finally we get

$$H^{\mathrm{LJ}}_{(3A+i),(3B+j)} = \rho_{\mathrm{AB}} \frac{\partial^2 E(R_0^{AB})}{\partial R_i^A \partial R_j^B}|_{R_i^{AB}=R_{0i}^{AB}, R_j^{AB}=R_{0j}^{AB}} \tag{5.16}$$

This scheme is implemented in GenSec and available with use of the flag "vdW" for preconditioning of the geometry optimization. The scheme was tested on model LJ Ar_n clusters, where n reflects the number of atoms in the cluster, the local minima of which were taken from the database [307]. For all the minima random displacements of 0.01 Å were applied for each atom. The BFGS TRM method was used for geometry optimization. The geometry optimizations were carried out with the vdW preconditioning scheme and with the scaled identity matrix using 70 as the scaling factor (default in ASE) as initial Hessian (which is also will be noted as unpreconditioned case). The performance gain is calculated as the number of steps required to reach the local minima for unpreconditioned case divided by the number of steps required to reach the same local minima with use of initial preconditioned vdW Hessian matrix, and shown as a function of cluster size in Fig. 5.6.

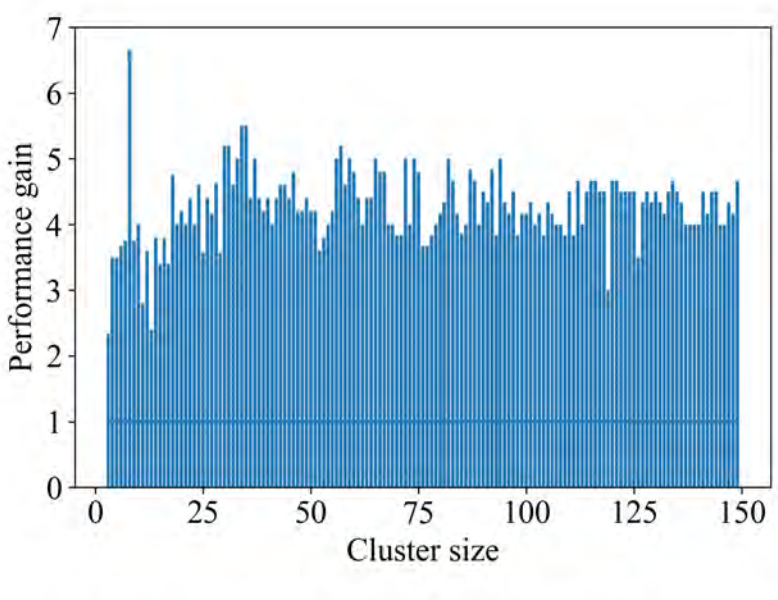


Figure 5.6 – Performance gain for the geometry optimization of LJ clusters of different sizes using vdW preconditioning scheme, compared to the unpreconditioned case.

Identical structures are obtained with and without the application of the preconditioner, and our preconditioning scheme shows significant performance gains for these systems, where the only force acting on the atoms is LJ force. Now we proceed to the combination

of the different Hessian schemes, and apply the combined Hessians to model interface systems.

Next we adapted the Exponential and Lindh preconditioning schemes described in Sec. 3.2.4 into the workflow of GenSec. To test the performance of the Exponential preconditioner, we optimized bulk $N \times N \times N$ fcc Cu unit cells were optimized using Effective Medium Theory (EMT) potential implemented in ASE [308], and to test the Lindh preconditioning scheme we used PBE with light settings implemented in FHI-aims to relax different Alanine dipeptide conformers obtained with GenSec. The results are shown in Fig. 5.7 - in both cases a significant performance gain is observed.

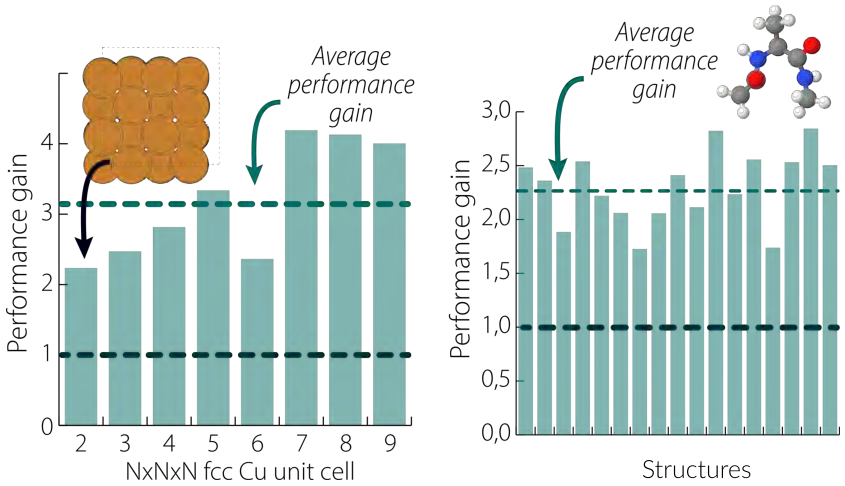


Figure 5.7 – Performance gain for geometry optimization with Exponential preconditioning scheme applied to Cu bulk systems (left) and performance gain of the Lindh preconditioning scheme applied to geometry optimization of different conformers of Alanine dipeptide structures (right).

Randomly generated geometries can be far away from any local minima. Especially for flexible molecular systems, the local environments can change dramatically during geometry optimization due to torsional rotations. In this case, the local PES cannot be approximated quadratically. To overcome this issue, one of the approaches could be to restart the BFGS procedure and reinitialize the Hessian matrix during the geometry optimization. One way to do this is to "reset" Hessian matrix after some fixed number of steps. By contrast, we restart and update the Hessian matrix depending on the change of the root mean square displacement (RMSD) value between snapshots in the geometry optimization trajectory:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \mathrm{d}_i^2}, \tag{5.17}$$

where d_i is the distance between the atomic positions. Randomly created flexible molecules are usually far away from a local minimum which means that harmonic approximation of quasi-Newton procedure that was initially made will not be valid after several optimization

steps and reinitialization of the Hessian matrix allows the BFGS algorithm to find local minima faster. For the same set of conformers of Alanine dipeptide presented in Fig. 5.7 we applied this scheme, where the Hessian matrix was reinitialized after the RMSD exceeded the specified value. The definition of the RMSD value is system specific and should be chosen with caution in order to obtain the best performance results - choosing the value to be too small will reinitialize the Hessian update too often, which could lead to a decrease of performance of BFGS algorithm. Harmonic approximation could be valid if the atom displacements are within 0.2 Å from their equilibrium positions [234]. The results in Fig. 5.8 show that this strategy can be twice as efficient compared to the case where preconditioning was applied only at the initialization step.

5.6.2 Combining the preconditioners

Having all of the preconditioning schemes implemented in GenSec, we created the model system of one hexane molecule adsorbed on Rh surface to test the performance of the combined preconditioner illustrated in Fig. 5.9. The system can be clearly separated into molecular and surface parts, and the strategy for applying the different preconditioning schemes is the following: the constructed initial Hessian can be obtained for the whole system using Exponential or Lindh. One can apply different preconditioning schemes to different parts, i.e, Exponential for the substrate part and Lindh for the molecular part. For the Hessian matrix elements that correspond to off-block-diagonal elements, that do not correspond solely to molecular or substrate parts, one can apply the vdW preconditioning scheme, or simply set those elements to 0.

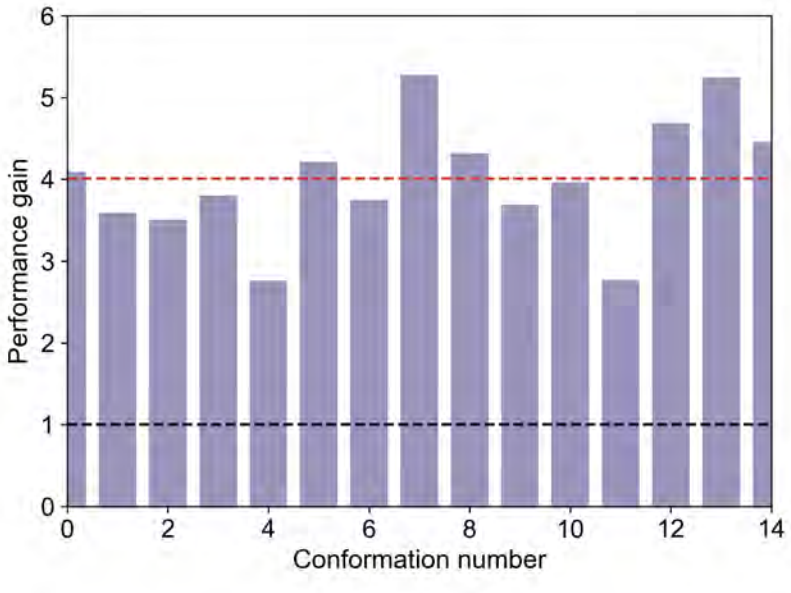


Figure 5.8 – Performance gain for geometry optimization of different randomly generated conformers of Alanine dipeptide with reinitialization of the Hessian after the conformational change exceeds 0.1 Å.

$$H_{(3A+i),(3B+j)} = \begin{cases} \text{Lindh term,} & \text{if A, B are in molecule} \\ \text{vdW or 0} & \text{if A, B belong to different parts of the system} \\ \text{Exponential} & \text{if A, B are in surface} \end{cases} \quad (5.18)$$

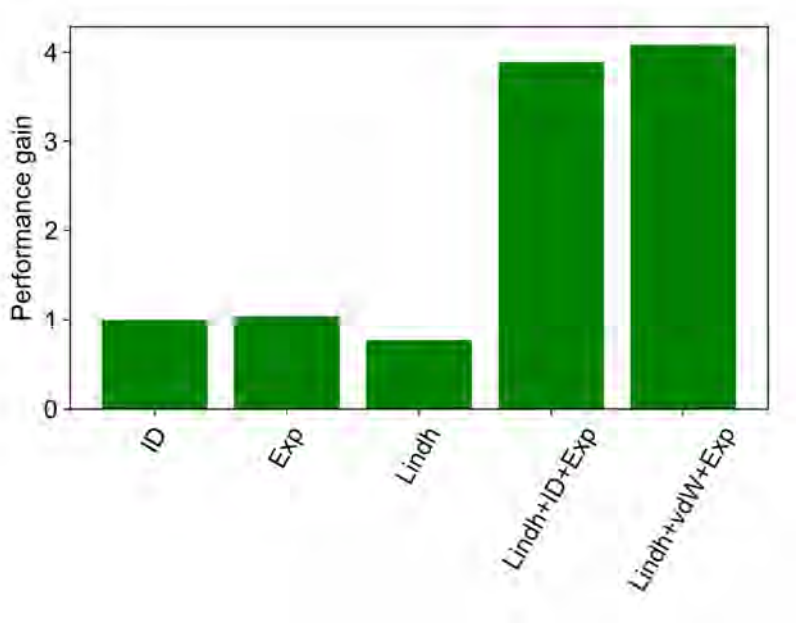


Figure 5.9 – Performance gain for geometry optimization with different preconditioning schemes applied to geometry optimization of hexane on Rh surface.

For the model system the effects of applying the different preconditioning schemes are shown in Fig. 5.9. The PES was constructed using adaptive intermolecular reactive bond order (AIREBO) potentials [309, 310] for carbohydrates, embedded atom model (EAM) interatomic potential for Rh atoms [311, 312] and LJ potential for interactions between molecule and surface. It is clear that applying a combined preconditioner is more efficient than applying a single preconditioning scheme to the whole system. Inclusion of the vdW preconditioning scheme doesn't give a significant performance gain in this case. This is likely because the vdW forces are never the largest forces in the optimization path. Nevertheless, this strategy can be efficient, and applying it to the broader range of systems with different potentials will be the scope of future investigations.

The package is open-source and ready for usage. Tutorials and documentation can be found at https://github.com/sabia-group/GenSec

5.7 Application to di-L-alanine on Cu(110)

Having presented the GenSec package, we now provide an example of how it can be applied to a system that has been previously investigated experimentally, namely di-L-alanine adsorbed on the Cu(110) surface. STM was utilized to investigate the sub-monolayer formation of

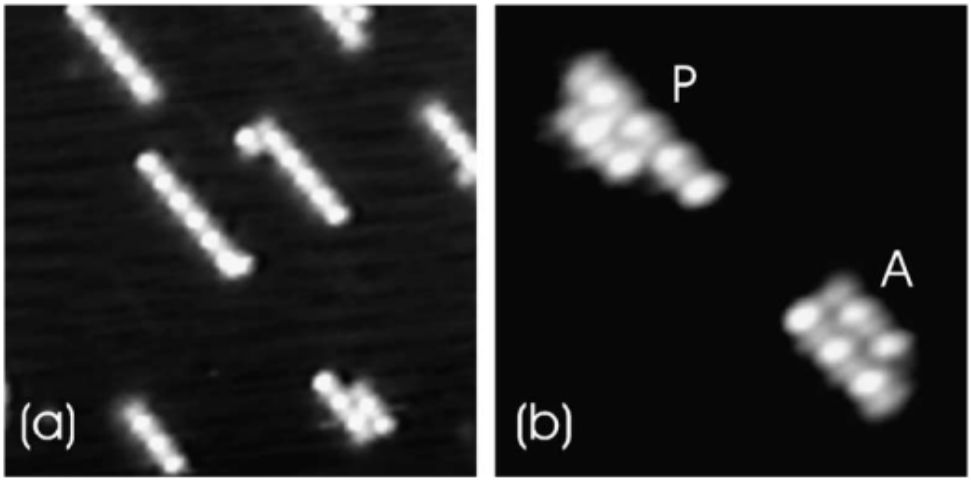


Figure 5.10 – Two STM images of di-L-alanine on Cu(110) at low coverage. The molecules were evaporated at a sample temperature of 248 K and scanning took place at 208 K to freeze out diffusion: (a) 160 Å × 160 Å , $V_1 = -2.10$ V, $I_1 = -0.34$ nA. (b) Two islands with parallel (P) or anti-parallel (A) di-L-alanine molecules in adjacent rows: 90 Å × 90 Å , $V_1 = -1.68$ V, $I_1 = -0.34$ nA. Reprinted from Surface Science, Volume 545, Issues 1–2, Ivan Stensgaard, Adsorption of di-L-alanine on Cu(110) investigated with scanning tunneling microscopy, Pages L747-L752, Copyright 2003, with permission from Elsevier.

this peptide, which is the smallest possible chiral peptide consisting of two AAs (L-alanine), on Cu (110) [313]. At low coverages, these molecules nucleate along the $[\bar{3}32]$ direction, forming small, predominantly one-dimensional islands. Coverage increase results in forming elongated, $[\bar{3}32]$-directed islands. At higher coverages, up to one monolayer, the islands merge to form phase barriers across domains with opposite orientations. In Fig. 5.10 and Fig. 5.11, we reproduce the experimental STM images from Ref.[313].

We investigated the adsorption of di-L-alanine on Cu(110) at DFT level of theory. In order to compare experimental and theoretical results we proceeded with comparing of the STM images obtained for the lowest energy structures obtained during the structure search. We analyzed the characteristics of the structures found by the random search, which one seems to be the experimental structure, and how it compares with the structure originally proposed in Ref. [313] and can be found in Fig 5.12(a).

5.7.1 Computational details

The electronic structure calculations were carried out using the numeric atom-centered orbital all-electron code FHI-aims [183, 184]. We used the standard *light* settings of FHI-aims for all species. For modeling the adsorbed molecules, a surface $1 \times 1 \times 2$ unit cell with $6 \times 6 \times 1$ k-point sampling was employed. The fcc(110) copper slab was produced using ASE package with lattice vectors directions $[\bar{3}32]$, $[\bar{1}1\bar{1}]$ and [110] that resulted in 4 layers in the slab with parameters a = 8.52 Å and b = 6.29 Å compared to experimental 8.48 Å and 6.29 Å lattice

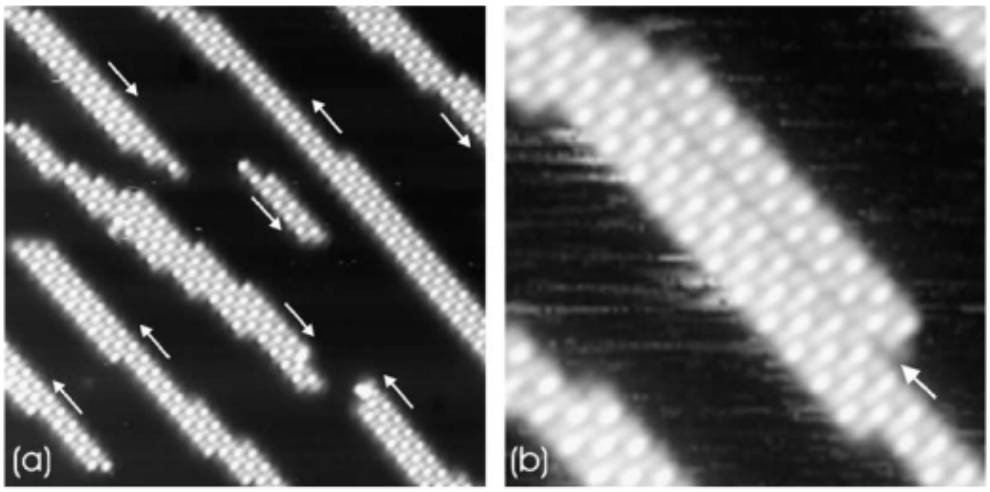


Figure 5.11 – (a) STM image of di-L-alanine on Cu(110). All molecules in an island are oriented parallel of antiparallel to the [$\bar{3}$32] direction as indicated by the two directions of the arrows. The di-L-alanine was evaporated at a sample temperature of 363 K and imaged at 198 K. Area: 250 Å $\times$ 250 Å , $V_1 = -1.25$ V, $I_1 = -0.65$ nA. (b) Formation of a domain boundary (marked with an arrow) between two antiparallel domains. Adsorption temperature: 363 K, imaged at 268 K, 100 Å $\times$ 100 Å , $V_1 = -1.68$ V, $I_1 = -1.52$ nA. Reprinted from Surface Science, Volume 545, Issues 1–2, Ivan Stensgaard, Adsorption of di-L-alanine on Cu(110) investigated with scanning tunneling microscopy, Pages L747-L752, Copyright 2003, with permission from Elsevier.

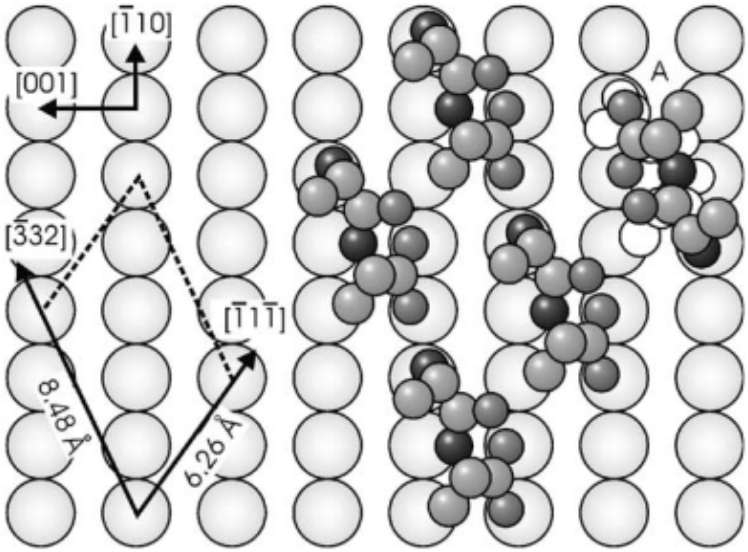


Figure 5.12 – Schematic model of the di-L-alanine surface layer on a Cu(110) substrate. The size and orientation of the unit cell is indicated. The atoms of the molecules are shown in shades of grey going from N (darkest) via O to C (lightest). Hydrogen atoms are left out. The molecule marked A in the upper right corner has been rotated by 180° and shifted slightly to adopt the same local adsorption geometry as the unrotated molecules. The position of the molecule before rotation is shown as an outline. Reprinted from Surface Science, Volume 545, Issues 1–2, Ivan Stensgaard, Adsorption of di-L-alanine on Cu(110) investigated with scanning tunneling microscopy, Pages L747-L752, Copyright 2003, with permission from Elsevier.

vectors lengths in [$\bar{3}32$], [$\bar{1}1\bar{1}$] respectively. The lattice parameter employed was 3.63 Å as in our previous works [83]. In order to isolate periodic images we added a 100 Å vacuum in the z direction and also employed the dipole correction. We employed the PBE+vdW^{surf} functional [130] which contains an effective screening of the vdW interactions optimized for metallic surfaces. The two bottom layers of the surface was constrained and a geometry optimization was performed until all forces in the system were below 0.01 eV/Å .

STM images were produced with Tersoff-Hamman approximation [197] with modelled applied voltage of -2 eV. This voltage was chosen based on the experimental values of the applied voltage for STM picture recording.

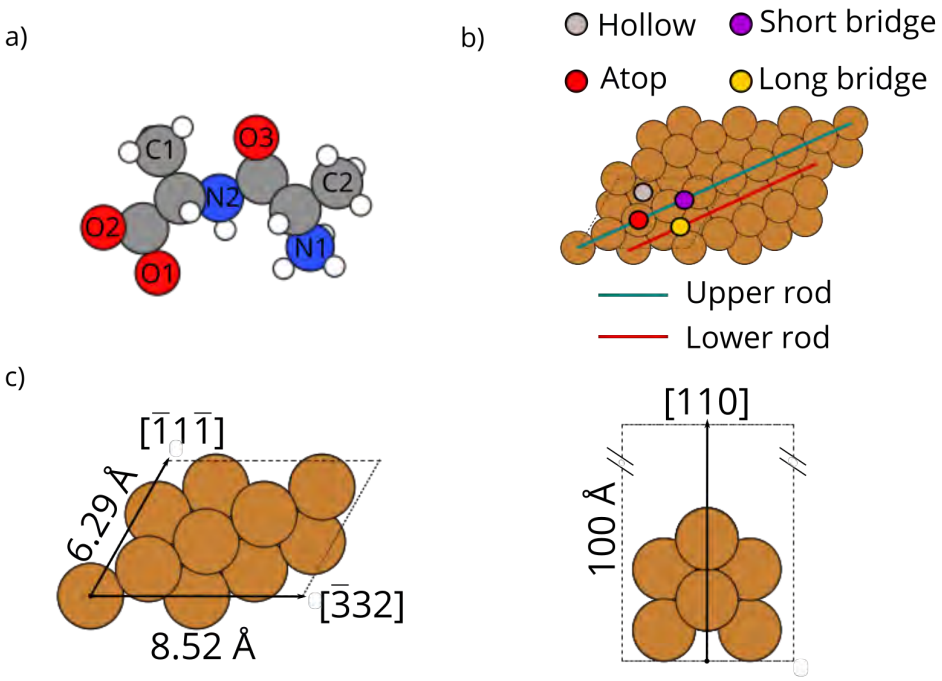


Figure 5.13 – a) Schematic representation of the di-L-alanine amino acid in its zwitterionic configuration. Red atoms are oxygen; blue atoms are nitrogen; white atoms are hydrogen, and grey atoms are carbon. b-d) Schematic representation of Cu(110).

An example of the di-L-alanine molecule and of the Cu(110) unit cell surface that we used for structure search can be found in Fig. 5.13.

5.7.2 Generation of trial structures

Trial structures were randomly produced using GenSec package with one molecule per unit cell. From the experimental study, we learned that we could apply a few constraints in the search. We restrict trial structures to be extended along the [$\bar{3}32$] direction. The structures were generated in zwitterionic state since Fig. 5.10(a) shows evidence that the molecules within a single-row island are aligned in the same direction at low coverage. This evidence points to a model in which the terminal carboxylic group of one molecule forms a hydrogen

bond with the terminal amino group of another molecule. The zwitterionic character of alanine in its solid-state [68], would be a good match for this type of relationship. However, it is impossible to rule out the possibility of deprotonation during the adsorption process, which would result in the formation of an anionic molecule. Investigations of tri-L-alanine for low coverage adsorption on Cu(110) revealed that the AA was bonding in the anionic form [314].

Ten searches were conducted in parallel, sharing the databases that blacklist trial candidates and geometry optimization trajectories obtained from different searches. Machinery implemented in GenSec allowed to perform such a structure search in a high-performance computing infrastructure by utilizing SQLite3 database features in ASE. After sampling of 500 structures we stop the structure search and select all the structures that fall within 1 eV energy range relative to the lowest energy structure and proceed to analysis of the results.

5.7.3 Analysis of the search

The structure that was proposed in Ref. [313] would bind with O1 and O2 oxygen atoms at the atop position to the same upper rod of the Cu(110) surface and atoms C1, N1 and N2 would adsorb also at atop positions on the neighbouring upper rod. Oxygen atom O3 and C2 from methyl group should not be connected to the surface. We manually prepared this structure and performed geometry optimization. This structure is depicted in Fig. 5.12 together with its STM image. As one can see, the patterns on STM images recorded experimentally and theoretically produced do no match: there are no interweaving bright and weak spots and their connectivity between neighbouring strands is absent.

After we performed a structure search only 23 unique structures in our database fall within 1 eV from the lowest energy structure. The structures either remain in the zwitterionic state or undergo deprotonation and adopt an anionic state. For all the 23 structures, we modelled STM images and created repeated images for easier visual comparison. One can clearly see that the patterns can differ considerably from each other. The particular pattern observed in the experiment (interweaving of bright and faded spots along a strand, with connections between strands that reminds of a tadpole) is very similar to the ones obtained for structure 7 (Fig. 5.14). All the lowest energy structures together with their STM images can be found in Appendix B.1-B.5 and we proceed to a more detailed analysis of the eight lowest energy structures found during the search (Fig. 5.14). The exact structure that was proposed in Ref. [313] was not found during the structure search. We prepared this structure manually and performed geometry optimization for it, which results in structure 8 (Fig. 5.14) but higher by 30 mEv in energy from it due to a slightly different adsorption pattern (Fig. 5.15). During geometry optimization C1 atom does not bind to the surface and points towards the vacuum region and thus, we can conclude that structure originally proposed in Ref. [313] is not stable.

The structures denoted 1,2,3 and 17 undergo deprotonation of the molecule adsorbed on the surface during geometry optimization. Most of the structures bind to the surface with at

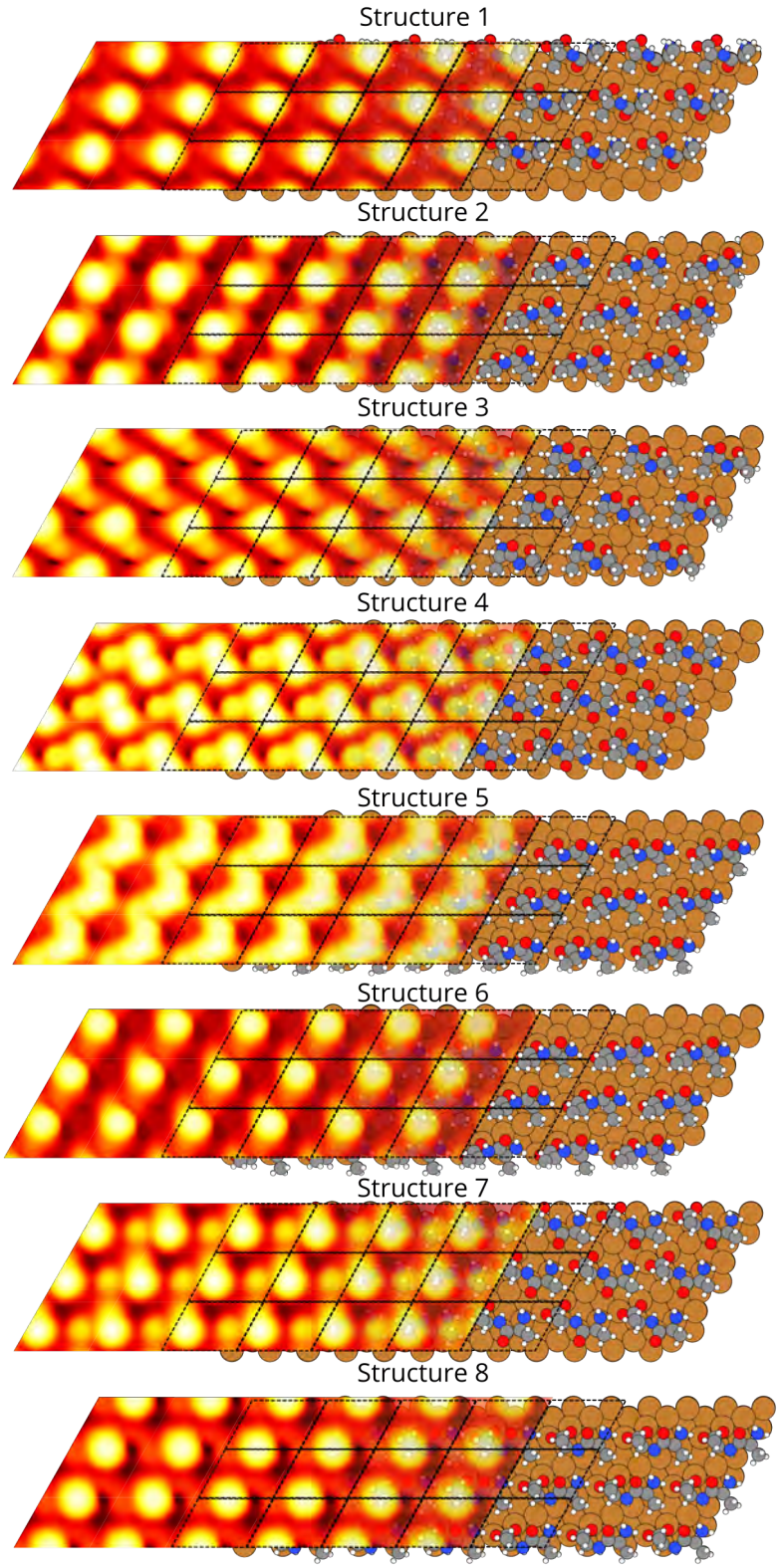


Figure 5.14 – Modelled STM images and structures 1-8 of di-L-alanine molecules adsorbed on Cu(110) surface together with unit cell represented with black dashed lines.

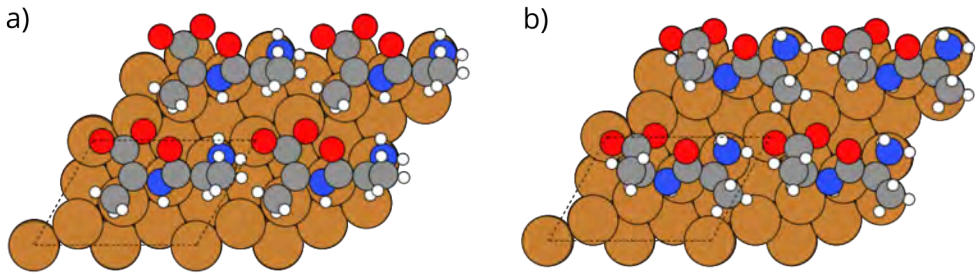


Figure 5.15 – proposed and relaxed structures.

least one oxygen atom from the carboxyl group and amino group attached to the different rods. The second oxygen atom from the carboxyl group can be attached to the same rod as the first oxygen atom (structures 1, 2, 4, 5, 6, 7, 8, 9, 11, 13, 16, 17, 23), to the same rod as an amino group (structures 3, 12, 14, 15, 19) or not attached to the surface. Almost in all cases the interstrand connectivity is done via carboxyl and amino groups, except for structures 3 and 17. Only in case of structure 22 both methyl groups are parallel to the surface - in this structure amino group is also not attached to the surface, in all other cases one or both methyl groups point towards the vacuum. We present in Fig. 5.14 a detailed visualization of structures 1, 2, 3, 7.

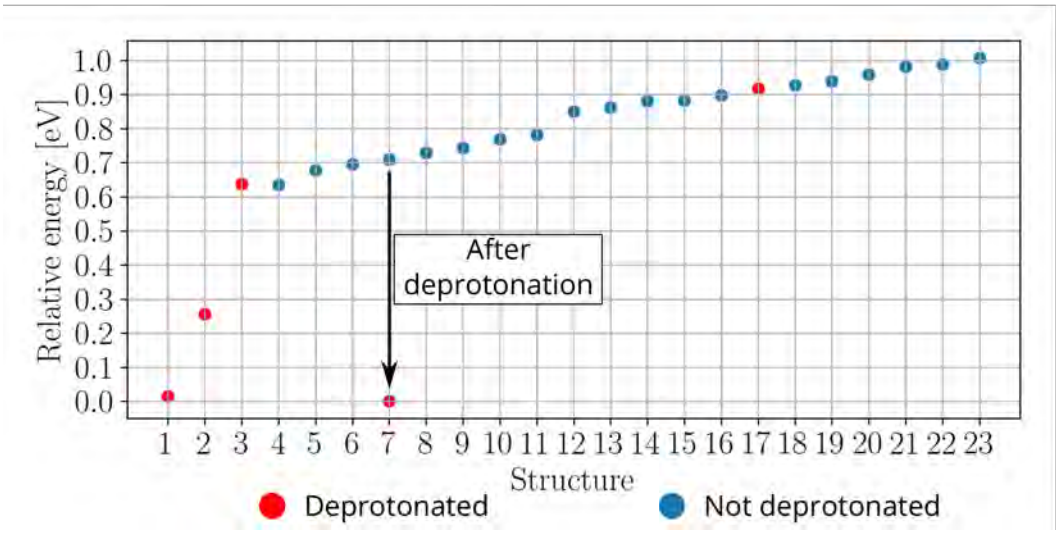


Figure 5.16 – Energy hierarchy of the obtained structures within 1 eV relative energy range.

Structure 7 binds with both O1 and O2 oxygens to the same rod of a surface at atop positions and N1 atom binds to another rod which is the same binding proposed in Ref. [313]. This structure differs from the one originally proposed by Stensgaard [313] by orientation of the N2 and O3 atoms and the orientation of the C1 and C2 atoms that do not interact with the surface in Structure 7. Moreover, the amino group interacts not only with carboxyl from the same strand, but also with O3 oxygen atom from another strand.

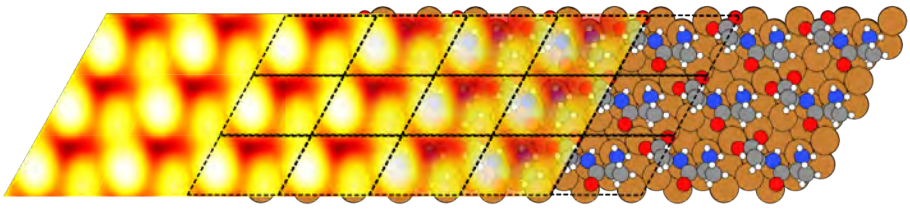


Figure 5.17 – Modelled STM image and structure of structure 7 after deprotonation together with unit cell represented with black dashed lines.

Since deprotonation is possible and the three lowest energy structures found in this search were deprotonated, we removed the hydrogen atom from the amino group that was pointing towards the surface and performed a new geometry optimization. The resulting structure is 10 mEv lower in energy than the lowest energy structure. The pattern from the modelled STM for this system is more pronounced and seems to agree even better with the experimental one, reproduced in Fig. 5.17.

Performing an analysis of the different stabilizing interactions of these self-assembled structures, shown in Appendix A.1, we find, interestingly, that the only structures for which the intrastrand and interstrand interactions are almost identical in energy are structures 7 and 11. These structures would likely fall into the exact same minimum if an optimization with a better basis set and accuracy threshold were performed.

The case of di-L-alanine on Cu(110) could be studied much further, but the results presented here show that a random search like the one performed here with GenSec can be a powerful ally of such STM experiments. Because it is based on first principles geometry optimizations, it also automatically identified the propensity for deprotonation of these molecules on such a reactive surface as Cu(110). This is an important point to consider when dealing with other types of theoretical approaches such as FF and schemes that keep molecules "rigid" or "whole".

We conclude that the best candidate theoretically predicted structure neither one that was proposed in the paper [313] nor the lowest energy structure found during the structure search. This supports the idea, that in a particular experiment the global minimum found theoretically may not always the most relevant structure. A random structure search strategy that covers the broadest possible parts of the conformational space (within a few constraints) can be quite effective.

One of the intriguing and yet not completely understood results is that the best candidate structure stands out among other lowest energy structures by having the perfect balance between intrastrand and interstrand interaction energies that could be relevant for the understanding of the self-assembly processes on surfaces.

The final result of comparing of simulated and experimental STM images can be found in Fig. 5.18.

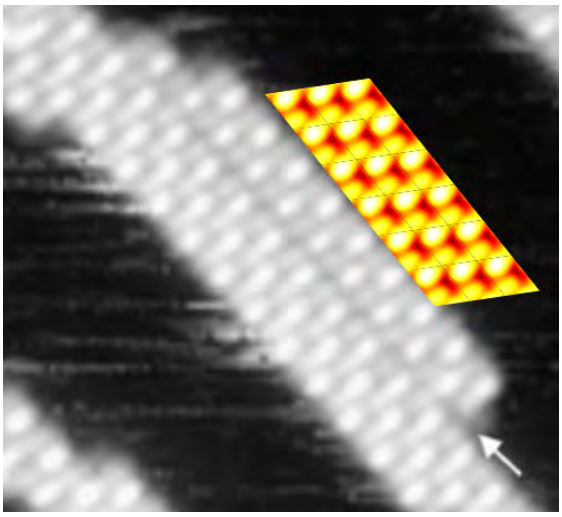


Figure 5.18 – Modelled STM image colored in oranges and experimental STM image colored in greys of di-L-alanine on Cu(110) aligned in direction of strand grow. Reprinted from Surface Science, Volume 545, Issues 1–2, Ivan Stensgaard, Adsorption of di-L-alanine on Cu(110) investigated with scanning tunneling microscopy, Pages L747-L752, Copyright 2003, with permission from Elsevier.

5.8 Conclusions

GenSec showed satisfactory results for structure search of di-L-alanine adsorbed on Cu(110) surface in both efficient utilization of resources (multiple structure searches were carried out at the same time connected to the shared database) and for an unbiased sampling of the conformational space of flexible molecule. The created infrastructure allows easily specify constraints of the search according to the experimental input and choose from multiple electronic structure codes available to connect through ASE package. Created databases contain not only the lowest energy structure but also all the intermediate steps together with energies and forces in a unified format that is convenient to share and reuse.

5.9 Outlook

The workflow of the package already allows an investigation of an arbitrary amount of adsorbates per unit cell with respect to specified surroundings. One of the package's strengths is that it can produce data in a parallel fashion, optimally utilizing available computational resources. The resulting data has general formatting independent of the electronic structure package used for structure search, making it reusable and easy to handle for further processing.

The main directions for further development of the package should be:

- Connection of the package to the ML packages that allow training cheap potentials on the fly for further exhaustive search procedure;
- Connection to the packages that allow to automatically generate low dimension representation of the conformational spaces and visualize them

Many people asked me what would I
do if I didn't finish the thesis.
We will never know it.

6 Conclusions

In this thesis, we have characterized the conformational space of the arginine amino acid in its neutral and protonated form in different non-biological environments, i.e. in isolation and in contact with metallic surfaces. In particular, we have analyzed how and why different parts of the conformational space become accessible or are excluded, depending on the protonation state and the environment, showing the importance of bond formation and charge rearrangement in these systems.

This study included the construction of a database based on thousands of structures optimized by density-functional theory including dispersion interactions. The construction of this database is a result in itself and we hope that in future systems, that are investigated will be also available for everyone. The analysis of complex systems is still far from being fully automated, and requires a human's creative approach and a tremendous amount of effort and time to be invested in the identification of structure-property relationships. Even the application of modern dimensionality reduction and visualization techniques should be considered only as a first step that can give inspiration for further analysis. Regarding the investigation of interface systems, we found, for example, that it is advantageous to start from a comprehensive sampling of the conformational space of the least-constrained molecular form, which in our case was the neutral Arg amino acid in the gas-phase. This is evidenced by the fact that in our low-dimensional projections, all low-energy conformers we observe on the surfaces for both Arg and Arg-H^+, lie among structural conformations that were already present in the gas-phase sampling of Arg, albeit often with high relative energies. This is not the general case though, and the environment can alter conformational

space in an unpredictable manner - this can be seen from the sampling of Arg-H^+ structures, the flexibility of which increased after adsorption. In addition, we find that for Cu, Ag, and Au surfaces, the energy hierarchies of different conformers are largely preserved when changing the substrate.

We illustrate that while INTERFACE-FF can sample relevant areas of conformational space, it is not able to capture consistent energy hierarchies. Additionally, the molecular chemical groups show a preference to adsorb on different surface sites, which could have considerable impact on self-assembly studies. Databases such as those we created will serve as an important source of data for further parametrization and improvement of these potentials.

Regarding the structural space of Arg and Arg-H^+ adsorbed on (111) surfaces of Cu, Ag and Au, we have learned the following: The adsorption of Arg leads to the formation of strong bonds to the surface that involve mostly the carboxyl and amino groups. This stabilizes the protomer that we label **P3** in this work, where the carboxyl group is deprotonated and the side chain is protonated. This is different to the dominant protomer in the gas phase, with the label **P1**. The bonds to the surface sterically constrain the conformations of this molecule, thus decreasing the number of observed structures with respect to the numbers observed in the gas phase. When adsorbed, Arg donates electrons to the surface, becoming slightly positively charged. We do not observe fully extended structures lying on the surface, and most conformers exhibit intramolecular H-bonds. The majority of conformers of Arg in the low-energy region adsorb with the $C_\alpha H$ chiral center pointing the hydrogen atom away from the surfaces.

Arginine in its protonated form, i.e. Arg-H^+, is the most abundant form of this amino-acid in biological environments, where it typically adopts the zwitterionic protomer **P7**. In the gas-phase, we observe that the non-zwitterionic state **P6** is dominant and that the addition of a proton decreases the number of allowed conformations with respect to isolated Arg due to the added electrostatic interactions, and the neutralization of the carboxyl group that would otherwise be involved in intramolecular H-bonds. Upon adsorption to metallic surfaces, we observe that the protomer **P6** is still dominant and that there are no strong bonds formed to the surface. In addition, this molecule receives electrons from the surface, thus becoming less positively charged. Both effects conspire to yield a homogeneous (flat) molecule-surface interaction, and a relatively high population of different structures in the low-energy range. Contrary to Arg, most low-energy conformers of Arg-H^+ adsorb with the C_α-H chiral center pointing the hydrogen atom towards to the surfaces. Finally, through the calculation of dissociation energies, we also conclude that the deprotonation of Arg-H^+ is energetically favorable only on Cu(111).

Our observations regarding the preferred protomers and deprotonation propensities discussed above are consistent with the observations in the literature that the adsorption of amino acids in their anionic and deprotonated form is common on reactive metals like Cu(111) [61]. One pronounced difference that we find among surfaces is the average adsorp-

tion height of the molecules: They follow the trend Cu(111) < Ag(111) < Au(111), and Arg is always closer than Arg-H^+ to the same respective surface.

The set of electronic-structure calculations presented here show that a flexible amino-acid like Arginine presents a rich conformational space involving different protomeric states and molecule orientations with respect to the surface, allied to a complex charge rearrangement. Going forward, it is clear that the likes of this study based solely on DFT cannot become a routine method due to the elevated computational cost. Addressing the whole breadth of amino acids as well as self assembly of these structures on surfaces will profit from this study as a benchmark and a means to develop models, possibly based on different machine-learning techniques that can bypass the cost of thousands of DFT structure optimizations. Unfortunately, to the best of our knowledge there is still no experimental results available for Arg and Arg-H^+ adsorbed on coinage metals are available.

With the approaching technology of exascale computing, we need to develop software that can efficiently utilize the available computational resources. We developed the GenSec package as a step further in automatising the kind of structure search described above. This can help reduce the effort required to carry out these kinds of investigations, and opens the path for routinely perform high-throughput calculations of interface systems, and also for modelling of self-assemblies formed on inorganic substrates. Many tasks that previously required manually setting parameters are automated in the package, such as the identification of the internal degrees of freedom of the flexible part of the interface. By setting up periodic boundary conditions, the package can produce arbitrary amount of molecules per unit cell that are obtained from the template. The construction of the database in a standardized form will make it possible to efficiently share data between researchers using modern material science repositories, facilitating the general understanding of the processes at the atomic level. Data produced with GenSec is suitable for parametrizing FFs and applying to machine learning methods that would allow the investigation of thermodynamical properties of systems and carry out calculations at longer time scales. The geometry optimization schemes together with their preconditioning schemes increase the efficiency for the most important part in the database generation procedure. The random search strategy implemented in the GenSec can be seen as robust foundation for other global search techniques such as evolutionary algorithms or Bayesian optimization methods, that rely on random generation to some extent. Further development would involve automated schemes for producing low dimensional representations using the procedures used in this thesis. The preconditioning schemes described in the thesis should be tested on the wide range of the different systems, leading to further optimisation of these techniques and increasing the efficiency of databases generation.

As a result of the efficient utilization of resources (multiple structure searches were carried out at the same time connected to the shared database) and the unbiased sampling of the conformational space of a flexible molecule, GenSec provided satisfactory results for the structure search for di-L-alanine adsorbed on Cu(110) surface. As a result of the newly devel-

oped infrastructure, it is now possible to establish search constraints based on experimental input and choose from a large number of electronic structure codes that are available to connect through the ASE package. The databases that have been created contain the lowest energy structure and all of the intermediate steps and their energies and forces, all in a single format that is easy to share and reuse.

A

Additional information on Arg and Arg-H^+ on metallic surfaces

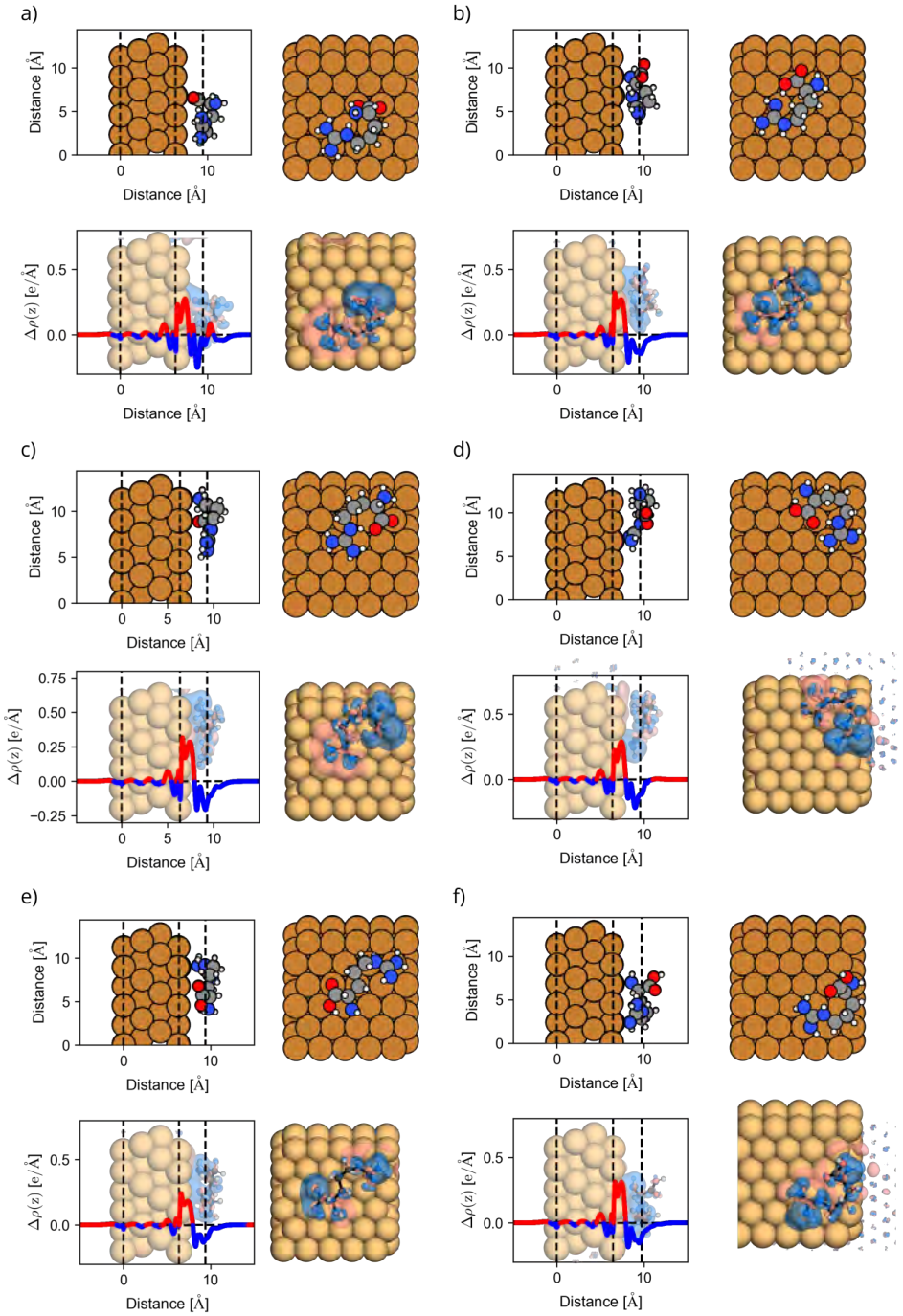


Figure A.1 – Side and top views of the adsorbed structures of Arg on Cu(111). Dashed black lines correspond to: average z position of the atoms in the lowest layer of the surface (left), average z position of atoms in the highest layer of the surface (middle), centre of the mass of the molecule (right). Red/blue solid lines (and also red/blue regions) correspond to the electron density accumulation/depletion.

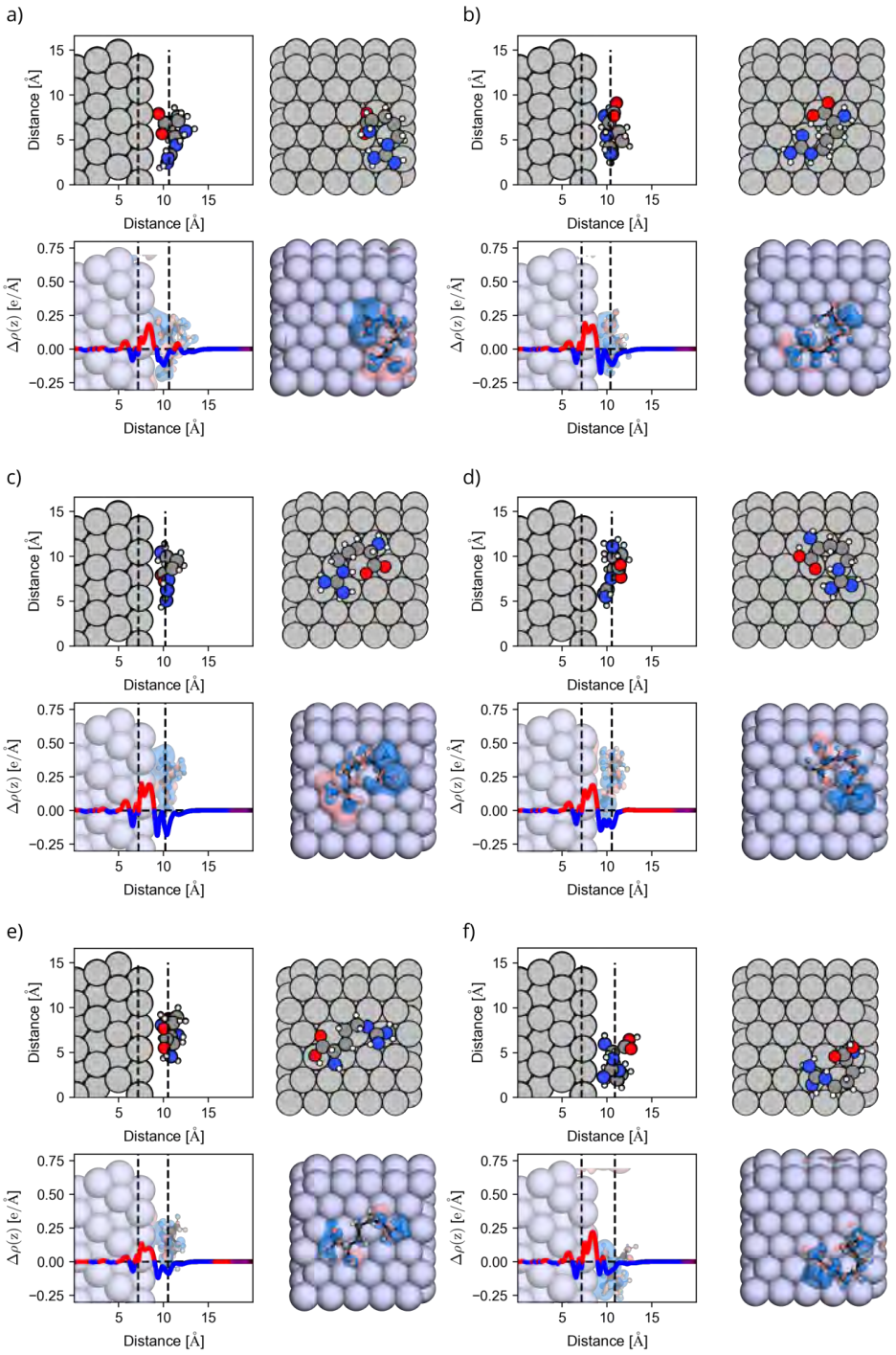


Figure A.2 – Side and top views of the adsorbed structures of Arg on Ag(111). Dashed black lines correspond to: average z position of the atoms in the lowest layer of the surface (left), average z position of atoms in the highest layer of the surface (middle), centre of the mass of the molecule (right). Red/blue solid lines (and also red/blue regions) correspond to the electron density accumulation/depletion.

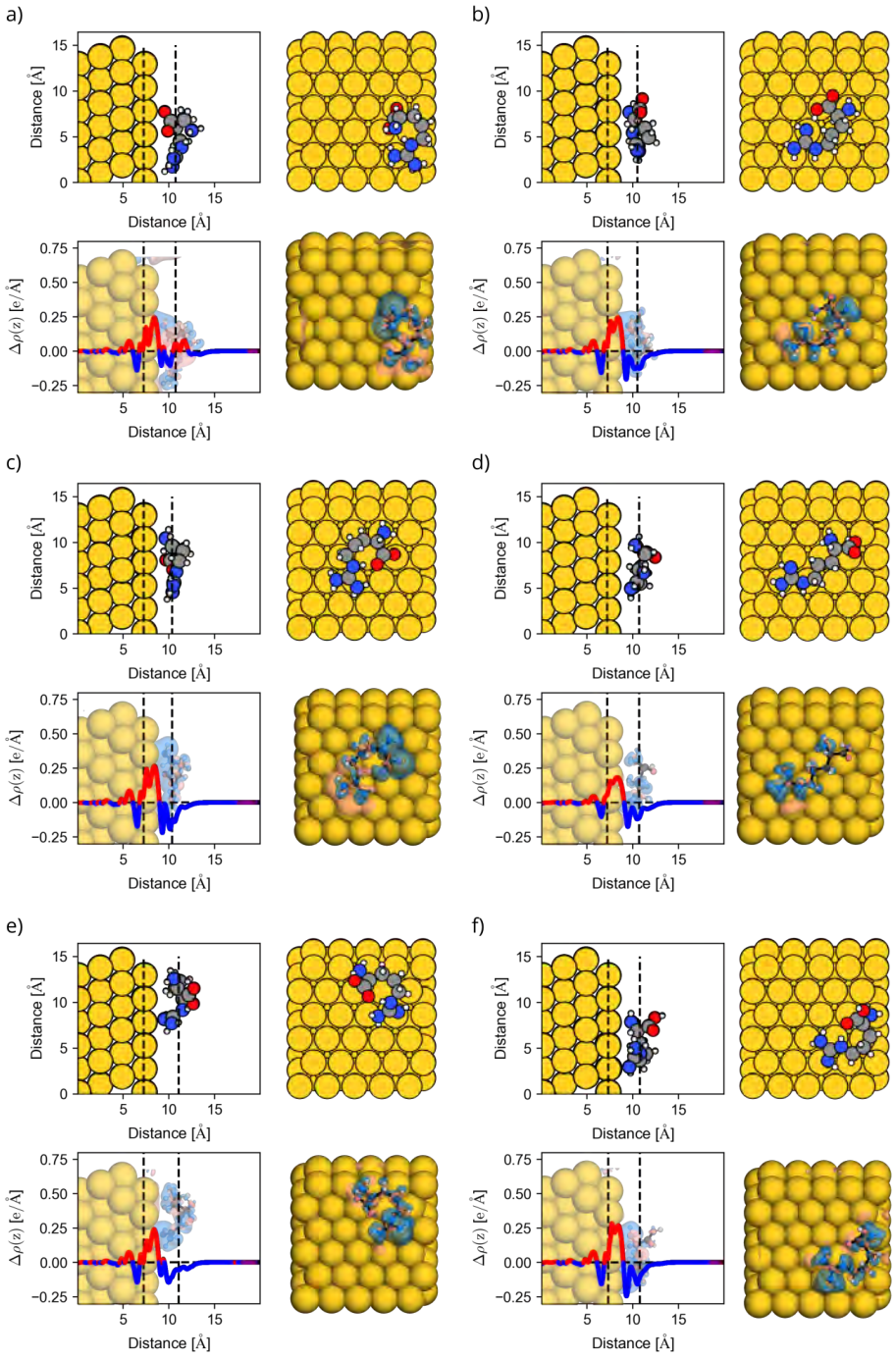


Figure A.3 – Side and top views of the adsorbed structures of Arg on Au(111). Dashed black lines correspond to: average z position of the atoms in the lowest layer of the surface (left), average z position of atoms in the highest layer of the surface (middle), centre of the mass of the molecule (right). Red/blue solid lines (and also red/blue regions) correspond to the electron density accumulation/depletion.

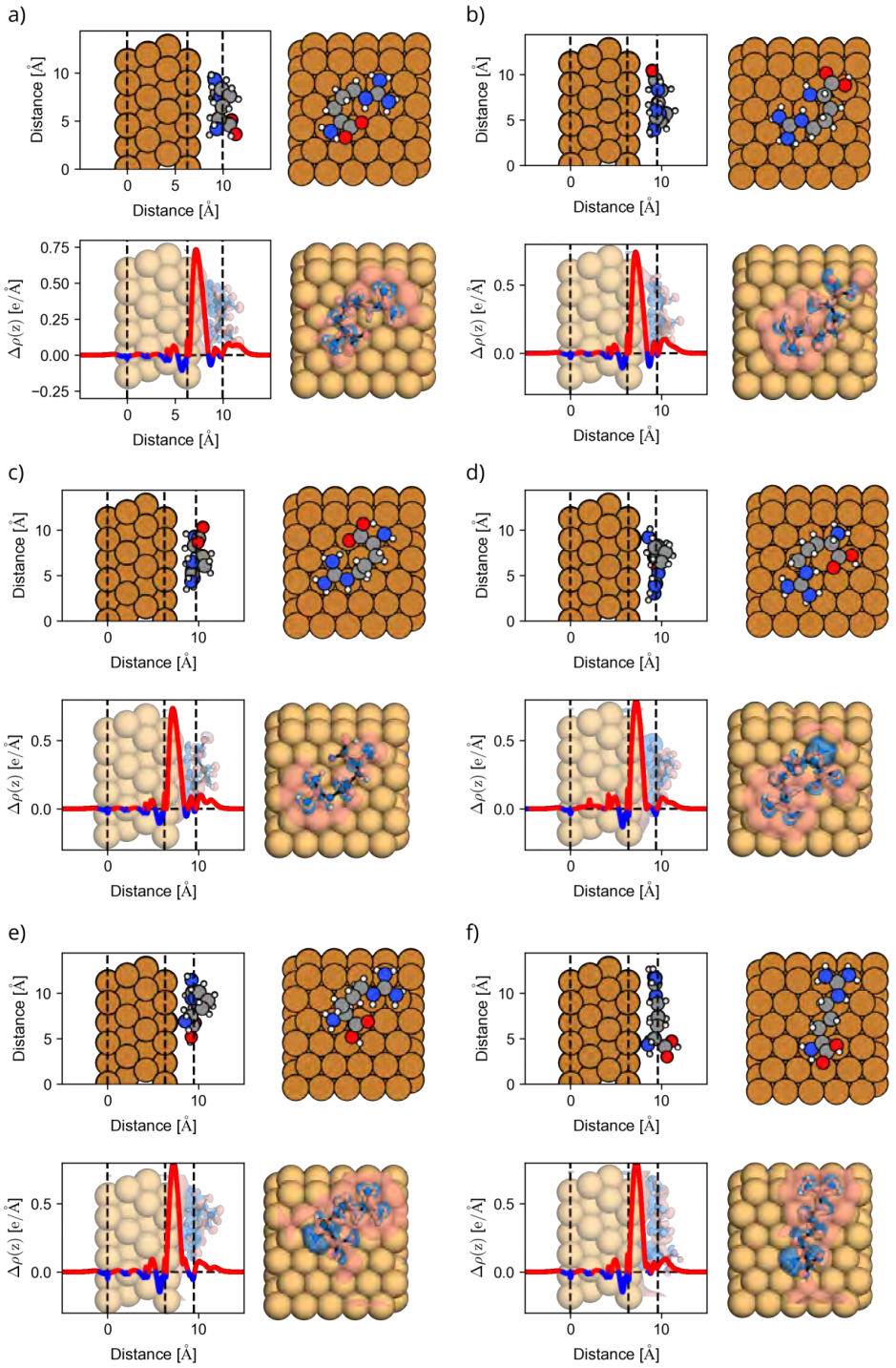


Figure A.4 – Side and top views of the adsorbed structures of Arg-H^+ on Cu(111). Dashed black lines correspond to: average z position of the atoms in the lowest layer of the surface (left), average z position of atoms in the highest layer of the surface (middle), centre of the mass of the molecule (right). Red/blue solid lines (and also red/blue regions) correspond to the electron density accumulation/depletion.

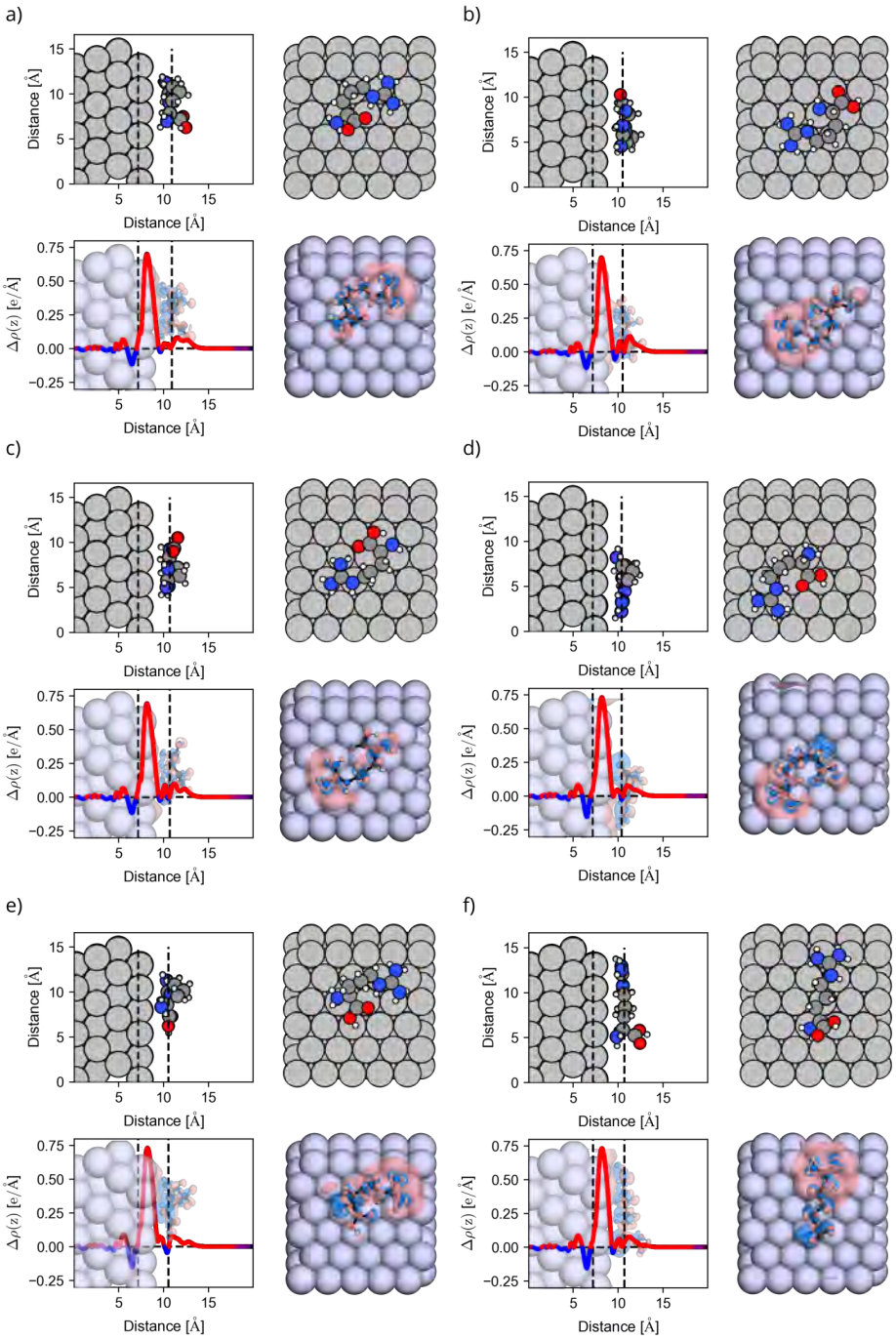


Figure A.5 – Side and top views of the adsorbed structures of Arg-H$^+$ on Ag(111). Dashed black lines correspond to: average z position of the atoms in the lowest layer of the surface (left), average z position of atoms in the highest layer of the surface (middle), centre of the mass of the molecule (right). Red/blue solid lines (and also red/blue regions) correspond to the electron density accumulation/depletion.

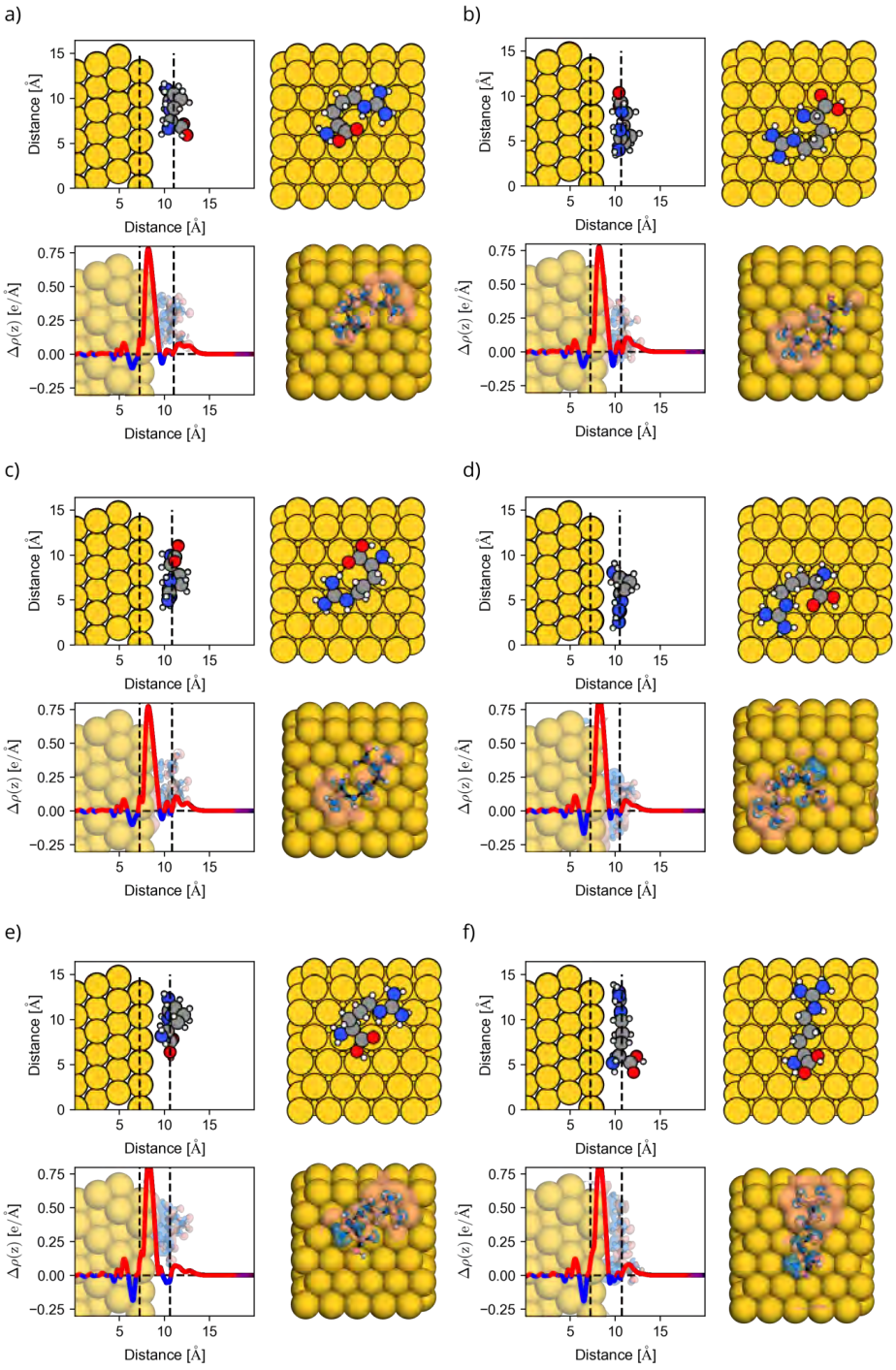


Figure A.6 – Side and top views of the adsorbed structures of Arg-H^+ on Au(111). Dashed black lines correspond to: average z position of the atoms in the lowest layer of the surface (left), average z position of atoms in the highest layer of the surface (middle), centre of the mass of the molecule (right). Red/blue solid lines (and also red/blue regions) correspond to the electron density accumulation/depletion.

B

Additional information on di-L-alanine molecule on Cu(110)

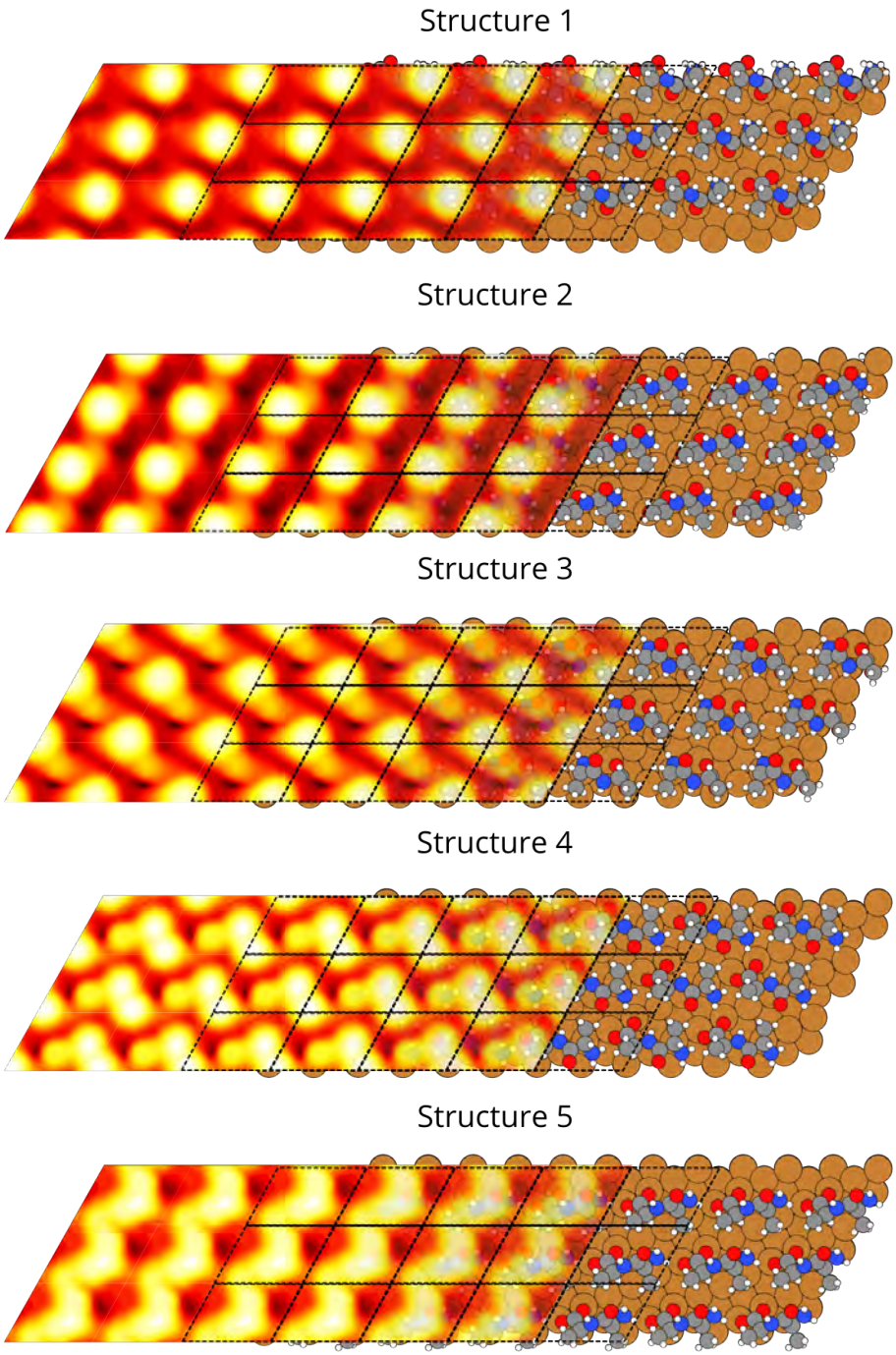


Figure B.1 – Modelled STM images and structures 1-5 of di-L-alanine molecules adsorbed on Cu(110) surface together with unit cell represented with black dashed lines.

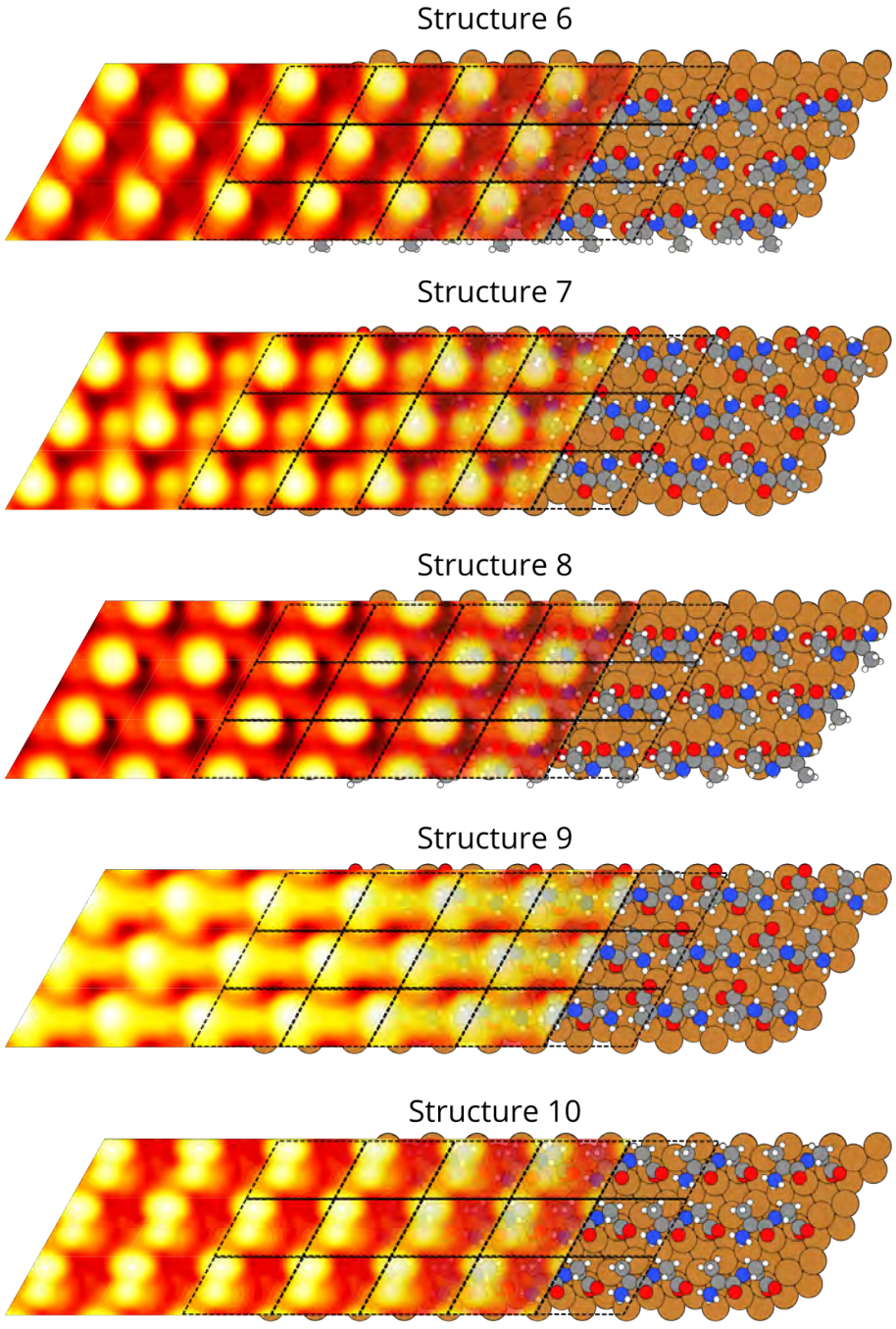


Figure B.2 – Modelled STM images and structures 6-10 of di-L-alanine molecules adsorbed on Cu(110) surface together with unit cell represented with black dashed lines.

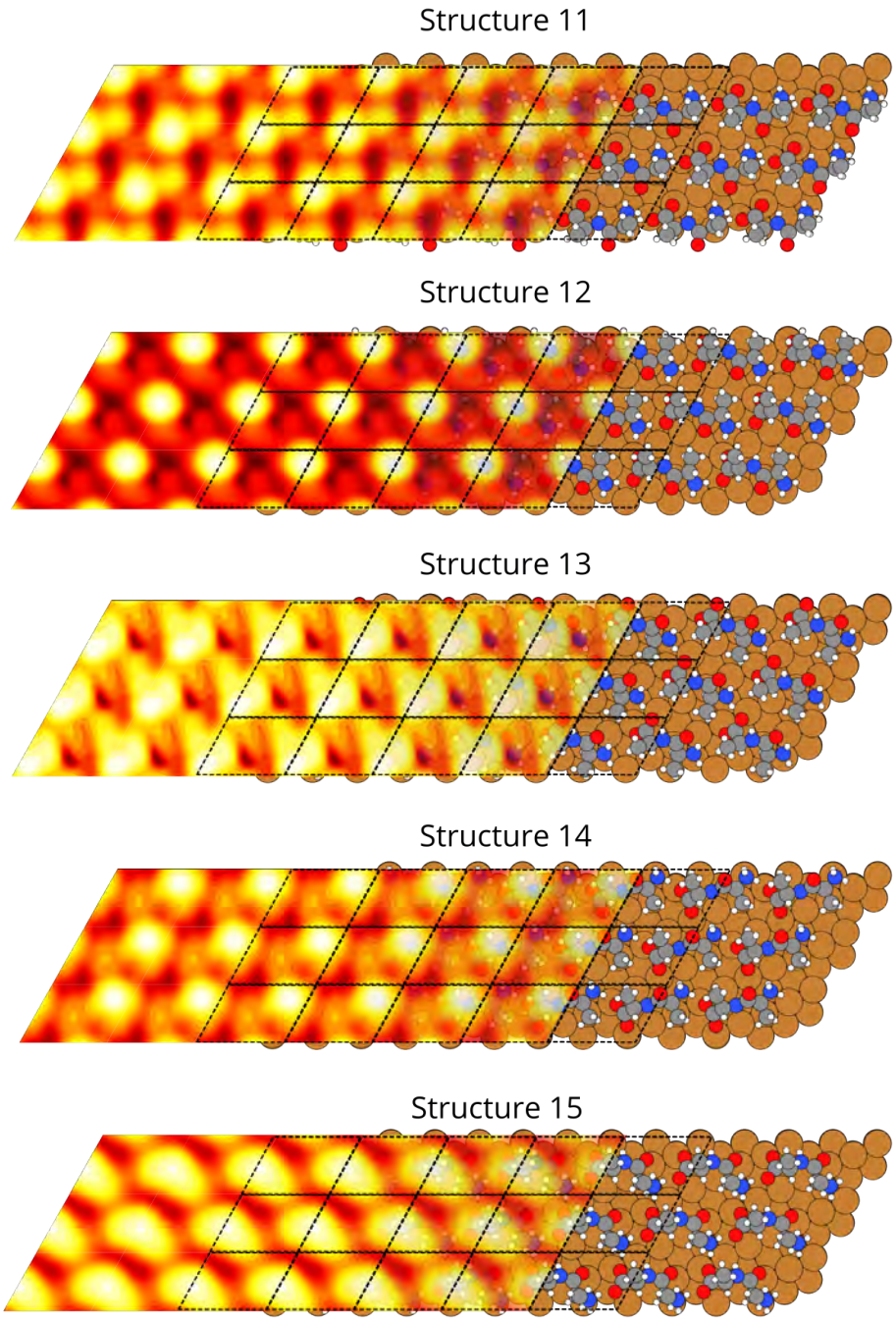


Figure B.3 – Modelled STM images and structures 11-15 of di-L-alanine molecules adsorbed on Cu(110) surface together with unit cell represented with black dashed lines.

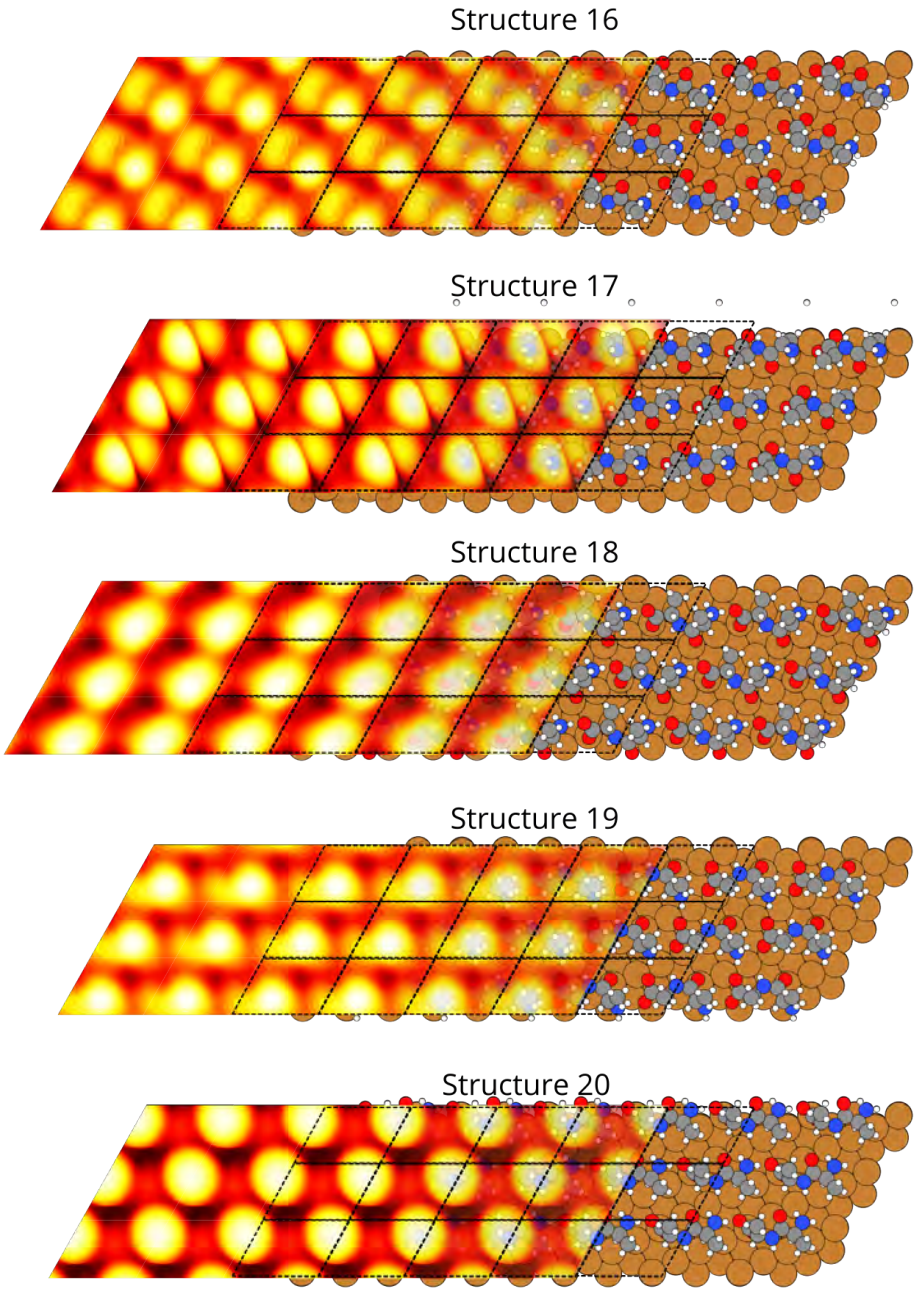


Figure B.4 – Modelled STM images and structures 16-20 of di-L-alanine molecules adsorbed on Cu(110) surface together with unit cell represented with black dashed lines.

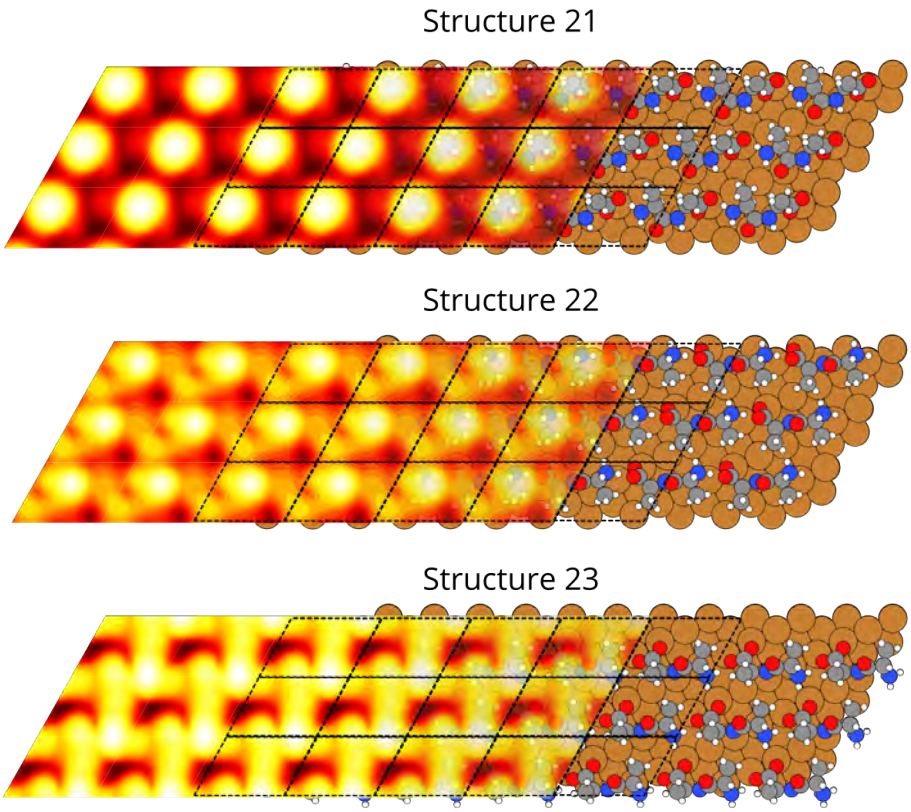


Figure B.5 – Modelled STM images and structures 21-23 of di-L-alanine molecules adsorbed on Cu(110) surface together with unit cell represented with black dashed lines.

A

Estimation of stabilizing interactions for di-L-alanine on Cu(110)

In order to address the question about molecular inter- and intrastrand interactions we needed to create bigger systems that would isolate only one strand that would allow to compare it to the fully periodic system, where molecule-molecule and molecule-surface interactions could be separated. For that the 6×2 unit cell slab system was prepared on which the different length of one strand will be calculated. For one unit cell sized systems and for bigger systems we applied $10 \times 10 \times 1$ and $1 \times 5 \times 1$ k-point sampling correspondingly. Now we would like to introduce some notations:

E - Energy
EB - Binding energy
S_1 - fully periodic single unit cell system
S_2^n - isolated system with n molecules on large surface

The binding energy can be calculated as follows:

$$EB_{S_1} = E_{S_1} - E_{\text{mol}} - E_{\text{surf}} \tag{A.1}$$

where E_{mol} is the energy of the isolated molecule in the same configuration and E_{surf} - energy of the small surface taken from S_1. This binding energy EB_{S_1} contains all the contributions between molecules and surfaces, molecular intrastrand interactions (within the strand) and

interstrand (between strands).

$$EB^{n}_{S_2} = E^{n}_{S_2} - nE_{\text{mol}} - E_{\text{surf}}, \tag{A.2}$$

where $EB^{1}_{S_2}$ is just molecules-surface interaction that can be subtracted in order to isolate molecule-molecule interactions. Now we create incremental function that calculates increase of molecule-molecule interaction with increasing of amount of molecules:

$$\Delta EB(n) = (E^{n}_{S_2} - E^{n-1}_{S_2}) - E_{\text{mol}}. \tag{A.3}$$

After that we need to subtract the molecule-surface interaction and we get the energy gain when we add one molecule to the strand:

$$E_{\text{interstrand}}(n) = \Delta EB(n) - EB^{1}_{S_2}. \tag{A.4}$$

Finally, the molecule-molecule interaction is also given by:

$$EB_{S_1} - EB^{1}_{S_2} = E_{\text{mol-mol}} = E_{\text{interstrand}}(\infty) + E_{\text{intrastrand}}(\infty), \tag{A.5}$$

using which one can calculate convergence with large n.

For all the structures that are not deprotonated on surface with the formulas obtained above we decompose the interactions and the results can be found in Fig. A.1. The results should be further processed in order to draw more precise conclusions, but what one can immediately see is that the only structure for which the intrastrand and interstrand interactions are almost identical in energy are structures 7 and 11 that are very similar (probably will fall to the same local minima if we perform geometry optimization with tighter settings).

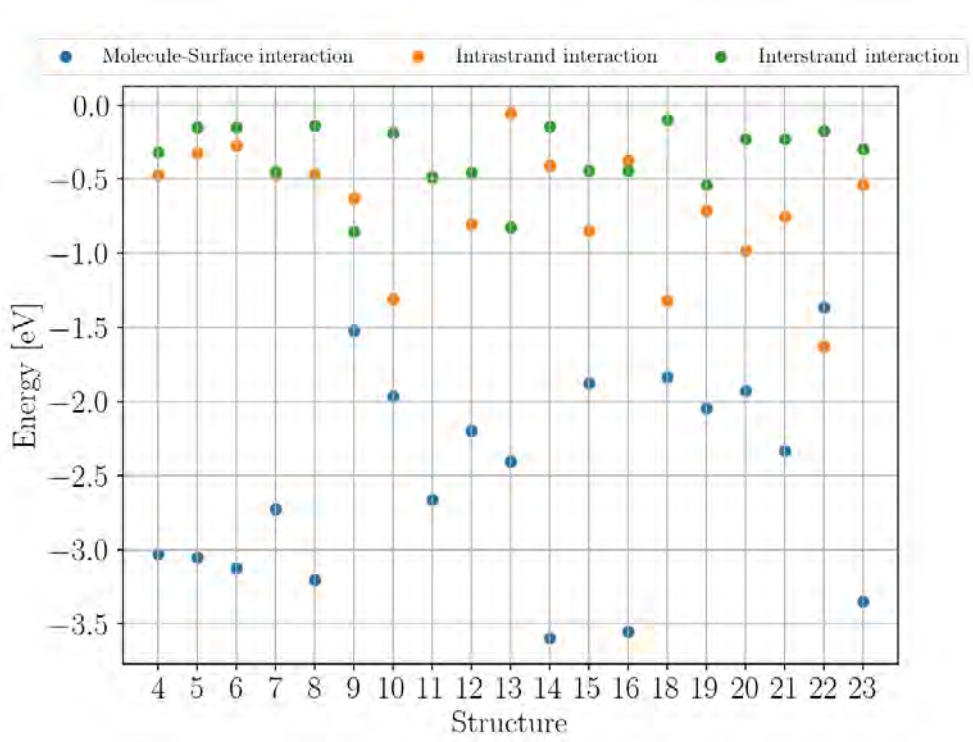


Figure A.1 – Molecule-surface, intrastrand and inerstrand interactions for the lowest energy structures of di-L-alanine adsorbed on Cu(110) surface

Bibliography

[1] Tiffany R Walsh and Marc R Knecht. Biointerface Structural Effects on the Properties and Applications of Bioinspired Peptide-Based Nanomaterials. *Chemical Reviews*, 117(20):12641–12704, 2017.

[2] 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. *Scientific Data*, 3:160009, 2016.

[3] Ludwig Bartels. Tailoring molecular layers at metal surfaces. *Nature Chemistry*, 2010.

[4] Bin Yang, Dave J. Adams, Maria Marlow, and Mischa Zelzer. Surface-mediated supramolecular self-assembly of protein, peptide, and nucleoside derivatives: From surface design to the underlying mechanism and tailored functions. *Langmuir*, 34(50):15109–15125, 2018. PMID: 30032622.

[5] E. J. Wood. Fundamentals of biochemistry: Life at the molecular level (Third Edition) by D. Voet, J. Voet, and C. W. Pratt. *Biochemistry and Molecular Biology Education*, 2008.

[6] Philip J. Cowen and Michael Browning. What has serotonin to do with depression? *World Psychiatry*, 2015.

[7] Emanuela Gatto, Lorenzo Stella, Fernando Formaggio, Claudio Toniolo, Leandro Lorenzelli, and Mariano Venanzi. Electroconductive and photocurrent generation properties of self-assembled monolayers formed by functionalized, conformationally-constrained peptides on gold electrodes. *Journal of Peptide Science*, 2008.

[8] Emanuela Gatto, Alessia Quatela, Mario Caruso, Roberto Tagliaferro, Marta De Zotti, Fernando Formaggio, Claudio Toniolo, Aldo Di Carlo, and Mariano Venanzi. Mimicking nature: A novel peptide-based bio-inspired approach for solar energy conversion. *ChemPhysChem*, 2014.

[9] Emanuela Gatto, Mario Caruso, Alessandro Porchetta, Claudio Toniolo, Fernando Formaggio, Marco Crisma, and Mariano Venanzi. Photocurrent generation through peptide-based self-assembled monolayers on a gold surface: Antenna and junction effects. *Journal of Peptide Science*, 2011.

[10] Shiro Yasutomi, Tomoyuki Morita, Yukio Imanishi, and Shunsaku Kimura. A molecular photodiode system that can switch photocurrent direction. *Science*, 2004.

[11] Riming Nie, Aiyuan Li, and Xianyu Deng. Environmentally friendly biomaterials as an interfacial layer for highly efficient and air-stable inverted organic solar cells. *Journal of Materials Chemistry A*, 2014.

[12] Maayan Matmor and Nurit Ashkenasy. Modulating semiconductor surface electronic properties by inorganic peptide-binders sequence design. *Journal of the American Chemical Society*, 2012.

[13] Maayan Matmor, George A. Lengyel, W. Seth Horne, and Nurit Ashkenasy. Peptide-functionalized semiconductor surfaces: Strong surface electronic effects from minor alterations to backbone composition. *Physical Chemistry Chemical Physics*, 2017.

[14] Xiaoyin Xiao, Bingqian Xu, and Nongjian Tao. Conductance Titration of Single-Peptide Molecules. *Journal of the American Chemical Society*, 2004.

[15] Lior Sepunaru, Sivan Refaely-Abramson, Robert Lovrinčić, Yulian Gavrilov, Piyush Agrawal, Yaakov Levy, Leeor Kronik, Israel Pecht, Mordechai Sheves, and David Cahen. Electronic transport via homopeptides: The role of side chains and secondary structure. *Journal of the American Chemical Society*, 2015.

[16] Cunlan Guo, Xi Yu, Sivan Refaely-Abramson, Lior Sepunaru, Tatyana Bendikov, Israel Pecht, Leeor Kronik, Ayelet Vilan, Mordechai Sheves, and David Cahen. Tuning electronic transport via hepta-alanine peptides junction by tryptophan doping. *Proceedings of the National Academy of Sciences of the United States of America*, 2016.

[17] Y. Pennec, W. Auwärter, A. Schiffrin, A. Weber-Bargioni, A. Riemann, and J. V. Barth. Supramolecular gratings for tuneable confinement of electrons on metal surfaces. *Nature Nanotechnology*, 2007.

[18] Ken Kanazawa, Atsushi Taninaka, Hui Huang, Munenori Nishimura, Shoji Yoshida, Osamu Takeuchi, and Hidemi Shigekawa. Scanning tunneling microscopy/spectroscopy on self-assembly of a glycine/cu(111) nanocavity array. *Chem. Commun.*, 47:11312–11314, 2011.

[19] Clara Shlizerman, Alexander Atanassov, Inbal Berkovich, Gonen Ashkenasy, and Nurit Ashkenasy. De novo designed coiled-coil proteins with variable conformations as components of molecular electronic devices. *Journal of the American Chemical Society*, 2010.

[20] Shengfu Chen, Zhiqiang Cao, and Shaoyi Jiang. Ultra-low fouling peptide surfaces derived from natural amino acids. *Biomaterials*, 2009.

[21] Sivan Nir and Meital Reches. Bio-inspired antifouling approaches: The quest towards non-toxic and non-biocidal materials. *Current Opinion in Biotechnology*, 2016.

[22] Mehmet Sarikaya, Candan Tamerler, Alex K.Y. Jen, Klaus Schulten, and François Baneyx. Molecular biomimetics: nanotechnology through biology. *Nature Materials*, 2(9):577–585, 2003.

[23] Carlos Mas-Moruno, Roberta Fraioli, Fernando Albericio, José María Manero, and F. Javier Gil. Novel peptide-based platform for the dual presentation of biologically active peptide motifs on biomaterials. *ACS Applied Materials and Interfaces*, 2014.

[24] Mareen Pagel, Rayk Hassert, Torsten John, Klaus Braun, Manfred Wießler, Bernd Abel, and Annette G. Beck-Sickinger. Multifunctional Coating Improves Cell Adhesion on Titanium by using Cooperatively Acting Peptides. *Angewandte Chemie - International Edition*, 2016.

[25] S.M. Barlow and R. Raval. Complex organic molecules at metal surfaces: bonding, organisation and chirality. *Surface Science Reports*, 50(6):201–341, 2003.

[26] Magalí Lingenfelder, Giulia Tomba, Giovanni Costantini, Lucio Colombi Ciacchi, Alessandro De Vita, and Klaus Kern. Tracking the chiral recognition of adsorbed dipeptides at the single-molecule level. *Angewandte Chemie - International Edition*, 2007.

[27] Magalí Lingenfelder. *Chiral recognition and supramolecular self-assembly of adsorbed amino acids and dipeptides at the submolecular level.* PhD thesis, École Polytechnique Fédèrale de Lausanne, 2008.

[28] Dmitriy Khatayevich, Christopher R. So, Yuhei Hayamizu, Carolyn Gresswell, and Mehmet Sarikaya. Controlling the surface chemistry of graphite by engineered self-assembled peptides. *Langmuir*, 2012.

[29] Lihi Adler-Abramovich, Daniel Aronov, Peter Beker, Maya Yevnin, Shiri Stempler, Ludmila Buzhansky, Gil Rosenman, and Ehud Gazit. Self-assembled arrays of peptide nanotubes by vapour deposition. *Nature Nanotechnology*, 2009.

[30] P. Beker and G. Rosenman. Bioinspired nanostructural peptide materials for supercapacitor electrodes. *Journal of Materials Research*, 2010.

[31] Dharmendr Kumar, Nimesh Jain, Vinay Jain, and Beena Rai. Amino acids as copper corrosion inhibitors: A density functional theory approach. *Applied Surface Science*, 514:145905, 2020.

[32] Anton Kasprzhitskii, Georgy Lazorenko, Tatiana Nazdracheva, Aleksandr Kukharskii, Victor Yavna, and Andrei Kochur. Theoretical evaluation of the corrosion inhibition performance of aliphatic dipeptides. *New Journal of Chemistry*, 2021.

[33] Maryam Dehdab, Mehdi Shahraki, and Sayyed Mostafa Habibi-Khorassani. Theoretical study of inhibition efficiencies of some amino acids on corrosion of carbon steel in acidic media: Green corrosion inhibitors. *Amino Acids*, 2016.

[34] Jia Jun Fu, Su Ning Li, Ying Wang, Lin Hua Cao, and Lu De Lu. Computational and electrochemical studies of some amino acid compounds as corrosion inhibitors for mild steel in hydrochloric acid solution. *Journal of Materials Science*, 2010.

[35] Sanjoy Satpati, Aditya Suhasaria, Subhas Ghosal, Abhijit Saha, Sukalpa Dey, and Dipankar Sukul. Amino acid and cinnamaldehyde conjugated Schiff bases as proficient corrosion inhibitors for mild steel in 1 M HCl at higher temperature and prolonged exposure: Detailed electrochemical, adsorption and theoretical study. *Journal of Molecular Liquids*, 2021.

[36] Manu S. Mannoor, Hu Tao, Jefferson D. Clayton, Amartya Sengupta, David L. Kaplan, Rajesh R. Naik, Naveen Verma, Fiorenzo G. Omenetto, and Michael C. McAlpine. Graphene-based wireless bacteria detection on tooth enamel. *Nature Communications*, 2012.

[37] Hye Jin Hwang, Myung Yi Ryu, Chan Young Park, Junki Ahn, Hyun Gyu Park, Changsun Choi, Sang Do Ha, Tae Jung Park, and Jong Pil Park. High sensitive and selective electrochemical biosensor: Label-free detection of human norovirus using affinity peptide as molecular binder. *Biosensors and Bioelectronics*, 2017.

[38] Ning Xia, Xin Wang, Jie Yu, Yangyang Wu, Shuchao Cheng, Yun Xing, and Lin Liu. Design of electrochemical biosensors with peptide probes as the receptors of targets and the inducers of gold nanoparticles assembly on electrode surface. *Sensors and Actuators, B: Chemical*, 2017.

[39] Giwan Seo, Geonhee Lee, Mi Jeong Kim, Seung Hwa Baek, Minsuk Choi, Keun Bon Ku, Chang Seop Lee, Sangmi Jun, Daeui Park, Hong Gi Kim, Seong Jun Kim, Jeong O. Lee, Bum Tae Kim, Edmond Changkyun Park, and Seung Il Kim. Rapid Detection of COVID-19 Causative Virus (SARS-CoV-2) in Human Nasopharyngeal Swab Specimens Using Field-Effect Transistor-Based Biosensor. *ACS nano*, 2020.

[40] Javad Payandehpeyman, Neda Parvini, Kambiz Moradi, and Nima Hashemian. Detection of sars-cov-2 using antibody–antigen interactions with graphene-based nanomechanical resonator sensors. *ACS Applied Nano Materials*, 0(0):null, 0.

[41] Owen J. Guy, Gregory Burwell, Zari Tehrani, Ambroise Castaing, Kelly Ann Walker, and S.H. Doak. Graphene Nano-Biosensors for Detection of Cancer Risk. *Materials Science Forum*, 711:246–252, 2012.

[42] Dmitriy Khatayevich, Tamon Page, Carolyn Gresswell, Yuhei Hayamizu, William Grady, and Mehmet Sarikaya. Selective detection of target proteins by peptide-enabled graphene biosensor. *Small*, 10(8):1505–1513, 2014.

[43] Vincent Humblot, Christophe Méthivier, Rasmita Raval, and Claire Marie Pradier. Amino acid and peptides on Cu(1 1 0) surfaces: Chemical and structural analyses of l-lysine. *Surface Science*, 2007.

[44] M. Smerieri, L. Vattuone, T. Kravchuk, D. Costa, and L. Savio. (S)-glutamic acid on Ag(100): Self-assembly in the nonzwitterionic form. *Langmuir*, 2011.

[45] S. M. Barlow, S. Louafi, D. Le Roux, J. Williams, C. Muryn, S. Haq, and R. Raval. Polymorphism in supramolecular chiral structures of R- and S-alanine on Cu(1 1 0). *Surface Science*, 2005.

[46] Li Ping Xu, Yibiao Liu, and Xueji Zhang. Interfacial self-assembly of amino acids and peptides: Scanning tunneling microscopy investigation. *Nanoscale*, 2011.

[47] A. Kühnle, L. M. Molina, T. R. Linderoth, B. Hammer, and F. Besenbacher. Growth of unidirectional molecular rows of cysteine on Au(110)-(1 × 2) driven by adsorbate-induced surface rearrangements. *Physical Review Letters*, 2004.

[48] Angelika Kühnle, Trolle R. Linderoth, Michael Schunack, and Flemming Besenbacher. L-cysteine adsorption structures on Au(111) investigated by scanning tunneling microscopy under ultrahigh vacuum conditions. *Langmuir*, 2006.

[49] Sybille Fischer, Anthoula C. Papageorgiou, Matthias Marschall, Joachim Reichert, Katharina Diller, Florian Klappenberger, Francesco Allegretti, Alexei Nefedov, Christof Wöll, and Johannes V. Barth. L -Cysteine on Ag(111): A combined STM and X-ray spectroscopy study of anchorage and deprotonation. *Journal of Physical Chemistry C*, 2012.

[50] Christian Engelbrekt, Renat R. Nazmutdinov, Tamara T. Zinkicheva, Dmitrii V. Glukhov, Jiawei Yan, Bingwei Mao, Jens Ulstrup, and Jingdong Zhang. Chemistry of cysteine assembly on Au(100): Electrochemistry,: in situ STM and molecular modeling. *Nanoscale*, 2019.

[51] Feng Gao, Zhenjun Li, Yilin Wang, Luke Burkholder, and W. T. Tysoe. Chemistry of Alanine on Pd(1 1 1): Temperature-programmed desorption and X-ray photoelectron spectroscopic study. *Surface Science*, 2007.

[52] Julia Laskin, Peng Wang, and Omar Hadjar. Soft-landing of peptide ions onto self-assembled monolayer surfaces: An overview. *Physical Chemistry Chemical Physics*, 2008.

[53] Stephan Rauschenbach, Frank L. Stadler, Eugenio Lunedei, Nicola Malinowski, Sergej Koltsov, Giovanni Costantini, and Klaus Kern. Electrospray ion beam deposition of clusters and biomolecules. *Small*, 2006.

[54] Zoltán Takáts, Justin M. Wiseman, Bogdan Gologan, and R. Graham Cooks. Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science*, 2004.

[55] Stephan Rauschenbach, Ralf Vogelgesang, N. Malinowski, Jürgen W. Gerlach, Mohamed Benyoucef, Giovanni Costantini, Zhitao Deng, Nicha Thontasen, and Klaus

Kern. Electrospray ion beam deposition: Soft-landing and fragmentation of functional molecules at solid surfaces. *ACS Nano*, 2009.

[56] Stephan Rauschenbach, Markus Ternes, Ludger Harnau, and Klaus Kern. Mass Spectrometry as a Preparative Tool for the Surface Science of Large Molecules. *Annual Review of Analytical Chemistry*, 9(1):473–498, 2016.

[57] G. Binnig and H. Rohrer. SCANNING TUNNELING MICROSCOPY G. BINNIG and H. ROHRER. *Surface Science*, 1983.

[58] Zhitao Deng, Nicha Thontasen, Nikola Malinowski, Gordon Rinke, Ludger Harnau, Stephan Rauschenbach, and Klaus Kern. A close look at proteins: Submolecular resolution of two- and three-dimensionally folded cytochrome c at surfaces. *Nano Letters*, 12(5):2452–2458, 2012. PMID: 22530980.

[59] Gordon Rinke, Stephan Rauschenbach, Ludger Harnau, Alyazan Albarghash, Matthias Pauly, and Klaus Kern. Active conformation control of unfolded proteins by hyperthermal collision with a metal surface. *Nano Letters*, 14(10):5609–5615, 2014. PMID: 25198655.

[60] Stephan Rauschenbach, Gordon Rinke, Rico Gutzler, Sabine Abb, Alyazan Albarghash, Duy Le, Talat S. Rahman, Michael Duy, Ludger Harnau, and Klaus Kern. Two-Dimensional Folding of Polypeptides into Molecular Nanostructures at Surfaces. *ACS Nano*, 11(3):2420–2427, 2017.

[61] Dominique Costa, Claire-Marie Pradier, Frederik Tielens, and Letizia Savio. Adsorption and self-assembly of bio-organic molecules at model surfaces: A route towards increased complexity. *Surface Science Reports*, 70(4):449–553, 2015.

[62] Marian L. Clegg, Leonardo Morales De La Garza, Sofia Karakatsani, David A. King, and Stephen M. Driver. Chirality in amino acid overlayers on Cu surfaces. *Topics in Catalysis*, 2011.

[63] Yeliang Wang, Magalí Lingenfelder, Stefano Fabris, Guido Fratesi, Riccardo Ferrando, Thomas Classen, Klaus Kern, and Giovanni Costantini. Programming hierarchical supramolecular nanostructures by molecular design. *The Journal of Physical Chemistry C*, 117(7):3440–3445, 2013.

[64] K. E. Wilson, H. A. Früchtl, F. Grillo, and C. J. Baddeley. (s)-lysine adsorption induces the formation of gold nanofingers on au111. *Chem. Commun.*, 47:10365–10367, 2011.

[65] Xiubo Zhao, Fang Pan, and Jian R. Lu. Recent development of peptide self-assembly. *Progress in Natural Science*, 18(6):653–660, 2008.

[66] Joachim Reichert, Agustin Schiffrin, Willi Auwärter, Alexander Weber-Bargioni, Matthias Marschall, Martina Dell'Angela, Dean Cvetko, Gregor Bavdek, Albano Cossaro, Alberto Morgante, and Johannes V. Barth. L-tyrosine on Ag(111): Universality of the amino acid 2D zwitterionic bonding scheme? *ACS Nano*, 2010.

[67] Vitally Feyer, Oksana Plekan, Tomáš Skála, Vladimír Cháb, Vladimír Matolín, and Kevin C. Prince. The electronic structure and adsorption geometry of L-histidine on Cu(110). *Journal of Physical Chemistry B*, 2008.

[68] J. Williams, S. Haq, and R. Raval. The bonding and orientation of the amino acid L-alanine on Cu {110} determined by RAIRS. *Surface Science*, 1996.

[69] T. E. Jones, C. J. Baddeley, A. Gerbi, L. Savio, M. Rocca, and L. Vattuone. Molecular ordering and adsorbate induced faceting in the Ag{110}-(S)-glutamic acid system. *Langmuir*, 2005.

[70] V. De Renzi, L. Lavagnino, V. Corradini, R. Biagi, M. Canepa, and U. Del Pennino. Very low energy vibrational modes as a fingerprint of H-bond network formation: L-cysteine on Au(111). *Journal of Physical Chemistry C*, 2008.

[71] M. Smerieri, L. Vattuone, M. Rocca, and L. Savio. Spectroscopic evidence for neutral and anionic adsorption of (S)-glutamic acid on Ag(111). *Langmuir*, 2013.

[72] Paul Adrien Maurice Dirac and Ralph Howard Fowler. Quantum mechanics of many-electron systems. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 123(792):714–733, 1929.

[73] Rosa Di Felice, Annabella Selloni, and Elisa Molinari. DFT Study of Cysteine Adsorption on Au(111). *The Journal of Physical Chemistry B*, 107(5):1151–1156, 2003.

[74] Rosa Di Felice and Annabella Selloni. Adsorption modes of cysteine on Au(111): Thiolate, amino-thiolate, disulfide. *The Journal of Chemical Physics*, 120(10):4906–4914, 2004.

[75] Luca M Ghiringhelli, Pim Schravendijk, and Luigi Delle Site. Adsorption of alanine on a Ni(111) surface: A multiscale modeling oriented density functional study. *Phys. Rev. B*, 74(3):35437, 2006.

[76] Corinne Arrouvel, Boubakar Diawara, Dominique Costa, and Philippe Marcus. DFT Periodic Study of the Adsorption of Glycine on the Anhydrous and Hydroxylated (0001) Surfaces of α-Alumina. *The Journal of Physical Chemistry C*, 111(49):18164–18173, 2007.

[77] Francesco Iori, Stefano Corni, and Rosa Di Felice. Unraveling the Interaction between Histidine Side Chain and the Au(111) Surface: A DFT Study. *The Journal of Physical Chemistry C*, 112(35):13540–13545, 2008.

[78] Wei Liu, Alexandre Tkatchenko, and Matthias Scheffler. Modeling adsorption and reactions of organic molecules at metal surfaces. *Accounts of Chemical Research*, 2014.

[79] Musa Ozboyaci, Daria B. Kokh, Stefano Corni, and Rebecca C. Wade. Modeling and simulation of protein-surface interactions: Achievements and challenges. *Quarterly Reviews of Biophysics*, 2016.

[80] Sung Sik Lee, Bongsoo Kim, and Sungyul Lee. Structures and bonding properties of Gold-Arg-Cys complexes: DFT study of simple peptide-coated metal. *Journal of Physical Chemistry C*, 2014.

[81] R. R. Nazmutdinov, I. R. Manyurov, T. T. Zinkicheva, J. Jang, and J. Ulstrup. Cysteine adsorption on the Au(111) surface and the electron transfer in configuration of a scanning tunneling microscope: A quantum-chemical approach. *Russian Journal of Electrochemistry*, 2007.

[82] José L.C. Fajín, José R.B. Gomes, and M. Natália D.S. Cordeiro. DFT study of the adsorption of d-(l-)cysteine on flat and chiral stepped gold surfaces. *Langmuir*, 2013.

[83] Dmitrii Maksimov, Carsten Baldauf, and Mariana Rossi. The conformational space of a flexible amino acid at metallic surfaces. *International Journal of Quantum Chemistry*, 121(3):e26369, 2021.

[84] Gongyi Hong, Hendrik Heinz, Rajesh R Naik, Barry L Farmer, and Ruth Pachter. Toward Understanding Amino Acid Adsorption at Metallic Interfaces: A Density Functional Theory Study. *ACS Applied Materials & Interfaces*, 1(2):388–392, 2009.

[85] Louise B. Wright, P. Mark Rodger, Stefano Corni, and Tiffany R. Walsh. GolP-CHARMM: First-principles based force fields for the interaction of proteins with Au(111) and Au(100). *Journal of Chemical Theory and Computation*, 2013.

[86] Zak E. Hughes, Louise B. Wright, and Tiffany R. Walsh. Biomolecular adsorption at aqueous silver interfaces: First-principles calculations, polarizable force-field simulations, and comparisons with gold. *Langmuir*, 2013.

[87] Hendrik Heinz, R. A. Vaia, B. L. Farmer, and R. R. Naik. Accurate simulation of surfaces and interfaces of face-centered cubic metals using 12-6 and 9-6 lennard-jones potentials. *Journal of Physical Chemistry C*, 112(44):17281–17290, 2008.

[88] Andrea Grisafi and Michele Ceriotti. Incorporating long-range physics in atomic-scale machine learning. *Journal of Chemical Physics*, 2019.

[89] Alexandre Tkatchenko. Machine learning for chemical discovery. *Nature Communications*, 2020.

[90] Jörg Behler. Four Generations of High-Dimensional Neural Network Potentials. *Chemical Reviews*, 2021.

[91] Zdenek Futera and Jochen Blumberger. Adsorption of Amino Acids on Gold: Assessing the Accuracy of the GolP-CHARMM Force Field and Parametrization of Au-S Bonds. *Journal of Chemical Theory and Computation*, 2019.

[92] Thomas P. Senftle, Sungwook Hong, Md Mahbubul Islam, Sudhir B. Kylasa, Yuanxia Zheng, Yun Kyung Shin, Chad Junkermeier, Roman Engel-Herbert, Michael J. Janik,

Hasan Metin Aktulga, Toon Verstraelen, Ananth Grama, and Adri C.T. Van Duin. The ReaxFF reactive force-field: Development, applications and future directions. *npj Computational Materials*, 2016.

[93] Susanna Monti, Cui Li, and Vincenzo Carravetta. Reactive dynamics simulation of monolayer and multilayer adsorption of glycine on Cu(110). *Journal of Physical Chemistry C*, 2013.

[94] Dominik Marx and Jürg Hutter. Ab initio molecular dynamics: Theory and implementation. *Modern methods and algorithms of quantum chemistry*, 2000.

[95] Alessandro Motta, Marie Pierre Gaigeot, and Dominique Costa. AIMD evidence of inner sphere adsorption of glycine on a stepped (101) boehmite AlOOH surface. *Journal of Physical Chemistry C*, 2012.

[96] Frederik Tielens, Vincent Humblot, and Claire Marie Pradier. Elucidation of the low coverage chiral adsorption assembly of l-lysine on Cu(1 1 0) surface: A theoretical study. *Surface Science*, 2008.

[97] Julian Schneider and Lucio Colombi Ciacchi. Specific material recognition by small peptides mediated by the interfacial solvent structure. *Journal of the American Chemical Society*, 2012.

[98] Arrigo Calzolari, Giancarlo Cicero, Carlo Cavazzoni, Rosa Di Felice, Alessandra Catellani, and Stefano Corni. Hydroxyl-rich β-sheet adhesion to the gold surface in water by first-principle simulations. *Journal of the American Chemical Society*, 2010.

[99] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95–99, 1963.

[100] O. Anatole von Lilienfeld, Raghunathan Ramakrishnan, Matthias Rupp, and Aaron Knoll. Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *International Journal of Quantum Chemistry*, 115(16):1084–1093, 2015.

[101] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[102] Stephen R. Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics*, 2015.

[103] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. PMID: 20426451.

[104] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108:058301, Jan 2012.

[105] Katja Hansen, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O. Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The Journal of Physical Chemistry Letters*, 6(12):2326–2331, 2015. PMID: 26113956.

[106] Haoyan Huo and Matthias Rupp. Unified representation of molecules and crystals for machine learning. *arXiv*, 2017.

[107] Marcel F. Langer, Alex Goeßmann, and Matthias Rupp. Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *arXiv*, 2020.

[108] Bing Huang and O. Anatole von Lilienfeld. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *The Journal of Chemical Physics*, 145(16):161102, 2016.

[109] Albert P. Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R. Kermode, Gábor Csányi, and Michele Ceriotti. Machine learning unifies the modeling of materials and molecules. *Science Advances*, 3(12), 2017.

[110] Sandip De, Felix Musil, Teresa Ingram, Carsten Baldauf, and Michele Ceriotti. Mapping and classifying molecules from a high-throughput structural database. *Journal of Cheminformatics*, 9(1):6, 2017.

[111] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87:184115, May 2013.

[112] M. Ceriotti, G. A. Tribello, and M. Parrinello. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences*, 108(32):13023–13028, 2011.

[113] Gareth A Tribello, Michele Ceriotti, and Michele Parrinello. Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 109(14):5196–5201, 2012.

[114] Michele Ceriotti, Gareth A Tribello, and Michele Parrinello. Demonstrating the Transferability and the Descriptive Power of Sketch-Map. *Journal of Chemical Theory and Computation*, 9(3):1521–1532, 2013.

[115] Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.*, 18:13754–13769, 2016.

[116] Lauri Himanen, Marc O.J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, and Adam S. Foster. Dscribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020.

[117] Felix Mayr and Alessio Gagliardi. Global property prediction: A benchmark study on open-source, perovskite-like datasets. *ACS Omega*, 6(19):12722–12732, 2021.

[118] Michele Ceriotti. Unsupervised machine learning in atomistic simulations, between predictions and understanding. *Journal of Chemical Physics*, 2019.

[119] Jerzy Leszczynski, Anna Kaczmarek-Kedziera, Tomasz Puzyn, Manthos G. Papadopoulos, Heribert Reis, and Manoj K. Shukla. Handbook of computational chemistry. *Handbook of Computational Chemistry*, 2017.

[120] Mario Barbatti, Matthias Ruckenbauer, Jaroslaw J. Szymczak, Adélia J.A. Aquino, and Hans Lischka. Nonadiabatic excited-state dynamics of polar π-systems and related model compounds of biological relevance. *Physical Chemistry Chemical Physics*, 2008.

[121] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

[122] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965.

[123] L. H. Thomas. The calculation of atomic fields. *Mathematical Proceedings of the Cambridge Philosophical Society*, 1927.

[124] Enrico Fermi. Statistical method to determine some properties of atoms. *Rend. Accad. Naz. Lincei*, 1927.

[125] M. Levy. Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem. *Proceedings of the National Academy of Sciences of the United States of America*, 1979.

[126] Matt Probert. Electronic Structure: Basic Theory and Practical Methods, by Richard M. Martin. *Contemporary Physics*, 2011.

[127] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A), 1965.

[128] D. M. Ceperley and B. J. Alder. Ground state of the electron gas by a stochastic method. *Physical Review Letters*, 1980.

[129] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 1996.

[130] Victor G. Ruiz, Wei Liu, and Alexandre Tkatchenko. Density-functional theory with screened van der Waals interactions applied to atomic and molecular adsorbates on close-packed and non-close-packed surfaces. *Physical Review B*, 93(3):035118, 2016.

[131] Axel D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 1993.

[132] Attila Szabo and Neil Ostlund. Szabo and Ostlund - Modern Quantum Chemistry. *SERBIULA (sistema Librum 2.0)*, 1996.

[133] Susi Lehtola, Conrad Steigemann, Micael J.T. Oliveira, and Miguel A.L. Marques. Recent developments in LIBXC — A comprehensive library of functionals for density functional theory. *SoftwareX*, 2018.

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

[135] Eberhard Engel and Reiner M. Dreizler. Density Functional Theory: An Advanced Course. *Theoretical and Mathematical Physics*, 2011.

[136] Matthias Ernzerhof and Gustavo E. Scuseria. Assessment of the Perdew-Burke-Ernzerhof exchange-correlation functional. *Journal of Chemical Physics*, 1999.

[137] Carlo Adamo and Vincenzo Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *Journal of Chemical Physics*, 1999.

[138] John P. Perdew and Yue Wang. Accurate and simple analytic representation of the electron-gas correlation energy. *Physical Review B*, 1992.

[139] Jorge Kohanoff. Electronic structure calculations for solids and molecules: Theory and computational methods, 2006.

[140] John P. Perdew, Matthias Ernzerhof, and Kieron Burke. Rationale for mixing exact exchange with density functional approximations. *Journal of Chemical Physics*, 1996.

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

[142] Robert L. Baldwin. Energetics of protein folding. *Journal of Molecular Biology*, 371(2):283–301, 2007.

[143] Robert A. DiStasio, O. Anatole von Lilienfeld, and Alexandre Tkatchenko. Collective many-body van der waals interactions in molecular systems. *Proceedings of the National Academy of Sciences*, 109(37):14791–14795, 2012.

[144] Mariana Rossi, Wei Fang, and Angelos Michaelides. Stability of complex biomolecular structures: van der waals, hydrogen bond cooperativity, and nuclear quantum effects. *The Journal of Physical Chemistry Letters*, 6(21):4233–4238, 2015. PMID: 26722963.

[145] Anthony M. Reilly, Richard I. Cooper, Claire S. Adjiman, Saswata Bhattacharya, A. Daniel Boese, Jan Gerit Brandenburg, Peter J. Bygrave, Rita Bylsma, Josh E. Campbell, Roberto Car, David H. Case, Renu Chadha, Jason C. Cole, Katherine Cosburn, Herma M. Cuppen, Farren Curtis, Graeme M. Day, Robert A. DiStasio Jr, Alexander Dzyabchenko, Bouke P. van Eijck, Dennis M. Elking, Joost A. van den Ende, Julio C. Facelli, Marta B. Ferraro, Laszlo Fusti-Molnar, Christina-Anna Gatsiou, Thomas S. Gee, René de Gelder, Luca M. Ghiringhelli, Hitoshi Goto, Stefan Grimme, Rui Guo, Detlef W. M. Hofmann, Johannes Hoja, Rebecca K. Hylton, Luca Iuzzolino, Wojciech Jankiewicz, Daniël T. de Jong, John Kendrick, Niek J. J. de Klerk, Hsin-Yu Ko, Liudmila N. Kuleshova, Xiayue Li, Sanjaya Lohani, Frank J. J. Leusen, Albert M. Lund, Jian Lv, Yanming Ma, Noa Marom, Artëm E. Masunov, Patrick McCabe, David P. McMahon, Hugo Meekes, Michael P. Metz, Alston J. Misquitta, Sharmarke Mohamed, Bartomeu Monserrat, Richard J. Needs, Marcus A. Neumann, Jonas Nyman, Shigeaki Obata, Harald Oberhofer, Artem R. Oganov, Anita M. Orendt, Gabriel I. Pagola, Constantinos C. Pantelides, Chris J. Pickard, Rafal Podeszwa, Louise S. Price, Sarah L. Price, Angeles Pulido, Murray G. Read, Karsten Reuter, Elia Schneider, Christoph Schober, Gregory P. Shields, Pawanpreet Singh, Isaac J. Sugden, Krzysztof Szalewicz, Christopher R. Taylor, Alexandre Tkatchenko, Mark E. Tuckerman, Francesca Vacarro, Manolis Vasileiadis, Alvaro Vazquez-Mayagoitia, Leslie Vogt, Yanchao Wang, Rona E. Watson, Gilles A. de Wijs, Jack Yang, Qiang Zhu, and Colin R. Groom. Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallographica Section B*, 72(4):439–459, Aug 2016.

[146] Noa Marom, Robert A. DiStasio Jr., Viktor Atalla, Sergey Levchenko, Anthony M. Reilly, James R. Chelikowsky, Leslie Leiserowitz, and Alexandre Tkatchenko. Many-body dispersion interactions in molecular crystal polymorphism. *Angewandte Chemie International Edition*, 52(26):6629–6632, 2013.

[147] Javier Carrasco, Wei Liu, Angelos Michaelides, and Alexandre Tkatchenko. Insight into the description of van der waals forces for benzene adsorption on transition metal (111) surfaces. *The Journal of Chemical Physics*, 140(8):084704, 2014.

[148] Wei Liu, Friedrich Maaß, Martin Willenbockel, Christopher Bronner, Michael Schulze, Serguei Soubatch, F. Stefan Tautz, Petra Tegeder, and Alexandre Tkatchenko. Quantitative prediction of molecular adsorption: Structure and binding of benzene on coinage metals. *Phys. Rev. Lett.*, 115:036104, Jul 2015.

[149] Reinhard J. Maurer, Victor G. Ruiz, Javier Camarillo-Cisneros, Wei Liu, Nicola Ferri, Karsten Reuter, and Alexandre Tkatchenko. Adsorption structures and energetics of molecules on metal surfaces: Bridging experiment and theory. *Progress in Surface Science*, 91(2):72–100, 2016.

[150] Wei Liu, Victor G. Ruiz, Guo Xu Zhang, Biswajit Santra, Xinguo Ren, Matthias Scheffler, and Alexandre Tkatchenko. Structure and energetics of benzene adsorbed on

transition-metal surfaces: Density-functional theory with van der Waals interactions including collective substrate response. *New Journal of Physics*, 2013.

[151] Wei Liu, Javier Carrasco, Biswajit Santra, Angelos Michaelides, Matthias Scheffler, and Alexandre Tkatchenko. Benzene adsorbed on metals: Concerted effect of covalency and van der Waals bonding. *Phys. Rev. B*, 86(24):245405, 2012.

[152] W A Al-Saidi, Haijun Feng, and Kristen A Fichthorn. Adsorption of Polyvinylpyrrolidone on Ag Surfaces: Insight into a Structure-Directing Agent. *Nano Letters*, 12(2):997–1001, 2012.

[153] Jan van Ruitenbeek. Dispersion forces unveiled. *Nature Materials*, 11:834–835, 2012.

[154] C. Wagner, N. Fournier, F. S. Tautz, and R. Temirov. Measurement of the binding energies of the organic-metal perylene-teracarboxylic-dianhydride/au(111) bonds by molecular manipulation using an atomic force microscope. *Phys. Rev. Lett.*, 109:076102, 2012.

[155] Jan Hermann and Alexandre Tkatchenko. Density-functional model for van der Waals interactions: Unifying atomic approaches with nonlocal functionals. *arXiv e-prints*, 2019.

[156] Jiří Klimeš and Angelos Michaelides. Perspective: Advances and challenges in treating van der waals dispersion forces in density functional theory. *The Journal of Chemical Physics*, 137(12):120901, 2012.

[157] Stefan Grimme. Accurate description of van der Waals complexes by density functional theory including empirical corrections. *Journal of Computational Chemistry*, 2004.

[158] Stefan Grimme. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of Computational Chemistry*, 2006.

[159] 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. *Journal of Chemical Physics*, 2010.

[160] Axel D. Becke and Erin R. Johnson. Exchange-hole dipole moment and the dispersion interaction. *Journal of Chemical Physics*, 2005.

[161] Axel D. Becke and Erin R. Johnson. A density-functional model of the dispersion interaction. *Journal of Chemical Physics*, 2005.

[162] Erin R. Johnson and Axel D. Becke. A post-Hartree-Fock model of intermolecular interactions. *Journal of Chemical Physics*, 2005.

[163] Alexandre Tkatchenko and Matthias Scheffler. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Physical Review Letters*, 2009.

[164] Takeshi Sato and Hiromi Nakai. Density functional method including weak interactions: Dispersion coefficients based on the local response approximation. *Journal of Chemical Physics*, 2009.

[165] John F. Dobson, Angela White, and Angel Rubio. Asymptotics of the dispersion interaction: Analytic benchmarks for van der Waals energy functionals. *Physical Review Letters*, 2006.

[166] Marcus Elstner, Pavel Hobza, Thomas Frauenheim, Sándor Suhai, and Efthimios Kaxiras. Hydrogen bonding and stacking interactions of nucleic acid base pairs: A density functional-theory based treatment. *Journal of Chemical Physics*, 2001.

[167] Axel D. Becke and Erin R. Johnson. Exchange-hole dipole moment and the dispersion interaction revisited. *Journal of Chemical Physics*, 2007.

[168] Petr Jurečka, Jiří Černý, Pavel Hobza, and Dennis R. Salahub. Density functional theory augmented with an empirical dispersion term. Interaction energies and geometries of 80 noncovalent complexes compared with ab initio quantum mechanics calculations. *Journal of Computational Chemistry*, 2007.

[169] Urs Zimmerli, Michele Parrinello, and Petros Koumoutsakos. Dispersion corrections to density functionals for water aromatic interactions. *Journal of Chemical Physics*, 2004.

[170] Marcus A. Neumann and Marc Antoine Perrin. Energy ranking of molecular crystals using density functional theory calculations and an empirical van der waals correction. *Journal of Physical Chemistry B*, 2005.

[171] H. B. G. Casimir and D. Polder. The influence of retardation on the london-van der waals forces. *Phys. Rev.*, 73:360–372, Feb 1948.

[172] K. T. Tang and M. Karplus. Padé-approximant calculation of the nonretarded van der waals coefficients for two and three helium atoms. *Phys. Rev.*, 171:70–74, Jul 1968.

[173] X. Chu and A. Dalgarno. Linear response time-dependent density functional theory for van der waals coefficients. *The Journal of Chemical Physics*, 121(9):4083–4088, 2004.

[174] Tore Brinck, Jane S. Murray, and Peter Politzer. Polarizability and volume. *The Journal of Chemical Physics*, 98(5):4305–4306, 1993.

[175] F. L. Hirshfeld. Bonded-atom fragments for describing molecular charge densities. *Theoretica Chimica Acta*, 1977.

[176] Petr Jurečka, Jiří Šponer, Jiří Černý, 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. *Physical Chemistry Chemical Physics*, 2006.

[177] Victor G. Ruiz, Wei Liu, Egbert Zojer, Matthias Scheffler, and Alexandre Tkatchenko. Density-functional theory with screened van der Waals interactions for the modeling of hybrid inorganic-organic systems. *Physical Review Letters*, 2012.

[178] E M Lifshitz. The theory of molecular attractive forces between solids. *Journal of Experimental and Theoretical Physics*, 1956.

[179] E. Zaremba and W. Kohn. Van der Waals interaction between an atom and a solid surface. *Physical Review B*, 1976.

[180] Guo Xu Zhang, Alexandre Tkatchenko, Joachim Paier, Heiko Appel, and Matthias Scheffler. Van der Waals interactions in ionic and semiconductor solids. *Physical Review Letters*, 2011.

[181] Todd Raeker. Physical Adsorption: Forces and Phenomena (Bruch, L.W.; Cole, Milton W.; Zaremba, Eugene). *Journal of Chemical Education*, 1998.

[182] Wolfgang S.M. Werner, Kathrin Glantschnig, and Claudia Ambrosch-Draxl. Optical constants and inelastic electron-scattering data for 17 elemental metals. *Journal of Physical and Chemical Reference Data*, 2009.

[183] 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. *Computer Physics Communications*, 180:2175–2196, 2009.

[184] V. Havu, V. Blum, P. Havu, and M. Scheffler. Efficient o(n) integration for all-electron electronic structure calculation using numeric basis functions. *Journal of Computational Physics*, 228(22):8367–8379, 2009.

[185] Igor Ying Zhang, Xinguo Ren, Patrick Rinke, Volker Blum, and Matthias Scheffler. Numeric atom-centered-orbital basis sets with valence-correlation consistency from H to Ar. *New Journal of Physics*, 2013.

[186] Jörg Neugebauer and Matthias Scheffler. Adsorbate-substrate and adsorbate-adsorbate interactions of Na and K adlayers on Al(111). *Physical Review B*, 1992.

[187] Norina A. Richter, Sabrina Sicolo, Sergey V. Levchenko, Joachim Sauer, and Matthias Scheffler. Concentration of vacancies at metal-oxide surfaces: Case study of MgO(100). *Physical Review Letters*, 2013.

[188] Daniel Berger, Andrew J. Logsdail, Harald Oberhofer, Matthew R. Farrow, C. Richard A. Catlow, Paul Sherwood, Alexey A. Sokol, Volker Blum, and Karsten Reuter. Embedded-cluster calculations in a numeric atomic orbital density-functional theory framework. *Journal of Chemical Physics*, 2014.

[189] Wolfram Steurer, Shadi Fatayer, Leo Gross, and Gerhard Meyer. Probe-based measurement of lateral single-electron transfer between individual molecules. *Nature Communications*, 2015.

[190] Daniel Hernangómez-Pérez, Jakob Schlör, David A. Egger, Laerte L. Patera, Jascha Repp, and Ferdinand Evers. Reorganization energy and polaronic effects of pentacene on NaCl films. *Physical Review B*, 2020.

[191] Hannu Pekka Komsa and Alfredo Pasquarello. Finite-size supercell correction for charged defects at surfaces and interfaces. *Physical Review Letters*, 2013.

[192] Ismaila Dabo, Boris Kozinsky, Nicholas E. Singh-Miller, and Nicola Marzari. Electrostatics in periodic boundary conditions and real-space corrections. *Physical Review B - Condensed Matter and Materials Physics*, 2008.

[193] M. Otani and O. Sugino. First-principles calculations of charged surfaces and interfaces: A plane-wave nonrepeated slab approach. *Physical Review B - Condensed Matter and Materials Physics*, 2006.

[194] Ofer Sinai, Oliver T. Hofmann, Patrick Rinke, Matthias Scheffler, Georg Heimel, and Leeor Kronik. Multiscale approach to the electronic structure of doped semiconductor surfaces. *Physical Review B - Condensed Matter and Materials Physics*, 2015.

[195] Christoph Freysoldt, Arpit Mishra, Michael Ashton, and Jörg Neugebauer. Generalized dipole correction for charged surfaces in the repeated-slab approach. *Physical Review B*, 2020.

[196] J. Bardeen. Tunnelling from a many-particle point of view. *Physical Review Letters*, 1961.

[197] J. Tersoff and D. R. Hamann. Theory and application for the scanning tunneling microscope. *Phys. Rev. Lett.*, 50:1998–2001, Jun 1983.

[198] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B*, 1998.

[199] Hendrik Heinz, Barry L. Farmer, Ras B. Pandey, Joseph M. Slocik, Soumya S. Patnaik, Ruth Pachter, and Rajesh R. Naik. Nature of molecular interactions of peptides with gold, palladium, and Pd-Au bimetal surfaces in aqueous solution. *Journal of the American Chemical Society*, 2009.

[200] Jie Feng, Ras B. Pandey, Rajiv J. Berry, Barry L. Farmer, Rajesh R. Naik, and Hendrik Heinz. Adsorption mechanism of single amino acid and surfactant molecules to Au {111} surfaces in aqueous solution: Design rules for metal-binding molecules. *Soft Matter*, 2011.

[201] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with namd. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[202] 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. *Journal of Chemical Information and Modeling*, 2005.

[203] David H. Wolpert and William G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1997.

[204] Ruxi Qi, Guanghong Wei, Buyong Ma, and Ruth Nussinov. Replica exchange molecular dynamics: A practical application protocol with solutions to common problems and a peptide aggregation and self-assembly example. 2018.

[205] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 1977.

[206] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. THE weighted histogram analysis method for free [U+2010] energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 1992.

[207] Christian Bartels. Analyzing biased Monte Carlo and molecular dynamics simulations. *Chemical Physics Letters*, 2000.

[208] Erik G. Brandt and Alexander P. Lyubartsev. Molecular Dynamics Simulations of Adsorption of Amino Acid Side Chain Analogues and a Titanium Binding Peptide on the TiO2 (100) Surface. *Journal of Physical Chemistry C*, 2015.

[209] Daan Frenkel, Berend Smit, Jan Tobochnik, Susan R. McKay, and Wolfgang Christian. Understanding Molecular Simulation. *Computers in Physics*, 1997.

[210] Martin Hoefling, Francesco Iori, Stefano Corni, and Kay Eberhard Gottschalk. The conformations of amino acids on a gold(111) surface. *ChemPhysChem*, 2010.

[211] Zdenek Futera. Amino-acid interactions with the Au(111) surface: Adsorption, band alignment, and interfacial electronic coupling. *Physical Chemistry Chemical Physics*, 2021.

[212] Carsten Baldauf and Mariana Rossi. Going clean: Structure and dynamics of peptides in the gas phase and paths to solvation. *Journal of Physics Condensed Matter*, 2015.

[213] Hendrik Heinz and Hadi Ramezani-Dakhel. Simulations of inorganic–bioorganic interfaces to discover new materials: insights, comparisons to experiment, challenges, and opportunities. *Chem. Soc. Rev.*, 45(2):412–448, 2016.

[214] Chris J. Pickard and R. J. Needs. High-pressure phases of silane. *Physical Review Letters*, 2006.

[215] Georg Schusteritsch and Chris J. Pickard. Predicting interface structures: From SrTiO3 to graphene. *Physical Review B - Condensed Matter and Materials Physics*, 2014.

[216] Miri Zilka, Dmytro V. Dudenko, Colan E. Hughes, P. Andrew Williams, Simone Sturniolo, W. Trent Franks, Chris J. Pickard, Jonathan R. Yates, Kenneth D.M. Harris, and Steven P. Brown. Ab initio random structure searching of organic molecular solids: Assessment and validation against experimental data. *Physical Chemistry Chemical Physics*, 2017.

[217] Sandro E Schönborn, Stefan Goedecker, Shantanu Roy, and Artem R Oganov. The performance of minima hopping and evolutionary algorithms for cluster structure prediction. *The Journal of Chemical Physics*, 130(14):144108, 2009.

[218] Thomas Bäck. Evolutionary Algorithms in Theory and Practice. *Evolutionary Algorithms in Theory and Practice*, 1996.

[219] Bernd Hartke. Efficient Global Geometry Optimization of Atomic and Molecular Clusters. 2006.

[220] Adriana Supady, Volker Blum, and Carsten Baldauf. First-Principles Molecular Structure Search with a Genetic Algorithm. *Journal of Chemical Information and Modeling*, 55(11):2338–2348, 2015.

[221] Farren Curtis, Xiayue Li, Timothy Rose, Álvaro Vázquez-Mayagoitia, Saswata Bhattacharya, Luca M. Ghiringhelli, and Noa Marom. GAtor: A First-Principles Genetic Algorithm for Molecular Crystal Structure Prediction. *Journal of Chemical Theory and Computation*, 2018.

[222] 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. *Journal of Physical Chemistry A*, 1997.

[223] Stefan Goedecker. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *Journal of Chemical Physics*, 2004.

[224] Konstantin Krautgasser, Chiara Panosetti, Dennis Palagin, Karsten Reuter, and Reinhard J. Maurer. Global structure search for molecules on surfaces: Efficient sampling with curvilinear coordinates. *Journal of Chemical Physics*, 2016.

[225] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 2010.

[226] J. Behler. Representing potential energy surfaces by high-dimensional neural network potentials. *Journal of Physics Condensed Matter*, 2014.

[227] Milica Todorović, Michael U. Gutmann, Jukka Corander, and Patrick Rinke. Bayesian inference of atomistic structure in functional materials. *npj Computational Materials*, 2019.

[228] Jari Järvi, Patrick Rinke, and Milica Todorović. Detecting stable adsorbates of (1 S)-camphor on Cu(111) with Bayesian optimization . *Beilstein Journal of Nanotechnology*, 2020.

[229] Jari Järvi, Benjamin Alldritt, Ondřej Krejčí, Milica Todorović, Peter Liljeroth, and Patrick Rinke. Integrating Bayesian Inference with Scanning Probe Experiments for Robust Identification of Surface Adsorbate Configurations. *Advanced Functional Materials*, 2021.

[230] Lincan Fang, Esko Makkonen, Milica Todorović, Patrick Rinke, and Xi Chen. Efficient Amino Acid Conformer Search with Bayesian Optimization. *Journal of Chemical Theory and Computation*, 2021.

[231] Jorge Nocedal and Stephen J. Wright. Numerical optimization 2nd edition. *International ADAMS user conference*, 2000.

[232] C. G. Broyden. Quasi-Newton methods and their application to function minimisation. *Mathematics of Computation*, 1967.

[233] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 1970.

[234] Oliver T. Hofmann, Egbert Zojer, Lukas Hörmann, Andreas Jeindl, and Reinhard J. Maurer. First-principles calculations of hybrid inorganic-organic interfaces: From state-of-the-art to best practice. *Physical Chemistry Chemical Physics*, 2021.

[235] Letif Mones, Christoph Ortner, and Gábor Csányi. Preconditioners for the geometry optimisation and saddle point search of molecular systems. *Scientific Reports*, 2018.

[236] Roland Lindh, Anders Bernhardsson, Gunnar Karlström, and Per Åke Malmqvist. On the use of a hessian model function in molecular geometry optimizations. *Chemical Physics Letters*, 241(4):423 – 428, 1995.

[237] David Packwood, James Kermode, Letif Mones, Noam Bernstein, John Woolley, Nicholas Gould, Christoph Ortner, and Gábor Csányi. A universal preconditioner for simulating condensed phase materials. *Journal of Chemical Physics*, 2016.

[238] Andrew A. Peterson. Acceleration of saddle-point searches with machine learning. *Journal of Chemical Physics*, 2016.

[239] Olli Pekka Koistinen, Freyja B. Dagbjartsdóttir, Vilhjálmur Ásgeirsson, Aki Vehtari, and Hannes Jónsson. Nudged elastic band calculations accelerated with Gaussian process regression. *Journal of Chemical Physics*, 2017.

[240] José A. Garrido Torres, Paul C. Jennings, Martin H. Hansen, Jacob R. Boes, and Thomas Bligaard. Low-Scaling Algorithm for Nudged Elastic Band Calculations Using a Surrogate Machine Learning Model. *Physical Review Letters*, 2019.

[241] Nongnuch Artrith and Jörg Behler. High-dimensional neural network potentials for metal surfaces: A prototype study for copper. *Physical Review B - Condensed Matter and Materials Physics*, 2012.

[242] Estefanía Garijo Del Río, Jens Jørgen Mortensen, and Karsten Wedel Jacobsen. Local Bayesian optimizer for atomic structures. *Physical Review B*, 2019.

[243] Estefanía Garijo Del Río, Sami Kaappa, José A. Garrido Torres, Thomas Bligaard, and Karsten Wedel Jacobsen. Machine learning with bond information for local structure optimizations in surface science. *The Journal of chemical physics*, 2020.

[244] Muhammed Shuaibi, Saurabh Sivakumar, Rui Qi Chen, and Zachary W Ulissi. Enabling robust offline active learning for machine learning potentials using simple physics-based priors. *Machine Learning: Science and Technology*, 2021.

[245] Ryosuke Jinnouchi, Kazutoshi Miwa, Ferenc Karsai, Georg Kresse, and Ryoji Asahi. On-the-Fly Active Learning of Interatomic Potentials for Large-Scale Atomistic Simulations. *Journal of Physical Chemistry Letters*, 2020.

[246] WOLFE P. Convergence Conditions for Ascent Methods. *SIAM Review*, 1969.

[247] Philip Wolfe. Convergence Conditions for Ascent Methods II: Some Corrections. *SIAM Review*, 1971.

[248] Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 1966.

[249] David Packwood, James Kermode, Letif Mones, Noam Bernstein, John Woolley, Nicholas Gould, Christoph Ortner, and Gábor Csányi. A universal preconditioner for simulating condensed phase materials. *The Journal of Chemical Physics*, 144(16):164109, 2016.

[250] Thomas H. Fischer and Jan Almlöf. General methods for geometry and wave function optimization. *Journal of Physical Chemistry*, 1992.

[251] Geza Fogarasi, Xuefeng Zhou, Patterson W. Taylor, and Peter Pulay. The Calculation of ab Initio Molecular Geometries: Efficient Optimization by Natural Internal Coordinates and Empirical Correction by Offset Forces. *Journal of the American Chemical Society*, 1992.

[252] Letif Mones, Christoph Ortner, and Gábor Csányi. Preconditioners for the geometry optimisation and saddle point search of molecular systems. *Scientific Reports*, 2018.

[253] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 1955.

[254] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

[255] P. W.A. Howe. Principal components analysis of protein structure ensembles calculated using NMR data. *Journal of Biomolecular NMR*, 2001.

[256] Lee Wei Yang, Eran Eyal, Ivet Bahar, and Akio Kitao. Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): Insights into functional dynamics. *Bioinformatics*, 2009.

[257] Katrijn Van Deun, Elise A.V. Crompvoets, and Eva Ceulemans. Obtaining insights from high-dimensional data: Sparse principal covariates regression. *BMC Bioinformatics*, 2018.

[258] Andrea Anelli, Edgar A. Engel, Chris J. Pickard, and Michele Ceriotti. Generalized convex hull construction for materials discovery. *Physical Review Materials*, 2018.

[259] Mazen Ahmad, Volkhard Helms, Olga V. Kalinina, and Thomas Lengauer. Relative Principal Components Analysis: Application to Analyzing Biomolecular Conformational Changes. *Journal of Chemical Theory and Computation*, 2019.

[260] Benjamin A Helfrecht, Rose K Cersonsky, Guillaume Fraux, and Michele Ceriotti. Structure-property maps with Kernel principal covariates regression. *Machine Learning: Science and Technology*, 2020.

[261] Piero Gasparotto, Maria Fischer, Daniele Scopece, Maciej O. Liedke, Maik Butterling, Andreas Wagner, Oguz Yildirim, Mathis Trant, Daniele Passerone, Hans J. Hug, and Carlo A. Pignedoli. Mapping the Structure of Oxygen-Doped Wurtzite Aluminum Nitride Coatings from Ab Initio Random Structure Search and Experiments. *ACS Applied Materials and Interfaces*, 2021.

[262] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000.

[263] Vin De Silva and Joshua B Tenenbaum. Sparse multidimensional scaling using landmark points. *Tech Report*, 2004.

[264] Laurens Van Der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008.

[265] Gareth A. Tribello and Piero Gasparotto. Using Data-Reduction Techniques to Analyze Biomolecular Trajectories. 2019.

[266] Lauri Himanen, Amber Geurts, Adam Stuart Foster, and Patrick Rinke. Data-Driven Materials Science: Status, Challenges, and Perspectives. *Advanced Science*, 2019.

[267] Teng Zhou, Zhen Song, and Kai Sundmacher. Big Data Creates New Opportunities for Materials Research: A Review on Methods and Applications of Machine Learning for Materials Design. *Engineering*, 2019.

[268] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[269] E. Mateo Marti, Ch Methivier, P. Dubot, and C. M. Pradier. Adsorption of (S)-histidine on Cu(110) and oxygen-covered Cu(110), a combined Fourier transform reflection absorption infrared spectroscopy and force field calculation study. *Journal of Physical Chemistry B*, 2003.

[270] Luiza Buimaga-Iarinca, Calin G. Floare, Adrian Calborean, and Ioan Turcu. DFT study on cysteine adsorption mechanism on Au(111) and Au(110). *AIP Conference Proceedings*, 2013.

[271] S. Blankenburg and W. G. Schmidt. Glutamic acid adsorbed on Ag(110): Direct and indirect molecular interactions. *Journal of Physics Condensed Matter*, 2009.

[272] Tanja Deckert-Gaudig, Eva Rauls, and Volker Deckert. Aromatic amino acid monolayers sandwiched between gold and silver: A combined tip-enhanced Raman and theoretical approach. *Journal of Physical Chemistry C*, 2010.

[273] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Physical Review Letters*, 78(7):1396–1396, 1997.

[274] Wheeler P. Davey. Precision measurements of the lattice constants of twelve common metals. *Physical Review*, 1925.

[275] Philipp Haas, Fabien Tran, and Peter Blaha. Calculation of the lattice constant of solids with semilocal functionals. *Phys. Rev. B*, 79(8):85104, 2009.

[276] J.H. van Lenthe, S. Faas, and J.G. Snijders. Gradients in the ab initio scalar zeroth-order regular approximation (zora) approach. *Chemical Physics Letters*, 328(1):107–112, 2000.

[277] Christoph van Wüllen. Molecular density functional calculations in the regular relativistic approximation: Method, application to coinage metal diatomics, hydrides, fluorides and chlorides, and comparison with first-order relativistic calculations. *The Journal of Chemical Physics*, 109(2):392–399, 1998.

[278] Matti Ropo, Volker Blum, and Carsten Baldauf. Trends for isolated amino acids and dipeptides: Conformation, divalent ion binding, and remarkable similarity of binding to calcium and lead. *Scientific Reports*, 6:35772, 2016.

[279] Michele Ceriotti, Sandip De, and Felix Musil. Glosim package. Code assessed in 2020-01-01.

[280] Michael B. Bolger. Chapter 9–computational techniques in macromolecular structural analysis. pages 433–490. Academic Press, San Diego, 1995.

[281] Donald A. McQuarrie. *Statistical Mechanics*. University Science Books, 2000.

[282] Brent Fultz. Vibrational thermodynamics of materials. *Progress in Materials Science*, 55(4):247–352, 2010.

[283] A Togo and I Tanaka. First principles phonon calculations in materials science. *Scr. Mater.*, 108:1–5, 2015.

[284] Karen Fidanyan. Development version of phonopy. Accessed: 2019-03-01.

[285] Mariana Rossi, Matthias Scheffler, and Volker Blum. Impact of Vibrational Entropy on the Stability of Unsolvated Peptide Helices with Increasing Length. *Journal Of Physical Chemistry B*, 117(18):5574–5584, 2013.

[286] Franziska Schubert, Mariana Rossi, Carsten Baldauf, Kevin Pagel, Stephan Warnke, Gert von Helden, Frank Filsinger, Peter Kupser, Gerard Meijer, Mario Salwiczek, Beate Koksch, Matthias Scheffler, and Volker Blum. Exploring the conformational preferences of 20-residue peptides in isolation: Ac-Ala 19 -Lys + H + vs. Ac-Lys-Ala 19 + H + and the current reach of DFT . *Physical Chemistry Chemical Physics*, 17(11):7373–7385, 2015.

[287] David A Egger, Zhen-Fei Liu, Jeffrey B Neaton, and Leeor Kronik. Reliable Energy Level Alignment at Physisorbed Molecule-Metal Interfaces from Density Functional Theory. *Nano Letters*, 15:2448–2455, 2015.

[288] Zhen-Fei Liu, David A Egger, Sivan Refaely-Abramson, Leeor Kronik, and Jeffrey B Neaton. Energy level alignment at molecule-metal interfaces from an optimally tuned range-separated hybrid functional. *The Journal of Chemical Physics*, 146(9):092326, February 2017.

[289] S.M Barlow, K.J Kitching, S Haq, and N.V Richardson. A study of glycine adsorption on a cu110 surface using reflection absorption infrared spectroscopy. *Surface Science*, 401(3):322–335, 1998.

[290] Susan M. Barlow, Souheila Louafi, Delphine Le Roux, Jamie Williams, Christopher Muryn, Sam Haq, and Rasmita Raval. Supramolecular assembly of strongly chemisorbed size- and shape-defined chiral clusters: S- and r-alanine on cu(110). *Langmuir*, 20(17):7171–7176, 2004.

[291] Christophe Méthivier, Vincent Humblot, and Claire-Marie Pradier. l-methionine adsorption on cu(110), binding and geometry of the amino acid as a function of coverage. *Surface Science*, 632:88–92, 2015.

[292] E Mateo Marti, S.M Barlow, S Haq, and R Raval. Bonding and assembly of the chiral amino acid s-proline on cu(110): the influence of structural rigidity. *Surface Science*, 501(3):191–202, 2002.

[293] E. Mateo Marti, Alida Quash, Ch. Methivier, P. Dubot, and C.M. Pradier. Interaction of s-histidine, an amino acid, with copper and gold surfaces, a comparison based on rairs analyses. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 249(1):85–89, 2004.

[294] Tugce Eralp, Andrey Shavorskiy, Zhasmina V. Zheleva, Georg Held, Nataliya Kalashnyk, Yanxiao Ning, and Trolle R. Linderoth. Global and local expression of chirality in serine on the cu110 surface. *Langmuir*, 26(24):18841–18851, 2010.

[295] Dmitrii Maksimov, Carsten Baldauf, and Mariana Rossi. Database of arg and arg-h^+ adsorbed on cu(111), ag(111) and au(111) in the NOMAD repository.

[296] Ziheng Lu, Bonan Zhu, Benjamin W.B. Shires, David O. Scanlon, and Chris J. Pickard. Ab initio random structure searching for battery cathode materials. *Journal of Chemical Physics*, 2021.

[297] Ask Hjorth Larsen, Jens JØrgen Mortensen, Jakob Blomqvist, Ivano E. Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N. Groves, BjØrk Hammer, Cory Hargus, Eric D. Hermes, Paul C. Jennings, Peter Bjerre Jensen, James Kermode, John R. Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob SchiØtz, Ole Schütt, Mikkel Strange, Kristian S. Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W. Jacobsen. The atomic simulation environment - A Python library for working with atoms. *Journal of Physics Condensed Matter*, 2017.

[298] Encyclopedia of physical science and technology. *Choice Reviews Online*, 2002.

[299] Vincent Frappier, Madeleine Duran, and Amy E. Keating. Erratum to: PixelDB: Protein-peptide complexes annotated with structural conservation of the peptide binding mode: PixelDB: Protein—Peptide Complexes (Protein Science, (2018), 27, 1, (276-285), 10.1002/pro.3320). *Protein Science*, 2018.

[300] Donald B. Johnson. Finding All the Elementary Circuits of a Directed Graph. *SIAM Journal on Computing*, 1975.

[301] Charles F.F. Karney. Quaternions in molecular modeling. *Journal of Molecular Graphics and Modelling*, 25(5):595 – 604, 2007.

[302] Soohyung Park, Haiyuan Wang, Thorsten Schultz, Dongguen Shin, Ruslan Ovsyannikov, Marios Zacharias, Dmitrii Maksimov, Matthias Meissner, Yuri Hasegawa, Takuma Yamaguchi, Satoshi Kera, Areej Aljarb, Mariam Hakami, Lain Jong Li, Vincent

Tung, Patrick Amsalem, Mariana Rossi, and Norbert Koch. Temperature-Dependent Electronic Ground-State Charge Transfer in van der Waals Heterostructures. *Advanced Materials*, 2021.

[303] Lincan Fang, Esko Makkonen, Milica Todorovic, Patrick Rinke, and Xi Chen. Efficient cysteine conformer search with bayesian optimization. 2020.

[304] Letif Mones, Christoph Ortner, and Gábor Csányi. Preconditioners for the geometry optimisation and saddle point search of molecular systems. *Scientific Reports*, 2018.

[305] X. Chu and A. Dalgarno. Linear response time-dependent density functional theory for van der waals coefficients. *The Journal of Chemical Physics*, 121(9):4083–4088, 2004.

[306] J Mitroy, M S Safronova, and Charles W Clark. Theory and applications of atomic and ionic polarizabilities. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 43(20):202001, oct 2010.

[307] Database of lennard-jones clusters. http://doye.chem.ox.ac.uk/jon/structures/LJ.html.

[308] K. W. Jacobsen, P. Stoltze, and J. K. Nørskov. A semi-empirical effective medium theory for metals and alloys. *Surface Science*, 1996.

[309] Steven J. Stuart, Alan B. Tutein, and Judith A. Harrison. A reactive potential for hydrocarbons with intermolecular interactions. *Journal of Chemical Physics*, 2000.

[310] Donald W. Brenner, Olga A. Shenderova, Judith A. Harrison, Steven J. Stuart, Boris Ni, and Susan B. Sinnott. A second-generation reactive empirical bond order (REBO) potential energy expression for hydrocarbons. *Journal of Physics Condensed Matter*, 2002.

[311] Murray S. Daw and M. I. Baskes. Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals. *Physical Review B*, 1984.

[312] Murray S. Daw, Stephen M. Foiles, and Michael I. Baskes. The embedded-atom method: a review of theory and applications. *Materials Science Reports*, 1993.

[313] I. Stensgaard. Adsorption of di-L-alanine on Cu(1 1 0) investigated with scanning tunneling microscopy. *Surface Science*, 2003.

[314] S. M. Barlow, S. Haq, and R. Raval. Bonding, organization, and dynamical growth behavior of tripeptides on a defined metal surface: Tri-L-alanine and Tri-L-leucine on Cu{100}. *Langmuir*, 2001.

Dmitrii Maksimov

Github: maksimovdmitrii
LinkedIn: dmitrii-maksimov-7ab0a887
Google Scholar: Dmitrii Maksimov
ResearchGate: Dmitrii_Maksimov
ORCID: 0000-0003-4448-8848
ScopusID: 56850160300
ResearcherID: M-7271-2013

Education

- **ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE (EPFL)** — Lausanne, Switzerland
  PhD Materials Science and Engineering — *June 2018 – July 2022*
- **St. Petersburg State University** — St. Petersburg, Russia
  Master of Physics - Biophysics — *September 2013 – July 2015*
- **Ural Federal University** — Yekaterinburg, Russia
  Bachelor of Physics - Theoretical Physics — *September 2009 – July 2013*

Working experience

- **Max Planck Institute for the Structure and Dynamics of Matter** — Hamburg, Germany
  Visiting researcher — *February 2020 – July 2022*
- **Fritz Haber Institute of the Max Planck Society** — Berlin, Germany
  Doctoral student — *February 2017 – July 2022*
- **St. Petersburg State University** — St. Petersburg, Russia
  Researcher — *May 2014 – January 2017*

Projects

- **Generation and Search (GenSec) package**: Created from scratch open-source structure search code for organic/inorganic hybrid materials. The package is designed to unify many electronic structure programs and produce material databases in a unified format suitable for further machine learning projects and training of the empirical potentials. One of the package's strengths is the parallel utilizing of resources on high-performance computational platforms. Tech: Python, ASE (July 2021 - Present time). Publication is in preparation.
- **Structure search of organic molecules adsorbed on vacancies of MoS_2 surface**: Preparation of candidate structures with GenSec followed by *ab initio* molecular dynamics simulation. Expected theoretical results should correspond to experimental scanning tunelling microscope (STM) images (Jan 2022 - Present time). Publication is in preparation.
- **Structure search of triglycine molecule with different protonation states**: Different conformers with different protonation states of triglycine molecule in isolation were investigated using *ab initio* structure search. Modelled theoretical properties for low-energy conformers are investigated corresponding to experimental near edge X-ray absorption fine structure (NEXAFS) spectra (September 2021 - Present time). Publication is in preparation.
- **Temperature-Dependent Electronic Ground-State Charge Transfer in van der Waals Heterostructures**: Investigation of self-assembly candidates of F4-TCNQ and F6-TCNQ molecules adsorbed on MoS_2 surface (July 2020 – July 2021).
- **The conformational space of a flexible amino acid at metallic surfaces**: Generation of the *ab initio* (PBE+vdW^{surf}) database of the Arg and Arg-H$^+$ molecules adsorbed on different metallic surfaces and analysis of the structure-property relationships employing unsupervised dimensionality reduction techniques (multidimensional scaling) using smooth overlap of atomic positions kernel molecular descriptors (March 2018 – June 2020).
- **Investigation of stacking geometries of nucleobases and their complexes with metal clusters**: Analysis of available databases and investigation of excited-state properties of nucleobases complexes with metal clusters (March 2015 – December 2016).

Publications

Temperature-Dependent Electronic Ground-State Charge Transfer in van der Waals Heterostructures, S. Park et al., Advanced Materials 33 (29), 2008677, (2021)
The conformational space of a flexible amino acid at metallic surfaces, D. Maksimov, C. Baldauf, M. Rossi, International Journal of Quantum Chemistry 121 (3), e26284, (2021)
Excitation spectra of Ag3–DNA bases complexes: a benchmark study, D. Maksimov, V. Pomogaev, A. Kononov, Chemical Physics Letters 673, 11-18, (2017)
Ag–DNA emitter: metal nanorod or supramolecular complex?, R. Ramazanov et al., The Journal of Physical Chemistry Letters 7 (18), 3560-3566, (2016)
Noncanonical stacking geometries of nucleobases as a preferred target for solar radiation, R. Ramazanov, D. Maksimov, A. Kononov, Journal of the American Chemical Society 137 (36), 11656-11665, (2015)

Skills Summary

- **Languages**: Python, Bash
- **Frameworks**: Scikit, Matplotlib, Pandas, Plotly
- **Tools**: GIT, ASE, PyMol, Blender, Jmol
- **Electronic structure packages**: FHI-aims, ORCA, Turbomole, GAMESS, Firefly

Teaching

- **Summer schools tutoring** 2017 – 2019
  Tutor at workshops
  - **Hands-on Workshop Density-Functional Theory and Beyond: Accuracy, Efficiency and Reproducibility in Computational Materials Science**: Humboldt University, Berlin, Germany, July 31 to August 11, 2017
  - **Hands-on DFT and Beyond: Frontiers of advanced electronic structure and Molecular Dynamics Methods**: Peking University, Beijing, China, July 30th to August 10th, 2018
  - **Hands-on DFT and Beyond: High-Throughput Screening and Big-Data Analytics, Towards Exascale Computational Materials Science**: University of Barcelona, Barcelona, Spain, August 26th to September 6th, 2019