Machine Learning for Structure Search of Ligand-protected Nanoclusters

Lincan Fang
Machine Learning for Structure Search of Ligand-protected Nanoclusters

Lincan Fang

A doctoral thesis completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall U3 (Otakaari 1) of the school on 23 February 2024 at 13:00.

Aalto University
School of Science
Department of Applied Physics
Supervising professor
Professor Patrick Rinke, Aalto University, Finland

Thesis advisor
Professor Chen Xi, Lanzhou University, China

Preliminary examiners
Academy Professor Hannu Häkkinen, University of Jyväskylä, Finland
Full Professor Jaakko Eemeli Akola, Norwegian University, Norway

Opponent
Associate Professor Olga López Acevedo, Institute of Physics Universidad de Antioquia, Colombia

Aalto University publication series
DOCTORAL THESES 45/2024

© 2024 Lincan Fang

ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Unigrafia Oy
Helsinki 2024

Finland
Abstract

Understanding the atomic structures of ligand-protected nanoclusters is essential for their application in various fields. These structures not only determine the physical and chemical properties of ligand-protected nanoclusters but also play a crucial role in their stability and reactivity. Knowing the precise atomic structures allows us to tailor nanoclusters for specific functions. However, because of the extraordinarily high dimensionality of the search space which encompasses an exceptionally large number of all potential structures, it is difficult to use quantum mechanical methods, such as the density functional theory, to find the low-energy structures of ligand-protected nanoclusters. On this point, the structure search of ligand-protected nanoclusters could be more efficient and accurate by utilizing machine learning methods.

In this dissertation, I developed machine learning methods to search the atomic structures of ligand-protected nanoclusters by decomposing the problem into three steps. For the first step, I developed a molecular conformer search procedure based on Bayesian optimization to search the structures of isolated molecules. Using four amino acids as examples, I showed that the procedure is both efficient and accurate. For the second step, I modified the procedure to search the structures of a single ligand on a nanocluster. I also developed and tested strategies to avoid steric clashes between a ligand and cluster parts. Moreover, I tested and demonstrated our modified procedure by searching structures for a cysteine molecule on a well-studied gold-thiolate cluster. As a result, I found that cysteine conformers in a cluster inherit the hydrogen bond types from isolated conformers, while the energy rankings and spacings between the conformers are reordered. In the final step, I applied a machine learning method based on kernel rigid regression (KRR) models to relax the structures of ligand-protected nanoclusters. Moreover, I used an active learning workflow to enhance the relaxation performance of the KRR models. To test and demonstrate our method, I applied it to search structures of Au_{18}(Cys)_{18}. We found that the low-energy structures with II-type hydrogen bonds (OH-\cdot\cdot\cdot N, OH from trans-COOH and N from NH_{2}) are dominant and the different configurations of the ligand layer indeed influence the properties of the clusters.

Keywords machine learning, Bayesian optimization, density-functional theory, active learning, nanocluster, structure search
Preface

The research presented in this doctoral thesis was carried out in the Computational Electronic Structure Theory (CEST) group at the Department of Applied Physics of Aalto University. I would like to express my deepest gratitude to my supervisor, Professor Patrick Rinke, for his invaluable guidance and assistance during this research project. Thank you for the guidance, advice, and encouragement when I face difficult problems in this research project. I am grateful to my advisor, Professor Xi Chen, who always helps, encourages, and guides me through my research.

I wish to thank Professor Olga López Acevedo for acting as my opponent. I also thank the pre-examiners of this thesis, Professor Hannu Häkkinen and Professor Jaakko Eemeli Akola, for examining and commenting on my thesis.

I thank all the CEST group members for helping me with my research and creating a good work environment for me. I am grateful to the kind people on the 4th floor for making the pleasant environment.

I am grateful to Finland for the nice people, the good summertime, and the fact that it's peaceful.

Aalto, February 6, 2024,

Lincan Fang
Contents

Preface 3

Contents 5

List of Publications 9

Author's Contribution 11

1. Introduction 13
   1.1 Ligand-protected nanoclusters ......................... 13
   1.2 Machine learning and structure search ................. 16
   1.3 Search objectives .................................. 17
      1.3.1 Isolated molecular structure search ............. 18
      1.3.2 Structure search for one single ligand on a cluster 19
      1.3.3 Search structures for the ligand-protected nanocluster .......... 20
   1.4 Thesis structure .................................. 21

2. Methods 23
   2.1 Density functional theory ........................... 23
      2.1.1 The Hamiltonian of many-body system .......... 23
      2.1.2 Born-Oppenheimer approximation ................. 24
      2.1.3 Hohenberg-Kohn theorem ........................ 25
      2.1.4 The Kohn-Sham method ........................... 26
   2.2 Classical force field ............................... 27
      2.2.1 AMBER force field ............................. 28
      2.2.2 Energy minimization ............................ 29
   2.3 Bayesian optimization ............................... 29
   2.4 Kernel rigid regression ............................. 32
      2.4.1 Rigid regression ............................... 32
      2.4.2 Kernel ridge regression ......................... 33
3. Computational details 37
   3.1 FHI-aims code usage 37
   3.2 GROMACS code 38
   3.3 BOSS code 38

4. Efficient amino acid conformer search with Bayesian optimization 41
   4.1 BOSS-based molecular conformer search 42
   4.2 2-D and 5-D cysteine conformer search 43
   4.3 Conformational energy hierarchy of cysteine 45
   4.4 Conformer search for serine, aspartic, and tryptophan 47
   4.5 Computational efficiency of the search procedure 49

5. Searching the structures of a cysteine adsorbate on a gold cluster with Bayesian optimization 51
   5.1 Two systems of a cysteine adsorbate on a Au_{25} cluster 51
   5.2 Three strategies for addressing steric clashes between the ligand and the cluster 52
   5.3 Evaluating the performance of the three strategies 54
   5.4 The predicted low-energy structures 55

6. Applying machine learning to accelerate the structure search for the ligand-protected nanoclusters 59
   6.1 Constructing the data pool of the Au_{25}(Cys)_{18}^{-1} 59
   6.2 Using KRR model to optimize the structures of Au_{25}(Cys)_{18}^{-1} 61
   6.3 Fixing the bond angles of cysteine by using force field methods 62
   6.4 Increasing the diversity of training data by active learning methods 63
   6.5 The effects of the active learning workflow and adding force field simulations 64
   6.6 Evaluating the reliability of the optimal KRR model relaxations 65
   6.7 Results and discussion 67
      6.7.1 Structure analysis of the low-energy structures 68
      6.7.2 Electronic property analysis of the low-energy structures 69

7. Summary and Outlook 71
   7.1 Summary 71
   7.2 Outlook 72

References 75
This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.


Author’s Contribution

Publication I: “Efficient Amino Acid Conformer Search with Bayesian Optimization”

X. C. and P. R. conceptualized the work. L. F. carried out all the DFT simulations under the supervision of P. R. and X. C.. E. M. performed CCSD(T) simulations and M. T. contributed to the method development. All authors participated in the manuscript writing.

Publication II: “Exploring the Conformers of an Organic Molecule on a Metal Cluster with Bayesian Optimization”

X. C. and P. R. conceptualized the work. L. F. carried out all the simulations under the supervision of P. R. and X. C.. X. G. contributed to the data analysis, and M. T. contributed to the method development. All authors participated in the manuscript writing.

Publication III: “Machine-learning accelerated structure search for ligand-protected clusters”

X. C. and P. R. conceptualized the work. L. F. carried out all the simulations under the supervision of P.R. and X. C.. J. L. contributed to the method development. All authors participated in the manuscript writing.
1. Introduction

The world is full of diverse materials that have been used since the beginning of human civilization. These materials have played a crucial role in shaping history, with each era marked by the predominant materials utilized at that time [1]. As humanity has evolved, our ability to manipulate materials has also developed, leading to the creation of various technologies. Moreover, materials have been instrumental in transforming the natural world and enabling humans to equip themselves from the Stone Age to the Digital Age.

In modern times, humans are perpetually developing and improving materials with an extensive range of functional properties to tackle various issues and apply them in different fields such as nanotechnology [2], healthcare [3], automobile industries manufacturing [4], chips [5], biotechnology [6], clean energy [7], food engineering [8], chemical engineering [9], and medicine production [10]. As these applications grow more intricate and challenging, there is an escalating demand for more effective and efficient approaches to designing materials with the desired functions. Thus, understanding the atomic structures of materials is essential in the design of materials. Given the importance of this topic, my research focuses on developing machine learning (ML) methods to search the atomic structures of molecules and molecule-metal clusters.

1.1 Ligand-protected nanoclusters

Nanoclusters usually contain tens to hundreds of atoms and have a size that is typically less than two nanometers. They act as a bridge between single atoms and bulk materials and are smaller than nanoparticles [11, 12, 13]. Inorganic metal nanoclusters can have one or multiple elements, such as gold, silver, copper, or other metallic elements [14, 15]. Moreover, organic-inorganic hybrid nanoclusters, such as ligand-protected nanoclusters, are composed of not only metal elements but also some organic ligand molecules. Due to their small size, the electrons of nan-
Nanoclusters are confined within a limited space, referred to as the quantum confinement effect. The confinement creates discrete energy levels and thus gives nanoclusters unique electronic, optical, and chemical properties distinct from those of bulk materials [16, 17, 18, 19, 20, 21, 22, 23].

Bare metal nanoclusters are usually unstable because they have many exposed surface atoms with a surface energy that is higher than the energy of the atoms in the interior of the cluster, which causes the whole cluster to be energetically unfavorable [24, 25, 26, 27, 28, 29]. To reduce their surface energy, the metal atoms in the bare cluster tend to reconstruct themselves, causing the atoms to rearrange and form new bonds with neighboring atoms. These reconstructions lead to the formation of new structures, such as a crystal or a solid, that are more stable than the bare nanoclusters [30, 31, 32]. In addition, because of the non-bonding orbitals of the surface atoms, bare nanoclusters usually are highly reactive, making it easy for them to react with molecules or other atoms. Bonding with molecules or atoms may lead to the reconstruction of surfaces or even the dissolution of the nanocluster.

The instability in bare nanoclusters restricts their practical application. Hence, metal nanoclusters have been stabilized by surfaces, proteins, polymers, or ligand molecules [33, 34, 35, 36, 37]. Ligand molecules are small organic molecules, such as amino acids [38], phosphides [39], and thiolates [40], which bind to the metal clusters through chemical bonds. These ligands can form a protective monolayer around the bare nanocluster, preventing the metal core from agglomeration, chemical reactions, structure reconstructions, and oxidation. These hybrid nanoclusters are referred to as ligand-protected nanoclusters. The initial synthesis of ligand-protected nanoclusters was achieved by Malatesta et al. using phosphine to stabilize Au$_{11}$ and obtained stable crystal structures in the 1960s [41]. Subsequently, more and more ligand-protected nanoclusters were synthesized and characterized, such as phosphine (PR$_3$)- and thiolate (SR)-protected nanoclusters [42, 43, 44, 45, 46]. A pivotal moment in cluster studies came when Kornberg’s group obtained large single crystal structures of the thiolate-protected Au$_{102}$ cluster, which opened the way to X-ray structure determination [47].

The physical and chemical properties of ligand-protected nanoclusters are greatly affected by the metal core and ligands [48]. For instance, it has been reported that electronic structures of gold ligand-protected nanoclusters are highly sensitive to the symmetry of gold cores which affects the optical absorption [49]. A further example shows that utilizing specific capping ligands can effectively control the size, shape, and exposure surface of ligand-protected nanoclusters [50, 51]. Moreover, previous studies have shown that using ligands can tune the catalytic activities of ligand-protected nanoclusters by generating different coordination strengths between the metal atoms in the core and ligands [52].
Introduction

By combining different ligands and cores, ligand-protected nanoclusters exhibit diverse properties that can be precisely tailored for specific functions (Fig. 1.1). Utilizing the catalytic activity of the high surface area and customizable composition in ligand-protected nanoclusters, effective catalysts have been designed and developed for chemical transformations [53, 54]. Owing to the unique electrical and optical properties of ligand-protected nanoclusters, these nanoclusters hold promise as next-generation optical devices [55, 56]. Moreover, ligand-protected nanoclusters find utility in biological applications, where their biocompatibility and tunable surface properties facilitate their use for biolabeling [57], cancer treatment [58], drug screening [59], drug delivery [60], etc. Applications of ligand-protected nanoclusters involving energy conversion [61], environmental remediation [62], and sensing [63] show their potential to promote the development of numerous fields and new multifunctional nanomaterials.

One interesting aspect of ligand-protected nanoclusters is that stable ligand-protected nanoclusters have certain numbers of metal atoms and ligands [64, 65, 66]. For example, \( \text{Au}_{25}(\text{SR})_{18}^{-1}, \text{Au}_{38}(\text{SR})_{24}, \text{Au}_{102}(\text{SR})_{44}, \text{Au}_{144}(\text{SR})_{60} \) and \( \text{Au}_{111}(\text{PR}_3)_7\text{X}_3 \) (SR: thiolate, PR\(_3\): phosphine, and X:halide) show high stability. Walter et al. provide a theory to explain this stability in ligand-protected nanoclusters [67]. In their theory, the ligands may either withdraw electrons (i.e., localize electrons into covalent bonds) from the metal core or attach as weak Lewis base ligands that coordinate to the core surface by dative bonds (i.e., they do not withdraw electrons from the core metal atoms). The number of delocalized electrons in ligand-protected nanoclusters can be calculated by \( n^* = N v_A - M - z \), where \( v_A \) is the atomic valence of the metal atom, \( M \) is the number of ligands that withdraw electrons, and \( z \) is the charge of the whole system. The delocalized electrons in the clusters form electronic shells that are similar to the electronic arrangement in atoms. If \( n^* \) fills one electronic shell, the
ligand-protected nanoclusters have enhanced stability. For the spherical ligand-protected nanoclusters \( n^* = 2, 8, 18, 34, 58, \ldots \), and the superatomic orbital is \( 1S^21P^61D^{10}2S^21F^{14}2P^61G^{18} \ldots \) (\( S - P - D - F - G - \) represent the angular-momentum characters). For example, the \( \text{Au}_{102}(SR)_{44} \) corresponds to \( n^* = 58 \) closing the \( 1G \) shell, and \( \text{Au}_{25}(SR)_{18}^{-1} \) corresponds to \( n^* = 8 \) closing the \( 1P \) shell.

Thiolate-protected gold nanoclusters are the most well-studied ligand-protected nanoclusters [46, 68, 69, 70]. They usually have a high-symmetrical gold core and a ligand-protected layer containing several staple structures. The two most common staples are -RS-Au-RS-Au-RS- and -RS-Au-RS-. The larger thiolate-protected gold nanoclusters usually have more short staples -RS-Au-RS-, whereas the smaller ones tend to have long staples. For example, \( \text{Au}_{102}(SR)_{44} \) has 19 -RS-Au-RS- and two -RS-Au-RS-Au-RS-, but \( \text{Au}_{25}(SR)_{18}^{-1} \) contains six longer staples -RS-Au-RS-Au-RS-. Various small organic ligand molecules, such as \( \text{SC}_{12} \text{H}_{25} \), \( \text{C}_{2} \text{H}_{4} \text{Ph} \) (Ph: phenyl group), \( -\text{C}_{2} \text{H}_{4} \text{Ph} \), 4-mercaptobenzoic acid (p-MBA), and glutathione, can bind with sulfur atoms to generate diverse protected monolayers, which makes thiolate-protected gold nanoclusters an exciting material to study.

Up until now, the design and optimization of functional ligand-protected nanoclusters remain fundamentally challenging because the desired properties often depend on the microscopic structure of the system, which is typically unknown. The structures of the metal part and metal-molecule interface can be obtained from experimental data, previously reported crystal structures, chemical intuition, or data-driven methods [71, 72, 73, 74, 75, 76]. However, the detection and determination of the structure of the organic ligand molecules are the main obstacles [77]. Since ligands in ligand-protected nanoclusters are usually organic small molecules and their structures are sensitive to the chemical environment surrounding them, directly detecting their atomic structures with traditional experimental and theoretical methods is difficult.

1.2 Machine learning and structure search

In recent years, artificial intelligence and ML methods, such as kernel ridge regression (KRR) [78], Gaussian process regression (GPR) [79, 80], genetic algorithms [81, 82], and artificial neural networks [83] have been increasingly applied in material science, such as the prediction of new stable materials [84, 85], the calculation of numerous properties [86, 87], and the speeding up of first-principle calculations [88].

A multitude of already successful ML applications in nanoclusters can be found. For instance, Pihlajamäki et al. employed distance-based ML methods to investigate the structural and energetic properties of \( \text{Au}_{38}(\text{SCH}_3)_{24} \) [89]. Malola et al. developed a data-driven approach that uses the informa-
Introduction

Figure 1.2. Illustration of the three systems for my work. Cysteine, \( \text{Au}_{25}(\text{CH}_3)_{17}\text{Cys}^{-1} \) (Cys: cysteine), and \( \text{Au}_{25}(\text{Cys})_{18}^{-1} \) are shown here as examples for the three systems. The gold color is used for gold atoms, red for oxygen, yellow for sulfur, gray for carbon, blue for nitrogen, and white for hydrogen.

tion about the local chemical environments of atoms in experimental data to predict the metal-ligand interface of gold and silver ligand-protected nanoclusters [76]. Jäger and his colleagues applied KRR methods to study hydrogen adsorption on nanoclusters [90]. Subramanian et al. employed active learning methods based on a neural network model to investigate the structures of bare gold clusters [91].

However, my work focuses on a different research topic, the atomic structure of ligand layers in the ligand-protected nanoclusters (Sec. 1.3). Therefore, the ML methods mentioned in the examples are unsuitable for my research. In this thesis, I developed new ML methods to tackle our studied systems.

1.3 Search objectives

This thesis aims to develop ML methods for searching the atomic structures of ligand layers in ligand-protected nanoclusters. My approach is from simple to complex, splitting the whole system of ligand-protected nanoclusters into three: isolated molecules, a single ligand molecule on one cluster, and the ligand-protected layer of ligand-protected nanoclusters (Fig. 1.2). I first developed a molecular conformer search procedure to efficiently and accurately search structures for isolated molecules (Publication I). Then, I modified this molecular conformer search procedure to search the structures of a single ligand molecule on a cluster (Publication II). Finally, based on the results of the single ligand structure search and combined with active learning methods, I trained a KRR model and utilized this model to explore the ligand-protected nanocluster structures (Publication III). Here, I introduce the challenges and objectives of each study.
1.3.1 Isolated molecular structure search

A conformer is a distinct conformation (i.e., the arrangement of atoms of a molecule) at a minimum of the potential energy surface (PES) of the molecule. Generally, one molecule has several stable molecular conformers, each with different properties\cite{92, 93, 94}. The task of molecular conformer search is to explore the conformational space of the molecule. However, identifying the low-energy conformers and determining their energy ranking is a persistent challenge in molecular modeling because of the high-dimensional configurational phase as well as the flexibility of molecules.

The first challenge of molecular conformer search is sufficient sampling of the configurational space. Even for a relatively small molecule, the conformational space (i.e., bond lengths, bond angles, and dihedral angles) are enormous \cite{95, 96, 97}. For this reason, a reasonable dimensionality reduction is required. Because the bond lengths and angles are rigid and the different conformers mainly originate from the flexible rotational groups, most search methods focus on sampling the dihedral angles while keeping the bond lengths and angles fixed \cite{98}.

In recent years, various methods and tools have been developed to efficiently generate diverse molecular structures\cite{99, 100, 101, 102, 103, 104}. These methods can be classified into systematic and stochastic. Systematic methods use a grid to sample all possible dihedral angles. They are intuitive and deterministic, but limited to small molecules. Stochastic methods, on the other hand, are based on random algorithms, such as distance selection, genetic algorithms \cite{105}, Monte Carlo annealing\cite{106, 107}, minima hopping\cite{108}, and basin hopping\cite{109, 110}. They can be applied to the sampling of dihedral angles for larger molecules, but the predicted conformers may vary each time. Additionally, knowledge-based methods from libraries\cite{111, 112}, such as the Cambridge Structure Database\cite{113}, and the Protein Data Bank\cite{114}, can increase the efficiency of dihedral angle sampling; however, they need to be combined with systematic or stochastic methods.

The second challenge in molecular structure search is to calculate the energy for the sampling structures with sufficient accuracy. In this respect, it is difficult to ensure adequate accuracy with a reasonable computational cost. For example, force field-based methods are efficient, but the accuracy is often not high enough for conformer identification and ranking. By contrast, quantum chemistry methods are accurate but too computationally expensive to provide to provide energies for all configurations produced in the search. To balance efficiency and accuracy, hierarchical methods have been developed\cite{115, 116, 117}. This approach uses low-accuracy calculations to scan the configurational space, identifying candidate structures that then must be refined with a method with higher accuracy. To
avoid missing the true low-energy conformers, a large portion of the conformational space must be sampled by low-accuracy methods, and many structures need to be optimized at a higher level.

To address these two challenges, in the study of search structures for isolated molecules (Publication I), I developed a molecular conformer search procedure that combines active learning Bayesian optimization (BO) and quantum chemistry methods to accurately and efficiently identify molecular conformers and predict their energies. I used four amino acids (cysteine, serine, tryptophan, and aspartic acid) to test and demonstrate the efficiency, accuracy, generalizability, and transferability of the developed procedure. More details regarding this can be found in Chapter 4 and Publication I.

I chose the four amino acid molecules in my thesis for several reasons. Firstly, the chosen amino acids play critical roles in biological functions. Secondly, these amino acids have four to six rotational groups that are accessible for the BO ML method. Thirdly, previous studies have reported the structures and energy rankings of these amino acids, providing reference results for our predicted results of the structures search for isolated molecules.

### 1.3.2 Structure search for one single ligand on a cluster

When a molecule bonds to the metal core of a cluster, it is surrounded by metal atoms and other ligands (as shown in Fig. 1.3a). Searching the structure of a ligand on a cluster with active learning may lead to steric clashes between the ligand and cluster (Fig. 1.3b), presenting technical challenges for active learning. Therefore, the procedure I developed for the isolated molecular structure search (Publication I) can not be directly applied to search the structure of a ligand in such systems. For a steric
clash structure, the DFT energy calculation returns an error or extremely high energy. The error means no energy input in the active learning procedure, which directly causes the procedure to stop. Introducing very high energies into the energy landscape produces a large variance in the model, which causes the active learning algorithm repeatedly sample the unimportant high-energy region in the search space and lead to an inaccurate prediction of the low-energy PES.

To address the issue of steric clashes, I developed three strategies to avoid sampling nonphysical structures during active learning (Publication II). I used a cysteine molecule on a well-studied thiolate-protected $\text{Au}_{25}$ cluster as the model system to present and evaluate the three strategies. Next, I investigated how a cluster affects the structures and energies of the absorbed single molecule. More details regarding this can be found in Chapter 5 and Publication II.

I chose the thiolate-protected $\text{Au}_{25}$ cluster for several reasons. Firstly, the thiolate-protected $\text{Au}_{25}$ cluster has been well-studied. Several studies for Au$_{25}$ nanoclusters can be found in the literature. For example, Akola et al. first theoretically predicted the atomic structure of $\text{Au}_{25}(\text{SR})_{18}^{-1}$ [72], and the structure was experimentally confirmed by Heaven et al. [71]. Rojas-Cervellera et al. investigated how a water solvent and glutathione ligands affect the structure and electronic properties of $\text{Au}_{25}(\text{SR})_{18}^{-1}$ using a QM/MM method [118]. They also revealed the molecular mechanism of the ligand exchange reaction of an antibody on a glutathione-coated $\text{Au}_{25}$ clusters [119]. Lopez-Acevedo et al. studied the ambient CO oxidation catalyzed by ligand-protected $\text{Au}_{25}$ clusters [120]. From these experimental and theoretical studies, we can easily obtain the geometric structures of the $\text{Au}_{25}\text{S}_{18}$ part. Secondly, the cluster offers an ideal situation for cysteine in a complicated, confined environment. Third, the cluster has two inequivalent $S$ sites, and this allows us to study how different surrounding environments affect the adsorbate structures of cysteine. Lastly, we can compare the adsorbate structures of cysteine with the conformers in the gas phase to elucidate confinement and proximity effects.

1.3.3 Search structures for the ligand-protected nanocluster

The ligand-protected layer which contains multiple ligands bound to the metal core, has a complex structure. In contrast to an isolated molecule or a single ligand on a cluster, the conformational space of the ligand layer has a much higher number of dimensions, which exacerbates the challenges in the structure search, such as sufficient sampling, accurate energy calculations, and avoidance of steric clashes. For instance, in $\text{Au}_{25}(\text{Cys})_{18}^{-1}$ (Cys: cysteine), if five rotational groups are used to define the search space of Cys, the entire search space for the ligand-protected layer is 90 dimensions. The structure search for such a high-dimensional system
is beyond the ability of BO. To search the structure of the ligand-protected layer, other ML structure search methods need to be developed.

I use eighteen cysteine molecules to replace SCH$_3$ in Au$_{25}$(SCH$_3$)$_{18}^{-1}$ to construct the study model Au$_{25}$(Cys)$_{18}^{-1}$. To sufficiently sample the configurational space of such a high-dimensional system, many structures need to be generated and relaxed. Then, I trained a KRR model to relax these structures. After obtaining low-energy structures, I analyzed their energy distributions, structure features, and electronic properties. More details regarding this can be found in Chapter 6 and Publication III.

1.4 Thesis structure

The dissertation is structured as follows. Chapter 2 reviews the various computational methods employed in the study, including density functional theory, force-field theory, Bayesian optimization, and kernel rigid regression. Chapter 3 outlines the keyword settings used in the codes/programs and offers a brief introduction to each code. Chapter 4 discusses the results obtained in Publication I where a molecular conformer search procedure was demonstrated using four amino acids. In Chapter 5, I present the three strategies for avoiding steric clashes during active learning and the results of the cysteine ligand structure search. Chapter 6 focuses on the results from Publication III, which demonstrates how to apply a machine learning method to accelerate the structure search for the ligand-protected nanoclusters. Finally, a general summary and outlook are presented in Chapter 7.
2. Methods

This chapter provides an overview of the computational methods utilized in this thesis. Firstly, I review the methodologies of density functional theory (DFT) and force fields, which were employed for the energy and force calculation. Following that, a concise explanation is provided for the machine learning techniques applied, namely Bayesian optimization (BO) and kernel ridge regression (KRR). This chapter aims to familiarize the reader with the fundamental computational methods employed throughout the research.

2.1 Density functional theory

The non-relativistic Schrödinger equation is the foundation for investigating the electronic structures of matter. Among the computational quantum mechanical simulation methods, DFT has been widely employed to determine the ground state electron density and equilibrium geometry. DFT is a first-principle method as it only requires fundamental physical quantities. In this section, a brief review of the density functional theory is provided [121, 122, 123, 124].

2.1.1 The Hamiltonian of many-body system

The Schrödinger equation for a non-relativistic and time-independent system is

$$\hat{H} \Psi(r_1, r_2, ..., r_n; \vec{R}_1, \vec{R}_2, ..., \vec{R}_n) = E \Psi(r_1, r_2, ..., r_n; \vec{R}_1, \vec{R}_2, ..., \vec{R}_n),$$

(2.1)

where $\Psi$ and $E$ are the wave function and the total energy of the system, respectively. $r_1, r_2, ..., r_n$ and $\vec{R}_1, \vec{R}_2, ..., \vec{R}_n$ respectively denote the Cartesian coordinates of electrons and nuclei in the system, which consists of $n$
Methods

electrons and \( N \) nuclei. \( \hat{H} \) is the Hamiltonian of the system and defined as

\[
\hat{H} = \hat{T}_{\text{nuc}} + \hat{T}_e + V_{\text{e-e}} + V_{\text{nuc-nuc}} + V_{\text{nuc-e}}
\]

\[
= -\sum_{i=1}^{n} \frac{\hbar^2}{2m_i} \nabla_i^2 - \sum_{i=1}^{N} \frac{\hbar^2}{2M_i} \nabla_i^2 + \frac{1}{2} \sum_{i,j \neq i}^{n} \frac{1}{4\pi\epsilon_0} \frac{e^2}{|\vec{r}_i - \vec{r}_j|} + \frac{1}{2} \sum_{i,j}^{N} \frac{1}{4\pi\epsilon_0} \frac{(Z_i e)(Z_j e)}{|\vec{R}_i - \vec{R}_j|} - \sum_{i=1}^{n} \sum_{j=1}^{N} \frac{1}{4\pi\epsilon_0} \frac{Z_i e^2}{|\vec{R}_i - \vec{R}_j|},
\]

(2.2)

where \( \hat{T}_{\text{nuc}}, \hat{T}_e, V_{\text{e-e}}, V_{\text{nuc-nuc}}, \) and \( V_{\text{nuc-e}} \) are the kinetic energy of \( N \) nuclei, the kinetic energy of \( n \) electrons, the interaction among electrons, the interaction among nuclei, and the interaction between nuclei and electrons, respectively. The reduced Planck constant \( \hbar \) and the permittivity of vacuum \( \epsilon_0 \) are also involved. The masses of the \( i \)th atomic nucleus and \( i \)th electron are denoted as \( M_i \) and \( m_i \), respectively. \( \vec{R}_i \) and \( \vec{r}_i \) represent the coordinates of the \( i \)th nucleus and the \( i \)th electron. Solving the many-body Schrödinger equation allows us to determine the ground-state properties of a system, but it is limited to small systems in practice. Practical approximations become necessary.

2.1.2 Born-Oppenheimer approximation

Nuclei are \( \sim 1800 \) times heavier than electrons, and hence electrons move much faster than nuclei due to momentum conservation. The Born-Oppenheimer approximation assumes that the nuclei are stationary at the moment of electronic motion, allowing the electronic motion and nuclear motion to be separated. As a result, the wavefunction of the system can be written as

\[
\Psi(\vec{r}_1, \vec{r}_2, ..., \vec{r}_n; \vec{R}_1, \vec{R}_2, ..., \vec{R}_n) = \Psi^{\text{nuc}} \Psi^e,
\]

(2.3)

where \( \Psi^{\text{nuc}} \) is the nuclear wavefunction and \( \Psi^e \) is the electronic wavefunction. \( \Psi^{\text{nuc}} \) is dependent on the positions of the nuclei and can be expressed as \( \Psi^{\text{nuc}}(\vec{R}_1, \vec{R}_2, ..., \vec{R}_n) \). By contrast, \( \Psi^e \) depends on the coordinates of both electrons and nuclei and can be represented as \( \Psi^e(\vec{r}_1, \vec{r}_2, ..., \vec{r}_n; \vec{R}_1, \vec{R}_2, ..., \vec{R}_n) \). Thus, Equation 2.1 can be written as

\[
\hat{H}_e \Psi^e = E_e \Psi^e
\]

\[
\hat{H}_{\text{nuc}} \Psi^{\text{nuc}} = E_{\text{nuc}} \Psi^{\text{nuc}},
\]

(2.4)

where \( \hat{H}_e \) and \( \hat{H}_{\text{nuc}} \) are the Hamiltonian of electrons and the Hamiltonian of nuclei, respectively. \( E_e \) and \( E_{\text{nuc}} \) are the total energy of electrons and the total energy of nuclei, respectively.

The Hamiltonian of electrons includes three terms: the kinetic energy of electrons, the potential energy from the interaction of electron-electron, and the potential energy from the interaction of electron-nucleus. In
accordance with the Born-Oppenheimer approximation, the nuclei are considered "fixed", and the electrons move in a static background. Hence, the interaction from static nuclei can be considered to be an external potential for electrons. The Hamiltonian of electrons is

\[ \hat{H}_e = \hat{T}_e + \hat{U}_{e-e} + \hat{V}_{\text{ext}} \]

\[ = -\sum_{i=1}^{n} \frac{\hbar^2}{2m_i} \nabla_i^2 + \frac{1}{2} \sum_{i,j(\neq i)}^{n} \frac{e^2}{4\pi\varepsilon_0 |r_i - r_j|} + \sum_{i=1}^{n} \sum_{j=1}^{N} \frac{1}{4\pi\varepsilon_0} \frac{Z_n e^2}{|R_i - R_j|}. \]  

(2.5)

Thus, the Schrödinger equation of electrons in Equation 2.4 becomes

\[ \hat{H}_e \Psi_e = (\hat{T}_e + \hat{U}_{e-e} + \hat{V}_{\text{ext}}) \Psi_e = E_e \Psi_e. \]  

(2.6)

The electronic wavefunction \( \Psi_e \) in Equation 2.6 is only related to the position of electrons if the electron spin is neglected. The \( \Psi_e \) that gives the lowest \( E_e \) is termed the ground state.

After \( E_e \) is obtained, inserting \( E_e \) into the Schrödinger equation of nuclei, we get

\[ (-\sum_{i=1}^{N} \frac{\hbar^2}{2M_i} \nabla_i^2 + E_e + \frac{1}{2} \sum_{i,j(\neq i)}^{N} \frac{(Z_i e)(Z_j e)}{4\pi\varepsilon_0 |R_i - R_j|}) \Psi_{\text{nuc}} = E_{\text{nuc}} \Psi_{\text{nuc}}. \]  

(2.7)

Compared with Equation 2.7, solving the Schrödinger equation of electrons (Equation 2.6) is problematic. Consequently, the next section describes how to solve the Schrödinger equation of electrons.

2.1.3 Hohenberg-Kohn theorem

DFT method has been developed to tackle Equation 2.6. It focuses on quantities in the real, three-dimensional coordinate space, principally on the electron density \([125]\). The concept of DFT was originally provided in the work of Thomas and Fermi in 1927 \([126, 127]\). In the Thomas-Fermi method and its refinements, the electron density \( n(r) \), found in Equation 2.8, plays a key role, where \( f_k \) is the occupation number.

\[ n(r) = \sum_k f_k |\psi_k(r)|. \]  

(2.8)

The Thomas-Fermi method neglects the exchange and correlation (XC) of the electrons and was not accurate enough, but the basic scheme of DFT has been determined in their work. In 1930, Dirac derived a first expression for the exchange energy in the context of Thomas-Fermi theory \([122]\). Recently, the widely applied DFT is based on the work of Hohenberg and Kohn. Hohenberg and Kohn proved the external potential in the system is a unique functional of \( n(r) \).

The basis of the Hohenberg and Kohn work is governed by two theorems \([124]\):
Methods

Theorem 1: The ground state density \( n_{gs}(r) \) uniquely determines the potential up to an arbitrary constant.

Theorem 2: A universal functional for the energy \( E[n] \) in terms of the density \( n(r) \) can be defined, valid for any external potential \( v_{ext}(r) \). For any given \( v_{ext}(r) \), the exact ground state energy of the system is the global minimum of this functional, and the density that minimizes the functional is the exact ground state density.

\[
E[n(r)] > E[n_{gs}(r)],
\]

where \( n_{gs}(r) \) is the electron density of the ground state. Minimizing \( E[n(r)] \) yields the ground state energy of the system. From these two theorems, energy \( E[n(r)] \) can be defined as

\[
E[n(r)] = T[n(r)] + U[n(r)] + V[n(r)] + E_{xc},
\]

where \( T[n(r)] \) is the kinetic energy functional of electrons, \( U[n(r)] \) is the functional describing the interaction of electron-electron in the system, and \( V[n(r)] \) is the external potential term. The ground state energy can be obtained by minimizing \( E[n(r)] \).

2.1.4 The Kohn-Sham method

If the XC interaction between the electron-electron pairs of the system is treated as an unknown term, Equation 2.10 contains two parts: the known terms that can be directly calculated and the unknown term that needs to be approximated. The known terms include the kinetic energy of electrons \( T[n(r)] \), the Hartree potential \( U_H[n(r)] \) that defines the electrostatic electron-electron interactions, and the external potential \( V_{ext}[n(r)] \) due to the nuclei and external field. The unknown term is the XC energy \( E_{xc}[n(r)] \). Equation 2.10 can be written as

\[
E[n(r)] = T[n(r)] + U_H[n(r)] + V_{ext}[n(r)] + E_{xc} = -\frac{\hbar^2}{2m} \sum_i^n \int \psi_i^* \nabla^2 \psi_i^* d^3r + \frac{e^2}{2} \int \int \frac{n(r)n(r')}{|r-r'|} d^3rd^3r' + \int V(r)n(r)dr + E_{xc}.
\]

In Equation 2.11, \( \psi_i^e \) is the wavefunction of the \( i \)th electron, \( V(r) = \sum I \frac{Z_i r_i^2}{|r-r_i|^3} \), and the kinetic energy is the sum of the electronic kinetic energy of all electrons. \( E_{xc}[n] \), the XC energy includes the exchange energy between the indistinguishable electrons and the energy of their correlated motion. The exact form of XC is unknown, hence we need the approximation of XC.

To solve Equation 2.11, Kohn and Sham considered an equivalent system. The particles in this system move in an effective field without any interaction with each other.
Then for the single electron, the Schrödinger equation is
\[
(-\frac{\hbar^2}{2m} \nabla^2 + V_{\text{eff}}(r))\psi_i(r) = \varepsilon_i \psi_i(r),
\]
(2.12)
where \(V_{\text{eff}}(r)\) is the effective field. The effective potential is
\[
V_{\text{eff}}(r) = V_{\text{nuc-e}}(r) + V_{\text{Hartree}}(r) + V_{\text{xc}}(r),
\]
(2.13)
where \(V_{\text{nuc-e}}(r)\) is the potential from the interaction between an electron and atomic nuclei. \(V_{\text{Hartree}}(r)\) is the Hartree potential that describes the static electron-electron interactions calculated from the electron density. The \(V_{\text{xc}}(r)\) is the potential from the XC energy \(V_{\text{xc}}(\mathbf{r}) = \frac{\delta E_{\text{xc}}(\mathbf{r})}{\delta n(\mathbf{r})}\). Then the solution of Equation 2.12 can be obtained by the self-consistent method:

1. Set an initial electron density \(n(r)\).

2. Use the electron density \(n(r)\) to solve Equation 2.12 to obtain the single electron wavefunction \(\psi_i(r)\).

3. Use the single electron wavefunction in Step 2 to obtain a new electron density \(n_{\text{new}}(r)\).

4. Compare \(n_{\text{new}}(r)\) obtained in Step 3 with \(n(r)\) that we used in Step 2. If the difference between these densities is below a threshold, the density is the ground-state electron density. Otherwise, return to Step 2, and use \(n_{\text{new}}(r)\) obtained in Step 3 to update the electron density.

In a self-consistent solution, the XC energy cannot be calculated precisely and must be approximated. One commonly used approximation is the local density approximation (LDA) [128, 129], which calculates the XC energy at each point in space based on the corresponding density of a homogeneous electron gas. LDA is the simplest approximation for the XC energy. The generalized gradient approximation (GGA) is an approximation that is more accurate than LDA. GGA takes into account the gradient of the electron density at each point to estimate the XC energy [130, 131]. The Perdew-Burke-Ernzerhof (PBE) GGA functional is the most common and widely used in DFT simulations for condensed matter physics today. Hybrid functionals, such as PBE0 [132] and B3LYP [133], are utilized in DFT simulations for obtaining more accurate results but they can be improved further.

### 2.2 Classical force field

A force field is employed to approximate the potential energy of a system using a set of predefined specific parameters that characterize diverse
interactions among atoms in the system. Interactions in a system are typically approximated using relatively simple forms, such as harmonic, Lennard-Jones, and Coulomb potentials. A force-field simulation can offer information on the total energy, force, velocity of each atom, system temperature, etc. Since the force field simulation is inexpensive, it can simulate extensive systems in long-time molecular dynamic simulations.

A typical classical force field includes five terms: bond stretching, bond bending, bond torsion, electrostatic Coulomb potential, and Van der Waals interaction. These five interaction potentials can be classified into bonded \( U_{\text{bonded}} \) and non-bonded \( U_{\text{non-bonded}} \) potentials, and the total potential is \( U_{\text{tot}} = U_{\text{bonded}} + U_{\text{non-bonded}} \) [134]. The force of the \( i \)th atom in the system is defined as
\[
F_i = -\frac{\partial U_{\text{tot},i}}{\partial r_i}.
\] (2.14)

Using Newton’s laws of motion, the motion parameters of the \( i \)th atom can be mathematically calculated.

### 2.2.1 AMBER force field

In this thesis, the assisted model building with energy refinement (AMBER) force field was used [135]. The simplest form of the AMBER force field (neglecting implicit solvent or polarization terms) utilizes the following Hamiltonian:

\[
U_{\text{tot}} = \sum_{\text{bonds}} k_b (r - r_0)^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 + \sum_{\text{dihedrals}} V_n [1 + \cos (n \phi - \delta)]
+ \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[ \frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^6} + \frac{q_i q_j}{4 \pi \epsilon_0 \epsilon r_{ij}} \right].
\] (2.15)

In Equation 2.15, \( k_b, k_{\theta}, \) and \( V_n \) are the bond force constant, the angle force constant, and the dihedral angle constant, respectively. In these harmonic functions, \( r, \theta, \) and \( \phi \) are the bond length, the bond angle, and the dihedral angle, respectively. \( r_0, \theta_0, \) and \( \phi_0 \) are the reference values of the bond length, bond angle, and dihedral angle from experiments or theoretical studies. \( \delta \) is the phase shift and \( n \) is the multiplicity. For the non-bonded interactions, parameters \( C_{ij}^{(12)} \) and \( C_{ij}^{(6)} \) are the coefficients of the repulsion term and the dispersion term of the Lennard-Jones potential, respectively. These two parameters rely on the types of the \( i \)th and \( j \)th atoms. The distance between two atoms is \( r_{ij} \). In practical simulation, a distance cutoff \( r_{\text{cut}} \) is set for \( r_{ij} \). When \( r_{ij} > r_{\text{cut}} \) the Lennard-Jones poten-
tial equals zero. Parameters \( q_i \) and \( q_j \) are respectively the charges of atoms \( i \) and \( j \). \( \epsilon_0 \) is the vacuum permittivity, and \( \epsilon_r \) is the relative permittivity. The value of \( q_i, q_j, \epsilon_0, \) and \( \epsilon_r \) are used to calculate the electrostatic potential. As with the Lennard-Jones potential, if \( r_{ij} > r_{\text{cut}} \), the electrostatic potential equals zero. These AMBER parameters, such as \( k_b, k_\theta, \) and \( V_n \), are optimized or created by fitting to quantum chemical calculations.

### 2.2.2 Energy minimization

To minimize the energy of a system, the steepest descent algorithm was employed in this work [134]. In this algorithm, the position of step \( n + 1 \) is obtained from the minimization result of step \( n \) as

\[
r_{n+1} = r_n + \frac{F_n}{\max(|F_n|)} h_n,
\]

(2.16)

where \( F_n, r_n, \) and \( h_n \) are the force, position, and maximum displacement in the \( n \)th step. After obtaining the new position \( r_{n+1} \), the new potential energy \( U_{n+1} \) is calculated. If \( U_{n+1} < U_n \), the new position is accepted, and the maximum displacement in step \( n + 1 \) is \( 1.2h_n \). If the new position is rejected, \( h_{n+1} \) is set to \( 0.2h_n \). The minimization process is stopped when the number of steps exceeds the preset maximum number of steps or when the absolute value of the maximum force falls below the convergence criterion for minimization.

### 2.3 Bayesian optimization

Bayesian optimization (BO) is a machine learning method based on Bayes’ theorem [136], which states that the posterior probability of a model \( M \) given evidence \( E \) is proportional to the likelihood of \( E \) given \( M \) multiplied by the prior probability of \( M \) [137], that is:

\[
P(M|E) \propto P(E|M)P(M),
\]

(2.17)

where \( P(M) \) is the prior beliefs about the space of possible objective functions, and \( P(E|M) \) is the likelihood of \( E \) given \( M \). \( P(M|E) \) is the updated belief about the unknown objective function.

Let \( x_i \) be the \( i \)th sample, and \( f(x_i) \) be the observation of the objective function at \( x_i \). As observations accumulate, \( D_{1:t} = \{x_{1:t}, f(x_{1:t})\} \), and the posterior distribution can be written as follow:

\[
P(f|D_{1:t}) \propto P(D_{1:t}|f)P(f),
\]

(2.18)

where \( P(D_{1:t}|f) \) is the likelihood function and \( P(f) \) is the prior distribution.

When it is difficult to evaluate the objective function, using a surrogate model to estimate the objective is a smart strategy. BO utilizes a Gaussian
Methods

process (GP) surrogate model [138], because the prior and posterior are then easy to calculate. Given data set \( D = \{x_{1:t}, f_{1:t}\} \), the resulting prior distribution on \( f_{1:t} \) is

\[
  f_{1:t} \sim N(\mu(x), K),
\]

where \( \mu \) is the mean value and is usually set as 0. In addition, \( K \) is the covariance matrix, which can be written as

\[
  K = \begin{bmatrix}
  k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_t) \\
  k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_t) \\
  \vdots & \vdots & \ddots & \vdots \\
  k(x_t, x_1) & k(x_t, x_2) & \cdots & k(x_t, x_t)
\end{bmatrix},
\]

(2.20)

where \( k(x_a, x_b) \) is a kernel function that is used to calculate the similarity (or distance in the feature space) between \( x_a \) and \( x_b \). Kernel function \( k(x_a, x_b) \) is frequently a squared exponential function (radial basis function kernel) and is defined as

\[
  k(x_a, x_b) = \sigma^2 \exp\left(-\frac{1}{2l^2}||x_a - x_b||^2\right).
\]

(2.21)

Moreover, for periodic dimensions, a sine function with periods \( p_n \) as a standard periodic kernel is expressed as

\[
  k(x_a, x_b) = \sigma^2 \exp\left(-\frac{1}{2l^2} \sin^2 \frac{|x_a - x_b|}{p_n} \pi\right),
\]

(2.22)

where \( \sigma^2 \) and \( l \) are the hyperparameters for variance and length scale, respectively.

Given a new point \( x_{t+1} \) the jointly Gaussian distribution of \( f_{1:t} \) and \( f_{t+1} \) is

\[
  \begin{bmatrix}
  f_{1:t} \\
  f_{t+1}
\end{bmatrix} \sim N(0, \begin{bmatrix} K & k \\
  k^T & k(x_{t+1}, x_{t+1}) \end{bmatrix}),
\]

(2.23)

where \( k = [k(x_{t+1}, x_1) k(x_{t+1}, x_2) \cdots k(x_{t+1}, x_t)] \). If the observations are modelled as noisy, i.e., \( y_i = f(x_i) + \epsilon_i \), where \( \epsilon_i \sim N(0,\sigma_{\text{noise}}^2) \), the new GP model needed to be predicted is

\[
  P(f_{t+1}|D, x_{t+1}) = N(\mu_t(x_{t+1}), \sigma_t^2(x_{t+1}) + \sigma_{\text{noise}}^2),
\]

(2.24)

where

\[
  \mu_t(x_t + 1) = k^T(K + \sigma_{\text{noise}}^2I)^{-1}y_{1:t},
\]

(2.25)

\[
  \sigma_t^2(x_{t+1}) = k(x_{t+1}, x_{t+1}) - k^T(K + \sigma_{\text{noise}}^2I)^{-1}k.
\]

(2.26)

The new GP model is determined by the posterior mean and variance. The location of the global minimum point \( \hat{x} \) is given by

\[
  \hat{x} = \arg\min(\mu_t(x_t + 1)).
\]

(2.27)
Methods

To sample the next most informative point in the next step, an acquisition function is constructed based on the current GP model. In this project, we employed the exploratory Lower Confidence Bond (eLCB) acquisition function [139, 140], defined as

$$A_i(x) = \mu_i(x) - \sqrt{\eta_i^2 \sigma_i^2(x)},$$

(2.28)

where $\mu_i(x)$ and $\sigma_i^2(x)$ are the mean and variance. $\eta_i$ is the balancing parameter, and it can be written as

$$\eta_i^2 = 2\log\left(\frac{\epsilon_i^2 + 2\pi^2}{3\epsilon_n}\right),$$

(2.29)

where $\epsilon_n$ is a small constant. After minimizing the acquisition function, the location of the next sample point $x_{\text{next}}$ can be obtained by:

$$x_{\text{next}} = \text{argmin} A_i(x).$$

(2.30)

In Equation 2.28, the $\mu_i(x)$ and $-\sqrt{\eta_i^2 \sigma_i^2(x)}$ terms correspond to exploration and exploitation. The balance of exploration and exploitation ensures the objective function is learned using a small size of data while exploring the entire space.

I illustrate how the BO algorithm selects the next sampling point in Figure 2.1. The objective function is $f(x) = (6x - 2)^2 \sin(12x - 4) - 5x$ on the interval $0 \leq x \leq 1$. In Fig. 2.1a, the black line is the true function ($f(x)$), and the BO algorithm fits five data points (red circle) and constructs a GP model $\mu(x)$ (blue line). The gray shaded area respects the uncertainty ($v(x)$); the higher and lower boundaries of the gray shaded area are $\mu(x) + v(x)$ and $\mu(x) - v(x)$, respectively. Fig. 2.1b shows the acquisition function calculated based on the GP model in Fig. 2.1a. The global minimum prediction by the GP model is marked by the red vertical line. The dashed green vertical line marks the location of the minimum value of the acquisition function, which is the location of the next sampling point.

The next sampling point is used for the surrogate model updating, and the BO search process is as follows:

1. Use some initial points to build an initial surrogate model.

2. Minimize the acquisition function $A(x)$ of the surrogate model. The next sample point is $x_{\text{next}} = \text{argmin} A_i(x)$. Add this sample point to previous samples and update the surrogate model.

3. If the parameter of the new surrogate model satisfies the convergence criterion, stop the BO process and obtain the global minimum from the final surrogate model. Otherwise, return to Step 2 until convergence.

In my study, I use the Bayesian Optimization Structure Search (BOSS) package [141] to implement BO to search low-energy structures of isolated
molecules (Publication I), to search low-energy structures of a ligand on a cluster (Publication II), and to optimize a KRR model (Publication III). More details of this are demonstrated in the next chapter.

2.4 Kernel rigid regression

Kernel rigid regression (KRR) is a machine learning method [142]. It is based on linear regression, regularization, and kernel tricks [143, 144, 145, 146, 147]. In this section, I first briefly review linear regression, regularization, and kernel tricks. I then introduce a molecular descriptor, the many-body tensor representation (MBTR), which is used in the trained KRR model.

2.4.1 Rigid regression

Given an input set \( X = [x_1, \ldots, x_d] \) and its corresponding response variables \( y = [y_1, \ldots, y_d] \), the linear function between them can be written as

\[
y = w^T X,
\]

where \( w \) is the weight vector that has the same dimension with \( y \) and \( X \).

To obtain the optimal weight \( w \), a quadratic cost function is minimized and is defined as

\[
L = \frac{1}{2} \sum_{i=1}^{d} (y_i - w^T x_i)^2.
\]

However, this approach may cause an overfitting problem. When overfitting exists in a model, the training data can be fitted very well, while the model performs poorly on new data. Regularizing the cost function can usually solve the overfitting problem. The general approach is to add a penalizing term in the cost function. Therefore, the cost function (Equation

![Figure 2.1](image.png)

**Figure 2.1.** (a) GP surrogate model. (b) The corresponding acquisition function.
2.32) becomes
\[
L = \frac{1}{2} \sum_{i=1}^{d} (y_i - w^T x_i)^2 + \frac{1}{2} \alpha \|w\|^2, \quad (2.33)
\]
where \(\alpha\) is the hyperparameter for the regularization determination. Taking the derivatives of Equation 2.33 and letting them equal zero gives
\[
\sum_{i=1}^{d} (y_i - w^T x_i) x_i = \alpha w, \quad (2.34)
\]
then, based on Equation 2.34, \(w\) is
\[
w = (\alpha I + \sum_{i} x_i x_i^T)^{-1} (\sum_{i} y_i x_i), \quad (2.35)
\]
where \(I\) is an identity matrix with dimension \(d\). The model that adds a penalizing term to the cost function of a linear regression model is called rigid regression. However, the rigid regression can only be applied to linearly separable data sets.

### 2.4.2 Kernel ridge regression

To make the model applicable for more diverse data, such as non-linear separable data sets, one can project the data to a higher feature space and make it separable. If the feature vector \(x_i\) becomes \(\phi(x_i)\), the cost function (Equation 2.33) can be written as
\[
L(w) = \frac{1}{2} \sum_{i=1}^{d} (\phi(x_i)^T w - y_i)^2 + \frac{1}{2} \alpha \|w\|^2. \quad (2.36)
\]
If we define \(y = (y_1, \cdots, y_d)\), \(X = (x_1, \cdots, x_d)\), and \(\Phi = (\phi(x_1), \cdots, \phi(x_d))\), then Equation 2.36 can be written as
\[
L(w) = \frac{1}{2} (\Phi w - y)^T (\Phi \Phi^T + \alpha I)^{-1} (\Phi w - y) + \frac{1}{2} \alpha \|w\|^2. \quad (2.37)
\]
Setting the gradient of Equation 2.37 to zero, the following analytic solution for weights is
\[
w = \Phi^T (\Phi \Phi^T + \alpha I)^{-1} y, \quad (2.38)
\]
For a new test data \(x\), the prediction is
\[
f(x) = \phi(x)^T \Phi^T (\Phi \Phi^T + \alpha I)^{-1} y, \quad (2.39)
\]
if we define a kernel function \(k\):
\[
k(x, y) = \phi(x)^T \phi(y), \quad (2.40)
\]
then \([\Phi \Phi^T]_{ij} = \phi(x_i)^T \phi(x_j) = k(x_i, x_j)\), and \([\phi(x)^T \phi]^T]_{ij} = \phi(x)^T \phi(x_j) = k(x, x_j)\). The prediction of Equation 2.39 is
\[
f(x) = k(x, X)(K + \alpha I)^{-1} y, \quad (2.41)
\]
where $K$ is the kernel matrix and $K_{ij} = k(x_i, x_j)$. For this, some common kernels can be used, such as the Gaussian kernel, the Laplacian kernel, and the radial basis function kernel. Determining the selection of kernel for KRR relies on the nature of the data and the problem that needs to be solved.

### 2.4.3 Many-body tensor representation

In the study of applying a ML method to accelerate the structure search for ligand-protected nanoclusters (Publication III), I applied KRR models to predict the energies and forces of $\text{Au}_{25}(\text{Cys})^{-1}_{18}$. During the predicted process, each structure needs to be translated to a vector representation for the kernel construction. In Publication III, the employed representation was MBTR. Here, a brief review of MBTR is given based on Ref. [78].

MBTR is a structural descriptor that captures the essence of a given structure by considering a variety of structural motifs. These motifs include elemental contents ($k = 1$), interatomic distances ($k = 2$), and bond angles ($k = 3$). In Publication III, I employed the motif with $k = 2$ to describe each structure and to form a vector. Specifically, the term $k = 2$ encodes the distance of each element pair $(Z_1, Z_2)$ in a structure as a sum of Gaussians

$$\sum_{l \in Z_1} \sum_{m \in Z_2} \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x - |\mathbf{R}_l - \mathbf{R}_m|^{-1})^2}{2\sigma}\right),$$

where the sum proceeds over all atom pairs of elements $Z_1$ and $Z_2$ in the structure, and $\mathbf{R}$ denotes the position of these atoms. The weighting function $w^{l,m}$ is

$$w^{l,m} = \exp(-s|\mathbf{R}_l - \mathbf{R}_m|),$$

where $s$ is a parameter that controls the magnitude of the weight. The weighting function is required for the sum to converge in an infinite lattice of atoms. Here, we set a cutoff distance $r_{\text{cutoff}}$, and if $|\mathbf{R}_l - \mathbf{R}_m|$ is larger than $r_{\text{cutoff}}$, the $w^{l,m}$ equals zero. The MBTR function is vectorized by evaluating it at $N_{\text{grid}}$ discrete grid points spanning from $x_{\text{min}}$ to $x_{\text{max}}$. In my research, $x_{\text{min}} = -0.1$, $x_{\text{max}} = 1.30$, and $N_{\text{grid}} = 50$. The MBTR vector of a $\text{Au}_{25}(\text{Cys})^{-1}_{18}$ structure os shown in Fig. 2.2.

The energy can be predicted using KRR by the equation:

$$E_{\text{KRR}} = \sum_{i}^{N} b_i k(s, s_i),$$

where $b_i$ is the fitting coefficients, $k$ is kernel function, and $s_i$ forms a set of $N$ reference structures. When using the MBTR as the descriptor,

$$k(s, s') = \exp(-\gamma ||\mathbf{M}(s) - \mathbf{M}(s')||_2^2),$$

where $\mathbf{M}(s)$ is the MBTR vector of structure $s$, and $\gamma$ is the parameter that controls the width of the kernel distribution. Based on Equation 2.41, $b_i$ is
Methods

Figure 2.2. A MBTR vector representation \((k = 2)\) of a \(\text{Au}_{25}(\text{Cys})_{13}^{-}\) structure.

determined by

\[
\mathbf{b} = (K + \alpha I)^{-1} E_{\text{train}},
\]

(2.46)

where \(E_{\text{train}}\) denotes the energies of structures in the training data.

In Publication III, I utilized KRR to predict the energies of structures but also their force by using the KRR model derivation. More details of this can be found in Ref. [78]. The code of KRR utilized in Publication III was provided by Jarno Laakso and encompasses several parts, including the many-body tensor representation (MBTR), MBTR derivation, KRR model, and derivation of KRR [148]. This ML-based approach offers computational efficiency and enables efficient scanning of the structures while mitigating the computational demands associated with employing DFT.
3. Computational details

3.1 FHI-aims code usage

This thesis utilizes the Fritz Haber Institute Ab Initio Molecular Simulation (FHI-aims) package for all DFT calculations. FHI-aims is an all-electron code that employs numeric atom-centered basis functions [149, 150, 151]. FHI-aims provides several numerical settings and basis sets that balance accuracy and efficiency. Moreover, FHI-aims scales efficiently on supercomputers having thousands of computing cores and can simulate large systems containing hundreds to thousands of atoms with high accuracy. FHI-aims provides DFT simulations with different functionals, such as PBE [131] and PBE0 [132]. It also provides a description of van der Waals (vdW) interactions, such as Tkatchenko-Scheffler (TS) [152] vdW and many-body dispersion (MBD) [153] in my work.

In the study of structure search for isolated molecules (Publication I), the "Tight" numerical settings and "tier2" basis sets were utilized throughout. The PBE + TS, PBE + MBD, PBE0 + TS, and PBE0 + MBD functionals were employed in the molecular search process. For structural relaxations, the geometry was considered to be converged when the maximum residual force (fmax) was below 0.01 eV/Å. To obtain more accurate energies, the vibrational free energies were computed using the finite-difference method within the harmonic approximation. A finite-difference displacement length of $\delta = 0.0025 \text{Å}$ was used. The vibrationally free energy $F_{vib}$ was calculated by

$$F_{vib} = \sum_{i} \left[ \frac{\hbar \omega_i}{2} + k_B T \ln(1 - \exp(-\frac{\hbar \omega_i}{k_B T})) \right], \quad (3.1)$$

where $N$ is the number of atoms in the molecule and $T$, $\omega_i$ and $k_B$ are the temperature, vibrational frequencies, and Boltzmann constant, respectively.

In the study of the structure search for the single ligand on a gold cluster
(Publication II), the system contains 25 gold atoms, 17 \(-\text{CH}_3\) groups, and a cysteine ligand, making DFT calculations with the PBE0 functional computationally too expensive. Hence, we used the PBE + MBD functional to calculate single-point energies and optimize structures. We utilized "Tight" numerical settings and "tier 2" basis sets for S, C, O, N, and H, and the "tier 1" set for Au. To ensure convergence during structure optimization, we set the fmax threshold to 0.01 eV/Å.

In the study of applying a ML method to accelerate the structure search for ligand-protected nanoclusters (Publication III), the PBE + TS functional was employed throughout. To generate the training data for building ML models, I used the "light" setting to optimize the structures. In the data generation process, the basis sets for all atoms in the system were "tier 1", and a structure was considered converged when fmax was below 0.05 eV/Å. During relaxation, all gold and sulfur atoms in the system were kept fixed.

### 3.2 GROMACS code

GROMACS, or GROningen MAchine for Chemical Simulations, is an open-source package used to perform molecular dynamic (MD) simulations [134]. Initially developed at the University of Groningen in the Netherlands, GROMACS is now supported and developed by an international community of researchers. The MD simulations in GROMACS utilize classical mechanics and employ a force field to describe the interactions between atoms in a given system. Additionally, GROMACS provides inexpensive ways for optimizing structures and calculating the PES of systems.

In this thesis, I only utilized GROMACS to perform all force field simulations in the study of applying ML to accelerate the structure search for ligand-protected nanoclusters (Publication III). A modified AMBER99 force field that was specially designed for thiolate-protected gold clusters [154] was utilized with the GROMACS code. I employed the steepest descent algorithm in GROMACS to optimize the random structures. Relaxation was complete when the maximum force (fmax) was below 100 kJ/mol/nm, or the number of relaxed steps exceeded 50,000. During the GROMACS relaxations, I employed the constrained method to keep the gold and sulfur atoms fixed in the systems.

### 3.3 BOSS code

Bayesian Optimization Structure Search (BOSS) [139, 155, 156, 157] is a Python program used to implement BO for various optimization tasks, both experimental and computational. BOSS uses the GPy software package [158] to build a GP model as a surrogate model for the target function.
Once the BOSS-predicted surrogate model has converged, the final GP model can be analyzed to extract the global and local minima locations. Further details of the BOSS approach can be found in Publication I and related references.

In the study of structure search for isolated molecules (Publication I), I used BOSS to learn the PES of isolated molecules. Using the BOSS post-process to analyze the predicted PES, I extracted the local minima structures within a low-energy window. During the BOSS sampling, due to the relatively rigid bond lengths and angles, I only selected the most informative dihedral angles of each molecule to define the search space.

In the study of the structure search for the single ligand on a gold cluster (Publication II), I utilized BOSS to learn the PES of the cysteine ligand to obtain both global and local minimum structures of the ligand. To facilitate the search, I selected the five dihedral angles of the cysteine ligand as a phase space and fixed other degrees of freedom in the system.

In the study of applying a ML method to accelerate the structure search for ligand-protected nanoclusters (Publication III), to obtain the optimal ML model, I used BOSS to optimize four hyperparameters, two for MBTR ($\sigma$ and $r_{\text{cutoff}}$) and two for KRR ($\alpha$ and $\gamma$). For each set of parameters, a corresponding KRR model was constructed for the mean absolute error (MAE) evaluation. After the BOSS learned model converged, I obtained the optimal KRR model with the lowest MAE.

For all optimized works of BOSS, I employed the eLCB acquisition function. I used the "rbf" kernel for the non-periodic sampling parameters and the "stdp" kernel for the periodic sampling parameters.
4. Efficient amino acid conformer search with Bayesian optimization

In this research, I developed an active learning procedure to predict the structures and energies of molecular conformers and used cysteine as an example to present the developed procedure. To test the generalizability and transferability of the procedure, I also studied the conformers of three other amino acids: tryptophan, serine, and aspartic acid.

This chapter focuses on the results from Publication I. In Sec. 4.1, I first briefly introduce the procedure. In Sec. 4.2, I demonstrate the efficiency and accuracy of the procedure with a 2-dimensional (2-D) search case of cysteine. Then I show how I search the conformers of cysteine in 5-D. In Sec. 4.3, I analyze the conformer search results of cysteine. I present the results of the conformer search for three other amino acids in Sec. 4.4. In the last section, I demonstrate the efficiency of the molecular conformer search procedure.

![Figure 4.1. Ball-and-stick model of the cysteine molecule. Red is used for oxygen, white for hydrogen, gray for carbon, blue for nitrogen, and yellow for sulfur. \(d_1, d_2, d_3, d_4,\) and \(d_5\) label the five dihedral angles of cysteine that we use to define our search space. Adapted from Publication I](image-url)
Efficient amino acid conformer search with Bayesian optimization

4.1 BOSS-based molecular conformer search

The procedure, BOSS-based molecular conformer search, includes four steps (Fig. 4.2): (i) system Preparation, (ii) Bayesian optimization conformer search, (iii) refinement, and (iv) validation.

In Step (i), I obtained the xyz-coordinates of cysteine from the drug bank online database [159] and optimized this structure with DFT. Then, I transformed the optimized xyz-coordinates to a z-matrix and obtained all dihedral angles. I selected the most informative degree of freedom dihedral angles $d_n$ to describe the different conformers. For example, for cysteine, five dihedral angles were used to define a 5-D search space (Figure 4.1).

In Step (ii), I employed BOSS to actively sample the search space defined in Step (i) to learn a molecular PES. In this process, only $d_n$ was sampled, and the other degrees of freedom were kept fixed at their DFT-optimized value.

After the BOSS-predicted PES converged, in Step (iii), I analyzed the PES and extracted the local minimum locations and related structures. Then, I employed DFT to relax these structures and obtain optimized structures and energies. To further refine the energies of optimized structures, I added vibrational entropy corrections for all relaxed local minimum structures. Thereafter, I calculated the zero-point energies and the vibrational free energies at 300 $K$ for these optimized structures. We also used coupled cluster (CC) to calculate the energies for the local minimum structures that we obtained within a low-energy window. To compare our prediction references, I calculated the MP2 or MP4 energies of the local minimum structures.

In Step (iv), I validated our results by comparing the low-energy conformers with the references.
### 4.2 2-D and 5-D cysteine conformer search

To test the accuracy and efficiency of Step (ii) in our procedure, I compared the 2-D ($d_1$ and $d_2$) BOSS-predicted PES with the reference PES that I generated using a $30 \times 30$ grid. At 60 iterations, BOSS-predicted PES accurately captures the minima and maxima using only 6% cost of the grid method. After 120 iterations, the BOSS-predicted PES resembles the reference map very well, featuring six energy minima of similar depth. This indicates the high accuracy and efficiency of BOSS-predicting PES.

After demonstrating the 2-D case of cysteine, I then employed BOSS to learn the 5-D PES of cysteine. It is impossible to check the convergence by visualizing the 5-D BOSS-predicted PES as in the 2-D case. Hence, I monitored the convergence in 5-D from the BOSS-predicted global minima and local minima at different BOSS iterations.

The BOSS-predicted global minimum energy (Fig. 4.4a) decreased continuously and the corresponding dihedral angles (Fig. 4.4b) changed throughout the whole procedure. Both energy and dihedral angles converged at 830 BOSS iterations, and further changes were negligible ($\Delta E < 0.025$, $\Delta d < 10^\circ$). The average computed energy of the sample points (red dashed line in Fig. 4.4a) is approximately 0.4 eV above the BOSS-predicted global minimum energy, which suggests that BOSS not only explores the region around the global minimum points but also explores the higher energy region to discover the local minima of the PES.

I show the relative energy of the local minima in different BOSS iterations in Fig. 4.5a. As the number of iterations increases, the number of local minima found tends to increase, and the curves rise slowly and gradually approach the curve for 1200 BOSS iterations. The curve for 1000 BOSS iterations is very similar to that of 1200 BOSS iterations within a low-energy region (< 0.25 eV), which suggests that the low-energy region of BOSS-predicted PES is sufficiently converged at 1000 iterations. I then relaxed the local minimum structures from 1000, 1200, 1400, and 1600
Figure 4.4. (a) Convergence of the global minimum energy computed from the BOSS-predicted global minimum configuration (black line). The average computed energy of the sampled conformers is shown with a red dashed line. (b) Value of the dihedral angles $d_n$ of the BOSS-predicted global minimum as a function of the number of sampled points. Adapted from Publication I.

Figure 4.5. (a) Progression of the relative energy of predicted local minima for a PBE0 + MBD BOSS run with a total number of 1600 iterations. Shown are intermediate curves at 400, 600, 800, 1000, 1200, 1400, and 1600 iterations. 0 eV is set to be the lowest energy found in the 1000 BOSS iterations. (b) We took the conformers from 1000, 1200, 1400, and 1600 iterations and did the DFT structure optimization with PBE0 + MBD. The conformers are reordered from the lowest to the highest energy. Adapted from Publication I.
iterations and obtained their DFT-optimized energies (Fig. 4.5b). The different energy curve lines almost overlap with each other below 0.25 eV, which also indicates that BOSS-predicted PES in the low energy region is sufficiently converged after 1000 BOSS iterations.

4.3 Conformational energy hierarchy of cysteine

I performed cysteine conformer search using the method introduced in Sec. 4.1 with four functionals: PBE + TS, PBE + MBD, PBE0 + TS, and PBE0 + MBD. To obtain more accurate energies of conformers, we applied the CCSD(T) single-point corrections to the 15 lowest energy conformers obtained from the PBE0 + MBD calculations.

I selected two reference papers for the comparison. In Ref. [160], cysteine molecules were produced in the gas-phase by laser ablation of a solid sample, and the structures of these cysteine molecules were characterized by their rotational spectrum. Ref. [160] also reported the theoretically predicted conformers of cysteine molecules using the MP4 calculation with the 6-311++G(d,p) basis set. Ref. [117] reported the computational results predicted by the MP2(FC)/cc-pVTZ method.

Here, I first introduce the local minimum structures that we obtained. I demonstrate the importance of different contributions to the energy hierarchy. Next, I briefly show and compare the predicted results from the four functional simulations. Finally, I compare our predicted conformers with references.

For the PBE0 + MBD simulation, I obtained 15 low-energy conformers within 0.2 eV from the global minimum. Among them, 11 conformers have the same geometric features as the conformers reported in Ref. [160]. Hence, I assigned them the same labels with Ref. [160], while naming the newly found four conformers as N1, N2, N3, and N4 (Fig. 4.6).

To illustrate the importance of different contributions to the energy hierarchy, I plotted the relative stability of the PBE0 + MBD conformers (Fig 4.7a), which not only includes the final energy ranking but also intermediate steps. The corresponding energy stability figures for the other three functionals are provided in Publication I.

Fig 4.7a shows that DFT optimizations play a major role in refining their energy ranking. The largest energy changes and reordering happens in this step. In the PBE0+MBD simulation, after DFT optimizations, the average energy change of these 15 structures is 0.095 eV, whereas the dihedral angles change on average by $\Delta d_1 = 16.9^\circ$, $\Delta d_2 = 20.9^\circ$, $\Delta d_3 = 8.9^\circ$, $\Delta d_4 = 26.1^\circ$, and $\Delta d_5 = 11.9^\circ$. The entropy corrections further refine the energies of the low-energy conformers. The zero-point energy calculation compresses the energy spacing but does not reorder any conformer. After adding finite temperature corrections (+ VE (300 K)), the energy spacing
Efficient amino acid conformer search with Bayesian optimization

Figure 4.6. Predicted low energy conformers of cysteine from the PBE0+MBD search. Following Ref. 47, conformers are named as I (NH–O=C), II (OH–N), and III (NH–OH) depending on the type of the hydrogen bonds, and as a, b, or c depending on the configuration of the –CH$_2$SH side chain. The experimentally detected conformers are marked in red and other conformers marked in black. The colour scheme of the atoms is the same as in Fig. 4.1. Adapted from Publication I.

Figure 4.7. Relative stability for all steps of the PBE0 + MBD-based search. (a) From left to right: BOSS prediction, after structure optimization, after adding the vibrational energy at 0 K (+VE (0 K)), and after adding the vibration energy at 300 K (+VE (300 K)). The two farthest right ones are +VE (300 K) and the energy order of the CCSD(T) result but enlarged two times. For each step, the energy of the most stable structure defines the zero of energy for that column. (b) From the left to right: BOSS prediction, after optimization, and after MP4 energy calculations. The last two columns show an enlarged version of the MP4 results in comparison with the MP4 results of Ref 47. Adapted from Publication I.
was further compressed and the energy order (above 0.1 eV) of a couple of the conformers was switched. The CCSD(T) corrections, our most accurate conformer energy hierarchy, are sensitive to the conformer geometry. The corrections continuously reduce the energy spacing and only switch the energy order of conformers IIa and IIc.

Among the top 10 most stable structures, Ref. [117] reported eight stable conformers (IIb, IIa, Ib, I’b, I’a, IIIb, IIIc and IIIα,b) that were also found in this work and Ref. [160]. I compared the corresponding structures of PBE0 + MBD with those in Ref. [117], and I found that the average difference of the dihedral angles between the results and geometries in Ref. [117] are small (from 0.7° to 4.6°). Moreover, our CCSD(T) results show an energy ranking that is very similar to the MP2 results in Ref. [117] (Table 4.1). This suggests that we indeed found the right conformers.

Ref. [160] does not provide atomic coordinates for the reported conformers. Therefore, I compare our results by considering the energy. Our CCSD(T) results have a very similar energy ranking to the MP4 results in Ref. [160]. To directly compare the energies of our predicted conformers with the energies in Ref. [160], I also used the same basis sets (6-311++G(d,p)) to calculate the single-point energies for our PBE0 + MBD predicted local minimum conformers and found that the two results agree within 4 meV for each conformer. This close match indicates that our conformer geometries agree very well with those of Ref. [160]. Furthermore, compared with the experimental results in Ref. [160], our CCSD(T) results identified the three most stable conformers are IIb, Ib, and I’a, which match the three most abundant experimentally detected conformers in the reference, although the ranking is slightly different (Table 4.1). The next three stable conformers of our CCSD(T) results are IIa, IIIb, and IIIc, which also agree with the experimental results in Ref. [160], but with a different energy order (Table 4.1).

In our simulations, PBE + TS, PBE + MBD, PBE0 + TS, and PBE0 + MBD all found the correct global minimum structure IIb. The four simulations predicted the six experimental conformers among the top-eight most stable structures (Table 4.1). The average energy difference between these four simulations is tiny, which suggests that the PBE + TS functional is sufficient for cysteine. For large molecules, an economical search process might be to use the BOSS sampling to learn PES first at the PBE + TS level and then use PBE0 + MBD to post-relax a certain number of low-energy conformers.

### 4.4 Conformer search for serine, aspartic, and tryptophan

I applied our conformer search procedure with the PBE0 + MBD functional to serine, aspartic, and tryptophan. As in the cysteine conformer search, I
Table 4.1. The energy order of the ten most stable conformers of cysteine from our DFT, MP4 and CCSD(T) computations, Ref. [160] and Ref. [117]. Our CCSD(T) and MP4 results are based on PBE0+MBD structures. b1: 6-311++G(d,p) basis set, b2: aug-cc-pvtz basis set, *: vibrational energy corrections not included, Exp(ratio): Abundance ratio of experimental detect conformers [160].

<table>
<thead>
<tr>
<th>Energy order</th>
<th>PBE+TS</th>
<th>PBE+MBD</th>
<th>PBE0+TS</th>
<th>PBE0+MBD</th>
<th>MP4 (b1)*</th>
<th>MP4 (b1)[160]</th>
<th>MP4 (b2)</th>
<th>CCSD(T) (b2)</th>
<th>CCSD(T)[117]</th>
<th>Exp[160]</th>
<th>Exp(ratio)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IIb</td>
<td>IIa</td>
<td>Ib</td>
<td>Ia</td>
<td>III βb</td>
<td>IIb</td>
<td>Ib</td>
<td>IIc</td>
<td>III βc</td>
<td>IIa</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>IIa</td>
<td>IIa</td>
<td>Ib</td>
<td>III βb</td>
<td>IIb</td>
<td>III βb</td>
<td>IIa</td>
<td>III βc</td>
<td>III aβ</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Ib</td>
<td>Ib</td>
<td>III βb</td>
<td>IIa</td>
<td>III βb</td>
<td>IIa</td>
<td>III aβ</td>
<td>III aβ</td>
<td>III aβ</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Ia</td>
<td>III βb</td>
<td>IIa</td>
<td>III βb</td>
<td>III aβ</td>
<td>III aβ</td>
<td>III aβ</td>
<td>III aβ</td>
<td>III aβ</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>III b</td>
<td>III βb</td>
<td>IIa</td>
<td>III aβ</td>
<td>III aβ</td>
<td>III aβ</td>
<td>III aβ</td>
<td>III aβ</td>
<td>III aβ</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

relaxed all the local minima in the BOSS-predicted PES, included entropy corrections at 300 K, and added CC corrections to the 15 lowest energy conformers. I selected reference results from previous studies for comparison [161, 162, 163]. Again, to directly compare our results with those reported in the reference, I calculated the MP2 or MP4 energies of three molecules, using our PBE0+MBD optimized structures and the same basis sets as in the references.

In Ref. [161], 11 serine conformers were theoretically predicted by MP2 calculation with a 6–311++G(d,p) basis set. They were classified into I, II, and III types based on the H-bonding type of the amino group and the carboxyl functional group. Conformers of Ia, IIb, I’a, IIc, III βb, and III βc were detected experimentally. Here, I found the seven experimentally detected serine conformers among the top nine most stable structures. Our MP4 results are very similar to that of Ref. [161] and the average difference is 6 meV.

Ref. [162] shows 15 aspartic conformers theoretical predicted by MP2 calculation with the 6-311++G(d,p) basis set. Using the H-bonding type of the amino group and the carboxyl functional group, these conformers were classified into I, II, and III types. Ref. [162] reported results of the six experimentally detected conformers (IIb-I, IIa-I, Ia-II, Ib-I, III βb-I, and Ia-I.) In my work, I found the six experimental reported conformers among the top eight most stable conformers. Compared with Ref. [162], our MP4 results switch the energy order of four conformers. And the average energy difference is 0.003 eV if we only consider experimental detected structures.
Ref. [163] reported 45 tryptophan conformers theoretical predicted by MP2 calculation with the 6-311++G** basis set. Based on the H-bonding type of the amino group and the carboxyl functional group, these conformers were classified into A, B, C, D, E, F, G, and H types. Among these structures, the dominant conformer is of type A. Other relevant conformers are of type B. Our CCSD(T) results found the same order. Our MP2 results showed that the A- and B-types are more stable than other types, which agrees with Ref. [163]. The average energy difference is 0.010 eV if we only consider A- and B-type conformers.

4.5 Computational efficiency of the search procedure

In our research, BOSS needs to sample about 1000 data points to predict a physically meaningful PES for the four amino acids with 5–6 degrees of freedom. Considering that FHI-aims requires on average 30 geometry optimization steps to relax the structure of an organic molecule, the computational cost of 1000 single-point energy calculations during BOSS sampling is equivalent to approximately 30 DFT relaxations.

For cysteine, approximately 80 local minimum structures were optimized with DFT. This means that the total computational expense for a complete conformer search of cysteine is about 100 DFT geometry optimizations (or 3400 DFT single-point energy calculations). Similar DFT steps were used to search the conformers of serine, aspartic, and tryptophan. This is a very small computational budget, compared to systematic [117] or stochastic [105] conformer search methods that need to relax thousands of structures. A genetic algorithm (GA)-based conformer search method [105] for molecules (4 to 6 degrees of freedom) requires between 20,000 and 60,000 single-point DFT calculations. In contrast, our procedure is a factor of 10 more efficient.
5. Searching the structures of a cysteine adsorbate on a gold cluster with Bayesian optimization

In the study of the structure search for isolated molecules, I developed a molecular conformer search procedure and applied it to four amino acids. I modified it to search the conformers of an organic molecule on a cluster. In particular, I developed three strategies to avoid the steric clashes between a molecule and a cluster during active learning. To test and demonstrate our search method, I chose to study a cysteine on a well-studied gold–thiolate Au$_{25}$ cluster.

This chapter focuses on the results from Publication II. In Sec. 5.1, I introduce two search systems. Then, I briefly review the strategies I developed in Sec. 5.2. In Sec. 5.3, I evaluated the performance of three strategies. All the stable structures of the two search systems were analyzed in Sec. 5.4.

5.1 Two systems of a cysteine adsorbate on a Au$_{25}$ cluster

In this work, I used the Au$_{25}$(SCH$_3$)$_{18}^{-1}$ as the model cluster. SCH$_3$ was chosen here to reduce the computational cost and complexity. The Au$_{25}$(SCH$_3$)$_{18}^{-1}$ cluster has an Au$_{13}$ icosahedral protected by six SCH$_3$ – Au – SCH$_3$ – Au – SCH$_3$ V-shaped staples. Hence, a cysteine can bind to two inequivalent

Figure 5.1. Ball-stick model of cysteine on an Au$_{25}$(SCH$_3$)$_{17}^{-1}$ cluster: (a) system A and (b) system B. Gold color is used for gold atoms, red for oxygen, white for hydrogen, gray for carbon, blue for nitrogen, and yellow for sulfur. $d_1$, $d_2$, $d_3$, $d_4$, and $d_5$ label the five dihedral angles of cysteine that we use to define our search space. Adapted from Publication II.
Searching the structures of a cysteine adsorbate on a gold cluster with Bayesian optimization

Figure 5.2. Two-dimensional ($d_1$ and $d_2$) maps of the shortest atomic pair distance $D_{\text{min}}$ between cysteine and the cluster for (a) system A and (b) system B. In the plots, a light color means the structures have a large $D_{\text{min}}$, while a dark color means the structures have a small $D_{\text{min}}$, where the steric clash may happen. The restricted sampling region of $d_i$ in strategy i is between the two red dashed lines. Adapted from Publication II.

S sites, one is on the top of the staple and the other is on the side. Correspondingly, I constructed two search systems with the same formula $\text{Au}_{25}(\text{SCH}_3)_{17}^- (\text{Cys})^-$ (Cys: cysteine): System A and System B (Figure 5.1). The systems were built by replacing one SCH$_3$ in $\text{Au}_{25}(\text{SCH}_3)_{18}$ with one deprotonated cysteine. The two systems allow us to study how the different surrounding environments affect the structures of cysteine. When I employed BOSS to search conformers for ligands, the five dihedral angles of the cysteine ligand on System A and System B were used to define the search space (Figure 5.1), and other degrees of freedom were kept fixed.

5.2 Three strategies for addressing steric clashes between the ligand and the cluster

When a cysteine bonds to a gold cluster, it is surrounded by gold atoms and other ligands. Searching the structure of the cysteine with active learning may lead to steric clashes between the cysteine and the gold cluster. To address challenges from the steric clash, I developed three strategies: i sampling in the limited phase space, ii safe distance selection, and iii energy transformation.

The most intuitive way to avoid steric clashes is to select a "safe" region in the search space, where the steric clash will not happen during the active learning. In Strategy i, I first randomly generated 100,000 $d_i$ ($i=1,2,3,4,5$) structures for both system A and B. For each random structure, I calculated the shortest atomic pair distance $D_{\text{min}}$ between cysteine and $\text{Au}_{25}(\text{SCH}_3)_{17}$ and plotted $D_{\text{min}}$ against $d_i$ (shown in Fig. S1–S3 in Publication II). For the $d_1$ dihedral angle of the two systems, I found a continuous region with relatively high values of $D_{\text{min}}$ that can be defined as a "safe" region. For other dihedral angles, no such "safe" region can be found. I then plotted
the $d_1-d_2$ 2-D distribution of $D_{\text{min}}$ for systems A and B (Fig. 5.2), which shows that a "safe" sample region can be defined by restricting $d_1$ to $[70^\circ, 210^\circ]$ for system A or $[140^\circ, 240^\circ]$ for system B.

In Strategy ii, I introduced a safe distance $D_0$ to distinguish between "safe" and "unsafe" structures and treat them differently. For each structure in the BOSS sampling, the shortest atomic distance $D_{\text{min}}$ between the cysteine ligand and Au$_{25}$(SCH$_3$)$_{17}$ is calculated. If $D_{\text{min}} > D_0$, the structure is identified as physically meaningful, and its energy is its DFT energy $E$. Otherwise, if $D_{\text{min}} < D_0$, it is considered to be a nonphysical structure, and a constant energy $E_0$ is used as its energy. The energy $E_{\text{new}}$ for updating the GP in the BOSS sampling is:

$$E_{\text{new}} = \begin{cases} E & D_{\text{min}} > D_0 \\ E_0 & D_{\text{min}} \leq D_0 \end{cases}.$$  \hspace{1cm} (5.1)

In this work, I chose $D_0 = 1.4$ Å and $E_0 = 6.0$ eV.

In Strategy iii, I introduced an energy cutoff ($E_{\text{cut}} = 2.0$ eV) to determine whether or not the DFT energy needs to be transformed. For a structure with a high DFT energy $E > E_{\text{cut}}$, I used a logarithmic energy transformation to attenuate the high energy of the nonphysical structure during the BOSS sampling. Additionally, BOSS may sample nonphysically meaningful structures that cannot obtain DFT energy. For these structures, I applied a penalty energy ($E_p = 4.5$ eV) to update the GP during the BOSS sampling. The method by which I obtained the energy $E_{\text{new}}$ in the data point for a BOSS-sampled structure can be concluded as

$$E_{\text{new}} = \begin{cases} E & E \leq E_{\text{cut}} \\ E_{\text{cut}} + \log_{10}(E - E_{\text{cut}} + 1) & E > E_{\text{cut}} \\ E_p & \#E \end{cases}.$$  \hspace{1cm} (5.2)

The three strategies all have their advantages and limitations. Strategy i is easy to apply and unlikely to sample a nonphysically meaningful structure. Moreover, it does not change any DFT energy order of these BOSS-sampled structures. However, Strategy i may miss parts of the phase space that correspond to low-energy structures, such as the yellow regions outside of the dashed line in Fig. 5.2. Strategy ii samples the entire phase space, and there is no DFT calculation for nonphysical structures. However, the disadvantages of Strategy ii are that the surrogate PES is not accurate for structures with $D_{\text{min}} < D_0$, and the energy-structure relation is discontinuous around $D_0$ which could lead to sub-optimal surrogate PES model fits. Since a structure with $D_{\text{min}}$ smaller or close to $D_0$ typically has a high energy, these disadvantages should not affect the low-energy PES prediction. Compared with Strategies i and ii, Strategy iii not only searches the entire space but also does not change the energy order of the
5.3 Evaluating the performance of the three strategies

I used the system A to evaluate the three strategies. The PES model using the three strategies converged within 1000 iterations. Firstly, I extracted all local minimum structures from the PES at the 1000th iteration. Then, I applied DFT to optimize these structures. After removing duplicate structures and only keeping unique structures, using Strategies \( i \), \( ii \), and \( iii \), I finally obtained 37, 45, and 39 unique structures within energy windows of 0.40 eV, 0.68 eV, and 0.63 eV from their global minimum, respectively.

I evaluated the strategies from two aspects: the accuracy of the surrogate PES prediction and the final local minima structures. I selected the 30 lowest energy structures after DFT relaxation for the evaluation. The accuracy of the surrogate PES prediction is measured by the differences between the BOSS-predicted energies and the DFT single-point energies of the 30 structures before relaxation. As shown in Fig. 5.3a-c, the DFT energy is generally higher than the BOSS energy. The mean differences are 0.29 eV, 1.14 eV, and 0.52 eV for Strategies \( i \)–\( iii \), respectively. The smaller difference indicates the BOSS-predicted PES is more accurate. Hence, Strategies \( i \) and \( iii \) are better than Strategy \( ii \) from this perspective.

Sampled structures. The only limitation with it is that the prediction of the high-energy region in the PES is inaccurate.
Searching the structures of a cysteine adsorbate on a gold cluster with Bayesian optimization

Figure 5.4. Types of hydrogen bonds between the amino group and the carboxyl group in cysteine. Adapted from Publication II.

The relaxed energy curves of these structures (Fig. 5.3d) show that the curves generated by Strategy ii and Strategy iii are very close to each other. This suggests that Strategy ii and Strategy iii have obtained very similar local minima structures. In the higher energy region, the three curves lie almost on top of each other, suggesting the three strategies found the same structures. The curve generated by Strategy iii lies below the other two in the low-energy region, indicating that Strategy iii found lower energy structures. Strategy i missed the global minimum structure, while Strategies ii and iii found the global one. Hence, from the two measured aspects, Strategy iii exhibits the best performance, and it was used to study system B.

5.4 The predicted low-energy structures

Using BOSS combined with Strategy iii to sample the PES of systems A and B, I obtained 39 unique local minima structures for system A and 30 unique local minima structures for system B. The energy window of these structures is 0.63 eV for system A and 0.40 eV for system B.

An isolated cysteine can form four types of hydrogen bonds (Fig. 5.4). The I, II, and III hydrogen bonds can be found in the low-energy structures of isolated cysteine molecules, whereas the IV hydrogen bond has not been found in the low-energy conformers of isolated cysteine. To investigate the geometric feature of the predicted ligand structures, I calculate the similarity between the predicted ligand structures for the two systems and the 11 reference cysteine conformers identified both in Ref. [160] and Publication I, as shown in Fig. 5.5. Given two structures a and b, the
 Searching the structures of a cysteine adsorbate on a gold cluster with Bayesian optimization

Figure 5.5. The similarity between the eleven reference cysteine molecules and the cysteine in the local minimum structures of (a) system A and (B) system B. The y-axis represents the eleven references, and the x-axis lists the local minima structures of system A or system B in the increasing order of energy. Adapted from Publication II.

The similarity of them can be calculated using the following equation:

\[ S_{\cos} = \frac{v_a \cdot v_b}{|v_a||v_b|}, \]  

where \( v_a \) and \( v_b \) are the vectors \([\cos(d_2), \cos(d_3), \cos(d_4), \cos(d_5)]\) of the two structures (the definition of \( d_i \) is shown in Fig. 5.1). The high similarity (i.e., the red color in Fig. 5.5a–b) suggests that cysteine may form I, II, and III hydrogen bonds when it adsorbs on the cluster. The low similarity (i.e., the blue color in Fig. 5.5a–b) means that the other local minimum structures are not similar to any reference molecular conformers. After analysis, I found that they mainly have type IV hydrogen bonds, suggesting that the cluster indeed affects the atomic structures of the absorbing cysteine.

Next, I selected the 10 lowest-energy predicted structures of the two systems for analysis. For system A, eight out of 10 structures have type II hydrogen bonds and two have type IV (Fig. 5.6a). The shortest distance between the cysteine ligand and cluster is either between a hydrogen atom in the cysteine and a gold atom in the cluster, or between a hydrogen atom in the cysteine and a sulfur atom in the cluster. This suggests that the weak H-S and H-Au interactions help to stabilize the system.

In system B, the four lowest energy structures have type II hydrogen bonds followed by four structures with type IV hydrogen bonds and one
Searching the structures of a cysteine adsorbate on a gold cluster with Bayesian optimization

**Figure 5.6.** Predicted top ten low energy structures of system A (a) and B (b) from the BOSS search. Adapted from Publication II.

**Figure 5.7.** Relative energies of the ten most stable structures of isolated cysteine molecules, system A and system B. The red line means the structures contain a type II hydrogen bond in the cysteine ligand, while the green line means other types of hydrogen bonds in the cysteine ligand. Adapted from Publication II.
with a type I hydrogen bond (Fig. 5.6b). As in system A, the H–Au and H–S interactions help stabilize the low-energy structures in system B. Comparing system A and B, it is clear that the local environment significantly affects the energy ranking of the cysteine conformers on the cluster.

I plot the energy of the structures in Fig. 5.6 and Fig. 5.7 and include the top ten most stable gas-phase cysteine conformers. Type II hydrogen bonds are dominant in system A and B. In contrast, for isolated cysteine, most conformers in the same energy window have other types of hydrogen bonds (type I and III). Comparing the energy ranking, system A and B exhibit a larger energy spacing between different configurations than the isolated cysteine molecular conformations.
6. Applying machine learning to accelerate the structure search for the ligand-protected nanoclusters

In the study of the structure search for the single ligand on a gold cluster (Publication II), I identified the low-energy conformers of a cysteine on a Au$_{25}$(CH$_3$)$_{17}$ cluster. To search the structure of the ligand-protected nanocluster, I directly utilized these lower-energy conformers to construct the initial ligand-layer structures of Au$_{25}$(Cys)$_{18}^{-1}$. Given the high dimensionality of the configurational space, using BO to search the atomic structures of the ligand layer is not suitable. Therefore, I trained a KRR model that can optimize the structures of Au$_{25}$(Cys)$_{18}^{-1}$. To increase the diversity of training data and improve the performance of the KRR model, I developed an active learning workflow. After obtaining a reliable KRR model, I used the model to optimize a considerable number of Au$_{25}$(Cys)$_{18}^{-1}$ structures. For all relaxations, the gold and sulfur atoms are constrained due to obtaining Au$_{25}$S$_{18}$ geometric structure from references [72, 71, 74]. Finally, I analyzed the structural and electronic features of the low-energy ones.

This chapter focuses on the results from Publication III. In Sec. 6.1, I show how to construct the initial structures of the Au$_{25}$(Cys)$_{18}^{-1}$ ligand-protected nanocluster. Then, I present the method for building a KRR model for structure relaxation in Sec. 6.2. In Sec. 6.3, I briefly introduce the addition of GROMACS in KRR models to improve the performance of relaxations. I then introduce our active learning workflow in Sec. 6.4. The effects of active learning and adding the GROMACS energies and forces are presented in Sec. 6.5. In Sec. 6.6, I analyze the low-energy structures.

6.1 Constructing the data pool of the Au$_{25}$(Cys)$_{18}^{-1}$ cluster

I first constructed a data pool for ML, which served three purposes. Firstly, the initial dataset for training a KRR model was chosen from the data pool. Secondly, the structures added to active learning were picked from this data pool. Thirdly, once the KRR model had converged, all the remaining data pool structures were optimized with the final model to
Applying machine learning to accelerate the structure search for the ligand-protected nanoclusters

**Figure 6.1.** Ball-stick models of $\text{Au}_{25}(\text{CH}_3)_{18}^-$ and $\text{Au}_{25}(\text{Cys})_{18}^-$. The gold color is used for gold atoms, red for oxygen, yellow for sulfur, gray for carbon, blue for nitrogen, and white for hydrogen. Adapted from Publication III.

identify the low-energy structures.

I obtained the structure of $\text{Au}_{25}(\text{SCH}_3)_{18}^-$ from the experiments [74]. To construct the initial ligand-layer structures I directly utilized local minimum structures of cysteine from Publication II to replace the 18 $-\text{SCH}_3$ in $\text{Au}_{25}(\text{SCH}_3)_{18}^-$ (Fig. 6.1). The 18 sulfur atoms in a $\text{Au}_{25}(\text{SCH}_3)_{18}^-$ can be classified into two types: six A site atoms located on the top of six V-shaped S–Au–S–Au–S structures, and 12 B site atoms located on the sides of these V-shaped structures (Fig. 6.2a). It was assumed here that the same type of sulfur atoms bind with identical cysteine structures.

Here, I chose the ten most stable conformers from system A in Publication II to bind with the A site and the ten most stable conformers from system B of Publication II to bind with the B site, yielding 100 combinations. For each combination, I generated 1,000,000 $\text{Au}_{25}(\text{Cys})_{18}^-$ structures by sampling dihedral angles around the S–C bonds (Fig. 6.2b). There are 18 dihedral angles, and because of the central symmetry of $\text{Au}_{25}\text{S}_{18}$, only nine dihedral angles needed to be sampled. Since most random sampling structures lead to steric clashes, I computed the shortest atomic distance ($D_{\text{min}}$) between atom pairs in each structure to pre-select physically reasonable structures. I obtained the 1000 structures with the largest $D_{\text{min}}$ values. I then utilized a force field method to relax these 1000 structures and yield the 1000 force field-relaxed structures. Since I generated 100 combinations, I obtained 100,000 force field-relaxed structures in total. I used these structures to build a "data pool".

In Publication II, we divided the H-bond of the cysteine ligand into two groups: "II-type" and "O-type" (Figure 5.7). Hence, the structures in the "data pool" can be divided into four groups: II–II, II–O, O–II, and O–O. The first letter labels the H-bond type of cysteine binding with the A site,
Applying machine learning to accelerate the structure search for the ligand-protected nanoclusters

Figure 6.2. (a) Ball-stick model of \( \text{Au}_{25}(\text{S})_{18} \). (B) Ball-stick model of \( \text{Au}_{25}(\text{Cys})_{18}^{-1} \). The gold color is used for gold atoms, red for oxygen, yellow for sulfur, gray for carbon, blue for nitrogen, and white for hydrogen. Adapted from Publication III.

and the second letter is for cysteine binding with the B site.

6.2 Using KRR model to optimize the structures of \( \text{Au}_{25}(\text{Cys})_{18}^{-1} \)

I randomly selected 100 force field-relaxed structures from the "data pool" and then employed DFT to optimize these structures. I found that the force field energy ranking of these structures is totally different from their DFT-relaxed energy ranking. Suggesting that the force field simulation is not sufficiently accurate to discover the low-energy structures.

Furthermore, since each structure is non-periodic and has 259 atoms, using DFT methods to relax these 100,000 structures would be extremely expensive. Hence, I employed an ML approach based on KRR to relax all 100,000 structures in the "data pool" to identify the low-energy structures.

To obtain data for training the KRR models, I employed DFT to optimize 1000 random structures from the "data pool". I then selected 5000 data points randomly from the 1000 DFT relaxation trajectories as the training set to construct a KRR model. I then picked 1000 data points from these trajectories as the test data. Subsequently, I employed BOSS to optimize the KRR hyperparameters (\( \alpha \), and \( \gamma \)) as well as the MBTR hyperparameters (\( \sigma \) and \( r_{\text{cutoff}} \)). Finally, I obtained the optimal hyperparameters (\( r_{\text{cutoff}} = 15.6 \, \text{Å} \), \( \sigma = 10^{-1.66} \), \( \alpha = 10^{-4.0} \), and \( \gamma = 10^{-5.95} \)) from the BOSS optimization and the corresponding KRR model. These parameters were used throughout this work.

I utilized the trained KRR model (referred to as the MBTR-KRR model) to predict the energies and forces of the 1000 structures in the test data. The mean absolute error for energy prediction (MAE\(_E\)) is 0.29 eV, and
Applying machine learning to accelerate the structure search for the ligand-protected nanoclusters

Figure 6.3. The main bond lengths and angle changes of cysteine ligand for structures in the MBTR-KRR model relaxed trajectory. The solid line is the average value of all bond lengths or angles in a structure. (a) Bond lengths, (b) Bond angles. Adapted from Publication III.

the mean absolute error for force prediction (MAE\(_F\)) is 0.61 eV/Å. I then applied this model to relax 100 new random structures from the data pool. I found that the high MAE\(_F\) causes the KRR model to perform poorly when relaxing the structures. Fig. 6.3a and Fig. 6.3b show the changes in bond lengths and bond angles when the MBTR-KRR model optimizes a structure. The changes in the bond lengths are tiny, suggesting that the MBTR-KRR model can capture the structure features of the bond lengths. By contrast, the bond angles change too drastically, which leads to incorrect bond angles during the structure relaxation.

The poor performance of the MBTR-KRR model can be attributed to three reasons. Firstly, the structure has 259 atoms (i.e., 777 degrees of freedom), resulting in an extremely complicated PES that is very difficult to explore. Secondly, the training data for the MBTR-KRR model were selected from DFT relaxed trajectories, which only cover a small portion of the entire PES. Thirdly, only the interatomic distances in a structure were used to construct the MBTR vector for the MBTR-KRR model, which leads to the lack of bond angle information in the MBTR-KRR model. I attempted to improve the MBTR-KRR model by adding more training data and testing different model hyperparameters values. However, MAE\(_E\) remained larger than 0.29 eV, and MAE\(_F\) remained larger than 0.61 eV/Å.

6.3 Fixing the bond angles of cysteine by using force field methods

To address the nonphysical changes of bond angles that occur during relaxation, I added the forces and energies information from force field simulations. The KRR model with force field simulation aims to learn the energy difference between DFT and force field results. Hence, a single data point contains the structure (MBTR vector) and the energy difference between DFT and force field simulations. The predicted energy and forces
Applying machine learning to accelerate the structure search for the ligand-protected nanoclusters

Figure 6.4. The active learning workflow. KRR: Kernel rigid regression ML model. GPR: Gaussian process regression ML model. Adapted from Publication III.

used for comparison with the DFT simulations are

\[ E_{\text{predicted}} = E_{\text{FF}} + \Delta E_{\text{KRR}}, \]

\[ F_{\text{predicted}} = F_{\text{FF}} + \Delta F_{\text{KRR}}, \]

where \( E_{\text{FF}} \) and \( F_{\text{FF}} \) are the force field-calculated energy and forces for a structure, respectively. \( \Delta E_{\text{KRR}} \) and \( \Delta F_{\text{KRR}} \) are the KRR model predicted energies and forces.

6.4 Increasing the diversity of training data by active learning methods

To increase the diversity of the training data and improve the performance of the KRR model, I developed an active learning workflow. The key idea is to actively update the KRR model by adding new data points obtained from the KRR structure relaxation. The workflow of active learning is depicted in Fig. 6.4.

In the active learning (Fig. 6.4), I employed the Gaussian processes regression (GPR) model to evaluate the uncertainty of the KRR-relaxed structure. When a preset uncertainty threshold was reached, the KRR structure relaxation would be stopped. Here, the Python package used for the GPR model is the scikit-learn software package [164]. The same kernel and regularization parameters were used for KRR and GPR. In this work, when the uncertainty was larger than 0.4 eV or the number of relaxation steps was more than 40, the KRR relaxation was stopped.

To monitor the convergence of the overall active learning, I analyzed the energy distribution of the newly added data in each iteration. Active learning was considered to have converged when the mean energy of all the newly added structures stopped changing.

Our active learning method followed five steps:

1. Randomly picking 5000 data points from the 1000 DFT relaxation trajectories as the initial dataset (training data) to train the initial KRR
and GPR models.

2. Using the KRR model to relax 1000 random structures from the "data pool", while employing the GPR model to calculate the uncertainties of the KRR predictions. The relaxation was stopped when a predetermined uncertainty threshold or relaxation step count was reached, and the last structure of each relaxation trajectory was chosen for Step 3.

3. Calculating the DFT single-point energies of the 1000 structures from Step 2.

4. Adding the 1000 new structures and their DFT energies to the training data, and then refitting the KRR and GPR models.

5. If the active learning is converged, output the final KRR model as the final model; otherwise, repeat Step 2.

### 6.5 The effects of the active learning workflow and adding fore field simulations

I used our active learning workflow to generate training data for two different KRR models: the KRR model with force field simulation (AL1) and the KRR model without force field simulation (AL2).

Fig. 6.5a shows the energy distribution of the newly added structures in each active learning iteration for AL1. The peak of the 5000 initial data points is located at -15 eV. For the first iteration, the distribution peak of all the newly added structures is located at -2.16 eV. As the number of active learning iterations increases, the energy distribution peak decreases and is finally located at -6.28 eV. Fig. 6.5b shows the mean energy value of all the newly added structures in each active learning iteration of AL1. The curve continuously decreases about 4 eV, then becomes stable after the tenth active learning iteration, and only has negligible change ($\Delta E < 0.05$ eV).

From Fig. 6.5a,b, I concluded that the active learning for AL1 converged at the tenth active learning iteration. Hence, I chose the KRR model in the tenth iteration as the final KRR model (referred to as MBTR-KRR-FF-AL) for AL1. Similarly, for AL2, I concluded that the active learning converged at the eighth active learning iteration (Fig. 6.5b,d). I then chose the KRR model in the eighth iteration as the final KRR model (referred to as MBTR-KRR-AL) for AL2.

To compare the performance of the final models of AL1 and AL2, I utilized the MBTR-KRR-FF-AL model and the MBTR-KRR-AL model to relax 100 random structures from the "data pool" for 30 steps. Subse-
Applying machine learning to accelerate the structure search for the ligand-protected nanoclusters

Figure 6.5. The DFT single-point energy distribution of the 5000 initial structures and the newly added structures in each active learning iteration, (a) with force field simulation, (C) without force field simulations. “n-iter” in the figure (a) and (c) means the nth iteration of active learning iteration and “in” in these two figures respects “initial data”. The mean energy curve of all the newly added structures during the active learning, (b) with force field simulations, (d) without force field simulations. $E_0 = -13745440$. Adapted from Publication III.

Consequently, we employed DFT to calculate the single-point energies of the relaxation trajectories. Fig. 6.6 depicts the average DFT energy curve of the MBTR-KRR-FF-AL model and the MBTR-KRR-AL model relaxations. The average energy curve (blue line) of the MBTR-KRR-AL model continuously decreases about 5.4 eV, then increases after the 9th relaxation step. For the MBTR-KRR-FF-AL model relaxation, the curve (orange line) continuously decreases about 7 eV, then becomes stable after 18 relaxation steps. This suggests that adding force field energies and forces to the KRR model improved the model performance.

6.6 Evaluating the reliability of the optimal KRR model relaxations

After obtaining the final KRR model (the MBTR-KRR-FF-AL model), I calculated the correlation between different levels of simulations. I randomly selected 100 force field-relaxed structures ($S_{FF}$) from the "data pool" and determined their corresponding force field energies ($E_{FF}(S_{FF})$). Subsequently, I calculated the DFT single-point energies ($E_{DFT}(S_{FF})$) of the structures in $S_{FF}$. Using the MBTR-KRR-FF-AL model, I relaxed $S_{FF}$ to generate 100 KRR-relaxed structures ($S_{KRR}$) and obtained their corresponding ML-predicted energies ($E_{KRR}(S_{KRR})$). I also calculated the DFT
Applying machine learning to accelerate the structure search for the ligand-protected nanoclusters

Figure 6.6. (a) The average energy curve of the relaxation trajectories from MBTR-KRR-FF-AL in AL1 and MBTR-KRR-AL in AL2. Adapted from Publication III.

Table 6.1. Correlation between energies. $S_1$: $E_{FF}$; $S_2$: $E_{KRR}$; $S_3$: $E_{KRR} \rightarrow DFT$; $S_4$: $E_{DFT}$. $E_{FF}$: Energy calculated by force field methods. $E_{DFT}$: Energy calculated by DFT. $E_{KRR}$: Energy predicted by the KRR model. Adapted from Publication III.

<table>
<thead>
<tr>
<th>Correlation</th>
<th>$E_{FF}(S_1)$</th>
<th>$E_{DFT}(S_1)$</th>
<th>$E_{KRR}(S_2)$</th>
<th>$E_{DFT}(S_2)$</th>
<th>$E_{DFT}(S_3)$</th>
<th>$E_{DFT}(S_4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{FF}(S_1)$</td>
<td>1.00</td>
<td>0.82</td>
<td>0.76</td>
<td>0.73</td>
<td>0.56</td>
<td>0.54</td>
</tr>
<tr>
<td>$E_{DFT}(S_1)$</td>
<td>0.82</td>
<td>1.00</td>
<td>0.84</td>
<td>0.87</td>
<td>0.66</td>
<td>0.59</td>
</tr>
<tr>
<td>$E_{KRR}(S_2)$</td>
<td>0.76</td>
<td>0.84</td>
<td>1.00</td>
<td>0.78</td>
<td>0.63</td>
<td>0.58</td>
</tr>
<tr>
<td>$E_{DFT}(S_2)$</td>
<td>0.73</td>
<td>0.87</td>
<td>0.78</td>
<td>1.00</td>
<td>0.84</td>
<td>0.80</td>
</tr>
<tr>
<td>$E_{DFT}(S_3)$</td>
<td>0.56</td>
<td>0.66</td>
<td>0.63</td>
<td>0.84</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>$E_{DFT}(S_4)$</td>
<td>0.54</td>
<td>0.59</td>
<td>0.58</td>
<td>0.80</td>
<td>0.96</td>
<td>1.00</td>
</tr>
</tbody>
</table>

energies ($E_{DFT}(S_{KRR})$) of $S_{KRR}$. After this, I employed DFT to further relax $S_{KRR}$, obtaining new relaxed structures ($S_{KRR} \rightarrow DFT$) and their respective DFT energies ($E_{DFT}(S_{KRR} \rightarrow DFT)$). Finally, I applied DFT to relax $S_{FF}$, obtaining the final set of relaxed structures ($S_{DFT}$) and their corresponding DFT energies ($E_{DFT}(S_{DFT})$).

Given two energy sets $X = [x_1, \cdots, x_i, \cdots, x_n]$ and $Y = [y_1, \cdots, y_i, \cdots, y_n]$, the correlation between them can be calculated using the following equation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{(x_i - \overline{X})^2}\sqrt{(y_i - \overline{Y})^2}}, \quad (6.3)$$

where $\overline{X}$ and $\overline{Y}$ are the mean of $X$ and $Y$. I summarized the calculated results in Table 6.1. The correlation between $E_{FF}(S_{FF})$ and $E_{DFT}(S_{DFT})$ is 0.54, whereas the correlation between $E_{DFT}(S_{FF})$ and $E_{DFT}(S_{DFT})$ is 0.59. These findings suggest that we cannot directly obtain the correct low-energy candidate structures from force field-relaxed structures. Similarly, the correlation between $E_{KRR}(S_{KRR})$ and $E_{DFT}(S_{DFT})$ is 0.58, indicating that we cannot directly select the correct low-energy candidate structures based on the KRR-predicted energies. However, the correla-
Applying machine learning to accelerate the structure search for the ligand-protected nanoclusters

Figure 6.7. The DFT energy distribution of relaxed structures (a) 100% selecting, (b) top 1% lowest-energy structures in each. Adapted from Publication III.

The correlation between $E_{\text{DFT}}(S_{\text{KRR}})$ and $E_{\text{DFT}}(S_{\text{DFT}})$ is significantly higher at 0.80, indicating that the correct low-energy structures can be obtained from the DFT energies of the KRR-relaxed structures. This suggests that selecting the correct low-energy candidate structures based on the DFT energies of KRR-relaxed structures is reliable. Additionally, the correlation between $E_{\text{DFT}}(S_{\text{KRR}} \rightarrow \text{DFT})$ and $E_{\text{DFT}}(S_{\text{DFT}})$ is 0.96, which means relaxing a structure with KRR then followed by DFT of directly with DFT will obtain very similar final structures.

6.7 Results and discussion

After confirming the reliability of the MBTR-KRR-FF-AL model relaxations, I utilized this model to relax all the remaining structures (about 90,000) in the data pool that were not used for training ML models. From each trajectory, I selected the structure obtained at the 40th relaxation step as the final relaxed structure. Subsequently, I employed DFT calculations to compute the single-point energies of all final relaxed structures and summarized their energy distribution in Fig. 6.7a. We defined 0 eV to be the energy of the global minimum structure. The lowest energy of the structures is 0 eV for the II–II group, 2.16 eV for the O–II group, 3.13 eV for the II–O group, and 5.11 eV for the O–O group. The energy ranges for the II-II, O–II, II–O, and O–O groups are 41.31 eV, 18.04 eV, 39.95 eV, and 17.99 eV, respectively.

To focus on lower-energy structures and remove irrelevant structures, I selected only the top 1% of the lowest-energy structures from each group for further analysis. Fig. 6.7b illustrates the energy distribution of these structures. The energy ranges for the top 1% most stable structures in the II–II, O–II, II–O, and O–O groups are 1.33 eV, 1.14 eV, 1.75 eV, and 1.71 eV, respectively. Notably, in Fig. 6.7b, I observed that the distribution order of the groups from low energy to high energy is as follows: II–II, O–II, II–O, and O–O.
Applying machine learning to accelerate the structure search for the ligand-protected nanoclusters

6.7.1 Structure analysis of the low-energy structures

After analyzing the energy distribution of the top 1% most stable structures in each group, I proceeded to analyze their structural features. Specifically, for each selected structure, I computed the distance of the inner H-bond within each ligand, the hydrogen-oxygen (H–O) distance between ligands (where hydrogen and oxygen atoms are not from the same ligand), the hydrogen-sulfur (H–S) distance between ligands (where hydrogen and sulfur atoms are not from the same ligand), and the hydrogen-gold (H–Au) distance. Here, the shorter H–O distance indicates a stronger interaction between ligands, while smaller H–S and H–Au distances suggest stronger interactions between ligands and cluster parts. I also computed the hydrogen-carbon (H-C) distance between ligands (for cases when hydrogen and carbon atoms are not from the same ligand). I did not find hydrogen atoms bonded to carbons.

An initial structure in groups II–II, O–II, II–O, and O–O has 18, 12, 6, and 0 II-type H-bonds, and 0, 6, 12, and 18 O-type H-bonds, respectively. For the top 1% most stable structures, I counted the average number of H-bonds in one structure, as shown in Table 6.2. If I use a distance cutoff of $r_{HB} = 3.0 \, \text{Å}$ to define a H-bond, after KRR model relaxations, one structure on average retains 17.75, 11.58, 5.97, and 0 II-type H-bonds and 0, 5.76,
Table 6.2. The average number of H-bonds in one KRR-relaxed structure.

<table>
<thead>
<tr>
<th>Group</th>
<th>II–II</th>
<th>O–II</th>
<th>II–O</th>
<th>O–O</th>
</tr>
</thead>
<tbody>
<tr>
<td>II H-bond</td>
<td>17.75</td>
<td>11.58</td>
<td>5.97</td>
<td>0</td>
</tr>
<tr>
<td>O H-bond</td>
<td>0</td>
<td>5.76</td>
<td>10.77</td>
<td>16.94</td>
</tr>
</tbody>
</table>

10.77, and 16.94 O-type H-bonds in groups II–II, O–II, II–O, and O–O. Moreover, the cumulative distribution curves of the H-bonds a given in Fig. 6.8a, and the flat in each curve again suggests that most structures contain 18 H-bonds. I concluded that the top 1% most stable structures retain the features of H-bonds in cysteine ligands.

Fig. 6.8b shows the cumulative distribution curves of the H–O distance. Notably, the curves generated from groups II–O and O–O lie below the curves from the other two groups, particularly in the distance range from 2.2 to 2.4 Å. This reveals that when the B site binds with II-type H-bond ligands, the average distance of H–O in the structure is smaller compared to binding with O-type H-bond ligands, which may lead to strong interactions between ligands.

Fig. 6.8c, d show the cumulative distribution curves of the H-S and H-Au distances, respectively. The curves of the four groups in the two figures are very close to each other. From these figures, I did not observe any apparent influence of the type of hydrogen bonds on the interactions between ligands and cluster parts.

### 6.7.2 Electronic property analysis of the low-energy structures

To study the electronic properties, I calculate the density of state (DOS) of the top 1% most stable structures in each group. To obtain the total DOS of one group, I summed all DOS in each group with the Boltzmann distribution. For one group with \( N \) structures, the total DOS of a group is

\[
DOS = \frac{1}{Z} \sum_{i}^{N} [dos_i \exp\left(\frac{E_{\text{min}} - E_i}{k_B T}\right)],
\]

where \( dos_i \) and \( E_i \) are the DOS and the DFT energy of the \( i \)th structure, respectively. \( E_{\text{min}} \) is the minimum energy in this group, \( k_B \) is the Boltzmann constant, and \( T \) is the temperature. Here, \( \frac{1}{Z} \) is the normalization constant, and \( Z = \sum_{i}^{N} \exp\left(-\frac{E_i}{k_B T}\right) \).

Fig. 6.9a shows the total DOS of four groups at 300 K. I also summarized the energy of the highest occupied molecular orbital (HOMO), the lowest unoccupied molecular orbital (LUMO), and the HOMO-LUMO gap in Table 6.3. There is a shift in the HOMO/LUMO energies if the structures in the four groups, but the different configurations have negligible effects on the HOMO-LUMO gap. Previous DFT studies [118, 72] reported that \( \text{Au}_{25}(\text{SCH}_3)_1^{18} \) is an 8-electron superatom whose orbitals were mainly
Table 6.3. The HOMO, LUMO, and HOMO-LUMO gap if the structures in the four groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>II-II</th>
<th>O–II</th>
<th>II–O</th>
<th>O–O</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOMO/eV</td>
<td>-3.33</td>
<td>-3.44</td>
<td>-3.00</td>
<td>-2.98</td>
</tr>
<tr>
<td>LUMO/eV</td>
<td>-2.14</td>
<td>-2.29</td>
<td>-1.82</td>
<td>-1.79</td>
</tr>
<tr>
<td>HOMO-LUMO gap/eV</td>
<td>1.19</td>
<td>1.15</td>
<td>1.18</td>
<td>1.19</td>
</tr>
</tbody>
</table>

Figure 6.9. (a) The total DOS of four groups of II-II, O–II, II–O, and O–O. Using the same color dash line as the color of the DOS solid line, the HOMO/LUMO of each group is marked. (b) The DOS and the projected DOS of a structure from the II–II group. Adapted from Publication III.

contributed by Au and S atoms. The HOMO-LUMO gap of the cluster is 1.19 eV, consistent with our results. Near the HOMO and LUMO, the DOSs of the four groups are similar to each other. I checked the electronic properties of all computed structures and found that the electronic states of $\text{Au}_{25}\text{(Cys)}^{-1}_{18}$ are mainly contributed by the Au and S atoms as shown in Fig. 6.9b. I, and therefore infer that the $\text{Au}_{25}\text{(Cys)}^{-1}_{18}$ in this work is also an 8-electron superatom.

The similar DOS shape but different HOMO/LUMO energies indicate the cysteine ligands shifted the Kohn-Sham eigenvalues of the Au and S atoms in the clusters. In Fig. 6.9a, I observed that the HOMO/LUMO states of II–O and O–O are very close and located in a higher energy region. I also noted that the HOMO/LUMO states of II–II and O–II are close and located in a lower energy region. These suggest that the type of H-bonds in ligands on the B site is the main factor shifting the DOS of the structures.
7. Summary and Outlook

This chapter summarizes my research objectives and achieved results. An outlook to future work concludes this dissertation.

7.1 Summary

The first chapter of the thesis motivated the importance of the monolayer-protected nanocluster. In this chapter, I also acknowledged that it is impossible to directly search the structure of ligand layers with currently available methods. In the thesis, I, therefore, split this problem into three: (i) developing a molecular conformer search procedure for isolated molecules, (ii) improving the procedure to search structures for a single ligand molecule on a cluster, (iii) developing an active learning workflow to improve the performance of using KRR model to relax structures of the ligand-protected layer in ligand-protected nanoclusters.

I developed a molecular conformer search procedure based on BO to accurately and efficiently search structures for four amino acid molecules (Chapter 4). In our procedure, I used the BOSS code to predict the PES model of molecules. I monitored the convergence of the BOSS-predicted PES model from the global minimum but also from the local minimum. Our results show that DFT optimization plays a major role in refining the energy ranking of BOSS-predicted local minimum conformers. Moreover, our search results agree well with the previously studied results in reference papers. For searching structures for an amino acid with 5–6 degrees of freedom, our procedure only needs about 3400 DFT single-point energy calculations. This is a small computational cost, compared to other search methods.

Because of the steric clashes, our procedure cannot directly search molecules on a cluster. Hence, in Chapter 5, I developed three strategies to address the steric clashes during structure search by our procedure. The three strategies all have their advantages and limitations. From using the accuracy of the surrogate PES prediction and the final local minima
structures to evaluate the three strategies. Using a logarithmic energy transformation to suppress high energies and applying an energy penalty to nonphysical structures for which DFT cannot return the energy, Strategy iii, exhibits the best performance. Our search results show that each predicted local minimum structure has one hydrogen bond and the weak $H-S$ and $H-Au$ interaction in the structure help to stabilize the system. Among the top 10 most stable structures, those structures with type II hydrogen bonds are dominant.

In Chapter 6, I used the predicted low-energy structures in Publication II to generate 100,000 structures of $Au_{25}(Cys)_{18}^{-1}$. Based on the H-bond type of ligands, these structures were classified into four groups: II–II, O–II, II–O, and O–O. I trained a KRR model to efficiently optimize these structures. By combining GROMACS simulations, I addressed the problem of the nonphysical changing of bond angles during relaxations. I developed an active learning workflow to generate new data from the KRR structure relaxations and iteratively updated the KRR model. After this, I applied the model to relax about 90,000 structures and selected the top 1% most stable structures in each group for analyzing the energy distribution, the structural features, and the electronic properties. The analysis shows the structures in groups II–II are the most stable ones and the type II H-bond in cysteine is an important low-energy feature in $Au_{25}(Cys)_{18}^{-1}$. The H-bond type of ligands binding with the B site is the main factor causing the difference in interactions (between ligands and between ligands and the cluster). The different configurations of the ligand layer influence the structural and electronic properties of $Au_{25}(Cys)_{18}^{-1}$.

7.2 Outlook

In the thesis, I developed ML methods to search structures for the isolated molecules, a molecule on a nanocluster, and the ligand-protected layer of nanoclusters. These methods enable the exploration of isolated molecules and complicated nanocluster systems with atomic resolution. By utilizing these ML methods, it becomes possible to computationally obtain the global and local minimum structures of systems, providing valuable insights into the structure-property mechanism of these systems. It is important to note that the theoretical work was carried out for molecules and clusters in vacuum. Therefore, if other researchers want to compare our results with experiments, they should keep in mind that the thiolated-protected Au clusters are produced in solution and the solution may affect the structure and the properties of the clusters.

Our methods have achieved great progress, but there remains a need for improvement in several aspects. Based on my work, I discuss some open questions that might be worth addressing in future research.
The first one is how to extend our ML methods for molecules with higher dimensions. Our procedure can search structures for molecules with five to six dimensions. Collaborating with our co-author, Guo Xiaomi, we developed a search method based on a generative model named variational auto-encoder (VAE) to search structures for molecules with seven to nine dimensions [165]. In future work, I expect that the machine-learning approaches that combine BO with generative models can be used to search the conformers of the higher dimensional molecules.

The second aspect is how to obtain more precise KRR models for ligand-protected nanoclusters. In Publication III, I used a global representation, MBTR, to describe high symmetry structures. In future work, it would be interesting to test whether using local descriptors (such as atom-centered symmetry functions (ACSFs) and the smooth overlap of atomic positions (SOAP)) or combining global and local descriptors can enhance the precision of the KRR models.

Currently, machine-learned potentials, such as neural network potentials [166], Gaussian approximation potentials [167], spectral neighbor analysis potentials [168], moment tensor potentials [169], and atomic cluster expansion [170], are widely used in material science. However, the machine-learned potentials for ligand-protected nanoclusters remain to be developed. The accurate and general machine-learned potentials, which can be used for the molecular dynamic simulations of ligand-protected nanoclusters, will be extremely useful in studying the structural properties of nanoclusters.
References


References


References


References


References


References


References


References


