JYU DISSERTATIONS 531

Antti Pihlajamäki

Machine Learning Approach to Atomic Simulations of Protected Gold Nanoclusters

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF MATHEMATICS
AND SCIENCE

JYU DISSERTATIONS 531

Antti Pihlajamäki

Machine Learning Approach to Atomic Simulations of Protected Gold Nanoclusters

Esitetään Jyväskylän yliopiston matemaattis-luonnontieteellisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi Ylistönrinteen auditoriossa FYS1
kesäkuun 20. päivänä 2022 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Mathematics and Science of the University of Jyväskylä,
in building Ylistönrinne, Auditorium FYS1, on June 20, 2022, at 12 o'clock.

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2022

Editors
Ilari Maasilta
Department of Physics, University of Jyväskylä
Päivi Vuorio
Open Science Centre, University of Jyväskylä

ISBN 978-951-39-9309-2 (PDF)
URN:ISBN:978-951-39-9309-2
ISSN 2489-9003

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-9309-2

ABSTRACT

Pihlajamäki, Antti
Machine learning approach to atomic simulations of protected gold nanoclusters
Jyväskylä: University of Jyväskylä, 2022, 86 p. (+included articles)
(JYU Dissertations
ISSN 2489-9003; 531)
ISBN 978-951-39-9309-2 (PDF)

In the nanometer lengthscale, the boundaries between physics, chemistry and biology disappear and all phenomena are reduced to the level of atomic interactions. Technological advancement has provided means to measure what happens at the atomic level but there are limitations and experiments cannot tell everything. Here computational studies provide further insight. The most accurate computational methods are based on the quantum mechanics, which explains atomic interactions at the level of electrons. However, these methods are computationally demanding, which limits their usage. One can also compromise the accuracy and use efficient force field methods. During the last two decades a third type of method, machine learning (ML), has become increasingly popular. ML methods utilize data from other computational methods or measurements to "learn" underlying trends. This way they reproduce the behavior of the high-level methods with significantly reduced computational cost. Their usage is not restricted to imitate other methods but they can also be used for data analysis. In this thesis, four studies demonstrate three different applications of ML methods in studies of gold nanoclusters protected by organic ligands. Wavelet-based image comparison method was used to analyze experimental and theoretical transmission electron microscope (TEM) images of the crystal lattice made of nanoclusters. The analysis ruled out the possible structural isomer and shed light to the cluster orientation in TEM images. So-called distance-based ML methods were utilized for dynamic simulations of the similar clusters. Based on the given configuration the ML method predicted potential energies, which were used to run Monte Carlo simulations emulating the dynamics of the clusters. After this, a new distance-based ML method was designed to estimate forces affecting to the individual atoms of the nanoclusters. Estimated force vectors enabled ML assisted structure optimization of the goldthiolate systems. The results showed the great potential of the distance-based methods on simulations of the complex nanostructures.

Keywords: Machine learning, Computational, Clusters, Gold, Thiolate

TIIVISTELMÄ

Pihlajamäki, Antti
Koneoppimismenetelmät suojattujen kultananoklusterien atomistisissa simulaatioissa
Jyväskylä: Jyväskylän yliopisto, 2022, 86 s. (+alkuperäiset artikkelit)
(JYU Dissertations
ISSN 2489-9003; 531)
ISBN 978-951-39-9309-2 (PDF)

Nanometrimittaskaalassa fysiikan, kemian ja biologian rajat hamartyvat ja kaikki ilmiot voidaan kasittaa atomien valisten vuorovaikutusten kautta. Teknologinen kehitys on mahdollistanut yha tarkemmat mittaukset atomistiselta tasolta. Kokeellisilla menetelmilla on kuitenkin rajansa eika pienimpia mekanismeja pystyta havainnoimaan suoraan. Tata varten tarvitaan laskennallisia menetelmia. Tarkimmat ja luotetuimmat menetelmat pohjautuvat suoraan kvanttimekaniikkaan ja kasittelevat atomien vuorovaikutuksia elektronien tasolla. Nama menetelmat vaativat valtavasti laskennallisia resursseja, mika rajoittaa niiden kayttoa. Vaihtoehtona on kayttaa voimakenttia, jotka tinkivat tarkkuudesta, mutta parantavat laskujen tehokkuutta. Viimeisen kahdenkymmenen vuoden aikana ns. koneoppimismenetelmat ovat nousseet suureen suosioon. Koneoppiminen hyodyntaa dataa muista menetelmista tai kokeista ja pyrkii loytamaan riippuvuussuhteita. Nain koneoppimismenetelma "oppii" jaljittelemaan korkea tasoisten menetelmien tuloksia, mutta kayttaen vain murto-osan laskennallista resursseista. Koneoppiminen ei rajoitu vain muiden menetelmien matkimiseen, vaan se on erittain tehokas tyokalu mm. suurten datamaarien analysointiin. Tassa vaitoskirjassa esitetyt nelja tutkimusta havainnollistavat kolmea erilaista kayttotarkoitusta koneoppimismenetelimille. Tutkimuskohteena ovat suojatut kultananopartikkelit. Nama partikkelit koostuvat metallisesta ytimesta ja orgaanisesta suojaavasta ligandikerroksesta, joten ne ovat fysikaalisesti ja kemiallisesti hyvin kompleksisia. Ensimmaisessa tutkimuksessa aaltopakettipohjaista kuvien vertailumenetelmaa hyodynnettiin analysoimaan kokeellisia ja simuloituja lapaisyelektronimikroskooppikuvia, joissa havaittiin nanopartikkeleista muodostuneita kiteita. Taman pohjalta partikkelin rakenneisomeeri saatiin rajattua pois jatkoanalyysista. Seuraavaksi ns. etaisyyspohjaisia koneoppimismentelmia hyodynnettiin kultananopartikkelien dynaamisiin simulaatioihin. Ensin nailla menetelmilla ennustettiin systeemin potentiaalienergiaa partikkelin rakenteen pohjalta. Ennustettujen energioita kaytettiin Monte Carlo -simulaatioissa mallintamaan partikkeleiden dynamiikkaa. Seuraavaksi uusi etaisyyspohjainen koneoppimismenetelma kehitettiin arvioimaan yksittaisten atomien voimia. Naiden voimavektoreiden pohjalta suoritettiin rakenneoptimointeja erilaisille kulta-tioli-systeemeille. Tulokset osoittivat, etta etaisyyspohjaiset mentelmat soveltuvat oivallisesti monimutkaisten nanorakenteiden mallitukseen.

Author

Antti Pihlajamäki
Department of Physics
Nanoscience Center
University of Jyväskylä
Jyväskylä, Finland

Supervisor

Professor Hannu Häkkinen
Department of Physics
Department of Chemistry
Nanoscience Center
University of Jyväskylä
Jyväskylä, Finland

Co-Supervisor

Staff Scientist Sami Malola
Department of Physics
Nanoscience Center
University of Jyväskylä
Jyväskylä, Finland

Reviewers

Associate Professor Mie Andersen
Aarhus Institute of Advanced Studies
Aarhus University
Aarhus, Denmark

Assistant Professor Badri Narayanan
J.B. Speed School of Engineering
University of Louisville
Louisville, KY, USA

Opponent

Assistant Professor Milica Todorović
Department of Mechanical and Materials Engineering
University of Turku
Turku, Finland

PREFACE

The work presented in this thesis was carried out in the Nanoscience Center, Department of Physics at University of Jyväskylä during 2018-2022. I acknowledge the financial support from the Faculty of Mathematics and Science Doctoral school and from the Jenny and Antti Wihuri rahasto. I am also grateful to Academy of Finland for initiating the AIPSE funding program, which enabled our HNP-AI project in Jyväskylä. Despite initial scepticism the project proved to be a great success. Without this project, I would not have had this opportunity to carry out the research, which led to this thesis.

I express my utmost gratitude to my supervisors Hannu Häkkinen and Sami Malola. First and foremost, I am grateful that in 2018 I got the opportunity to be part of the project applying machine learning to nanoscience. Their guidance on nanoclusters, computational nanoscience and scientific writing have greatly helped me to grow as a scientist. I also want to thank Tommi Kärkkäinen and his group: Joakim Linja, Joonas Hämäläinen and Paavo Nieminen. They have been excellent collaborators and co-workers. From our discussions, I have learned immensely about machine learning, data mining and computations in general. Our group's Slack channel have also been a great source of support, especially from Joakim and Joonas. I also thank the current and past members of our Computational Nanoscience group for creating an open and inclusive working environment.

Even if some people might have not been directly involved in the work in this thesis, their role is not any lesser. I thank Pekka Koskinen for participating in my PhD thesis follow-up group and, most importantly, for giving an initial interest on computational nanoscience during my Masters studies 2016-2018. The work of Vesa Apaja and Juhani Forsman must be acknowledged. They have kept the computing clusters, Puck and Oberon, of our university running. Those machines have been in hard use during these four years.

Finally, I thank my family and friends for their support. My parents Pirjo and Jorma Pihlajamäki and my brother Timo Pihlajamäki have been a valuable counterbalance in my otherwise work-filled life. It has always been a pleasure to visit home, which is secluded from pretty much everything. I also want to mention the influence of my grandfather Antti Lehtinen. I have always reminisced with smile his questions about when I'll be a licentiate. Thank you for my two close friends Tero Lähderanta and Henri Nerg. Sharing our experiences about our own thesis work and life have been a joy during these years. In the end, I am especially grateful for Liu Yan, who has been my closest support and comfort. We have accompanied each other through the pandemic and without her it would have been a very lonely period of time.

Jyväskylä, April 2022

Antti Pihalajamäki

LIST OF INCLUDED ARTICLES

PI Qiaofeng Yao, Lingmei Liu, Sami Malola, Hongyi Xu, Zhenna Wu, Tiankai Chen, Yitao Cao, María Francisca Matus, **Antti Pihlajamäki**, Shuangquan Zang, Yu Han, Hannu Häkkinen and Jiangping Xie. Engineering Colloidal Crystals of Atomically Precise Gold Nanoparticles Promoted by Particle Surface Dynamics. *Accepted in Nature Chemistry*, 2022.

PII **Antti Pihlajamäki**, Joonas Hämäläinen, Joakim Linja, Paavo Nieminen, Sami Malola, Tommi Kärkkäinen, and Hannu Häkkinen. Monte Carlo Simulations of $Au_{38}(SCH_3)_{24}$ Nanocluster Using Distance-Based Machine Learning Methods. *Journal of Physical Chemistry A, 124 (23), 4827–4836*, 2020.

PIII **Antti Pihlajamäki**, Joakim Linja, Joonas Hämäläinen, Paavo Nieminen, Sami Malola, Tommi Kärkkäinen, and Hannu Häkkinen. Orientation adaptive minimal learning machine for directions of atomic forces. *ESANN 2021: Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning Online event (Bruges, Belgium), October 06 - 08, 529—534*, 2021.

PIV **Antti Pihlajamäki**, Sami Malola, Tommi Kärkkäinen, and Hannu Häkkinen. Orientation Adaptive Minimal Learning Machine: Application to Thiolate-Protected Gold Nanoclusters and Gold-Thiolate Rings. *arXiv: 2203.09788*, 2022.

In article [PI], the author implemented an image comparison method for experimental and simulated TEM images, ran the image comparisons, analyzed comparison results, reported them in the Supplementary Information of the article, and participated in the finalizing the main article.

In [PII], the author wrote the code to run Monte Carlo simulations, adjusted structure description parameters, trained Extreme Minimal Learning Machine (EMLM) method, ran Monte Carlo simulations, analyzed results and wrote the first draft of the article.

In [PIII], the author derived the new Orientation Adaptive Minimal Learning Machine (OAMLM) method for atomic force direction estimation on the basis of the original Minimal Learning Machine (MLM). The author also generated test data using Density Functional Tight-Binding method, analyzed the results and wrote the manuscript of the article.

In [PIV], the author refined OAMLM method, did extensive parameter testing for the method and structural descriptions, trained machine learning methods, implemented optimization algorithm compatible with the machine learning, ran all computations, analyzed results, wrote the first draft of the article and collected published material into the Gitlab (https://gitlab.jyu.fi/aneepihl/oamlm_forces.git).

Author also contributed in a following article:

API Sami Malola, Paavo Nieminen, **Antti Pihlajamäki**, Joonas Hämäläinen, Tommi Kärkkäinen and Hannu Häkkinen. A method for structure prediction of metal-ligand interfaces of hybrid nanoparticles. Nature Communications, 10, Article number: 3973 (2019), doi: https://doi.org/10.1038/s41467-019-12031-w

LIST OF ABBREVIATIONS

AI	Artificial intelligence
ANN	Artificial neural network
ASE	Atomic Simulation Environment
BFGS	Broyden-Fletcher-Goldfarb-Shanno
BoB	Bag of Bonds
CNN	Convolutional neural network
CPU	Central processing unit
CW-SSIM	Complex Wavelet Structural Similarity
ELM	Extreme Learning Machine
EMLM	Extreme Minimal Learning Machine
DFT	Density Functional Theory
DFTB	Density Functional Tight-Binding
GPU	Graphics processing unit
KRR	Kernelized Ridge Regression
MBTR	Many-Body Tensor Representation
MC	Monte Carlo
MD	Molecular dynamics
ML	Machine learning
MLM	Minimal Learning Machine
MPC	Mono-layer protected cluster
OAMLM	Orientation Adaptive Minimal Learning Machine
PBE	Pedrew-Burke-Ernzerhof
PCA	Principal component analysis
PET	Phenyl ethyl thiolate
p-MBA	*p*-mercaptobenzoic acid
RMSD	Root mean square displacement
RMSE	Root mean square error
SOAP	Smooth Overlap of Atomic Positions
SSIM	Structural Similarity
STM	Scanning tunneling microscope
TEM	Transmission electron microscope

CONTENTS

1 INTRODUCTION

Nanoscience is an extremely diverse research field, which focuses on the phenomena taking place in nanometer lengthscale. At this level, the boundaries of physics, chemistry and biology start to disappear. It does not matter whether one studies electronic excitations, plasmonics, catalytic reactions, protein folding etc., everything is dictated by the interactions between individual atoms and their electronic structure. It is possible to measure these phenomena experimentally and with tools like atomic force and helium ion microscopes, one could get a glimpse how matter actually looks like. However, chemical reactions are fast and imaging methods have their limitations, therefore all information cannot be acquired experimentally. Here computational methods can give further insight about what is happening at atomic level. There are numerous tools with varying complexity. Methods relying directly on quantum mechanics are accurate but they require large computational resources. In order to simulate large systems with more than thousands of atoms, one needs to compromise accuracy and emphasize efficiency, which could be done with force field methods. There does not exist a single method, which could handle everything with satisfying accuracy and high computational efficiency. Novel approaches and tools are needed to push the boundaries of the nanoscience even further.

This thesis focuses on the development of the computational methods for the simulations and analysis of so-called monolayer protected clusters (MPCs) using machine learning (ML) approaches. MPCs are metal nanoparticles, which consist of metallic core covered by protecting organic ligands such as thiolates, phosphines, alkynyls or carbenes [1]. The core size might vary from a few metal atoms to hundreds of metal atoms. Between the metallic core and the ligand layer, there is a metal-ligand interface, which often contains sulfur or phosphorus creating a linking structure between these chemically and physically very different environments. Ligands have a central role on the formation of the MPCs, because they passivate the metal particles and enable the synthesis of nanoparticles with exact size [1]. However, ligands are not just for stabilizing the MPCs but they can also be used for MPC functionalization. The possibility to tune both the metallic core and the ligands enable MPCs to be used in many applications such as nanomedicine, catalysis and biological imaging [1–8].

MPCs can be synthesized in the lab but there is no practical way to explicitly see how they behave at the atomic level, because their size is just around 1-5 nm and timescales of the chemical reactions are very fast. Here the computational methods can bring the needed insight by "making experiments in silico". Before anything can be calculated one needs a full atomic structure of the MPC in question. A common way the solve the atomic structure of the MPCs is to crystallize the sample and analyze its X-ray diffraction. One of the earliest crystal structures of the MPCs was $Au_{39}(PPh_3)_{14}Cl_6$ (Ph = phenyl) published in 1992 [9] and after discovery of the $Au_{102}(p{-}MBA)_{44}$ (p-MBA = *p*-mercaptobenzoic acid) in 2007 [10] there has been a boom of new crystal structures with various ligands and metals [1].

In this thesis I focus on two well-known gold MPCs: $[Au_{25}(SR)_{18}]^-$ and $Au_{38}(SR)_{24}$. Here, R denotes organic part of the thiolate ligand. The atomic structure of $[Au_{25}(PET)_{18}]^-$ (PET = phenyl ethyl thiolate) was initially predicted theoretically [11] and later in 2008 the crystal structure was determined about the same time by two groups [12, 13]. Afterwards its structure was also successfully determined with charge states of 0 and 1+ [14, 15]. $Au_{38}(PET)_{24}$, on the other hand, is a peculiar nanocluster as it has two experimentally found stable structural isomers. The first structure was determined in 2010 by Qian *et al.* [16] and that is why I refer it as a Q isomer. The another crystal structure was found by Tian *et al.* 2015 [17], referred as isomer T. The Q isomer is cylinder shaped and isomer T is oblate-like as seen in figures 1.1 (b) and (c). From these two structural isomers, Q is thermodynamically more stable than T, which has been shown both experimentally and computationally [17–19]. The structural differences are not limited only to the shape but their ligand shells also differ significantly. This can be highlighted by writing their chemical formula using the "divide and protect" idea [20]. This means that the metallic core and protecting layer can be thought as separate entities and naturally notation should emphasize it. This way Q isomer could be written as Au_{23}@$[SR\text{-}Au\text{-}SR\text{-}Au\text{-}SR]_6[SR\text{-}Au\text{-}SR]_3$ and T isomer Au_{23}@$[SR\text{-}Au\text{-}SR\text{-}Au\text{-}SR\text{-}Au\text{-}SR]_2[SR\text{-}Au\text{-}SR\text{-}Au\text{-}SR]_3[SR\text{-}Au\text{-}SR]_3[SR]_1^b$, where superscript *b* refers to a bridge site. This notation tells us that Q isomer has three short gold-thiolate units and six long units. T isomer, on the other hand, has two even longer units and a single bridge site thiolate. $[Au_{25}(SR)_{18}]^-$ in 1.1(a) is significantly more simple structure than $Au_{38}(SR)_{24}$. It has a icosahedral core with 13 gold atoms and six gold-thiolate unit. This can also be written as Au_{13}@$[SR\text{-}Au\text{-}SR\text{-}Au\text{-}SR]_6$.

After an atomic structure is available, the computational work can start. There are

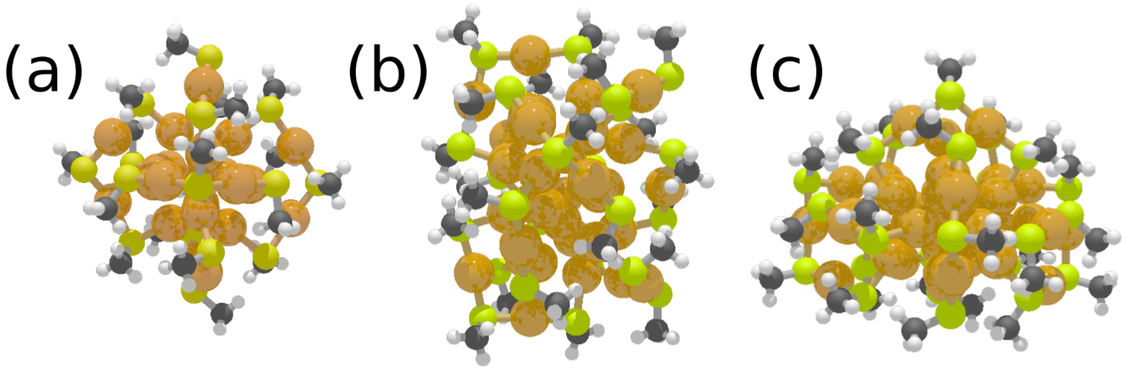


FIGURE 1.1 (a) The structure of $[Au_{25}(SCH_3)_{18}]^-$ [12, 13]. Two structural isomers of the $Au_{38}(SCH_3)_{24}$ nanocluster: (b) the Q isomer [16] and (c) the T isomer [17]. The $[Au_{25}(SR)_{18}]^-$ has 13 gold atom icosahedral core and six gold-thiolate units. The $Au_{38}(SCH_3)_{24}$ structures consist of 23 gold atom core and various gold-thiolate units. The protecting units are $[\text{Au-SCH}_3]_x$ oligomers with different lengths. The organic parts of the thiolate ligands are simplified to be methyls. Colors: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen. Panels (b) and (c) are reprinted from the article [PIV] (arXiv: 2203.09788).

numerous different computational methods and approaches, which all have their own advantages and drawbacks. The most trusted computational methods are based on quantum mechanics and among them the density functional theory (DFT) is one of the most commonly used ones. The grounds for DFT were established 1964 by Hohenberg and Kohn, when they showed that wavefunctions can be replaced by electron density [21]. Since then, several different DFT codes have been developed. However, the accuracy comes with a cost. Quantum mechanical methods require lots of computational resources, which limits their accessibility and the size of the simulated systems.

If one wants to simulate large systems containing even more than thousands of atoms, force fields are a viable option. They are computationally lighter than quantum mechanical methods due to the their simplified mathematical form. However, they require adjustment of parameters to fit their potential functions to suit the system at hand. For MPCs, there are existing ReaXFF [22] and AMBER-GROMACS [23] forces fields readily available. The drawback of the force fields is that simplifications lose some properties and fitting parameters is a laborious process. The third emerging approach is to use ML methods, which are the main focus in this thesis.

The idea of ML methods is to use data from other sources, find underlying trends from it and then reproduce the behaviour according to "learned" relations. ML methods can never exist on their own, because they rely on data, which has to be generated either experimentally or computationally. However, they are extremely versatile tools as they can be trained to emulate the behavior of the high-level computational methods with significantly reduced computational cost. One could consider them to belong to the midground of the *ab initio* methods and force fields, because in optimal case they have capabilities to reach accuracy almost equal to *ab initio* methods but their computational efficiency often does not reach the force fields [24]. The ML algorithms are general in the sense that they are not restricted to any specific task but one could modify them to "learn" almost anything. There are numerous examples where ML methods have been used successfully to create accurate ML force fields and potentials for atomic simulations [24–26]. They also have many application in material informatics [27, 28], catalysis research [29, 30], spectroscopy [31], study phase transitions of water [32, 33] and they can even be trained to build materials [34–36].

Before proceeded to explain the contents of the thesis, some terminology should be clarified. In every day language, term artificial intelligence (AI), machine learning (ML), deep learning etc. are entangled or even used interchangeably. However, they have certain differences and should not be confused with each other. AI is the high-level term and it can be considered to include several methods such as ML, computer vision, image recognition, clustering etc.. These methods can also have certain degree of overlap. For example, sometimes image recognition can be consider to be a ML method and ML methods can be used to enhance image recognition. AI, as term itself, can also be divided into subgroups. There are two main categories "weak AI" and "strong AI" [37]. "Weak AI" is goal oriented and works only within certain well defined field. "Strong AI" is supposedly very human like or could even exceeded human capabilities. It should be general, able to adapt into many problems and have a level of strategic consideration. There is no real "strong AI" developed currently but all AI methods available belong to "weak AI". In this thesis,the term AI is not discussed any further. The lower level term ML is much more important term to the thesis than AI.

It was just mentioned that ML algorithms are designed to learn underlying trends and to improve according to "past experiences". However, there are number of different ML methods and each one of them have their own properties and flavors. In order to get further perspective on ML, it is useful to divide it to three main categories: supervised ML, unsupervised ML and reinforcement learning [38]. Supervised ML is the most important concept for this thesis. The idea is that there is some labeled data, for example atomic structures and corresponding potential energies, and the method will fit itself to the data. The most rudimentary example would be a simple curve fitting. One has some values on x-axis and their corresponding values y. Fitting a curve to this data would be strictly speaking the most primitive form of supervised ML. In practise, fitting of a ML model is done in high-dimensional space, dimensionality of which is related to the degrees of freedom of the studied system. Unsupervised ML works with data without labels (or mostly without labels). These methods try to find inner trends from the data without further guidance. Many data clustering methods fall into this category. Reinforcement learning is collection of methods where machine learns to perform tasks according to given feedback. These methods are often used in robotics and games. For example, if a computer wins a game, it will get positive feedback guiding it towards some action over others, and from losing it will get negative feedback.

One could roughly point out two major categories of ML: kernel-based methods and methods based on artificial neural networks (ANNs). Kernel-based methods use reference data, which is usually a subset of the training data, to measure similarities/differences between data points using some kernel function and the output is predicted according to these similarities. This is called a "kernel trick" and it allows one to project data from the original data space to kernel space where regression or classification is done. These similarity-based approaches form a basis for all the methods presented in this thesis. However, let us also briefly consider the other option. ANNs consist of layers of nodes or neurons, which are connected via weights. The operation of ANNs is based on collections of large weight matrices, which perform regression from one layer to another. When the number of layers increases the methods are referred as deep ANNs, which are currently very popular as they have shown great potential to master very difficult tasks, such as playing board game Go [39–41] and simulating protein folding [42]. ANNs are flexible tools because of their large number of fitted weights. However, this is also their major drawback. In order to optimize all the weights and parameters of deep ANNs, one is required to spend significant amount of computational resources often accompanied with GPUs. There is also a high risk of overfitting due to the flexibility. Furthermore, ANNs work as "black boxes" and user does not really know what is happening inside the method, because deep ANNs are complex and extremely difficult to visualize with our limited human visualization capabilities. By using kernel-based methods instead, we aim to develop more interpretable and understandable models than ANNs and also to add further customization possibilities.

There are already many ML packages and approaches for simulations of atomic systems [43–53], therefore one could ask why one would need more. A common practise to validate these methods is to use datasets of relatively small molecules [54–59] or homogeneous/symmetric systems such as metals [60, 61]. Datasets with similar atomic systems can also be combined in order to increase the diversity of the data and to improve the generalization capabilities of the ML models [62]. MPCs, on the other hand, are a challenging systems for ML, because they consist of chemically very different parts. In addition to the studies presented in this thesis, there have been a few successful examples of applying ML methods to the research of MPCs. For example, the synthesis and properties of MPCs have been studied with ANNs [63] and support vector machines [64], and rule-based methods have been applied to the construction of metal-ligand interface [65]. Due to the chemical and physical complexity of the MPCs, the ML model has to be able to handle a wide range of interactions or to split the problem into smaller parts to simplify the task. This is not a trivial task to do and ML methods with special functionalities are required.

Having discussed a general picture about MPCs and ML methods, let us go through the contents of this thesis. The thesis demonstrates three cases of the usage of data-driven and ML methods. The first part focuses on linking experimental results to computational analysis in article [PI]. In the study, $[Au_{25}(p{-}MBA)_{18}]^{-}$ nanocluster was experimentally discovered within a crystalline structure, where nanoclusters were linked together. A wavelet-based image comparison method was applied to the analysis of experimental and simulated transmission electron microscope (TEM) images. The second article [PII] demonstrates how distance-based ML methods can be used to simulate atomic dynamics of the $Au_{38}(SCH_3)_{24}$ nanocluster. According to our knowledge, this was the first time when distance-based ML methods were used in the simulations of MPCs. ML methods were trained to predict potential energy values for the configurations of the two structural isomers of $Au_{38}(SCH_3)_{24}$. Monte Carlo algorithm was used to emulate the dynamics of the clusters at different simulation temperatures. In the the third and fourth articles [PIII] and [PIV] the focus is shifted from handling of potential energies to atomic forces. In [PIII], the Orientation Adaptive Minimal Learning Machine (OAMLM) method for force direction estimation is introduced for the first time and its operation was demonstrated with simple carbohydrate chains. In [PIV], the force directions were combined with predicted norms, when distance-based ML methods were used to estimate force vector subjecting to the individual atoms in $Au_{38}(SCH_3)_{24}$. The task was split into two part: predicting the norm of the force and estimation of the direction. Handling directional information is not a simple task for the ML methods, therefore the existing distance-based ML methods were heavily modified creating a novel approach to the atomic force estimation. The performance of the method was validated with structure optimization. The test system included not only $Au_{38}(SCH_3)_{24}$ structures but also gold-thiolate rings and $Au_{25}(SCH_3)_{18}$, which were not explicitly included into training data.

This thesis thesis is structured in a following manner. In section 2 the theoretical background of the used methods is presented. The image comparison method is presented first and then structural descriptors, which were used to create representations about the atomic structures/environments. Then the theory behind the distance-based ML methods is introduced going from the simplest method to the most complex one. The theory part is finished with the description how ML methods were applied to the Monte Carlo simulations and structure optimization. Section 3 discusses the results and finding from articles [PI], [PII], [PIII] and [PIV].

In the end, everything is summarized in Conclusions section 4

2 THEORETICAL BACKGROUND

In this section the theoretical backgrounds of the used computational tools are introduced. First wavelet-based image comparison is introduced. After this the focus shifts towards the actual ML methods by explaining the structural descriptors used in this thesis. There are two different descriptors discussed: Many-Body Tensor Representation (MBTR) and Smooth Overlap of Atomic Positions (SOAP). The descriptors are used to predict potential energy values and atomic force vectors, therefore naturally the third part of this section explains the underlying ideas of the distance-based ML methods. This followed by the discussion on how the ML methods were applied to simulations of the MPCs in Monte Carlo dynamics and structure optimization. In the end, the used DFT methods are summarized briefly.

2.1 Wavelet-based image comparison

Image recognition and computer vision are fundamental and well-known applications of AI and ML methods. At the same time they are also probably the most common type of ML, which is encountered in everyday life as many people are familiar with facial recognition and QR code reading. There are numerous ways to approach the task every method having their own characteristics. The method used here to compare experimental and simulated TEM images is originally proposed by Simoncelli *et al.* and it is called Complex Wavelet Structural Similarity (CW-SSIM), which utilizes wavelet transformations to compare two images [66, 67].

Before going into the details of the CW-SSIM, we should consider what kind of data TEM images are, how the simulated images are acquired and what factors have to be taken into account during the comparison. Experimental TEM images are taken by focusing an electron beam through the sample, which in this case would be packed $[Au_{25}(p{-}MBA)_{18}]^-$ cluster lattice. The regions, where electrons are absorbed by the sample, are seen as gray or black and the background is white due to the lack of absorption. Only relatively heavy atoms are seen, because they have the highest capability to absorb electrons. Light atoms, such as carbon and hydrogen, are not visible in the TEM images.

This process can be imitated computationally in a simple manner, if the atomic structure is available. First the direction of an imaginary electron beam is decided and all atomic coordinates are projected to the plane perpendicular to the beam direction. Then the projected atoms are considered as two-dimensional Gaussian-style functions within the plane. One can calculate the value of the ith pixel by summing up the contributions as

$$\phi_i = \sum_{j=1}^{N_{atoms}} Z_j^{1.5} e^{\frac{-|\mathbf{r}_j - \mathbf{p}_i|^2}{\eta^2}} . \tag{2.1}$$

Here $\mathbf{p}_i$ denotes the position of the ith pixel and $\mathbf{r}_j$ is the position of the atom j projected to the plane. Width parameter η was set to be 1.0 Å. Z_j is an atomic number, which is used to emphasize heavy atoms over light ones. The scaling based on the previous studies on the dependence of image intensity on atomic number [68].

There is a fundamental difference between experimental and simulated TEM images. In the experiments, when the electron beam is fully absorbed in some region of the sample, anything behind it is hidden and is seen as a flat black area in the image. The computational approach does not have this kind of "cut-off behavior" but everything contributes to the pixel values. Hence, when the simulated data is transformed to the actual images, all pixel values are scaled resulting into a smooth images with inner structure more visible than in the experimental images. This difference is visualized in figure 2.1. In figure 2.1(a) the image has flat black regions due to the "cut-off behavior" but in 2.1(b) all peaks are visible, when cut-off is not present. As a side note, image comparison methods have been used to analyze scanning tunneling microscope (STM) images of a $Ag_{374}(TBBT)_{113}Br_2Cl_2$ (TBBT = tert-butyl benzenethiolate) cluster [69]. However, the method relying on minima and maxima of the STM images cannot be used, because the fundamental difference between TEM and STM is similar to the difference between experimental and simulated TEM images.

The origin of the CW-SSIM is in the Structural Similarity (SSIM) method, which uses a sliding window over compared images and calculates similarity using luminescence, contrast and structure of the images [70]. However, it is very sensitive to noise and misalignment of the images [66, 67]. This makes it ill-suited

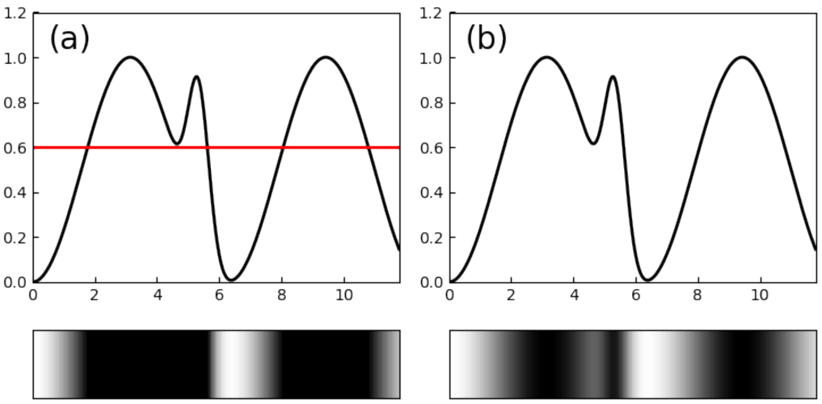


FIGURE 2.1 The conceptual difference between experimental and simulated TEM images is visualized in (a) and (b). Vertical axis represents the thickness of the sample lying on the horizontal axis. The panel (a) visualizes experimental situation, where everything is absorbed if region is thicker than 0.6 shown with horizontal red line. The simulated TEM in (b) shows inner features caused by the center peak. The bars at bottom side give an example how TEM images would look like.

for comparing TEM images, because the images are very likely aligned differently and there are fundamental differences between experimental and simulated TEM images as mentioned above. The motivation of the CW-SSIM is that by using different wavelets one could pick out wanted features for the comparison. The authors of the method also showed that wavelet-based comparison is less sensitive to noise and possible misalignment of the images [66]. Similar idea is also seen in convolutional neural networks (CNNs), where different filters are used to analyze certain features in data by calculating convolutions [71]. However, CNNs usually learn their filters instead of using predefined ones.

The first task is to calculate a wavelet transformation of the images to be compared. Originally Simoncelli *et al.* used so-called Steerable Pyramid decomposition method, which relies on Fourier transformations and series of high-pass, low-pass and orientation filters [66, 72–74]. However, in this study the conventional wavelet transformation was done using convolutions. The basic convolution of two functions is calculated as

$$[f * g](t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)\mathrm{d}\tau \tag{2.2}$$

Moving in to discrete 2D representation, convolution can be written as

$$[f * g](\mathbf{t}) \approx \sum_{i=-N_h}^{N_h} \sum_{j=-N_v}^{N_v} f((i \cdot \Delta_h, j \cdot \Delta_v))g(\mathbf{t} - (i \cdot \Delta_h, j \cdot \Delta_v))\Delta_h \Delta_v \tag{2.3}$$

Here N_h and N_v denote the number of steps included in the calculation in each direction. Constants Δ_h and Δ_v are the discrete step sizes in horizontal and vertical directions. In the actual wavelet transformation $f(\cdot)$ would be the image, $g(\cdot)$ complex conjugate of the wavelet and **t** would be a vector pointing to the center of the wavelet [75]. In practise, convolutions are done via implementation in the Scipy package [76].

There are many choices for possible wavelets and one might even want to use several ones to gain more statistics. Here we used so-called Rickers wavelet, which is also called Mexican Hat wavelet, because of its shape. The 2D Rickers wavelet is a negative normalized second derivative of the Gaussian function written as [75]

$$g(\mathbf{r}) = \frac{2}{\sqrt{3\beta}\pi^{1/4}} \left(1 - \left(\frac{|\mathbf{r}|}{\beta}\right)^2\right) e^{\frac{-|\mathbf{r}|^2}{2\beta^2}}, \tag{2.4}$$

where β determines the width of the wavelet. Rickers wavelets have been used in computer vision [77], therefore they are a safe first choice for this image comparison.

After two images are wavelet transformed the similarity measure is calculated. The CW-SSIM measure is computed as [66]

$$\begin{aligned} S(\mathbf{w}_x, \mathbf{w}_y) \;=\; & \left(\frac{2\sum_{i=1}^{N} |w_{x,i}||w_{y,i}| + K}{\sum_{i=1}^{N} |w_{x,i}|^2 + \sum_{i=1}^{N} |w_{y,i}|^2 + K}\right) \\ & \cdot \left(\frac{2|\sum_{i=1}^{N} w_{x,i} w_{y,i}^*| + K}{2\sum_{i=1}^{N} |w_{x,i} w_{y,i}^*| + K}\right) \end{aligned} \tag{2.5}$$

Here x and y denote two images to be compared. Vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ contain the values of corresponding transformed images, and $w_{x,i}$ and $w_{y,i}$ are their vector elements. Here it is important to realize that the compared values should be taken from the same positions of the analyzed images, therefore the order of the indices is crucial. Parameter K is a small number used to stabilize the calculation and avoid division by zero. In this study it was set to 0.01. The similarity value varies between 0 and 1, where 1 indicates identical images and 0 completely different ones.

2.2 Structural descriptors

When it comes to handling atomic structures, ML methods and conventional computational tools have some major differences. Quantum mechanical methods calculate electronic structure according to atomic positions, and force fields locate their potentials similarly to corresponding spatial coordinates. However, if one uses coordinates as an input to ML method, there will be many complications. As a simple example, let's consider two atom system, where atoms are moving along the x axis. By changing the distance one can gather data used to train the ML model. If one uses coordinates as an input to the model, it probably works in the beginning. Problems start to emerge when the system is translated or rotated in the space. This changes the input and ML model naturally assumes that the output also changes, even though the distance between the atoms has not changed. The order of atoms is also a factor in this case, because input AB is different than BA.

In order to cope with this issue, it is a good practise to use so-called descriptors to create representations of the systems. These mathematical methods are representing an atomic system in translation, rotation and permutation invariant way. Continuity and uniqueness are also desired properties of a good descriptor [78]. In the simple example case earlier, the pairwise distance between the atoms would be a good descriptor, because it fulfills all requirements mentioned above.

In the focus of this thesis there are two often used descriptors called Many-Body Tensor Representation (MBTR)[79] and Smooth Overlap of Atomic Positions (SOAP)[80]. MBTR is an global descriptor, which creates a single representation for a full atomic structure. It has been used in the article [PII]. SOAP is a local descriptor, which is used to describe a chemical environment of a single atom or a point around an atomic structure. SOAP is in an important role in the article [PIV]. Even if these two descriptors are in the focus of this thesis, they are far from being the only descriptors available. Depending on the application there are numerous options for descriptors [81], for example Coulomb matrix [82], Ewald sum matrix [83], Atom-centered symmetry functions [49], Atom-density representation [84], zernike descriptors [85, 86] and Bag of Bonds [87] to name a few.

MBTR takes the ideas of so-called Coulomb Matrix [82] and Bag of Bonds (BoB) [87] descriptors taking them step further. Coulomb Matrix takes all N atoms in the system forming an $N \times N$ matrix were every matrix element contains a number calculated from the distance between the atoms and their atomic numbers. As such this matrix is not a permutation invariant description but it has to be diagonalized [82]. In the BoB, on the other hand, atom pairs from the Coulomb Matrix are grouped in to "bags" according to their element pairs and put in to descending order. This leads us to the roots of MBTR.

MBTR groups atoms according to their element and goes through all atom pairs or triplets measuring some properties between them. The basic approach is to measure inverse distance between the atom pairs. Every measured inverse distance is saved, weighted and Gaussian broadened. The main calculation is written as

$$f(x, Z_1, Z_2) = \sum_{i=1}^{N_{Z_1}} \sum_{j=1}^{N_{Z_2}} w(\mathbf{r}_i, \mathbf{r}_j) D(x, g(\mathbf{r}_i, \mathbf{r}_j)). \tag{2.6}$$

Here we used an exponential weight $w(\mathbf{r}_i, \mathbf{r}_j) = \exp(-\alpha|\mathbf{r}_i - \mathbf{r}_j|)$. Function $g(\mathbf{r}_i, \mathbf{r}_j)$ represents the measured property i.e. inverse distance $1/|\mathbf{r}_i - \mathbf{r}_j|$. N_{Z_i} is a number of atoms with atomic number of Z_i. The Gaussian broadening is applied with

$$D(x, g) = (\sigma\sqrt{2\pi})^{-1} \text{exp}\left(\frac{(x-g)^2}{2\sigma^2}\right). \tag{2.7}$$

Parameter α in weighting and σ in broadening are constants adjusted according to system and task at hand. The variable x is the heart of MBTR description. It is a sweeping variable, which is used to probe the values of function (2.7). Ideally it would be continuous but in practise it is discrete with n_x values between a given interval. Now it becomes clear where MBTR has gotten its name. If system has $N_{element}$ number of different elements, MBTR forms all pair of elements (originally in both orders), probes the n_x values from the function (2.6) with all element pairs and in the end the final representation is $N_{element} \times N_{element} \times n_x$ tensor[78, 79].

In order to generate MBTR description in practise one has to define set of parameters $\{\text{min}, \text{max}, n_x, \sigma, \alpha, \text{cutoff}\}$. Here "min" and "max" refer to the minimum and maximum values of the sweeping variable x. The parameter α is the multiplier in the exponential weighting term and "cutoff" determines how small values are included into the summation in equation (2.6). This "cutoff" parameter basically determines the lower boundary for how small faraway contributions are considered meaningful.

It has to be mentioned that this is just a single form of MBTR. There at least two other forms, which are often used. The simplest one contains just one summation and it measures a number of atoms of specific elements. It is usually not used alone but along other descriptions. The third one has three summations and measures the angles formed by the atom triplets [78]. This is computationally significantly more demanding than the previous two as it has to go through three nested loops. These are the three most commonly used forms of MBTR but in principle there is no definite rules how it has to be implemented. For example, the property depicted on function $g(\cdot)$ does not have to be an inverse distance or angle but, in principle, one is free to modify it to suit ones purposes. I will not go into details of other types of MBTR but shall retain the focus on the pairwise representation, which have been used in this thesis.

SOAP is an elegant local descriptor, which approaches the task from totally different perspective than MBTR. I will focus on the implementation within DScribe package [78] as it is the one used in this thesis and it is more intuitive than the original version by Bartók *et al.* [80]. Instead of measuring distances or angles between atoms, SOAP uses atom density to describe the chemical environment of an atom. The atom density field is written as

$$\rho^Z(\mathbf{r}) = \sum_{i=1}^{N_Z} e^{-\frac{|\mathbf{r}-\mathbf{R}_i|^2}{2\sigma_{SOAP}^2}}. \tag{2.8}$$

Z denotes the atomic number and $\rho^Z(\mathbf{r})$ is a density of atoms with atomic number Z. The point or atom, which will be described, is located at $\mathbf{r}$. $\mathbf{R}_i$ is the position of atom i and σ_{SOAP} is a parameter, which defines the amount of broadening used in the description. Small σ_{SOAP} means that the atom densities are narrowly focused to the locations of atoms and this would generate very specific description. This is good, if one aims for very accurate and specific models. However, it will compromise transferability of ML models, because they will be very accurate close to their training region but perform poorly when the description changes. Large broadening makes chemical environments look more alike than in the case of small broadening. This naturally increases transferability of the ML methods but the cost is reduced accuracy. σ_{SOAP} is one of the most crucial parameters within the SOAP.

Atom density itself is not a useful description of a chemical neighborhood. It does not fulfil any requirements of a good descriptor. To form an actual description, SOAP uses a trick familiar from basic quantum mechanics: atom density is represented as a series of radial basis functions and spherical harmonics. Then the atom density can be written as

$$\rho^Z(\mathbf{r}) = \sum_{nlm} c_{nlm}^Z b_n(r) Y_{lm}(\theta, \phi), \tag{2.9}$$

where c_{nlm}^Z is a coefficient, $b_n(r)$ is a radial basis function and $Y_{lm}(\theta, \phi)$ represents spherical harmonics. In DScribe package authors have simplified the model by using only real spherical harmonics [78] instead of four-dimensional hyperspherical harmonics in the original model [80]. Next the coefficients are solved similarly as one would solve an eigenvalue problem in basic quantum mechanics, therefore

$$c^Z_{nlm} = \int\int\int dV\, b_n(r) Y_{lm}(\theta, \phi) \rho^Z(\mathbf{r}). \tag{2.10}$$

The coefficients c^Z_{nlm} still contain directional information as spherical harmonics have a spatial orientation. The clever trick to hide this information is to input them into a power spectrum. In DScribe package authors write the spectral elements as

$$p^{Z_1,Z_2}_{nn'l} = \pi\sqrt{\frac{8}{2l+1}} \sum_m \left(c^{Z_1}_{nlm}\right)^* c^{Z_2}_{n'lm}. \tag{2.11}$$

These power spectrum values form a rotation, translation and permutation invariant description. However, there are the radial basis functions, which are not defined yet and they are a significant difference between the original SOAP and the one implemented in DScribe. Bartók *et al.* used polynomial basis [78, 80] but Himanen *et al.* point out that this requires numerical integration to solve the coefficients [78]. Instead of polynomial basis, they utilize Gaussian type orbitals, hence radial the basis functions are written as

$$b_n(r) = \sum_{n'=1}^{n_{max}} \beta_{nn'l} \phi_{n'l}(r) \tag{2.12}$$

$$\phi_{n'l}(r) = r^l e^{\alpha_{n'l} r^2}. \tag{2.13}$$

The advantage of Gaussian type orbitals is that the integrations in (2.10) can be solved analytically. A coefficient α_{nl} is optimized during the description so that $\phi_{n'l}$ is close to zero when r approaches cut-off radius r_{cut} determined prior the description process. Coefficients $\beta_{nn'l}$ ensure that $\phi_{n'l}$ are orthonormal. Himanen *et al.* solve them by Löwdin orthogonalization [78, 88].

$$\boldsymbol{\beta} = \mathbf{S}^{\frac{1}{2}} \tag{2.14}$$

$$S_{nn'} = \langle \phi_{nl} | \phi_{n'l} \rangle = \int_0^\infty dr\, r^2 r^l e^{\alpha_{nl} r^2} r^l e^{\alpha_{n'l} r^2}, \tag{2.15}$$

where **S** is so-called overlap matrix.

In practise, none of the summations go to infinity but they go through values up to maxima n_{max} and l_{max}, which are parameters given by the user. They also play a crucial role on the description accuracy. The integers m are restricted by l similarly same way as in quantum mechanics, hence $m \in [-l, l]$. In addition to this, only atoms within cut-off radius r_{cut} are used in the description. This reduces the computational burden and localizes the description within some confined region. This confinement can be used to reduce the effect of far away atoms to the description. When determining the descriptions the fitted parameters are n_{max}, n_{max}, σ_{SOAP} and r_{cut}.

2.3 Distance-based machine learning methods

Distance-based machine learning methods are a type of kernel-based ML methods. Kernel-based methods contain reference points and they make output prediction according to the similarity measures between given input data and these references. There are several ways to measure the similarity and distance-based methods use an Euclidean distance. Measuring similarity and using that as a tool to perform regression or classification is often referred as a "kernel trick". It means that a method transforms a given input from the original data space into a kernel space representation, dimensionality of which is always equal to the number of reference points.

In this thesis I have used three different distance-based machine learning methods: Minimal Learning Machine (MLM) [89], Extreme Minimal Learning Machine (EMLM) [90] and Orientation Adaptive Minimal Learning Machine (OAMLM) [[PIII], [PIV]]. From these three, EMLM is the most simple one having references only in the input side. The idea behind EMLM and its name is based on a ML method called Extreme Learning Machine (ELM) [91–95]. ELM is one kind of an ANN or a perceptron with only one hidden layer with special training methods. MLM has references in both input and output sides and it is performing a regression between these two distance spaces. The output of the MLM is calculated by solving a multilateration problem, where one searches a data point based on the predicted distances between references and the target. OAMLM is a specialized version of the MLM designed especially for atomic force directions. Instead of performing regression between input spaces, it predicts angular information from the input space distances.

Let us go through the theoretical background of the distance-based ML methods starting from the simplest one: EMLM. The method starts with the input data $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times n_x}$ and corresponding output data $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times n_y}$. In the case of EMLM, K reference points $\mathbf{Q} = \{\mathbf{q}_j\}_{j=1}^{K} \in \mathbb{R}^{K\times n_x}$ from the input data $\mathbf{X}$. The training is done via regularized least-squares fitting [90]

$$\min_{\mathbf{W}\in\mathbb{R}^{K\times n_y}} J(\mathbf{W}) = \frac{1}{2N}\sum_{i=1}^{N}\left|\mathbf{d}_i^T\mathbf{W} - \mathbf{y}_i^T\right|^2 + \frac{\beta}{2K}\sum_{i=1}^{K}\sum_{j=1}^{n_y}|W_{ij}|^2. \tag{2.16}$$

Here the vector $\mathbf{d}_i \in \mathbb{R}^K$ contains Euclidean distances between the ith input data point and the references in $\mathbf{Q}$. The weight matrix $\mathbf{W} \in \mathbb{R}^{K \times n_y}$ performs the linear regression from kernel space to the output. Regularization can be adjusted with constant β. Usually it is a very small positive number but for noisy data it can be useful. The regularization makes the model less sensitive to outliers. If the model is though metaphorically to be a piece of wire, regularization controls its stiffness. In this thesis regularization is set to the square root of the machine epsilon.

The optimal solution for the weight is obtained by calculating the zero point of the first derivative of the equation (2.16). In a matrix form this is written as

$$\frac{1}{N}\mathbf{D}^T\left(\mathbf{D}\mathbf{W} - \mathbf{Y}\right) + \frac{\beta}{K}\mathbf{W} = 0, \tag{2.17}$$

which then yields

$$\left(\mathbf{D}^T\mathbf{D} + \frac{\beta}{K}\mathbf{I}\right)\mathbf{W} = \mathbf{D}^T\mathbf{Y}. \tag{2.18}$$

Matrix $\mathbf{D} \in \mathbb{R}^{N \times K}$ contains all Euclidean distances between the training input data points in $\mathbf{X}$ and references in $\mathbf{Q}$. Equation (2.18) is a practical representation of the solution, because it can be directly solved with any numerical optimization method. In order to predict output for an arbitrary input, one has to form a distance vector between the input and references $\mathbf{d} \in \mathbb{R}^K$. The output is computed via matrix multiplication $\mathbf{d}^T\mathbf{W}$. At this point one might realize that EMLM is fundamentally a Kernelized Ridge Regression (KRR) with Euclidean distance kernel function [38, 90].

MLM introduced by de Souza *et al.* [89] goes a step further by having references in both input and output spaces. In addition to input space references $\mathbf{Q}$, there are also corresponding references $\mathbf{T} = \{\mathbf{t}_j\}_{j=1}^K \in \mathbb{R}^{K \times n_y}$ in output space. The regression won't be done between input side distances and an output but MLM performs regression between two distance spaces as

$$\mathbf{D}_{out} = \mathbf{D}_{in}\mathbf{B} + \epsilon. \tag{2.19}$$

$\mathbf{D}_{in} \in \mathbb{R}^{N \times K}$ contains Euclidean distances between N input training data points in $\mathbf{X}$ and K reference points in $\mathbf{Q}$. $\mathbf{D}_{out} \in \mathbb{R}^{N \times K}$, on the other hand, consists of distances between training output data in $\mathbf{Y}$ and output references $\mathbf{T}$. $\mathbf{B} \in \mathbb{R}^{K \times K}$ is a weight matrix that performs the linear regression. The residual ϵ is assumed to be close to zero and it is here for the sake of completeness. The training of the MLM is done by solving the weight matrix $\mathbf{B}$. de Souza *et al.* show that the approximate solution is [89]

$$\mathbf{B} = \left(\mathbf{D}_{in}^{T}\mathbf{D}_{in}\right)^{-1}\mathbf{D}_{in}^{T}\mathbf{D}_{out}. \tag{2.20}$$

Output prediction with MLM is done in two parts. First output space distances are predicted from input space distances as

$$\mathbf{d}_{out}^{T} = \mathbf{d}_{in}^{T}\mathbf{B} \tag{2.21}$$

In the second part the actual output is acquired by solving the multilateration problem using $\mathbf{d}_{out}$ and $\mathbf{T}$. There is a variety of methods to solve this [96]. In the publication [PII] we are dealing with scalar output data, therefore the task is relatively straightforward. The idea is to minimize the objective function

$$\min_{y\in\mathbb{R}} J(y) = \sum_{k=1}^{K}\left((y-t_k)^2 - (\mathbf{d}_{in}^{T}\mathbf{B})_k^2\right)^2, \tag{2.22}$$

where $\mathbf{d}_{in} \in \mathbb{R}^K$ is a vector containing distances between an input $\mathbf{x}$ and the input reference points in $\mathbf{Q}$ and t_k is the kth output reference. One has to find an output y, which minimizes the objective function. As introduced by Mesquita *et al.*, differentiation leads in to a cubic equation and the minimum or minima are found from zero points of the first derivative [97]

$$Ky^3 - 3\sum_{k=1}^{K} t_k y^2 + \sum_{k=1}^{K}\left(3t_k^2 - (\mathbf{d}_{in}^{T}\mathbf{B})_k^2\right)y + \sum_{k=1}^{K}\left((\mathbf{d}_{in}^{T}\mathbf{B})_k^2 - t_k^3\right) = 0. \tag{2.23}$$

This is fundamentally a cubic equation of a form $ay^3 + by^2 + cy + d = 0$ and it has one to three unique roots. From the possible real valued roots the one that yields the smallest value of the objective function (2.22) is chosen as an output.

In order to predict vectorial output, for which the direction is crucial, one needs to develop an alternative approach. OAMLM utilizes the idea of handling input and output spaces separately similarly as introduced in the case of conventional MLM. However, instead of working with the distances in the output space, OAMLM works with angular information [PIII]-[PIV]. The idea is to align reference local atomic environments with the input environment and then predict angles between the corresponding reference unit force vectors and a target vector.

With previous derivations at our disposal, we can go through the foundations of the OAMLM. Input space training data $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times n_x}$ contains local descriptions of the chemical environments, such as SOAP descriptions introduced earlier. This kind of descriptor does not contain any information about the orientation of the system, because they are designed to be rotation invariant as discussed in the section 2.2. In order to enable OAMLM to adapt to the spatial orientation of the data, it also needs coordinates of neighboring atoms $\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times(1+M)\times 3}$. In $\mathbf{p}_i$ the first row is the position the atom of interest followed by M neighbors, which are selected with some suitable sampling scheme. Different sampling schemes are discussed later in the section 2.4. For every training data point there are also corresponding unit force vectors collected into $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times 3}$, where $|\mathbf{y}_i| = 1$ for all values of i. From this data, K reference data points are sampled into $\mathbf{Q} = \{\mathbf{q}_j\}_{j=1}^{K} \in \mathbb{R}^{K\times n_x}$ for chemical descriptors, $\mathbf{S} = \{\mathbf{s}_j\}_{j=1}^{K} \in \mathbb{R}^{K\times(1+M)\times 3}$ for coordinates of the neighboring atoms and $\mathbf{T} = \{\mathbf{t}_j\}_{j=1}^{K} \in \mathbb{R}^{K\times 3}$ for unit force vectors.

Atomic environments can be in any spatial orientation, therefore the coordinates of the neighboring atoms in **P** and **S** are used to align environments. This can be done in several ways depending on the system at hand. We used a method presented by Arun *et al.*, which utilizes Singular Value Decomposition (SVD) to calculate rotation matrix to align to sets of points [98]. For the OAMLM it is not crucial to get a perfect alignment of the atomic neighborhoods but getting systematic results is more important. First the atoms, which are subjected by the forces, are moved to the origin and their neighbors are translated together with them to preserve the general positioning. Then matrix $\mathbf{A}_{i,j} \in \mathbb{R}^{3\times 3}$ is formed by calculating it as

$$\mathbf{A}_{i,j} = \sum_{k=1}^{1+M} (\mathbf{p}_{i,k} - \mathbf{p}_{i,1})(\mathbf{s}_{j,k} - \mathbf{s}_{j,1})^T. \tag{2.24}$$

The indexing refers to the ith input and the jth reference neighborhood. With SVD one can split this matrix in a following manner $\mathbf{A}_{i,j} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T$. The rotation matrix is calculated as $\mathbf{R}_{i,j} = \mathbf{V}\mathbf{U}^T$. Arun *et al.* point out an interesting feature in this alignment scheme. If the determinant of the matrix $\mathbf{R}_{i,j}$ is -1 instead of 1, $\mathbf{R}_{i,j}$ contains an reflection [98]. This reflection can be fixed but fortunately this kind of behavior seldom happens. It is important to realize that the summation in equation (2.24) depends on the order of the atomic neighborhood coordinates. In a general case, one would have to test all permutations as done in the article [PIII]. However, by implementing specific selection rules for neighborhood points the number of tested permutations can be reduced immensely. This kind of rules were used in [PIV] and they are discussed in the section 2.4.

After the alignment is done, the accuracy is calculated as

$$g_{i,j} = \frac{1}{1+M} \sum_{k=1}^{1+M} |(\mathbf{p}_{i,k} - \mathbf{p}_{i,1}) - \mathbf{R}_{i,j}(\mathbf{s}_{j,k} - \mathbf{s}_{j,1})| \tag{2.25}$$

or

$$g'_{i,j} = \frac{1}{1+M}\sqrt{\sum_{k=1}^{1+M} |(\mathbf{p}_{i,k} - \mathbf{p}_{i,1}) - \mathbf{R}_{i,j}(\mathbf{s}_{j,k} - \mathbf{s}_{j,1})|^2}. \tag{2.26}$$

The rotation matrices $\mathbf{R}_{i,j}$ are not only used to align environments but they are also used to rotate reference unit force vectors in $\mathbf{T}$. This way reference force vectors are comparable with vectors in $\mathbf{Y}$. Dot products between these vectors are calculated as $\hat{\mathbf{y}}_i \cdot (\mathbf{R}_{i,j}\hat{\mathbf{t}}_j)$. This dot product represents two different aspects: the cosine of the angle between two vectors and the projection of the vector $\hat{\mathbf{y}}_i$ on to the rotated reference vectors. The projection implication is clarified in the output estimation part. The $g^{(\prime)}_{i,j}$ and dot products are used to form matrices $\mathbf{D}_g = \{g_{i,j}\} \in \mathbb{R}^{N\times K}$ and $\mathbf{D}_c = \{\hat{\mathbf{y}}_i \cdot (\mathbf{R}_{i,j}\hat{\mathbf{t}}_j)\} \in \mathbb{R}^{N\times K}$.

The training of the OAMLM is done the same way as the training of the MLM with equation (2.20). $\mathbf{D}_{in}$ is the Euclidean distances between datapoints in $\mathbf{X}$ and $\mathbf{Q}$ as before. However, $\mathbf{D}_{out}$ is different. In the article [PIII], two weight matrices are trained: $\mathbf{B}_c$ to predict dot products and $\mathbf{B}_g$ to predict alignment accuracy. By substituting $\mathbf{D}_{out}$ with either $\mathbf{D}_c$ or $\mathbf{D}_g$ one can solve corresponding weight matrices. The predictions done with $\mathbf{B}_g$ are used to estimate uncertainty and it might be useful on some applications. In the article [PIII] this was used to select the most reliable neighborhood environments to the final direction estimation. However, in the article [PIV], the alignment accuracy estimation is omitted to simplify the method.

In the article [PIII] the SOAP descriptions were sub-optimal, therefore Huber regression [99] was used to introduce statistical robustness into the training process in addition to one shown in equation (2.20). This is done in a similar fashion as in robust MLM method by Gomes *et al.* [100]. Huber regressor is a linear regression model, which weights outlying data points linearly and others with squared weight. The Huber parameter p_{Huber} determines how strictly data points are classified as outliers. Decreasing the p_{Huber} parameter means that more points are classified as outliers, which aims for a more robust model. Formally, the optimization is written as [99]

$$\min_{\mathbf{w},\mu,c} J_{Huber}(\mathbf{w},\mu,c) \;=\; \sum_{i=1}^{N}\left[\mu + g\left(\frac{\mathbf{x}_i^T\mathbf{w} + c - y_i}{\mu}\right)\mu\right] + \alpha\sum_{j=1}^{M}|w_j|^2, \tag{2.27}$$

where

$$g(z) = \begin{cases} z^2\,, & \text{if } |z| \leq p_{Huber} \\ 2p_{Huber}|z| - p_{Huber}^2\,, & \text{if } |z| > p_{Huber}. \end{cases} \tag{2.28}$$

Here $\mathbf{x} \in \mathbb{R}^{N\times M}$ are input data points and $\mathbf{y} \in \mathbb{R}^N$ are corresponding outputs. Weights $\mathbf{w} \in \mathbb{R}^M$, intercept c and parameter μ are optimized during the fitting process. The intercept values are used to improve linear fit, when the data is not centered to the origin. The parameter α determines the regularization same way as β in equation (2.16). Here it is set to be small. In practise, the training with Huber regression was done by optimizing every column of the weight matrix in $\mathbf{B}_{c/g}$. For every optimization run, matrix $\mathbf{D}_{in}$ and a column of $\mathbf{D}_{c/g}$ were used to substitute $\mathbf{x}$ and $\mathbf{y}$ in the Huber regression in equation (2.27).

Direction estimation with OAMLM has three parts: prediction of dot products (and alignment accuracies, if wanted), fitting of the reference environments with the input environment and estimation of the direction. The inputs are description $\mathbf{x}$ and neighborhood coordinates $\mathbf{p}$. Vector $\mathbf{d}_{in}$ is formed by calculating Euclidean distances between $\mathbf{x}$ and reference points in $\mathbf{Q}$. Weight matrices are used to predict dot products and alignment accuracies as $\mathbf{d}_{c/g}^T = \mathbf{d}_{in}^T\mathbf{B}_{c/g}$. Then reference neighborhood coordinates in $\mathbf{S}$ are aligned with $\mathbf{p}$ yielding real alignment accuracies. Using the corresponding rotation matrices reference unit vectors in $\mathbf{T}$ are rotated accordingly to match the orientation. The final task is to find vector $\hat{\mathbf{v}}$ pointing to the estimated force direction. As in the case of multilateration problem in MLM, the vector $\hat{\mathbf{v}}$ is found by minimizing the difference between the real and predicted dot products. In the articles [PIII] and [PIV] two different schemes are used to find the direction. The first way is to numerically optimize a cost function

$$\min_{\hat{\mathbf{v}}\in\mathbb{R}^3} J_1(\hat{\mathbf{v}}) = -\sum_{j=1}^{K} \exp\left(-\left(\frac{d_{c,j} - (\mathbf{R}_{i,j}\hat{\mathbf{t}}_j)\cdot\hat{\mathbf{v}}}{\sigma_1}\right)^2 - \left(\frac{g_{i,j}^{(\prime)}}{\sigma_2}\right)^2 \right), \tag{2.29}$$

We call this as a numeric loss function. In the article [PIII], not all data points were included into the loss function but only those reference environments j, for which $g_{i,j} \leq d_{g,j}$. In the article [PIV] this selection was omitted and all references were used. The minimization was done with Sequential Quadratic Programming (SQP) implemented in SciPy package [76].

The drawback of the loss function in equation (2.29) is that there are two weighting parameters σ_1 and σ_2, which need to be tested. In order to simplify the direction estimation, we also used a loss function with parabolic nature.

$$\min_{\hat{\mathbf{v}} \in \mathbb{R}^3} J_2(\hat{\mathbf{v}}) = \frac{1}{2} \sum_{j=1}^{K} \omega_{i,j} \left[\hat{\mathbf{v}} \cdot (\mathbf{R}_{i,j} \hat{\mathbf{t}}_j) - d_{c,j} \right]^2 , \tag{2.30}$$

where

$$\omega_{i,j} = \exp\left(- \left(\frac{g_{i,j}^{(\prime)}}{\sigma_2} \right)^2 \right) . \tag{2.31}$$

The advantage of the equation (2.30) is that it can be solved analytically by taking derivative respect to $\hat{\mathbf{v}}$ and as a result

$$\hat{\mathbf{v}}_i = \frac{\sum_{j=1}^{K} \omega_{i,j} d_{c,j} (\mathbf{R}_{i,j} \hat{\mathbf{t}}_j)}{\sum_{j=1}^{K} \omega_{i,j}} . \tag{2.32}$$

Now the projection nature of the dot products becomes clear, because the solution is a weighted average of predicted projections. In practise, the level of numeric error present in the values of $\mathbf{d}_c$ and $\omega_{i,j}$ causes that $\hat{\mathbf{v}}_i$ is not a unit vector. Before applying this estimation in the ML framework it has to be divided with its norm. Due to the analytic solution of the loss function, we call equation (2.30) as analytic loss function.

The different ways of EMLM, MLM and OAMLM to predict an output are intriguing. EMLM is the simplest, as it just measures the Euclidean distance distances and directly produces output. MLM and OAMLM need separate methods to find an optimal solution, because MLM predicts distances in output space and OAMLM yields dot products with respect to 3-dimensional output space directions. For example, in the 2D output space multilateration problem of the MLM can be thought as a search of a point, where the circles with predicted radii cross. This simple example is visualized in figure 2.2 (a). The direction estimation within OAMLM is actually a surprisingly similar problem. Initially it might seem like the method has to search a line, where the cones centered to the output reference vectors would cross. This is somewhat difficult to visualize. More natural way to approach this, is to think that every reference vector is a point on the surface of a unit sphere. Hence, the estimated output lies in the crossing of the circles on the spherical surface seen in figure 2.2(b). OAMLM carries certain amount of uncertainty and not all of these circles would cross at the same point, therefore weighting and averaging are needed.

There are numerous ML tools developed, so one might naturally ask why to use distance-based ML methods. First of all, the basic MLM and EMLM have only

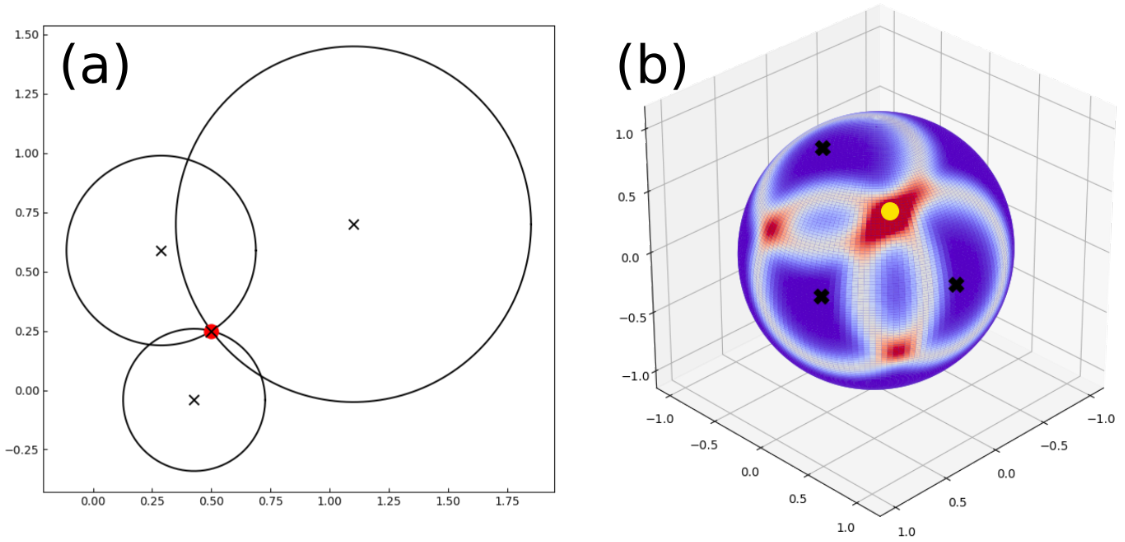


FIGURE 2.2 MLM solves its output via multilateration problem, where it has predicted Euclidean distances between output space references and yet unknown result. The simple 2D case of the multilateration is shown in (a). The "x"s are representing reference points, the lengths of the circle radii are predicted distances and the red dot is the result. OAMLM can be thought to estimate its output by doing similar search on the surface of the unit sphere in (b). Black "x"s are again references and the yellow dot is the optimal solution.

one hyperparameter, the number of reference points (the regularization of the EMLM is not considered here). The lack of many hyperparameters makes models easy to use. In contrast, Artificial Neural Networks (ANNs) have several ones, such as the widths and number of the hidden layers and learning rate [71]. When applying ML tools to the atomic structures, there are often several parameters in the structural descriptors. This means that the user has to optimize the way how the data is presented prior to the actual training of a ML method. Having few hyperparameters reduces the need for complex model fitting with different parametrizations of the descriptor. Furthermore, distance-based ML methods rarely overfit, if data is high-dimensional [101]. Similarly high-dimensional data might cause difficulties for ANNs. Linja *et al.* have shown that the construction of the distance matrix in MLM hides dimensionality and allows an exemplar performance compared to ANNs [102].

The dimensionality of the data is a crucial aspect to be considered, when choosing and developing ML methods. Verleysen and François have presented a splendid example about "the curse of dimensionality" [103]. They show that due to the dimensionality some geometric properties behave against our everyday understanding. For example, the volume ratio of the sphere and cube approaches zero in high-dimensions. These changes affect the data sampling and commonly used Gaussian style kernel functions, because increasing the dimensionality increases the expected distance between the data points, therefore data is sparser and Gaussian functions look flatter than in the low-dimensions. This is why Gaussian-based methods might encounter difficulties, when the dimensionality is high. However, distance-based methods should not be as vulnerable. The changes in Euclidean distances still exist but Verleysen and François point out that the relative differences between the distances is still decreasing and they recommend to use Minkowski distance instead.

Furthermore, it is easier to analyze what causes certain behavior of the distance-based ML methods than ANNs, because of the reference points. For example, one might analyze how densely references have been selected and what kind of distances given input possesses. In comparison, it is significantly more difficult to analyze connections within large ANNs. If one cannot interpret why model works as it works, there is a risk of "clever Hans effect" [104]. This means that the model does not learn expected features of the data but it might find some peculiar aspects. For example, in image recognition this kind of artificial features might be labels, text or background effects in the images instead of actual objects in the image. As a conclusion, distance-based ML methods are reliable and they are quick to be prepared for testing without tedious hyperparameter optimization.

2.4 Atomic neighborhood alignment for force estimation in gold-thiolate systems

This section presents the alignment schemes used in the OAMLM for gold-thiolate systems. In the article [PIV], the ML method was trained on the basis of $Au_{38}(SCH_3)_{24}$ nanocluster. This MPC contains four different elements with chemically versatile environments. These atoms can be divided into five different categories: core gold, unit gold, sulfur, carbon and hydrogen. All environments of these atoms have their own characteristics and the alignment should address this.

For hydrogen, the environment is dictated by the nearest carbon and two other hydrogen atoms bound to the same carbon atom. These atoms are used to align the environment of a hydrogen atom and there are only two permutations of the neighbor hydrogen atoms to be tested. This is shown in figure 2.3 (a). The neighborhood environment of a carbon atom in methyl thiolate ligand contains one sulfur atom and three hydrogen atoms as seen in figure 2.3 (b). Hence, there are six permutations of the hydrogen atoms to be tested during the alignment. In the case of sulfur atoms, the alignment uses the bound carbon atom and two nearest gold atoms, order of which has to be confirmed. The sulfur neighborhood is visualized in 2.3 (c).

Unit gold atoms are ideally bound to two sulfur atoms with somewhat covalently and the S-Au-S bond angle is approximately 180°. Because of this linearity, other atoms are also required to make reliable alignment. Hence, the nearest carbon and another gold atom are selected for both sulfur atoms. This way the unit gold alignment relies on two blocks of sulfur, gold and carbon atoms shown in figure 2.3 (d). There are some special cases, where unit gold atom has only one sulfur atom within 3.0 Å. These gold atoms are handled like "half unit" gold atoms, where the alignment is done with only one block. The alignment accuracy of hydrogen, carbon, sulfur and unit gold atoms is evaluated with equation (2.25).

The gold core is the most complicated environment to be aligned, because metallic core of the $Au_{38}(SCH_3)_{24}$ is flexible and the environments can be very homogeneous. There are two type of core gold atoms: atoms with one Au-S bond and atoms with only metallic interactions. For all core gold atoms, maximum of 12 nearest neighbors are selected within maximum distance of 5.0 Å. If there is a sulfur atom within 3.0 Å then it is selected first and then the rest of the neighbors are gold atoms.

If both core gold atom environments to be aligned contain sulfur atoms, then the alignment is relatively straightforward. For the input environment the sulfur and the nearest gold atom alongside the atom itself are used. For the reference environment the first two atoms are the same (gold atom itself and sulfur) but instead of using just the nearest neighbor gold atom, all gold atoms are tested one by one. Hence, every alignment is done with three points.

If either one of the core gold environments does not include sulfur, then the alignment is done using only gold atoms. For input environment two nearest gold atoms are used, but for reference environment all possible pairs of the neighbors are formed. This results into triangles that are first compared to find suitable candidates for alignment. The difference is estimated as

$$u_k = \sum_{i=1}^{3} [(l_{k,i} - l_{0,i})^2 + (\theta_{k,i} - \theta_{0,i})^2], \tag{2.33}$$

where $l_{k,i}$ and $l_{0,i}$ are the lengths of the ith side of the triangles from the data of kth reference and input respectively. Correspondingly, $\theta_{k,i}$ and $\theta_{0,i}$ are angles. Side lengths are given in ångstroms and angles in radians. From these triangles ten with lowest u_k were selected for actual alignment. The alignment accuracy was estimated with (2.26). An example of the core gold atom neighborhood is shown in the figure 2.3 (e).

Implementing physical and chemical information into the ML model is related to the earlier rule-based method for constructing metal-ligand interface based on the positions of the metal atoms [65]. This way the ML method does not have

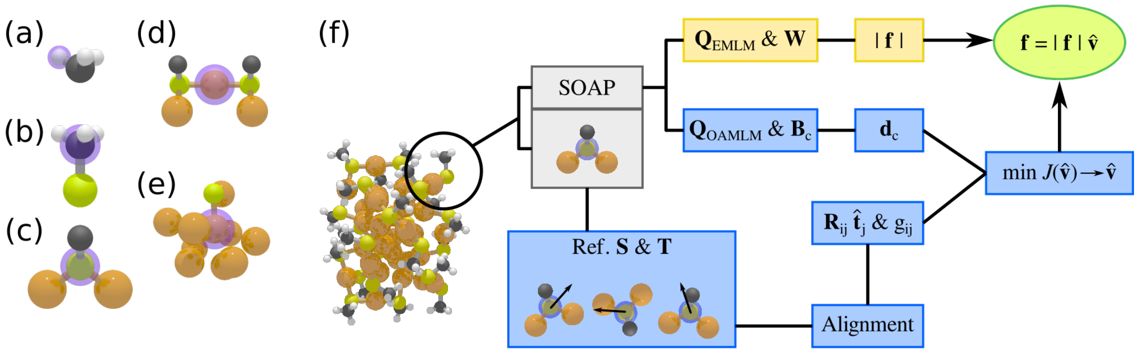


FIGURE 2.3 Atomic environments are used for alignment, which enables OAMLM to adjust itself to the spatial orientation of the input. Examples of the aligned atoms for the (a) hydrogen, (b) carbon, (c) sulfur, (d) unit gold and (e) core gold. The actual input atoms are highlighted with purple. Panel (f) demonstrates the ML framework for the full force estimation, where the description part is shown in grey boxes, norm prediction with EMLM in yellow and the direction estimation of the OAMLM in blue boxes. Colors for atoms: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen. Reprinted from the article [PIV] (arXiv: 2203.09788)

to fit everything based on the statistics but the user can guide it towards desired performance. The connection of the atomic environment handling and the force estimation is visualized in the figure 2.3 (f). It shows how the environmental data is first collected and divided into the different parts of the ML method. EMLM predicts force norms according to the SOAP description. OAMLM uses the same SOAP description as EMLM to predict dot products and via alignment the reference unit force vectors are rotated to correspond to the spatial orientation. The force direction is estimated and multiplied with the predicted norm resulting into a ML estimated force vector for the given atom.

2.5 Monte Carlo based dynamics for MPCs

In the article [PII], the EMLM and MLM were used to predict potential energy values according to the MBTR description and there were no forces estimated. In principle, the force vectors could have been computed with numeric differentiation, if analytic solution was proved impractical. However, there is always a concern whether the forces would conserve the energy and is the ML energy landscape smooth enough for differentiation. However, forces are required to run molecular dynamics (MD) simulations. MD simulations are often run using some imaginary heat bath such like in Berendsen [105], Langevin [106, 107] or Nosé-Hoover dynamics [108, 109]. This would be a risky approach to test new force fields, both conventional and ML based, because the interaction with imaginary heat bath via friction term and random noise-like additions might hide the effects of nonphysical forces. It would be more rigorous test to use Velocity Verlet algorithm [110] as it runs plain Newtonian equations of motion conserving energy but the drawback would be the lack of control over the simulation temperature.

In contrast to the MD, Monte Carlo (MC) methods do not need any information about the forces. They are just performing a random walk on the potential energy surface resulting into an imitation of the real dynamics. In the article [PII] we decided to simulate dynamics with MC instead of MD. The random walk was performed for $Au_{38}(SCH_3)_{24}$ by giving every part in the structure a possibility to move. There are three different moving parts: gold atoms, sulfur atoms and methyl groups. The gold atoms are moved randomly into any direction according to the given step size. The sulfur atoms are moved the same way but the alignment of the S-C bond is preserved by slightly rotating the methyl group visualized in the figure 2.4(a). The methyl groups are moved as a single block keeping C-H bonds fixed. The methyl groups were pivoted with S-C bond, the length of which can change according to the step size. This is shown in the figure 2.4(b). The methyl group can also be rotated around the S-C bond. The preservation of the S-C bond alignment is implemented to prevent hydrogen atoms from wandering between sulfur and carbon. This is not a physical or chemical phenomenon, therefore the ML method was not trained to handle this and it would not produce reliable potential energy predictions.

During a full MC step, every moving parts is gone through in random order and a movement is proposed. The probability of the proposal to be accepted is determined with Metropolis question [111]

$$P = \min\left\{1, \exp\left(\frac{-(E_{i+1}-E_i)}{k_B T}\right)\right\}. \tag{2.34}$$

E_i is the potential energy of the ith configuration, E_{i+1} is the potential energy of the configuration after a proposed move, k_B is the Boltzmann constant and T is a simulation temperature. Going downhill in energy landscape is always permitted but going uphill is accepted with certain probability defined by the energy difference and simulation temperature. The step size is adjusted during the simulation. Too large step size would cause all proposed moves to be rejected and too small step would lead acceptance of every move. Hence, the step is adjusted during the simulations so that the acceptance of the moves is between 40% and 60%. This step size is the same throughout the whole cluster and it is not affected by the type of the moved part.

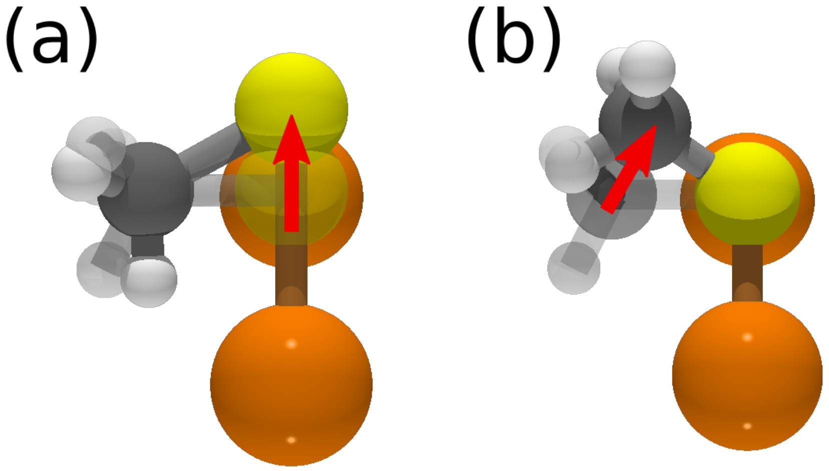


FIGURE 2.4 Here the preservation of the S-C bond orientation is visualized. When (a) sulfur atom or (b) methyl group is moved the hydrogen atoms of the methyl are adjusted according to the S-C bond alignment. Colors: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen. Reprinted with permission from *Journal of Physical Chemistry A*, 124 (23), pp. 4827–4836, **2020**. Copyright 2020 American Chemical Society.

2.6 Methods of structure optimization

In the article [PIV] the atomic forces were estimated with EMLM and OAMLM, but need some application to prove their usefulness. Earlier in the discussion of the MC methods, it was mentioned that running MD with any new force field has a concern about the conservation of energy. Here ML forces contain uncertainty in both norm and direction. Instead of running MD simulations, the structure optimization is a realistic and useful usage case of the ML forces. If the estimated force vectors could drive atoms close to some local energy minimum, one could save a significant amount of computational resources compared to the DFT. The ML forces could be used to run coarse optimization and then the fine minimum is finally searched with DFT or other high-level method.

The optimization method used here was the classic Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [112–115]. BFGS belongs to the category of quasi-Newton methods. The idea behind the Newton methods is to search a zero of the gradient of the function $f(\mathbf{x})$. If the current position of the optimization is $\mathbf{x}_0$ and next position $\mathbf{x}$ is assumed to be close to the $\mathbf{x}_0$, then it can be written that

$$f(\mathbf{x}) = f(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0) \cdot \nabla f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0) \cdot \mathbf{H}^{-1} \cdot (\mathbf{x} - \mathbf{x}_0). \tag{2.35}$$

Here $\mathbf{H}$ is the Hessian matrix of the function to be optimized. Because $(\mathbf{x} - \mathbf{x}_0) \rightarrow \mathbf{0}$, this can be simplified to

$$\nabla f(\mathbf{x}) = \nabla f(\mathbf{x}_0) + \mathbf{H}^{-1} \cdot (\mathbf{x} - \mathbf{x}_0). \tag{2.36}$$

The idea was to find the point, where $\nabla f(\mathbf{x}) = 0$. This can be substituted to the equation (2.36) and it yields a proposed optimization step [116]

$$\mathbf{x} - \mathbf{x}_0 = -\mathbf{H} \cdot \nabla f(\mathbf{x}_0) \tag{2.37}$$

The quasi-Newton methods use the same basic principles as the Newton methods, however the **H** is not an exact Hessian matrix but an approximation. The approximation is supposed to improve, when optimization progresses. To simplify the notation, the change in gradient is written as $\mathbf{g}_i \equiv \nabla f(\mathbf{x}_{i+1}) - \nabla f(\mathbf{x}_i)$ and the move $\mathbf{u} \equiv \mathbf{x}_{i+1} - \mathbf{x}_i$ for the optimization step number i. In the BFGS the approximation is often written as [116, 117]

$$\mathbf{H}_{i+1} = \mathbf{H}_i + \frac{\mathbf{g}_i \times \mathbf{g}_i}{\mathbf{g}_i \cdot \mathbf{u}_i} - \frac{(\mathbf{H}_i\mathbf{u}_i) \times (\mathbf{H}_i\mathbf{u}_i)}{\mathbf{u}_i \cdot (\mathbf{H}_i\mathbf{u}_i)}. \tag{2.38}$$

Here symbol "$\times$" denotes outer or cross product and "$\cdot$" is the inner or dot product. In the beginning of the simulation the approximation of the Hessian matrix is usually set to unit matrix multiplied by some constant. During the optimization process, it is not obligatory nor even convenient to take a full step proposed by the equation (2.37). The step is usually scaled according to some maximum step size given prior for the algorithm and this is taken into account when updating the Hessian matrix approximation. The actual implementation of the BFGS algorithm used in the study is based on the one included in Atomic Simulation Environment (ASE) package [118].

2.7 Density functional theory methods

In [PII] and [PIV] the ML method development was based on the data from the DFT MD simulations of $Au_{38}(SCH_3)_{24}$ by Juarez-Mosqueda *et al.* [18]. Hence, the level of used DFT methods is the same. Calculations were done using DFT code GPAW [119, 120] and the exchange-correlation functional was Pedrew-Burke-Ernzerhof (PBE) functional [121]. In GPAW, electron density is managed in real space grid, for which the grid spacing was set to be 0.2 Å.

In [PII], DFT was used to run MD simulations of $Au_{38}(SCH_3)_{24}$ to generate validation datasets, which were independent from original training data. In [PIV], there were two usage cases for DFT. It was used to calculate single point potential energies for atomic configurations yielded by the structure optimization using ML force. DFT was also used to run comparison BFGS structure optimization on some test structures. This way it was possible to compare how well geometries from ML and DFT optimization runs agreed. DFT-level optimization was done with BFGS implementation in ASE package [118]. The convergence criterion was set as $|\mathbf{f}_{max}| \leq 0.05$ eV/Å

3 RESULTS AND DISCUSSION

In this section the results from articles [PI], [PII], [PIII] and [PIV] are discussed. The underlying theme in these studies is to compare similarity/difference between data and use these measures to analyze features or calculate some useful properties for MPCs. In [PI], the experimental and simulated TEM two structural isomers of the $[Au_{25}(p-MBA)_{18}]^{-}$ nanocluster were compared. This way one of the structural isomers was excluded from the further computational analysis and the analysis gave some hints about the spatial orientation of the clusters in the observed crystal lattice.

According to the best knowledge of the author, the article [PII] is the first demonstration of applying distance-based ML methods on the simulations of MPCs. The studied system was $Au_{38}(SCH_3)_{24}$ nanocluster and ML methods were harnessed to predict potential energy values for different configurations according to MBTR descriptor. The potential energy predictions were used in Monte Carlo simulations. This approach lacked the force vectors, therefore in the articles [PIII] and [PIV] the focus was shifted from the potential energy of the whole structure to the estimation of the forces affecting to individual atoms. The challenge in ML force vectors is that forces have both norm and direction but conventional structural descriptors hide any directional information. In order to overcome this dilemma, the MLM framework was redesigned to predict directional information resulting into a new distance-based ML method OAMLM. In the article [PIII], the operation of the OAMLM framework was demonstrated with linear alkane molecules. In the article [PIV], both direction and norm estimates were combined and ML force vectors were used in structure optimization of the configurations of the gold-thiolate rings, $Au_{38}(SCH_3)_{24}$ and $Au_{25}(SCH_3)_{18}$ nanoclusters. The ML method does not handle charge information, therefore the neutral $Au_{25}(SCH_3)_{18}$ was used.

3.1 Comparison of experimental and simulated TEM images of $[\mathbf{Au_{25}(p{-}MBA)_{18}}]^-$

In the article [PI], Qiaofeng Yao and his co-workers observed experimentally that $[Au_{25}(p{-}MBA)_{18}]^-$ could organize into large colloidal crystals. The shape of the crystals and packing of the clusters depends on the used counter ions. Using either lithium cations (Li^+), tetra methyl ammonium (TMA), tetra ethyl ammonium (TEA) or tetra propyl ammonium (TPA), the packing of the clusters could be modified. In addition to just different packing, TEM images in figure 3.1 revealed that in the presence of the TEA counter ions $[Au_{25}(p{-}MBA)_{18}]^-$ clusters are linked together via partially opened protecting units along a single lattice direction. Afterwards this was confirmed with DFT calculations and GROMACS MD simulations by showing that the protecting ligands did open and the linked with another cluster along a single direction in the lattice. In order to run rigorous computational analysis of the structure, we needed to get information about the orientation of the $[Au_{25}(p{-}MBA)_{18}]^-$ cluster within the lattice and exclude the possibility of any structural isomers.

Before this study began, the new structural isomer for $[Au_{25}(PET)_{18}]^-$ (PET = phenyl ethyl thiolate) was observed via MD simulations [122] and later its existence was confirmed by experiments [123, 124]. This isomer had more outstretched protecting units than expected. The structural difference is visualized in figures 3.2 A and B with p-MBA ligand instead of PET. It was an interesting idea that maybe the new isomer might be more suitable to form the linking between the clusters. It has to be mentioned that there is no X-ray crystal structure of the $[Au_{25}(p{-}MBA)_{18}]^-$ but the structures used in the computational studies are based on crystal structure of $[Au_{25}(PET)_{18}]^-$ [12, 13] and the computationally observed isomer with PET ligands [122]. From now on, the structure based on the crystal structure of the $[Au_{25}(PET)_{18}]^-$ from 2008 [12, 13] is called isomer 1 and the new computationally discovered structure is called isomer 2.

The experimental TEM image of the crystal lattice shown in figure 3.1 contains several individual clusters. In order to compare the images, 13 structures were cut out from the full image. These are highlighted in the figure 3.1. The sampled images still contained background noise, which was cut away in circular manner. This resulted into sharp edges, therefore small amount of Gaussian smoothening was added to the edges. This is visualized in figures 3.2 C i and iii.

The simulated TEM images were generated in 200 evenly distributed directions around the both $[Au_{25}(p{-}MBA)_{18}]^-$ structures as mentioned in section 2.1. The dimensions (resolution) of the images was set to be the same as the experimental images. In the simulated images, there are vague remains of the organic ligands, which are not present in the experimental TEM images. The images are in 8-bit

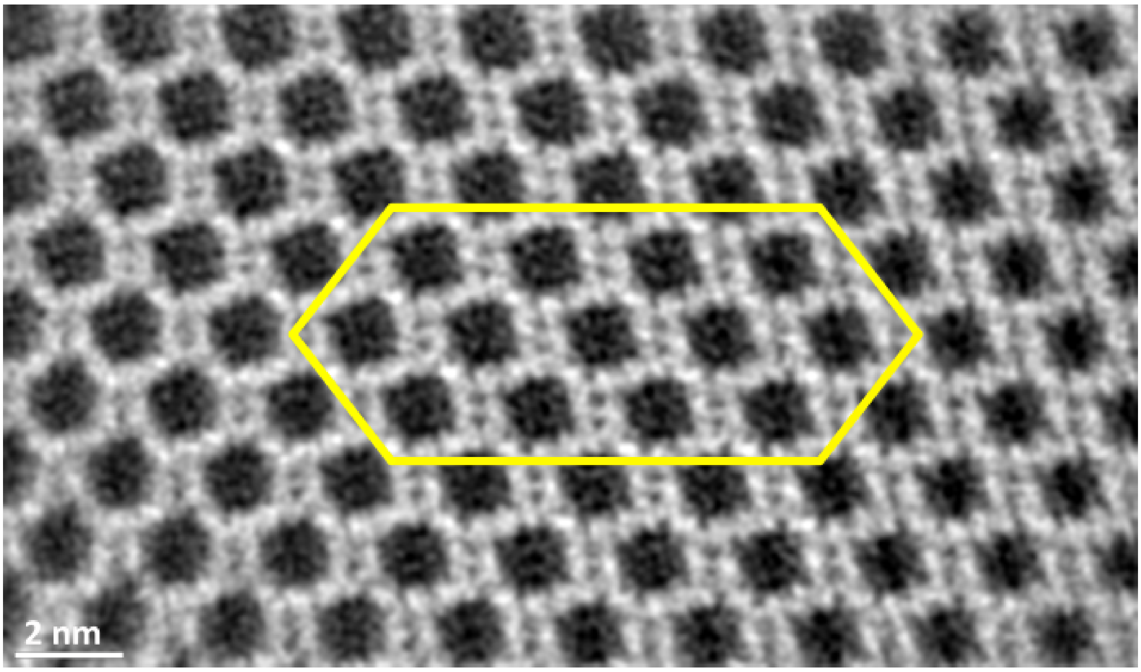


FIGURE 3.1 Contrast transfer function (CTF) corrected TEM image of $[Au_{25}(p{-}MBA)_{18}]^{-}$ nanoclusters in a lattice with TEA counter ions. The yellow outline shows 13 clusters, which were used in the CW-SSIM comparison. The figure is reprinted from the preprint of the article [PI].

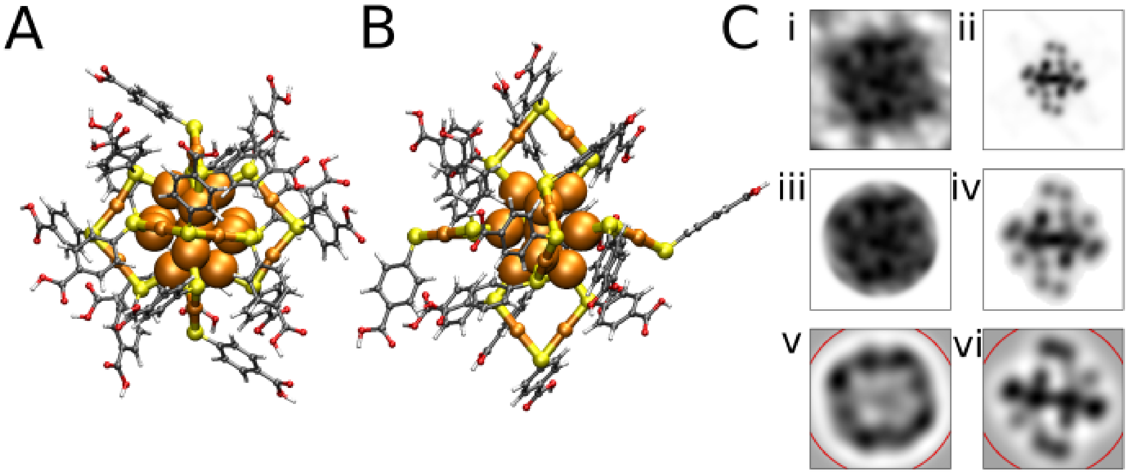


FIGURE 3.2 Structural isomers of the $[Au_{25}(p{-}MBA)_{18}]^{-}$ based on (A) crystal from 2008 [12, 13] and (B) MD simulations [122]. (C) Handling of the experimental and simulated TEM images: i-ii initial images, iii-iv removing background, v-vi wavelet transformed images and compared region highlighted with red. Colors for atoms: orange, gold; yellow, sulfur; grey, carbon; red, oxygen; white, hydrogen. The figure is reprinted from the preprint of the article [PI].

format meaning that every pixel value is between 0 (black) and 255 (white). In order to remove vague ligand effects, pixels with values 240 or higher were set to 255 as visualized in figures 3.2 C ii and iv.

Before the wavelet transformations took place, the values of the pixels were flipped so that black ones had value of 255 and white background was 0. This ensured that the convolution will really pick out the features with different shades and the convolution outside the dark regions is approximately zero. The wavelet transformations were done with ten wavelet widths, ranging from 5 to 30 pixels, of the 2D Rickers wavelet in equation (2.4). Different wavelet widths pick out different features of the images. The narrow wavelets are sensitive to edges and wide wavelets are favored to compare intensities. To get a fair comparison, the CW-SSIM in equation (2.5) is calculated with all 10 widths and the similarity values are then averaged.

One does not know beforehand how the orientations of the simulated images are related to the direction of the experimental TEM image. Hence, the wavelet transformed experimental TEM images were rotated 360° in 50 steps and the averaged CW-SSIM similarity measures were calculated for every rotation. However, the rotations might cause some extra numeric noise, which is the most severe in the corners of the convoluted images. Because this numeric issue, the corners are excluded from the comparison as visualized in figures 3.2 C v and vi.

In the end, all 13 experimental TEM images had 50 averaged CW-SSIM similarity measures for 200 simulated TEM images from the both isomers ($13 \times 50 \times 200 \times 2$ similarity measures in total). The figure 3.3 (A) i shows the highest CW-SSIM values for every simulated images in the case of a single experimental image. The average results are shown in 3.3 (A) ii for all experimental images. From these results it can already be concluded that it is very likely the isomer 1 seen in the experimental TEM images. However, this tells us only about the averages and general trends. In figure 3.3 (A) i it can be seen that in some cases isomer 2 images have got higher similarity values than the images of the isomer 1. It is difficult to determine unconditionally which simulated images are the best. Hence, a relative scoring method was developed. Every experimental TEM images is used to rank the simulated images according to their highest averaged CW-SSIM values. The most similar one gets $N-1$ points, where N is the number of simulated images compared, and the least similar one gets zero points. For every simulated image, their scores from different experimental images were summed and sums were divided with $13 \cdot (N-1)$. This way every simulated image gets a score from 0 to 1.

The scoring was done first with all 2×200 simulated images including both isomers, therefore $N = 400$ in a scoring scheme. The figure 3.3 (B) i shows the scores. Within the 25 highest scoring images there are just two simulated from the

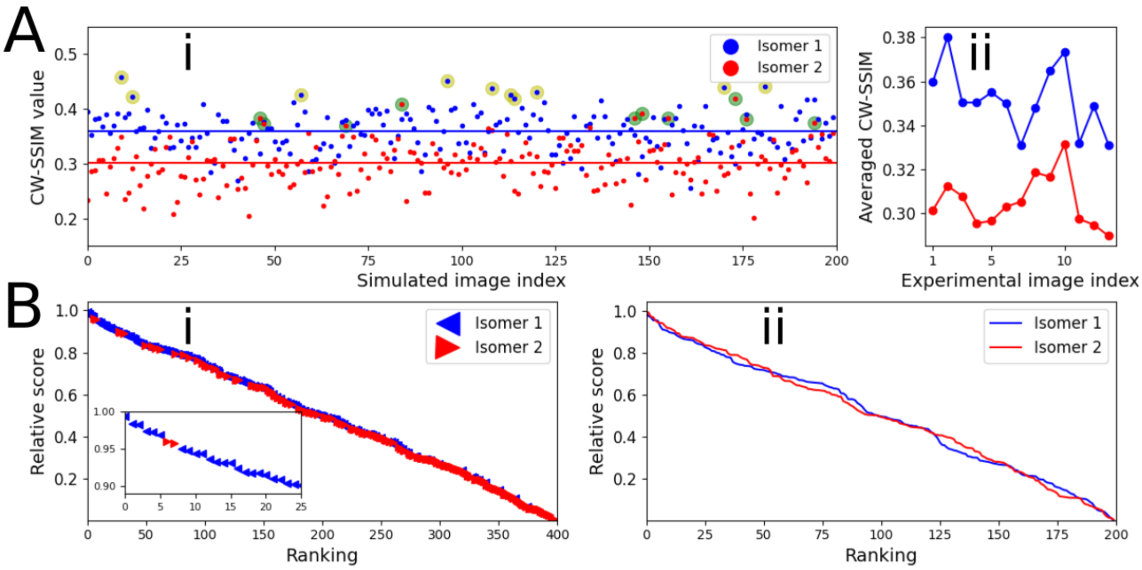


FIGURE 3.3 CW-SSIM comparison results. (A) (i) Example of a comparison between a single experimental image and two sets of 200 simulated images show that majority of isomer 1 images have higher similarity than isomer 2. Horizontal lines are averages and the highest similarities are highlighted with yellow and green. (A) (ii) average similarity results for 13 experimental TEM images. (B) The relative scores of the individual simulated images when (i) all data from both isomers are included and (ii) when isomers are scored separately. The figure is reprinted from the preprint of the article [PI].

isomer 2 and even the 100 highest scores are dominated by the isomer 1. These results further support the conclusion that the isomer 1 is seen in experimental TEM images. In figure 3.3 (B) ii the isomers were scored separately, therefore $N = 200$ for both scorings. This analysis was done to ensure the systematic behavior of the scoring. If the highest results would be significantly lower than 1 or the behaviour differs significantly from the linear relation, then it indicates that the comparison is not systematic. Fortunately, the graph shows that scores behave very closely like $y = (-1/N)x + 1$ line, which adds further evidence to the conclusion. The five highest scoring images are presented in figure 3.4 (B) i-v for isomer 1 and vi-x for isomer 2.

This far the analysis has focused only on determining, which isomer of the $[Au_{25}(p{-}MBA)_{18}]^-$ is seen in the experimental images. Next, the orientation of the clusters is studied. Figures 3.4 (A) and (C) visualize the rotation angles of transformed experimental images, when the highest similarity measures were yielded for the best five simulated images of the isomer 1 and isomer 2 respectively. The fitted lines on polar graphs highlight how systematic the similarity measures are. $[Au_{25}(p{-}MBA)_{18}]^-$ is a highly symmetric structure due to its icosahedral core and symmetric long protecting units, therefore it is expected that symmetry is reflected to similarity measures. Furthermore, the TEM images in 3.1 and 3.4 (B) possess rectangular features, which explain why the highest similarities are ac-

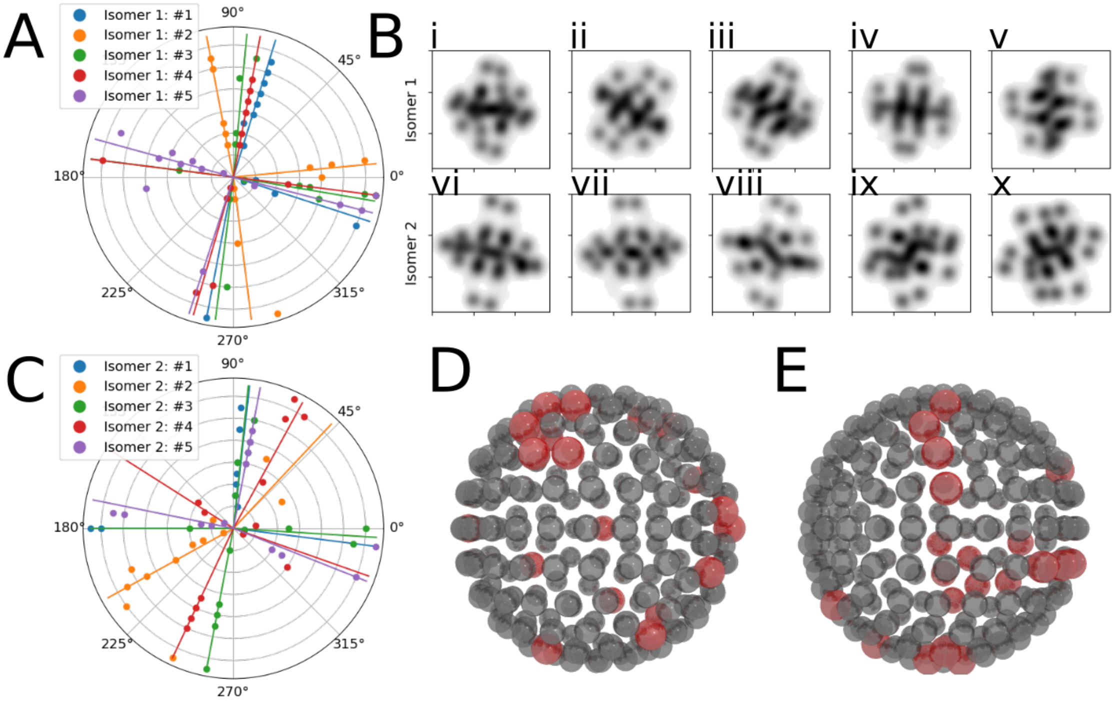


FIGURE 3.4 The orientation angles of the experimental images in the case of five highest scoring simulated images of (A) isomer 1 and (C) isomer 2. (B) the five highest scoring simulated TEM images (i)-(v) for isomer 1 and (vi)-(x) for isomer 2. Red spheres point out the directions from, which the top 20 simulated TEM images were calculated for (D) isomer 1 and (E) isomer 2. The figure is reprinted from the preprint of the article [PI].

quired on approximately 90° rotations of the experimental images. The symmetry of the cluster is also seen in the directions, along which the simulated TEM images are generated. This is seen in figures 3.4 (D) and (E), where the TEM simulation directions are shown for the best 20 images for isomer 1 and 2 respectively. The placement of these directions carries some resemblance with the way how the halves of a baseball are stitched together. This information could be used as a guideline to choose suitable orientations of the cluster to run simulations with the lattice.

As a summary, in this study CW-SSIM similarity measure was applied to the comparison of experimental and simulated TEM images of a $[Au_{25}(p{-}MBA)_{18}]^{-}$ nanocluster. The analysis reliably rules out the possibility of having isomer 2 present in the experimentally observed lattice structure. Images were not compared only by CW-SSIM similarity measures but the simulated images were also scored collectively with relative scoring scheme, which further justified the conclusion that isomer 1 is observed in TEM images. The orientation of the clusters were also analyzed and guidelines about the orientations were acquired. However, the results were not as clear as for ruling out the isomer 2.

3.2 Monte Carlo simulations of $Au_{38}(SCH_3)_{24}$ nanocluster using distance-based machine learning methods

In [PII], the distance-based ML methods were used to predict potential energies for the configurations of $Au_{38}(SCH_3)_{24}$ nanocluster. The training of the method relied on the DFT-level MD simulations of the two isomers of the $Au_{38}(SCH_3)_{24}$ by Juarez-Mosqueda *et al.* [18]. During those simulations, the structures were heated until they started to break. The simulations were long having 12413 configurations for the Q isomer and 12647 for the T isomer. Due to the lengths of the simulations and high temperatures, they managed to cover a significant portion of the configuration space.

The set of MBTR parameters is written as $\{\min, \max, n_x, \sigma, \alpha, \text{cutoff}\}$. The details of the parameters are discussed in section 2.2. Initially, the parameters were set to $\{0, 1.4, 100, 0.1, 0.5, 10^{-3}\}$. MLM was trained with all configurations as a training and reference data. The data was min-max scaled to the interval of $[0, 1]$ and constant variables were excluded from the input vectors. After this, initial tests were done by running MC simulations at various simulation temperatures. However, these simulations produced non-physical and broken structures. The MLM did not handle this phenomena properly, because MC produced structures, which lied too far away from the training region of the model. Hence, 1580 configurations for the Q and 2124 for the T isomer were taken from these simulations and single point DFT potential energies were calculated. These datapoints were used to expand the training set.

The MBTR parameter set was also updated to $\{0, 1.2, 100, 0.045, 0.8, 10^{-5}\}$ after initial tests. These parameters contain significantly less Gaussian broadening $(\sigma : 0.1 \rightarrow 0.045)$ and emphasizes shorter distance contributions due to the slightly increased α parameter $(\alpha : 0.5 \rightarrow 0.8)$, which affects exponential weighting of the MBTR. In the figure 3.5, the feature space of the $Au_{38}(SCH_3)_{24}$ with the second MBTR parameters is visualized via Principal Component Analysis (PCA). It shows clearly that MC simulations were outside the training original region.

After the training set was expanded with additional configurations and MBTR descriptions were updated, both MLM and EMLM models were trained the same way as before. They were validated by testing their performance on data generated by separate MD simulations. For the isomer Q, the average temperatures of three MD runs were 269 K (2000 steps), 475 K (2000 steps), and 795 K (3653 steps). Correspondingly for the T isomer the simulation temperatures were 273 K (2000 steps) and 486 K (2049 steps). The test results presented in figure 3.6 show that ML and DFT potential energies correlate. However, both ML methods are overestimating potential energy, especially for configurations at high energies. Root-mean-squared error (RMSE) was 2.98 eV for MLM and 2.67 eV for EMLM.

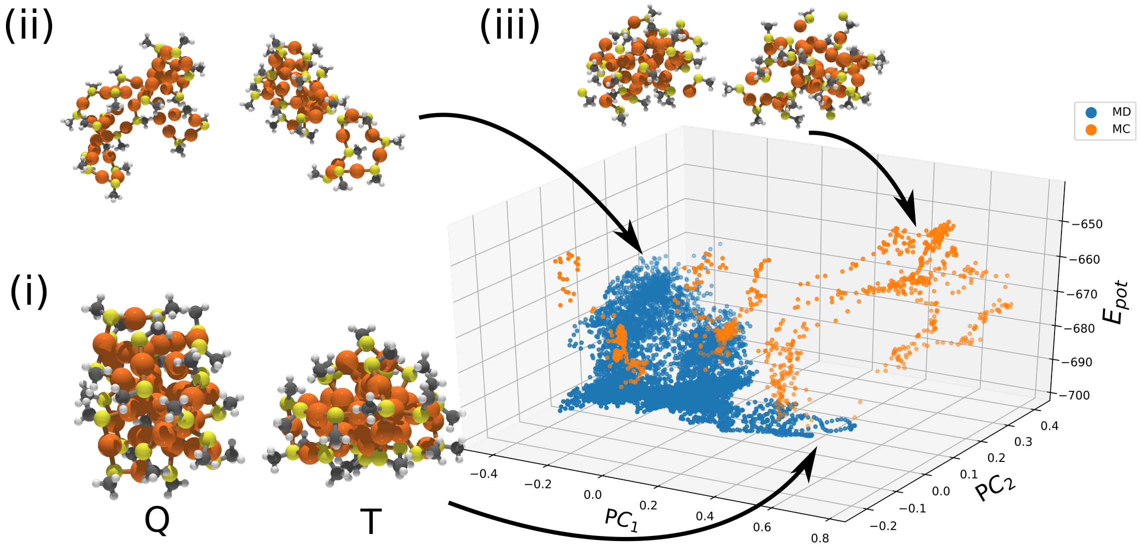


FIGURE 3.5 The Principal Component Analysis (PCA) allows us to visualize MBTR feature space and the potential energy values of the $Au_{38}(SCH_3)_{24}$. (i) shows the initial structures of the Q and T isomers of $Au_{38}(SCH_3)_{24}$. (ii) presents high-energy configurations from MD simulations for both isomers and (iii) contains examples about broken $Au_{38}(SCH_3)_{24}$ structures. Colors: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen. Reprinted with permission from *Journal of Physical Chemistry A*, 124 (23), pp. 4827–4836, **2020**. Copyright 2020 American Chemical Society.

At first glance RMSEs might seem quite large but they are still within reasonable limits. The difference between minimum and maximum potential energy is approximately 30 eV meaning that RMSE is less than 10% of the interval. Furthermore, the relative errors compared to the absolute values are just 0.38% for MLM and 0.33% for EMLM. However, the averaged error estimates do not tell the whole truth. This is the same logic as in the comparison of TEM images in the previous section. There one isomer of the $[Au_{25}(p{-}MBA)_{18}]^-$ produced higher similarities than other by average but there were still individual good images. Same way here low-energy regions are more accurate than high-energy ones, therefore the model can be used in applications if the simulations stay within its domain of applicability [125].

The true test for the ML methodology was to run MC simulations. However, one has to take into account where the model is the most reliable. MC simulations should not be run at too high temperatures, therefore 200 K, 250 K and 300 K were used. These simulations should stay within the harmonic vibration regime. It is also important to note that for MC simulations the absolute values of the potential energies are not important but the performance is determined by the reliability of the energy differences. MC simulations were run only with EMLM, because it is slightly more accurate and it is about one magnitude faster than MLM. Using a single thread in a single core of the Intel Xeon CPU E5–2680 v3 @ 2.50 GHz with

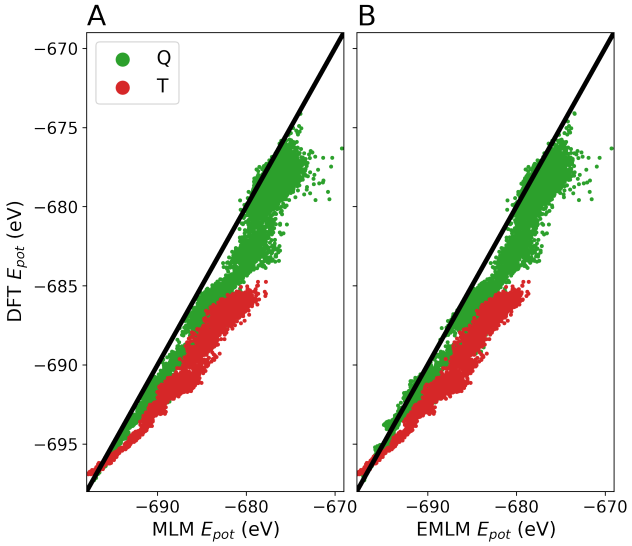


FIGURE 3.6 The potential energy values predicted with (A) MLM and (B) EMLM are compared to the corresponding DFT values from the MD calculations for Q and T isomers. When all data points were considered the (RMSE) was 2.98 eV for MLM and 2.67 eV for EMLM. Reprinted with permission from *Journal of Physical Chemistry A*, 124 (23), pp. 4827–4836, **2020**. Copyright 2020 American Chemical Society.

8GB memory, EMLM can predict one potential energy value in 0.05 s and MLM in 0.56 s. Computing the MBTR description takes about 0.07 s. As a comparison, a calculation of a single point potential energy value for $Au_{38}(SCH_3)_{24}$ in CSC Mahti supercomputer using 128 CPU cores (one computing node) requires over one minute.

In the figure 3.7 A, PCA shows that 300 K simulations are restricted into a small region in the MBTR feature space. EMLM predicted potential energy fluctuations are visualized in the figure 3.7 B. MC simulation show that the T isomer has about 1.5 eV higher potential energy than the Q isomer on average. This is also known from the experiments and DFT [17–19]. This displays that EMLM has learned realistic relative energetics. In order to get further confirmation, one could also analyze how potential energy change during the simulation steps. In the figure 3.8, the relative numbers of steps with corresponding potential energy fluctuations from both 300 K MC and DFT MD simulations are visualized. It shows clearly that the MC steps and MD steps are causing similar potential energy fluctuations further proving that the model works as expected.

Comparing only potential energy does not tell anything about structural changes that $Au_{38}(SCH_3)_{24}$ structures undergo. In order to get a deeper structural understanding, bond distances and bond angles from both MC and DFT MD were compared. C-H bonds were fixed, as mentioned in section 3.2, therefore there

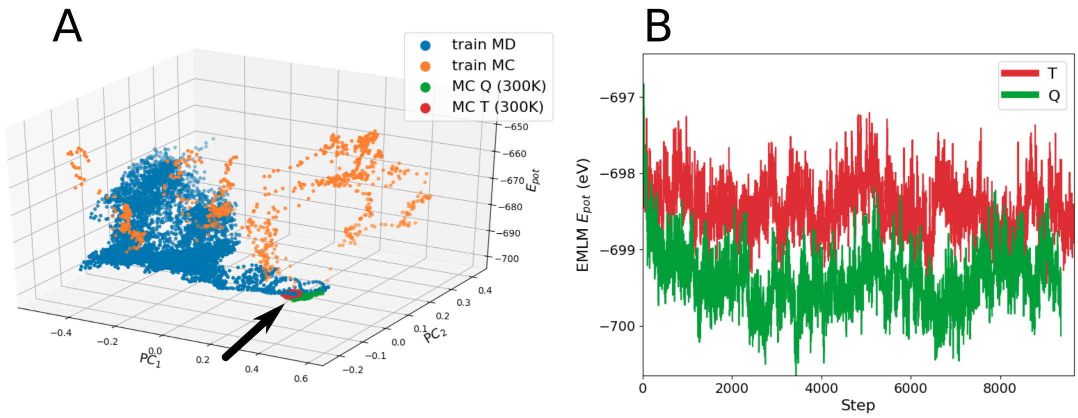


FIGURE 3.7 In the panel A PCA is used to visualize the region in the MBTR feature space, where 300 K MC simulations are moving. The panel B shows the potential energy fluctuations during the MC simulations. Reprinted with permission from *Journal of Physical Chemistry A*, 124 (23), pp. 4827–4836, **2020**. Copyright 2020 American Chemical Society.

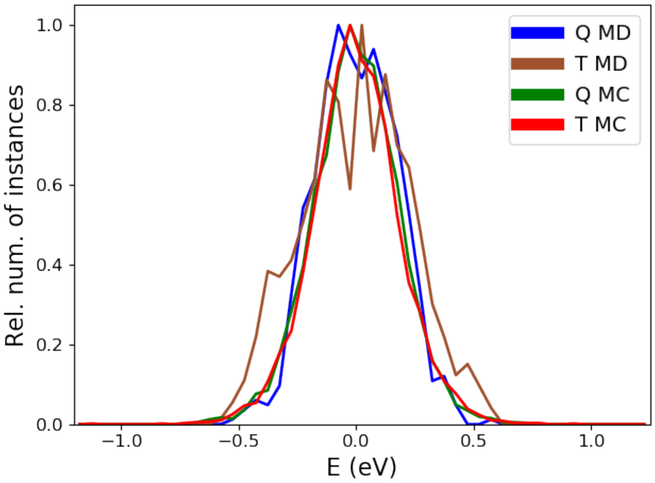


FIGURE 3.8 The distributions of the energy differences between simulations steps of DFT MD and EMLM MC show similar shape. Both types of simulations were run in 300 K (aimed temperature for MD). The width of the sampling step is 0.05 eV.

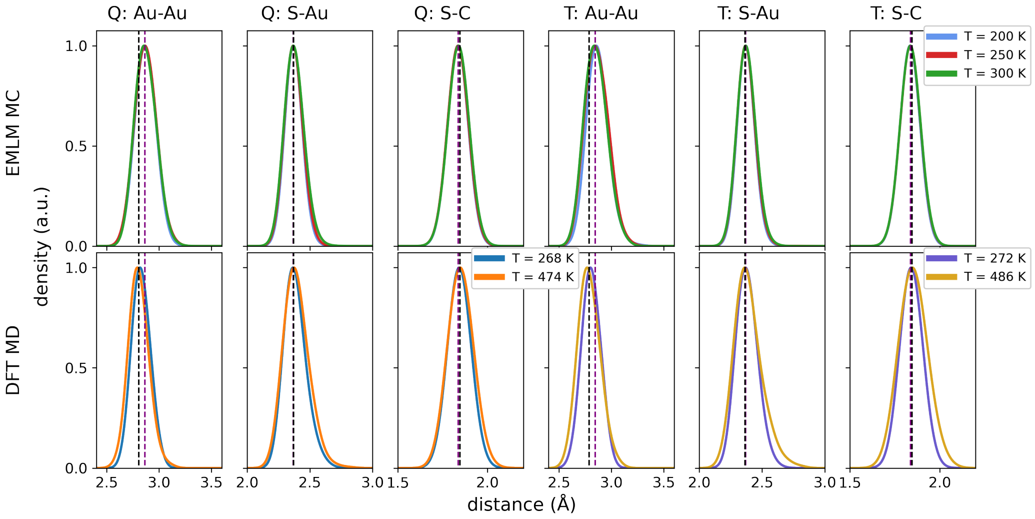


FIGURE 3.9 The top row of plots shows the distributions of the bond distances in the MC simulations. The bottom row shows corresponding distributions from DFT MD simulations. The vertical dashed lines highlight the average peak positions. The Gaussian smoothening $\sigma = 0.05$ Å. Reprinted with permission from *Journal of Physical Chemistry A*, 124 (23), pp. 4827–4836, **2020**. Copyright 2020 American Chemical Society.

were three types of fluctuating bonds: Au-Au, S-Au and S-C. The bond distance distributions in figure 3.9 show that MC and DFT MD results agree on bond distances. Only Au-Au bonds are slightly overestimated.

The bond angles within the protecting gold-thiolate units are special subjects of interest. There are two types of bond angles: Au-S-Au and S-Au-S. In this notation the angle is centered on the middle atom. Ideally Au-S-Au should be about 90° and S-Au-S 180° but in practise there are always some fluctuations. The angle distributions in figure 3.10 reveal that the EMLM has difficulties on producing bond angles. Especially S-Au-S angle distributions have been broadened significantly and their peaks have drifted towards smaller angles than expected. Au-S-Au bonds are still within reasonable limits but they also differ from the DFT MD simulations.

One has to ask what has caused such a behavior, where bond distances are accurate but angles are not. The issue lies in the description. The used MBTR description considered only pairwise distances within the systems and it does not include explicit angular information. In principle, knowing all of the atomic distance one could acquire some knowledge about the angles but this is not straightforward. Furthermore, the construction of the distance matrix used to make prediction in EMLM and MLM hides the individual features of the MBTR and prediction is done using only Euclidean distances between descriptions. It is not unexpected that the bond angles are not simulated accurately. The broadened angle distributions are

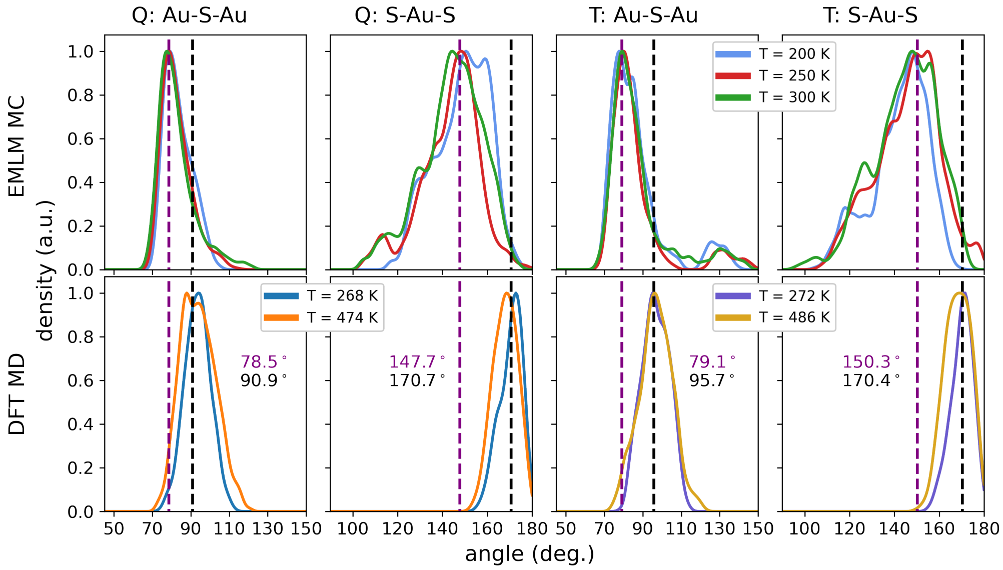


FIGURE 3.10 The top row of plots shows the distributions of the bond angles in the MC simulations. The bottom row shows corresponding distributions from DFT MD simulations. The vertical dashed lines highlight the average peak positions. Gaussian smoothening $\sigma = 1.75°$. Reprinted with permission from *Journal of Physical Chemistry A*, 124 (23), pp. 4827–4836, **2020**. Copyright 2020 American Chemical Society.

caused by the geometric change in protecting units, where the gold atoms within the units are pulled towards the core. This bends protecting units to the shape like letter "M".

In this study, the potential of the distance-based ML methods was demonstrated in the simulations of the $Au_{38}(SCH_3)_{24}$ cluster. The potential energy was predicted well with both MLM and EMLM. The faster EMLM method was then used to run MC simulations at various temperatures. Energetics were in agreement with DFT and the model produced realistic bond distances. However, the bond angles caused difficulties due to the lack of angular information in the structural descriptor.

3.3 Force direction estimation example with alkanes

Calculating the potential energy of an atomic system is a routine task for conventional simulation methods and for ML it is also a relatively straightforward problem to handle. One just has to form a model for regression from structural descriptions to scalar energy values. Forces, i.e. the negative gradients of the potential energy surface, are a different problem. The usual way to get force vectors from the ML force field is to take gradient over the model and the description. However, this can lead to specialized models with limited generalization capabilities. The models trained to predict how much every atom contributes to the potential energy of the system are often able to address this generalization problem. The drawback of these models is that they are expensive to train, because one atomic configuration has just one potential energy but it has N atoms. The model has to learn to divide potential energy into N local contributions, which is not a trivial task. Forces, on the other hand, are not constrained by the global structure of an atomic system but they are local properties dictated by the chemical environments of the individual atoms. If the model is trained only on forces, one is not bound to use full atomic structures. It requires only local environments around individual atoms, which could be sampled independently from various systems. That is why in this study the force vectors subjecting to individual atoms were predicted directly leaving the potential energy predictions out. This kind of local ML models assume strong localization of interactions in the spirit of nearsightedness approximation [126], which in many cases is a reasonable assumption.

Designing a ML vector field method for atomic forces has also a physical justification. We assume that Born-Oppenheimer approximation is valid *i.e.* atoms are moving slowly enough and electrons have time to adjust. Hence, Hellman-Feynman theorem states that the forces are true quantum mechanical observables [127, 128] and they can be, in principle, solved analytically separately from the energy calculation. This is also a relief for data generation. If there would not be an analytic way to compute atomic forces, then one would have to rely on numerical differentiation, which would require numerous energy calculations to get gradients of the potential energy surface.

In the ML point of view, the prediction of atomic force vectors means that the model has to create a mapping from structure/description space to a 3-dimensional vector field. To complicate the matters even more, the vector field has to be able to address the spatial orientation of the system. In high symmetry systems, this is easy as one can utilize symmetry to fix some coordinates. Unfortunately, MPCs are naturally low symmetry structures and symmetric features vary from structure to structure. If the spatial orientation of the atomic system is not reflected into the ML vector field, the model would have little use. Unke et al. point out in their extensive review article that not all vector fields are gradient fields, therefore extra care is required when constructing a model to estimate atomic forces [24]. It has to be mentioned that vector fields as such are not uncommon in ML applications. For example, they can be used in robotics [129]. However, often the problem setting can be restricted by defining some "lab coordinate system", which is not viable for MPCs or many other nanostructures.

In [PIII], the OAMLM force direction estimation framework was introduced for the first time. The performance of the method was demonstrated with alkane chain with two to seven carbon atoms. The datasets were generated by running 1000 simulation steps with Velocity Verlet MD simulations [110] using Density Functional Tight-Binding (DFTB) code Hotbit [130] to calculate forces and energies. In the beginning of MD simulations, every atom was given a random velocity vector sampled from the Maxwell-Boltzmann distribution corresponding the temperature of 750 K and the time step was 1.5 fs. The atomic environments were described with SOAP descriptor using parameters $n_{max} = 6$, $l_{max} = 1$, $\sigma_{SOAP} = 1.0$ Å and $r_{cut} = 3.0$ Å. The atomic environments were aligned by using four nearest neighbor atoms and going through all permutations. The accuracy of the alignment was measured with equation (2.25). These parameters were more or less arbitrary and in this demonstration they were not optimized. However, their effect to model accuracy is significant and the parameter optimization is done in the next section 3.4.

The training data was generated by running two separate MD simulations for all six alkane chains ($6 \times 2 \times 1000$ configurations, 2×27000 carbon environments, 2×66000 hydrogen environments). From this data 7500 points were sampled with RS-maximin method [101, 131] for both carbon and hydrogen. The idea of this sampling method is that the first selected data point is closest to the data mean and then the following points should maximize the distance to the previous points. This enables a good coverage over the whole dataset. The training was done in two ways: directly solving equation (2.20) with least-squares fitting and using Huber regression [99]. Because SOAP parameters are sub-optimal, Huber regression is used to make the model statistically robust similar way as in robust MLM method by Gomes *et al.* [100] as mentioned in the section 2.3. Robust OAMLM models were trained with Huber parameters of $p_{Huber} \in [1, 2]$ with steps of 0.1. The Huber parameter of 1 produces the most robust model and increasing it reduces the robustness. During the training of the OAMLM all 7500 data points were saved as references.

The test set was generated by running a third set of individual MD simulations. For carbon all data points were used in tests but for hydrogen only data points from the every third configuration were used. The performance was measured with the weighted average of the angles between predicted direction and DFTB force vectors. Weights were squared norms of the DFTB forces. This emphasizes the correct handling of the large forces over small ones. In the case of small forces, the direction is elusive and extremely sensitive to even slightest movement of atoms.

The figure 3.11 A shows the train errors for robust models and test errors for all models. The train errors for OAMLM with regular training for carbon was 7.4° and for hydrogen 1.3°. The results show that robustness improves the test results, when SOAP parameters are not optimized. In the 3.11 B and C the effect of robustness is visualized with 2D histogram plots. If the change in the angle between predicted and real force direction is negative it means that the robustness has improved the prediction. For carbon this is clear but for hydrogen improvement is not that visible. The test results for the regular models and for the robust models yielding the smallest weighted average angle are shown in figures 3.11 D-G. There the hydrogen results show that the improvements do happen on large forces. However, this is a compromise, because the directions of the small forces predicted with robust OAMLM are not as accurate as with regular OAMLM.

As a summary, this study demonstrated the usage of the OAMLM method for atomic forces directions for the very first time. The SOAP descriptor parameters were not optimized for alkanes, therefore the results left something to hope for. Adding robustness to the models via Huber regression could be used to improve the results, when input descriptor data is sub-optimal. The direction estimation framework is a promising approach for atomic simulations and its potential becomes more evident in the next section.

3.4 Machine learning for atomic forces and their application to gold-thiolate structures

In the study of article [PIV], EMLM and OAMLM frameworks, presented in section 2.3 and 2.4, were utilized to estimate atomic force vectors for gold-thiolate systems. During the model development the SOAP parameters were optimized in two phases. First a few most optimal parameter sets were selected based on EMLM norm predictions and then the optimal set was chosen according to the OAMLM direction performance. The main dataset used here is the same DFT MD simulation data from the reference [18], which was used in the [PII]. However, the amount data was restricted by sampling logarithmically 1000 configurations for both isomers. The ML methodology was applied to structure optimization of gold-thiolate rings, $Au_{25}(SCH_3)_{18}$ and two isomers of the $Au_{38}(SCH_3)_{24}$.

3.4.1 SOAP parameter selection via force norms and directions

In this section, the testing procedure of the SOAP parameters and the main test results are discussed. However, only final results are shown in detail and for the rest the process is explained verbally. Several SOAP parameters were tested: $n_{max} \in [2,7]$, $l_{max} \in [0,4]$, $r_{cut} \in \{4.0\text{ Å}, 5.0\text{ Å}\}$ and $\sigma_{SOAP} \in \{1.0, 0.75, 0.5, 0.25\}$. In total, this means 240 description sets for sulfur, carbon and hydrogen. For gold atoms we used only $\sigma_{SOAP} = 0.25$ value resulting 60 SOAP parameter sets. The test were first done with EMLM predicting norms of the forces. These results were used to limit the parameters to be tested with OAMLM. Force norm prediction

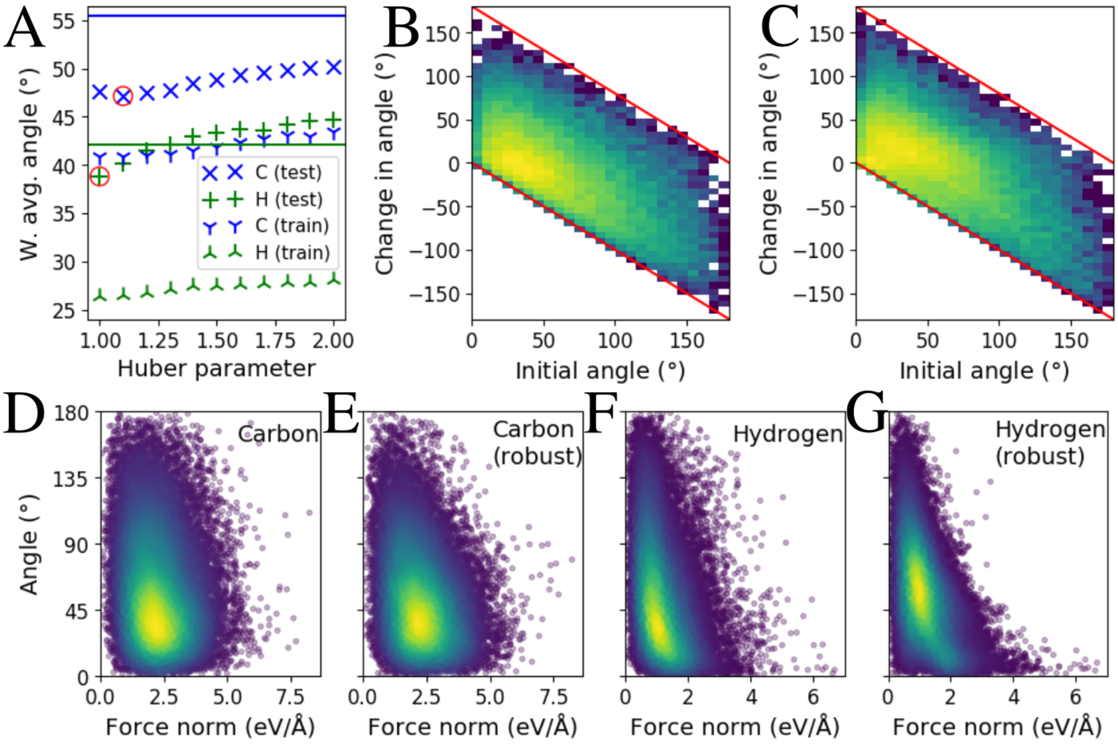


FIGURE 3.11 The performance of the OAMLM method with alkanes. Panel A shows the train and test errors of the robust OAMLM models with crosses. Horizontal lines correspond to the errors of the regular OAMLM models. Training errors of the regular OAMLM models are below the visualization range. The best results are highlighted with red circles. The effect of robustness is visualized B for carbon and C for hydrogen with 2D histograms. More negative change means more correction to the prediction. Colors are logarithmically normalized. Panels D-G show the test results for the regular models and for the robust models with the smallest test error. Colors in B-G present the density of the points: yellow, dense region; purple, sparse region. Reprinted with permission from *ESANN 2021: Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning Online event* (2021) pp. 529–534.

is an easier task to do than estimation of the direction, therefore it is justified to expect that if norms are not predicted with adequate accuracy then directions won't be either. Furthermore, the training and the testing of EMLM is significantly faster than for the OAMLM. All tests were done separately for all atom types present in the $Au_{38}(SCH_3)_{24}$ nanocluster: core gold, unit gold, sulfur, carbon and hydrogen. The discussion about these atom types and their atomic environment alignment schemes are presented in section 2.4.

During the SOAP parameter tests, Q and T isomers of the $Au_{38}(SCH_3)_{24}$ were handled separately. For every atom type models were first trained with data from one isomer and then tested with another. This enables one to evaluate the transferability of the model in a similar fashion as cross-validation. For the training data 2500 points were sampled with RS-maximin [101, 131] and all of them were saved as references. After the norm tests, the SOAP parameters were limited to $\sigma_{SOAP} = 0.25$, and $(n_{max}, l_{max}) \in \{(6,4),(7,3),(7,4)\}$ with both $r_{cut} = 4.0$ Å and $r_{cut} = 5.0$ Å.

The OAMLM direction test were run with the parameters, which were selected based on the EMLM tests. The training and testing were done in the same manner as for the force norm EMLM models. Both numeric and analytic loss functions shown in equations (2.29) and 2.30 were tested with parameters $\sigma_1 = 0.25$ and $\sigma_2 = 0.5$. The optimal compromise for the SOAP parameter was determined to be $\sigma_{SOAP} = 0.25$, $(n_{max}, l_{max}) = (7,4)$ and $r_{cut} = 4.0$ Å. Numeric loss function showed inferior results compered to the analytic one. With these SOAP parameters the analytic loss function parameter values $\sigma_2 \in \{0.25, 0.5, 0.75\}$ were tested for all atom types. Only for unit gold atoms $\sigma_2 = 0.25$ showed improvement, therefore we used that one for unit gold and for everything else $\sigma_2 = 0.5$.

With these optimal parameters at disposal, the final ML models were trained. This training used combination of data from both Q and T isomers. For EMLM 5000 points were selected with RS-maximin [101, 131] as training and reference data. The rest of the data was used as test data and the results with corresponding RMSE values are shown in figure 3.12. For OAMLM, only 2500 points were sampled as training and reference data. The direction estimation with OAMLM is relatively slow, because of the neighborhood alignment. Hence, fewer reference points result to faster models. The direction results are shown in figure 3.13. Here the performance of the OAMLM is evaluated with weighted average of angles between estimated direction and DFT force vectors The weights are squared norms of the real force vectors as in [PIII].

The test results for both EMLM and OAMLM show that the handling of unit gold, sulfur and hydrogen atoms is the most reliable. Core gold atoms contain the most uncertainty, which is expected. The metallic core undergoes many changes during the MD simulations, which makes neighborhood alignment in the OAMLM

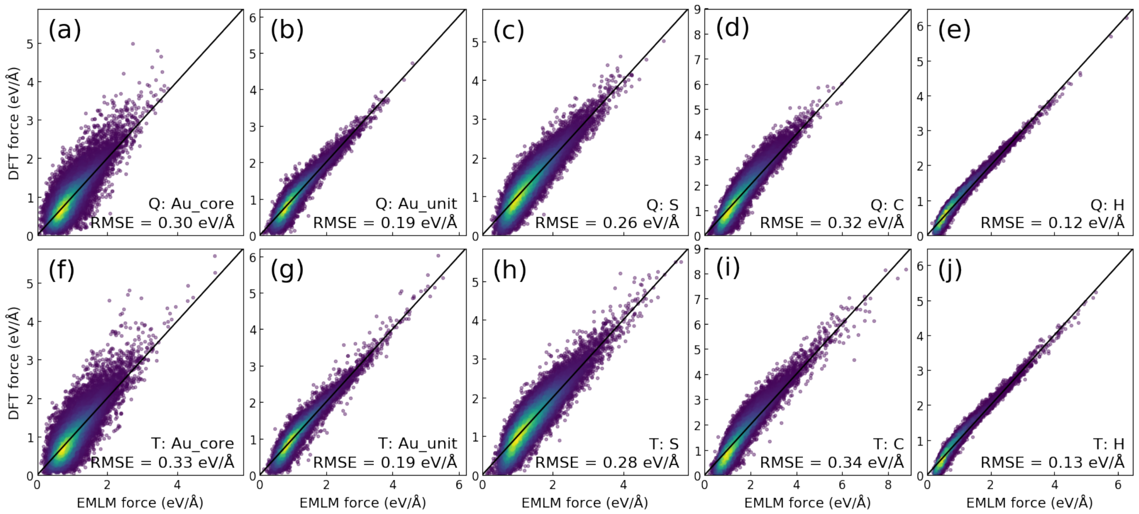


FIGURE 3.12 Panels (a)-(e) show the norm test results for the Q isomer and (f)-(j) for the T isomer of $Au_{38}(SCH_3)_{24}$. Here EMLM was trained with data from the both isomers. The tested element is written to the corner of every graph along with RMSE values. For hydrogen only third of the data points are plotted. The colors visualize the density of the points: yellow means dense region and purple sparse. Reprinted from the article [PIV] (arXiv: 2203.09788)

challenging. The same diversity also causes uncertainty into the EMLM norm prediction. Against initial expectations, the accuracy of the methyl carbon atoms is just slightly better than the accuracy of core gold atoms. The uncertainty of the carbon models is most likely originated from the SOAP description accuracy. The chemical environment of the carbon is dictated by three hydrogen atoms and a sulfur atom. The movement of hydrogen atoms is limited, therefore the descriptions of the atomic environments are very alike. The description accuracy could be improved by even smaller Gaussian broadening parameter σ_{SOAP} but this also introduces a risk of reduced transferability of the models. However, all showed models show reasonable accuracy.

3.4.2 Optimization of gold-thiolate rings

In the first application of the ML framework, the uncertain gold core was left out. ML forces were used to optimize gold-thiolate rings containing four, five or six gold atoms as seen in figure 3.14. Neither EMLM or OAMLM was explicitly trained to handle these rings, which makes this an interesting test of generalizability. It has to be noted that, the original DFT MD trajectory of the T isomer by Juarez-Mosqueda *et al.* contains a highly deformed seven gold atom ring [18]. The ring broke out from the cluster during the late stages of the heating. However, the data sampling is not guaranteed to select data points from this ring. Gold-thiolate rings are not just computational model structures but they have been detected

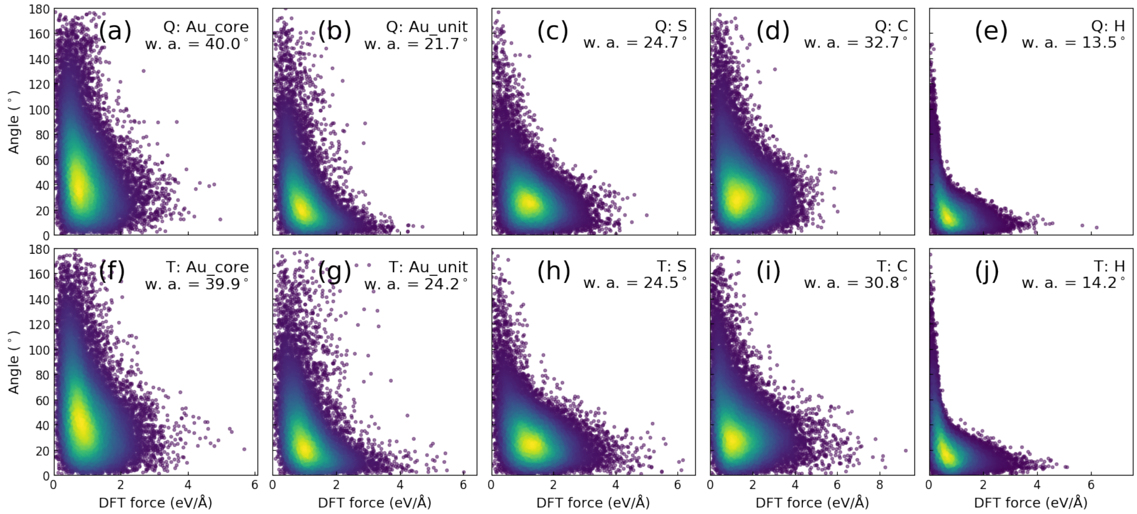


FIGURE 3.13 Panels (a)-(e) show the direction test results for the Q isomer and (f)-(j) for the T isomer of $Au_{38}(SCH_3)_{24}$. Here OAMLM was trained with both isomers. The OAMLM models used analytic loss function in equation (2.30). Vertical axes are the angle between the predicted direction and the DFT force vectors. Horizontal axes show corresponding DFT force norms. The tested element is written to the corner of every graph. For hydrogen only third of the data points are plotted. In the graphs, "w. a." stands for weighted average. The colors visualize the density of points: yellow means dense region and purple sparse. Reprinted from the article [PIV] (arXiv: 2203.09788)

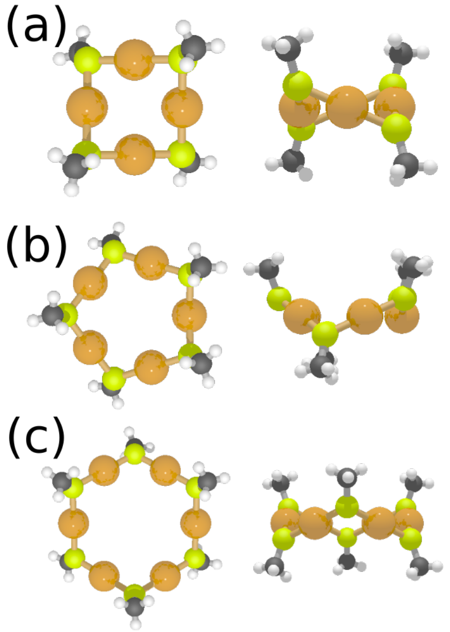


FIGURE 3.14 Top and side views of the initial structures for (a) four, (b) five and (c) six gold atom gold-thiolate rings. Colors: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen. Reprinted from the article [PIV] (arXiv: 2203.09788)

in experiments [132–134] and they have also been studied with DFT [135]. This further adds value to the test.

The optimization used BFGS algorithm presented in section 2.6. It was run using both ML and DFT force vectors. For ML optimization the maximum optimization step size was 0.1 Å and for DFT optimization it was the default 0.2 Å. Convergence criterion for ML optimization was set to $|\mathbf{f}_{max}| \leq 0.1$ eV/Å but due to the uncertainty of the ML method and the behavior of the Hessian matrix approximation in BFGS this criterion was not reached. After optimization the potential energy values were calculated with single point DFT for ML optimization configurations. The potential energies are shown in figures 3.15 (a)-(c). The final configurations for DFT optimization are shown in figures 3.15 (d)-(f) and for ML optimization in figures 3.15 (g)-(i).

The potential energies are decreasing almost monotonously for four and five gold atom rings. The six gold atom ring contains more empty space than any data point in the training set, therefore results are expected to have certain level of uncertainty. By looking the final configurations, one could realize interesting structural differences. DFT optimizations have preserved the clockwise twisting of the rings but ML optimization had reversed the twisting to be counter-clockwise. The twisting is especially clear for four and five gold atom rings. In order to ensure that this a realistic behavior the final ML optimization configurations were optimized again with DFT. These configurations are visualized in figures 3.15

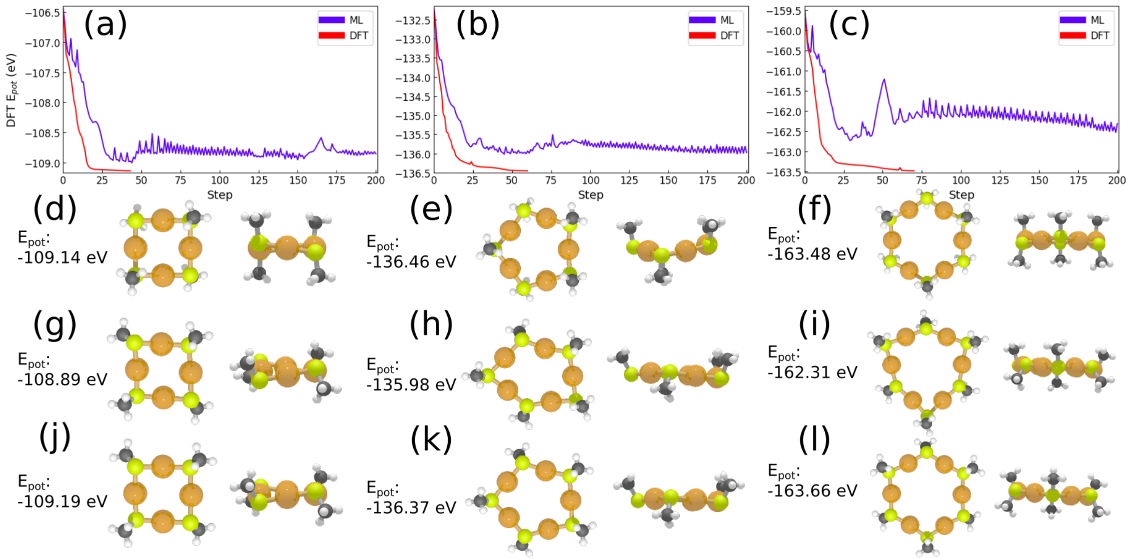


FIGURE 3.15 The DFT calculated potential energy evolution of the BFGS optimizations for (a) four, (b) five and (c) six gold atom gold-thiolate rings. The final structures from the DFT (d)-(f) and ML (g)-(i) optimizations viewed from top and side. The structures in (j)-(l) are DFT optimization results, which started from the corresponding ML optimized configurations. Colors: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen. Reprinted from the article [PIV] (arXiv: 2203.09788)

(j)-(l). This second round of optimization preserved the twisting and surprisingly the potential energies are slightly smaller for four and six gold atom rings than what DFT initially suggested. This shows the potential of the ML forces estimated with EMLM and OAMLM. They could be used in the coarse optimization to help DFT and reduce computational cost of the optimization.

3.4.3 Partial optimization of $Au_{38}(SCH_3)_{24}$ ligand shell

The next step to validate ML force framework is to optimize some part of the $Au_{38}(SCH_3)_{24}$ nanocluster. In this case, the test case was a single protecting gold-thiolate unit, which was outstretched by 2 Å. For Q isomer this unit lied on the corner of the cylindrical shape and for T isomer it was in the middle as seen in the figures 3.16 (a) and (b). This outstretched unit contained two core gold, two unit gold, three sulfur, three carbon and nine hydrogen atoms. During the optimization everything else expect these were fixed.

Different maximum step sizes for the BFGS optimization were tested. The step size affects significantly the optimization performance, because it determines how fast Hessian matrix approximation is updated and how much the uncertainty of the ML method affects the approximation. Because of the second order information, BFGS is vulnerable to the noise and inaccuracies of the gradient. This is caused by

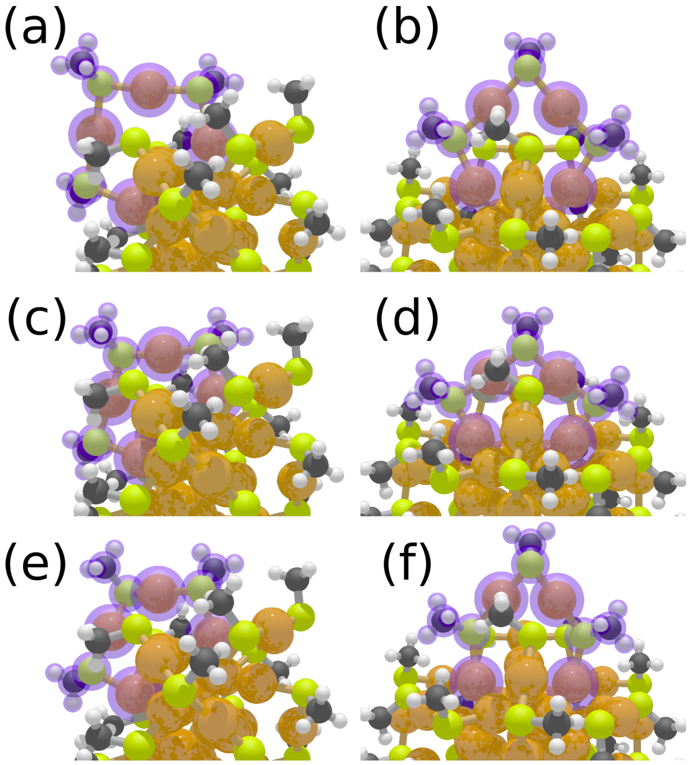


FIGURE 3.16 The stretched protecting unit of the $Au_{38}(SCH_3)_{24}$ Q isomer lies on the corner of the structure (a) and for T isomer it is in the middle (b). DFT constrained optimization results (c)-(d) are used for comparison. (e) and (f) are constrained ML optimized structures from the 150th optimization step with 0.05Å maximum BFGS step size. During the optimizations everything else was fixed except the parts highlighted with purple. Colors: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen. Reprinted from the article [PIV] (arXiv: 2203.09788)

the ill-posedness of the noisy derivatives [136]. Hence, it is a good idea to try to control how the uncertainty builds up to the Hessian matrix approximation.

The optimization performance was determined by comparing single point DFT calculated potential energies and root-mean squared displacement (RMSD) of the ML and DFT optimized structures. Here the terminology has to be clarified. Here RMSD refers to the structural difference between two atomic configurations and RMSE is used to describe the error of the ML method used to predict either potential energies as in [PII] or force norms in [PIV]. Here the RMSD is calculated using the moving atoms from the ML optimization configurations and the final DFT optimized structure. The hydrogen atoms are excluded from the RMSD.

The potential energy evolution of the Q isomer optimization is shown in the figure 3.17 (a). With all maximum step sizes the potential energy is decreasing effectively in monotonous fashion, even if it could not reach the potential energy produced by DFT. RMSD values, however, show more difference than potential energy evolution. The simulation with 0.05 Å gets closest to the DFT optimized structure. The difference is caused by the slightly different angle of the protecting unit, which is induced by the two core gold atoms. These gold atoms are not fitted as deep into their places as done by DFT, which is seen in figures 3.16 (c) and (e). Hence, the unit is left little outstretched and the orientation is changing, because optimization is trying to overcome this barrier. T isomer shows this behaviour even more clearly. The potential energy decreases as supposed to in the figure 3.17 (c) but after about 80 optimization steps the RMSD values diverge in the figure 3.17 (d). In the figures 3.16 (d) and (f), it is clear that two core gold atoms are not fully fitted in their places causing similar effect as in the case of Q isomer. Even tough the ML optimization did not reach the same results as DFT, it still shows a promising performance by reducing potential energy up to certain degree almost monotonously.

3.4.4 Optimization of the MD snapshots of the $Au_{25}(SCH_3)_{18}$ and $Au_{38}(SCH_3)_{24}$ nanoclusters

The greatest challenge for the ML method is to optimize arbitrary configurations of the thiolate-protected clusters. There were three different systems to be optimized. The first one was $Au_{25}(SCH_3)_{18}$, which was taken from the 1500th step of the 500 K DFT MD simulation of the $[Au_{25}(SCH_3)_{18}]^-$. The ML method does not recognize charged systems, therefore the optimization is done for the neutral one and this is also addressed in the single point DFT calculations. The other two test systems were $Au_{38}(SCH_3)_{24}$ configurations from the original DFT MD simulations from the reference [18]. For Q isomer the 1000th configuration was used and for T isomer the 600th. $Au_{25}(SCH_3)_{18}$ is a smaller and more well-defined system than $Au_{38}(SCH_3)_{24}$, therefore it works as an initial link between gold-thiolate

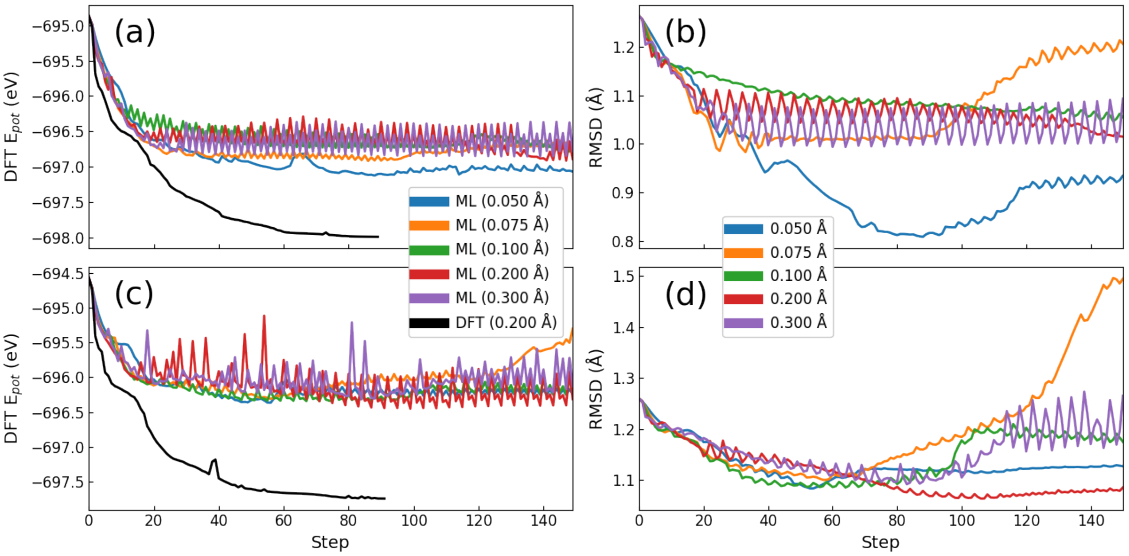


FIGURE 3.17 Different maximum step sizes were tested for the BFGS optimization of the stretched protecting units. The potential energy evolution for Q isomer is shown in (a) and the RMSD compared to the DFT optimized structure is in (b). (c) and (d) are corresponding plots for T isomer. Reprinted from the article [PIV] (arXiv: 2203.09788)

rings and larger gold-thiolate systems. Furthermore, this also demonstrates the generalizability of the method. The initial configurations are shown in the figure 3.18.

The partial optimization of the ligand shells in the previous section showed that 0.05 Å is a good compromise for the maximum BFGS step size, therefore all ML optimizations use this value. $Au_{25}(SCH_3)_{18}$ was optimized with four different schemes. The full optimization with all atoms free and optimization of the ligand shell with gold core fixed are standard approaches. The other two approaches optimize the structure in parts. First, outside ligand shell containing unit gold, sulfur, carbon and hydrogen atoms is optimized 24 steps keeping gold core fixed. After this ligand shell is fixed and core is optimized 12 steps. The first partwise optimization scheme simply runs the optimization in turns as described. The second one tries to minimize uncertainty effects by resetting the Hessian matrix approximation to the initial value after every optimization round (24 steps ligand shell, 12 steps core).

The single point DFT potential energy evolution in the figure 3.19 shows that all optimization schemes manage to lower the potential energy by about 5.0 eV. However, after about 70 steps the potential energy of the full optimization and ligand shell optimization start to increase. This suggest that the uncertainty of the ML forces is causing problems. The partwise optimization schemes perform better than the standard approaches. The potential energy rise is more modest. The approach using the resetting of the Hessian matrix appear to perform extremely

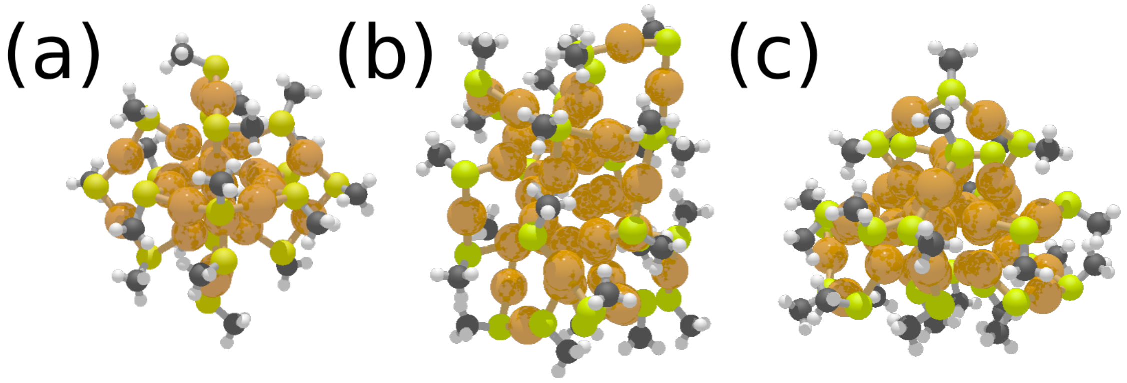


FIGURE 3.18 (a) 1500th configuration of the $[Au_{25}(SCH_3)_{18}]^-$ from 500 K DFT MD. (b) 1000th configuration of the $Au_{38}(SCH_3)_{24}$ Q isomer from the MD simulations from the reference [18]. (c) 600th configuration of the $Au_{38}(SCH_3)_{24}$ T isomer from the same source. Colors: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen. (b) and (c) are reprinted from the article [PIV] (arXiv: 2203.09788)

well. The only drawback is that the resetting induces some fluctuation.

The optimizations of the $Au_{25}(SCH_3)_{18}$ implicate that partwise schemes are the most reliable one. Hence, $Au_{38}(SCH_3)_{24}$ isomers were optimized with these two approaches. The potential energy evolution for the isomer Q in the figure 3.20 (a) shows a reasonable performance by reducing the potential energy by approximately 1.0 eV. The resetting of the Hessian matrix approximation is yet again reaching lower energies but the fluctuation is increased. The results for the isomer T in the figure 3.20 (b) show how challenging structure it is. The potential energy is reduced about 0.5 eV but soon it starts to rise. The problem most likely originates from the challenging direction estimation in the OAMLM. T isomer is much more dynamic than any other examples tested, therefore it can produce structures that are not within the reliable data space region of the OAMLM.

Let us summarize the results in the [PIV]. The ML approach using EMLM to predict atomic force norms and OAMLM to estimate force directions was developed. SOAP parameters were tested extensively with both EMLM and OAMLM. The resulting ML framework was applied to the structure optimization of the gold-thiolate rings, protecting units of the $Au_{38}(SCH_3)_{24}$ nanocluster, and DFT MD snapshots of $Au_{25}(SCH_3)_{18}$ and $Au_{38}(SCH_3)_{24}$. The performance was competent for thiolate-rings and $Au_{25}(SCH_3)_{18}$. Unexpectedly, the model performed better with $Au_{25}(SCH_3)_{18}$ than $Au_{38}(SCH_3)_{24}$, which was used to train it. The presented ML framework shows a great potential to be used in the coarse structure optimization, which would leave only fine tuning for computationally heavy DFT. Other usage case would be hybrid optimization, where most of the optimization is done with ML and a few DFT optimization steps are added time to time to correct the optimization.

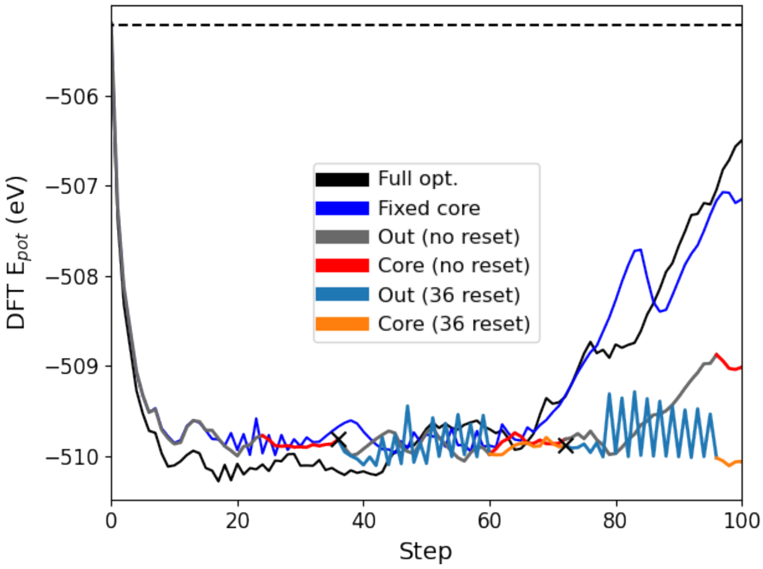


FIGURE 3.19 $Au_{25}(SCH_3)_{18}$ was optimized with four different ML BFGS schemes. The single point DFT potential energy values are decreasing efficiently. The later parts of the optimization runs show that the uncertainty builds up to the Hessian matrix approximation leading to increasing potential energies. The dashed line shows the potential energy of the initial configuration and the crosses on the curves show when the approximation of the Hessian matrix was reset.

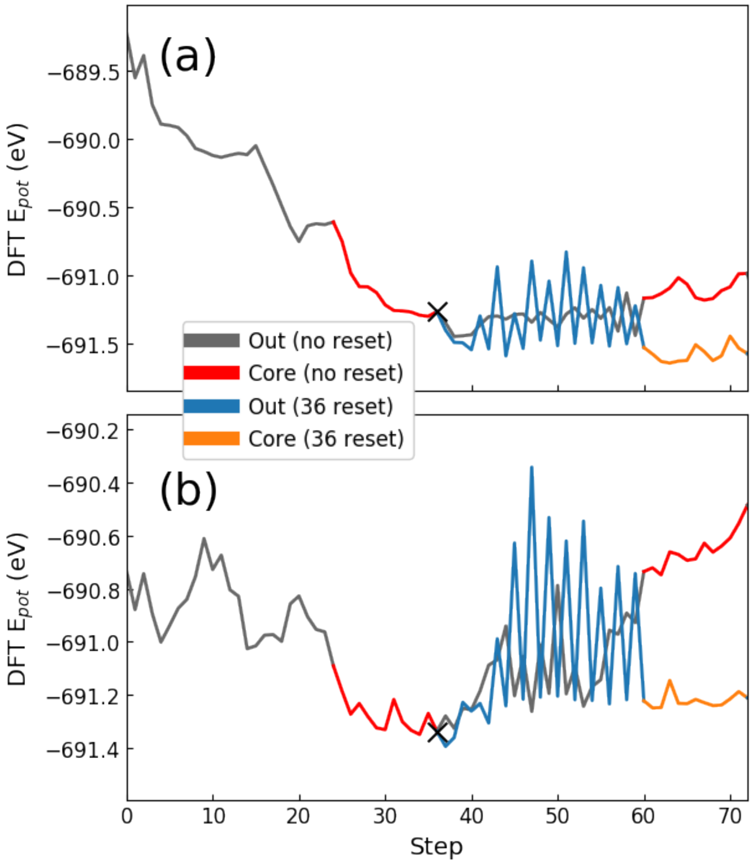


FIGURE 3.20 MD snapshots of the $Au_{38}(SCH_3)_{24}$ were optimized using ML forces. Single point DFT potential energy evolution for Q isomer is shown in (a) and for T isomer in (b). Optimization was done with partwise scheme. First protecting outer layer was optimized 24 steps and then gold core 12 steps. There were two different optimization approaches: normal BFGS and BFGS where Hessian matrix approximation was reset every 36 optimization step. Crosses on the curves show when the approximation of the Hessian matrix was reset. Reprinted from the article [PIV] (arXiv: 2203.09788)

4 CONCLUSIONS AND OUTLOOK

In this thesis, wavelet-based image comparison method and distance-based ML methods were applied to the analysis and simulations of the MPCs. The studies demonstrated three different usage cases of these data-driven methods: linking the analysis of experimental and computational data [PI], atomistic simulations with global [PII] and local ML models [PIII]-[PIV]. All of these are fundamental applications of the ML methods in the field of computational nanoscience.

In [PI], experimental and simulated TEM images of the $[Au_{25}(p-MBA)_{18}]^-$ nanocluster were compared with CW-SSIM similarity measure. The analysis reliably ruled out one of the two possible topological isomers of the nanocluster. This reduced the amount of further computational tests, that would have been required to make the conclusion about the isomers. Similarity measures were also used to generate guidelines about the orientation of the clusters in the lattice observed via experimental TEM.

In [PII], MLM and EMLM were utilized to predict potential energies for configurations of the $Au_{38}(SCH_3)_{24}$. They reached accuracy comparable to DFT but they used just a fraction of the computational resources. These ML potential energies were applied to the MC simulations, which managed to generate satisfying dynamics for the two structural isomers of the $Au_{38}(SCH_3)_{24}$. These potential energy models were considered as global ML potentials, which were specifically designed for $Au_{38}(SCH_3)_{24}$.

Poltavsky and Tkatchenko point out that global ML models are just special cases or subsets of some general model, therefore they can reach good accuracy for specific systems but they usually do not generalize well [137]. Models composed from smaller local partitions are more generalizable than global models, which is why in [PIII] and [PIV] the focus was shifted to local atomic forces. Previous potential energy methods did not contain atomic forces and numeric differentiation was not applied. Hence, it was desirable to develop a ML method, which could also calculate atomic forces. The challenge was to reliably estimate force directions and the first model, OAMLM, for this task was introduced in [PIII]. In [PIV], the force directions from OAMLM were combined with force norms predicted with EMLM. These ML estimated atomic forces were applied to the structure optimization of different gold-thiolate systems. The ML approach was very promising for coarse optimization, especially when the uncertainty of the ML forces was addressed by resetting the Hessian matrix approximation of the BFGS algorithm. Structure optimization is a routine task in computational nanoscience, therefore by using the ML optimization method, one could reduce significantly required CPU time in supercomputers. This way the resources could be saved for computationally more demanding tasks.

ML is a powerful tool in nanoscience. Its usefulness is not restricted on the atomistic simulations but it can even be used to analyze results and build a link between experimental and computational data. However, ML can never exist on its own but it requires data from other sources. It can be thought that ML lives in symbiosis with its data source. The data also determines the limitations of the method. If the data is inaccurate, such like sub-optimal SOAP descriptions in [PIII], the accuracy of the whole model is compromised. In some cases the model could be improved by introducing robustness but this is not always the case. There always exists some region in the data space, domain of applicability, where the model is the most accurate [125]. If one tries to use the model outside this region, the predicted output will naturally be uncertain. Hence, having a high quality data is vital for all ML applications.

Choosing a suitable ML method for the task at hand can be as crucial for the final result as the quality of the data. In this thesis, distance-based methods have been used, because they have very few hyperparameters, they seldom overfit in high-dimensions [101] and they can outperform deep ANNs with high-dimensional data [102]. They are also relatively interpretable, because they rely on reference data. Furthermore, the division of input and output spaces in MLM, enable unique modification possiblities as shown in [PIII] and [PIV]. However, even if the distance-based methods are versatile and reliable, it does not mean that other methods are not needed. For certain application other methods might have superior performance compared to others. For example, deep ANNs are shown to be the state of the art methods in the demanding image recognition task of microscopy [138–140]. ANNs are also only methods, which enable novel reinforcement learning approaches [141–143]. Other kernel-based ML method have also their own characteristic applications, such as Gaussian processes in Bayesian optimization [144, 145].

Emergence of the ML to many fields of expertise has not been smooth sailing but it has risen lots of discussion about the credibility of the methods. This is a welcomed discussion as ML methods are increasingly popular but understanding them is still limited. It is important to know what the method is doing and not just take method given, insert data and collect results from other end. ML has also been called as a modern day alchemy. Robbert Dijkgraaf in his column in Quanta Magazine argues that being "alchemy" is not bad at all [146]. In contrary, it may be even an essential path towards better understanding of the field. There would not be modern day chemistry, which has discovered even the MPCs focused in this thesis, if there had not been alchemists trying to turn lead into gold during the Middle ages. It is crucial to experiment with novel methods and via trial-and-error the field will mature.

REFERENCES

[1] T. Tsukuda and H. Häkkinen. *Protected metal clusters: from fundamentals to applications*. Amsterdam, Netherlands: Elsevier, 2015. ISBN: 9780444635020.

[2] M.-C. Daniel and D. Astruc. "Gold Nanoparticles: Assembly, Supramolecular Chemistry, Quantum-Size-Related Properties, and Applications toward Biology, Catalysis, and Nanotechnology". *Chem. Rev.* 104.1 (2004), pp. 293–346. DOI: 10.1021/cr030698+.

[3] K. D. M. Weerawardene, H. Häkkinen, and C. M. Aikens. "Connections Between Theory and Experiment for Gold and Silver Nanoclusters". *Annu. Rev. Phys. Chem.* 69.1 (2018), pp. 205–229. DOI: 10.1146/annurev-physchem-052516-050932.

[4] E. Boisselier and D. Astruc. "Gold nanoparticles in nanomedicine: preparations, imaging, diagnostics, therapies and toxicity". *Chem. Soc. Rev.* 38 (6 2009), pp. 1759–1782. DOI: 10.1039/B806051G.

[5] Y. Zhang, P. Song, T. Chen, X. Liu, T. Chen, Z. Wu, Y. Wang, J. Xie, and W. Xu. "Unique size-dependent nanocatalysis revealed at the single atomically precise gold cluster level". *PNAS* 115.42 (2018), pp. 10588–10593. DOI: 10.1073/pnas.1805711115.

[6] C. Sun, N. Mammen, S. Kaappa, P. Yuan, G. Deng, C. Zhao, J. Yan, S. Malola, K. Honkala, H. Häkkinen, B. K. Teo, and N. Zheng. "Atomically Precise, Thiolated Copper–Hydride Nanoclusters as Single-Site Hydrogenation Catalysts for Ketones in Mild Conditions". *ACS Nano* 13.5 (2019), pp. 5975–5986. DOI: 10.1021/acsnano.9b02052.

[7] V. Marjomäki, T. Lahtinen, M. Martikainen, J. Koivisto, S. Malola, K. Salorinne, M. Pettersson, and H. Häkkinen. "Site-specific targeting of enterovirus capsid by functionalized monodisperse gold nanoclusters". *PNAS* 111.14 (2014), pp. 1277–1281. DOI: 10.1073/pnas.1310973111.

[8] M. Azubel, S. D. Carter, J. Weiszmann, J. Zhang, G. J. Jensen, Y. Li, and R. D. Kornberg. "FGF21 trafficking in intact human cells revealed by cryo-electron tomography with gold nanoparticles". *eLife* 8 (2019), e43146. DOI: 10.7554/eLife.43146.

[9] B. K. Teo, X. Shi, and H. Zhang. "Pure gold cluster of 1:9:9:1:9:9:1 layered structure: a novel 39-metal-atom cluster [(Ph3P)14Au39Cl6]Cl2 with an interstitial gold atom in a hexagonal antiprismatic cage". *J. Am. Chem. Soc.* 114 (7 1992), pp. 2743–2745. DOI: 10.1021/ja00033a073.

[10] P. D. Jadzinsky, G. Calero, C. J. Ackerson, D. A. Bushnell, and R. D. Kornberg. "Structure of a thiol monolayer-protected gold nanoparticle at 1.1 Å resolution". *Science* 318 (5849 2007), pp. 430–433. DOI: 10.1126/science.1148624.

[11] J. Akola, M. Walter, R. L. Whetten, H. Häkkinen, and H. Grönbeck. "On the structure of thiolate-protected Au_{25}". *J. Am. Chem. Soc.* 130 (12 2008), pp. 3756–3757. DOI: 10.1021/ja800594p.

[12] M. W. Heaven, A. Dass, P. S. White, K. M. Holt, and R. W. Murray. "Crystal structure of the gold nanoparticle $[N(C_8H_{17})_4]$ $[Au_{25}(SCH_2CH_2Ph)_{18}]$". *J. Am. Chem. Soc.* 130 (12 2008), pp. 3754–3755. DOI: 10.1021/ja800561b.

[13] M. Zhu, C. M. Aikens, F. J. Hollander, G. C. Schatz, and R. Jin. "Correlating the crystal structure of a thiol-protected Au_{25} cluster and optical properties". *J. Am. Chem. Soc.* 130 (18 2008), pp. 5883–5885. DOI: 10.1021/ja801173r.

[14] M. Zhu, W. T. Eckenhoff, T. Pintauer, and R. Jin. "Conversion of Anionic $[Au_{25}(SCH_2CH_2Ph)_{18}]^-$Cluster to Charge Neutral Cluster via Air Oxidation". *J. Phys. Chem. C* 112.37 (2008), pp. 14221–14224. DOI: 10.1021/jp805786p.

[15] M. A. Tofanelli, K. Salorinne, T. W. Ni, S. Malola, B. Newell, B. Phillips, H. Häkkinen, and C. J. Ackerson. "Jahn–Teller effects in $Au_{25}(SR)_{18}$". *Chem. Sci.* 7 (3 2016), pp. 1882–1890. DOI: 10.1039/C5SC02134K.

[16] H. Qian, W. T. Eckenhoff, Y. Zhu, T. Pintauer, and R. Jin. "Total structure determination of thiolate-protected Au38 nanoparticles". *J. Am. Chem. Soc.* 132 (24 2010), pp. 8280–8281. DOI: 10.1021/ja103592z.

[17] S. Tian, Y.-Z. Li, M.-B. Li, J. Yuan, J. Yang, Z. Wu, and R. Jin. "Structural isomerism in gold nanoparticles revealed by x-ray crystallography". *Nat. Commun.* 6 (2015), p. 8667. DOI: 10.1038/ncomms9667.

[18] R. Juarez-Mosqueda, S. Malola, and H. Häkkinen. "Ab initio molecular dynamics studies of $Au_{38}(SR)_{24}$ isomers under heating". *Eur. Phys. J. D.* 73.3 (2019), p. 62. DOI: 10.1140/epjd/e2019-90441-5.

[19] M. G. Taylor and G. Mpourmpakis. "Thermodynamic stability of ligand-protected metal nanoclusters". *Nat. Commun.* 8 (2017), p. 15988. DOI: 10.1038/ncomms15988.

[20] H. Häkkinen, M. Walter, and H. Grönbeck. "Divide and Protect: Capping Gold Nanoclusters with Molecular Gold–Thiolate Rings". *J. Phys. Chem. B* 110.20 (2006), pp. 9927–9931. DOI: 10.1021/jp0619787.

[21] P. Hohenberg and W. Kohn. "Inhomogeneous Electron Gas". *Phys. Rev.* 136 (3B 1964), B864–B871. DOI: 10.1103/PhysRev.136.B864.

[22] G.-T. Bae and C. M. Aikens. "Improved ReaxFF force field parameters for Au-S-C-H systems". *J. Phys. Chem. A* 117 (40 2013), pp. 10438–10446. DOI: 10.1021/jp405992m.

[23] E. Pohjolainen, X. Chen, S. Malola, G. Groenhof, and H. Häkkinen. "A unified AMBER-compatible molecular mechanics force field for thiolate-protected gold nanoclusters". *J. Chem. Theory Comput.* 12 (3 2016), pp. 1342–1350. DOI: 10.1021/acs.jctc.5b01053.

[24] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller. "Machine Learning Force Fields". *Chemical Reviews* 121.16 (2021), pp. 10142–10186. DOI: 10.1021/acs.chemrev.0c01111.

[25] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi. "Machine Learning for Molecular Simulation". *Annual Review of Physical Chemistry* 71.1 (2020), pp. 361–390. DOI: 10.1146/annurev-physchem-042018-052331.

[26] P. Friederich, F. Häse, J. Proppe, and A. Aspuru-Guzik. "Machine-learned potentials for next-generation matter simulations". *Nat. Mater.* 20 (2021), pp. 750–761. DOI: 10.1038/s41563-020-0777-6.

[27] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques. "Recent advances and applications of machine learning in solid-state materials science". *npj Comput. Mater.* 5 (2019), p. 83. DOI: 10.1038/s41524-019-0221-0.

[28] G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, and A. Fazzio. "From DFT to machine learning: recent approaches to materials science–a review". *JPhys Materials* 2.3 (2019), p. 032001. DOI: 10.1088/2515-7639/ab084b.

[29] T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa, and K.-i. Shimizu. "Machine Learning for Catalysis Informatics: Recent Applications and Prospects". *ACS Cat.* 10.3 (2020), pp. 2260–2297. DOI: 10.1021/acscatal.9b04186.

[30] M. Andersen and K. Reuter. "Adsorption Enthalpies for Catalysis Modeling through Machine-Learned Descriptors". *Accounts of Chemical Research* 54.12 (2021), pp. 2741–2749. DOI: 10.1021/acs.accounts.1c00153.

[31] K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari, and P. Rinke. "Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra". *Adv. Sci.* 6.9 (2019), p. 1801367. DOI: https://doi.org/10.1002/advs.201801367.

[32] H. Chan, M. J. Cherukara, B. Narayanan, T. D. Loeffler, C. Benmore, S. K. Gray, and S. K. Sankaranarayanan. "Machine learning coarse grained models for water". *Nat. Commun.* 10 (2019), p. 379. DOI: 10.1038/s41467-018-08222-6.

[33] T. K. Patra, T. D. Loeffler, H. Chan, M. J. Cherukara, B. Narayanan, and S. K. R. S. Sankaranarayanan. "A coarse-grained deep neural network model for liquid water". *Appl. Phys. Lett.* 115 (19 2019), p. 193101. DOI: 10.1063/1.5116591.

[34] M. S. Jørgensen, H. L. Mortensen, S. A. Meldgaard, E. L. Kolsbjerg, T. L. Jacobsen, K. H. Sørensen, and B. Hammer. "Atomistic structure learning". *J. Chem. Phys.* 151.5 (2019), p. 054111. DOI: 10.1063/1.5108871.

[35] S. A. Meldgaard, H. L. Mortensen, M. S. Jørgensen1, and B. Hammer. "Structure prediction of surface reconstructions by deep reinforcement learning". *J. Phys. Condens. Mat.* 32.40 (2020), p. 404005. DOI: 10.1088/1361-648X/ab94f2.

[36] M.-P. V. Christiansen, H. L. Mortensen, S. A. Meldgaard, and B. Hammer. "Gaussian representation for image recognition and reinforcement learning of atomistic structure". *J. Chem. Phys.* 153.4 (2020), p. 044107. DOI: 10.1063/5.0015571.

[37] J. Kaplan. *Artificial intelligence: What everyone needs to know*. New York, United States of America: Oxford University Press, 2016.

[38] K. P. Murphy. *Machine learning: A probabilistic perspective*. Cambridge, Massachusetts: MIT Press, 2012. ISBN: 9780262305242.

[39] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. "Mastering the game of Go with deep neural networks and tree search". *Nature* 529 (2016), pp. 484–489. DOI: 10.1038/nature16961.

[40] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis. "Mastering the game of Go without human knowledge". *Nature* 550 (2017), pp. 354–359. DOI: 10.1038/nature24270.

[41] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". *Science* 362.6419 (2018), pp. 1140–1144. DOI: 10.1126/science.aar6404.

[42] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. "Highly accurate protein structure prediction with AlphaFold". *Nature* 596 (2021), pp. 583–589. DOI: 10.1038/s41586-021-03819-2.

[43] S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko. "sGDML: Constructing accurate and data efficient molecular force fields using machine learning". *Computer Physics Communications* 240 (2019), pp. 38–45. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2019.02.007.

[44] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller. "SchNetPack: A Deep Learning Toolbox For Atomistic Systems". *Journal of Chemical Theory and Computation* 15.1 (2019), pp. 448–455. DOI: 10.1021/acs.jctc.8b00908.

[45] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi. "Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons". *Phys. Rev. Lett.* 104 (13 2010), p. 136403. DOI: 10.1103/PhysRevLett.104.136403.

[46] L. Zhang, J. Han, H. Wang, R. Car, and W. E. "Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics". *Phys. Rev. Lett.* 120 (14 2018), p. 143001. DOI: 10.1103/PhysRevLett.120.143001.

[47] H. Wang, L. Zhang, J. Han, and W. E. "DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics". *Comp. Phys. Commun.* 228 (2018), pp. 178–184. DOI: 10.1016/j.cpc.2018.03.016.

[48] J. Behler and M. Parrinello. "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces". *Phys. Rev. Lett.* 98 (14 2007), p. 146401. DOI: 10.1103/PhysRevLett.98.146401.

[49] J. Behler. "Atom-centered symmetry functions for constructing high-dimensional neural network potentials". *J. Chem. Phys.* 134 (7 2011), p. 074106. DOI: 10.1063/1.3553717.

[50] O. T. Unke and M. Meuwly. "PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges". *Journal of Chemical Theory and Computation* 15.6 (2019), pp. 3678–3693. DOI: 10.1021/acs.jctc.9b00181.

[51] J. S. Smith, O. Isayev, and A. E. Roitberg. "ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost". *Chem. Sci.* 8 (4 2017), pp. 3192–3203. DOI: 10.1039/C6SC05720A.

[52] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, and A. E. Roitberg. "Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning". *Nat. commun.* 10.1 (2019), pp. 1–8.

[53] H. Chan, B. Narayanan, M. Cherukara, T. D. Loeffler, M. G. Sternberg, A. Avarca, and S. K. R. S. Sankaranarayanan. "BLAST: bridging length/timescales via atomistic simulation toolkit". *MRS Advances* 6 (2021), pp. 21–31. DOI: 10.1557/s43580-020-00002-z.

[54] L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond. "Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17". *J. Chem. Inf. Model.* 52.11 (2012), pp. 2864–2875. DOI: 10.1021/ci300415d.

[55] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld. "Quantum chemistry structures and properties of 134 kilo molecules". *Scientific Data* 1 (2014).

[56] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller. "Machine learning of accurate energy-conserving molecular force fields". *Sci. Adv.* 3.5 (2017), e1603015. DOI: 10.1126/sciadv.1603015.

[57] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko. "Quantum-chemical insights from deep tensor neural networks". *Nat. Commun.* 8 (2017), p. 13890. DOI: 10.1038/ncomms13890.

[58] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko. "Towards exact molecular dynamics simulations with machine-learned force fields". *Nat. Commun.* 9.1 (2018), pp. 1–10.

[59] A. Stuke, C. Kunkel, D. Golze, M. Todorović, J. T. Margraf, K. Reuter, P. Rinke, and H. Oberhofer. "Atomic structures and orbital energies of 61,489 crystal-forming organic molecules". *Scientific Data* 7 (2020), p. 58.

[60] D. Balcells and B. B. Skjelstad. "tmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes". *J. Chem. Inf. Model.* 60.12 (2020), pp. 6135–6146. DOI: 10.1021/acs.jcim.0c01041.

[61] V. Botu, R. Batra, J. Chapman, and R. Ramprasad. "Machine Learning Force Fields: Construction, Validation, and Outlook". *J. Phys. Chem. C* 121 (1 2017), pp. 511–522. DOI: 10.1021/acs.jpcc.6b10908.

[62] A. Stuke, M. Todorović, M. Rupp, C. Kunkel, K. Ghosh, L. Himanen, and P. Rinke. "Chemical diversity in molecular orbital energy predictions with kernel ridge regression". *J. Chem. Phys.* 150.20 (2019), p. 204121. DOI: 10.1063/1.5086105.

[63] J. Li, T. Chen, K. Lim, L. Chen, S. A. Khan, J. Xie, and X. Wang. "Deep learning accelerated gold nanocluster synthesis". *Adv. Intell. Syst.* 1 (3 2019), p. 1900029. DOI: 10.1002/aisy.201900029.

[64] S. M. Copp, S. M. Swasey, A. Gorovits, P. Bogdanov, and E. G. Gwinn. "General approach for machine learning-aided design of DNA-stabilized silver clusters". *Chem. Mater.* 32 (1 2020), pp. 430–437. DOI: 10.1021/acs.chemmater.9b04040.

[65] S. Malola, P. Nieminen, A. Pihlajamäki, J. Hämäläinen, T. Kärkkäinen, and H. Häkkinen. "A method for structure prediction of metal-ligand interfaces of hybrid nanoparticles". *Nat. Commun.* 10 (2019), p. 3973. DOI: 10.1038/s41467-019-12031-w.

[66] Z. Wang and E. Simoncelli. “Translation insensitive image similarity in complex wavelet domain”. In: *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* 2005, pp. 573–576. DOI: 10.1109/ICASSP.2005.1415469.

[67] Z. Wang and A. C. Bovik. “Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures”. *IEEE Signal Proc. Mag.* 26 (1 2009), pp. 98–117. DOI: 10.1109/MSP.2008.930649.

[68] Z. W. Wang, Z. Y. Li, S. J. Park, A. Abdela, D. Tang, and R. E. Palmer. “Quantitative Z-contrast imaging in the scanning transmission electron microscope with size-selected clusters”. *Phys. Rev. B* 84 (7 2011), p. 073408. DOI: 10.1103/PhysRevB.84.073408.

[69] Q. Zhou, S. Kaappa, S. Malola, H. Lu, D. Guan, Y. Li, H. Wang, Z. Xie, Z. Ma, H. Häkkinen, N. Zheng, X. Yang, and L. Zheng. “Real-space imaging with pattern recognition of a ligand-protected Ag_{374} nanocluster at submolecular resolution”. *Nat. Commun.* 9 (2018), p. 2948. DOI: 10.1038/s41467-018-05372-5.

[70] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. Simoncelli. “Image Quality Assessment: from Error Visibility to Structural Similarity”. *IEEE T. Image Process.* 13 (4 2004). DOI: 10.1109/TIP.2003.819861.

[71] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[72] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. “Shiftable multiscale transforms”. *IEEE Transactions on Information Theory* 38.2 (1992), pp. 587–607. DOI: 10.1109/18.119725.

[73] E. Simoncelli and W. Freeman. “The steerable pyramid: a flexible architecture for multi-scale derivative computation”. In: *Proceedings., International Conference on Image Processing*. 1995, pp. 444–447. DOI: 10.1109/ICIP.1995.537667.

[74] T. Briand, J. Vacher, B. Galerne, and J. Rabin. “The Heeger & Bergen Pyramid Based Texture Synthesis Algorithm”. *Image Processing On Line* 4 (2014), pp. 276–299. DOI: 10.5201/ipol.2014.79.

[75] S. Mallat and C. Mallat. *A Wavelet Tour of Signal Processing*. Elsevier Science & Tecnology, 1999. ISBN: 9780080520834.

[76] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt,

and SciPy 1.0 Contributors. "SciPy 1.0: fundamental algorithms for scientific computing in Python". *Nat. Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[77] A. P. Witkin. "Scale-Space Filtering". In: *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (II)*. 1983, pp. 1019–1022.

[78] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster. "DScribe: Library of descriptors for machine learning in materials science". *Comput. Phys. Commun.* 247 (2020), p. 106949. DOI: 10.1016/j.cpc.2019.106949.

[79] H. Huo and M. Rupp. *Unified Representation of Molecules and Crystals for Machine Learning*. 2017. DOI: 10.48550/ARXIV.1704.06439.

[80] A. P. Bartók, R. Kondor, and G. Csányi. "On representing chemical environments". *Phys. Rev. B* 87 (18 2013). DOI: 10.1103/PhysRevB.87.184115.

[81] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti. "Physics-Inspired Structural Representations for Molecules and Materials". *Chem. Rev.* 121.16 (2021), pp. 9759–9815. DOI: 10.1021/acs.chemrev.1c00021.

[82] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. "Fast and accurate modeling of molecular atomization energies with machine learning". *Phys. Rev. Lett.* 108 (5 2012), p. 058301. DOI: 10.1103/PhysRevLett.108.058301.

[83] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento. "Crystal structure representations for machine learning models of formation energies". *Int. J. Quantum Chem.* 115 (16 2015), pp. 1094–1101. DOI: 10.1002/qua.24917.

[84] M. J. Willatt, F. Musil, and M. Ceriotti. "Atom-density representations for machine learning". *J. Chem. Phys.* 150.15 (2019), p. 154110. DOI: 10.1063/1.5090481.

[85] N. Canterakis. "3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition". In: *11th Scandinavian Conference on Image Analysis*. 1999, pp. 85–93.

[86] M. Novotni and R. Klein. "Shape retrieval using 3D Zernike descriptors". *Computer-Aided Design* 36 (11 2004), pp. 1047–1062. DOI: 10.1016/j.cad.2004.01.005.

[87] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko. "Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space". *J. Phys. Chem. Lett.* 6 (12 2015), pp. 2326–2331. DOI: 10.1021/acs.jpclett.5b00831.

[88] P.-O. Löwdin. "On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals". *The Journal of Chemical Physics* 18.3 (1950), pp. 365–375. DOI: 10.1063/1.1747632.

[89] A. H. de Souza Júnior, F. Corona, G. A. Barreto, Y. Miche, and A. Lendasse. "Minimal Learning Machine: A novel supervised distance-based approach for regression and classification". *Neurocomputing* 164 (2015), pp. 34–44. DOI: 10.1016/j.neucom.2014.11.073.

[90] T. Kärkkäinen. "Extreme minimal learning machine: Ridge regression with distance-based basis". *Neurocomputing* 342 (2019), pp. 33–48. DOI: 10.1016/j.neucom.2018.12.078.

[91] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. "Extreme learning machine:A new learning scheme of feedforward neural networks". In: *Proc. IEEEInt. Joint Conf. Neural Netw.* Vol. 2. 2004, pp. 985–990.

[92] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. "Extreme learning machine: Theory and applications". *Neurocomputing* 70 (1-3 2006), pp. 489–501. DOI: 10.1016/j.neucom.2005.12.126.

[93] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang. "Extreme learning machine for regression and multiclass classification". *IEEE Trans. Syst., Man, Cybern. B, Cybern.* 42 (2 2012), pp. 513–529. DOI: 10.1109/TSMCB.2011.2168604.

[94] E. Cambria, G.-B. Huang, L. L. C. Kasun, H. Zhou, C. M. Vong, J. Lin, J. Yin, Z. Cai, Q. Liu, K. Li, V. C. Leung, L. Feng, Y.-S. Ong, M.-H. Lim, A. Akusok, A. Lendasse, F. Corona, R. Nian, Y. Miche, P. Gastaldo, R. Zunino, S. Decherchi, X. Yang, K. Mao, B.-S. Oh, J. Jeon, J. Jeon, K.-A. Toh, A. B. J. Teoh, J. Kim, and H. Yu. "Extreme learning machines [trends & controversies]". *IEEE Intelligent Systems* 28 (6 2013), pp. 30–59. DOI: 10.1109/MIS.2013.140.

[95] A. Akusok, K.-M. Björk, Y. Miche, and A. Lendasse. "High-performance extreme learning machines: A complete toolbox for big data applications". *IEEE Access* 3 (2015), pp. 1011–1025. DOI: 10.1109/ACCESS.2015.2450498.

[96] W. Navidi, W. S. M. Jr., and W. Hereman. "Statistical methods in surveying by trilateration". *Comput. Stat. Data Anal.* 27 (2 1998), pp. 209–227. DOI: 10.1016/S0167-9473(97)00053-4.

[97] D. P. P. Mesquita, J. P. P. Gomes, and A. H. Souza Junior. "Ensemble of efficient minimal learning machines for classification and regression". *Neural Process Lett.* (3 46), pp. 751–766. DOI: 10.1007/s11063-017-9587-5.

[98] K. S. Arun, T. S. Huang, and S. D. Blostein. "Least-Squares Fitting of Two 3-D Point Sets". *IEEE T. Pattern Anal.* PAMI-9 (5 1987), pp. 698–700. DOI: 10.1109/TPAMI.1987.4767965.

[99] P. J. Huber. *Robust Statistics*. New Jersey, USA: John Wiley & Sons, Inc, 1981. ISBN: 0-47141805-6.

[100] J. P. P. Gomes, D. P. P. Mesquita, A. L. Freire, A. H. SouzaJunior, and T. Kärkkäinen. "A Robust Minimal Learning Machine based on the M-Estimator". In: *ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. 2017, pp. 383–388.

[101] J. Hämäläinen, A. S. C. Alencar, T. Kärkkäinen, C. L. C. Mattos, A. H. Souza Júnior, and J. P. P. Gomes. "Minimal Learning Machine: Theoretical results and clustering-based reference point selection". *J. Mach. Learn. Res.* 21.239 (2020), pp. 1–29.

[102] J. Linja, J. Hämäläinen, P. Nieminen, and T. Kärkkäinen. "Do Randomized Algorithms Improve the Efficiency of Minimal Learning Machine?" *Mach. Learn. Knowl. Extr.* 2.4 (2020), pp. 533–557. DOI: 10.3390/make2040029.

[103] M. Verleysen and D. François. "The Curse of Dimensionality in Data Mining and Time Series Prediction". In: *Computational Intelligence and Bioinspired Systems*. Ed. by J. Cabestany, A. Prieto, and F. Sandoval. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 758–770. ISBN: 978-3-540-32106-4. DOI: 10.1007/11494669_93.

[104] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller. "Unmasking Clever Hans predictors and assessing what machines really learn". *Nat. Commun.* 10 (2019), p. 1096. DOI: 10.1038/s41467-019-08987-4.

[105] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. "Molecular dynamics with coupling to an external bath". *J. Chem. Phys.* 81.8 (1984), pp. 3684–3690. DOI: 10.1063/1.448118.

[106] D. S. Lemons and A. Gythiel. "Paul Langevin's 1908 paper "On the Theory of Brownian Motion" ["Sur la théorie du mouvement brownien," C. R. Acad. Sci. (Paris) 146, 530–533 (1908)]". *Am. J. Phys.* 65.11 (1997), pp. 1079–1081. DOI: 10.1119/1.18725.

[107] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. New York, USA: Oxford University Press, 1989.

[108] W. G. Hoover. "Canonical dynamics: Equilibrium phase-space distributions". *Phys. Rev. A* 31 (3 1985), pp. 1695–1697. DOI: 10.1103/PhysRevA.31.1695.

[109] W. G. Hoover. "Constant-pressure equations of motion". *Phys. Rev. A* 34 (3 1986), pp. 2499–2500. DOI: 10.1103/PhysRevA.34.2499.

[110] L. Verlet. "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules". *Phys. Rev.* 159 (1 1967), pp. 98–103. DOI: 10.1103/PhysRev.159.98.

[111] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. New York, USA: Springer-Verlag New York Inc., 1999. ISBN: 0-387-98707-X.

[112] C. G. Broyden. "The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations". *IMA Journal of Applied Mathematics* 6.1 (1970), pp. 76–90. ISSN: 0272-4960. DOI: 10.1093/imamat/6.1.76.

[113] R. Fletcher. "A new approach to variable metric algorithms". *The Computer Journal* 13.3 (1970), pp. 317–322. ISSN: 0010-4620. DOI: 10.1093/comjnl/13.3.317.

[114] D. Goldfarb. "A family of variable-metric methods derived by variational means". *Math. Comp.* 24.109 (1970), pp. 23–26. DOI: 10.1090/S0025-5718-1970-0258249-6.

[115] D. F. Shanno. "Conditioning of quasi-Newton methods for function minimization". *Math. Comp.* 24.111 (1970), pp. 647–656. DOI: 10.1090/S0025-5718-1970-0274029-X.

[116] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Cambridge: Cambridge University Press, 1988. ISBN: 0-521-43064-X.

[117] J. Nocedal and S. J. Wright. *Numerical Optimization*. New York: Springer cop., 1999. ISBN: 0-387-98793-2.

[118] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen. "The atomic simulation environment—a Python library for working with atoms". *J. Phys. Condens. Mat.* 29.27 (2017), p. 273002. DOI: 10.1088/1361-648x/aa680e.

[119] J. J. Mortensen, L. B. Hansen, and K. W. Jacobsen. "Real-space grid implementation of the projector augmented wave method". *Phys. Rev. B* 71 (3 2005), p. 035109. DOI: 10.1103/PhysRevB.71.035109.

[120] J. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Dułak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. A. Hansen, H. H. Kristoffersen, M. Kuisma, A. H. Larsen, L. Lehtovaara, M. Ljungberg, O. Lopez-Acevedo, P. G. Moses, J. Ojanen, T. Olsen, V. Petzold, N. A. Romero, J. Stausholm-Møller, M. Strange, G. A. Tritsaris, M. Vanin, M. Walter, B. Hammer, H. Häkkinen, G. K. H. Madsen, R. M. Nieminen, J. K. Nørskov, M. Puska, T. T. Rantala, J. Schøtz, K. S. Thygesen, and K. W. Jacobsen. "Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method". *J. Phys.: Condens. Matter* 22.25 (2010), p. 253202. DOI: 10.1088/0953-8984/22/25/253202.

[121] J. P. Perdew, K. Burke, and M. Ernzerhof. "Generalized gradient approximation made simple". *Phys. Rev. Lett.* 77 (18 1996), p. 3865. DOI: 10.1103/PhysRevLett.77.3865.

[122] M. F. Matus, S. Malola, E. Kinder Bonilla, B. M. Barngrover, C. M. Aikens, and H. Häkkinen. "A topological isomer of the $Au_{25}(SR)_{18}{}^{-}$ nanocluster". *Chem. Commun.* 56 (58 2020), pp. 8087–8090. DOI: 10.1039/D0CC03334K.

[123] E. Kalenius, S. Malola, M. F. Matus, R. Kazan, T. Bürgi, and H. Häkkinen. "Experimental Confirmation of a Topological Isomer of the Ubiquitous $Au_{25}(SR)_{18}$ Cluster in the Gas Phase". *J. Am. Chem. Soc.* 143.3 (2021), pp. 1273–1277. DOI: 10.1021/jacs.0c11509.

[124] Y. Cao, S. Malola, M. F. Matus, T. Chen, Q. Yao, R. Shi, H. Häkkinen, and J. Xie. "Reversible isomerization of metal nanoclusters induced by intermolecular interaction". *Chem* 7.8 (2021), pp. 2227–2244. DOI: 0.1016/j.chempr.2021.06.023.

[125] C. Sutton, M. Boley, L. M. Ghiringhelli, M. Rupp, J. Vreeken, and M. Scheffler. "Identifying domains of applicability of machine learning models for materials science". *Nat. Commun.* 11 (2020), p. 4428. DOI: 10.1038/s41467-020-17112-9.

[126] E. Prodan and W. Kohn. "Nearsightedness of electronic matter". *Proceedings of the National Academy of Sciences* 102.33 (2005), pp. 11635–11638. DOI: 10.1073/pnas.0505436102.

[127] H. Hellman. "Einführung in die Quantenchemie". *Franz Deuticke, Leipzig* 285 (1937).

[128] R. P. Feynman. "Forces in Molecules". *Phys. Rev.* 56 (4 1939), pp. 340–343. DOI: 10.1103/PhysRev.56.340.

[129] A. Lemme, K. Neumann, R. Reinhart, and J. Steil. "Neural learning of vector fields for encoding stable dynamical systems". *Neurocomputing* 141 (2014), pp. 3–14. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2014.02.012.

[130] P. Koskinen and V. Mäkinen. "Density-functional tight-binding for beginners". *Comp. Mater. Sci.* 47.1 (2009), pp. 237–253. ISSN: 0927-0256. DOI: 10.1016/j.commatsci.2009.07.013.

[131] T. F. Gonzalez. "Clustering to minimize the maximum intercluster distance". *Theor. Comput. Sci.* 38 (1985), pp. 293–306. DOI: 10.1016/0304-3975(85)90224-5.

[132] M. J. Hostetler, J. E. Wingate, C.-J. Zhong, J. E. Harris, R. W. Vachet, M. R. Clark, J. D. Londono, S. J. Green, J. J. Stokes, G. D. Wignall, G. L. Glish, M. D. Porter, N. D. Evans, and R. W. Murray. "Alkanethiolate Gold Cluster Molecules with Core Diameters from 1.5 to 5.2 nm: Core and Monolayer Properties as a Function of Core Size". *Langmuir* 14.1 (1998), pp. 17–30. DOI: 10.1021/la970588w.

[133] S. Chen, A. C. Templeton, and R. W. Murray. “Monolayer-Protected Cluster Growth Dynamics”. *Langmuir* 16.7 (2000), pp. 3543–3548. DOI: 10.1021/la991206k.

[134] M. K. Corbierre and R. B. Lennox. “Preparation of Thiol-Capped Gold Nanoparticles by Chemical Reduction of Soluble Au(I)-Thiolates”. *Chem. Mater.* 17.23 (2005), pp. 5691–5696. DOI: 10.1021/cm051115a.

[135] H. Grönbeck, M. Walter, and H. Häkkinen. “Theoretical Characterization of Cyclic Thiolated Gold Clusters”. *J. Am. Chem. Soc.* 128 (31 2006), pp. 10268–10275. DOI: 10.1021/ja062584w.

[136] Z. Wang, H. Wang, and S. Qiu. “A new method for numerical differentiation based on direct and inverse problems of partial differential equations”. *Appl. Math. Lett.* 43 (2015), pp. 61–67. ISSN: 0893-9659. DOI: 10.1016/j.aml.2014.11.016.

[137] I. Poltavsky and A. Tkatchenko. “Machine Learning Force Fields: Recent Advances and Remaining Challenges”. *J. Phys. Chem. Lett.* 12.28 (2021), pp. 6551–6564. DOI: 10.1021/acs.jpclett.1c01204.

[138] S. Helgadottir, A. Argun, and G. Volpe. “Digital video microscopy enhanced by deep learning”. *Optica* 6.4 (2019), pp. 506–513. DOI: 10.1364/OPTICA.6.000506.

[139] B. Midtvedt, S. Helgadottir, A. Argun, J. Pineda, D. Midtvedt, and G. Volpe. “Quantitative digital microscopy with deep learning”. *Appl. Phys. Rev.* 8.1 (2021), p. 011310. DOI: 10.1063/5.0034891.

[140] S. Helgadottir, B. Midtvedt, J. Pineda, A. Sabirsh, C. B. Adiels, S. Romeo, D. Midtvedt, and G. Volpe. “Extracting quantitative biological information from bright-field cell images using deep learning”. *Biophys. Rev.* 2.3 (2021), p. 031401. DOI: 10.1063/5.0044782.

[141] T. D. Loeffler, S. Banik, T. K. Patra, M. Sternberg, and S. K. Sankaranarayanan. “Reinforcement learning in discrete action space applied to inverse defect design”. *J. Phys. Commun.* 5.3 (2021), p. 031001. DOI: 10.1088/2399-6528/abe591.

[142] S. Banik, T. D. Loeffler, R. Batra, H. Singh, M. J. Cherukara, and S. K. R. S. Sankaranarayanan. “Learning with Delayed Rewards—A Case Study on Inverse Defect Design in 2D Materials”. *ACS Appl. Mater. Inter.* 13.30 (2021), pp. 36455–36464. DOI: 10.1021/acsami.1c07545.

[143] S. Sankaranarayanan, S. Manna, T. Loeffler, R. Batra, S. Banik, H. Chan, K. Sasikumar, M. Sternberg, T. Peterka, M. Cherukara, S. Gray, and B. Sumpter. *Learning in Continuous Action Space for Determination of High Dimensional Potential Energy Surfaces*. 2021. DOI: 10.21203/rs.3.rs-284625/v1.

[144] M. Todorović, M. U. Gutmann, J. Corander, and P. Rinke. "Bayesian inference of atomistic structure in functional materials". *NPJ Comput. Mater.* 5 (2019), p. 35. DOI: 10.1038/s41524-019-0175-2.

[145] S. Kaappa, E. G. del Río, and K. W. Jacobsen. "Global optimization of atomic structures with gradient-enhanced Gaussian process regression". *Phys. Rev. B* 103 (17 2021), p. 174114. DOI: 10.1103/PhysRevB.103.174114.

[146] R. Dijkgraaf. "The Uselessness of Useful Knowledge". *Quanta Magazine* (Oct. 20, 2021). URL: https://www.quantamagazine.org/science-has-entered-a-new-era-of-alchemy-good-20211020 (visited on 03/25/2022).

ORIGINAL PAPERS

PI

ENGINEERING COLLOIDAL CRYSTALS OF ATOMICALLY PRECISE GOLD NANOPARTICLES PROMOTED BY PARTICLE SURFACE DYNAMICS

by

Qiaofeng Yao, Lingmei Liu, Sami Malola, Hongyi Xu, Zhenna Wu, Tiankai Chen,
Yitao Cao, María Francisca Matus, **Antti Pihlajamäki**, Shuangquan Zang, Yu Han,
Hannu Häkkinen and Jiangping Xie 2022

Accepted in Nature Chemistry

Surface Dynamics Promoted Supercrystal Engineering of Atomically Precise Gold Nanoparticles

Qiaofeng Yao[1,7#], Lingmei Liu[2#], Sami Malola[3#], Meng Ge[4], Hongyi Xu[4], Zhennan Wu[1], Tiankai Chen[1], Yitao Cao[1], María Francisca Matus[3], Antti Pihlajamäki[3], Yu Han[5]*, Hannu Häkkinen[3,6]* and Jianping Xie[1,7]*

[1] Department of Chemical and Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117585.

[2] Multi-scale Porous Materials Center, Institute of Advanced Interdisciplinary Studies & School of Chemistry and Chemical Engineering, Chongqing University, Chongqing, 400044 P. R. China.

[3] Department of Physics, Nanoscience Center, University of Jyväskylä, FI-40014 Jyväskylä, Finland.

[4] Department of Materials and Environmental Chemistry, Stockholm University, SE-106 91 Stockholm, Sweden.

[5] Advanced Membranes and Porous Materials Center, Physical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia.

[6] Department of Chemistry, Nanoscience Center, University of Jyväskylä, FI-40014 Jyväskylä, Finland.

[7] Joint School of National University of Singapore and Tianjin University, International Campus of Tianjin University, Binhai New City, Fuzhou, China 350207.

* Corresponding author. Email: yu.han@kaust.edu.sa (Y.H.); hannu.j.hakkinen@jyu.fi (H.H.); chexiej@nus.edu.sg (J.X.)

[#] These authors contributed equally to this work.

Abstract

Controllable packing of functional nanoparticles (NPs) into supercrystals is of core interest in the development of NP-based metamaterials. Compared with the conventional crystallization method that treats NPs as hard spheres, here we demonstrate at the molecular level that the size, morphology, and symmetry of unary supercrystals can be tailored by using the surface dynamics of NPs. In the presence of excess tetraethylammonium cations, atomically precise $[Au_{25}(SR)_{18}]^-$ NPs (SR = thiolate) can be crystallized into micro-meter-sized hexagonal rod-like supercrystals. Experimental characterization and theoretical modeling reveal a R-3m space group, in which NPs are aligned into polymeric chains through a unique $SR-[Au(I)-SR]_4$ inter-particle linker. This linker is established by the asymmetric conjugation of the dynamically detached $SR-[Au(I)-SR]_2$ protecting motifs between neighbored NPs, which is made possible by intensive ion-pairing-*cum*-CH···π interactions between tetraethylammonium cations and SR ligands. By changing the dosage and type of tetraalkylammonium cations, the symmetry, morphology, and size of supercrystals can be systematically tuned. This work not only provides a convenient method for supercrystal engineering, but also highlights the importance of surface dynamics in dictating the assembly behavior of NPs.

Nature presents a great variety of crystalline materials through the orderly arrangement of atoms, ions, and molecules. In the past few decades, the scope of crystalline materials has been remarkably expanded by using functional inorganic nanoparticles (NPs) as "programmable atom equivalents (PAEs)"[1-6]. Assembling monodisperse NPs into supercrystals has proven an effective way to modulate their intrinsic optical, electronic, magnetic, and catalytic activities through inter-particle coupling and crystal order coherence[4, 7-12], which can be promoted by diverse inter-particle interactions, including electrostatic interaction[6], depletion force[13], metallophilicity[8, 14], H-bond[15], and biorecognition interaction[16, 17]. In these documented successful attempts, inorganic NPs are generally regarded as hard (or slightly deformable) spheres, and their stacking symmetry is determined by their "static" surface patterns. Intriguingly, recent advances in atomically precise nanoscience reveal marked structural dynamics in/between inorganic core and organic protecting shell of NPs[18-21], although such dynamics has not yet been utilized to regulate the assembly behavior of NPs.

Atomically precise thiolated gold NPs or "nanoclusters" with a specific chemical formula $[\mathrm{Au}_m(\mathrm{SR})_n]^q$ (m, n, and q are the number of gold atoms, thiolate ligands (SR), and net charge per particle, respectively) are an emerging family of ultra-small metal particles (core size <3 nm)[22, 23]. They can be synthesized and characterized with atomic precision. Recent advances in X-ray crystallography suggest a core-shell structure of $[\mathrm{Au}_m(\mathrm{SR})_n]^q$ NPs with well-ordered atomic (metal) and molecular (ligand) arrangement patterns at different structural hierarchies, reminiscent of biomolecules like proteins[3, 24]. Such well-ordered intra- and inter-particle arrangement patterns are largely sustained by supramolecular interactions, such as $\mathrm{CH}\cdots\pi$ interaction[3], metallophilicity[8, 14], and conformational matching of protecting motifs[24]. More intriguingly, the diffusion of intra-particle metal atoms and the migration of surface protecting motifs have been observed in many $[Au_m(SR)_n]^q$ or their alloy NPs, suggesting the existence of structural dynamics at the molecular and atomic levels[19-21]. In addition, $[Au_m(SR)_n]^q$ NPs also exhibit size- and structure-sensitive physicochemical properties (e.g., HOMO-LUMO transitions[25, 26], luminescence[27-29], and intrinsic chirality[30]), which provide a good channel to probe the growth fundamentals of supercrystals.

Herein, we demonstrate that the long-overlooked surface dynamics of Au NPs can pave an alternative way for regulating the structure (e.g., size, shape, and packing symmetry) of the NP supercrystals. Atomically precise $[Au_{25}(p\text{-MBA})_{18}]^-$ (p-MBA = *para*-mercaptobenzoic acid) are employed as model NPs, and tetraalkylammonium cations are used to regulate their surface dynamics via ion-pairing-*cum*-CH···π interactions (panel (vii), Fig. 1A). In the absence of any tetraalkylammonium cations, the deprotonated NPs tend to pack as hard spheres to cubic (face-centered-cubic or FCC-like) superlattices, forming a macroscopic octahedral supercrystal shape (panel (i)-(iii), Fig. 1A). However, the introduction of a suitable tetraalkylammonium cation (e.g., tetraethylammoniun (TEA^+)) as structure-directing agent will give rise to NP polymers connected by $SR\text{-}[Au(I)\text{-}SR]_4$ linkers (panel (iv), Fig. 1A), which are formed by the asymmetric conjugation of two dynamically detached $SR\text{-}[Au(I)\text{-}SR]_2$ motifs from neighbored NPs. Close-packing of as-formed NP polymers leads to micro-meter-sized hexagonal rod-like supercrystals (panel (v)-(vi), Fig. 1A). The NP packing symmetry and morphology of these NP metamaterials can be tuned by the dosage and size of related tetraalkylammonium cations. This work demonstrates a facile method for engineering the morphology and symmetry of crystalline

NP metamaterials at the micro-meter size regime and highlights the unconventional importance of surface dynamics of NPs in determining their assembly behavior.

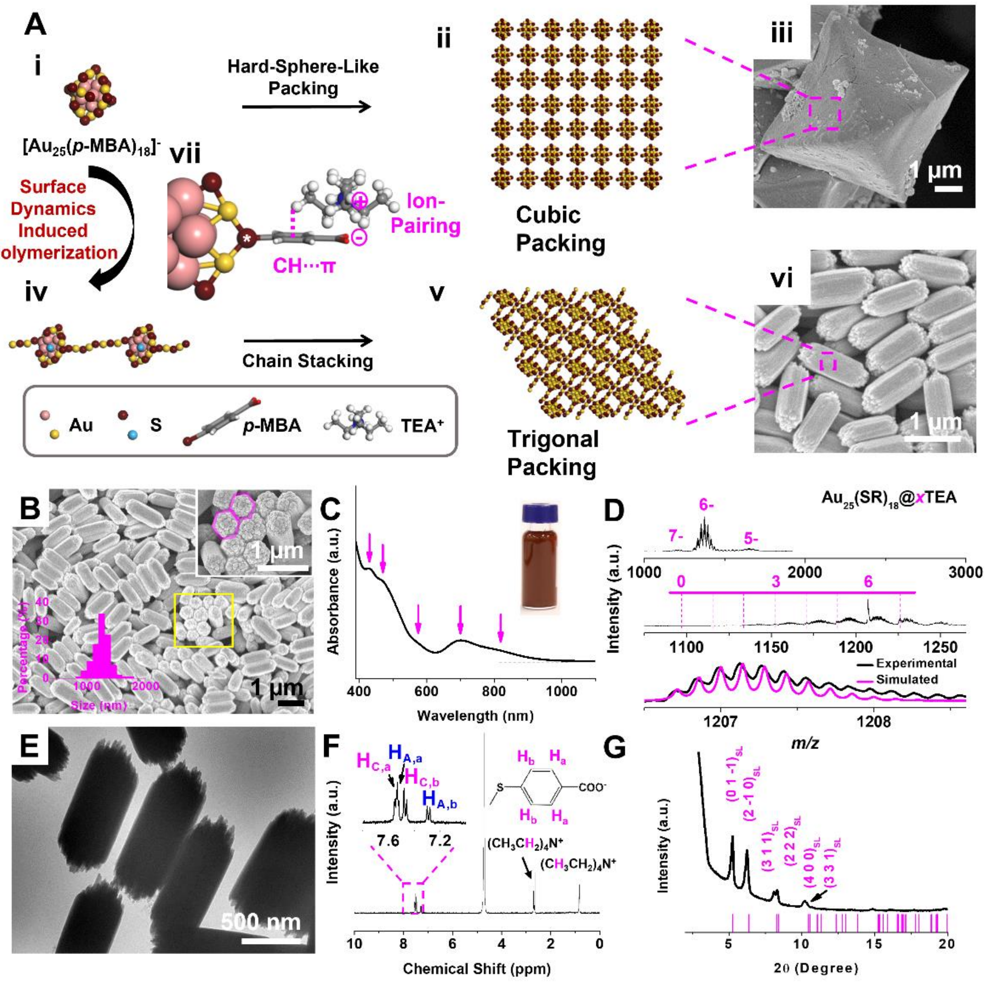


Figure 1. Crystallization of $[Au_{25}(p\text{-MBA})_{18}]^-$ nanoparticles into hexagonal rod-like supercrystals. (A) Schematic illustration of crystallizing $[Au_{25}(p\text{-MBA})_{18}]^-$ (*p*-MBA = *para*-mercaptobenzoic acid) NPs into hexagonal rod-like supercrystals in the presence of tetraethylammonium cations (TEA^+): (i) hard-sphere-like $[Au_{25}(p\text{-MBA})_{18}]^-$; (ii) face-centered-cubic (FCC) superlattice of $[Au_{25}(p\text{-MBA})_{18}]^-$; (iii) a typical field-emission

scanning electron microscopy (FESEM) image of octahedral supercrystals; (iv) TEA^+-induced one-dimensional alignment of $[Au_{25}(p\text{-MBA})_{18}]^-$; (v) R-3m superlattice of $[Au_{25}(p\text{-MBA})_{18}]^-$; (vi) a typical FESEM image of hexagonal rod-like supercrystals; and (vii) the ion-pairing-*cum*-CH···π interactions between TEA^+ and *p*-MBA ligand. (B) FESEM, (E) transmission electron microscopy (TEM) images, and (G) powder X-ray diffraction (P-XRD) pattern of hexagonal rod-like supercrystals formed at the molar ratio of TEA^+/Li^+, $R_{TEA/Li} = 3/1$. (C) Ultraviolet-visible (UV-vis) absorption, (D) electrospray ionization mass spectrometry (ESI-MS), and (F) ^1H nuclear magnetic resonance (^1H-NMR) spectra of the hexagonal rod-like supercrystals re-dissolved in (deuterated) water. The insets of (B) are longitudinal size histogram of supercrystals (bottom) and zoom-in view of the squared area (top), where the magenta hexagons outline the cross-sections of two rods. The inset of (C) is a digital photo of supercrystals dispersed in dimethyl sulfoxide (DMSO). The top panel of (D) is wide-range mass spectrum, where the charges of particle peaks are labelled; the middle panel of (D) is zoom-in view of particle peaks carrying 7-charge, where the number of TEA^+ bonded to individual $[Au_{25}(p\text{-MBA})_{18}]^-$ is indicated by the dashed droplines; the black and magenta lines in the bottom panel of (D) are experimental and simulated isotope patterns of $[Au_{25}(p\text{-MBA})_{18}@6TEA - 12H]^{7-}$, respectively. Left inset of (F) is zoom-in view of the aromatic region of ^1H-NMR spectrum, and right inset depicts the hydrogen atoms of *p*-MBA ligands in different chemical environments. The Miller indexes of superlattice (SL) planes are labeled in (G).

Results

Synthesis of $[Au_{25}(p\text{-}MBA)_{18}]^-$ Supercrystals. Molecularly pure $[Au_{25}(p\text{-}MBA)_{18}]^-$ NPs were synthesized by a carbon monoxide (CO)-reduction method reported elsewhere (Supplementary Fig. 1 and Supplementary Note 1)[31]. The patching of TEA^+ on the particle surface was conducted by cyclic cation exchange of freshly prepared $[Au_{25}(p\text{-}MBA)_{18}]^-$ with excess TEA^+ (small Li^+ was used as co-cations (with a molar ratio of TEA^+/Li^+, $R_{TEA/Li}$ = 3/1) to neutralize the surface charge of NPs), followed by crystallization via a selective evaporation approach in a dual-solvent system of water and dimethyl sulfoxide (DMSO)[32].

The precipitates produced by selective solvent evaporation can be re-dispersed in its mother liquid or fresh DMSO by shaking briefly (inset, Fig. 1C), which suggests the successful formation of micro-meter-sized supercrystals. Field-emission scanning electron microscopy (FESEM) and transmission electron microscopy (TEM) analyses (Fig. 1B and 1E) on the supercrystals manifest a rod-like morphology with a typical longitudinal size of 1.25 ± 0.15 µm (100 rods counted), an aspect ratio (*r*) of 2.45, and a hexagonal cross-section (top inset, Fig. 1B). Powder X-ray diffraction (P-XRD) pattern (Fig. 1G) of the rod-like supercrystals shows clearly discernable peaks in the 2θ regime of 2-20˚, indicating their highly crystalline nature.

In sharp contrast to the well-maintained structural architectures in DMSO, the as-obtained hexagonal rod-like supercrystals can be completely dissociated into discrete $[Au_{25}(p\text{-}MBA)_{18}]^-$ NPs in water. The combined UV-vis absorption spectroscopy (Fig. 1C) and electrospray ionization mass spectrometry (ESI-MS, Fig. 1D) analyses suggest that the size- and structure-uncompromised $[Au_{25}(p\text{-}MBA)_{18}]^-$ NPs are recovered from the dissociated supercrystals, with a typical recovery of 88.7% measured based on the characteristic molar extinction coefficient at ~690 nm (ε_{690}). The ESI-MS spectrum also suggests a superior structural stability for Au_{25}(*p*-MBA)$_{18}$@6TEA, where extraordinarily high population of Au_{25}(*p*-MBA)$_{18}$@6TEA is observed in the particle peaks carrying 7- charge. The quantitative ^{1}H nuclear magnetic resonance (^{1}H-NMR, Fig. 1F) analysis manifests 6.07 ± 0.11 (three independent samples tested) TEA$^+$ molecules bonded to individual particle (i.e., $x = 6$ in Au_{25}(*p*-MBA)$_{18}$@*x*TEA). More details about molecular characterization of the re-dissolved hexagonal rod-like supercrystals can be found in Supplementary Note 2 and Supplementary Fig. 2.

Packing and atomic structure of NPs in the supercrystals. The supercrystal structure formed by the TEA$^+$-bonded [Au_{25}(*p*-MBA)$_{18}$]$^-$ NPs (i.e., Au_{25}(*p*-MBA)$_{18}$@6TEA) was examined by 3D electron diffraction (3D ED) technique. The 3D ED datasets collected from hexagonal rod-like supercrystals (Fig. 2C and Supplementary Fig. 3) reveal a R-3m unit cell with cell parameters of $a = b = 27.7$ Å, $c = 23.4$ Å, $\alpha = \beta = 90^\circ$, and $\gamma = 120^\circ$. With the determined unit cell, the main diffraction peaks at $2\theta = 5.22^\circ$, 6.25°, 8.08°, 8.33°, 10.21°, and 10.37° can be assigned to the superlattice (SL) planes of $(0\ 1\ \text{-}1)_{SL}$, $(2\ \text{-}1\ 0)_{SL}$, $(3\ 1\ 1)_{SL}$, $(2\ 2\ 2)_{SL}$, $(4\ 0\ 0)_{SL}$, and $(3\ 3\ 1)_{SL}$, respectively (Fig. 1G).

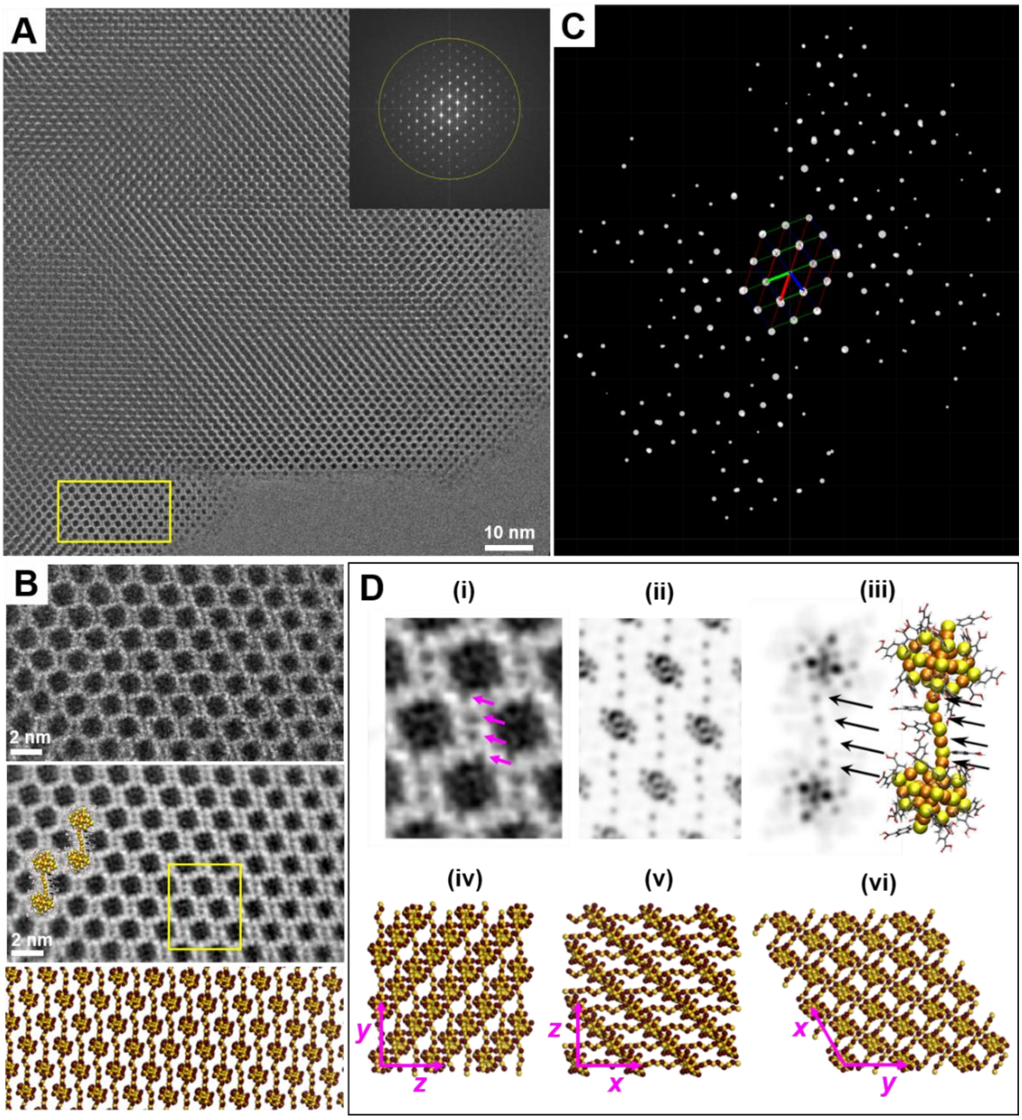


Figure 2. Packing structure determination of hexagonal rod-like supercrystals. (A, B) Bright-field ultralow-dose TEM image, (C) reconstructed 3D electron diffraction lattice, and (D) particle packing models of rod-like supercrystals formed at $R_{\mathrm{TEA/Li}}$ = 3/1 along [1 -1 -1]$_{\mathrm{SL}}$ zone axis. The inset of (A) shows the fast Fourier transfer pattern of the corresponding TEM image. The top panel of (B) is the zoom-in view of the squared area in (A), while the middle and bottom panels are the contrast transfer function (CTF)-corrected image and structure model of the top panel, respectively. Panel (i) of (D) is the

enlarged view of the squared area of (B), while panel (ii) is the simulated TEM image reproducing that in panel (i); panel (iii) is the simulated TEM image (left) and structure model (right) of the $[Au_{25}(SR)_{18}]^-$ dimer formed through SR-[Au(I)-SR]$_4$ inter-particle linker (SR = thiolate); panel (iv), (v) and (vi) are the R-3m superlattices of $[Au_{25}(p\text{-}MBA)_{18}]^-$ viewed along *x* (iv), *y* (v), and *z* (vi) axis, respectively. Color code: golden/orange, Au; purple/yellow, S; gray, C; red, O; light gray, H; the hydrocarbon tails of *p*-MBA ligands are only shown in panel (iii) of (D) for clarity purpose.

In order to furnish more structural details, we acquired the high-resolution images of the supercrystals consisting of $[Au_{25}(p\text{-}MBA)_{18}]^-$ NPs by using our recently developed ultralow-dose TEM (ULD-TEM) technique, which can effectively avoid electron beam-induced structural changes/damages[33, 34]. The bright-field ULD-TEM image (the total electron dose as low as ~15 $e^{-}\cdot Å^{-2}$) taken at the Scherzer focus along the $[1\ \text{-}1\ \text{-}1]_{SL}$ incidence (Fig. 2A) shows a highly ordered structure composed of monodisperse NPs. The corresponding fast Fourier transfer pattern is shown as an inset in Fig. 2A, and the yellow circle indicates the frequency of 2.5 Å. Based on the structure of the unit cell solved by the 3D ED, we indexed the projected direction of this typical ULD-TEM image as $[1\ \text{-}1\ \text{-}1]_{SL}$ zone axis. We then selected an ultrathin area from Fig. 2A (marked by the square) for image processing, which was performed by correcting the effect of contrast transfer function (CTF) of the objective lens. The CTF-corrected image approximately corresponds to the projected electrostatic potential of the structure and is therefore directly interpretable. In the CTF-corrected image (middle panel, Fig. 2B), black dots with a diameter of 0.98 nm are observed with a rhombus packing structure. The size of the black dots nicely matches with the core diameter of $[Au_{25}(SR)_{18}]^-$ measured by X-ray crystallography[25, 35], which means each black dot represents one $[Au_{25}(SR)_{18}]^-$ NP. Moreover, this conclusion is supported by high-angle annular dark field scanning transmission electron microscopy (HAADF-STEM) (Supplementary Fig. 4). More intriguingly, regular dark spots are observed between neighbored NPs (arrowed in panel (i) of Fig. 2D), suggesting an unusual packing mode in the supercrystals (*vide infra*).

The near-atomic-resolution ULD-TEM images (Fig. 2B) provide a good opportunity to investigate the atomic structure of individual $[Au_{25}(p\text{-MBA})_{18}]^-$ NP in the supercrystals. Since there is currently no atom-level crystal structure of $[Au_{25}(p\text{-MBA})_{18}]^-$, we analyze the structure based on two potential candidates: the crystal structure of $[Au_{25}(PET)_{18}]^-$ with an organothiolate ligand (i.e., 2-phenylethanethiolate (PET), hereinafter referred to as **isomer 1)**[25, 35], and the recently theoretically suggested[36] and experimentally observed (in gas-phase)[37] topological isomer of $[Au_{25}(PET)_{18}]^-$ (**isomer 2**). Theoretical models of the corresponding structures with the *p*-MBA ligand, optimized by density functional theory (DFT) calculations (using the GPAW software[38]; see technical details in the section Theoretical Simulations and Supplementary Note 3 in Supplementary Information), are shown in Supplementary Fig. 5B and 5C. Simulated TEM images of the models projected from a number of spatial directions were compared to the TEM data shown in Fig. 2B by using the Complex Wavelet Structural Similarity (CW-SSIM) method (details in Supplementary Note 3)[39]. The results (Supplementary Fig. 6 and 7) clearly indicate that the **isomer 2** could be excluded from further consideration to build a 3D atomic model of the Au_{25} NP supercrystals.

Using the $[Au_{25}(SR)_{18}]^-$ **isomer 1** structure, we built models implying a crystal of packed single-particle chains or polymers of Au_{25} NPs, where the linkers in the polymers consist of four Au atoms with bridging thiolates (yielding the observed dark spots in TEM images in Fig. 2B, 2D, and Supplementary Fig. 5). Formation of such four-Au-atom linker bridged by thiolates can be envisioned by considering the atomic structure of **isomer 1** (Supplementary Fig. 5B). It should be reminded that in the "divide-and-protect" scheme[40], the chemical composition of $[Au_{25}(SR)_{18}]^-$ can be written as $[Au_{13}@(SR-[Au(I)-SR]_2)_6]^-$, where the icosahedral Au_{13} core is protected by six $SR-[Au(I)-SR]_2$ motifs (Supplementary Fig. 1C). Considering the dynamics of their surface structure, two neighbored Au_{25} NPs can react by opening one end of the $SR-[Au(I)-SR]_2$ motif via breaking a Au-S bond on a core-type SR (i.e., SR_C as illustrated in Supplementary Fig. 2A), followed by conjugation of the opened motifs from the neighbored NP. Therefore, two geometries for such polymeric linkers exist: one having a RS-SR bond in the middle in a symmetric configuration (i.e., $[SR-Au(I)]_2-RS-SR-[Au(I)-SR]_2$), and the other one having an asymmetric geometry with a $SR-[Au(I)-SR]_4$ linker connecting the NPs (Supplementary Fig. 8 and 9C). DFT calculations on periodic NP polymer model (Supplementary Fig. 9A and 9B) imply that the asymmetrically linked NP polymer is energetically preferred over the symmetric one, with the energy difference of ~0.75 eV per simulation unit cell. Optimal inter-particle distance was estimated to be about 2.3 – 2.5 nm (Supplementary Fig. 9B), which is close to the experimentally observed inter-particle distance along the $[1\ 1\ 0]_{SL}$ direction (i.e., 2.75 nm). An experimental support for this result was obtained from the Raman scattering spectrum of the hexagonal rod-like supercrystals, where no S-S bond fingerprints were observed in the regime of 400-550 cm^{-1} (Supplementary Fig. 9D)[41]. It should be noted that the seminal work on the reaction chemistry of Au NPs has theoretically proposed different structural models based on the inter-particle SR-[Au(I)-SR$_2$] linkers, accounting for the Au$_{25}$ NP dimers captured by ESI-MS in gas phase[42]. We excluded those models in our DFT calculations due to their obviously shorter inter-particle distance.

Using the asymmetric linkage model, a 3D model crystal from the packed $[Au_{25}(SR)_{18}]^-$ NP polymers was built upon orienting the polymeric chains along the $[1\ 1\ 0]_{SL}$ direction. This 3D model (bottom panel, Fig. 2B) shows a good match with the CTF-corrected ULD-TEM image in Fig. 2B (middle panel). Close to the atomic resolution, the zoom-in comparison of experimental, simulated TEM images, and structural model of $[Au_{25}(SR)_{18}]^-$ NP dimer further strengthens the model, as shown in panel (i)-(iii) of Fig. 2D. Typical views of the R-3m superlattice along different axis are illustrated in Fig. 2D (panel (iv)-(vi)). The accuracy of as-proposed packing model has also been verified by the ULD-TEM images taken along the $[1\ 0\ 0]_{SL}$ direction (Supplementary Fig. 10).

Besides the formation of 1D polymeric chain along the $[1\ 1\ 0]_{SL}$ direction, the formation of R-3m superlattice is also prompted by the close packing of as-formed polymeric chains of $[Au_{25}(p\text{-MBA})_{18}]^-$, which does not involve the formation of the SR-[Au(I)-SR$_4$] linkers. As shown in Supplementary Fig. 11A, the hexagonal arrangement of NPs in the $(1\ \text{-}1\ \text{-}1)_{SL}$ plane can be formed by the close packing of $[1\ 1\ 0]_{SL}$ oriented NP chains in an ABAB stacking manner. In this 2D stacking pattern, each NP has four nearest-neighbored NPs with a characteristic inter-particle distance of 1.76 nm, which is the shortest inter-particle distance observed in the R-3m superlattice (Supplementary Fig. 11A). Subsequent layer-by-layer stacking of as-described 2D particle planes along the $[1\ \text{-}1\ \text{-}1]_{SL}$ direction with an inter-particle distance of 1.76 nm produces a 3D R-3m superlattice. Of note, the preferential growth direction of hexagonal rod-like supercrystals is revealed by large-scale ULD-TEM analysis as [1 -1 2]$_{SL}$ (Supplementary Fig. 11B), corresponding well to the chain closest-packing direction in the superlattice (Supplementary Fig. 11A). It should be noted that the ordered 1D alignment of atomically precise metal NPs have been previously made possible via metallophilic interactions (e.g., Au-Au and Ag-Au-Ag)[8, 14, 43], disulfide bonds[44], and atomic/molecular linkers[45, 46]. However, the alignment of Au NPs by virtue of their surface dynamics has not yet been reported. The aforementioned packing mode of NP polymers also suggests different packing density within and vertical to the (1 -1 -1)$_{SL}$ plane, which allows us to exfoliate the as-formed hexagonal rod-like supercrystals into layered NP assemblies (Supplementary Fig. 12 and Supplementary Note 4), reminiscent of the layer-by-layer exfoliation of 2D materials such as graphene and black phosphorus[47, 48].

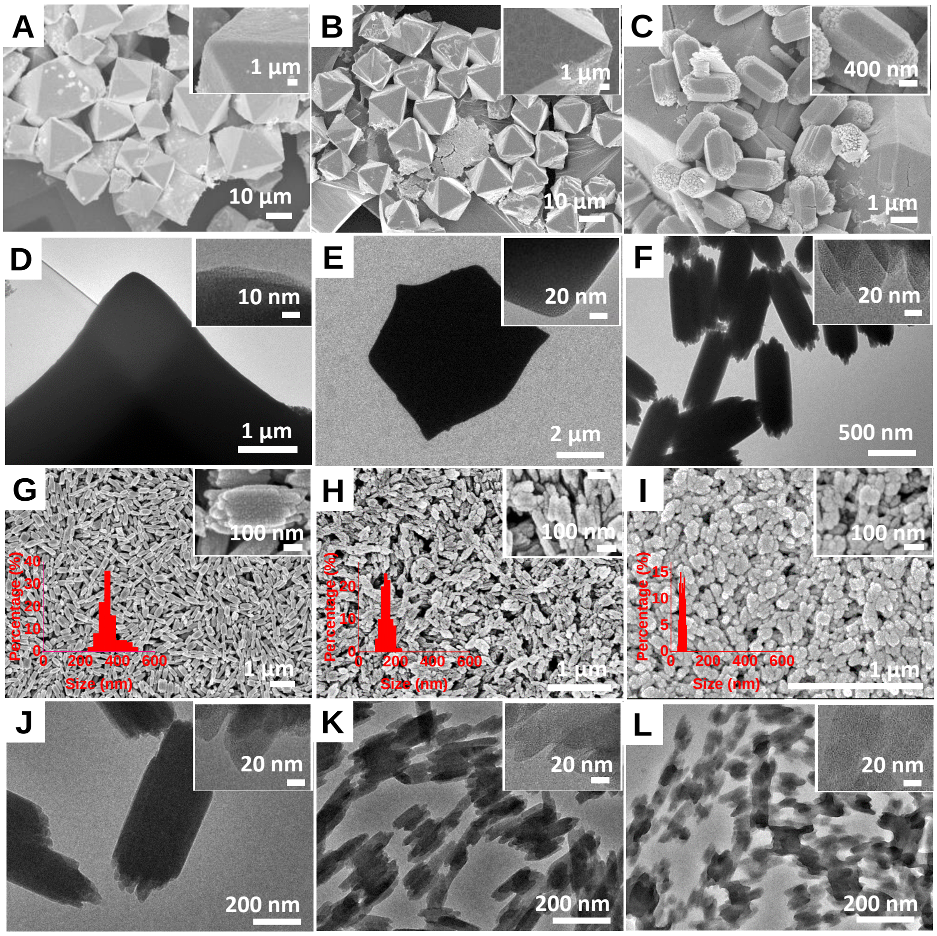


Figure 3. Engineering crystallization habit of [Au_{25}(*p*-MBA)$_{18}$]$^-$ nanoparticles by tetraalkylammonium cations. (A-C, G-I) FESEM and (D-F, J-L) TEM images of [Au_{25}(*p*-MBA)$_{18}$]$^-$ NPs crystallized at varied ratios of TEA$^+$ and Li$^+$, $R_{TEA/Li}$ = 0/4 (A, D), 1/3 (B, E), 2/2 (C, F), 3.5/0.5 (G, J), 3.75/0.25 (H, K), and 4/0 (I, L). Top insets are zoom-in views of the corresponding EM images, while bottom insets in (G-I) are longitudinal size histograms of the corresponding supercrystals. The ~1 nm dots observed in the insets of (D-F) and (J-L) indicate the supercrystals are packed by [Au_{25}(*p*-MBA)$_{18}$]$^-$ NPs.

Morphology engineering of the supercrystals via tetraalkylammonium cations. The use of structure-directing agent, TEA^+, is crucial for the formation of the hexagonal rod-like supercrystals. By changing the molar ratio of TEA^+ and alkaline metal co-cations (e.g., Li^+) while keeping other experimental conditions unchanged, the morphology of as-formed supercrystals could evolve from octahedron ($R_{TEA/Li}$ = 0/4, Fig. 3A and 3D; $R_{TEA/Li}$ = 1/3, Fig. 3B and 3E), via a mixture of octahedron and hexagonal rods ($R_{TEA/Li}$ = 2/2, Fig. 3C and 3F), to pure hexagonal rods ($R_{TEA/Li}$ = 3/1, Fig. 1B and 1E). This readily suggests that a threshold surface coverage of TEA^+ is required to trigger the formation of hexagonal rod-like supercrystals. Otherwise, the crystallization behavior of $[Au_{25}(p\text{-MBA})_{18}]^-$ is predictable by the typical hard sphere model in the Li^+-rich surroundings, where the entropy effect tends to pack NPs into the FCC supercrystals with an octahedral morphology (Supplementary Fig. 13, 14 and Supplementary Note 5)[32]. It should be noted that similar formation of octahedral or concave-octahedral supercrystals has been observed for Cs^+-deprotonated $[Ag_{44}(p\text{-MBA})_{30}]^{4-}$ NPs, which is governed by the entropy effects and electrostatic repulsion rather than the surface dynamics of NPs[32]. In contrast to modulating the crystalline phase, further increasing the dosage of TEA^+ can reduce the size of rod-like supercrystals while keeping their R-3m packing unchanged (Fig. 3G-3L, Supplementary Fig. 14, 15 and Supplementary Note 5).

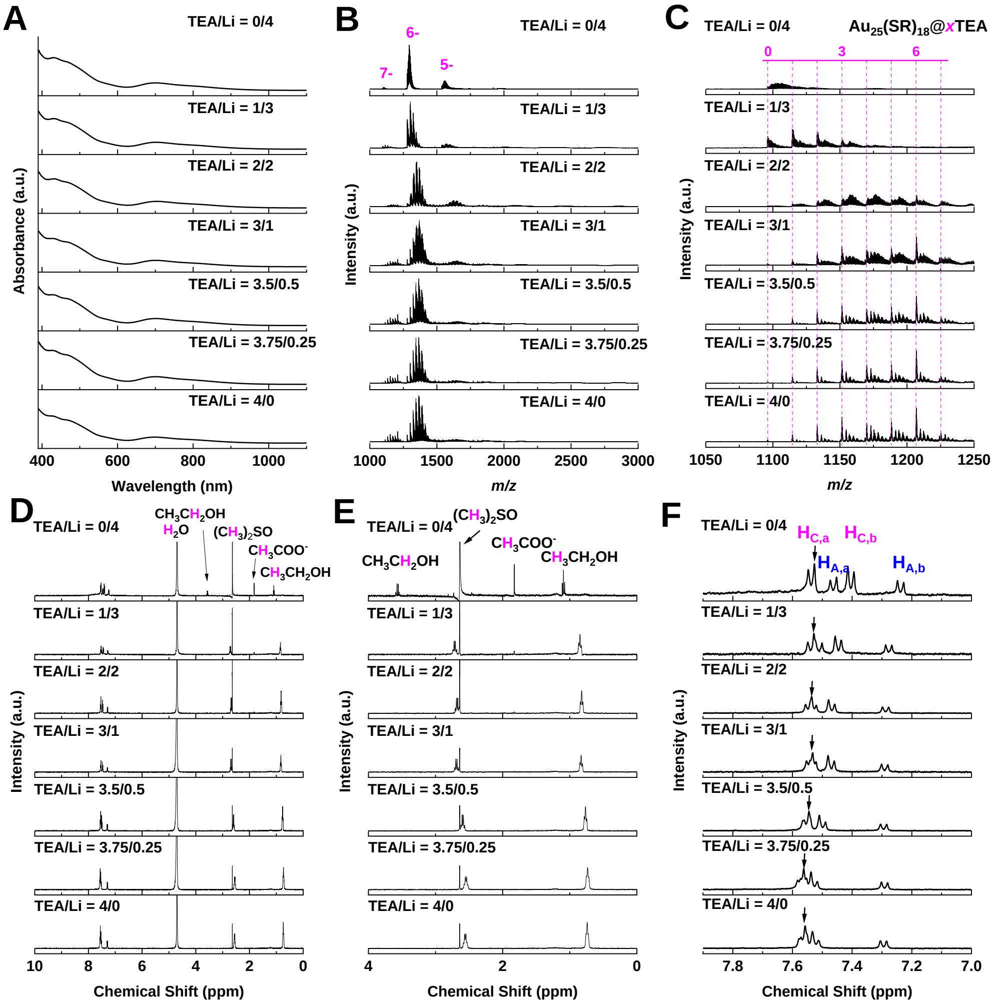


Figure 4. Molecule-level insights into building blocks of supercrystals. (A) UV-vis absorption, (B) wide-range and (C) zoom-in ESI-MS, and (D-F) ^{1}H-NMR spectra of $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs crystallized at varied ratios of TEA^+ and Li^+, $R_{TEA/Li}$ = 0/4, 1/3, 2/2, 3/1, 3.5/0.5, 3.75/0.25, and 4/0. The charge numbers of particle ions are labelled above the corresponding peaks in (B). (C) shows the zoom-in view of 7- cluster peaks in (B), and the dashed droplines in (C) are eye guides of x values in $Au_{25}(SR)_{18}@x$TEA. (E) and (F) exhibit the zoom-in views of the aliphatic and aromatic region of (D), respectively.

UV-vis absorption (Fig. 4A) and ESI-MS (Fig. 4B and 4C) spectra of the re-dissolved supercrystals confirm that the size of $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs in the supercrystals remains unchanged regardless of the $R_{TEA/Li}$ values. The typical recovery of $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs from octahedral supercrystals ($R_{TEA/Li}$ = 1/3) was measured to be 91.7%, which is similar to that of rod-like supercrystals (88.7% measured at $R_{TEA/Li}$ = 3/1). More intriguingly, the extensive formation of rod-like supercrystals coincides with the dominance of $Au_{25}(p\text{-MBA})_{18}$@6TEA species in the ESI-MS spectra (Fig. 4C). This again suggests the superior structural stability of $Au_{25}(p\text{-MBA})_{18}$@6TEA and its pivotal role in the formation of R-3m superlattice. The crucial role of $Au_{25}(p\text{-MBA})_{18}$@6TEA can also be verified by ^{1}H-NMR analysis (Fig. 4D). All the peaks identified in Fig. 4D can be attributed to *p*-MBA anchored on the surface of $[Au_{25}(p\text{-MBA})_{18}]^-$ (Fig. 4F), TEA^+, and residual solvent (i.e., DMSO and ethanol; Fig. 4E). Quantitative analysis based on the integral peak intensity suggests that the number of TEA^+ bonded to individual $[Au_{25}(p\text{-MBA})_{18}]^-$ NP increases with the increase of $R_{TEA/Li}$ and reaches a plateau of $Au_{25}(p\text{-MBA})_{18}$@6TEA at $R_{TEA/Li}$ = 3/1 or higher (Supplementary Fig. 16).

Molecular interaction between TEA^+ and $[Au_{25}(p\text{-MBA})_{18}]^-$. To reveal the preferential bonding sites of TEA^+ on the surface of $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs, we zoomed-in the aromatic region of the ^{1}H-NMR spectra. It can be seen from Fig. 4F that the four types of chemically distinct hydrogen of *p*-MBA ligands (see Supplementary Fig. 2A) exhibit different chemical shifts (δ) in response to increasing dosage of TEA^+. With the increase of $R_{TEA/Li}$, the resonances of $H_{A,a}$ (denoting H_a atom in SR_A), $H_{A,b}$, and $H_{C,b}$ exhibit significant downfield shifts, while those of $H_{C,a}$ only show marginal downfield shifts (arrows in Fig.

4F). The marginal downfield shifts of $H_{C,a}$ readily indicate that TEA^+ would preferentially bond to the $–COO^-$ groups of SR_A instead of SR_C. This ion-pairing-induced downfield shift of $H_{A,a}$ resonance is also supported by the gradual upfield shift of H resonances of TEA^+ with its elevating concentration (Fig. 4E). Therefore, the significant downfield shifts of $H_{A,b}$ and $H_{C,b}$ should be attributed to their close proximity to the Au(0) core of $[Au_{25}(p\text{-}MBA)_{18}]^-$ NPs, whose electronic structure is sensitive to TEA^+ bonding, as evidenced by the red-shifted absorption peak at ~690 nm (Supplementary Fig. 17). It has become increasingly known that the CH···π interactions are effective in maintaining the ligand arrangement patterns at the intra-particle level and assembly fashions at the inter-particle level of atomically precise metal NPs[3, 49]. Therefore, we hypothesized that the CH···π interaction is another important attribute (compared to the ion-pairing interaction) that can tether TEA^+ on the surface of $[Au_{25}(p\text{-}MBA)_{18}]^-$ NPs (panel (vii), Fig. 1A). This assertion is experimentally supported by 2D ^1H-^1H nuclear Overhauser effect spectroscopy (NOESY) analysis on the NP solution before crystallization (Supplementary Fig. 18), indicating that the alkyl chains of TEA^+ are spatially close to the phenyl rings of *p*-MBA ligands. Therefore, by combining the ion-pairing and CH···π (i.e., ion-pairing-*cum*-CH···π) interactions, six TEA^+ cations can selectively bind to the six SR_A in individual $[Au_{25}(p\text{-}MBA)_{18}]^-$ NP. Anchoring bulky TEA^+ on the surface of $[Au_{25}(p\text{-}MBA)_{18}]^-$ NPs can then induce strains in the SR-[Au(I)-SR]$_2$ protecting motifs, thereby enhancing their dynamics on the particle surface. The enhanced dynamics of the SR-[Au(I)-SR]$_2$ motifs promotes their partial detachment from the NP surface and further conjugation into the inter-particle SR-[Au(I)-SR]$_4$ linker, which can provide a unique mechanism for construction of polymeric NP chains in the self-assembly/crystallization scenarios. The enhanced surface dynamics induced by selective TEA$^+$ bonding is also experimentally supported by tandem mass spectrometry (MS/MS) analysis (Supplementary Fig. 19-69 and Supplementary Note 6).

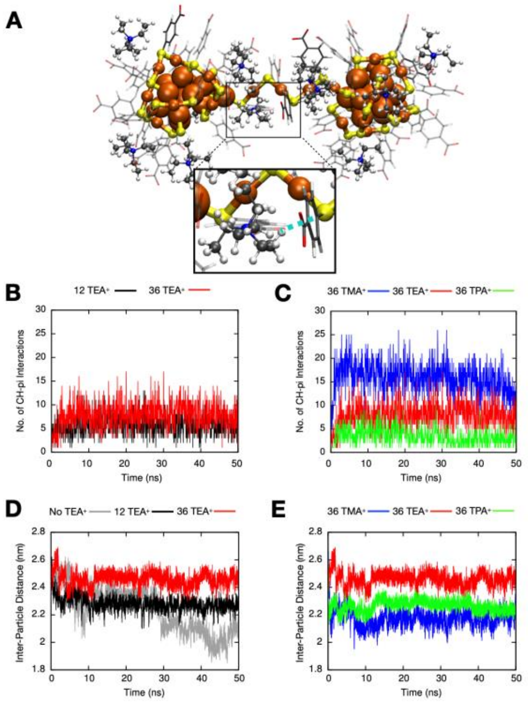


Figure 5. Effects of tetraalkylammonium cations on the stability of [Au$_{25}$(*p*-MBA)$_{18}$]$^-$ dimer. (A) Representative snapshot from 50 ns molecular dynamics (MD) trajectory of [Au$_{25}$(*p*-MBA)$_{18}$]$^-$ dimer formed by the assistance of TEA$^+$, showing the CH···π interactions between *p*-MBA$^-$ ligands and TEA$^+$. Color code and representation: large orange spheres, Au atoms in the Au$_{13}$ core; small orange spheres, Au atoms in the protecting motifs; thick sticks, *p*-MBA ligands; balls and sticks, TEA$^+$ cations; dashed cyan lines, CH···π interactions; yellow, S; gray, C; red, O; blue, N; light gray, H. (B, C) Total number of CH···π interactions and (D, E) inter-particle distances in [Au$_{25}$(*p*-MBA)$_{18}$]$^-$ dimer formed at varied dosages of TEA$^+$ (B, D) and varied tetraalkylammonium cations (C, E): tetramethylammonium (TMA$^+$), TEA$^+$, and tetrapropylammonium (TPA$^+$).

To reveal the molecular details of the interaction between the tetraalkylammonium cations and the linked $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs, we carried out extensive molecular dynamics (MD) simulations on short (dimeric and tetrameric) polymer models of $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs in the aqueous solution in the presence of tetramethylammonium (TMA$^+$)/TEA$^+$/tetrapropylammonium (TPA$^+$) using GROMACS software[50]. Fig. 5A shows a representative snapshot from 50 ns MD trajectory of $[Au_{25}(p\text{-MBA})_{18}]^-$ dimer with TEA$^+$ cations, which can visualize the CH···π interactions between *p*-MBA ligands and TEA$^+$. The total number of CH···π interactions in the simulation cell remains rather similar irrespective of the cation concentration (12 vs. 36 TEA$^+$ in the cell, Fig. 5B), but, remarkably, the TEA$^+$ has a clear effect on stabilizing the inter-particle distance (Fig. 5D, 5E and Supplementary Note 7). The inter-particle distance suitable for the growth of Au_{25} NP polymers was achieved in the presence of adequate TEA$^+$, despite TMA$^+$ gives rise to the greatest number of CH···π interactions (Fig. 5C). The optimal effects of TEA$^+$ on the formation of $[Au_{25}(p\text{-MBA})_{18}]^-$ NP polymer and thus rod-like supercrystals are supported by our attempts on crystallization of $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs in the presence of TMA$^+$, TPA$^+$, and tetrabutylammonium (TBA$^+$), which ubiquitously yield octahedral supercrystals (Supplementary Fig. 70-74). Finally, MD simulations of the longer tetrameric polymer model of $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs (Supplementary Fig. 75) show that the longer NP polymers are very flexible in solution, and the crystallization must proceed by gradually increasing (weak) interactions between neighboring polymeric chains when the solvent evaporates, eventually rigidifying the packed chains with a short inter-chain distance (1.76 nm; Supplementary Fig. 11A). Moreover, in order to evaluate the relative importance of the ion-pairing reactions and CH···π interactions in stabilizing the SR-[Au(I)-SR]$_4$ inter-particle linkers, we compared short-range Coulomb (as a proxy of ion-paring interactions) and Lennard-Jones (as a proxy of CH···π interactions) energies (Supplementary Fig. 76 and Supplementary Table 1). The consistently more negative value of the latter throughout the concerned simulation time (50 ns, with 12 or 36 TEA$^+$ in the simulation cell) unambiguously supports the dominant role of the CH···π interactions in stabilizing the as-mentioned inter-particle linkers.

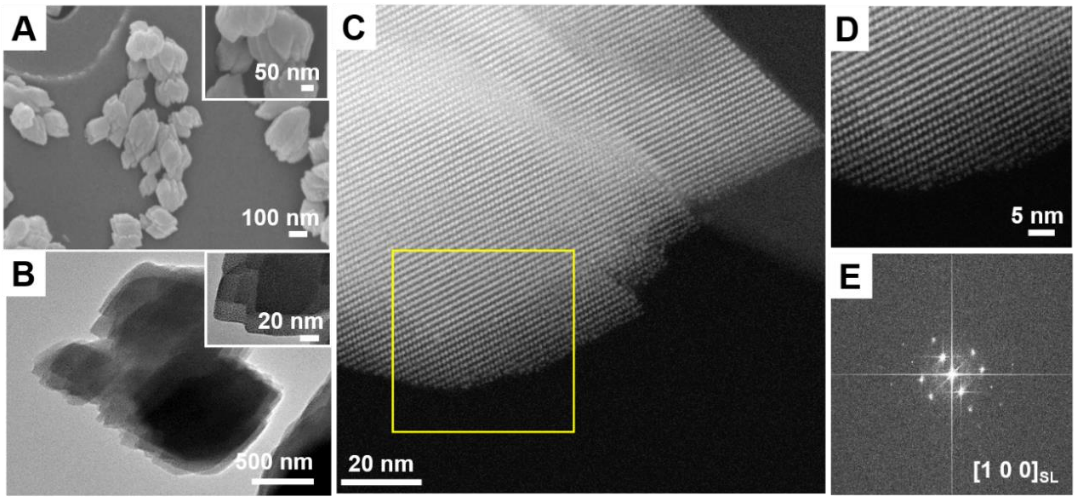


Figure 6. Shaping supercrystals into truncated rhombus flakes. (A) FESEM, (B) TEM, and (C) HAADF-STEM images viewed along [1 0 0]$_{SL}$ direction of [Au$_{25}$(*p*-MBA)$_{18}$]$^-$ supercrystals formed in the presence of TEA$^+$ and TMA$^+$ with a molar ratio of $R_{TEA/TMA}$ = 2/2. (D) Enlarged view and corresponding (E) fast Fourier transfer pattern of the squared area in (C).

Kinetically controlled evolution of truncated rhombus flake-like supercrystals. With a good understanding of the importance of TEA$^+$ in the formation of hexagonal rod-like supercrystals, we also studied the effects of co-cations on the crystallization of [Au_{25}(*p*-MBA)$_{18}$]$^-$ NPs. The data not only reveals a descending trend of supercrystal size with the increase of co-cation size, but also suggests that TMA^+ can competitively (against TEA^+) bond to [Au_{25}(*p*-MBA)$_{18}$]$^-$ during their crystallization process (Supplementary Fig. 77-85 and Supplementary Note 8). Our analysis of crystallization kinetics (Supplementary Fig. 86-92 and Supplementary Note 9) further suggests fast nucleation and slow growth kinetics for the hexagonal rod-like supercrystals ($R_{TEA/Li}$ = 3/1), and slow nucleation and fast growth kinetics for the octahedral supercrystals ($R_{TEA/Li}$ = 0/4).

Based on the above kinetics knowledge, we can further fine-tune the morphology of supercrystals. In the crystallization solution containing TEA^+ and TMA^+, increasing the ratio of competitive cation TMA^+ is expected to slow down the growth kinetics of rod-like supercrystals, allowing supercrystals to have more relaxing time to evolve into a well-defined shape. Therefore, we conducted the crystallization of [Au_{25}(*p*-MBA)$_{18}$]$^-$ with the molar ratio of TEA^+/TMA^+, $R_{TEA/TMA}$ = 2/2. The as-formed supercrystals exhibit a well-defined morphology of truncated rhombus flake (Fig. 6A and 6B). The HAADF-STEM image (Fig. 6C-6E) and P-XRD spectrum (Supplementary Fig. 93D) suggest that these flake-like supercrystals adopt R-3m superlattices, similar to the rod-like supercrystals formed at $R_{TEA/Li}$ = 3/1. UV-vis absorption (Supplementary Fig. 93A), ESI-MS (Supplementary Fig. 93B), and ^{1}H-NMR (Supplementary Fig. 93C) spectra of the re-dissolved supercrystals confirm that the building blocks for the flake-like supercrystals are size-unchanged [Au_{25}(*p*-MBA)$_{18}$]$^-$ NPs. More interestingly, reducing $R_{TEA/TMA}$ further to 1/3 can shape supercrystals into rhombus prisms (Supplementary Fig. 94I), while crystallization attempts using TPA^+ and TBA^+ as co-cations in a similar $R_{TEA/CoM}$ regime produce rod-like supercrystals exclusively (Supplementary Fig. 94), corroborating the delicate tunability of crystallization kinetics of $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs by TMA^+.

Discussion

In summary, we have developed a structure-directing agent assisted method for rational engineering of the symmetry, morphology, and size of Au NP supercrystals. This strategy utilizes the surface modulation capability of TEA^+ cation, where the unique ion-pairing-*cum*-CH···π interactions between *p*-MBA ligands and TEA^+ enhance the dynamic partial detachment of SR-[Au(I)-SR]$_2$ protecting motifs from the surface of $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs. The conjugation of such partially detached SR-[Au(I)-SR]$_2$ motifs between neighbored NPs gives rise to an SR-[Au(I)-SR]$_4$ inter-particle linker, aligning $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs into 1D polymeric chains. The close packing of as-formed NP polymeric chains produces hexagonal rod-like supercrystals. Such hexagonal rod-like supercrystals adopt a trigonal R-3m space group, which is in sharp contrast to the FCC octahedral supercrystals of $[Au_{25}(p\text{-MBA})_{18}]^-$ formed without TEA^+. The delicate control of crystallization kinetics by TMA^+ can further shape the supercrystals into truncated rhombus flakes and rhombus prisms. Extensive theoretical work has provided molecule-level understanding of the internal structure of the supercrystals, starting from image analysis of TEM data to an atom-level model of the linked $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs. This facilitates DFT calculations and MD simulations on the atomic structure, cation-ligand interactions, and the dynamic stabilization of the linked NP polymers, indicating TEA^+ as the optimal structure-directing cation. This work not only demonstrates the usefulness of tetraalkylammonium cations in tailoring the symmetry, morphology, and size of NP supercrystals, but also exemplifies the importance of molecule-level surface dynamics of Au NPs to their assembly and crystallization behavior.

Data availability

The authors declare that all the data supportive to the conclusion of this work are available in the paper and its Supplementary Information, and/or from the authors on a reasonable request basis.

References

1. Zhang, C., *et al.* A general approach to DNA-programmable atom equivalents. *Nat. Mater.* **12**, 741-746 (2013).

2. Shevchenko, E. V., Talapin, D. V., Kotov, N. A., O'Brien, S. & Murray, C. B. Structural diversity in binary nanoparticle superlattices. *Nature* **439**, 55-59 (2006).

3. Zeng, C., Chen, Y., Kirschbaum, K., Lambright, K. J. & Jin, R. Emergence of hierarchical structural complexities in nanoparticles and their assembly. *Science* **354**, 1580-1584 (2016).

4. Huang, R.-W., *et al.* Hypersensitive dual-function luminescence switching of a silver-chalcogenolate cluster-based metal–organic framework. *Nat. Chem.* **9**, 689-697 (2017).

5. Takano, S. & Tsukuda, T. Chemically modified gold/silver superatoms as artificial elements at nanoscale: design principles and synthesis challenges. *J. Am. Chem. Soc.* **143**, 1683-1698 (2021).

6. Kalsin, A. M., Fialkowski, M., Paszewski, M., Smoukov, S. K., Bishop, K. J. M. & Grzybowski, B. A. Electrostatic self-assembly of binary nanoparticle crystals with a diamond-like lattice. *Science* **312**, 420-424 (2006).

7. Boles, M. A., Engel, M. & Talapin, D. V. Self-assembly of colloidal nanocrystals: from intricate structures to functional materials. *Chem. Rev.* **116**, 11220-11289 (2016).

8. De Nardi, M., *et al.* Gold nanowired: a linear $(Au_{25})_n$ polymer from Au_{25} molecular clusters. *ACS Nano* **8**, 8505-8512 (2014).

9. Zhao, M., *et al.* Ambient chemical fixation of CO_2 using a robust Ag_{27} cluster-based two-dimensional metal–organic framework. *Angew. Chem. Int. Ed.* **59**, 20031-20036 (2020).

10. Ross, M. B., Ku, J. C., Vaccarezza, V. M., Schatz, G. C. & Mirkin, C. A. Nanoscale form dictates mesoscale function in plasmonic DNA–nanoparticle superlattices. *Nat. Nanotechnol.* **10**, 453-458 (2015).

11. Huang, J.-H., Wang, Z.-Y., Zang, S.-Q. & Mak, T. C. W. Spontaneous resolution of chiral multi-thiolate-protected Ag_{30} nanoclusters. *ACS Cent. Sci.* **6**, 1971-1976 (2020).

12. Chen, T., *et al.* Crystallization-induced emission enhancement: a novel fluorescent Au-Ag bimetallic nanocluster with precise atomic structure. *Sci. Adv.* **3**, e1700956 (2017).

13. Bodnarchuk, M. I., Kovalenko, M. V., Heiss, W. & Talapin, D. V. Energetic and entropic contributions to self-assembly of binary nanocrystal superlattices: temperature as the structure-directing factor. *J. Am. Chem. Soc.* **132**, 11967-11977 (2010).

14. Hossain, S., *et al.* Understanding and designing one-dimensional assemblies of ligand-protected metal nanoclusters. *Mater. Horiz.* **7**, 796-803 (2020).

15. Desireddy, A., *et al.* Ultrastable silver nanoparticles. *Nature* **501**, 399-402 (2013).

16. Tian, Y., Zhang, Y., Wang, T., Xin, H. L., Li, H. & Gang, O. Lattice engineering through nanoparticle–DNA frameworks. *Nat. Mater.* **15**, 654-661 (2016).

17. Auyeung, E., *et al.* DNA-mediated nanoparticle crystallization into Wulff polyhedra. *Nature* **505**, 73-77 (2014).

18. Cao, Y., *et al.* Reversible isomerization of metal nanoclusters induced by intermolecular interaction. *Chem* **7**, 2227-2244 (2021).

19. Zheng, K., Fung, V., Yuan, X., Jiang, D.-e. & Xie, J. Real time monitoring of the dynamic intracluster diffusion of single gold atoms into silver nanoclusters. *J. Am. Chem. Soc.* **141**, 18977-18983 (2019).

20. Salassa, G., Sels, A., Mancin, F. & Bürgi, T. Dynamic nature of thiolate monolayer in $Au_{25}(SR)_{18}$ nanoclusters. *ACS Nano* **11**, 12609-12614 (2017).

21. Barrabés, N., Zhang, B. & Bürgi, T. Racemization of chiral $Pd_2Au_{36}(SC_2H_4Ph)_{24}$: doping increases the flexibility of the cluster surface. *J. Am. Chem. Soc.* **136**, 14361-14364 (2014).

22. Jin, R., Zeng, C., Zhou, M. & Chen, Y. Atomically precise colloidal metal nanoclusters and nanoparticles: fundamentals and opportunities. *Chem. Rev.* **116**, 10346-10413 (2016).

23. Chakraborty, I. & Pradeep, T. Atomically precise clusters of noble metals: emerging link between atoms and nanoparticles. *Chem. Rev.* **117**, 8208-8271 (2017).

24. Li, Y., Zhou, M., Song, Y., Higaki, T., Wang, H. & Jin, R. Double-helical assembly of heterodimeric nanoclusters into supercrystals. *Nature* **594**, 380-384 (2021).

25. Zhu, M., Aikens, C. M., Hollander, F. J., Schatz, G. C. & Jin, R. Correlating the crystal structure of a thiol-protected Au_{25} cluster and optical properties. *J. Am. Chem. Soc.* **130**, 5883-5885 (2008).

26. Lopez-Acevedo, O., Kacprzak, K. A., Akola, J. & Häkkinen, H. Quantum size effects in ambient CO oxidation catalysed by ligand-protected gold clusters. *Nat. Chem.* **2**, 329-334 (2010).

27. Liao, L., *et al.* An unprecedented kernel growth mode and layer-number-odevity-dependent properties in gold nanoclusters. *Angew. Chem. Int. Ed.* **59**, 731-734 (2020).

28. Wu, Z., *et al.* Unraveling the impact of gold(I)–thiolate motifs on the aggregation-induced emission of gold nanoclusters. *Angew. Chem. Int. Ed.* **59**, 9934-9939 (2020).

29. Zheng, J., Nicovich, P. R. & Dickson, R. M. Highly fluorescent noble-metal quantum dots. *Annu. Rev. Phys. Chem.* **58**, 409-431 (2007).

30. Dolamic, I., Knoppe, S., Dass, A. & Bürgi, T. First enantioseparation and circular dichroism spectra of Au_{38} clusters protected by achiral ligands. *Nat. Commun.* **3**, 798 (2012).

31. Yao, Q., *et al.* Understanding seed-mediated growth of gold nanoclusters at molecular level. *Nat. Commun.* **8**, 927 (2017).

32. Yao, Q., *et al.* Counterion-assisted shaping of nanocluster supracrystals. *Angew. Chem. Int. Ed.* **54**, 184-189 (2015).

33. Zhang, D., *et al.* Atomic-resolution transmission electron microscopy of electron beam–sensitive crystalline materials. *Science* **359**, 675-679 (2018).

34. Liu, L., *et al.* Imaging defects and their evolution in a metal–organic framework at sub-unit-cell resolution. *Nat. Chem.* **11**, 622-628 (2019).

35. Heaven, M. W., Dass, A., White, P. S., Holt, K. M. & Murray, R. W. Crystal structure of the gold nanoparticle $[N(C_8H_{17})_4][Au_{25}(SCH_2CH_2Ph)_{18}]$. *J. Am. Chem. Soc.* **130**, 3754-3755 (2008).

36. Matus, M. F., Malola, S., Kinder Bonilla, E., Barngrover, B. M., Aikens, C. M. & Häkkinen, H. A topological isomer of the $Au_{25}(SR)_{18}^{-}$ nanocluster. *Chem. Commun.* **56**, 8087-8090 (2020).

37. Kalenius, E., Malola, S., Matus, M. F., Kazan, R., Bürgi, T. & Häkkinen, H. Experimental confirmation of a topological isomer of the ubiquitous $Au_{25}(SR)_{18}$ cluster in the gas phase. *J. Am. Chem. Soc.* **143**, 1273-1277 (2021).

38. Enkovaara, J., *et al.* Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method. *J. Phys. Condens. Matter.* **22**, 253202 (2010).

39. Wang, Z. & Simoncelli, E. P. Translation insensitive image similarity in complex wavelet domain. In: *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* **2**, 573-576 (2005).

40. Häkkinen, H., Walter, M. & Grönbeck, H. Divide and protect: capping gold nanoclusters with molecular gold−thiolate rings. *J. Phys. Chem. B* **110**, 9927-9931 (2006).

41. Price, R. C. & Whetten, R. L. Raman spectroscopy of benzenethiolates on nanometer-scale gold clusters. *J. Phys. Chem. B* **110**, 22166-22171 (2006).

42. Baksi, A., Chakraborty, P., Bhat, S., Natarajan, G. & Pradeep, T. $[Au_{25}(SR)_{18}]_2^{2-}$: a noble metal cluster dimer in the gas phase. *Chem. Commun.* **52**, 8397-8400 (2016).

43. Yuan, P., *et al.* Solvent-mediated assembly of atom-precise gold–silver nanoclusters to semiconducting one-dimensional materials. *Nat. Commun.* **11**, 2229 (2020).

44. Lahtinen, T., *et al.* Covalently linked multimers of gold nanoclusters $Au_{102}(p\text{-MBA})_{44}$ and $Au_{\sim 250}(p\text{-MBA})_n$. *Nanoscale* **8**, 18665-18674 (2016).

45. Liu, X., *et al.* $Ag_2Au_{50}(PET)_{36}$ nanocluster: dimeric assembly of $Au_{25}(PET)_{18}$ enabled by silver atoms. *Angew. Chem. Int. Ed.* **59**, 13941-13946 (2020).

46. Wen, Z.-R., Guan, Z.-J., Zhang, Y., Lin, Y.-M. & Wang, Q.-M. $[Au_7Ag_9(dppf)_3(CF_3CO_2)_7BF_4]_n$: a linear nanocluster polymer from molecular Au_7Ag_8 clusters covalently linked by silver atoms. *Chem. Commun.* **55**, 12992-12995 (2019).

47. Wang, C., *et al.* Monolayer atomic crystal molecular superlattices. *Nature* **555**, 231 (2018).

48. Bao, W., *et al.* Approaching the limits of transparency and conductivity in graphitic materials through lithium intercalation. *Nat. Commun.* **5**, 4224 (2014).

49. Huang, R.-W., *et al.* [$Cu_{81}(PhS)_{46}(t\mathrm{BuNH_2})_{10}(H)_{32}]^{3+}$ reveals the coexistence of large planar cores and hemispherical shells in high-nuclearity copper nanoclusters. *J. Am. Chem. Soc.* **142**, 8696-8705 (2020).

50. Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E. & Berendsen, H. J. C. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **26**, 1701-1718 (2005).

Acknowledgements

We acknowledge the financial support of the Ministry of Education, Singapore (Academic Research Grant R-279-000-580-112 and R-279-000-538-114) and National Natural Science Foundation of China (22071174). The theoretical work at University of Jyväskylä was supported by the Academy of Finland (grants 292352, 318905, 319208, and HH's Academy Professorship). The computations were done at the CSC computing center in Finland and in the FCCI - Finnish Grid and Cloud Infrastructure (persistent identifier urn:nbn:fi:research-infras-2016072533).

Author contributions

J.X. and Y.H. supervised the experimental work. J.X. and Q.Y. conceived the idea and designed the experiment. Q.Y., L.L. M.G. and H.X. carried out the experiments and characterizations. Z.W., T.C. and Y.C. contributed to data interpretation and theory development. H.H. supervised the theoretical and computational work. A.P. performed the image similarity analysis correlating Au_{25} NP models to the TEM data, S.M. performed DFT computations of the single NPs and linked NP models, M.F.M. performed the MD simulations of the linked NP models in aqueous solution. All authors contributed to manuscript writing.

Additional information

Supplementary information is available for this paper at http://

Methods

Synthesis of $[Au_{25}(p\text{-MBA})_{18}]^-$ Nanoparticles. $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs were prepared according to a reported protocol with some minor modifications[31]. Specifically, 10 mL of 50 mM *p*-MBA aqueous solution (in 150 mM NaOH) and 5 mL of 50 mM $HAuCl_4$ aqueous solution were sequentially added to 238.75 mL of ultrapure water, and the reaction mixture was stirred at 1,000 rpm for 5 min. After that, the pH value of the reaction mixture was adjusted to 10.5 by dropping 1 M NaOH aqueous solution. After stirring for another 30 min, a light-yellow solution of Au(I)-(*p*-MBA) complexes was formed. Subsequently, CO was bubbled into the reaction mixture at a flow rate of 100 mL per min for 2 min, to initiate the reduction of the Au(I)-(*p*-MBA) complexes. The reaction was allowed to proceed airtightly at room temperature (25 °C) under vigorous stirring (1,000 rpm) for 3 days. The reddish-brown solution obtained at the end of this procedure was collected as raw product.

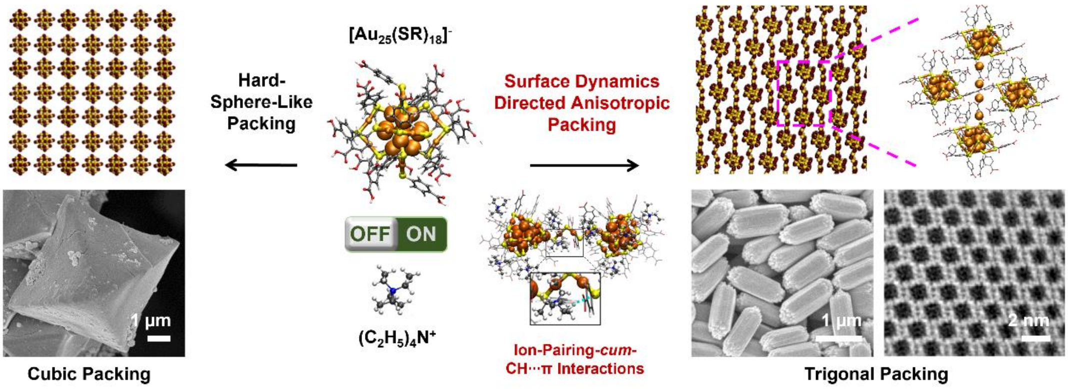
The raw product was first concentrated 10 times by rotary evaporation (water bath temperature 40 °C, cooling temperature 4 °C, and rotation rate 160 rpm). After that, ethanol (double the volume of the concentrated NP solution) was added, followed by centrifugation at 12,000 rpm for 5 min. The resultant pellet was washed twice with ethanol and re-dissolved in water for further characterization.

Cyclic Cation Exchange of $[Au_{25}(p\text{-MBA})_{18}]^-$ Nanoparticles. The freshly prepared $[Au_{25}(p\text{-MBA})_{18}]^-$ NP solution (raw product, 60 mL) was first concentrated 10 times by rotary evaporation (water bath temperature 40 °C, cooling temperature 4 °C, and rotation rate 160 rpm). After that, ethanol (double the volume of the concentrated NP solution) was added, followed by centrifugation at 12,000 rpm for 5 min. The precipitate was recovered and re-dissolved in 1 mL of aqueous solution of acetate salt (66.67 mM, pH = 11.85) of desired cations (e.g., Li^+, Na^+, K^+, Cs^+, TMA^+, TEA^+, and TBA^+). After incubating for 10 min under moderate stirring (600 rpm), five volumetric equivalent of ethanol was added, followed by centrifugation at 10,000 rpm for 5 min. The re-dissolution-centrifugation cycle was repeated two more times to complete the cation exchange process.

Synthesis of $[Au_{25}(p\text{-MBA})_{18}]^-$ Supercrystals. The supercrystals of $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs were grown in a dual-solvent system. Cation-exchanged $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs were re-dissolved in a mixture of DMSO/water (1/1, v/v) containing 33.33 mM (total cation concentration) of the designed cation or cation combination to form the crystallization solution (the target concentration of NPs was 0.50 mM). The crystallization solution was then placed in a vacuum oven at ~20 mbar and 50 °C to selectively remove water from the mixture. The evaporation usually lasted 1-2 days (depending on the cations used), and the solid supercrystals of $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs can be collected from the bottom of the crystallization tube at the end of the procedure. For a typical growth of the hexagonal rod-like supercrystals, a crystallization solution containing 33.33 mM (total cation concentration) of TEA^+/Li^+ (3/1, mol/mol) was subjected to vacuum treatment for 1 day. After that, the hexagonal rod-like supercrystals can be collected as dark red precipitates at the bottom of the crystallization tube.

Complete details about synthesis, characterization, and theoretical modeling of $[Au_{25}(p\text{-MBA})_{18}]^-$ supercrystals can be found in Supplementary Information.

TOC


$[Au_{25}(SR)_{18}]^-$
Hard-Sphere-Like Packing
Surface Dynamics Directed Anisotropic Packing
OFF
ON
$(C_2H_5)_4N^+$
Ion-Pairing-*cum*-CH···π Interactions
1 µm
1 µm
2 nm
Cubic Packing
Trigonal Packing

Supplementary Information for

Surface Dynamics Promoted Supercrystal Engineering of Atomically Precise Gold Nanoparticles

Qiaofeng Yao[1,7#], Lingmei Liu[2#], Sami Malola[3#], Meng Ge[4], Hongyi Xu[4], Zhennan Wu[1], Tiankai Chen[1], Yitao Cao[1], María Francisca Matus[3], Antti Pihlajamäki[3], Yu Han[5*], Hannu Häkkinen[3,6*] and Jianping Xie[1,7*]

[1] Department of Chemical and Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117585.

[2] Multi-scale Porous Materials Center, Institute of Advanced Interdisciplinary Studies & School of Chemistry and Chemical Engineering, Chongqing University, Chongqing, 400044 P. R. China.

[3] Department of Physics, Nanoscience Center, University of Jyväskylä, FI-40014 Jyväskylä, Finland.

[4] Department of Materials and Environmental Chemistry, Stockholm University, SE-106 91 Stockholm, Sweden.

[5] Advanced Membranes and Porous Materials Center, Physical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia.

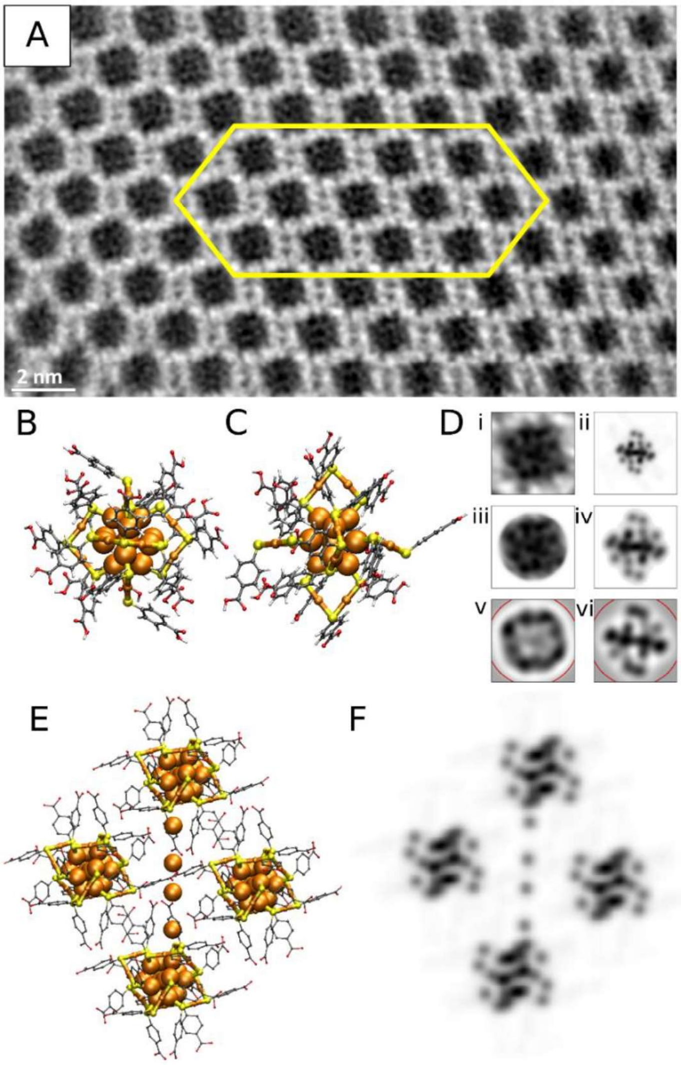
[6] Department of Chemistry, Nanoscience Center, University of Jyväskylä, FI-40014 Jyväskylä, Finland.

[7] Joint School of National University of Singapore and Tianjin University, International Campus of Tianjin University, Binhai New City, Fuzhou, China 350207.

* Corresponding author. Email: yu.han@kaust.edu.sa (Y.H.); hannu.j.hakkinen@jyu.fi (H.H.); chexiej@nus.edu.sg (J.X.)

# These authors contributed equally to this work.

Table of Contents



Supplementary Figure 5. (A) Contrast transfer function (CTF)-corrected TEM image from which 13 NPs within the yellow border were analyzed. (B, C) Atomic structures of **isomer 1** (B) and **2** (C) of $[Au_{25}(p\text{-MBA})_{18}]^-$. (D) Initial (i, ii), pre-processed (iii, iv), and transformed (v, vi) images from the experimental (i, iii, v) and computational (ii, iv, vi) TEM data, where red circles highlight the compared regions. (E) Proposed structure arrangement of $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs in the rod-like supercrystals. (F) TEM image calculated from the particle arrangement detailed in (E), which indicates that only metal atoms are seen in the TEM image. Color code: orange, Au; yellow, S; gray, C; light gray, H; red, O.

Image Similarity Measuring Scheme

In order to get further information about the structure of $[\mathrm{Au}_{25}(p\text{-MBA})_{18}]^-$ supercrystals and the orientation of NPs within the crystals, image recognition-based analysis method was devised in order to compare contrast transfer function (CTF)-corrected experimental TEM images and computationally generated TEM images of particle models (Supplementary Fig. 5A). The computational images were generated along 200 evenly spaced directions yielding statistical view about the orientations. Both **isomers 1** and **2** were used as candidates for the structure seen in the experimental TEM images. The isomers are visualized in Supplementary Fig. 5B and 5C. It should be noted that there is no reported experimental crystal structure of $[\mathrm{Au}_{25}(p\text{-MBA})_{18}]^-$. The published crystal structures of $[\mathrm{Au}_{25}(\mathrm{SR})_{18}]^-$ NPs are with organothiolates like PET[27, 28, 33], 1-naphthalenethiolate[36], and several alkyl thiolates[37-39]. Simulated TEM images were generated with simple scheme using projections and Gaussian functions[40]. The value for the pixel *i* is calculated as a summation over atoms.

$$\phi_i = \sum_{j=1}^{N_{atoms}} Z_j^{1.5} e^{|\mathbf{p}_j - \mathbf{p}_i|^2}$$

Here Z_j is the atomic number of atom *j* and $\mathbf{p}_i$ is the position of the pixel *i* in the images. Vector $\mathbf{p}_j$ is the position of an atom *j* projected to the two-dimensional plane on which TEM image is computed. The method emulates an imaginary electron beam hitting the plane of analysis, and the atomic number tells the weight of an atom resulting in the intensity, where heavy atoms are seen but light ones are not. The scaling is selected based on the previous studies on the dependence of image intensity on atomic number[41]. The spacing of pixels is chosen so that the experimental image and the computational one has the same resolution.

The basis of the image comparison is the Complex Wavelet Structural Similarity (CW-SSIM) method[42]. CW-SSIM can be seen as an improvement to the original Structural Similarity (SSIM) method, which uses sliding windows over two images to compare luminescence, contrast, and structure[43]. CW-SSIM, on the other hand, utilizes wavelet transformations to stabilize comparison making it less sensitive to small distortion and noise[42, 44]. Similarity value of CW-SSIM is formulated as

$$S(\mathrm{w}_x, \mathrm{w}_y) = \frac{2\sum_{i=1}^{N} |w_{x,i}||w_{y,i}| + K}{\sum_{i=1}^{N} |w_{x,i}|^2 + \sum_{i=1}^{N} |w_{y,i}|^2 + K} \cdot \frac{2|\sum_{i=1}^{N} w_{x,i}w_{y,i}^*| + K}{2\sum_{i=1}^{N} |w_{x,i}w_{y,i}^*| + K}$$

Here x and y denote two images to be compared. Vectors $\mathbf{w}_k$ contain the values of corresponding transformed images and $w_{k,i}$ are their vector elements. Here it is important to realize that the compared values should be taken from the same positions of the analyzed images. Order of indices is crucial. Parameter K is a small number used to stabilize the calculation. In this study it was set to 0.01. The similarity value varies between 0 and 1, where 1 indicates identical images and 0 points to completely different ones.

Originally Wang and Simoncelli used CW-SSIM to compare images transformed with Steerable Pyramid decomposition method[35], which relies on Fourier transform and series of high-pass, low-pass, and orientation filters[42, 45-47]. However, in this study a more straightforward approach was adopted. Wavelet transformations were done with discrete 2D convolutions using Rickers wavelets, also known as Mexican Hat Wavelets. This wavelet has been used in computer vision[48]. The basic convolution of two functions is formulated as

$$[f * g](t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)\mathrm{d}\tau$$

Moving into discrete 2D representation, the convolution can be written as

$$[f * g](\mathbf{t}) \approx \sum_{i=-N_h}^{N_h} \sum_{i=-N_v}^{N_v} f((i \cdot \Delta_h, j \cdot \Delta_v))g(\mathbf{t} - (i \cdot \Delta_h, j \cdot \Delta_v))\Delta_h \Delta_v$$

N_h and N_v are number of steps in horizontal and vertical direction, respectively. In practice they are determined by the dimensions of the computed material, *i.e.,* image size. The discrete step sizes are denoted by Δ_h and Δ_v. Vector **t** points to the center of the discrete wavelet $g(\cdot)$. Function $f(\cdot)$ presents the image to be transformed. Since image and wavelet are not infinite, the zero-padding approach was used. This means that outside the region defined by wavelet and image, functions will yield zero. The actual convolution is computed with ready implementation provided in SciPy[49]. The 2D Rickers wavelet is simply negative normalized second derivative of the Gaussian function written as[50]

$$g(\mathbf{r}) = \frac{2}{\sqrt{3\sigma}\pi^{1/4}} \left(1 - \left(\frac{|\mathbf{r}|}{\sigma}\right)^2\right) e^{\frac{-|\mathbf{r}|^2}{2\sigma^2}}$$

Here, **r** is a position vector and σ determines the width of the wavelet. Actual wavelet transformation is just a convolution, where $g(\cdot)$ would be a complex conjugate of the wavelet. In this study wavelet is real-valued and conjugation keeps it exactly the same.

In order to compare simulated TEM images with experimental ones, images have to be pre-processed. Thirteen NPs were chosen for comparison as highlighted in the Supplementary Fig. 5A. After cutting out an image of a single particle, the background was cut off spherically, edges were Gaussian smoothened, and images were centered as shown in panel (i) and (iii) of Supplementary Fig. 5D. Simulated image contains vague contributions from *p*-MBA ligands, which are not visible in experimental image. These ligand effects were erased by setting pixel values of 240 or higher (very light gray) to

255 (pure white). Then excess background was cut off and image was centered. The process is visualized in panel (ii) and (iv) of Supplementary Fig. 5D.

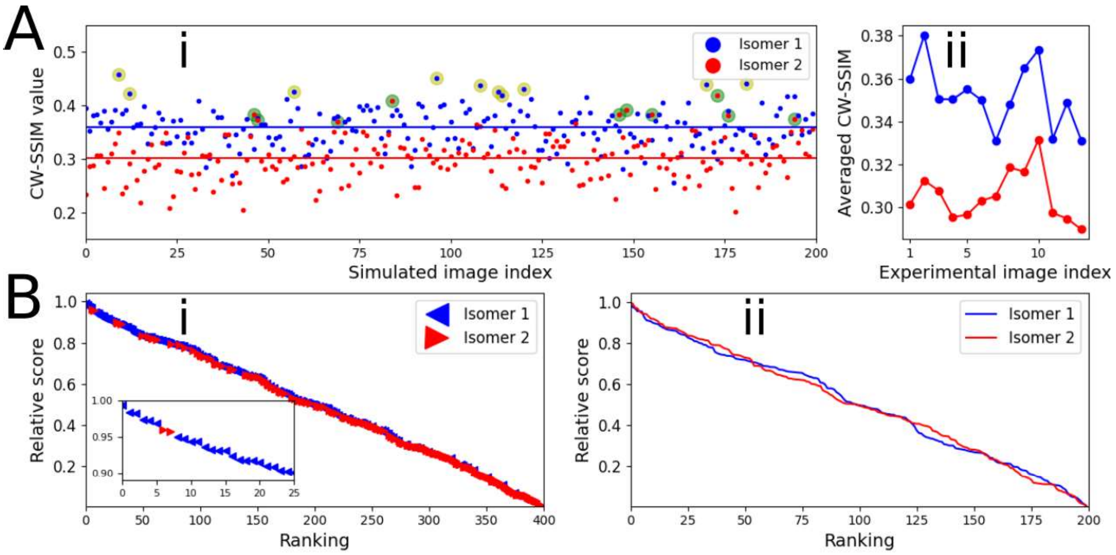
After all images were pre-processed, they went through comparison process. Convolutions were done with ten different widths of wavelets sampled evenly from 5 pixels to 30 pixels. Wavelets with different widths pick up different features. Small wavelets are sensitive for edges and large ones emphasize intensities. Transformed experimental images were rotated 360 degrees in 50 steps and for every rotation they were compared with simulated images, since the real particle orientation in the experimental data was unknown. CW-SSIM values were calculated with all wavelet widths and averaged. By this way, one gets a balanced measure over different features. During the comparison, corners of the transformed images were excluded from the comparison. One reason is that corners are not important, and another one is that they are the most vulnerable regions to small distortions during rotations. The compared regions and examples of wavelet transformations are visualized in panel (v) and (vi) of Supplementary Fig. 5D.

In order to get reasonable conclusions, results from different sets (different experimental images) have to be compared. As directly comparing similarity measures is not reliable, so-called scoring or voting approach was adapted. Every experimental image gives points to the simulated images according to their ranking. The worst image gets zero and the best one gets score of N-1, where N is the number of compared simulated images. Then scores given by experimental images are summed and divided with theoretical maximum score 13(N-1). This yields a relative score for every simulated image.

Structure and Orientation of $[Au_{25}(p\text{-MBA})_{18}]^-$ NPs in Supercrystals

Panel (i) of Supplementary Fig. 6A shows CW-SSIM values for all simulated images compared with a single experimental image. Every dot corresponds to the highest CW-SSIM value from the 50 rotations. Horizontal lines are showing average CW-SSIM values for corresponding isomers. From panel (ii) of Supplementary Fig. 6A, where all average CW-SSIM values are presented, **isomer 1** is yielding higher similarity values. This strongly indicates that the structures seen in the rod-like supercrystals, even with $SR\text{-}[Au(I)\text{-}SR]_4$ inter-particle linkers, are more closely related to **isomer 1** rather than **isomer 2**.

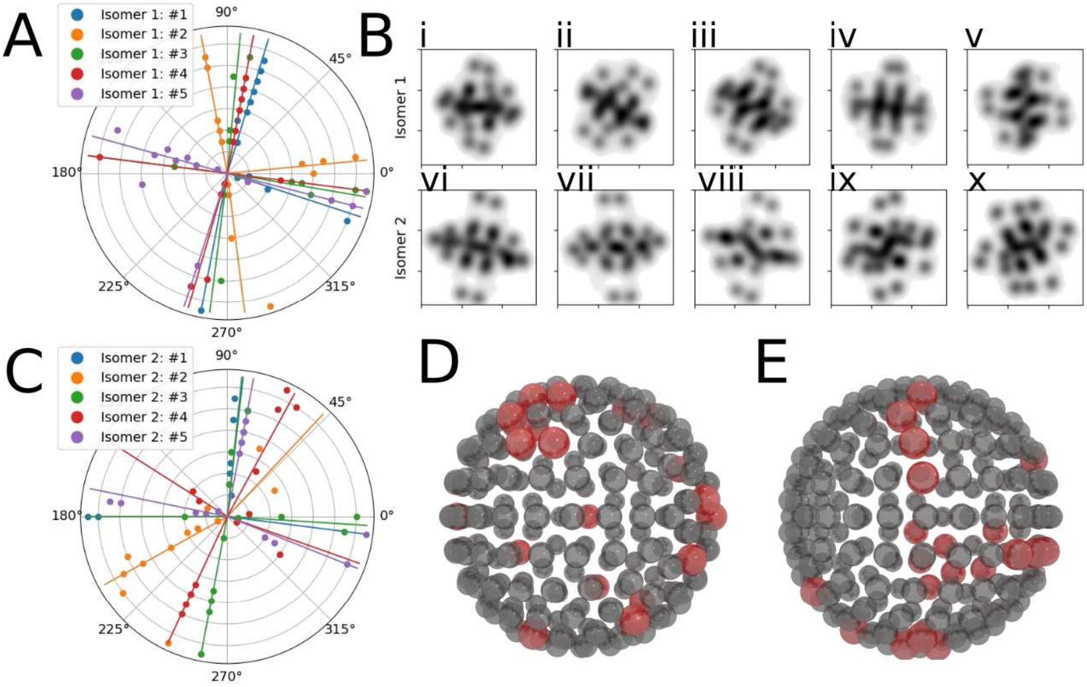
In panel (i) and (ii) of Supplementary Fig. 6B, the relative scores of simulated images are shown in a descending order. In panel (i), all simulated images from both isomers were included into scoring. The highest scoring images are clearly from the **isomer 1**. There are just a few images from **isomer 2** in the best 50 images and even fewer in the best 25 (inset of panel (i)). The panel (ii) shows the comparison where isomers are scored separately. Both images show consistent systematic behavior of the method. The images with highest CW-SSIM values are the same throughout the experimental images, leading to the relative scores close to 1. In panel (i)-(x) of Supplementary Fig. 7B, five highest scoring images of both isomers are shown in descending order of scoring. These images are chosen with scoring, where isomers are handled separately. They all have similar features such as relatively rectangular outline and dark linear region in the middle. The directions, along which the 20 highest scoring simulated images were computed, are visualized in Supplementary Fig. 7D and 7E. The symmetric behavior on how the points align can thus clearly be visualized by the red spheres, and the positioning of red spheres is similar to the way how the halves of baseball are stitched together.



Supplementary Figure 6. (A) CW-SSIM comparison results for a single experimental image (i) and average of 13 experimental images (ii). Horizontal lines in (i) show average similarity values for tested images. (B) Relative scores for simulated images. In (i), both isomers are taken into combined comparison, while isomers are scored separately in (ii); the inset of (i) depicts zoom-in view of the best-scored 25 images.

In order to get further insight into the orientation of the NPs in the supercrystals, the rotation angles yielding highest similarities were analyzed. The results for **isomer 1** and **2** are shown in Supplementary Fig. 7A and 7C. In the polar plot, the radial distance from the origin is referring to the index of a corresponding experimental image. As $[Au_{25}(SR)_{18}]^{-}$ is a symmetric NP, it is expected to see symmetric behavior also in the angles. For a single computational image, the best rotation angles are forming shape resembling a “+” sign. This is due to the shape of the images. Experimental and highest scoring simulated images are all showing some level of rectangular shape (Supplementary Fig. 7B). This makes algorithm to emphasize the orientations, where the "corners" are matching or the dark features in center are aligned. This shows that the algorithm is really picking up significant features during the comparison process.

With this basis of orientation information, one can build up more reliable models for DFT calculations and MD simulations.



Supplementary Figure 7. (A, C) Visualization of rotation angles, where the highest similarity values are yielded, with five highest scoring images for **isomer 1** (A) and **2** (C). (B) Five highest scoring images in descending order for **isomer 1** (i-v) and **2** (vi-x). (D, E) Visualization of directions (indicated by the red spheres) along which the best twenty simulated images are calculated for **isomer 1** (D) and **2** (E), respectively.

Structure of Isolated Linker Molecules

Supplementary Fig. 8A-8D show the optimized structures of linker Au(I)-SR polymer in the symmetric bonding mode, which is isolated [SR-Au(I)]$_2$-RS-SR-[Au(I)-SR]$_2$ (SR = *p*-MBA and SCH$_3$), with net charges of 0 and -2. Remarkable finding is that the polymer with *p*-MBA does not form a clear S-S bond, neither as neutral nor as -2 charged. In the neutral conformation, the S-S distance is larger than 3 Å. In contrary,

PII

MONTE CARLO SIMULATIONS OF AU$_{38}$(SCH$_3$)$_{24}$ NANOCLUSTER USING DISTANCE-BASED MACHINE LEARNING METHODS

by

Antti Pihlajamäki, Joonas Hämäläinen, Joakim Linja, Paavo Nieminen, Sami Malola, Tommi Kärkkäinen, and Hannu Häkkinen 2020

Journal of Physical Chemistry A, 124 (23), 4827–4836

THE JOURNAL OF PHYSICAL CHEMISTRY A

pubs.acs.org/JPCA Article

Monte Carlo Simulations of $Au_{38}(SCH_3)_{24}$ Nanocluster Using Distance-Based Machine Learning Methods

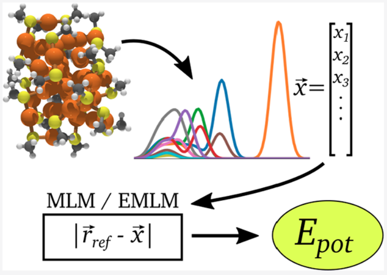
Published as part of The Journal of Physical Chemistry virtual special issue "Machine Learning in Physical Chemistry".

Antti Pihlajamäki, Joonas Hämäläinen, Joakim Linja, Paavo Nieminen, Sami Malola, Tommi Kärkkäinen, and Hannu Häkkinen*

Cite This: *J. Phys. Chem. A* 2020, 124, 4827−4836 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We present an implementation of distance-based machine learning (ML) methods to create a realistic atomistic interaction potential to be used in Monte Carlo simulations of thermal dynamics of thiolate (SR) protected gold nanoclusters. The ML potential is trained for $Au_{38}(SR)_{24}$ by using previously published, density functional theory (DFT) based, molecular dynamics (MD) simulation data on two experimentally characterized structural isomers of the cluster and validated against independent DFT MD simulations. This method opens a door to efficient probing of the configuration space for further investigations of thermal-dependent electronic and optical properties of $Au_{38}(SR)_{24}$. Our ML implementation strategy allows for generalization and accuracy control of distance-based ML models for complex nanostructures having several chemical elements and interactions of varying strength.



INTRODUCTION

Monolayer-protected clusters (MPCs) are small metal nanoparticles that have a metal core with size ranging from a few atoms to a few hundred atoms and a protecting surface layer of organic molecules such as thiols, phosphines, alkynyls, or carbenes.[1] MPCs are synthesized via wet chemistry by reducing metal salts in the presence of the protecting molecules. A variety of synthesis recipes and combination of metals and protecting molecules yields a rich chemistry and a large array of products in terms of size, shape, and composition of metal cores and the molecular overlayer. The wide range of synthetic parameters gives a unique possibility to study the fundamental structure−stability−property relations and to engineer the properties for applications such as catalysis, plasmonics, biosensing, and drug delivery.

The first crystallographically resolved MPCs were reported already over 50 years ago (such as the so-called undecagold Au_{11} cluster protected by phosphines[2]), and first advances in synthesis and structural characterization produced a series of mostly noble metal clusters protected by L-type (such as phosphine) and mixed L−X type (X being an electronegative ligand such as halide or thiolate) ligands. The largest such known cluster was the phosphine−halide protected Au_{39}, reported in 1992.[3]

Considerable steps forward were taken when Brust and coworkers[4] reported a synthesis that produced all-thiolate protected gold clusters for an average size of two nanometers. Several new chemical compositions of both organo-soluble and water-soluble clusters were reported soon after,[5−8] culminating to the breakthroughs of the first crystal structure of a large water-soluble all-thiol protected cluster, $Au_{102}(pMBA)_{44}$ (pMBA = *p*-mercaptobenzoic acid) by the Kornberg group in 2007[9] as well as the organo-soluble $Au_{25}(PET)_{18}$[−10−12] in 2008 and $Au_{38}(PET)_{24}$ (PET = phenyl ethyl thiolate)[13,14] clusters in 2008−2010. Up to date, atomic structures of at least 150 different compounds are crystallographically known, which facilitates detailed theoretical computations and dynamical simulations of the properties of MPCs and greatly helps to correlate structures to measured properties in experimental data.

Density functional theory (DFT) methods are the cornerstone for all computations that need to deal with details of the electronic structure, such as studies of optical absorption, optical excitation, fluorescence, and magnetism. However, while giving the most accurate and detailed information, DFT methods are also numerically the most demanding. DFT computations of some of the largest structurally known MPCs like the thiolate-protected Ag_{374}[15,16] have to deal with up to 13 000 valence electrons, and even a single-point DFT energy

Received: February 21, 2020
Revised: April 30, 2020
Published: May 15, 2020

https://dx.doi.org/10.1021/acs.jpca.0c01512
J. Phys. Chem. A 2020, 124, 4827−4836

calculation can take minutes and use hundreds or even thousands of CPU cores in a supercomputer. Force fields describing gold−thiolate MPCs have been developed to be used in molecular dynamics (MD) simulations, e.g., in the context of ReaxFF[17] and AMBER-GROMACS.[18] Effective but reliable methods to simulate the atomic dynamics of MPCs are needed, for instance, to study interactions of the clusters with the environment in the solvent phase, or with biomolecules and biological materials (viruses, proteins, lipid layers etc.).[19−21] However, developing such force fields may be time-consuming, system- or problem-specific, and suffer from poor transferability. Finally, understanding of nucleation processes in formation reactions of MPCs or reactions between two different MPCs are fundamental unsolved issues that are currently out of reach of any usable simulation method.

Machine learning (ML) and data-driven methods are emerging as a promising alternative to analyze structure−property correlations and make systematic predictions of physicochemical properties in materials science.[22,23] So far, ML has been applied to relatively small systems such as molecules with up to a few tens of atoms or systems where degrees of freedom can be limited such as binding of an atom to the surface.[24−28] A few homogeneous systems such as bulk water[29,30] or pure metal nanoparticles[31,32] have been studied as well. There has been very few studies of applying ML to MPCs. Recently deep neural networks and support vector machines were applied successfully to predict formation of MPCs in varying synthesis conditions.[33,34]

Systems with diverse chemical environments, such as MPCs, possess a large number of degrees of freedom, a range of chemical interactions of varying strength, and may require large training sets in order to cover the chemical space thoroughly enough. The most popular ML methods include neural networks, kernel ridge regression and Gaussian processes.[35] Neural networks have a great potential to learn very complicated data, because of their large number of parameters, weights, and network shapes to be adjusted during training. On the other hand, this flexibility also makes the method prone to overfitting. Kernel ridge regression and Gaussian processes are versatile tools, since one can define different kernel functions suiting a problem at hand. These kernels can easily transform the method to a complex one.

Here we demonstrate that even simple distance-based methods are applicable to complex systems such as MPCs. We use two methods, the so-called Minimal Learning Machine (MLM)[36] and the Extreme Minimal Learning Machine (EMLM)[37] and create a ML potential for a gold−thiolate $Au_{38}(SR)_{24}$ cluster. We utilize our previously published extensive DFT MD simulation data[38] based on two known structural isomers of $Au_{38}(PET)_{24}$[13,39] (Figure 1A,B) as the initial training set. We test the ML potential by performing Monte Carlo simulations up to 300 K and compare the cluster dynamics to that from DFT MD simulations. To our knowledge, this work reports the first successful demonstration of a ML potential for MPCs, suitable for fast explorations of the configurational space. An immediate application could be to combine the MLM/EMLM potential with the recently published algorithm[40] designed to build complete nanoparticle structures based only on information about the metal core, in order to accelerate structural discovery. Alternatively, the efficient probing of the configuration space at a desired temperature can be utilized to generate realistic cluster structures for further investigations of thermal-dependent electronic and optical properties of $Au_{38}(SR)_{24}$.

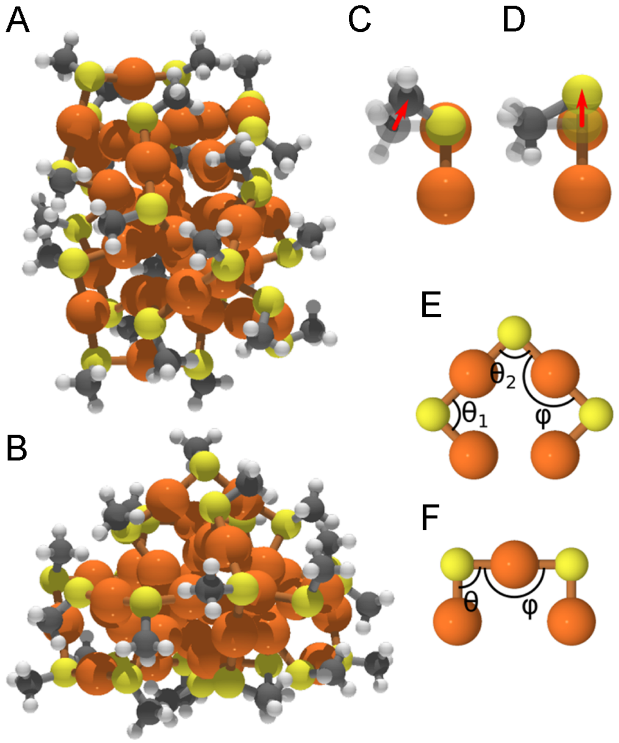


Figure 1. Initial structures of $Au_{38}(SCH_3)_{24}$ are visualized for Q and T isomers in parts A and B, respectively. While moving sulfur atoms and methyls the orientation of the S−C bond has to be preserved. Part C shows how alignment is preserved if methyl is moved. Part D shows the same when sulfur atom is moved. A long protecting unit is visualized in part E) and a short unit in part F. In parts E and F, methyls are omitted for the sake of clarity. Key: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen.

THEORETICAL METHODS

Here we discuss the necessary components of the development of the ML method to deal with dynamical simulations of thiolate protected gold nanoclusters. We introduce the used descriptor for the cluster structures, the general principles of the distance-based machine learning, and the Monte Carlo method to probe the configuration space.

Many-Body Tensor Representation. The Cartesian coordinates of atomic positions include the whole structural information about a single nanostructure, however one cannot use them to describe the system for a machine-learning method. If even a small rotation or translation is applied to the system, the coordinates would change, but physically, the situation is still the same. In order to overcome this problem, one needs to use a suitable structural descriptor, which are required to be invariant to translation, rotation, and permutation. Cartesian coordinates are not fulfilling any of these requirements. In addition to these requirements it is desirable that description would be continuous, unique in the sense of description−property correlation, and fast to be computed.[41] There have been several different approaches with a varying level of complexity to describe nanostructures for machine-learning methods. Frequently used descriptors in the field are atom-centered symmetry functions,[42] Coulomb matrices,[43] Ewald sum and sine matrices,[44] bag of bonds,[45] Zernike functions,[46,47] and smooth overlap of atomic positions (SOAP),[48] to name a few. These descriptors can be divided to local and global ones depending on whether they describe the environment around a single atom or the whole system as relationships between atoms. In this study, we used a global descriptor called many-body tensor representation (MBTR),[41] which is implemented in the DScribe package.[49] We chose to use a global descriptor instead of a local one, because it gives a straightforward and fast way to describe the system. It gives a single representation for a single configuration. A local descriptor, on the other hand, would have to be evaluated several times in order to describe every atom in the system. Since our system is quite large and has many different chemical interactions, a global descriptor such as the MBTR keeps the process simple and transparent.

The basic idea of the MBTR is based on a bag of bonds description. There, the system is first divided into the contributions of different element pairs and then described with pairwise distances between the atoms belonging to the elements of interest. Huo and Rupp used this as a starting point and formalized the basis of MBTR.[41] Afterward Jäger et al. simplified the theoretical presentation[50] and Himanen et al. implemented it into the DScribe package.[49] The backbone of the description is

$$f_k(x, z_1, z_2) = \sum_{i=1}^{N_{atoms,1}} \sum_{j=1}^{N_{atoms,2}} w_k(i, j) D(x, g_k(i, j)) \tag{1}$$

where

$$D(x, g) = (\sigma\sqrt{2\pi})^{-1} \exp\left(\frac{(x-g)^2}{2\sigma^2}\right) \tag{2}$$

In eq 1, summations are going through atoms with atomic (element) numbers of z_1 and z_2. Function $D(x, g)$ introduces broadening, which can be controlled by changing the parameter σ. Here x is sweeping variable, which probes the values produced by the function $g_k(i, j)$. Parameter k is the one defining the properties that are used to describe the system. In the theory, there is no limits for k; therefore, in principle, one can freely define a suitable property. Usually choices are $k = 1$ for atomic numbers, $k = 2$ for pairwise atomic distances (or the inverse of the distance), and $k = 3$ for angles formed by three different atoms. In this study, we chose to set $k = 2$ in order to use pairwise distances, therefore the weights are $w_2(i, j) = \exp(-dR_{ij})$ and the property measure is defined as $g_2(i, j) = R_{i,j}{}^{-1}$. Here d is a parameter, which is used to define the amount of weight for the contributions of atoms i and j if they are $R_{i,j}$ apart from each other.

As the name suggest, MBTR is a tensor with dimensions of $N_{elements} \times N_{elements} \times n_x$, when $k = 2$. $N_{elements}$ is the number of different elements in the system and n_x is the number of points that variable x can probe. Every element pair is described with their own summation but all pairs are using the same set of parameters. We list parameters as sets of {min, max, n_x, σ, d, cutoff}. First there are minimum and maximum values of the variable x. n_x is the number of points for x. As mentioned earlier, σ controls the broadening and d is used in weighting. DScribe package has also its own parameter to define cutoff. Only the values of the eq 1, which are greater than the cutoff, are used in summation for every value of x. This affects the sensitivity of the descriptor and also the speed of computations. A small cutoff value allows a large number of values to be included into the summation increasing the time spent for every element pair. On the other hand, a small cutoff would allow smaller changes in the structure to be visible in the description than a large cutoff. Using small cutoff values makes the descriptor sensitive but also very system-specific. Thus, there is a trade-off between accuracy and transferability.

Distance-Based Machine Learning Methods. *Minimal Learning Machine MLM.* Here we briefly introduce the theoretical background of the utilized distance-based machine-learning methods. First we go through the Minimal Learning Machine (MLM) formalized by de Souza Júnior et al.[36] In general, we assume that a set of N_d input points $X = \{\mathbf{x}_i\}_{i=1}^{N_d}$, $\mathbf{x}_i \in \mathbb{R}^n$, are given with the corresponding output points $Y = \{\mathbf{y}_i\}_{i=1}^{N_d}$, $\mathbf{y}_i \in \mathbb{R}^p$, to be predicted. We restrict here to univariate (nonlinear) regression problems. In supervised machine learning, one usually trains a model to map input points to certain output directly or through some kernel space. In that case the mapping f: $X \rightarrow Y$ between input and output spaces would be used to make the regression model as

$$\mathbf{Y} = f(\mathbf{X}) + \mathbf{E} \tag{3}$$

where $\mathbf{E}$ denotes residuals. MLM, on the other hand, determines the Euclidean distances between input and reference points and then uses these distances to construct a linear regression model to predict the Euclidean distances in the output space. These predicted distances with respect to the output space reference points form a multilateration problem from which the actual output is computed.

Reference points are defined as $\mathrm{M} = \{\mathbf{m}_k\}_{k=1}^{K}$ with $M \subseteq X$ and corresponding outputs are naturally $T = \{\mathbf{t}_k\}_{k=1}^{K}$ with $T \subseteq Y$. Then input space distances $d(\mathbf{x}_i, \mathbf{m}_k) = |\mathbf{x}_i - \mathbf{m}_k|$ are forming the distance matrix $\mathbf{D}_x \in \mathbb{R}^{N_d \times K}$. Analogously output space distances $\delta(\mathbf{y}_i, \mathbf{t}_k) = |\mathbf{y}_i - \mathbf{t}_k|$ are presented in a matrix $\Delta_y \in \mathbb{R}^{N_d \times K}$. In the notation, Greek letters are used for output space distances in order to distinguish them from input space notations. Next the mapping g is used to create regression model between distances in input and output spaces as

$$\Delta_y = g(\mathbf{D}_x) + \mathbf{E} \tag{4}$$

Next, de Souza Júnior et al. assume that the mapping g has a linear structure for each response. The model simplifies into a matrix product[36]

$$\Delta_y = \mathbf{D}_x \mathbf{B} + \mathbf{E} \tag{5}$$

In order to get the matrix $\mathbf{B}$ containing the coefficients for the K responses some approximations are needed. $\mathbf{B}$ is estimated from training data through minimizing the multivariate residual sum of squares. This provides a least-squares estimate of the matrix

$$\hat{\mathbf{B}} = (\mathbf{D}_x^T \mathbf{D}_x)^{-1} \mathbf{D}_x^T \Delta_y \tag{6}$$

Solving the $\hat{\mathbf{B}}$ corresponds to training of the model.

Now the last task is the multilateration problem in the output space. There is no single definite way to approach this problem, but many approaches can be applied.[51] The idea is to minimize the objective function of single output regression problem

$$J(y) = \sum_{k=1}^{K} ((y - t_k)^2 - (\mathbf{d}(\tilde{\mathbf{x}}, M)\hat{\mathbf{B}})_k^{\ 2})^2 \tag{7}$$

where $\mathbf{d}(\tilde{\mathbf{x}}, M) \in \mathbb{R}^{1\times K}$ is a vector containing distances between a new input $\tilde{\mathbf{x}}$ and all reference points M. The task is to find suitable output y, which minimizes the objective function. In our case we adopted cubic equation introduced by Mesquita et al.[52] The minimum or minima are found where the derivative equals zero. Differentiation yields

$$\begin{aligned} &Ky^3 - 3\sum_{k=1}^{K} t_k y^2 + \sum_{k=1}^{K} (3t_k^{\ 2} - (\mathbf{d}(\tilde{\mathbf{x}}, M)\hat{\mathbf{B}})_k^{\ 2})y \\ &\quad + \sum_{k=1}^{K} ((\mathbf{d}(\tilde{\mathbf{x}}, M)\hat{\mathbf{B}})_k^{\ 2} - t_k^{\ 3}) \\ &= 0. \end{aligned} \tag{8}$$

This can be thought as a cubic equation $ay^3 + by^2 + cy + d = 0$. From three possible roots, we choose the one that yields the smallest value of the objective function.

Extreme Minimal Learning Machine EMLM. Another distance-based machine-learning method, which was used in this study, is the Extreme Minimal Learning Machine (EMLM). The origin of the method lies in the so-called Extreme Learning Machine (ELM), which are single-layer perceptrons with special training and optimization methods.[53−57] When their training methods are combined with the Euclidean distance basis of MLMs, one gets EMLM.[37]

The first step is again to collect N_d input points into a matrix $\mathbf{X} \in \mathbb{R}^{n\times N_d}$. Corresponding outputs are in a matrix $\mathbf{Y} \in \mathbb{R}^{p\times N_d}$. Here n and p are the lengths of single input and output vectors $\mathbf{x}_i$ and $\mathbf{y}_i$. Input points $\mathbf{x}_i$ are first operated with a kernel function $\mathbf{h}(\cdot)$ forming new inputs $\mathbf{H} \in \mathbb{R}^{K\times N_d}$. Here $\mathbf{h}(\cdot)$ is a vector valued function, which is used to calculate the input vector in a kernel space. Due to the fact that we are using distance-based method, K is the number of reference points; therefore, the elements of $\mathbf{H}$ are defined as

$$\mathbf{H}_{i,j} = (\mathbf{h}(\mathbf{x}_j))_i = |\mathbf{m}_i - \mathbf{x}_j| \tag{9}$$

This is just the Euclidean distance between a reference point and an input point. We simplify the notation by writing $\mathbf{h}_j \equiv \mathbf{h}(\mathbf{x}_j)$. Now $\mathbf{h}_j \in \mathbb{R}^{K\times 1}$ and $\mathbf{H} \in \mathbb{R}^{K\times N_d}$. Then as Kärkkäinen states, the training of the model is done through regularized least-squares (RLS) optimization problem[37]

$$\min_{\mathbf{V}\in\mathbb{R}^{p\times K}} \frac{1}{2N_d}\sum_{i=1}^{N_d} |\mathbf{V}\mathbf{h}_i - \mathbf{y}_i|^2 + \frac{\alpha}{2K}\sum_{i=1}^{p}\sum_{j=1}^{K} |\mathbf{V}_{ij}|^2 \tag{10}$$

The parameter α is a small positive real number (square root of machine ϵ by default) used for regularization. $\mathbf{V}$ is a matrix containing the coefficients used for the actual regression and $\mathbf{V} \in \mathbb{R}^{p\times K}$. One could say, that $\mathbf{V}$ and reference points together form the actual machine-learning model. The minimum of the optimization problem lies on the zero point of the matrix derivative. The optimal solution $\mathbf{W} \equiv \mathbf{V}_{optimal}$ satisfies

$$\frac{1}{N_d}(\mathbf{WH} - \mathbf{Y})\mathbf{H}^T + \frac{\alpha}{K}\mathbf{I} = 0 \tag{11}$$

After getting the optimal solution for the RLS problem, one can use $\mathbf{W}$ to predict the output for a new arbitrary input $\tilde{\mathbf{x}}$. This is done as

$$f(\tilde{\mathbf{x}}) = \mathbf{Wh}(\tilde{\mathbf{x}}) \tag{12}$$

where $\mathbf{h}$ is the same vector valued kernel function as before. With input vector $\tilde{\mathbf{x}}$, it yields a $K \times 1$ vector. The elements of this vector are defined to be Euclidean distances as $|\mathbf{m}_i - \tilde{\mathbf{x}}|$.

We can see that the EMLM framework is fundamentally a kernel ridge regression with the Euclidean distance basis as a kernel. Because of the structural similarity to the linear radial basis function network, the EMLM model possesses the universal approximation capability.[58−60] MLM and EMLM have just one hyperparameter, which is the number of reference points. Overfitting is rarely an issue for distance based ML methods, therefore we can use all data points as reference points in training without worrying about overfitting.[37,61] There is no need for optimization of hyper- or metaparameters. This is a significant difference compared to the artificial neural networks, support vector machines, Gaussian processes or other popular ML methods. These methods require hyper- or metaparameter optimization through, for example, cross-validation.

Monte Carlo. We used Monte Carlo to simulate the dynamics of the $\mathrm{Au_{38}(SCH_3)_{24}}$ clusters with simplified methyl ligands. Clusters are divided to three different moving parts: gold, sulfur and methyl. Gold atoms are moved into a random direction according to the step size. Sulfur is moved in a similar fashion, but in order to preserve the orientation of sulfur−carbon bond, the methyl group is rotated making it to face the sulfur atom. The same principle is applied for the movement of the methyl groups. When methyl is moved according to the step size, the S−C bond orientation is preserved. In addition to this we allowed methyl group to rotate around the sulfur−carbon bond. The way how the alignment of sulfur−carbon bond is preserved is visualized in Figure 1, parts C and D. The stretching of carbon−hydrogen bond does not have a significant contribution to the total potential energy of the system; therefore, we decided to fix these bonds.

The acceptance of every move is decided according to the Metropolis question. The probability of the move to be accepted is defined as

$$P = \min\left\{1, \exp\left(\frac{-(E_{i+1} - E_i)}{k_B T}\right)\right\} \tag{13}$$

E_i is the potential energy of the ith configuration and E_{i+1} is the potential energy of the configuration after a proposed move. Going downhill in energy landscape is always permitted but going uphill is accepted with certain probability defined by the energy difference and simulation temperature T. In the exponent k_B is the Boltzmann constant. The step size of a single move is adjusted during the simulations so that the acceptance of the moves is between 40% and 60%. This step size is the same throughout the whole cluster and it is not affected by the type of the moved block. During a MC step, all moving parts are sampled randomly, and every one of them has an opportunity to move. This means that one MC step consists of 38 + 24 + 24 = 86 trial moves.

■ RESULTS AND DISCUSSION

Generating Training Data and Training the Models. The training data from the $\mathrm{Au_{38}(SCH_3)_{24}}$ clusters were

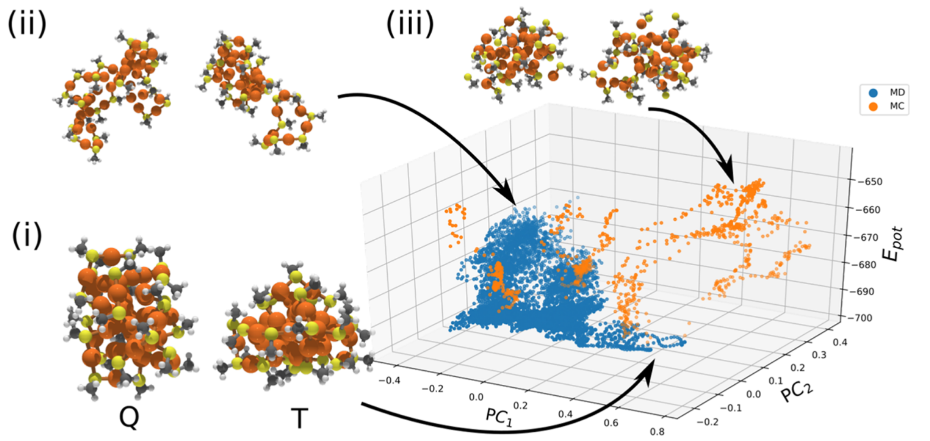


Figure 2. PCA visualization of MBTR descriptors of the training data. For the sake of clarity only 25% of the points are present in the graph. (i) the initial structures and (ii) high-temperature structures of the original MD simulations[38] (iii) snapshots from Monte Carlo simulations, where S−Au bonds have been broken. In parts ii and iii, left/right structures originate from Q/T isomers. Key: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen.

generated using density functional theory (DFT) run with GPAW code.[62,63] The major training data were published earlier by Juarez-Mosqueda et al.[38] In that work, Born−Oppenheimer *NVT* molecular dynamics simulations were run for the so-called Q[13] and T[39] isomers of $Au_{38}(SCH_3)_{24}$ at various temperatures between 400 and 1200 K. To be consistent with the training data we used same level of theory (the Perdew−Burke−Ernzerhof (PBE) exchange-correlation functional[64]). The DFT MD simulation trajectories of Juarez-Mosqueda et al.[38] contained 12413 configurations for the Q isomer and 12647 for the T isomer.

We used two different sets of MBTR parameters {min, max, n_x, σ, d, cutoff}. The first set was {0, 1.4, 100, 0.1, 0.5, 10^{-3}} and the second set was {0, 1.2, 100, 0.045, 0.8, 10^{-5}} (for a discussion on choosing the parameters, see the Supporting Information text and Figures S1 and S2). In the beginning, we trained MLM for the MBTR data corresponding to the first set of parameters. Minmax scaling was applied to the training data, so that descriptor values belonged to interval [0, 1]. As we mentioned earlier in the Theory section, overfitting is rarely an issue for MLM and EMLM. Therefore, we used the Full MLM and EMLM variants meaning that all data points were selected as reference points. We used MLM to predict potential energies during the Monte Carlo simulations in various simulation temperatures and with different starting structures taken from the training data. Monte Carlo frequently found the outer boundaries of the reference points pushing itself out of the working range of MLM. This resulted in erroneous potential energy values and nonphysical structures. In the Supporting Information text and Figure S3, we show that the MLM, which was trained only with the initial MD data,[38] is not able to handle configurations produced by the Monte Carlo. However, it can still find clear structure−energy correlation within the training data.

To cope with the erroneous behavior, we expanded the MLM training set including the MC-generated “unrealistic” configurations and their energies from DFT. The training set was expanded with 1580 new configurations for the Q and 2124 for the T isomer. After this we used the second set of MBTR parameters, which had improved descriptive possibilities (see Supporting Information). With the expanded training set and improved descriptor we trained both MLM and EMLM. In Figure 2, the principal component analysis (PCA) of the MBTR shows that the training set contains a large variety of configurations of both isomers spanning a large area of the feature space. Due to the fact that MLM/EMLM methods are using the Euclidean distances to measure the similarity of input point it is educative to visualize how the data points are arranged in the feature space.

Validation: Potential Energy MLM/EMLM vs DFT-MD. For validation, we created new independent DFT MD reference data sets for both Q and T isomers. For the Q isomer, we ran 2000 steps at 269 K, 2000 steps at 475 K, and 3653 steps at 795 K. For the T isomer we ran 2000 steps at 273 K and 2049 steps at 486 K. Potential energies were predicted for every configuration using both MLM and EMLM and compared to the actual DFT values from the MD run. The performance is seen in Figure 3. Generally, the predicted values correlate clearly with the DFT values, with the root-mean-

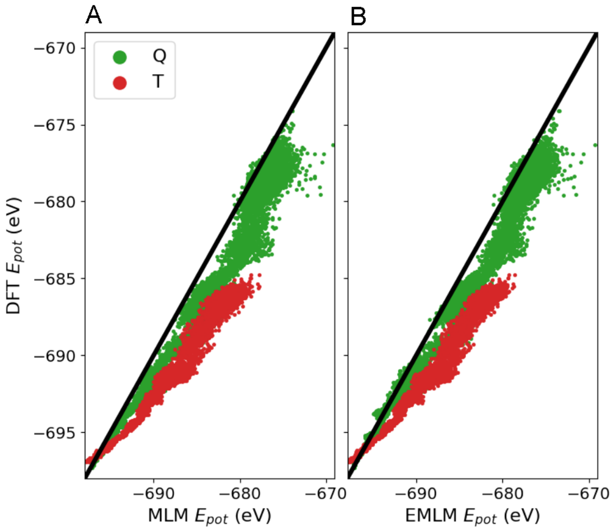


Figure 3. Correlation between the predicted potential energy from (A) MLM and (B) EMLM to the DFT energy from the MD calculations for Q and T isomers.

squared error (RMSE) being 2.98 eV for MLM and 2.67 eV for EMLM. The corresponding average relative errors are only 0.38% and 0.33%, respectively. The predicted energies are somewhat higher (less negative) than those from DFT. Our training set contains a lot of high energy configurations of $Au_{38}(SCH_3)_{24}$; therefore, the set might be biased. The visualization of PCA in Figure 4 indicates that the new MD simulations are rather far away from the points in the original training set. However, they are not outside of the working region of the MLM and EMLM like the first Monte Carlo simulations, which were used to expand the training set. This enables distance-based methods to predict well the potential energy values.

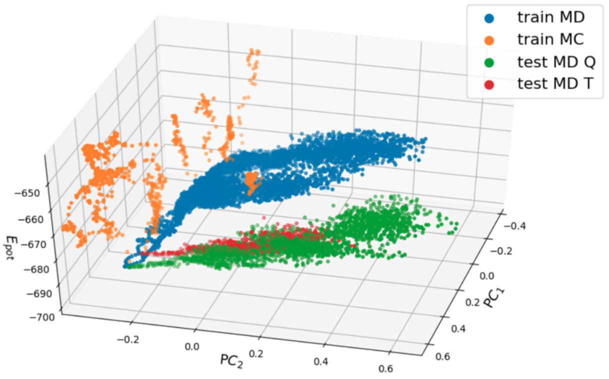


Figure 4. Visualization for PCA from training data and test MD data. Potential energies on *z* axis are computed with DFT. The graph is rotated with respect to Figure 2. In order to keep visualization clear, only 25% of the points are included.

MC Simulations with EMLM-Predicted Energies. As the most stringent test, we performed MC simulations of both Q and T isomers at temperatures of 200, 250, and 300 K, using the EMLM-predicted potential energy in the Metropolis criterion while advancing the dynamics. Typical simulations were run for 9000 to 10000 MC steps, one MC step consisting of 86 independent trial moves of the atoms (hence 86 EMLM energy evaluations per MC step). PCA of the runs at 300 K is shown in Figure 5A, indicating that the MC dynamics of both isomers are concentrated on a quite small region close to the *T* = 0 K local potential energy minimum, as expected for this rather low temperature. Figure 5(B) shows the evolution of the potential energy of both isomers at 300 K indicating that the potential energy of the Q isomer is consistently lower by about 1 eV than that of the T isomer. This result is consistent with the energetics known from DFT.

We analyzed the statistics of selected bond distances and bond angles for both isomers from the MC runs at 200, 250, and 300 K. The last 500 MC steps from each simulations were used for the analysis. Figure 6 shows the statistics for the nearest neighbor Au−Au bonds in the metal core as well as for the S−Au and S−C bonds, and compares them to the statistics obtained from DFT MD runs at 268 and 474 K for Q isomer and 272 and 486 K for T isomer. We observe that the EMLM-MC runs generally slightly overestimate the Au−Au bonds in both isomers as compared to DFT MD. The peaks of the distributions are at 2.862 Å (MC) and 2.805 Å (MD) for Q isomer, and 2.845 Å (MC) and 2.805 Å (MD) for T isomer. For S−Au and S−C bonds, EMLM-MC and DFT-MD produce very similar distributions both regarding the peak position and width. This analysis shows that the EMLM-MC runs indeed are able to simulate the bond dynamics of the atoms in the harmonic vibration regime.

Figure 7 shows the corresponding comparison between EMLM-MC and DFT-MD data for Au−S−Au and S−Au−S angles. In the crystal structures of these isomers the Au−S−Au angle is close to 90° and S−Au−S angle close to 170° (Figure 1). We observe that the maxima of Au−S−Au angles produced by EMLM-MC are slightly smaller than 90°, with a small side peak around 130° for the T isomer. We see a wider scatter in describing the S−Au−S angles in EMLM-MC as compared to DFT-MD, with the distributions having a maximum around 150° and tail extending close to 100°. MD simulations shows distributions peaked around 170°. We assign these slight discrepancies to the k2 description of the MBTR which does not take into account any angular information.

CONCLUSION

Distance-based machine-learning methods discussed in this study are conceptually straightforward and very simple to implement. We have shown here that they are suitable to

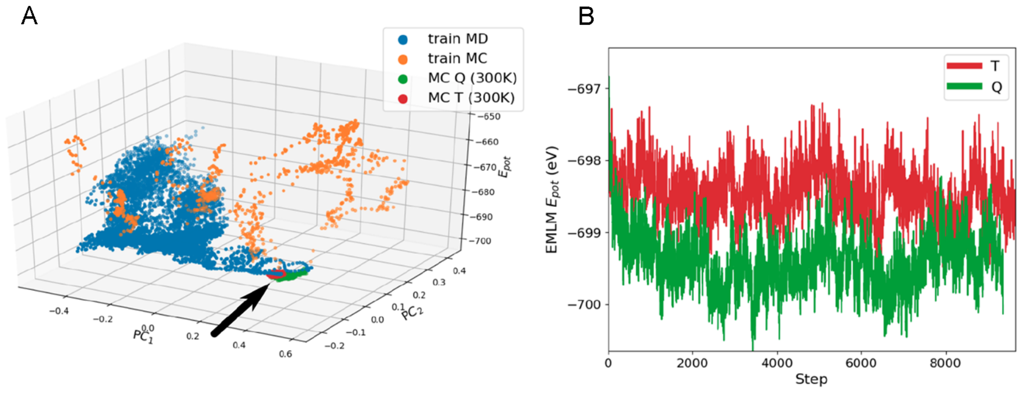


Figure 5. (A) Same as Figure 2, but including also the PCA analysis of EMLM MC runs at 300 K for isomers Q and T. The arrow highlights the region of the MC data. The analysis indicates that both of the isomers are vibrating close to their minima. Only 25% of the points are included into the Figure and PC1 values are multiplied with −1 to produce a comparable graph. (B) Evolution of potential energies of both isomers predicted by EMLM during MC.

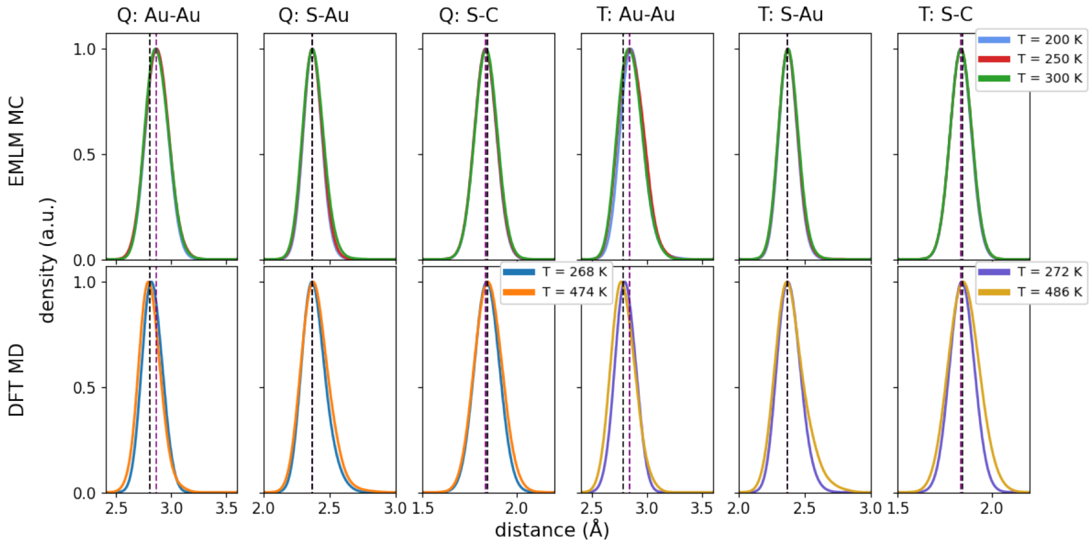


Figure 6. Top row: bond distance distributions from EMLM MC simulations at the indicted temperatures. Bottom row: the same data from DFT MD simulations at indicated temperatures. Labels on the top indicate the isomer and bond type. The vertical dashed lines indicate the average peak positions for every angle distribution in both MC and MD cases for every column (purple, MC; black, MD). Most of them are overlapping, and only black lines are visible. The statistics are summed from Gaussian-smoothened (σ = 0.05 Å) data points.

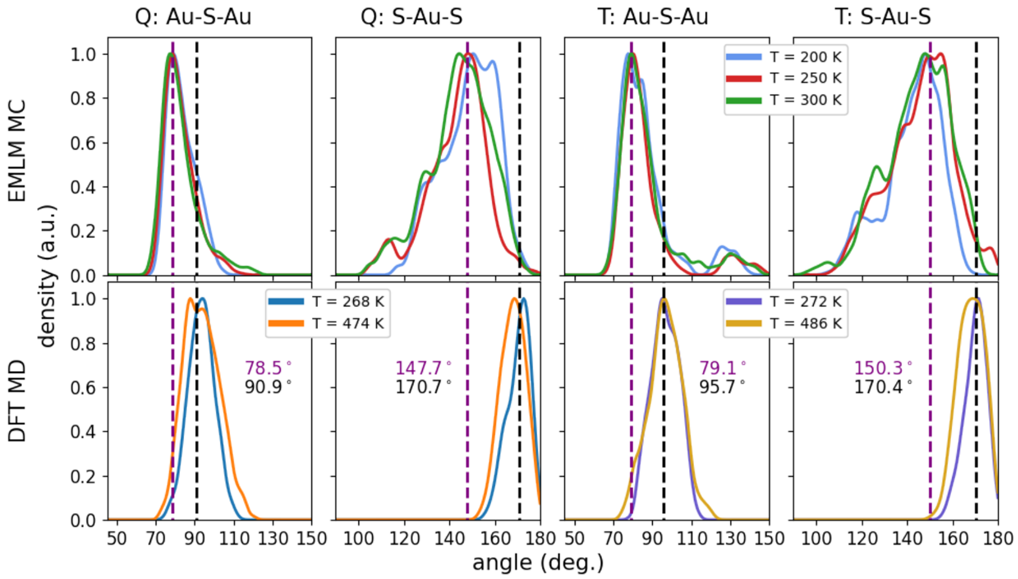


Figure 7. Top row: Selected bond angles distributions from EMLM MC simulations at the indicted temperatures. Bottom row: the same data from DFT MD simulations at indicated temperatures. Labels on the top indicate the isomer and type of the angle. The vertical dashed lines indicate the average peak positions for every angle distribution in both MC and MD cases for every column (purple, MC; black, MD). The colored numbers show the averages. The statistics are summed from Gaussian-smoothened (σ = 1.75°) data points.

simulate complex systems such as MPCs that have a number of chemical interactions with varying strength, while resulting in significantly reduced computational cost as compared to DFT. The CPU time to predict the energy by using MLM or EMLM with MBTR k2-level descriptors for the atomic structure is several magnitudes smaller than for DFT. For a comparison, MLM/EMLM energy predictions were run on a single core of Intel Xeon CPU E5−2680 v3 @ 2.50 GHz with 8GB memory. Computing MBTR k2 with our parameters took about 0.07 s for one atomic structure. Prediction of the potential energy using MBTR k2 took about 0.05 s with EMLM and 0.56 s with MLM. The order-of-magnitude difference between MLM and EMLM arises from the fact that the EMLM needs reference points only in the input space and is ready to give an output estimate from matrix and vector multiplication, while the MLM is predicting distances in the output space and solving a multilateration problem.

Excluding all angular information and using only pairwise distances to describe atomic structures with MBTR k2-level further helps to make these methods computationally light. The lack of angular information in MBTR k2 description does not mean, that our methods would not be able to reproduce reasonable bond angles. As shown in the Supporting Information, we could improve the description of the angles of protecting $RS(AuSR)_{n=1,2}$ units by tuning the parameters, although the MC simulations showed that the energy landscape produced by EMLM slightly differed from the one that DFT would yield.

Monte Carlo was shown to be an efficient strategy to study the energy landscape learned by MLM and EMLM. The method is not bound by any assumptions; therefore, it freely explores the feature space and gives useful insight of possible weaknesses of the machine-learning method. An important lesson learned in this work was that the initial MC simulations showed that our initial DFT-MD training set[38] was not extensive enough to train a comprehensive machine-learning method, since the DFT-MD produced atomistic configurations that were all “physical”. By enlarging the training data with the structures corresponding to the DFT energies of the “unphysical” configurations predicted by MLM/EMLM-MC back to the training data, we were able to teach the methods to avoid the unphysical regions of the configurational phase space.

Our future work involves further development of the models and descriptors for MPCs and other heterogeneous nanostructures. Here we used a global descriptor and predicted the potential energy of the system as a property of a whole system. Dividing the potential energy into atomic or molecular contributions creates in principle a way to get spatial insight into the energetics.[26] Fabrizio et al. have pointed out that it is reasonable to use global description when predicting global properties but it might cause size-dependence, which sometimes can be overcome with usage of local descriptions.[65] Our method is currently trained solely for $Au_{38}(SCH_3)_{24}$ with the goal to demonstrate that distance-based machine-learning methods can be used to handle complex systems such as MPCs. We aim to generalize the methods by including other MPCs (other metals and ligands) and other sizes of gold−thiolate clusters in the training set.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.0c01512.

Additional discussion on testing of the MBTR parameters (text and Figures S1 and S2) and performance of the MLM with the initial MD training data (Figure S3) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Hannu Häkkinen − *Department of Physics, Nanoscience Center and Department of Chemistry, Nanoscience Center, University of Jyväskylä, FI-40014 Jyväskylä, Finland*; orcid.org/0000-0002-8558-5436; Email: hannu.j.hakkinen@jyu.fi

Authors

Antti Pihlajamäki − *Department of Physics, Nanoscience Center, University of Jyväskylä, FI-40014 Jyväskylä, Finland*

Joonas Hämäläinen − *Faculty of Information Technology, University of Jyväskylä, FI-40014 Jyväskylä, Finland*; orcid.org/0000-0002-8466-9232

Joakim Linja − *Faculty of Information Technology, University of Jyväskylä, FI-40014 Jyväskylä, Finland*; orcid.org/0000-0003-2573-1240

Paavo Nieminen − *Faculty of Information Technology, University of Jyväskylä, FI-40014 Jyväskylä, Finland*

Sami Malola − *Department of Physics, Nanoscience Center, University of Jyväskylä, FI-40014 Jyväskylä, Finland*

Tommi Kärkkäinen − *Faculty of Information Technology, University of Jyväskylä, FI-40014 Jyväskylä, Finland*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpca.0c01512

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Academy of Finland through the AIPSE research program with Grant 315549 to H.H. and Grant 315550 to T.K., through the Universities Profiling Actions with Grant 311877 to T.K., and through H.H.’s Academy Professorship. MC simulations were done at the FGCI node at the University of Jyväskylä (persistent identifier: urn:nbn:fi:research-infras-2016072533) and the DFT MD simulations were run as part of PRACE Project 2018194723 at the Barcelona Supercomputing Centre.

■ REFERENCES

(1) Tsukuda, T.; Häkkinen, H. *Protected metal clusters: from fundamentals to applications*; Elsevier: Amsterdam, The Netherlands, 2015.

(2) McPartlin, M.; Mason, R.; Malatesta, L. Novel cluster complexes of gold(0)-gold(I). *J. Chem. Soc. D* **1969**, *0*, 334−334.

(3) Teo, B. K.; Shi, X.; Zhang, H. Pure gold cluster of 1:9:9:1:9:9:1 layered structure: a novel 39-metal-atom cluster [(Ph3P)14Au39Cl6] Cl2 with an interstitial gold atom in a hexagonal antiprismatic cage. *J. Am. Chem. Soc.* **1992**, *114*, 2743−2745.

(4) Brust, M.; Walker, M.; Bethell, D.; Schiffrin, D. J.; Whyman, R. Synthesis of thiol-derivatised gold nanoparticles in a two-phase liquid-liquid system. *J. Chem. Soc., Chem. Commun.* **1994**, *0*, 801−802.

(5) Schaaff, T. G.; Whetten, R. L. Giant gold-glutathione cluster compounds: intense optical activity in metal-based transitions. *J. Phys. Chem. B* **2000**, *104*, 2630−2641.

(6) Schaaff, T. G.; Shafigullin, M. N.; Khoury, J. T.; Vezmar, I.; Whetten, R. L. Properties of a ubiquitous 29 kDa Au:SR cluster compound. *J. Phys. Chem. B* **2001**, *105*, 8785−8796.

(7) Templeton, A. C.; Wuelfing, W. P.; Murray, R. W. Monolayer-protected cluster molecules. *Acc. Chem. Res.* **2000**, *33*, 27−36.

(8) Negishi, Y.; Nobusada, K.; Tsukuda, T. Glutathione-protected gold clusters revisited: bridging the gap between gold(I)-thiolate complexes and thiolate-protected gold nanocrystals. *J. Am. Chem. Soc.* **2005**, *127*, 5261−5270.

(9) Jadzinsky, P. D.; Calero, G.; Ackerson, C. J.; Bushnell, D. A.; Kornberg, R. D. Structure of a thiol monolayer-protected gold nanoparticle at 1.1 Åresolution. *Science* **2007**, *318*, 430−433.

(10) Heaven, M. W.; Dass, A.; White, P. S.; Holt, K. M.; Murray, R. W. Crystal structure of the gold nanoparticle $[N(C_8H_{17})_4]$-$[Au_{25}(SCH_2CH_2Ph)_{18}]$. *J. Am. Chem. Soc.* **2008**, *130*, 3754−3755.

(11) Zhu, M.; Aikens, C. M.; Hollander, F. J.; Schatz, G. C.; Jin, R. Correlating the crystal structure of a thiol-protected Au_{25} cluster and optical properties. *J. Am. Chem. Soc.* **2008**, *130*, 5883−5885.

(12) Akola, J.; Walter, M.; Whetten, R. L.; Häkkinen, H.; Grönbeck, H. On the structure of thiolate-protected Au_{25}. *J. Am. Chem. Soc.* **2008**, *130*, 3756−3757.

(13) Qian, H.; Eckenhoff, W. T.; Zhu, Y.; Pintauer, T.; Jin, R. Total structure determination of thiolate-protected Au38 nanoparticles. *J. Am. Chem. Soc.* **2010**, *132*, 8280−8281.

(14) Lopez-Acevedo, O.; Tsunoyama, H.; Tsukuda, T.; Häkkinen, H.; Aikens, C. M. Chirality and electronic structure of the thiolate-protected Au_{38} nanocluster. *J. Am. Chem. Soc.* **2010**, *132*, 8210−8218.

(15) Yang, H.; Wang, Y.; Chen, X.; Zhao, X.; Gu, L.; Huang, H.; Yan, J.; Xu, C.; Li, G.; Wu, J.; et al. Plasmonic twinned silver nanoparticles with molecular precision. *Nat. Commun.* **2016**, *7*, 12809.

(16) Zhou, Q.; Kaappa, S.; Malola, S.; Lu, H.; Guan, D.; Li, Y.; Wang, H.; Xie, Z.; Ma, Z.; Häkkinen, H.; et al. Real-space imaging with pattern recognition of a ligand-protected Ag_{374} nanocluster at sub-molecular resolution. *Nat. Commun.* **2018**, *9*, 2948.

(17) Bae, G.-T.; Aikens, C. M. Improved ReaxFF force field parameters for Au-S-C-H systems. *J. Phys. Chem. A* **2013**, *117*, 10438−10446.

(18) Pohjolainen, E.; Chen, X.; Malola, S.; Groenhof, G.; Häkkinen, H. A unified AMBER-compatible molecular mechanics force field for thiolate-protected gold nanoclusters. *J. Chem. Theory Comput.* **2016**, *12*, 1342−1350.

(19) Marjomäki, V.; Lahtinen, T.; Martikainen, M.; Koivisto, J.; Malola, S.; Salorinne, K.; Pettersson, M.; Häkkinen, H. Site-specific targeting of enterovirus capsid by functionalized monodisperse gold nanoclusters. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 1277−1281.

(20) Martikainen, M.; Salorinne, K.; Lahtinen, T.; Malola, S.; Permi, P.; Häkkinen, H.; Marjomäki, V. Hydrophobic pocket targeting probes for enteroviruses. *Nanoscale* **2015**, *7*, 17457−17467.

(21) Pohjolainen, E.; Malola, S.; Groenhof, G.; Häkkinen, H. Exploring strategies for labeling viruses with gold nanoclusters through non-equilibrium molecular dynamics simulations. *Bioconjugate Chem.* **2017**, *28*, 2327−2339.

(22) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **2019**, *5*, 83.

(23) Schleder, G. R.; Padilha, A. C. M.; Acosta, C. M.; Costa, M.; Fazzio, A. From DFT to machine learning: recent approaches to materials science-a review. *JPhys. Materials* **2019**, *2*, 032001.

(24) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(25) Sun, J.; Wu, J.; Song, T.; Hu, L.; Shan, K.; Chen, G. Alternative approach to chemical accuracy: A neural networks-based first-principles method for heat of formation of molecules made of H, C, N, O, F, S, and Cl. *J. Phys. Chem. A* **2014**, *118*, 9120−9131.

(26) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.

(27) Chen, X.; Jørgensen, M. S.; Li, J.; Hammer, B. Atomic energies from a convolutional neural network. *J. Chem. Theory Comput.* **2018**, *14*, 3933−3942.

(28) Kolsbjerg, E. L.; Peterson, A. A.; Hammer, B. Neural-network-enhanced evolutionary algorithm applied to supported metal nanoparticles. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2018**, *97*, 195424.

(29) Chan, H.; Cherukara, M. J.; Narayanan, B.; Loeffler, T. D.; Benmore, C.; Gray, S. K.; Sankaranarayanan, S. K. Machine learning coarse grained models for water. *Nat. Commun.* **2019**, *10*, 379.

(30) Patra, T. K.; Loeffler, T. D.; Chan, H.; Cherukara, M. J.; Narayanan, B.; Sankaranarayanan, S. K. R. S. A coarse-grained deep neural network model for liquid water. *Appl. Phys. Lett.* **2019**, *115*, 193101.

(31) Artrith, N.; Kolpak, A. M. Grand canonical molecular dynamics simulations of Cu-Au nanoalloys in thermal equilibrium using reactive ANN potentials. *Comput. Mater. Sci.* **2015**, *110*, 20−28.

(32) Zeni, C.; Rossi, K.; Glielmo, A.; Fekete, Á.; Gaston, N.; Baletto, F.; De Vita, A. Building machine learning force fields for nanoclusters. *J. Chem. Phys.* **2018**, *148*, 241739.

(33) Li, J.; Chen, T.; Lim, K.; Chen, L.; Khan, S. A.; Xie, J.; Wang, X. Deep learning accelerated gold nanocluster synthesis. *Adv. Intell. Syst.* **2019**, *1*, 1900029.

(34) Copp, S. M.; Swasey, S. M.; Gorovits, A.; Bogdanov, P.; Gwinn, E. G. General approach for machine learning-aided design of DNA-stabilized silver clusters. *Chem. Mater.* **2020**, *32*, 430−437.

(35) Murphy, K. P. *Machine learning: A probabilistic perspective*; MIT Press: Cambridge, MA, 2012.

(36) de Souza Júnior, A. H.; Corona, F.; Barreto, G. A.; Miche, Y.; Lendasse, A. Minimal Learning Machine: A novel supervised distance-based approach for regression and classification. *Neurocomputing* **2015**, *164*, 34−44.

(37) Kärkkäinen, T. Extreme minimal learning machine: Ridge regression with distance-based basis. *Neurocomputing* **2019**, *342*, 33−48.

(38) Juarez-Mosqueda, R.; Malola, S.; Häkkinen, H. Ab initio molecular dynamics studies of $Au_{38}(SR)_{24}$ isomers under heating. *Eur. Phys. J. D* **2019**, *73*, 62.

(39) Tian, S.; Li, Y.-Z.; Li, M.-B.; Yuan, J.; Yang, J.; Wu, Z.; Jin, R. Structural isomerism in gold nanoparticles revealed by x-ray crystallography. *Nat. Commun.* **2015**, *6*, 8667.

(40) Malola, S.; Nieminen, P.; Pihlajamäki, A.; Hämäläinen, J.; Kärkkäinen, T.; Häkkinen, H. A method for structure prediction of metal-ligand interfaces of hybrid nanoparticles. *Nat. Commun.* **2019**, *10*, 3973.

(41) Huo, H.; Rupp, M. Unified representation of molecules and crystals for machine learning. *arXiv* 2017, 1704.06439v3 [physics.chem-ph].

(42) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106.

(43) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(44) Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **2015**, *115*, 1094−1101.

(45) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.

(46) Canterakis, N. 3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition. *11th Scandinavian Conference on Image Analysis*; 1999; pp 85−93.

(47) Novotni, M.; Klein, R. Shape retrieval using 3D Zernike descriptors. *Computer-Aided Design* **2004**, *36*, 1047−1062.

(48) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.

(49) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **2020**, *247*, 106949.

(50) Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Himanen, L.; Foster, A. S. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Comput. Mater.* **2018**, *4*, 37.

(51) Navidi, W.; Murphy, W. S., Jr.; Hereman, W. Statistical methods in surveying by trilateration. *Comput. Stat. Data Anal.* **1998**, *27*, 209−227.

(52) Mesquita, D. P. P.; Gomes, J. P. P.; Souza Junior, A. H. Ensemble of efficient minimal learning machines for classification and regression. *Neural Process Lett.* **2017**, *46*, 751−766.

(53) Huang, G.-B.; Zhu, Q.-Y.; Siew, C.-K. Extreme learning machine: A new learning scheme of feedforward neural networks. *Proc. IEEEInt. Joint Conf. Neural Netw.* **2004**, 985−990.

(54) Huang, G.-B.; Zhu, Q.-Y.; Siew, C.-K. Extreme learning machine: Theory and applications. *Neurocomputing* **2006**, *70*, 489−501.

(55) Huang, G.-B.; Zhou, H.; Ding, X.; Zhang, R. Extreme learning machine for regression and multiclass classification. *IEEE Trans. Syst., Man, Cybern. B, Cybern.* **2012**, *42*, 513−529.

(56) Cambria, E.; Huang, G.-B.; Kasun, L. L. C.; Zhou, H.; Vong, C. M.; Lin, J.; Yin, J.; Cai, Z.; Liu, Q.; Li, K.; et al. Extreme learning machines [trends & controversies]. *IEEE Intelligent Systems* **2013**, *28*, 30−59.

(57) Akusok, A.; Björk, K.-M.; Miche, Y.; Lendasse, A. High-performance extreme learning machines: A complete toolbox for big data applications. *IEEE Access* **2015**, *3*, 1011−1025.

(58) Poggio, T.; Girosi, F. Networks for approximation and learning. *Proc. IEEE* **1990**, *78*, 1481−1497.

(59) Park, J.; Sandberg, I. W. Universal approximation using radial-basis-function networks. *Neural Comput* **1991**, *3*, 246−257.

(60) Liao, Y.; Fang, S.-C.; Nuttle, H. L. Relaxed conditions for radial-basis function networks to be universal approximators. *Neural Networks* **2003**, *16*, 1019−1028.

(61) Hämäläinen, J.; Alencar, A. S. C.; Kärkkäinen, T.; Mattos, C. L. C.; Souza Júnior, A. H.; Gomes, J. P. P. Minimal Learning Machine: Theoretical results and clustering-based reference point selection. *arXiv* 2019, 1909.09978v1 [cs.LG].

(62) Enkovaara, J.; Rostgaard, C.; Mortensen, J. J.; Chen, J.; Dułak, M.; Ferrighi, L.; Gavnholt, J.; Glinsvad, C.; Haikola, V.; Hansen, H. A.; et al. Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method. *J. Phys.: Condens. Matter* **2010**, *22*, 253202.

(63) Mortensen, J. J.; Hansen, L. B.; Jacobsen, K. W. Real-space grid implementation of the projector augmented wave method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2005**, *71*, 035109.

(64) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.

(65) Fabrizio, A.; Grisafi, A.; Meyer, B.; Ceriotti, M.; Corminboeuf, C. Electron density learning of non-covalent systems. *Chem. Sci.* **2019**, *10*, 9424−9432.

Supporting Information for: Monte Carlo Simulations of $Au_{38}(SCH_3)_{24}$ Nanocluster Using Distance-Based Machine Learning Methods

Antti Pihlajamäki,[†] Joonas Hämäläinen,[‡] Joakim Linja,[‡] Paavo Nieminen,[‡] Sami Malola,[†] Tommi Kärkkäinen,[‡] and Hannu Häkkinen[*,†,¶]

[†]*Department of Physics, Nanoscience Center, University of Jyväskylä, FI-40014 Jyväskylä, Finland*

[‡]*Faculty of Information Technology, University of Jyväskylä, FI-40014 Jyväskylä, Finland*

[¶]*Department of Chemistry, Nanoscience Center, University of Jyväskylä, FI-40014 Jyväskylä, Finland*

E-mail: hannu.j.hakkinen@jyu.fi

1 Effects of MBTR k2 parameters to Monte Carlo simulation

The theory behind the MBTR descriptor is explained in the main article. It has a few parameters, which affect its descriptive effectiveness. One can define minimum and maximum values of the sweeping variable x, adjust Gaussian broadening with σ, increase or decrease the effect of long distance terms with d and adjust the summation of distributions with cut-off.[1,2] We used the MBTR descriptor with k = 2 (MBTR k2), which contains only pairwise distances in the description. However, it does not totally neglect angular information. The parametrization actually affects the angles in protecting units during Monte Carlo simulations.

We present parameters as sets of $\{\text{min},\text{max},n_x,\sigma,d,\text{cutoff}\}$. The first used set was $\{0, 1.4, 100, 0.1, 0.5, 10^{-3}\}$ and the second one was $\{0, 1.2, 100, 0.045, 0.8, 10^{-5}\}$. The MBTR is visualized in the top row of Figures S1 and S2. The most significant pairwise term is S-Au, which is drawn with thick red line in the Figures. When the first parameter set is used, the S-Au curve is dominated by the peak at about $x \approx 0.2$. This corresponds to the distance of 5.0 Å. This is not the bond distance between the closest neighboring S and Au atoms. On the other hand, the MBTR shows two separate peaks when the second parameter set is used. One is at $x \approx 0.2$ and another one at $x \approx 0.4$. The second peak corresponds to the region close to 2.5 Å, which is close to the bond length of S-Au bond. This shows clearly that using the second parameter set descriptor can distinguish closest and second closest S and Au neighbors.

We trained EMLMs and ran Monte Carlo simulations at 200 K, 250 K and 300 K using both parameter sets. The mechanics of Monte Carlo are explained in the main article. Simulations were run for 9000 to 10000 steps and the last 500 were used for the analysis. In the Figures S1 and S2, the angle distributions are presented for the corresponding MBTR parameters. It is clear that the distributions are much more well defined, when the second parameter set is used. Especially the S-Au-S angle improves when the method uses the latter parameter set.

2 Pitfalls of using molecular dynamics as a training data

Molecular dynamics (MD) simulations are always deterministic. They create a path in a configuration space as a function of time or simulation steps. This creates a pitfall for machine learning methods, especially for those whose construction relies on the actual observations, like the reference points with the distance-based methods.

In the beginning we used the first set of MBTR parameters to describe the structures of $Au_{38}(SCH_3)_{24}$ clusters. The structures were from the publication of Juarez-Mosqueda *et al.*[3] The data set contained 25060 configurations in total. This data was then used to train the Minimal Learning Machine (MLM). From the whole data set 80% was randomly chosen to the training set. All training data points were selected as reference points during the training process. Finally the remaining 20% of the MD data was used for testing. The test results can be seen in Figure S3. It seems that the predicted potential energies of MLM follow accurately the results of DFT. In other words, the MLM could find a clear structure-property correlation from the data set. Unfortunately the MLM is greatly restricted to the path that MD had made. It is not a difficult task to predict the property (in our case potential energy) between two similar data points along the path but predicting what is outside the path is much more difficult. This was seen in Monte Carlo simulations. They frequently broke the structures and made non-physical configurations, when this MLM was used to predict potential energies. In the main article principal component analysis (PCA) of the MBTR descriptors in Figure 2 shows how much the predicted structures differ from the original training data. In order to improve the generalization capability of the method we used configurations from these Monte Carlo simulations to expand the training set.

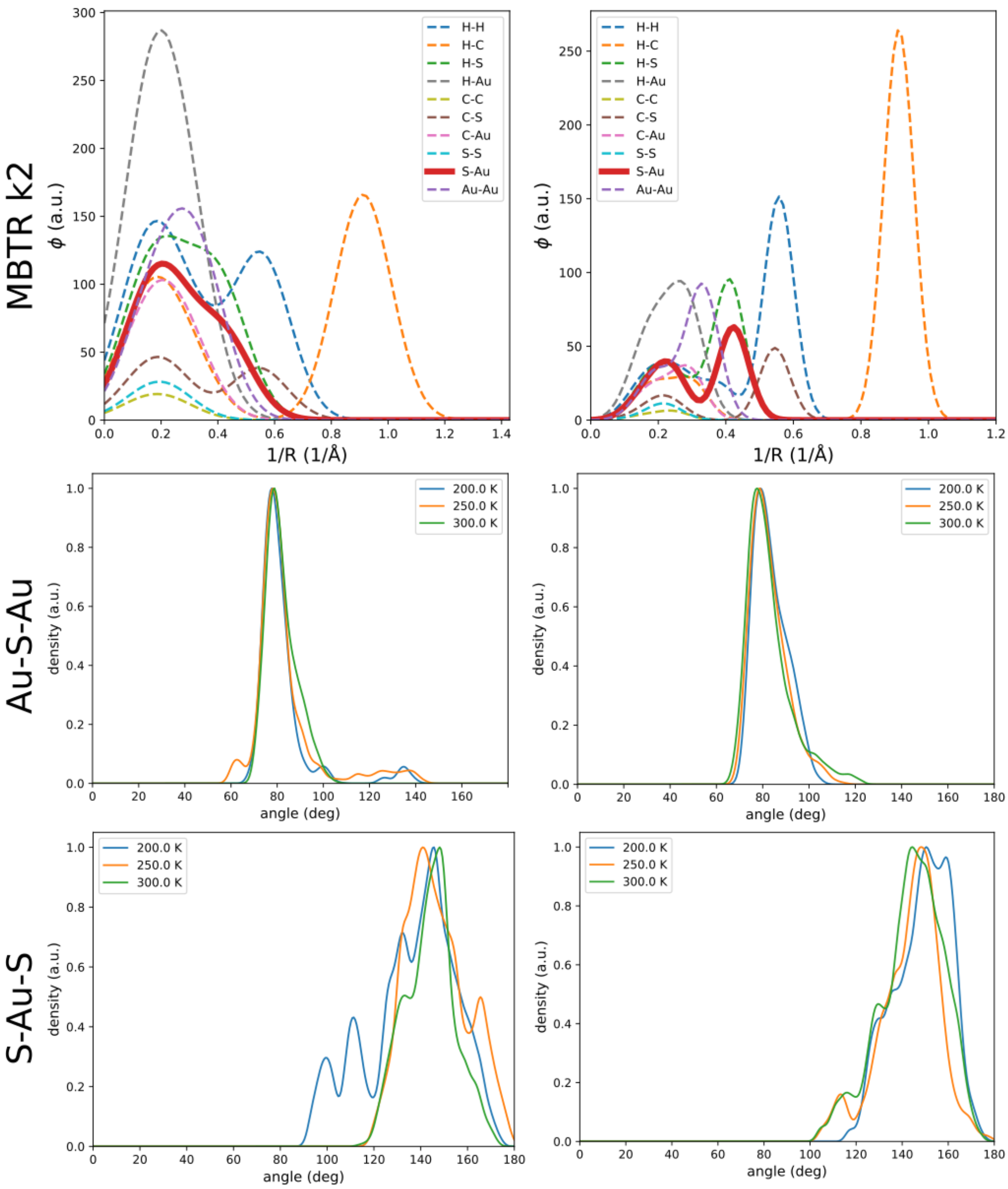


Figure S1: Here we visualize the effect of MBTR k2 to the angles of protecting units during the Monte Carlo simulations of $Au_{38}(SCH_3)_{24}$ Q. In the top row MBTR k2 is shown for different element pairs. Left side shows the results for the first parameter set and right side for the second set. The statistics of angles are summed from gaussian-smoothened ($\sigma = 1.75°$) data points.

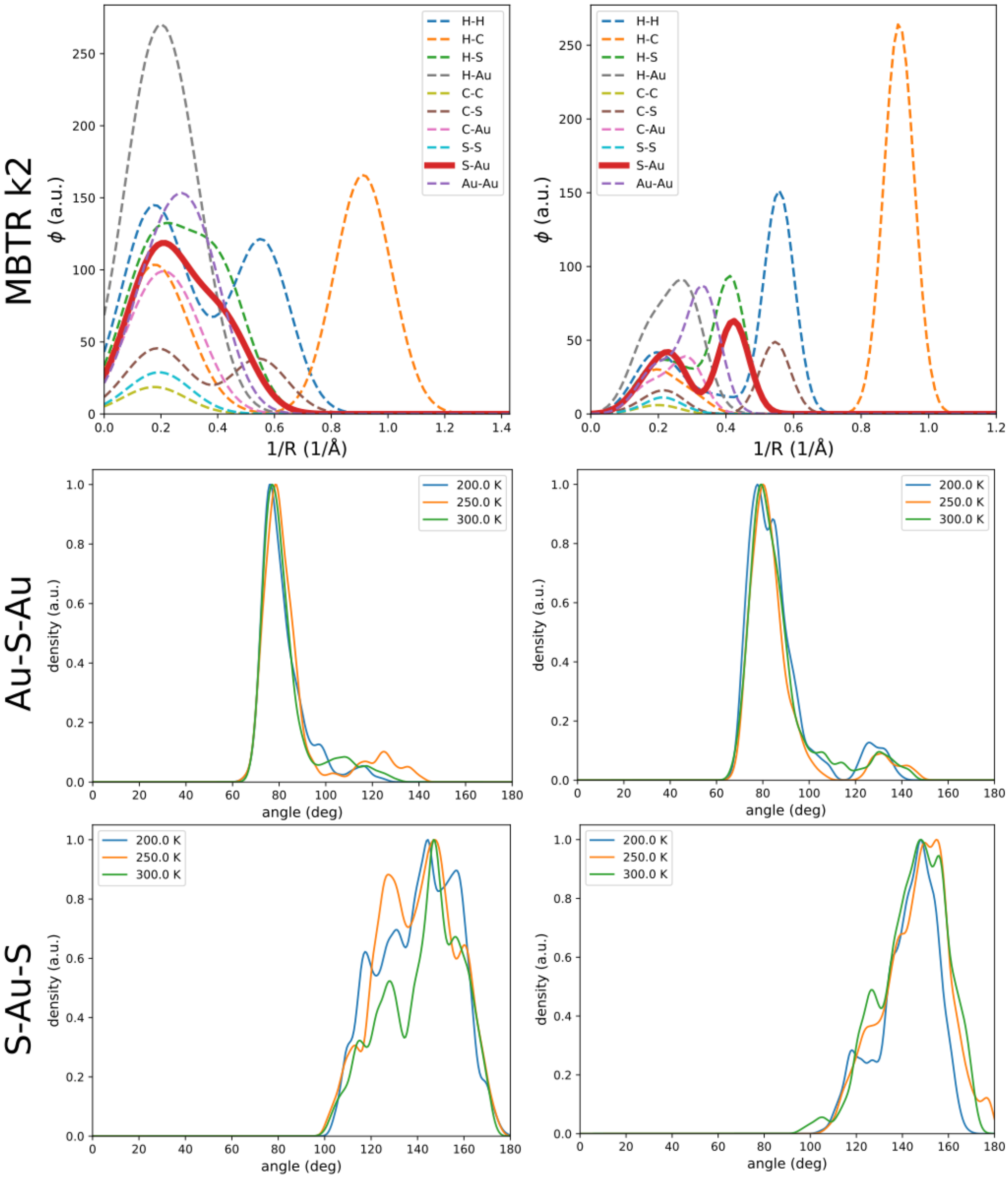


Figure S2: Here we visualize the effect of MBTR k2 to the angles of protecting units during the Monte Carlo simulations of $Au_{38}(SCH_3)_{24}$ T. In the top row MBTR k2 is shown for different element pairs. Left side shows the results for the first parameter set and right side for the second set. The statistics of angles are summed from gaussian-smoothened ($\sigma = 1.75°$) data points.

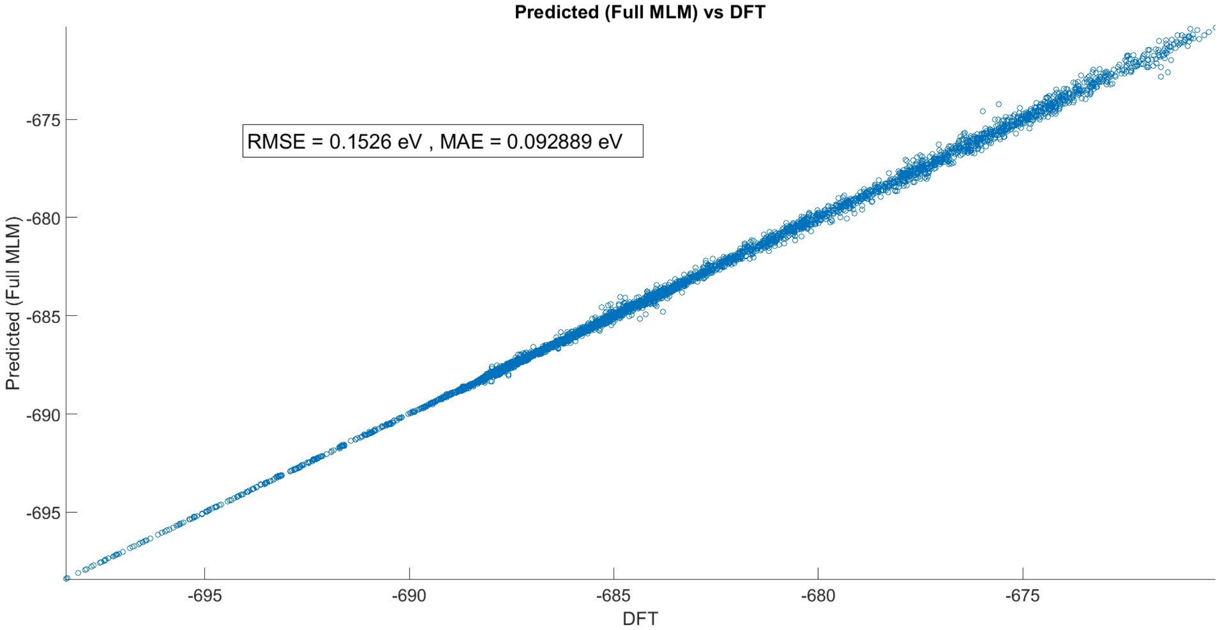


Figure S3: Here the potential energies predicted by MLM are shown as a function of real DFT level potential energies. Data set contains both Q and T isomers of $Au_{38}(SCH_3)_{24}$. From the data 80% is used as reference structures and the remaining 20% are used for testing. When MLM is interpolating within the set its predictive power is excellent. RMSE = root mean squared error, MAE = mean absolute error

References

(1) Huo, H.; Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning, arXiv.org/1704.06439v3. **2017**,

(2) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **2020**, *247*, 106949.

(3) Juarez-Mosqueda, R.; Malola, S.; Häkkinen, H. Ab initio molecular dynamics studies of $Au_{38}(SR)_{24}$ isomers under heating. *Eur. Phys. J. D.* **2019**, *73*, 62.

PIII

ORIENTATION ADAPTIVE MINIMAL LEARNING MACHINE FOR DIRECTIONS OF ATOMIC FORCES

by

Antti Pihlajamäki, Joakim Linja, Joonas Hämäläinen, Paavo Nieminen, Sami Malola, Tommi Kärkkäinen, and Hannu Häkkinen 2021

ESANN 2021: Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning Online event (Bruges, Belgium), October 06 - 08, 529—534

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7. Available from http://www.i6doc.com/en/.

Orientation Adaptive Minimal Learning Machine for Directions of Atomic Forces

Antti Pihlajamäki[1], Joakim Linja[2], Joonas Hämäläinen[2], Sami Malola[1], Paavo Nieminen[2], Tommi Kärkkäinen[2] and Hannu Häkkinen[1,3] *

1- University of Jyväskylä - Department of Physics, Nanoscience Center
FI-40014 Jyväskylä, Finland

2- University of Jyväskylä - Faculty of Information Technology
FI-40014 Jyväskylä, Finland

3- University of Jyväskylä - Department of Chemistry, Nanoscience Center
FI-40014 Jyväskylä, Finland

Abstract. Machine learning (ML) force fields are one of the most common applications of ML in nanoscience. However, commonly these methods are trained on potential energies of atomic systems and force vectors are omitted. Here we present a ML framework, which tackles the greatest difficulty on using forces in ML: accurate prediction of force direction. We use the idea of Minimal Learning Machine to device a method which can adapt to the orientation of an atomic environment to estimate the directions of force vectors. The method was tested with linear alkane molecules.

1 Introduction

In computational studies of atomic and molecular systems there are two fundamental quantities: potential energy of the system and force vectors subjecting to the atoms. In general, atomistic simulations produce output for both of these quantities. The most accurate way to compute them is to use *ab initio* methods, which are directly based on quantum mechanics. However, they are computationally demanding, which has risen the popularity of machine learning (ML) tools. This is due to the ability of ML to imitate the results of the high-level theoretical methods with lowered computational cost. Especially popular tools are ML force fields, which estimate high-dimensional potential energy surfaces of atomic systems [1]. These energy surfaces can be differentiated to get forces but the training the methods focuses on potential energies and forces are omitted.

Training a ML method to predict forces, instead of potential energies, is not simple. Chemical environments of the atoms are often presented using so-called descriptors. They produce translation, rotation and permutation invariant representations of the environment according to the chemical composition and geometry of the system [2]. They make regression tasks more feasible than in

*This work was supported by Academy of Finland through the AIPSE research program with grant 315549 to H.H. and 315550 to T.K., through the Universities Profiling Actions with grant 311877 to T.K., and through H.H.'s Academy Professorship. Work was also supported by "Antti ja Jenny Wihurin rahasto" via personal funding to A.P.. Computations were done at the FCCI node in the University of Jyväskylä (persistent identifier: urn:nbn:fi:research-infras-2016072533).

the case of using the atomic coordinates. Descriptors are highly useful but they are not suitable for predicting a rotation variant output, such as force directions, without major adjustments. However, if the description is rotation variant, the model would require large amounts of data to cover the orientation space.

We tackle the challenges above by utilizing the Minimal Learning Machine (MLM) [3] framework to create an orientation adaptive method. The main input is still an invariant description of a chemical environment. Output half of the method adjusts the spacial orientation of the reference data, which enables it to estimate force directions without having to cover the orientation space. Here we focus on the directions of the forces. Predicting the norms of the forces is a normal regression task, which can be handled with conventional ML methods such as Ridge regression or artificial neural networks.

2 Theoretical basis of orientation adaptive MLM

The general idea of the method is similar to the original MLM, which relies on separate handling of input and output spaces[3]. The force direction prediction splits into three parts, which use the descriptions of the chemical environments, coordinates of the atoms and unit force vectors. First, reference atomic coordinates are fitted on top of input coordinates, rotating reference unit force vectors respectively. Next, Euclidean distances are measured between input and reference descriptions forming a distance matrix, which is used to predict the cosines of angles between rotated reference force vectors and a force vector to be predicted. Finally, by minimizing the difference between real and predicted angles, the direction of the force is found.

2.1 Training orientation adaptive MLM

Three types of data are used in training: described chemical environments $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N \in \mathbb{R}^{N \times d_x}$, cartesian coordinates of the atoms itselves and their M nearest neighbors $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N \in \mathbb{R}^{N \times (1+M) \times 3}$, and unit vectors pointing to the directions of the forces $\mathbf{V} = \{\hat{\mathbf{v}}_i\}_{i=1}^N \in \mathbb{R}^{N \times 3}$. From this data K references are selected forming $\mathbf{Q} = \{\mathbf{q}_j\}_{j=1}^K \in \mathbb{R}^{K \times d_x}$, $\mathbf{S} = \{\mathbf{s}_j\}_{j=1}^K \in \mathbb{R}^{K \times (1+M) \times 3}$ and $\mathbf{T} = \{\mathbf{t}_j\}_{j=1}^K \in \mathbb{R}^{K \times 3}$ respectively. In $\mathbf{y}_i$ and $\mathbf{s}_i$ the first rows are the positions of the analyzed atoms itselves.

The angle between force vectors can be measured reliably only if associated atomic neighborhoods are in the same spatial orientation, therefore reference neighborhoods in $\mathbf{S}$ are aligned with the ones in $\mathbf{Y}$. We used fitting method introduced by Arun *et al.*[4]. From point sets one calculates

$$\mathbf{H}_{i,j} = \sum_{k=1}^{1+M} (\mathbf{y}_{i,k} - \mathbf{y}_{i,1})^T (\mathbf{s}_{j,k} - \mathbf{s}_{j,1}) \tag{1}$$

where $\mathbf{H}_{i,j} \in \mathbb{R}^{3 \times 3}$. Using Singular Value Decomposition (SVD) $\mathbf{H}_{i,j} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{W}^T$ one can form a rotation matrix $\mathbf{R}_{i,j} = \mathbf{W}\mathbf{U}^T$. This aligns neighbor-atoms, when the analyzed atom itself is translated to the origin. The optimal order of M

neighbor-atoms is not known, therefore permutations are tested and the success of fitting is estimated as $g_{i,j} = \frac{1}{1+M}\sum_{k=1}^{1+M} |(\mathbf{y}_{i,k} - \mathbf{y}_{i,1}) - (\mathbf{s}_{j,k} - \mathbf{s}_{j,1})\mathbf{R}_{i,j}|$. The permutation yielding the smallest $g_{i,j}$ is selected.

After SVD fitting, the matrices needed to train the model are formed. The basic training of the weight matrix $\mathbf{B} \in \mathbb{R}^{K\times K}$ is written as

$$\mathbf{B} = \left(\mathbf{D}_{\text{in}}^T \mathbf{D}_{\text{in}}\right)^{-1} \mathbf{D}_{\text{in}}^T \mathbf{D}_{\text{out}} \tag{2}$$

where $\mathbf{D}_* \in \mathbb{R}^{N\times K}$ are originally distance matrices in input and output spaces [3]. In our case there are three different $\mathbf{D}_*$ matrices: $\mathbf{D}_x = \{|\mathbf{x}_i - \mathbf{q}_j|\}$ containing Euclidean distances between chemical descriptions, $\mathbf{D}_g = \{\frac{1}{1+M}\sum_{k=1}^{1+M} |(\mathbf{y}_{i,k} - \mathbf{y}_{i,1}) - (\mathbf{s}_{j,k} - \mathbf{s}_{j,1})\mathbf{R}_{i,j}|\}$ with the goodness values of the SVD fittings and $\mathbf{D}_c = \{\hat{\mathbf{v}}_i \cdot (\hat{\mathbf{t}}_j \mathbf{R}_{i,j})\}$ having the cosines of the angles between force vectors in $\mathbf{V}$ and the rotated vectors of $\mathbf{T}$. We use equation (2) to train two weight matrices, $\mathbf{B}_g$ from goodness values of fittings and $\mathbf{B}_c$ from cosines. In both training processes $\mathbf{D}_{in} = \mathbf{D}_x$ but the output side $\mathbf{D}_{out}$ matrix is substituted with $\mathbf{D}_c$ or $\mathbf{D}_g$ respectively. The purpose of two weight matrices is to use one to predict angles and another is used to determine reliability of the data points.

SVD fitting causes variation to $\mathbf{D}_c$ and $\mathbf{D}_g$ matrices, because configurations might be difficult to fit together. This variation is behaving as a semi-random noise, distribution of which is unclear. Hence, we used Huber regression to make the model robust to outliers [5]. The idea is similar to the robust MLM by Gomes *et al.* [6]. The columns of matrices $\mathbf{B}_c$ and $\mathbf{B}_g$ are optimized by giving $\mathbf{D}_x$ and columns of $\mathbf{D}_c$ or $\mathbf{D}_g$ to Huber regressor. The regressor also produces intercept values c_j for every column of $\mathbf{B}$ to ensure that data is centered to origin. The robustness of the method is determined by the Huber parameter $\epsilon \in [1, 2]$, where 1 is producing statistically the most robust model.

2.2 Prediction of output direction

The prediction takes a description $\mathbf{x}$ of the chemical environment and $\mathbf{y}$ set of coordinates of neighbor-atoms as an input. Euclidean distances between $\mathbf{x}$ and reference descriptions in $\mathbf{Q}$ are measured forming distance vector $\mathbf{d}_x$, which is used to estimate cosines and SVD fitting successes. With normally trained model this is simply $\mathbf{d}_* = \mathbf{d}_x \mathbf{B}_*$ and with robust trained model it is $\mathbf{d}_* = \mathbf{d}_x \mathbf{B}_* + \mathbf{c}_*$, where $\mathbf{c}_*$ contains intercept values. The reference points sets in $\mathbf{S}$ are SVD fitted to $\mathbf{y}$ producing success values of fittings, which are saved to vector $\mathbf{g}$, and rotation matrices to operate reference vectors in $\mathbf{T}$.

Only references for which $g_j - d_{g,j} \leq 0$ are used to predict the direction. Otherwise accuracy of the SVD fitting is not enough. Unit vector $\hat{\mathbf{u}}$ pointing to the predicted direction is found by minimizing loss function

$$\min_{\hat{\mathbf{u}}\in\mathbb{R}^3} J(\hat{\mathbf{u}}) = -\sum_{k\in\Gamma} \exp\left(-\left(\frac{d_{c,k} - (\hat{\mathbf{t}}_k \mathbf{R}_k)\cdot\hat{\mathbf{u}}}{\sigma_1}\right)^2 - \left(\frac{g_k}{\sigma_2}\right)^2\right). \tag{3}$$

Here σ_1 and σ_2 are parameters defining the width and the depth of the contributions of the included references. Γ contains the indeces of the accepted references.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7. Available from http://www.i6doc.com/en/.

The optimization is done via Sequential Quadratic Programming (SQP). Initial guess is always a vector pointing from the atom itself to its nearest neighbor.

3 Testing with alkanes

Linear alkane molecules with number of carbon atoms ranging from two to seven were used as test systems. Thermal vibrations of the molecules were simulated by running molecular dynamics (MD) using Density Functional Tight-Binding (DFTB) code Hotbit to compute potential energies and forces [7]. For every atom initial velocities were generated from Maxwell-Boltzmann distribution with temperature of 750 K. A single run was 1000 MD steps (1 step = 1 configuration of the molecule) with 1.5 fs time step. Chemical environments were described using the Smooth Overlap of Atomic Positions (SOAP) [8] implemented in DScribe package [2]. SOAP parameters were set to $n_{max} = 6$, $l_{max} = 1$ and cut-off radius was 3.0 Å (for further details see corresponding references). For the SVD fitting four nearest neighbors were used. The dataset was produced by two separate MD runs from all molecules ($6 \times 2 \times 1000$ configurations, 2×27000 carbon environments, 2×66000 hydrogen environments). Features in SOAP descriptions were min-max scaled into $[0, 1]$ and training data of 7500 points was sampled for both elements using RS-maximin [9, 10]. All training data points were saved as reference points. The descriptions of the hydrogen and carbon atoms were sampled separately and separate models were trained for both elements. For both elements one regular and eleven robust models were trained. For robust models Huber parameters were sampled evenly from the range $[1, 2]$ with steps of 0.1. The third set of MD runs was used as test data. For carbon models, test data contained all data points from the third MD runs. For hydrogen, data points from the every second configurations of the molecules were used. The parameters in loss function (3) were $\sigma_1 = 0.25$ and $\sigma_2 = 0.5$. The performance was measured with weighted averages of the angles between predicted and real force vectors. The squared norms of the real force vectors were used as weights.

In Figure 1 A the weighted average angles are shown for different models. Training errors of regular models are out of visualization range. For hydrogen this training error is 1.3° and for carbon 7.4°. Horizontal lines, representing the test errors of regular models, lie at 55.4° for carbon and 42.1° for hydrogen. Adding robustness increases training errors but generality is improved. For carbon all robust models are working better than the regular model, the best one producing the weighted average angle of 47.2° with Huber parameter of 1.1. For hydrogen the effect of robustness is not as clear as for carbon. Only three most robust models show improvement and the best result with Huber parameter 1.0 is giving the weighted average angle of 38.3°. Panels D-G in Figure 1 show the test results for the regular models and the best robust models. In the case of carbon the effect of robustness is not clear. The main improvements lie in the region of large forces. For corresponding results for hydrogen effect is evident. The directions of large forces have improved. The directions of the small forces are difficult to handle, because even the tiniest movement of the atoms might

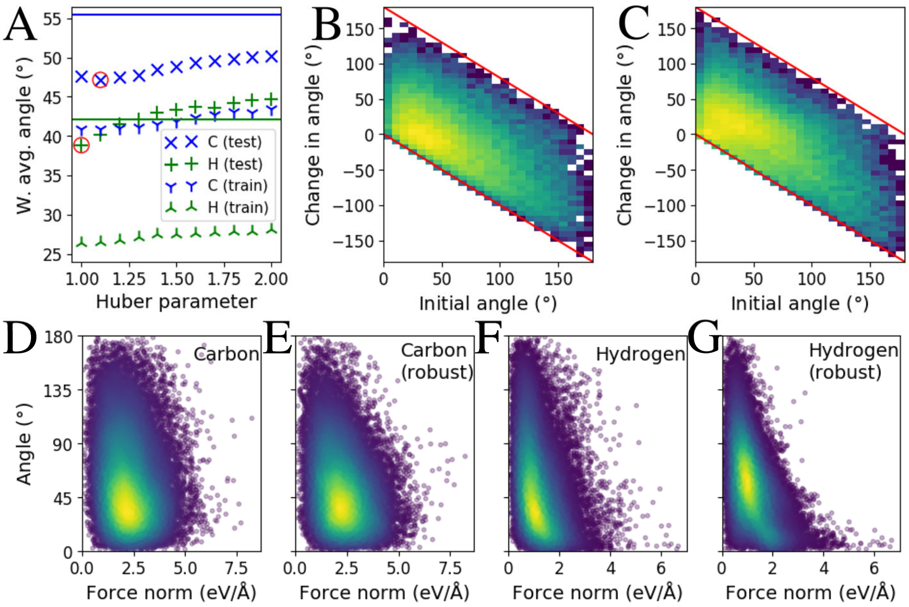


Fig. 1: **Performance of the method.** Panel A shows the weighted average angles between predicted and real directions. Lines show the test error of the regular models and crosses correspond to the robust ones with the best test results circled. Training errors of the regular models are below the visualization range. B and C show the effect of robustness from the best models of carbon and hydrogen respectively as 2D histograms. Colors are logarithmically normalized. Panels D-G show the test results. Colors present the density of the points.

totally change it. It is significantly more important to get correct directions for the large forces than for small ones. For hydrogen the robust model is starting to perform less well for small forces, which is seen as a different position of the density maximum.

In Fig. 1 B and C the effect of added robustness is visualized. Horizontal axes are the angles from the predictions with regular models. Vertical axes show the difference $\phi_j - \theta_j$, where $j \in [1, N_{test}]$, θ_j is a angle produced by the regular model and ϕ_j is a corresponding angle from the best robust model. Negative values correspond to improved predictions. Lower side red line shows the optimal correction. Data is focusing to the lower region showing that robustness is mostly improving predictions. For carbon in panel B this is clear, because data is distributed close to the lower red line. For hydrogen improvement is modest. The maximum region is spread significantly along horizontal direction (no effect) and for small initial angles prediction have worsened but the main trend is improving. A similar behavior can be seen in Fig 1 F and G.

4 Conclusions

Orientation adaptive MLM shows great promise on force direction prediction. Its advantage is that it is not bound to full atomic structures but local chemical environments are enough. The shown accuracy is not perfect but it could be improved by optimizing its several parameters such as the ones of the SOAP descriptor, the loss function and the number of reference points. We are also working to improve the fitting of atomic neighborhoods. A beautiful aspect of the method is that even after training the model, there are possibilities to affect its accuracy by tailoring the loss function and the optimization method. The future applications of the method lie in atomic structure optimization and MD simulations. However, direction estimation is not only important in nanoscience but also in, for example, engineering wind power[11] and predicting stock market[12]. Our method adds a new adjustable tool to tackle directional tasks.

References

[1] A. Pihlajamäki, J. Hämäläinen, J. Linja, *et al.* Monte Carlo Simulations of $Au_{38}(SCH_3)_{24}$ Nanocluster Using Distance-Based Machine Learning Methods. In *J. Phys. Chem. A*, vol. 124 p. 4827–4836, 2020. doi:10.1021/acs.jpca.0c01512.

[2] L. Himanen, M. O. J. Jäger, E. V. Morooka, *et al.* DScribe: Library of descriptors for machine learning in materials science. In *Comput. Phys. Commun.*, vol. 247 p. 106949, 2020. doi:10.1016/j.cpc.2019.106949.

[3] A. H. de Souza Júnior, F. Corona, G. A. Barreto, *et al.* Minimal Learning Machine: A novel supervised distance-based approach for regression and classification. In *Neurocomputing*, vol. 164 pp. 34 – 44, 2015. doi:10.1016/j.neucom.2014.11.073.

[4] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-Squares Fitting of Two 3-D Point Sets. In *IEEE T. Pattern Anal.*, vol. PAMI-9 pp. 698 – 700, 1987. doi: 10.1109/TPAMI.1987.4767965.

[5] P. J. Huber. *Robust Statistics.* John Wiley & Sons, Inc, New Jersey, USA, 1981. ISBN 0-47141805-6.

[6] J. P. P. Gomes, D. P. P. Mesquita, A. L. Freire, *et al.* A Robust Minimal Learning Machine based on the M-Estimator. In *ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 383–388. 2017.

[7] P. Koskinen and V. Mäkinen. Density-functional tight-binding for beginners. In *Comp. Mater. Sci.*, vol. 47(1) pp. 237 – 253, 2009. ISSN 0927-0256. doi: 10.1016/j.commatsci.2009.07.013.

[8] A. P. Bartók, R. Kondor, and G. Csányi. On representing chemical environments. In *Phys. Rev. B*, vol. 87, 2013. doi:10.1103/PhysRevB.87.184115.

[9] T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. In *Theor. Comput. Sci.*, vol. 38 pp. 293–306, 1985. doi:10.1016/0304-3975(85)90224-5.

[10] J. Hämäläinen, A. S. C. Alencar, T. Kärkkäinen, *et al.* Minimal Learning Machine: Theoretical Results and Clustering-Based Reference Point Selection. In *J. Mach. Learn. Res.*, vol. 21(239) pp. 1–29, 2020. URL `http://jmlr.org/papers/v21/19-786.html`.

[11] Z. Tang, G. Zhao, and T. Ouyang. Two-phase deep learning model for short-term wind direction forecasting. In *Renew. Energ.*, vol. 173 pp. 1005–1016, 2021. doi: 10.1016/j.renene.2021.04.041.

[12] M. Ballings, D. V. den Poel, N. Hespeels, *et al.* Evaluating multiple classifiers for stock price direction prediction. In *Expert Syst. Appl.*, vol. 42 pp. 7046–7056, 2015. doi: 10.1016/j.renene.2021.04.041.

PIV

ORIENTATION ADAPTIVE MINIMAL LEARNING MACHINE: APPLICATION TO THIOLATE-PROTECTED GOLD NANOCLUSTERS AND GOLD-THIOLATE RINGS

by

Antti Pihlajamäki, Sami Malola, Tommi Kärkkäinen, and Hannu Häkkinen 2022

arXiv: 2203.09788

Orientation Adaptive Minimal Learning Machine: Application to Thiolate-Protected Gold Nanoclusters and Gold-Thiolate Rings

Antti Pihlajamäki and Sami Malola

Department of Physics, Nanoscience Center,
University of Jyväskylä, FI-40014 Jyväskylä, Finland

Tommi Kärkkäinen

Faculty of Information Technology, University of Jyväskylä, FI-40014 Jyväskylä, Finland

Hannu Häkkinen

Department of Physics, Nanoscience Center,
University of Jyväskylä, FI-40014 Jyväskylä, Finland
Department of Chemistry, Nanoscience Center,
University of Jyväskylä, FI-40014 Jyväskylä, Finland[*]

(Dated: 17.03.2022)

Abstract

Machine learning (ML) force fields are one of the most common applications of ML methods in the field of physical and chemical science. In the optimal case, they are able to reach accuracy close to the first principles methods with significantly lowered computational cost. However, often the training of the ML methods rely on full atomic structures alongside their potential energies, and applying the force information is difficult especially in the case of kernel-based methods. Here we apply distance-based ML methods to predict force norms and estimate the directions of the force vectors of the thiolate-protected $Au_{38}(SCH_3)_{24}$ nanocluster. The method relies only on local structural information without energy evaluations. We apply the atomic ML forces on the structure optimization of the gold-thiolate rings and partial optimization of two known structural isomers of the $Au_{38}(SCH_3)_{24}$ nanocluster. The results demonstrate that the method is well-suited for the structural optimizations of the gold-thiolate systems, where the atomic bonding has a covalent nature in the ligand shell and at the metal-ligand interface.

I. INTRODUCTION

Monolayer-protected clusters (MPCs) are chemically diverse nanostructures consisting of metallic core, protecting organic ligand layer and an interface structure between [1]. The ligand layer stabilizes the metal particles, which would otherwise agglomerate or react with outside environment. Stabilization enables MPCs to have atomically precise structures. This chemically complex yet atomically well-defined nature of MPCs makes them an interesting research subject, where possible applications vary from catalysis and biological imaging to nanomedicine [1, 2]. Understanding the operational mechanisms of the MPCs in these applications requires development of efficient and reliable novel computational strategies.

Density functional theory (DFT) was introduced over half a century ago by Hohenberg and Kohn [3] and it has developed into the main tool in the field of computational nanoscience. However, DFT often requires lots of computational resources to be run in a reasonable amount of time. This has lead into development of various force fields, which accelerate the computations. For MPCs there have been developed, for example, ReaxFF [4] and AMBER-GROMACS [5] force fields. The drawback of these methods is that one has

* hannu.j.hakkinen@jyu.fi

to compromise accuracy and often one still needs to do extensive parameter optimization. The introduction of machine learning (ML) methods to physical and chemical sciences have offered alternative approaches to atomic simulations. ML methods are not strictly bound by predefined mathematical functions imitating physical and chemical behavior but they are used to find underlying trends on given data. This has lead into numerous ML force fields, which are able to produce similar behavior of atoms as DFT in well-defined cases with fewer computational resources [6–8]. However, even if ML force fields are one of the most common application ML methods in the research field, underlying algorithms are general and their applications are not restricted on force fields. They also have many application in material informatics [9, 10], catalysis research [11] and they can even be trained to build materials [12–14].

MPCs form a challenging nanomaterial class for ML methods, because of their chemical complexity and general low-symmetry molecular structure. However, there have been some successful studies on the subject. For example, artificial neural networks and support vector machine have been used to study synthesis and properties of MPCs [15, 16], a rule-based method has been utilized to compare local atomic environments and to construct metal-ligand interfaces [17], and distance-based ML methods have been used to predict potential energies of $Au_{38}(SCH_3)_{24}$ nanocluster for finite temperature Monte Carlo simulations of their dynamical properties [18]. $Au_{38}(SCH_3)_{24}$ is also the focus of this study. This MPC has two known isomers: a cylindrical Q isomer [19] and an oblate-like T [20]. The structures are visualized in FIG. 1.

The structural difference of these two isomers can be highlighted by writing their chemical formula using the "divide and protect" idea [21]. This means that the metallic core and protecting layer can be thought as separate entities and naturally notation should emphasize it. This way Q isomer could be written as Au_{23}@$[SR\text{-}Au\text{-}SR\text{-}Au\text{-}SR]_6[SR\text{-}Au\text{-}SR]_3$ and T isomer Au_{23}@$[SR\text{-}Au\text{-}SR\text{-}Au\text{-}SR\text{-}Au\text{-}SR]_2[SR\text{-}Au\text{-}SR\text{-}Au\text{-}SR]_3[SR\text{-}Au\text{-}SR]_3[SR]_1^b$, where the superscript b refers to a bridge site and R denotes the organic part of the thiolate. In this notation it is clear that both isomers have 23 gold atom core and protecting layers consisting of gold-thiolate oligomers or units of varying lengths. Both isomers have been found experimentally and Q isomer is thermodynamically more stable than T isomer as shown both by experiments and DFT calculations [20, 22, 23]. Having two distinct structural isomers makes this MPC a very appealing testing ground for ML methods, because one can use the data from both isomers to test the generalizability of the method.

In this study we present a local force-based ML approach to simulate atomic systems. Instead of training a ML method to predict potential energies for given configurations and then taking a gradient to obtain forces, we train our method to predict directly force vectors subjecting to individual atoms. According to the Hellman-Feynman theorem, if the Born-Oppenheimer approximation is valid, the forces are true quantum mechanical observables [24, 25]. Hence, they can be solved analytically separately from the energy calculation, which justifies the approach to use ML to predict forces directly. The goal is to create a model that handles atoms locally, which gives it a great potential to be generalized over different systems with similar local features. This kind of an generalizability has shown to be achievable at least for methods predicting electron density [26, 27]. It is relatively straightforward to predict potential energies and other scalar values. However, the potential energy is a global property of the system in the quantum mechanical point of view and it cannot be unambiguously separated from the full structure. Hence, the training of a ML method requires full structure as a single input or a collection of smaller parts but this often produces very specialized models. If a model is trained to predict potential energies for one system, it very likely will not work for another one with slight modifications. Hence, it is an attractive idea to train a model with truly local properties, such as force vectors in our case. It has to be noted that practicality of the usage of local contributions depends on the chosen ML method. With artificial neural networks it is possible to get information how much a single atom is contributing to the system [28, 29] but analyzing and using local features in kernel-based methods is more difficult.

The ML approach that we present here predicts force vectors subjecting to individual atoms without any given knowledge about the potential energy of the system. There have been attempts to predict directly force vectors from atomic data of metal nanoparticles and surfaces [30–32]. However, these attempts use rotation variant representations of atomic environments instead of conventional rotation, translation and permutation invariant descriptors. This enables one to use conventional machine learning tools but introduces a new drawback: one has to somehow cover the orientation space. This is still a viable for lattice based systems with high symmetry. For low symmetry systems, this kind of an approach requires alignment of atomic environments and/or large amounts of rotated data. Our method, on the contrary, uses conventional invariant descriptors and the ML method itself is made orientation adaptive. The approach enables fair comparison of chemical environments as the commonly used descriptors, such as Smooth Overlap of Atomic Positions (SOAP) [33], Atom-Centered Symmetry Functions (ACSF) [34] and Many-Body Tensor Representation (MBTR) [35], are already well-known and tested. Our method breaks the force prediction task into two parts: (i) prediction of the norm of the force and (ii) estimation of the direction. Both parts utilize the so-called distance-based ML, which also enables elegant prediction of different attributes from the same similarity matrix. The similarity measure, as the name suggests, is the Euclidean distance.

We trained and tested the method by using the data previously generated from DFT-level molecular dynamics (MD) simulations of the two structural isomers of $\mathrm{Au}_{38}(\mathrm{SCH}_3)_{24}$ nanocluster [22]. This data has already been used to predict potential energies using distance-based ML [18], therefore this study provides a logical continuation to the previous research. We have tested extensively different parameters related the method and applied it to the structure optimization of three different systems: gold-thiolate rings, $\mathrm{Au}_{38}(\mathrm{SCH}_3)_{24}$ with outstretched protecting units in its ligand shell and arbitrary configurations of the previously mentioned MD simulations. Gold-thiolate rings are especially interesting test case as they are not explicitly included into the training data, hence they demonstrate the generalization possibilities of our ML approach. Furthermore, their existence in cluster synthesis have been verified experimentally [36–38] and they have also been studied theoretically [39]. The tests demonstrate the usefulness of our method for coarse optimization. It can guide optimization to the close vicinity of the local minimum, which can then be reached with finer optimization via DFT. The method allows breaking and making of chemical bonds, hence in the future it could be applied to the dynamic simulations where chemical reactions can take place.

II. COMPUTATIONAL METHODS

Here we go through the theoretical background of the ML approach. First the SOAP descriptor is presented briefly to explain its parameters, which are tested during the model development. Then the background of the distance-based ML methods is introduced and how they are applied to our systems at hand.

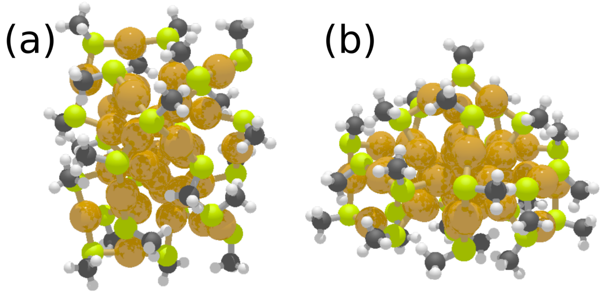


FIG. 1. Two structural isomers of the $Au_{38}(SCH_3)_{24}$ nanocluster: (a) the Q isomer [19] and (b) the T isomer [20]. The structures consist of metallic 23 gold atom core and protecting ligand layer. Metal-ligand interface is constructed from $[Au\text{-}SCH_3]_x$ oligomers or units of various lengths. Colors: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen.

A. Smooth Overlap of Atomic Positions

SOAP is a local descriptor, which means that it is used to describe a local chemical environment of an atom or a single point. The basic idea is to present every atom as a 3D Gaussian function, then present these functions as a series expansion using radial basis functions and spherical harmonics and, finally, collecting coefficient from the expansion into a power spectrum [33, 40]. We used the version implemented in DScribe package by Himanen *et al.* [40] and we follow their formalism to introduce main aspects of the SOAP.

The starting point of the SOAP is to represent every atom with a three dimensional Gaussian function. Every element is handled separately and the environment of the point $\mathbf{r}$ is written as

$$\rho^Z(\mathbf{r}) = \sum_i^{\{Z\}} e^{-\frac{|\mathbf{r}-\mathbf{r}_i|^2}{2\sigma_{SOAP}^2}}. \tag{1}$$

Here Z is an atomic number and the summation goes over all atoms of that type. The positions of these atoms are denoted with $\mathbf{r}_i$. The elegant idea behind SOAP is to use radial basis functions b_n and spherical harmonics Y_{lm} to form a series expansion of the form

$$\rho^Z(\mathbf{r}) = \sum_{nlm} c^Z_{nlm} b_{nl}(r) Y_{lm}(\theta, \phi). \tag{2}$$

The coefficients c^Z_{nlm} are the heart of the whole description. They are solved via integration

$$c^Z_{nlm} = \int\int\int dV \; b_{nl}(r) Y_{lm}(\theta, \phi) \rho^Z(\mathbf{r}) \tag{3}$$

and then collected into a power spectrum

$$p^{Z_1,Z_2}_{nn'l} = \pi\sqrt{\frac{8}{2l+1}} \sum_m \left(c^{Z_1}_{nlm}\right)^* c^{Z_2}_{n'lm}. \tag{4}$$

The values $p^{Z_1,Z_2}_{nn'l}$ are stored into a vector, which works as a local description of the point $\mathbf{r}$. The equation (4) is slightly different than the one in the original publication of Bartók *et al.* [33]. In the DScribe package Himanen *et al.* use real (tesseral) spherical harmonics instead of complex ones and, in addition to this, they replace polynomial radial basis functions with Gaussian type orbitals

$$b_{nl}(r) = \sum_{n'=1}^{n_{\max}} \beta_{nn'l} \; r^l \; e^{\alpha_{n'l} r^2}. \tag{5}$$

This simplifies the theory and makes programming the descriptor efficient. In practise, the summation in the series does not include all indices n and l but they are restricted to maximum values n_{max} and l_{max}, which are parameters of the descriptor. The index l restricts the values integer m, because $m \in [-l, l]$ same way as side quantum number restrict magnetic quantum numbers. Furthermore, only atoms within some pre-defined cut-off radius r_{cut}, which also is a parameter, are included in to the summation in 1. For further details, see references [33, 40]. In this study, we tested the effects of four SOAP parameters: n_{max}, l_{max}, r_{cut} and Gaussian broadening σ_{SOAP}.

B. Distance-based ML tools

The basic construct in the distance-based machine learning is to use Euclidean distances between reference and input data as a measure of similarity and to predict an output using these distances. There are two main distance-based ML methods: Minimal Learning Machine (MLM) [41] and Extreme Minimal Learning Machine (EMLM) [42]. Both of them are general ML tools and they have been used successfully to predict potential energies for $\mathrm{Au_{38}(SCH_3)_{24}}$ nanoclusters [18]. Distance-based methods are especially appealing methods to study complex nanostructures, because they have been shown to work well with high-dimensional data and even out-perform deep neural networks in some cases [43]. This is due to the distance matrix, which effectively hides the dimensionality of the data. The same feature also makes distance-based ML methods resistant to overfitting [44]. In addition to this, distance-based ML methods usually have only one hyperparameter: the number of reference points, which reduces parameter testing. When applying ML methods to nanosystems, there are often several parameters to tune, such as the ones of the descriptors. This means that the user have to optimize the way how the data is presented prior the actual model can be trained. The lack of hyperparameters reduces the need for complex model fitting with different parametrizations of the descriptor.

Recently, a variation of MLM, which specifically addresses the directions of the atomic forces, was proposed: Orientation Adaptive Minimal Learning Machine (OAMLM) [45]. It takes the concept of using Euclidean distances as an input space similarity measure to perform predictions but instead of predicting corresponding distances to output space references, as MLM does [41], it predicts cosines of angles between reference vectors and a target vector. It can also produce estimates for the uncertainty of the predictions to provide interesting opportunities for different applications, where uncertainty might play a role.

We go through the theory behind the distance-based ML methods to form a basis for the discussion on OAMLM and the full force prediction framework. All of these methods start with the input data $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N \in \mathbb{R}^{N\times n_x}$ and corresponding output data $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N \in \mathbb{R}^{N\times n_y}$. In our case, $\mathbf{X}$ contains SOAP descriptions of the chemical environments of the atoms and $\mathbf{Y}$ information about the forces, either norms or unit vectors pointing to the directions of the force vectors. Let's first consider the simplest method EMLM, which is used to predict the norms of the forces given in $\mathbf{Y}$. From the input data $\mathbf{X}$, K reference points are sampled forming a reference set $\mathbf{Q} = \{\mathbf{q}_j\}_{j=1}^K \in \mathbb{R}^{K\times n_x}$. The training of EMLM is done via regularized least-squares optimization problem, which is used to find optimal weights to perform regression from Euclidean distances between points in $\mathbf{X}$ and $\mathbf{Q}$ to predict $\mathbf{Y}$ [42].

$$\min_{\mathbf{W}\in\mathbb{R}^{K\times n_y}} J(\mathbf{W}) = \frac{1}{2N}\sum_{i=1}^{N}\left|\mathbf{d}_i^T\mathbf{W} - \mathbf{y}_i^T\right|^2 + \frac{\beta}{2K}\sum_{i=1}^{K}\sum_{j=1}^{n_y}|W_{ij}|^2. \tag{6}$$

Vector $\mathbf{d}_i \in \mathbb{R}^K$ contains Euclidean distances between ith input data point and K references. $\mathbf{W} \in \mathbb{R}^{K\times n_y}$ is a weight matrix, which does a linear regression from kernel space to output. Constant β is used for regularization, which might be useful if one has noisy data. In our case, it is fixed to the square root of machine epsilon.

The minimum of the equation 6 can be found by writing it with full matrices and finding the zero point of the first derivative.

$$\frac{1}{N}\mathbf{D}^T\left(\mathbf{D}\mathbf{W}-\mathbf{Y}\right)+\frac{\beta}{K}\mathbf{V}=0 \tag{7}$$

$$\left(\mathbf{D}^T\mathbf{D}+\frac{\beta}{K}\mathbf{I}\right)\mathbf{W}=\mathbf{D}^T\mathbf{Y} \tag{8}$$

Matrix $\mathbf{D} \in \mathbb{R}^{N\times K}$ contains all Euclidean distances between training data and references. The equation (8) is now a simple representation of the training of EMLM and it can be easily solved numerically. To predict output for an arbitrary input, one has to calculate distances between the input and references forming $\mathbf{d} \in \mathbb{R}^K$ and then compute matrix multiplication $\mathbf{d}^T\mathbf{W}$. This is analogous to Kernelized Ridge Regression (KRR), where one has a variety of choices for kernel functions [46].

Next we shall go through the framework of the MLM presented by de Souza *et al.* [41] and proceed step by step to the direction prediction scheme of the OAMLM. The main difference between MLM and EMLM is that in addition to references $\mathbf{Q}$ in input space MLM also has references $\mathbf{T}=\{\mathbf{t}_j\}_{j=1}^K \in \mathbb{R}^{K\times n_y}$ in output space. The idea is not to predict directly output for certain input but to form regression between the two distance spaces.

$$\mathbf{D}_{out}=\mathbf{D}_{in}\mathbf{B}+\epsilon. \tag{9}$$

Here $\mathbf{D}_{in} \in \mathbb{R}^{N\times K}$ contains Euclidean distances between the N input training data points in $\mathbf{X}$ and K reference points in $\mathbf{Q}$. $\mathbf{D}_{out} \in \mathbb{R}^{N\times K}$, on the other hand, consists of distances between training output data in $\mathbf{Y}$ and the output reference set $\mathbf{T}$. $\mathbf{B} \in \mathbb{R}^{K\times K}$ is a weight matrix that performs the linear regression and ϵ is a residual, which is assumed to be small. It is shown that the approximate solution for the weight matrix is [41]

$$\mathbf{B}=\left(\mathbf{D}_{in}^T\mathbf{D}_{in}\right)^{-1}\mathbf{D}_{in}^T\mathbf{D}_{out}. \tag{10}$$

In order to calculate output with MLM, one first predicts distances between still unknown result and output space references using input space distances and just solved weights as $\mathbf{d}_{out}^T=\mathbf{d}_{in}^T\mathbf{B}$. The result is found by solving multilateration problem, for which there are several methods [44, 47].

With these derivations at our disposal, let us proceed to the OAMLM. To remind, the input space training data $\mathbf{X}=\{\mathbf{x}_i\}_{i=1}^N \in \mathbb{R}^{N\times n_x}$ contains SOAP descriptions of chemical environments, which do not include directional information. For this reason OAMLM also needs coordinates of neighboring atoms $\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times(1+M)\times 3}$ as an accompanying data. In $\mathbf{p}_i$ the first row is the position of the studied atom itself followed by M neighbors. For every training data point there are also their unit force vectors collected into $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times 3}$, where $|\mathbf{y}_i| = 1$ for all values of i. Similar to MLM, OAMLM also uses references both in input and output spaces. The K reference data points used are sampled into $\mathbf{Q} = \{\mathbf{q}_j\}_{j=1}^{K} \in \mathbb{R}^{K\times n_x}$ for chemical descriptors, $\mathbf{S} = \{\mathbf{s}_j\}_{j=1}^{K} \in \mathbb{R}^{K\times(1+M)\times 3}$ for coordinates of the neighboring atoms and $\mathbf{T} = \{\mathbf{t}_j\}_{j=1}^{K} \in \mathbb{R}^{K\times 3}$ for corresponding unit force vectors.

Atomic environments can be in any spatial orientation, therefore the directions of the forces cannot be compared directly. As a solution, the coordinates of neighboring atoms in $\mathbf{P}$ and $\mathbf{S}$ are used to align atomic environments. In this study we used Singular Value Decomposition (SVD) based method presented originally by Arun *et al.* [48]. First the atoms, for which forces are predicted, are moved to the origin and their neighbors are translated together with them to preserve the general positioning. Then matrix $\mathbf{A}_{i,j} \in \mathbb{R}^{3\times 3}$ is formed by calculating it as

$$\mathbf{A}_{i,j} = \sum_{k=1}^{1+M} (\mathbf{p}_{i,k} - \mathbf{p}_{i,1})(\mathbf{s}_{j,k} - \mathbf{s}_{j,1})^T. \tag{11}$$

Index i refers to the ith input and j stands for the jth reference. With SVD one can split this matrix as $\mathbf{A}_{i,j} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T$. These can be further used to get a rotation matrix $\mathbf{R}_{i,j} = \mathbf{V}\mathbf{U}^T$, which will align points in $\mathbf{S}$ and $\mathbf{P}$ as well as possible, when the atoms are moved to the origin as in equation (11). It is important to notice that this alignment approach depends on the order of given neighborhood points, therefore one needs to form certain rules how the environments are aligned or go through all permutations. However, OAMLM is not restricted to the alignment approach used here. In principle, it is possible to define any alignment scheme suited for specific problems.

The rotation matrices are used to align atomic neighborhoods together, which then yields estimates of alignment accuracy as

$$g_{i,j} = \frac{1}{1+M} \sum_{k=1}^{1+M} |(\mathbf{p}_{i,k} - \mathbf{p}_{i,1}) - \mathbf{R}_{i,j}(\mathbf{s}_{j,k} - \mathbf{s}_{j,1})| \tag{12}$$

or

$$g'_{i,j} = \frac{1}{1+M}\sqrt{\sum_{k=1}^{1+M} |(\mathbf{p}_{i,k} - \mathbf{p}_{i,1}) - \mathbf{R}_{i,j}(\mathbf{s}_{j,k} - \mathbf{s}_{j,1})|^2}. \tag{13}$$

The same rotation matrices are also used to rotate reference unit force vectors in $\mathbf{T}$ to be comparable with data in $\mathbf{Y}$. Dot products between these vectors are calculated as $\hat{\mathbf{y}}_i \cdot (\mathbf{R}_{i,j}\hat{\mathbf{t}}_j)$. This dot product is the cosine of the angle between two vectors, as we are working with unit vectors, and it is evaluated by OAMLM during the prediction phase [45]. The $g^{(\prime)}_{i,j}$ and dot products are used to form matrices $\mathbf{D}_g = \{g^{(\prime)}_{i,j}\} \in \mathbb{R}^{N\times K}$ and $\mathbf{D}_c = \{\hat{\mathbf{y}}_i \cdot (\mathbf{R}_{i,j}\hat{\mathbf{t}}_j)\} \in \mathbb{R}^{N\times K}$ respectively.

Now one has everything needed to train the OAMLM using the same training scheme as for MLM in equation (10). $\mathbf{D}_{in}$ is the same as before: Euclidean distances between datapoints in $\mathbf{X}$ and $\mathbf{Q}$. However, $\mathbf{D}_{out}$ is different. As mentioned in the reference [45], OAMLM has two weight matrices: $\mathbf{B}_c$ to predict dot products and $\mathbf{B}_g$ to predict alignment successes. To acquire those $\mathbf{D}_{out}$ in equation (10) is substituted with $\mathbf{D}_c$ or $\mathbf{D}_g$ correspondingly. However, in this study we do not use $\mathbf{B}_g$, which could be used for uncertainty estimation. We use only $\mathbf{B}_c$ to predict dot products.

The output prediction procedure with OAMLM is similar to the methods in MLM. The schematic picture of the full force prediction process is shown in the FIG. 2. As an input, the method takes description $\mathbf{x}_i$ and its neighborhood coordinates $\mathbf{p}_i$. Vector $\mathbf{d}_{in}$ is formed by calculating Euclidean distances between $\mathbf{x}$ and the reference points in $\mathbf{Q}$. The weight matrix $\mathbf{B}_c$ is used to predict dot products as $\mathbf{d}_c^T = \mathbf{d}_{in}^T\mathbf{B}_c$. Then reference neighborhood coordinates in $\mathbf{S}$ are are aligned with $\mathbf{p}$ yielding alignment accuracies $g^{(\prime)}_{i,j}$ and with corresponding rotation matrices reference unit vectors in $\mathbf{T}$ are rotated accordingly. The last part of the prediction is similar to the multilateration problem. However, instead of minimizing the distance differences we minimize the difference between the predicted dot products and dot products of the rotated reference unit force vectors $\mathbf{R}_{i,j}\hat{\mathbf{t}}_j$ and yet unknown vector $\hat{\mathbf{v}}$. There is no specific method to do this. In the reference [45] the $\hat{\mathbf{v}}$ was found numerically by using Sequential Quadratic Programming (SQP) to optimize cost function

$$\min_{\hat{\mathbf{v}}_i\in\mathbb{R}^3} J_1(\hat{\mathbf{v}}_i) = -\sum_{j=1}^{K} \exp\left(-\left(\frac{d_{c,j} - (\mathbf{R}_{i,j}\hat{\mathbf{t}}_j)\cdot\hat{\mathbf{v}}_i}{\sigma_1}\right)^2 - \left(\frac{g^{(\prime)}_{i,j}}{\sigma_2}\right)^2\right), \tag{14}$$

We call this a numeric loss function. Here we do not make initial selection of the used reference data as in the original paper [45] but we simply use all references.

In this study we decided to also use more simple cost function as a comparison:

$$\min_{\hat{\mathbf{v}}_i \in \mathbb{R}^3} J_2(\hat{\mathbf{v}}_i) = \frac{1}{2} \sum_{j=1}^{K} \omega_{i,j} \left[\hat{\mathbf{v}}_i \cdot (\mathbf{R}_{i,j} \hat{\mathbf{t}}_j) - d_{c,j} \right]^2 , \tag{15}$$

where

$$\omega_{i,j} = \exp \left(- \left(\frac{g_{i,j}^{(\prime)}}{\sigma_2} \right)^2 \right) . \tag{16}$$

The advantage of equation (15) is that it can be solved analytically by taking a derivative over $\hat{\mathbf{v}}_i$ and as a result

$$\hat{\mathbf{v}}_i = \frac{\sum_{j=1}^{K} \omega_{i,j} d_{c,j} (\mathbf{R}_{i,j} \hat{\mathbf{t}}_j)}{\sum_{j=1}^{K} \omega_{i,j}} . \tag{17}$$

The result is interestingly a weighted average of predicted projections. In practise, $\hat{\mathbf{v}}_i$ is not a unit vector, because there is always numeric error present in the values of $\mathbf{d}_{c,j}$ and $\omega_{i,j}$, therefore one has to remember to divide it with its norm before using the result. We call equation (15) as an analytic loss function. In these two loss functions, σ_1 and σ_2 are parameters of the ML model and they are also tested during the model development.

C. Atomic force prediction scheme for $Au_{38}(SCH_3)_{24}$

$Au_{38}(SCH_3)_{24}$ nanocluster, which is shown in FIG. 1, contains four different elements and has chemically various environments. There are covalently bound methyl thiolate ligands. There is a metallic gold core, where gold atoms are interacting with each other. On the surface of the core some gold atoms can also form bonds to the sulfur atoms. Within the metal-ligand interface structure, sulfur and gold atoms are bound with relatively covalent nature forming protecting units. Within these units the gold atoms are bound only to sulfur atoms, ideally forming two Au-S bonds. There are very diverse features determining the interactions between atoms, therefore it is a good idea to split the problem into smaller parts.

We classify the atoms into five categories: core gold atoms inside the metallic core, unit gold atoms in protecting units, sulfur, carbon and hydrogen. For every atom type

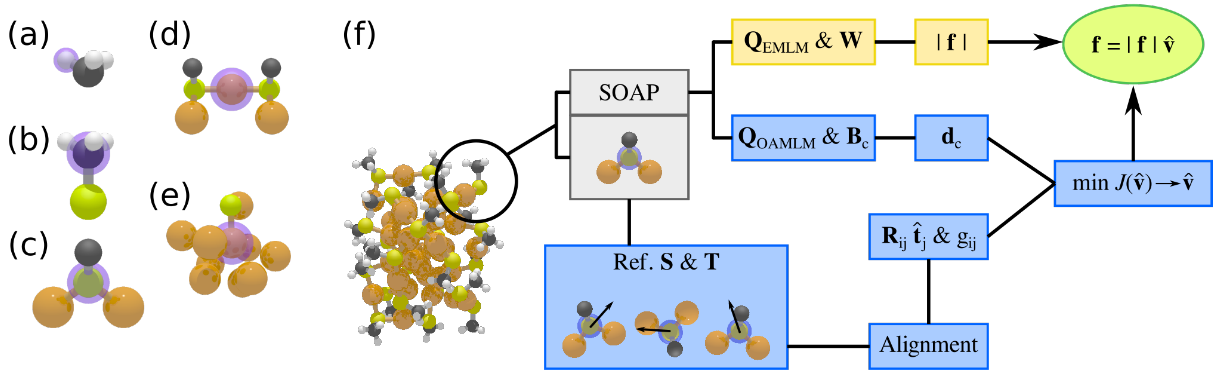


FIG. 2. The atomic force prediction framework. Examples of atomic environments used in alignment for (a) hydrogen, (b) carbon, (c) sulfur, (d) unit gold and (e) core gold. The atoms, for which the alignment is done, are highlighted with purple. Panel (f) demonstrates the full force prediction scheme. Description part is shown in grey boxes, norm prediction with EMLM in yellow and the direction estimation of the OAMLM in blue boxes. Colors for atoms: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen.

we train one EMLM for force norms and one OAMLM for force directions. The norm prediction part is a straightforward standard ML problem, where the method predicts a scalar output according to a given input and the references. For the direction scheme, we have to define, which neighborhood atoms are used to align reference environments to an input environment. In principle, one could just select n nearest neighbors and go through all permutations. However, this wastes computational resources by attempting many unfavorable permutations. Hence, we need to define certain rules according to physical and chemical understanding.

The most simple alignment scheme is for hydrogen. It uses only the nearest carbon, and two other nearest hydrogen atoms bound to the carbon as seen in the FIG. 2 (a). There are only two permutations of the hydrogen atoms to test. Aligning carbon is similar to the hydrogen scheme. It uses the nearest sulfur atom and three hydrogen atoms, as shown in the FIG. 2 (b), which results into six permutations of hydrogen atoms to be tested. The alignment of a sulfur atom neighborhood uses the nearest carbon and two nearest gold atoms shown in the FIG. 2 (c). There are only two permutations of the gold atoms to be tested.

The gold atoms have the most versatile chemical environments of all atoms in the cluster. Unit gold atoms use two blocks of atoms for alignment. The blocks contain the nearest sulfur atom and two other atoms bound to it: a carbon and another gold atom. Hence, there are two sulfur, two carbon and two gold atoms used to do the alignment. An example of the neighborhood structure is visualized in FIG. 2 (d). These atoms are handled as blocks, due to the linear nature of the S-Au-S bonding, therefore there are only two permutations to test.

The MD data used in model development is extremely dynamic and the nature of the Au-S bonds might change significantly. Hence, if a gold atom has only one sulfur within 3.0 Å and there is no another gold atom within the same distance, the gold atom is considered to be just a half of an unit. This corresponds to a transition state where old unit is broken and new is going to be formed. In this case alignment is done by using only one block of sulfur, carbon and gold atoms. This kind of alignment is much more unstable than the standard way but fortunately breaking of S-Au bond is not a common phenomenon. For hydrogen, carbon, sulfur and unit gold atoms the alignment accuracy is calculated using the equation (12).

The environments in the metallic core gold atoms can be very homogeneous making it difficult to be aligned. Within the core the gold atoms can be bound to a single sulfur atom and the rest of the interactions are metallic or another scenario is that all interactions are metallic. For every core gold there can be maximum of twelve neighboring atoms selected. If there is a sulfur atom within 3.0 Å, it will be selected first. Then the rest are nearest gold atoms within 5.0 Å from the nearest to the furthest. There can be less than twelve neighbors selected for a core gold atom, if there are not so many fulfilling the requirements as seen in the FIG. 2 (e). It is clear that there are too many neighboring atoms to go through all possible permutations in a reasonable amount of time. There can be maximum $12! = 479001600$ permutations for a single atomic neighborhood. In order to make the task feasible, we device two alignment schemes depending on whether the aligned gold atoms are bound to a sulfur atom or not.

The first scenario for core gold is that both input and reference gold atoms have a sulfur atoms within their immediate vicinity. In this case, the alignment is done by using three points: gold atom itself, sulfur atom and one neighboring gold atom. For input environment we select the nearest neighboring gold atom as the third point. For reference environment the selection is the same except that in addition to the nearest neighboring gold atom we also go through all other possible neighboring gold atoms. These three atoms are used to make alignments and the accuracy is evaluated with equation (13).

The second scenario is that at least one of the environments does not contain sulfur. Here the alignment uses only three atoms similarly to the previous core gold scenario. For the input environment the three points are the atom itself and its two nearest neighbor gold atoms. For the reference environment, we use the atom itself and all possible pairs of the neighbors. Here the order does play a role, therefore one would get maximum of $12 \cdot 11 = 132$ pairs.

These pairs together with the main gold atom form triangles, which are used to rule out some permutations. The difference between the kth triangle of the reference environment and the triangle formed from input data is measured as

$$u_k = \sum_{i=1}^{3} [(l_{k,i} - l_{0,i})^2 + (\theta_{k,i} - \theta_{0,i})^2]. \tag{18}$$

Here $l_{k,i}$ is the length of the ith side of the triangle in Ångstroms and $\theta_{k,i}$ is an angle of the ith corner in radians. The lower index k refers to the reference data triangle and lower index 0 to the input data triangle. Then n triangles, for which the difference u_k is the smallest, are selected. We decided to use $n = 10$. These triangles are used to make SVD alignments and the one yielding the smallest value of the equation (13) is selected.

As mentioned earlier, the number of neighborhood atoms for the core gold atoms is not constant. Hence, when the alignment success is estimated, it is required that every atom has some nearest neighbor distance. Let us clarify this via an example. If input environment has 6 neighbors and reference has 10, then after the alignment we measure the nearest neighbor distance for all 10 atoms in the reference environment and use them in the equation (13). It does not matter whether reference or input has more atoms but the accuracy is always estimated with the largest number of nearest neighbor distances. This is used to emphasize the differences between the atomic environments of the core gold atoms.

Implementing chemical rules and primary knowledge into the algorithm resembles the approach to construct metal-ligand interfaces by Malola *et al.* [17]. There authors used distances and angles to compare environments between reference structures and the environments of arbitrary points within unprotected metal clusters. This comparison enabled them to determine whether or not those points were suitable for interface atoms. Our force prediction method shows similar philosophy to the task but here we have to use actual spatial alignment in order to capture the orientation information.

D. Structure optimization via ML forces

As an usage example of ML forces, we perform structure optimization in different scenarios. The model does not yield values for potential energy of the system but the optimization is run solely with ML estimated forces. We used classic quasi-Newton method Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [49–52] to run structure optimization. The challenge is that ML predicted forces have always some level of uncertainty, which is seen in the optimization algorithm as a noise. In this study we do not explicitly address the uncertainty in the optimization algorithm but it is an aspect that should be considered in the future studies. The used BFGS implementation is based on the one included in Atomic Simulation Environment (ASE) package [53].

E. DFT methods

For reference calculations, we used the DFT code GPAW [54] as it was also used in the original MD simulations of $Au_{38}(SCH_3)_{24}$ by Juarez-Mosqueda *et al.* [22]. The exchange-correlation functional was Perdew-Burke-Ernzerhof functional (PBE) [55] and we used 0.2 Å real space grid spacing. BFGS structure optimization using GPAW computed potential energies and forces were run with the original implementation in ASE package [53]. The DFT-level BFGS optimizations were considered to be converged if the maximum force of the atoms was ≤ 0.05 eV/Å.

III. RESULTS AND DISCUSSION

The results are divided into six parts. First the effect of SOAP parameters to norm and direction prediction are shown. This way the optimal description parameters are found. They are used in the next two parts were full EMLM models for norms and OAMLM models for directions are trained and tested. The last three parts focus on structure optimization. The used test cases are gold-thiolate rings, $Au_{38}(SCH_3)_{24}$ cluster structures with outstretched protecting units and snapshots from the MD simulations of the both isomers of the $Au_{38}(SCH_3)_{24}$ nanocluster.

The training and testing of the models relies heavily on the DFT-level MD simulation data of the $Au_{38}(SCH_3)_{24}$ nanocluster from reference [22]. In that study, authors run long MD simulations on both isomers of the $Au_{38}(SCH_3)_{24}$, where the systems were heated so that they broke down. The less stable T isomer started to undergo significant structural changes very early and in the later stages highly deformed seven gold atom gold-thiolate ring broke out of the structure. These simulations resulted into over 12 000 configurations for both isomers, which serve as an ideal dataset for our study here.

A. Data and SOAP parameter selection

The data used to train and test our model was extracted from the DFT level MD simulations of $Au_{38}(SCH_3)_{24}$ published in the reference [22]. For both isomers we sampled logarithmically 1000 configurations. This guaranteed that we got denser sampling from the high temperature region, where there are more changes in the structure, than from the low temperature region. This data contains 24 000 local environments for carbon and sulfur, 72 000 for hydrogen from both isomers. Q isomer data contains 22 836 core, 15 123 unit and 41 half unit gold atoms. T isomer data contains 22 055 core, 15 888 unit and 57 half unit gold atoms.

The importance of the level of description cannot be emphasized too much. If description is not accurate enough the prediction will be poor. However, if description is overly accurate, it will lead to a highly specialized model, which cannot be generalized and the risk of overfitting increases. We tested several SOAP parameters: $r_{cut} \in \{4.0\text{Å},\ 5.0\text{Å}\}$, $\sigma_{SOAP} \in \{1.0, 0.75, 0.5, 0.25\}$, $n_{max} \in [2, 7]$ and $l_{max} \in [0, 4]$. This totals 240 description sets for sulfur, carbon and hydrogen. For gold atoms we used only $\sigma_{SOAP} = 0.25$ sets resulting 60 SOAP parameter sets. In this article and its Supplemental Material, we show only a selected portion of the tests. The complete analysis of the parameter tests is available in (`https://gitlab.jyu.fi/aneepihl/oamlm_forces.git`).

First these sets were used to predict norms of the forces and to restrict the number of parameters to be tested in the direction prediction scheme. It is easier to predict norms than directions, therefore it is reasonable to assume that if norms are predicted inaccurately directions won't be any better. For every parameter set we trained one EMLM with Q isomer data and one EMLM with T isomer data. Then we used Q model to predict norms from T set and vice versa. This is close to so-called cross validation approach often used when testing ML methods.

From every data set of a single isomer, 2500 points were selected with RS-maximin sampling [44]. This data was used as a training data and all points were saved as references into the models. The SOAP data points were minmax scaled between 0 and 1. The performance was measured with root mean squared error (RMSE). Tests showed the most promising parameters to be $\sigma_{SOAP} = 0.25$, and $(n_{max}, l_{max}) \in \{(6,4),(7,3),(7,4)\}$ with both $r_{cut} = 4.0$ Å and $r_{cut} = 5.0$ Å resulting to only six parameter sets to be tested with OAMLM. The results with $\sigma_{SOAP} = 0.25$ and $r_{cut} = 4.0$ Å are shown in the Supplemental Material figures S1 − S4 for sulfur, S5 − S8 for carbon, S9 − S12 for hydrogen, S13 − S16 for unit gold and S17 − S20 for core gold.

Testing with OAMLM was done in a similar fashion as with EMLM: models were trained with one isomer and then tested with another. As mentioned in the section II B, the output direction estimation can be done via numeric or analytic loss function by using either equation (14) or (15). The initial tests were ran with both output estimation methods and their parameters were set as $\sigma_1 = 0.25$ and $\sigma_2 = 0.5$. The performance was measured with weighted average of the angles between the estimated force directions and the corresponding DFT calculated force vectors. The squared norms of the DFT force vectors worked as weights. This emphasizes the handling of the large forces, for which it is more important to get directions correct than for the small ones. When the norm decreases the direction of the force vector becomes more and more elusive and sensitive to changes in the chemical environment.

The analytic loss function was performing better than the numeric one, which showed unstable performance. Parameters $\sigma = 0.25$, $n_{max} = 7$, $n_{max} = 4$ and $r_{cut} = 4.0$ Å showed satisfying performance for all atom types. The results with these parameters using numeric loss function are shown in Supplemental Material figure S21 and analytic loss function results are shown in S22. The longer cut-off radius did not show a significant improvement compared to the selected one, therefore it is natural to use shorter one. There will be less atoms included into the description making it slightly faster to compute and it is more likely to results in generalizable method.

After finding the optimal SOAP parameters, we also tested how σ_2 parameter affects the performance of the analytic loss function. In addition to the previously used value of 0.5, we also tested values 0.25 and 0.75 with previously acquired SOAP parameters. The results with these parameters are shown in Supplemental Material figures S23 and S24. The tests do not show any significant effect to better or worse. For unit gold atoms the $\sigma_2 = 0.25$ seem to be slightly better option than 0.5, because the weighted average angles were previously approximately 29° (Q $\rightarrow$ T) and 25° (T $\rightarrow$ Q), and with smaller σ_2 parameter the values decreased to about 25° (Q $\rightarrow$ T) and 24° (T $\rightarrow$ Q). We settled on $\sigma_2 = 0.25$ for unit gold atoms and for everything else $\sigma_2 = 0.5$.

B. Full EMLM force norm models

After determining suitable parameters for the SOAP description, we trained EMLM with combination of data from both Q and T isomers. From the combined data set, 5000 points for each atom type were selected with RS-maximin sampling [44]. This data was used as a training data and all points were saved as references. All descriptions were minmax scaled between 0 and 1. The models were tested with the remaining data from both isomers. The predictions are visualized in FIG. 3 along with RMSE values.

For unit gold, sulfur and hydrogen RMSEs are lower than 0.3eV/Å and predictions correlate well with DFT force norms as seen in FIG. 3 (b), (c), (e), (g), (h) and (j). The largest RMSE values belong to core gold and methyl carbon model. It is expected that gold core is difficult to handle as it undergoes great deal of structural changes. Against expectations, methyl carbon proved to be difficult for the EMLM. The chemical environment of the carbon is mostly determined by its neighboring hydrogen atoms and a sulfur, therefore the changes are quite small due to the rigid covalent bonds. A logical explanation would be that the carbon needs more exact SOAP description with high sensitivity to small changes. This could be achieved by using even smaller value for Gaussian broadening parameter σ_{SOAP}. However, the acquired accuracy is reasonable and can be used as a part of the simulations.

C. Full OAMLM force direction models

The OAMLM models were trained in same manner as EMLM but only 2500 data points were used in training and as references. The alignment of atomic environments is a relatively

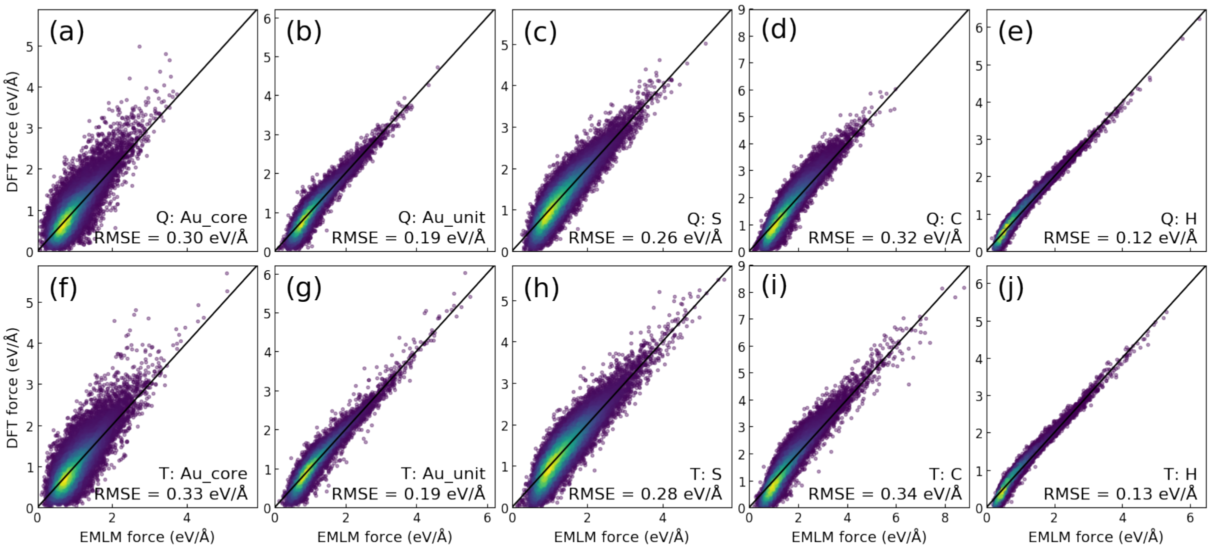


FIG. 3. Performance of different full EMLM models in comparison to DFT level forces. Panels (a)-(e) show the test results for the Q isomer and (f)-(j) for the T isomer. The tested element is written to the corner of every graph along with RMSE values. For hydrogen only third of the data points are plotted. The colors visualize the density of the points: yellow means dense region and purple sparse.

slow process, therefore having fewer references makes the model more feasible to use. During the predictions the weighting parameter in analytic loss function (15) was set as $\sigma_2 = 0.25$ for unit gold atoms and for everything else $\sigma_2 = 0.5$. As an error measurement we used weighted average of angles between predicted force directions and DFT level force vectors. As a weights, we used the squared norms of the DFT forces the same way as before.

The results are plotted in the FIG. 4. The effects of small forces are visible in all plots. When the norm of the force is small, the direction is extremely difficult to be estimated, which leads to the increased deviation close to the zero. The weighted averages show similar trends as the RMSEs in the case of force norms. Unit gold, sulfur and hydrogen are the easiest to handle as seen in FIG. 4 (b), (c), (e), (g), (h) and (j). From these three atom types the largest the largest weighted average angle 24.7° belongs to sulfur atoms of the isomer Q. The unit gold data contains some individual points, for which the angle is not as accurate as for the rest. This uncertainty is most likely caused by the inclusion of "half unit" gold atoms and possible classification difficulties. The classification rules mentioned in the section II C are approximate and especially the T isomer data might contain instances

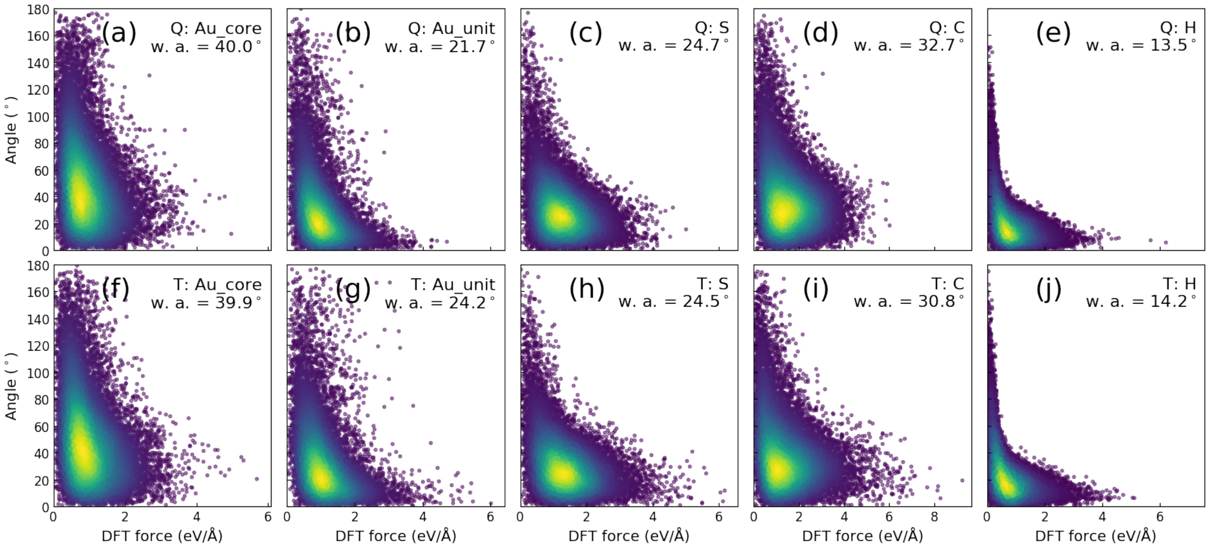


FIG. 4. Performance of the full OAMLM models using analytic loss function in equation (15). Vertical axises are the angle between the predicted direction and the DFT force vectors. Horizontal axises show corresponding DFT force norms. Panels (a)-(e) show the test results for the Q isomer and (f)-(j) for the T isomer. The tested element is written to the corner of every graph. For hydrogen only third of the data points are plotted. In the graphs, "w. a." stands for weighted average. The colors visualize the density of points: yellow means dense region and purple sparse.

where classification is not clear.

For core gold atoms in the FIG. 4 (a) and (f) the points are more spread than the other atom types. This hints that the alignment of the core environment is not straightforward, which leads into difficult estimation of the direction. However, OAMLM still manages to yield reasonable estimates even with highly complex alignment situations. For the methyl carbons in the FIG. 4 (d) and (i), the origin of the uncertainty is likely the same as in the case of force norm prediction. It needs more exact SOAP description with small gaussian broadening parameter σ_{SOAP}. Hydrogen atoms do not make extreme movements, therefore all their permutations yield very similar alignments, which are difficult to distinguish without highly specialized structural descriptors.

D. Application to structure optimization

As we now have a full force estimation method combined from EMLMs and OAMLMs, the next step is to apply it to the structure optimization with BFGS. In the first test we leave the complicated metallic core out and focus on covalently bound parts by optimizing gold-thiolate rings. The second case is to optimize a stretched protecting unit attached to the $Au_{38}(SCH_3)_{24}$ cluster. The third one is the most difficult test, where we use our model and BFGS to optimize snapshots from the original MD simulation trajectory.

1. Gold-thiolate rings

Testing the model with gold-thiolate rings is an interesting test case, because the model is not explicitly trained with them. In the MD trajectory of the T isomer there is an seven gold atom ring breaking out from the structure in the end [22] but there is no guarantee how much it has been sampled and the ring in MD is highly deformed. The starting structures were generated by making even geometric shapes, where sulfur atoms lie in the corners. Sulfur atoms were displace from the plane 1.0Å up and down in turns. We focus on the rings containing four, five or six gold atoms. These structures are shown in FIG. 5.

Structures were optimized by both DFT and ML model using BFGS algorithm. Optimization with DFT used the default 0.2 Å maximum step size of the ASE package. For ML forces the step size was set to half smaller value of 0.1 Å. ML-based optimization ran 200 optimization step, which was its maximum number of iterations. The stopping criterion was that if maximum force is ≤ 0.1 eV/Å, the optimization would stop. However, due to the uncertainty in the model optimizations did not reach this. After optimizations, potential energies were computed for ML optimization trajectories via single point DFT calculations.

The potential energies in FIG. 6 panels (a), (b) and (c) are decreasing during the optimization as supposed to. For four and five gold atom rings, the descending of the potential energy is effectively monotonous. With six gold atoms, the ML optimization initially manages to decrease the potential energy the same manner as before but after about 50 steps it adopts a geometry, which does not fully agree with DFT. The six gold atom ring contains more empty space in the middle of the ring than in the any configuration used to train the ML model, therefore it is expected that increasing the ring size increases the uncertainty of

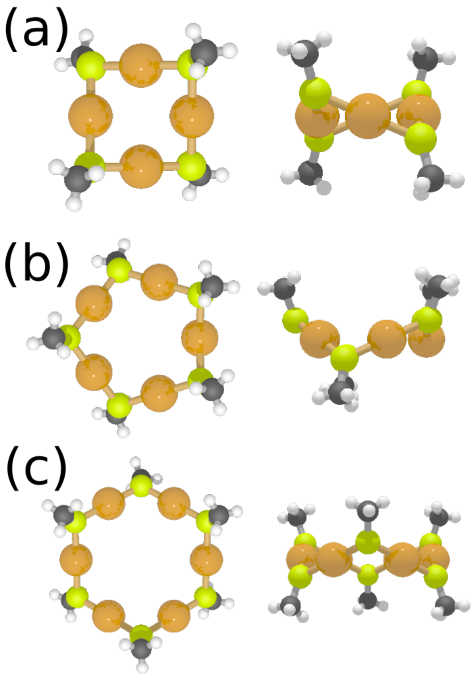


FIG. 5. Top and side views of the initial structures for (a) four, (b) five and (c) six gold atom gold-thiolate rings. Colors: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen.

the model.

The comparison of the structures from DFT and ML optimization reveals intriguing differences. The four gold ring configuration, into which DFT optimization converged, is just slightly twisted clockwise out of the plane as seen in FIG. 6 (d). However, ML optimization has been twisted on the opposite direction in FIG. 6 (g). Similar trend is also seen with five gold atom ring in FIG. 6 panels (e) and (h). For six gold atom ring, the twisting is not very clear in FIG. 6 panels (f) and (i). There the hexagonal ring shape has been deformed towards the triangle, which is likely caused by the method preference to produce 90° Au-S-Au angles locally as ML methods do not see the whole structure.

Due to the differences in the DFT and ML optimization results, we decided to optimize the final structures from the ML optimization with DFT. The results are shown in FIG. 6 panels (j), (k) and (l). It is surprising that in the case of four and six gold atom rings the potential energies of these newly optimized structures are slightly better than the ones from direct DFT optimization. The twisting has also been preserved, which indicates that the structural differences in plain DFT and ML optimizations are not defects but features of

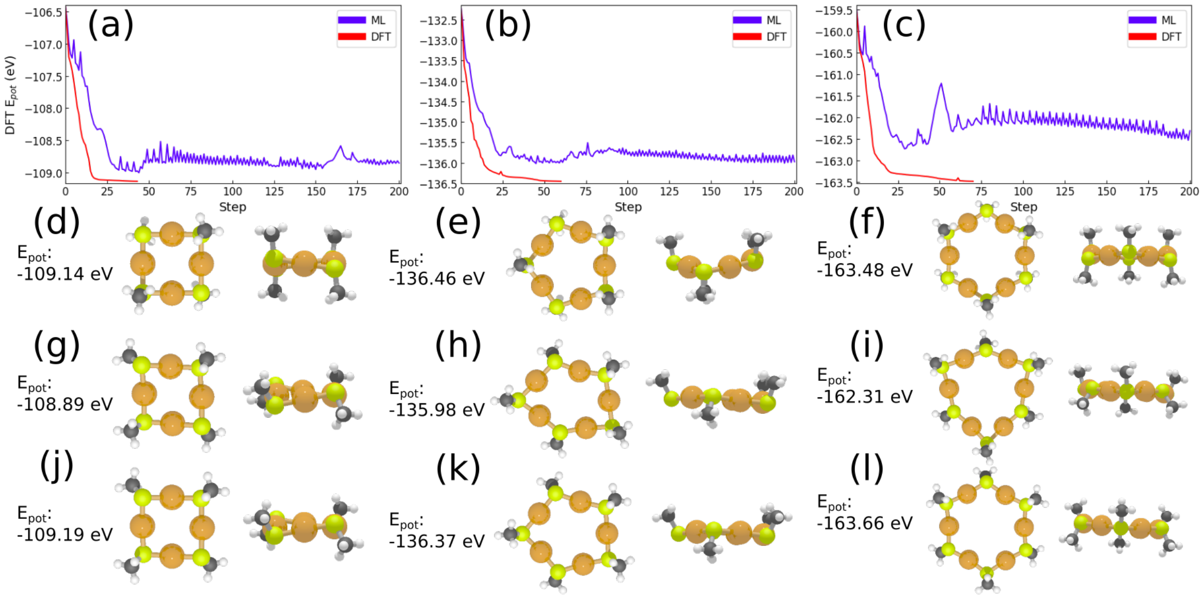


FIG. 6. (a)-(c) show the DFT calculated potential energy evolution during the DFT and ML BFGS optimization for four, five and six gold atom gold-thiolate rings respectively. (d)-(f) the final structures from the DFT optimization viewed from top and side. Correspondingly (g)-(i) are the final structures from ML optimization. Structures in (j)-(l) are DFT optimization results, which started from the corresponding ML optimization results. Colors: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen.

realistic local energy minima.

The structures optimized only with DFT settled to a local energy minimum close to the initial structures. ML method on the other hand passed this minimum and continued into another one resulting into an opposite twisting of the structure. It is likely that the firstly mentioned energy minimum is shallow compared to its surrounding potential energy landscape. The ML method either has not learned this kind of profile or the minimum was hidden by the uncertainty in the model. However, this behavior enabled the optimization to proceed close to an alternative energy minimum, which could possibly be even better. This demonstrates that our ML methodology can be utilized as a hybrid optimization tool, where ML executes coarse optimization and DFT is used in fine tuning.

2. *Partial optimization of the $Au_{38}(SCH_3)_{24}$ nanocluster*

The second test case is to optimize $\mathrm{Au}_{38}(\mathrm{SCH}_3)_{24}$ structures, which are otherwise DFT optimized except one long protecting unit is pulled outwards 2.0 Å. This is done for both isomers. As seen in the FIG. 7 (a) for Q isomer the pulled unit lies on the corner of the cylindrical shape and for T isomer the pulled unit is in the middle of the structure presented in 7 (b). As a comparison to the ML optimization, we optimized the structure also with DFT forces and BFGS. During the optimization process only atoms belonging to stretched unit were allowed to move and others were fixed, therefore there were four gold, three sulfur, three carbon and nine hydrogen atoms that are moving. Two of the gold atoms were classified as belonging to the unit and two to the core.

Different maximum step sizes of the ML BFGS algorithm were compared by calculating single point DFT potential energies and by comparing structures with root-mean squared deviation (RMSD). We use term RMSE to refer prediction error in the case of testing force norm prediction with EMLM and with term RMSD we refer to the structural difference of atomic configurations. They are essentially the same but with with terminology we want to distinguish that they are measuring two different kinds of differences. Here the RMSD is calculated between the final structure from the DFT level optimization and configurations of interest from ML optimization. Only moving atoms, except hydrogen atoms, are included into the RMSD calculation.

As the ML method has always some level of uncertainty in both force norms and directions, the maximum step size might affect the convergence. If for one element the force is overestimated, the optimization would scale all requested steps collectively letting the atom affected by the largest force be moved the most and the rest are moved just slightly. Hence, too large step size might lead to back and forth movement, when the atom with overestimated force overshoots and passes a minimum. A small step size reduces the possibility of overshooting and the BFGS approximation of Hessian matrix is updated with more modest rate than with a large maximum step size.

The potential energy comparison is shown in FIG. 8 (a) for Q isomer and (c) for T isomer. Some differences in the convergence and the fluctuation of the potential energy are observed between different step sizes. However, all curves have converged in the similar energy level and potential energy is decreasing with a good rate. In the FIG. 8 (a) maximum step size

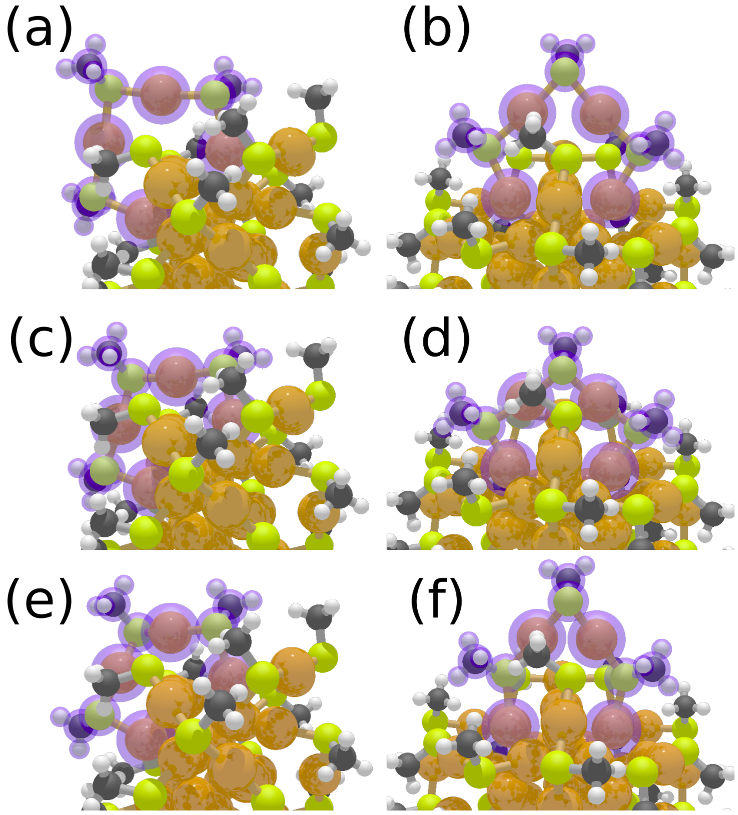


FIG. 7. (a) and (b) show the stretched protecting units the $Au_{38}(SCH_3)_{24}$ Q and T isomer respectively. (c) and (d) are DFT constrained optimization result starting from the structures (a) and (b). (e) and (f) are constrained ML optimized structures from 150th optimization step with 0.05Å maximum BFGS step size. During the optimization everything else is fixed expect the part highlighted with purple. Colors: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen.

0.05 Å is giving the most stable performance and it reaches the lowest energy value, even tough it is higher than what DFT optimization yields. The optimization for T isomer shows more fluctuation in FIG. 8 (c) and the energy differences between DFT and ML optimizations are larger than in the case of Q isomer.

By looking at the structures and comparing them with RMSD, we can get some insight about the behavior of the ML optimization, which are not visible in potential energy. For Q isomer, the RMSD evolution in FIG. 8 (b) indicates that ML optimizations with different maximum step sizes converge to somewhat different configurations. Maximum step size 0.05 Å manages to get closest to the DFT optimization results. This can also be seen in FIG. 7 (c) and (e), where the structures are visualized. They have very close resemblance.

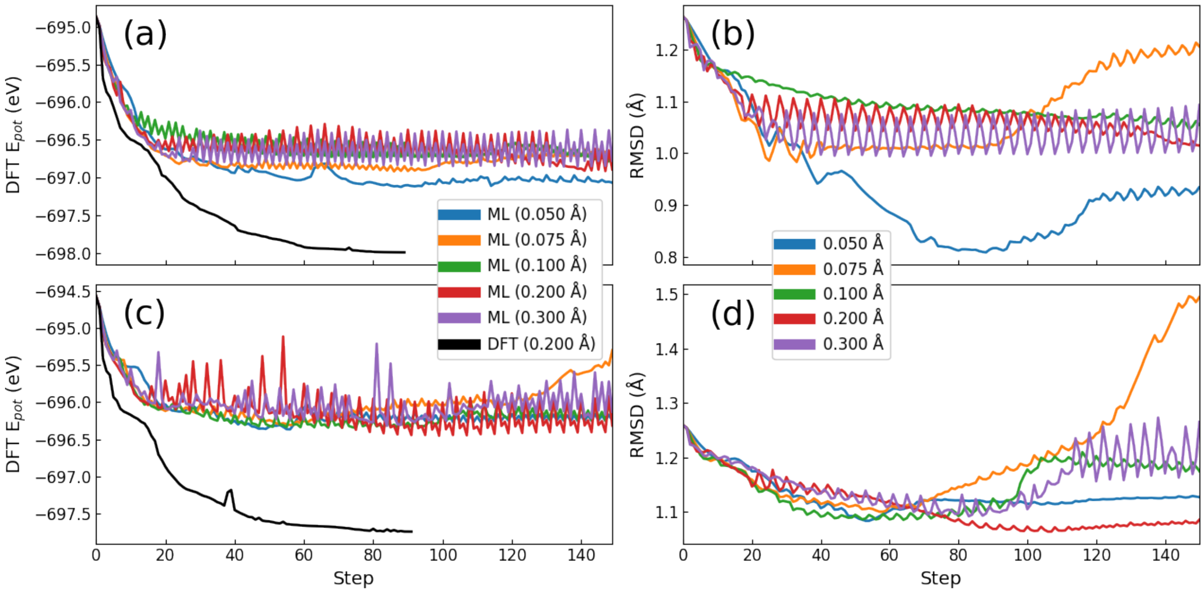


FIG. 8. The evolution of the potential energy and RMSD during the optimization with different BFGS maximum step sizes. (a) shows the potential energy evolution for Q isomer and (b) the RMSD compared to the DFT optimized structure. (c) and (d) are corresponding plots for T isomer.

The optimizations of the T isomers are seen to converge into very similar RMSD values in FIG. 8 (d). After about 80 optimization steps the differences start to emerge. This is caused mostly caused by the two core gold atoms. As seen in the FIG. 7 (d) and (f), during the ML optimization two core gold atoms are not placed as deep into the core as with DFT, which leaves protecting unit protruding from the cluster. As the convergence criterion is not reached and optimization continues, BFGS forces this unit to bend while trying to minimize the potential energy. However, even if DFT and ML optimizations lead to somewhat different structures, the potential energy is shown to be surprisingly stable.

3. *MD configurations*

The most challenging task is to optimize arbitrary configurations from the MD runs, which were also used to extract training and testing data. For Q isomer we used 1000th step and for T isomer 600th step from the corresponding trajectories. The structures are visualized in FIG. 9 panels (a) and (b).

The most uncertain part of our ML framework is the gold core, therefore we decided

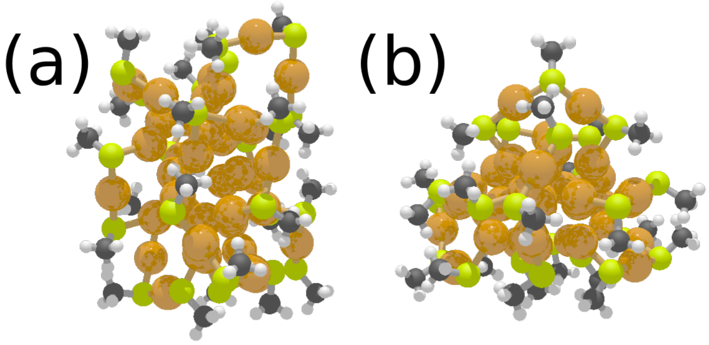


FIG. 9. (a) 1000th configuration of the $Au_{38}(SCH_3)_{24}$ Q isomer from the MD simulations from the reference [22]. (b) 600th configuration of the $Au_{38}(SCH_3)_{24}$ T isomer from the same source. Colors: orange, gold; yellow, sulfur; gray, carbon; white, hydrogen.

to run the optimization in parts to simplify the situation. First the outer layer containing unit gold, sulfur, carbon and hydrogen atoms is optimized 24 steps and core gold atoms are fixed. Next outer layer is fixed and core gold atoms are optimized 12 steps. This way the uncertainty inside the core does not affect directly the steps on outer layer and vice versa. The maximum step size was 0.05 Å as it was shown to result into stable optimizations in the previous section.

Another way that we used to minimize the uncertainty effects to the BFGS optimization, was resetting the Hessian matrix approximation. Here resetting means that the Hessian matrix approximation is returned to the initial value. Optimization of MD configurations drives the ML method to its limits, therefore there is a risk that the simulations reach regions where the reliability of the method is compromised. This can affect the performance of the BFGS algorithm, because the usage second order information via Hessian matrix approximation, makes it maximally affected by the noise and inaccuracies of the gradient. This is due to the ill-posedness of the noisy derivatives [56]. Hence, readjusting the optimization might help to cope with uncertainty. We used two different resetting schemes: conventional BFGS with no resetting and resetting after every 36 optimization steps (one round for both outer layer and core).

After the optimization was run with ML forces, potential energy values were computed via single point DFT calculations as before. The results for the Q isomer are shown in FIG. 10 (a) and for the T isomer in FIG. 10 (b). The optimization of the Q isomer shows almost monotonous decreasing of the potential energy. However, without resetting the Hessian matrix approximation the potential energy start a slight increase on the second round of the core optimization. Resetting seems to improve the optimization but it introduces fluctuation to the outer layer optimization.

The optimization of the T isomer configuration is again more unstable than Q isomer as expected. The first optimization round decreased the potential energy by about 0.5 eV but then the effects from the uncertainty accumulated into the Hessian matrix approximation start to emerge. This is seen as an increasing potential energy. Resetting the Hessian matrix minimizes the increase, but it introduces signifigant fluctuation to the outer layer optimization.

The results in 10 demonstrate the complexity of the optimization of the arbitrary $Au_{38}(SCH_3)_{24}$ configurations. However, our method combining EMLMs and OAMLMs manages to decrease the potential energy by about 1.0 eV for Q isomer and 0.5 eV for T isomer. Furthermore, tests show that the effect of uncertainty accumulated into Hessian matrix approximation could be reduced by resetting. This is valuable practical information, if one desires to use the method for real applications. Straightforward way to improve the optimization would be to add DFT-level optimization steps between the ML optimization rounds. If ML method is steered towards non-physical configurations because of the accumulated uncertainty or inputs outside the training region, the DFT optimization steps could help the overall process to converge towards a better configuration.

IV. CONCLUSIONS

In this study we applied a novel concept of ML forces to optimize chemically complex protected $Au_{38}(SCH_3)_{24}$ nanocluster and gold-thiolate rings. The methodology was based on distance-based ML methods. The prediction of the atomic forces was divided into two parts. The prediction of the norms was done with conventional EMLM method, and the estimation of the force directions used a newly developed OAMLM method. Different parameters were tested rigorously utilizing the two structural isomers of the $Au_{38}(SCH_3)_{24}$ nanocluster. First we tested the performance of the model by training it with the data from one isomer and then tested it with the other. After this, another training dataset was collected using both isomers and both norm and direction prediction methods were tested.

As an application of the ML method, we used a BFGS structure optimization algorithm to

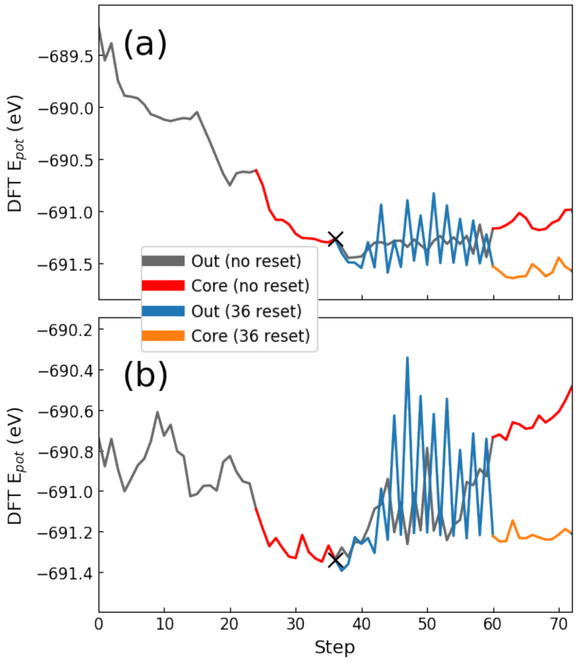


FIG. 10. Evolution of DFT potential energy during the ML optimization of MD snapshots for (a) Q isomer and (b) T isomer. Optimization is done in turns first optimizing 24 steps of protecting outer layer and the 12 steps of gold core. There are two different optimization approaches: normal BFGS and BFGS where Hessian matrix approximation is reset every 36 optimization step. Crosses on the curves show when the approximation of the Hessian matrix is reset.

utilize atomic forces estimated with EMLM and OAMLM. The optimization was first tested with gold-thiolate rings, which showed surprisingly good performance as these structures were not explicitly included in the training data. Here the method shows a great promise of generalizability. The second testing case was to optimize stretched protecting units on both isomers of the $Au_{38}(SCH_3)_{24}$. Especially the results of the isomer Q were in good agreement with the DFT. The greatest challenge was to optimize MD snapshots with ML forces with different approaches to BFGS. The method managed to reasonably reduce the potential energies of these systems. The same tests also demonstrated that resetting of the Hessian matrix approximation is an effective approach to minimize the uncertainty effects.

Overall, the results are promising and suggest that the method could be useful for hybrid optimization method, where coarse optimization is done with ML and fine tuning with DFT. This approach already was already briefly shown to work for gold-thiolate rings. Furthermore, the method managed to handle $Au_{38}(SCH_3)_{24}$ nanocluster, which is an encouraging result suggesting that our methodology could be utilized on optimization of complex nanostructures.

ACKNOWLEDGMENTS

This work was supported by Academy of Finland through the AIPSE research program with grant 315549 to H.H. and 315550 to T.K., through the Universities Profiling Actions with grant 311877 to T.K. This work was also supported by "Antti ja Jenny Wihurin rahasto" via personal funding to A.P. ML computations were done at the FCCI node in the University of Jyväskylä (persistent identifier: urn:nbn:fi:research-infras-2016072533) and DFT computations at the CSC supercomputing center in Finland. We acknowledge J. Linja, J. Hämäläinen and P. Nieminen for numerous discussions on ML methods. J. Hämäläinen provided the basis for EMLM and RS-maximin codes.

V. ASSOCIATED CONTENT

A. Supplemental Material

The Supplemental Material is available free of charge at [URL will be inserted by publisher]. It contains detailed results for the SOAP parameter testing with EMLM (figures S1 – S20) and OAMLM (figures S21 – S24).

B. Code and its availability

The whole method is written in Python 3.6 and it relies on Numpy [57], Scikit-learn [58], Atomic Simulation Environment [53], DScribe [40] and Scipy [59] packages. The parallelization of the testing and training of the methods and the BFGS optimization are done via mpi4py package [60–63]. The code, optimization data and complete parameter test visualizations are available at Gitlab `https://gitlab.jyu.fi/aneepihl/oamlm_forces.git`.

VI. AUTHOR INFORMATION

A. Corresponding Author

Hannu Häkkinen – Departments of Physics and Chemistry, Nanoscience Center, University of Jyväskylä, FI-40014 Jyväskylä, Finland; Email: hannu.j.hakkinen@jyu.fi; Orcid: 0000-0002-8558-5436

B. Notes

The authors declare no competing financial interest.

[1] T. Tsukuda and H. Häkkinen, *Protected metal clusters: from fundamentals to applications* (Elsevier, Amsterdam, Netherlands, 2015).

[2] S. Malola and H. Häkkinen, Prospects and challenges for computer simulations of monolayer-protected metal clusters, Nat. Commun. **12**, 2197 (2021).

[3] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, Phys. Rev. **136**, B864 (1964).

[4] G.-T. Bae and C. M. Aikens, Improved reaxff force field parameters for au-s-c-h systems, J. Phys. Chem. A **117**, 10438 (2013).

[5] E. Pohjolainen, X. Chen, S. Malola, G. Groenhof, and H. Häkkinen, A unified amber-compatible molecular mechanics force field for thiolate-protected gold nanoclusters, J. Chem. Theory Comput. **12**, 1342 (2016).

[6] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, Machine learning for molecular simulation, Annual Review of Physical Chemistry **71**, 361 (2020).

[7] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, Machine learning force fields, Chemical Reviews **121**, 10142 (2021).

[8] P. Friederich, F. Häse, J. Proppe, and A. Aspuru-Guzik, Machine-learned potentials for next-generation matter simulations, Nat. Mater. **20**, 750–761 (2021).

[9] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, Recent advances and applications of machine learning in solid-state materials science, npj Comput. Mater. **5**, 83 (2019).

[10] G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, and A. Fazzio, From dft to machine learning: recent approaches to materials science–a review, JPhys Materials **2**, 032001 (2019).

[11] T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa, and K.-i. Shimizu, Machine learning for catalysis informatics: Recent applications and prospects, ACS Cat. **10**, 2260 (2020).

[12] M. S. Jørgensen, H. L. Mortensen, S. A. Meldgaard, E. L. Kolsbjerg, T. L. Jacobsen, K. H. Sørensen, and B. Hammer, Atomistic structure learning, J. Chem. Phys. **151**, 054111 (2019).

[13] S. A. Meldgaard, H. L. Mortensen, M. S. Jørgensen1, and B. Hammer, Structure prediction of surface reconstructions by deep reinforcement learning, J. Phys. Condens. Mat. **32**, 404005 (2020).

[14] M.-P. V. Christiansen, H. L. Mortensen, S. A. Meldgaard, and B. Hammer, Gaussian representation for image recognition and reinforcement learning of atomistic structure, J. Chem. Phys. **153**, 044107 (2020).

[15] J. Li, T. Chen, K. Lim, L. Chen, S. A. Khan, J. Xie, and X. Wang, Deep learning accelerated gold nanocluster synthesis, Adv. Intell. Syst. **1**, 1900029 (2019).

[16] S. M. Copp, S. M. Swasey, A. Gorovits, P. Bogdanov, and E. G. Gwinn, General approach for machine learning-aided design of dna-stabilized silver clusters, Chem. Mater. **32**, 430 (2020).

[17] S. Malola, P. Nieminen, A. Pihlajamäki, J. Hämäläinen, T. Kärkkäinen, and H. Häkkinen, A method for structure prediction of metal-ligand interfaces of hybrid nanoparticles, Nat. Commun. **10**, 3973 (2019).

[18] A. Pihlajamäki, J. Hämäläinen, J. Linja, P. Nieminen, S. Malola, T. Kärkkäinen, and H. Häkkinen, Monte carlo simulations of $Au_{38}(SCH_3)_{24}$ nanocluster using distance-based machine learning methods, J. Phys. Chem. A **124**, 4827 (2020).

[19] H. Qian, W. T. Eckenhoff, Y. Zhu, T. Pintauer, and R. Jin, Total structure determination of thiolate-protected au38 nanoparticles, J. Am. Chem. Soc. **132**, 8280 (2010).

[20] S. Tian, Y.-Z. Li, M.-B. Li, J. Yuan, J. Yang, Z. Wu, and R. Jin, Structural isomerism in gold nanoparticles revealed by x-ray crystallography, Nat. Commun. **6**, 8667 (2015).

[21] H. Häkkinen, M. Walter, and H. Grönbeck, Divide and Protect: Capping Gold Nanoclusters with Molecular Gold–Thiolate Rings, J. Phys. Chem. B **110**, 9927 (2006).

[22] R. Juarez-Mosqueda, S. Malola, and H. Häkkinen, Ab initio molecular dynamics studies of $Au_{38}(SR)_{24}$ isomers under heating, Eur. Phys. J. D. **73**, 62 (2019).

[23] M. G. Taylor and G. Mpourmpakis, Thermodynamic stability of ligand-protected metal nanoclusters, Nat. Commun. **8**, 15988 (2017).

[24] H. Hellman, Einführung in die quantenchemie, Franz Deuticke, Leipzig **285** (1937).

[25] R. P. Feynman, Forces in molecules, Phys. Rev. **56**, 340 (1939).

[26] A. Fabrizio, A. Grisafi, B. Meyer, M. Ceriotti, and C. Corminboeuf, Electron density learning of non-covalent systems, Chem. Sci. **10**, 9424 (2019).

[27] A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, and M. Ceriotti, Transferable machine-learning model of the electron density, ACS Cent. Sci. **5**, 57 (2019).

[28] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, Quantum-chemical insights from deep tensor neural networks, Nat. Commun. **8**, 13890 (2017).

[29] X. Chen, M. S. Jørgensen, J. Li, and B. Hammer, Atomic energies from a convolutional neural network, J. Chem. Theory Comput. **14**, 3933 (2018).

[30] V. Botu and R. Ramprasad, Learning scheme to predict atomic forces and accelerate materials simulations, Phys. Rev. B **92**, 094306 (2015).

[31] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, Machine learning force fields: Construction, validation, and outlook, J. Phys. Chem. C **121**, 511 (2017).

[32] P. Pattnaik, S. Raghunathan, T. Kalluri, P. Bhimalapuram, C. V. Jawahar, and U. D. Priyakumar, Machine learning for accurate force calculations in molecular dynamics simulations, J. Phys. Chem. A **124**, 6954 (2020).

[33] A. P. Bartók, R. Kondor, and G. Csányi, On representing chemical environments, Phys. Rev. B **87**, 10.1103/PhysRevB.87.184115 (2013).

[34] J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, J. Chem. Phys. **134**, 074106 (2011).

[35] H. Huo and M. Rupp, Unified representation of molecules and crystals for machine learning, (2017), arXiv:1704.06439v3 [physics.chem-ph].

[36] M. J. Hostetler, J. E. Wingate, C.-J. Zhong, J. E. Harris, R. W. Vachet, M. R. Clark, J. D. Londono, S. J. Green, J. J. Stokes, G. D. Wignall, G. L. Glish, M. D. Porter, N. D. Evans, and R. W. Murray, Alkanethiolate gold cluster molecules with core diameters from 1.5 to 5.2 nm: Core and monolayer properties as a function of core size, Langmuir **14**, 17 (1998).

[37] S. Chen, A. C. Templeton, and R. W. Murray, Monolayer-protected cluster growth dynamics, Langmuir **16**, 3543 (2000).

[38] M. K. Corbierre and R. B. Lennox, Preparation of thiol-capped gold nanoparticles by chemical reduction of soluble au(i)-thiolates, Chem. Mater. **17**, 5691 (2005).

[39] H. Grönbeck, M. Walter, and H. Häkkinen, Theoretical characterization of cyclic thiolated gold clusters, J. Am. Chem. Soc. **128**, 10268–10275 (2006).

[40] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, Dscribe: Library of descriptors for machine learning in materials science, Comput. Phys. Commun. **247**, 106949 (2020).

[41] A. H. de Souza Júnior, F. Corona, G. A. Barreto, Y. Miche, and A. Lendasse, Minimal learning machine: A novel supervised distance-based approach for regression and classification, Neurocomputing **164**, 34 (2015).

[42] T. Kärkkäinen, Extreme minimal learning machine: Ridge regression with distance-based basis, Neurocomputing **342**, 33 (2019).

[43] J. Linja, J. Hämäläinen, P. Nieminen, and T. Kärkkäinen, Do randomized algorithms improve the efficiency of minimal learning machine?, Mach. Learn. Knowl. Extr. **2**, 533 (2020).

[44] J. Hämäläinen, A. S. C. Alencar, T. Kärkkäinen, C. L. C. Mattos, A. H. Souza Júnior, and J. P. P. Gomes, Minimal learning machine: Theoretical results and clustering-based reference point selection, J. Mach. Learn. Res. **21**, 1 (2020).

[45] A. Pihlajamäki, J. Linja, J. Hämäläinen, P. Nieminen, S. Malola, T. Kärkkäinen, and H. Häkkinen, Orientation adaptive minimal learning machine for directions of atomic forces, in *ESANN 2021: Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning Online event* (2021) pp. 529–534.

[46] K. P. Murphy, *Machine learning: A probabilistic perspective* (MIT Press, Cambridge, Massachusetts, 2012).

[47] W. Navidi, W. S. M. Jr., and W. Hereman, Statistical methods in surveying by trilateration, Comput. Stat. Data Anal. **27**, 209 (1998).

[48] K. S. Arun, T. S. Huang, and S. D. Blostein, Least-squares fitting of two 3-d point sets, IEEE T. Pattern Anal. **PAMI-9**, 698 (1987).

[49] C. G. Broyden, The convergence of a class of double-rank minimization algorithms 1. general considerations, IMA Journal of Applied Mathematics **6**, 76 (1970).

[50] R. Fletcher, A new approach to variable metric algorithms, The Computer Journal **13**, 317 (1970).

[51] D. Goldfarb, A family of variable-metric methods derived by variational means, Math. Comp. **24**, 23 (1970).

[52] D. F. Shanno, Conditioning of quasi-newton methods for function minimization, Math. Comp. **24**, 647 (1970).

[53] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, The atomic simulation environment—a python library for working with atoms, J. Phys. Condens. Mat. **29**, 273002 (2017).

[54] J. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Dułak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. A. Hansen, H. H. Kristoffersen, M. Kuisma, A. H. Larsen, L. Lehtovaara, M. Ljungberg, O. Lopez-Acevedo, P. G. Moses, J. Ojanen, T. Olsen, V. Petzold, N. A. Romero, J. Stausholm-Møller, M. Strange, G. A. Tritsaris, M. Vanin, M. Walter, B. Hammer, H. Häkkinen, G. K. H. Madsen, R. M. Nieminen, J. K. Nørskov, M. Puska, T. T. Rantala, J. Schøtz, K. S. Thygesen, and K. W. Jacobsen, Electronic structure calculations with gpaw: a real-space implementation of the projector augmented-wave method, J. Phys.: Condens. Matter **22**, 253202 (2010).

[55] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. **77**, 3865 (1996).

[56] Z. Wang, H. Wang, and S. Qiu, A new method for numerical differentiation based on direct and inverse problems of partial differential equations, Appl. Math. Lett. **43**, 61 (2015).

[57] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, Array programming with numpy, Nature **585**, 357–362 (2020).

[58] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res.

12, 2825 (2011).

[59] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İlhan Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . Contributors, Scipy 1.0: fundamental algorithms for scientific computing in python, Nat. Methods **17**, 261–272 (2020).

[60] L. Dalcín, R. Paz, and M. Storti, Mpi for python, J. Parallel Distr. Com. **65**, 1108 (2005).

[61] L. Dalcín, R. Paz, M. Storti, and J. D'Elía, Mpi for python: Performance improvements and mpi-2 extensions, J. Parallel Distr. Com. **68**, 655 (2008).

[62] L. D. Dalcín, R. R. Paz, P. A. Kler, and A. Cosimo, Parallel distributed computing using python, Adv. Water Resour. **34**, 1124 (2011), new Computational Methods and Software Tools.

[63] L. Dalcín and Y.-L. L. Fang, mpi4py: Status update after 12 years of development, Comput. Sci. Eng. **23**, 47 (2021).

Supplemental Material for "Orientation Adaptive Minimal Learning Machine: Application to Thiolate-Protected Gold Nanoclusters and Gold-Thiolate Rings"

Antti Pihlajamäki and Sami Malola
Department of Physics, Nanoscience Center,
University of Jyväskylä, FI-40014 Jyväskylä, Finland

Tommi Kärkkäinen
Faculty of Information Technology, University of Jyväskylä, FI-40014 Jyväskylä, Finland

Hannu Häkkinen
Department of Physics, Nanoscience Center,
University of Jyväskylä, FI-40014 Jyväskylä, Finland
Department of Chemistry, Nanoscience Center,
University of Jyväskylä, FI-40014 Jyväskylä, Finland[*]

(Dated: 17.03.2022)

I. SOAP PARAMETER TESTING VIA FORCE NORMS

The force norms were predicted separately for every element using EMLM. With every parameter set 2500 points were selected using RS-maximin method [1] from both structural isomers of the $Au_{38}(SCH_3)_{24}$ labeled as Q and T [2, 3]. These sets were used as training data, which was also saved into the model as reference data. Models were trained with training data from a single isomer only and then tested with all data from the another one without selection. Tested SOAP parameters were $n_{\max} \in [2, 7]$, $l_{\max} \in [0, 4]$, $r_{cut} \in \{4.0, 5.0\}$ Å and $\sigma_{SOAP} \in \{1.0, 0.75, 0.5, 0.25\}$ Å in the case of sulfur, carbon and hydrogen. For unit and core gold atoms tests were the same except only $\sigma_{SOAP} = 0.25$ Å sets were tested. In this Supplementary Information document we show all test with $\sigma_{SOAP} = 0.25$ Å and $r_{cut} = 4.0$ Å. For the sulfur tests are visualized in FIG. S1-S4, for carbon in FIG. S5-S8, for hydrogen in FIG. S9-S12, for unit gold in FIG. S13-S16 and for core gold in FIG. S17-S20 The complete tests are available at (`https://gitlab.jyu.fi/aneepihl/oamlm_forces.git`).

* hannu.j.hakkinen@jyu.fi

A. Sulfur

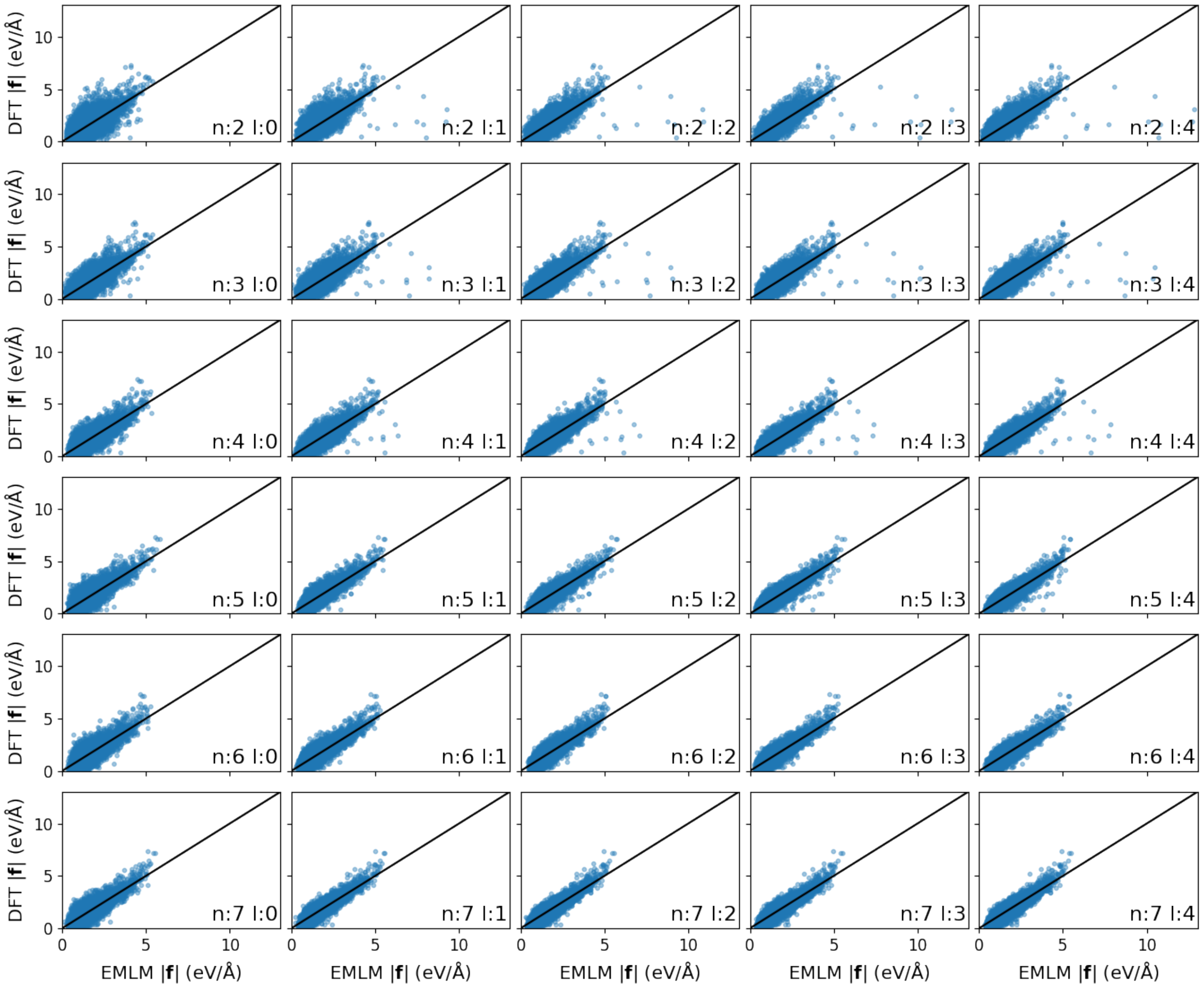


FIG. S1. EMLM force norm predictions are compared to the corresponding DFT values in the case of sulfur. The models were trained with Q isomer data and tested with T. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

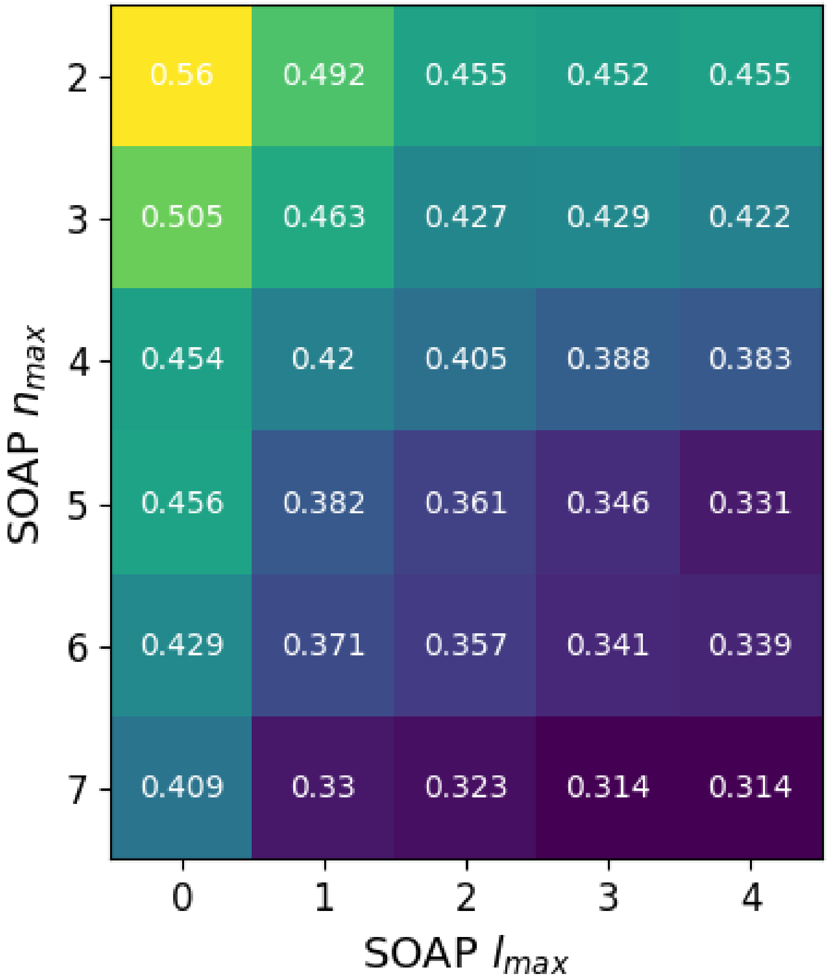


FIG. S2. RMSE values for the testing of EMLM models in FIG. S1. The models were trained with Q isomer data and tested with T. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

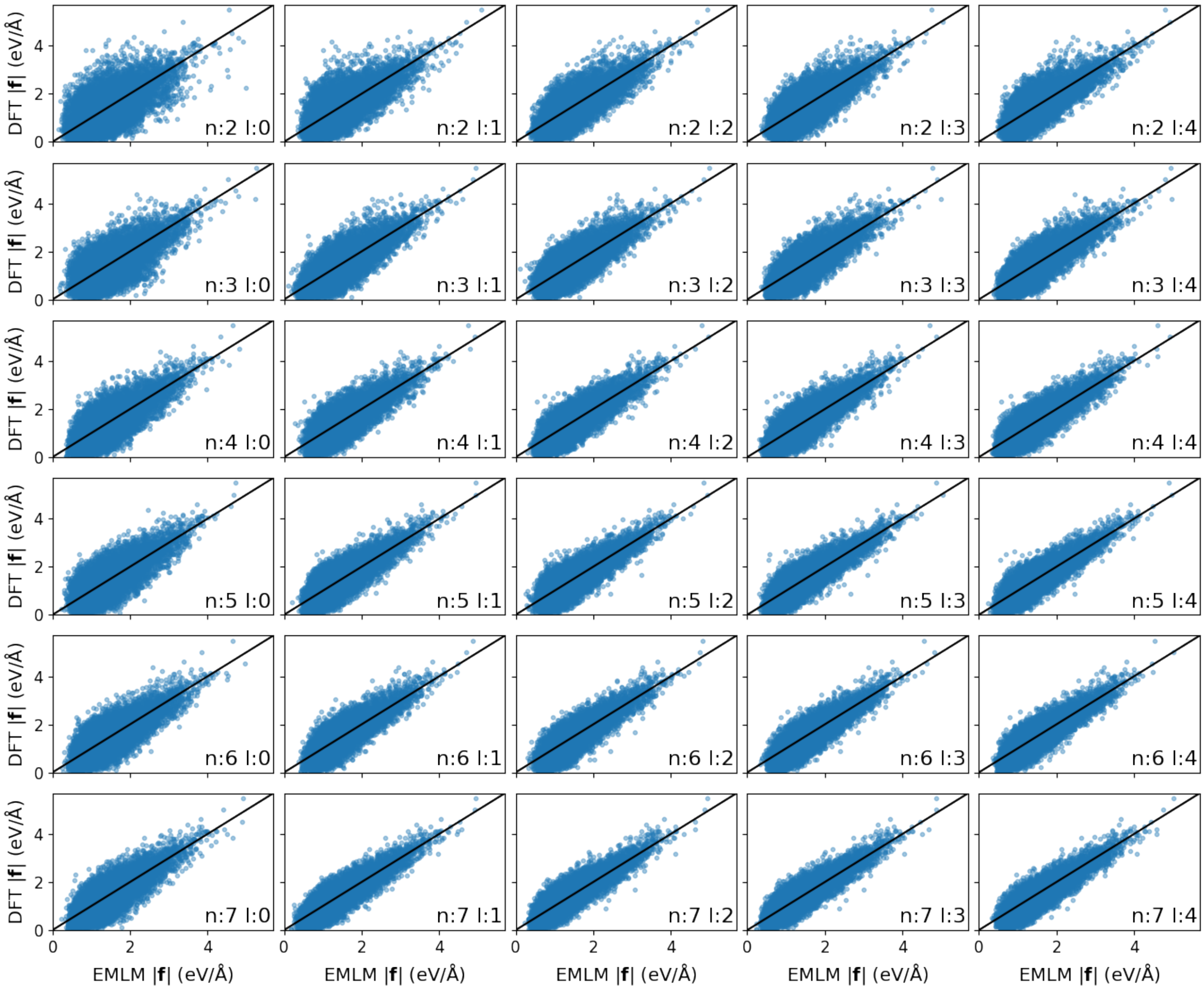


FIG. S3. EMLM force norm predictions are compared to the corresponding DFT values in the case of sulfur. The models were trained with T isomer data and tested with Q. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

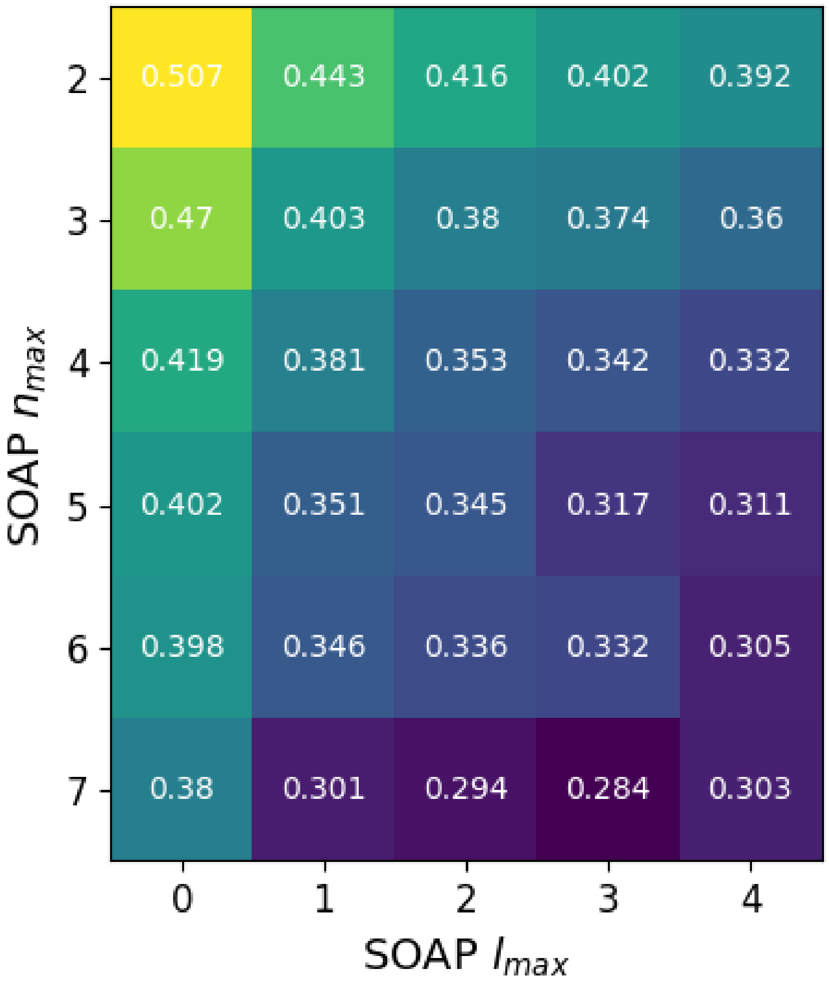


FIG. S4. RMSE values for the testing of EMLM models in FIG. S3. The models were trained with T isomer data and tested with Q. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

B. Carbon

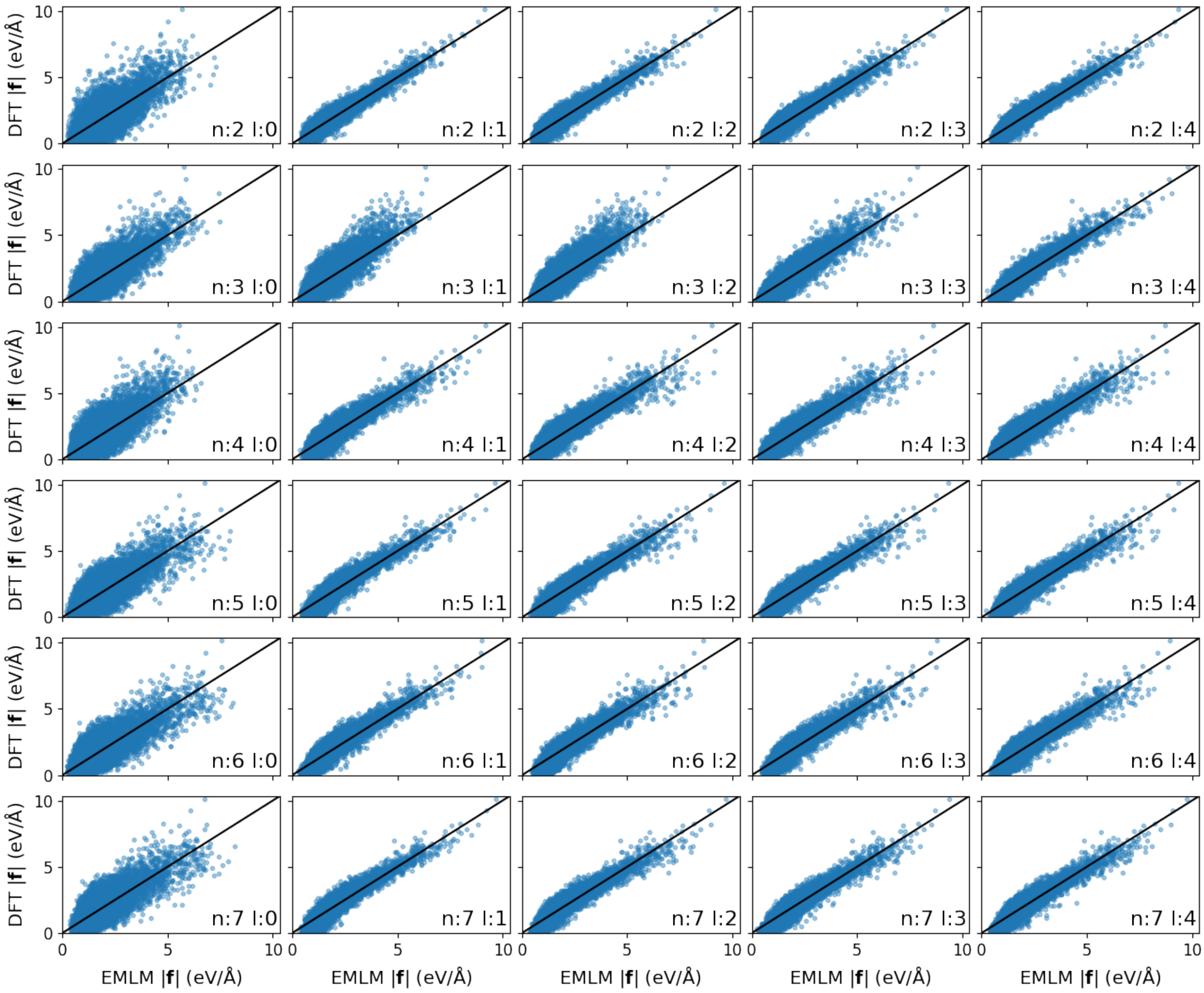


FIG. S5. EMLM force norm predictions are compared to the corresponding DFT values in the case of carbon. The models were trained with Q isomer data and tested with T. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

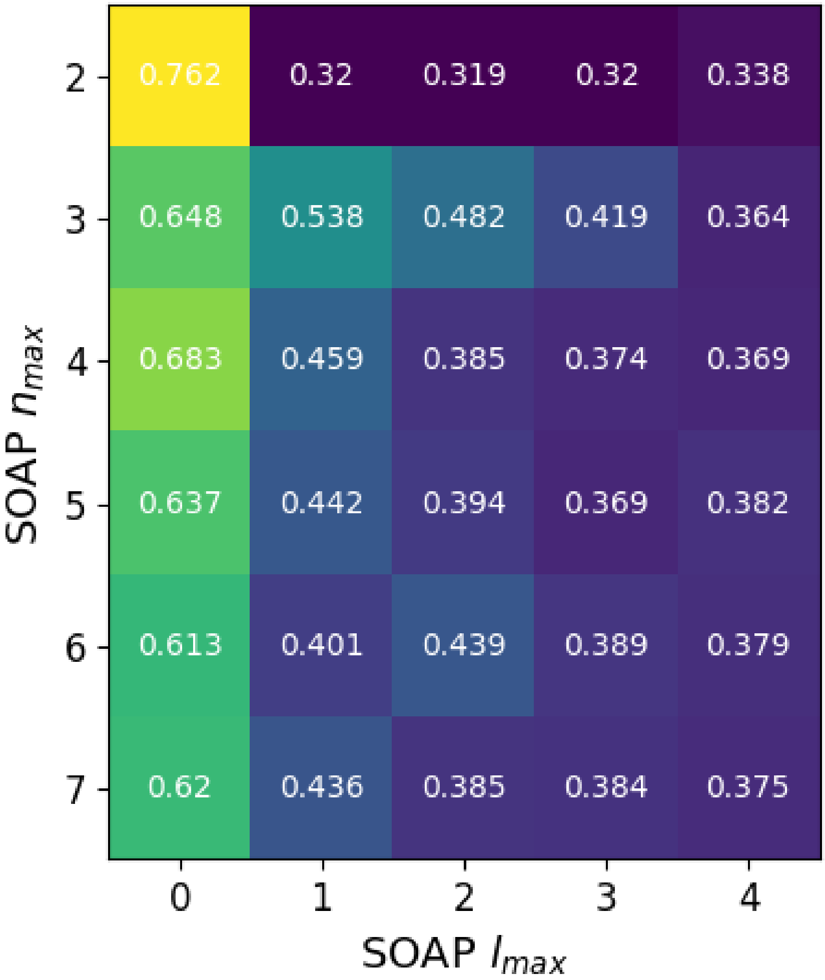


FIG. S6. RMSE values for the testing of EMLM models in FIG. S5. The models were trained with Q isomer data and tested with T. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

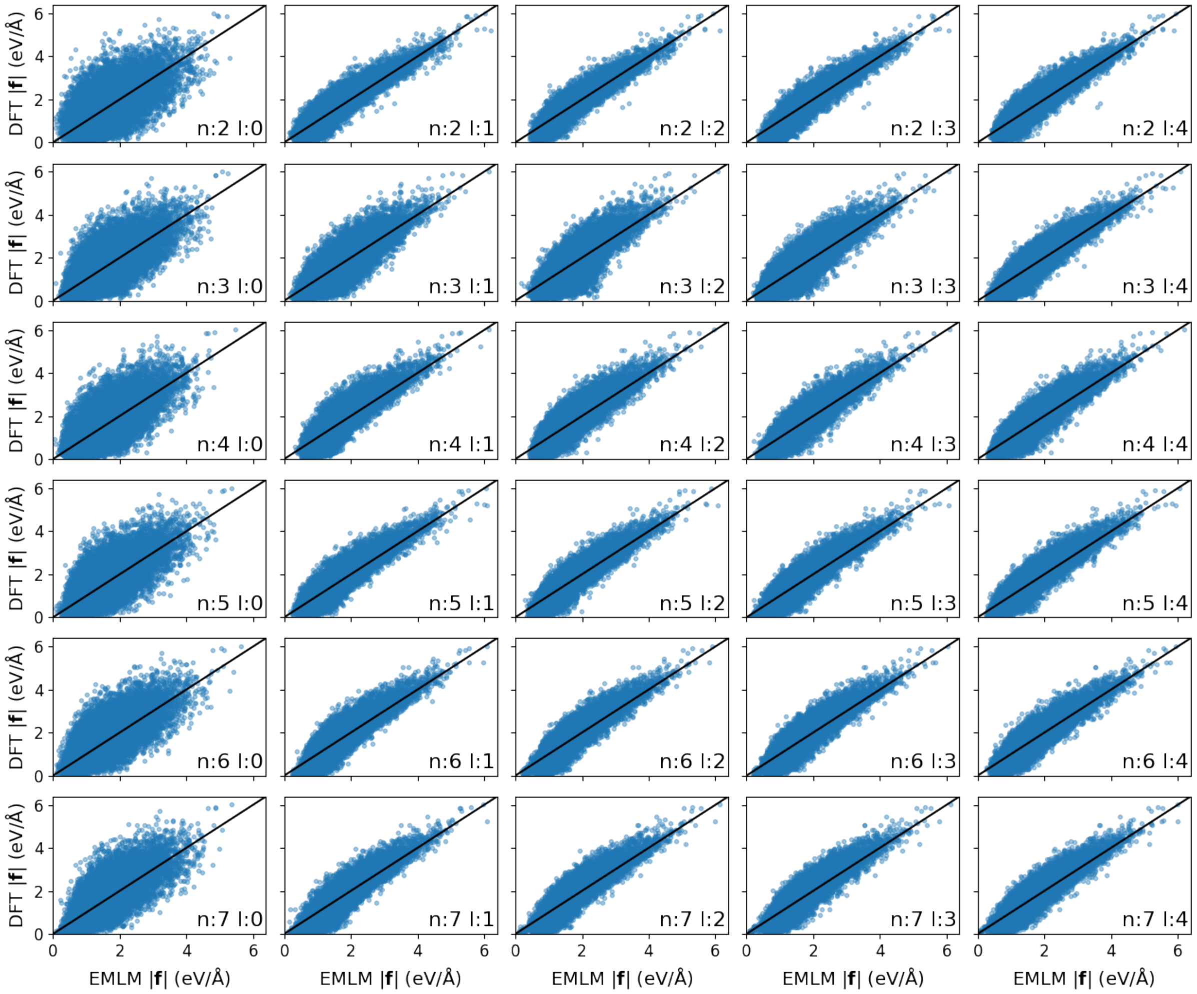


FIG. S7. EMLM force norm predictions are compared to the corresponding DFT values in the case of carbon. The models were trained with T isomer data and tested with Q. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

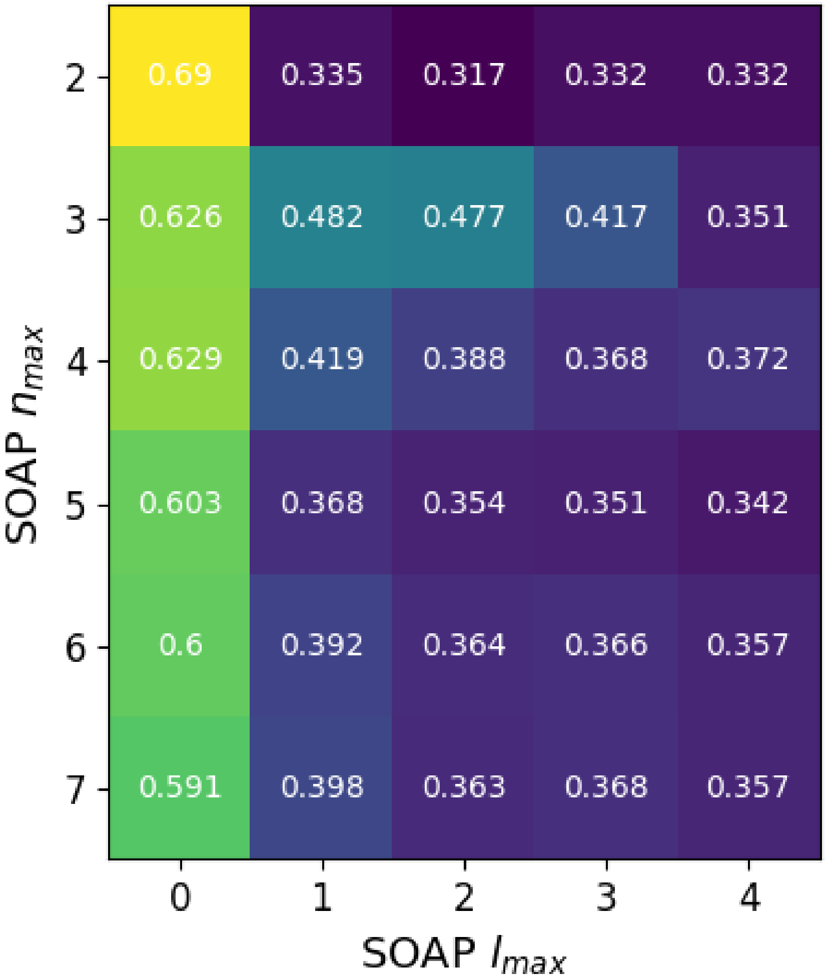


FIG. S8. RMSE values for the testing of EMLM models in FIG. S7. The models were trained with T isomer data and tested with Q. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

C. Hydrogen

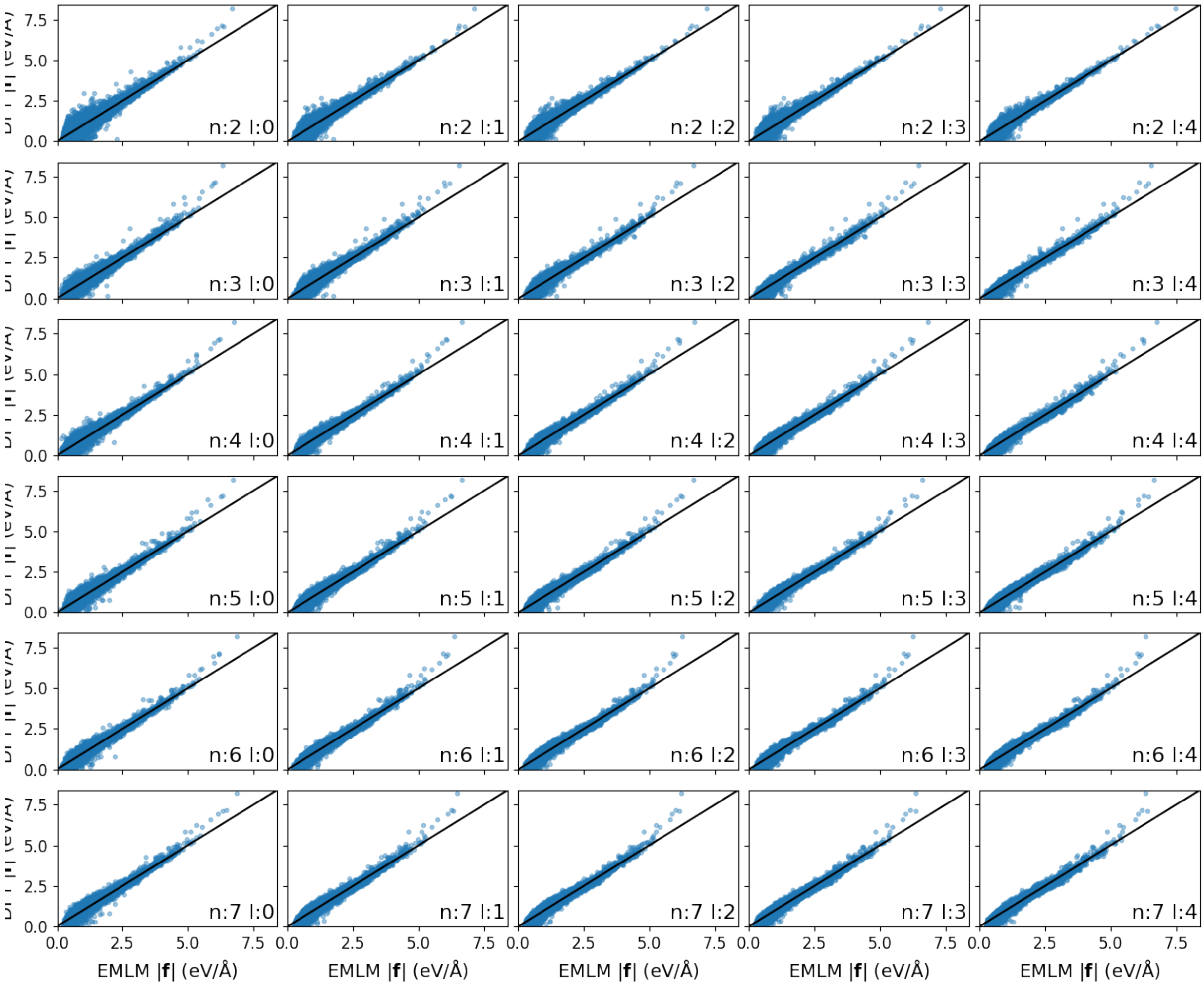


FIG. S9. EMLM force norm predictions are compared to the corresponding DFT values in the case of hydrogen. The models were trained with Q isomer data and tested with T. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

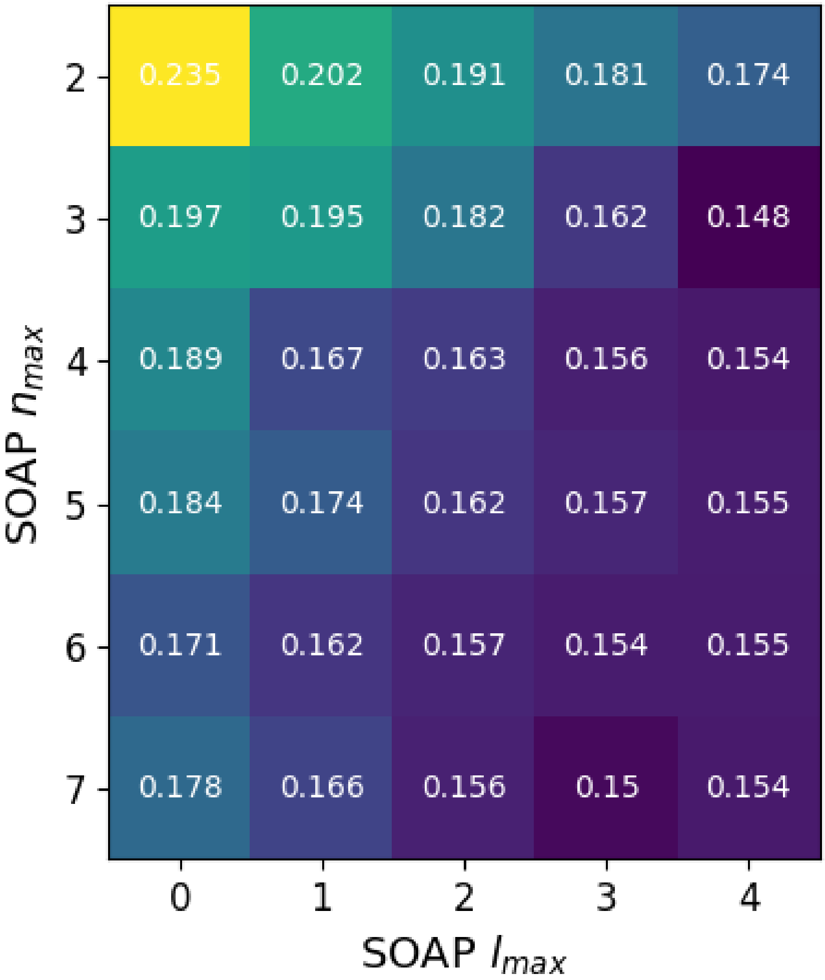


FIG. S10. RMSE values for the testing of EMLM models in FIG. S9. The models were trained with Q isomer data and tested with T. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

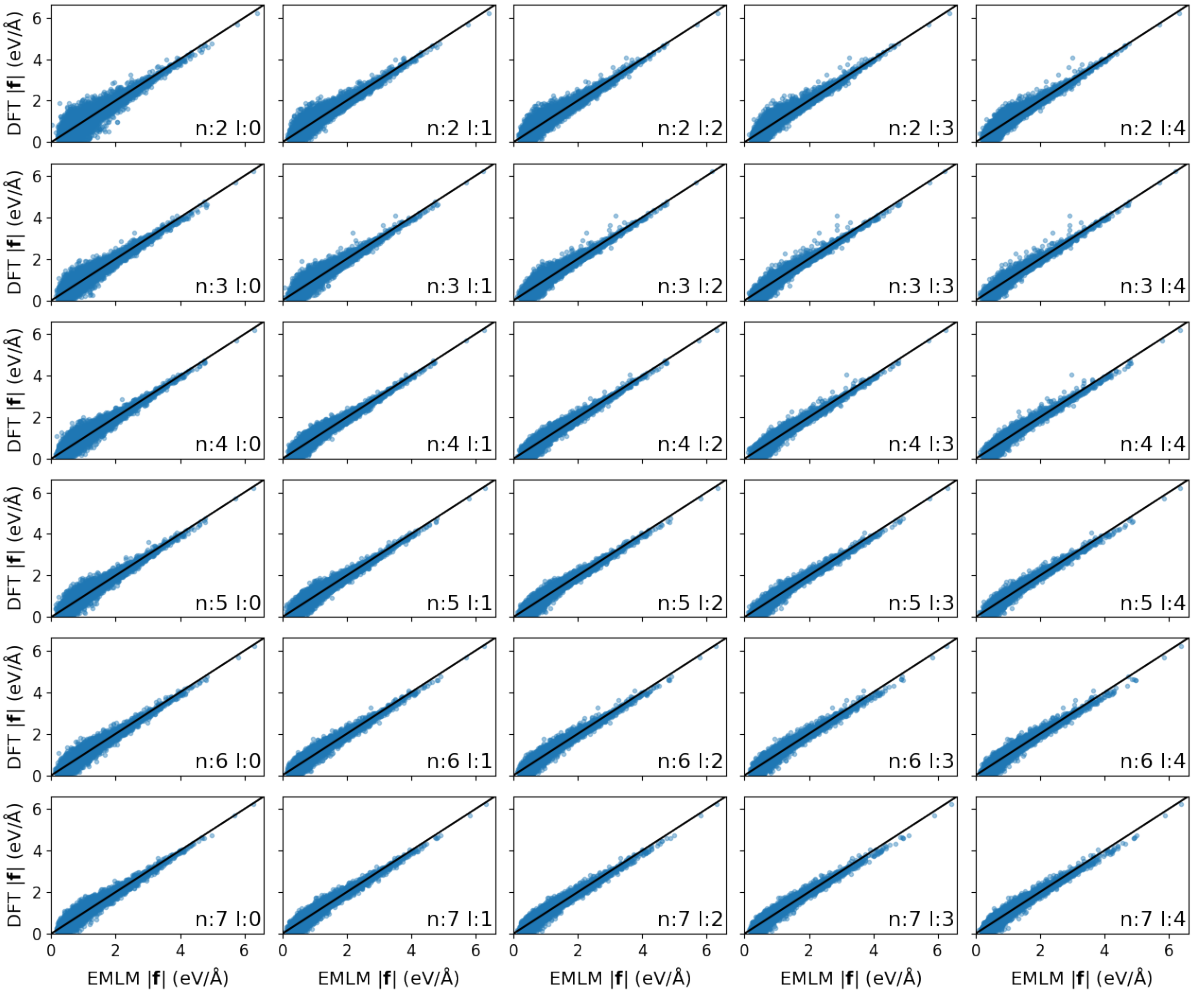


FIG. S11. EMLM force norm predictions are compared to the corresponding DFT values in the case of hydrogen. The models were trained with T isomer data and tested with Q. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

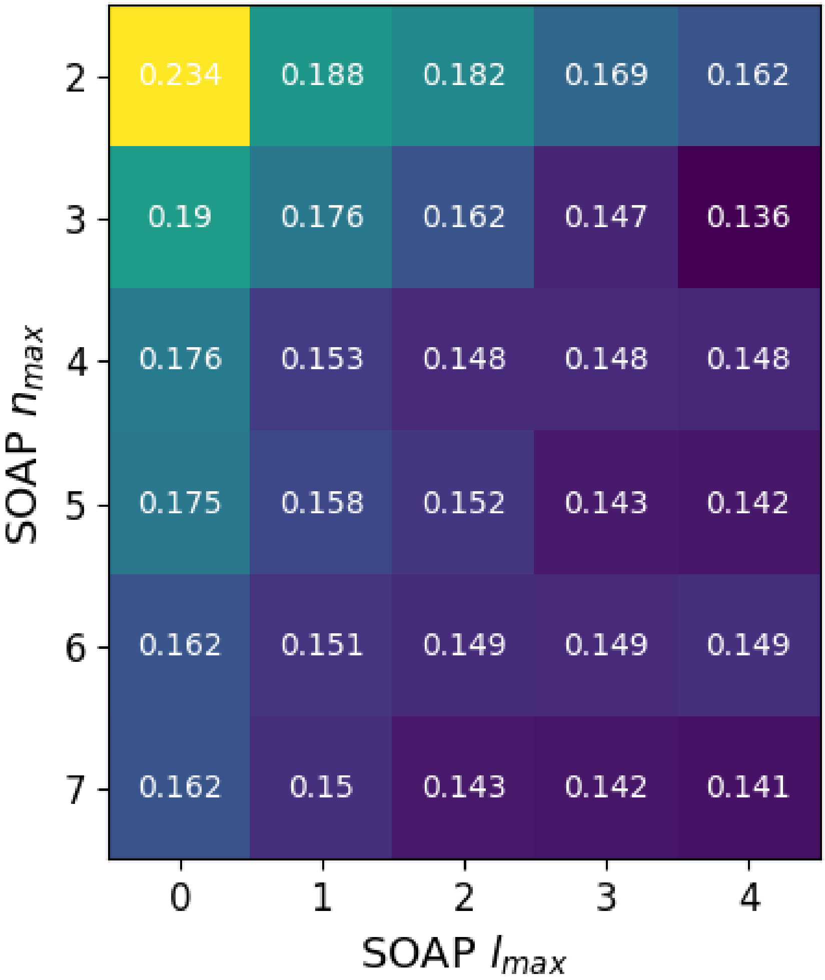


FIG. S12. RMSE values for the testing of EMLM models in FIG. S11. The models were trained with T isomer data and tested with Q. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

D. Unit gold

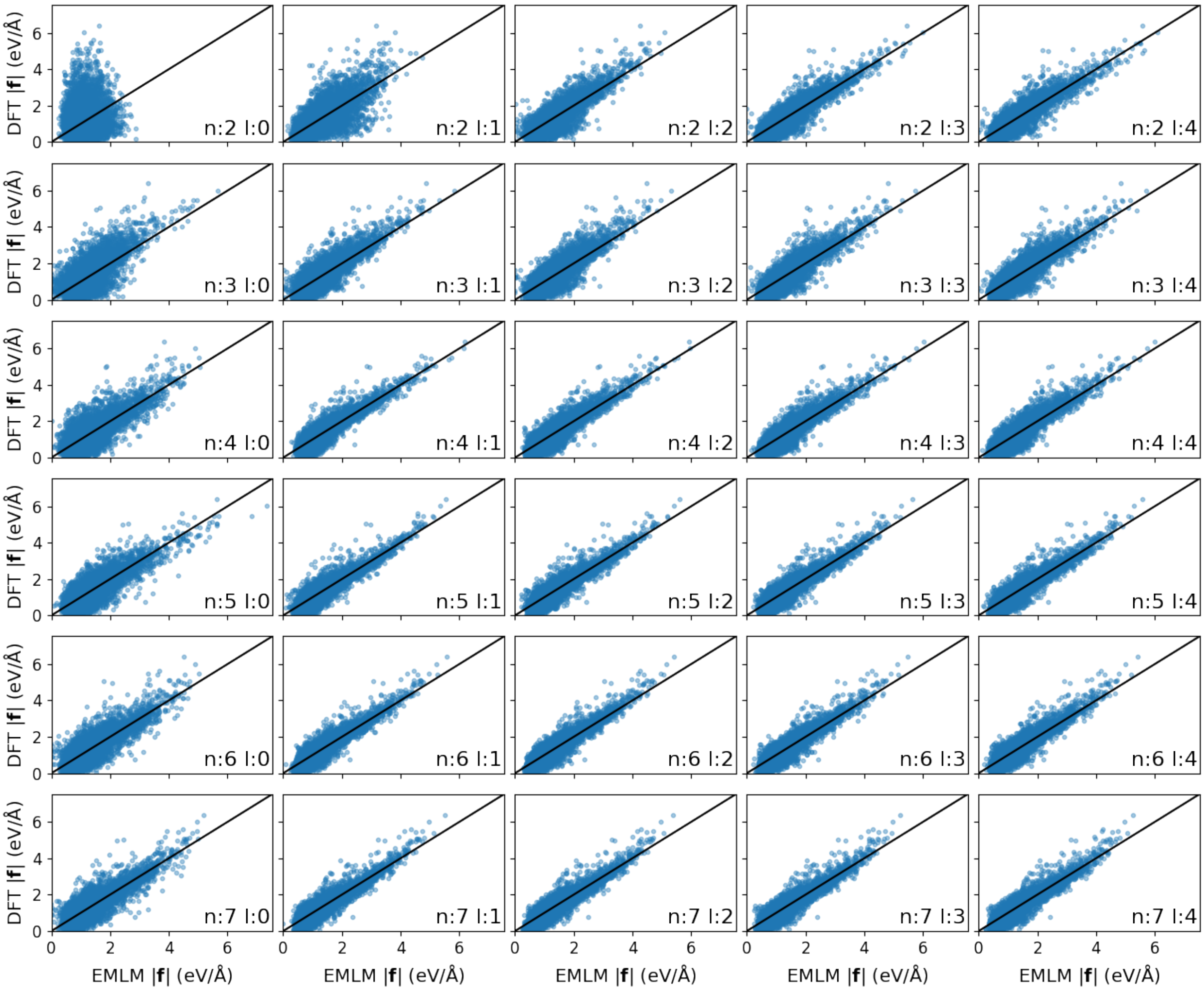


FIG. S13. EMLM force norm predictions are compared to the corresponding DFT values in the case of unit gold. The models were trained with Q isomer data and tested with T. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

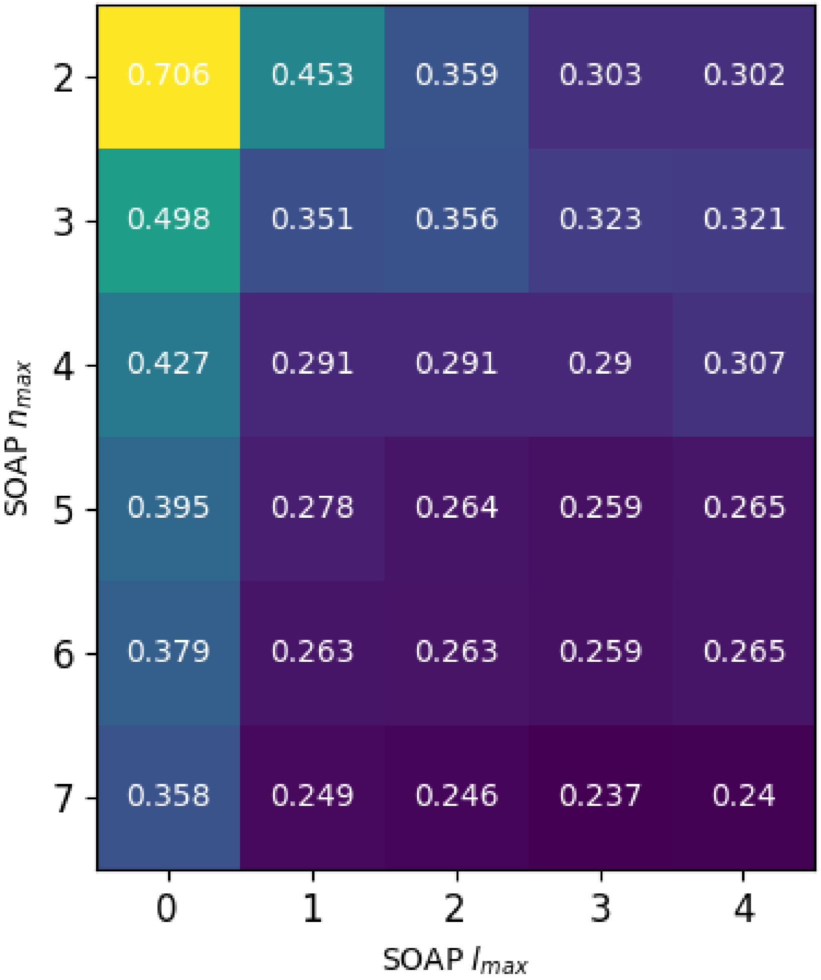


FIG. S14. RMSE values for the testing of EMLM models in FIG. S13. The models were trained with Q isomer data and tested with T. For SOAP descriptions cut-off radius r_{cut} = 4.0 Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

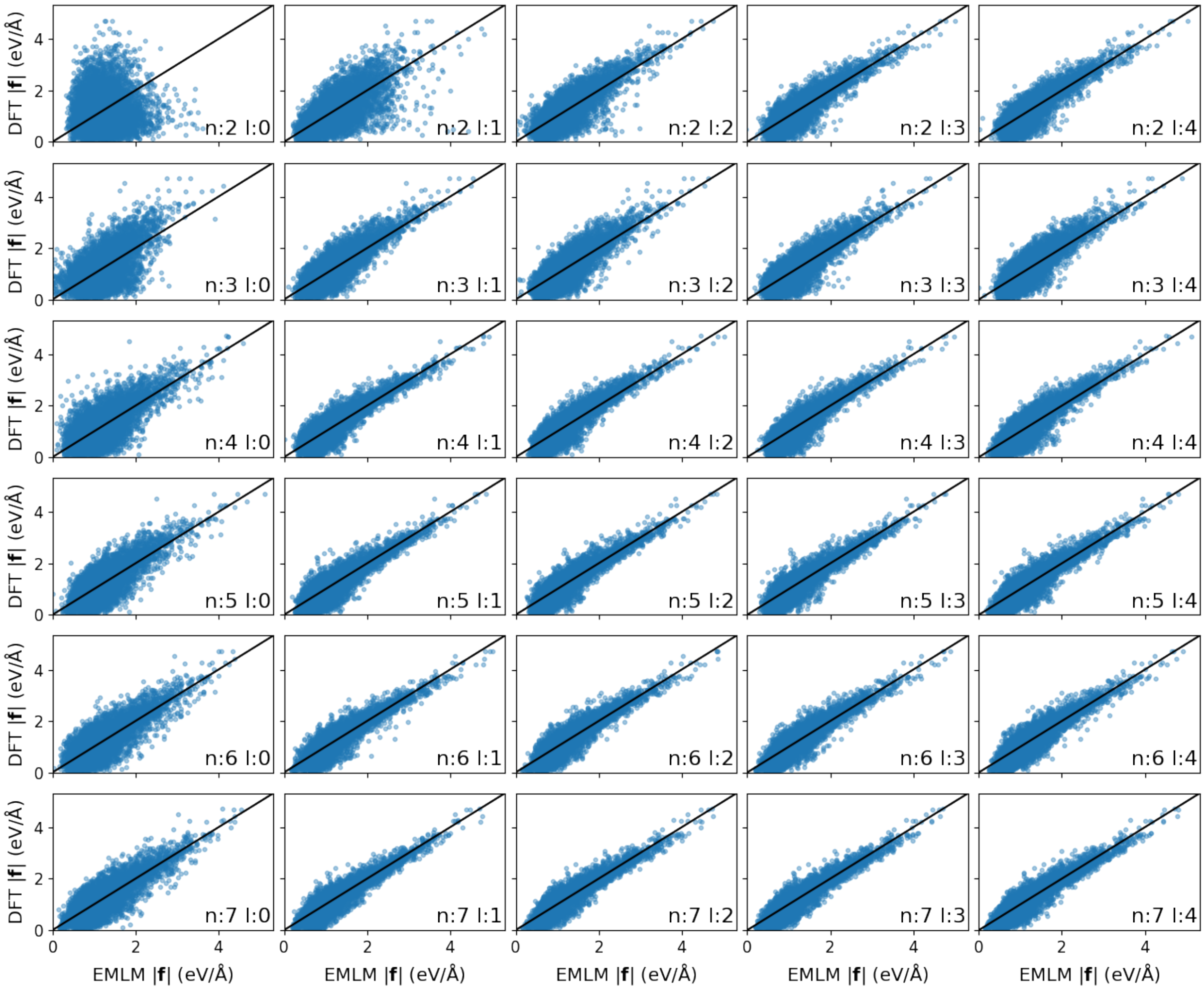


FIG. S15. EMLM force norm predictions are compared to the corresponding DFT values in the case of unit gold. The models were trained with T isomer data and tested with Q. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

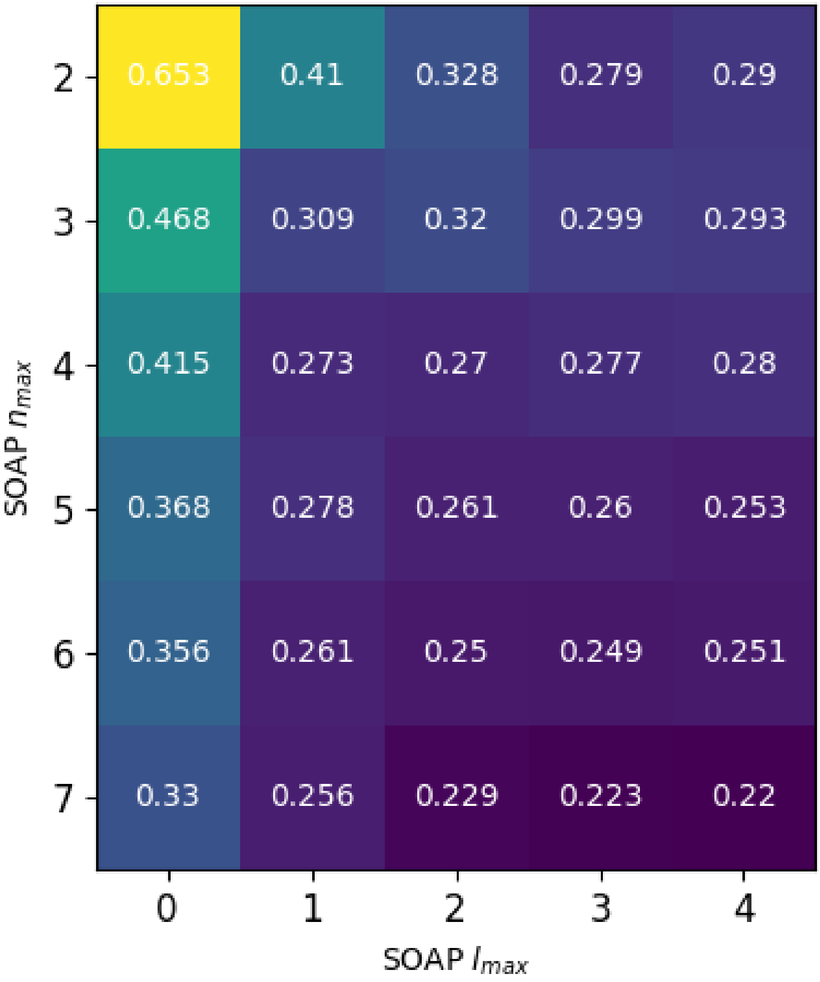


FIG. S16. RMSE values for the testing of EMLM models in FIG. S15. The models were trained with T isomer data and tested with Q. For SOAP descriptions cut-off radius r_{cut} = 4.0 Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

E. Core gold

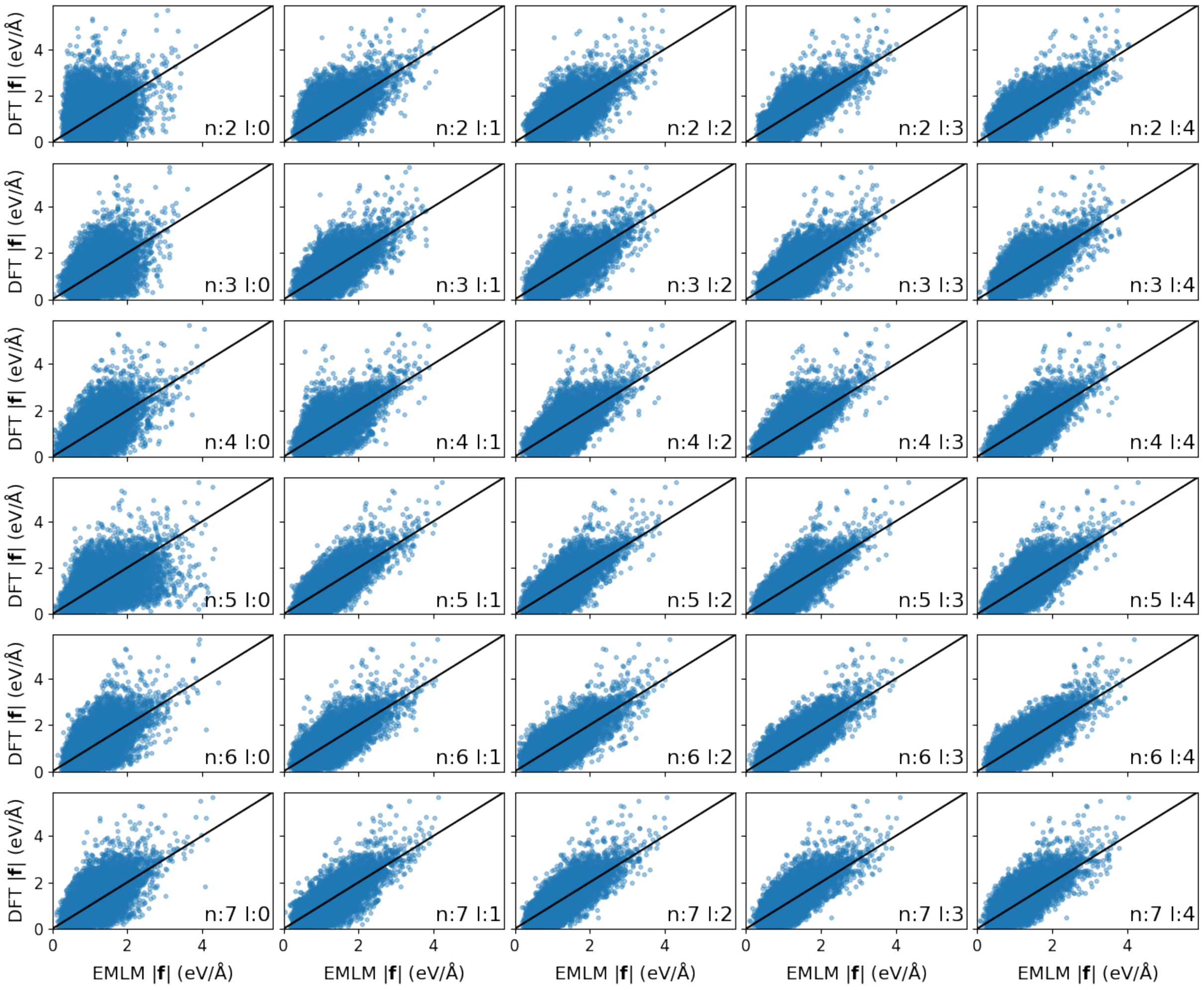


FIG. S17. EMLM force norm predictions are compared to the corresponding DFT values in the case of core gold. The models were trained with Q isomer data and tested with T. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

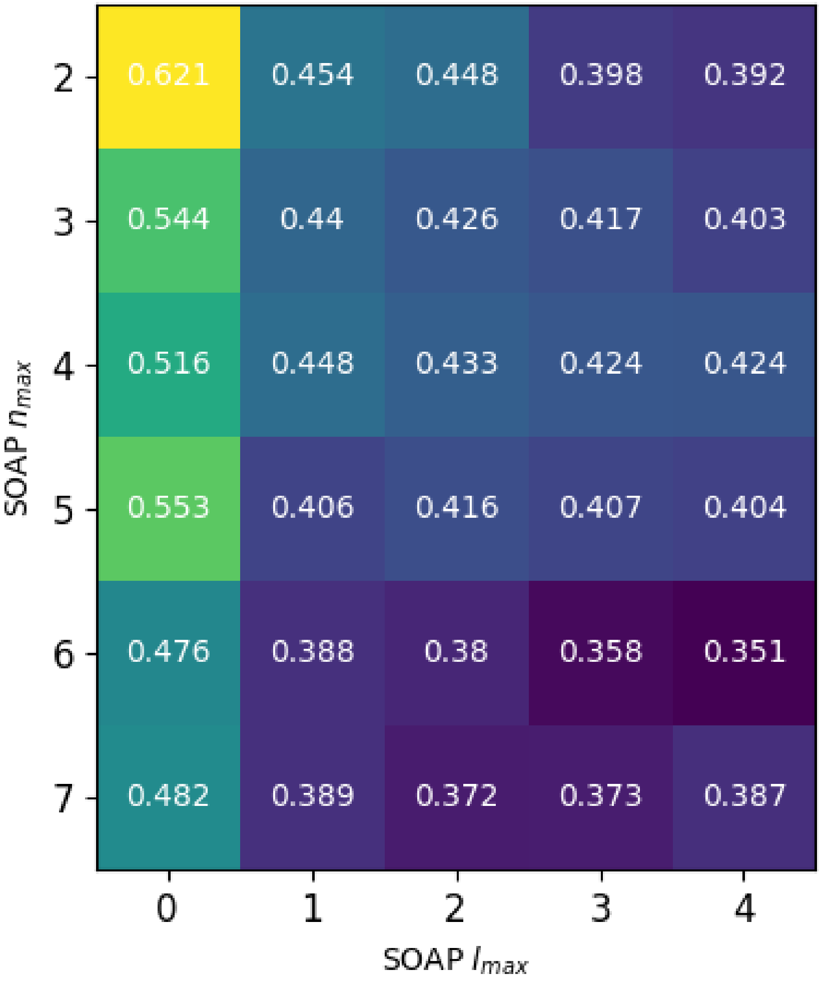


FIG. S18. RMSE values for the testing of EMLM models in FIG. S17. The models were trained with Q isomer data and tested with T. For SOAP descriptions cut-off radius r_{cut} = 4.0 Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

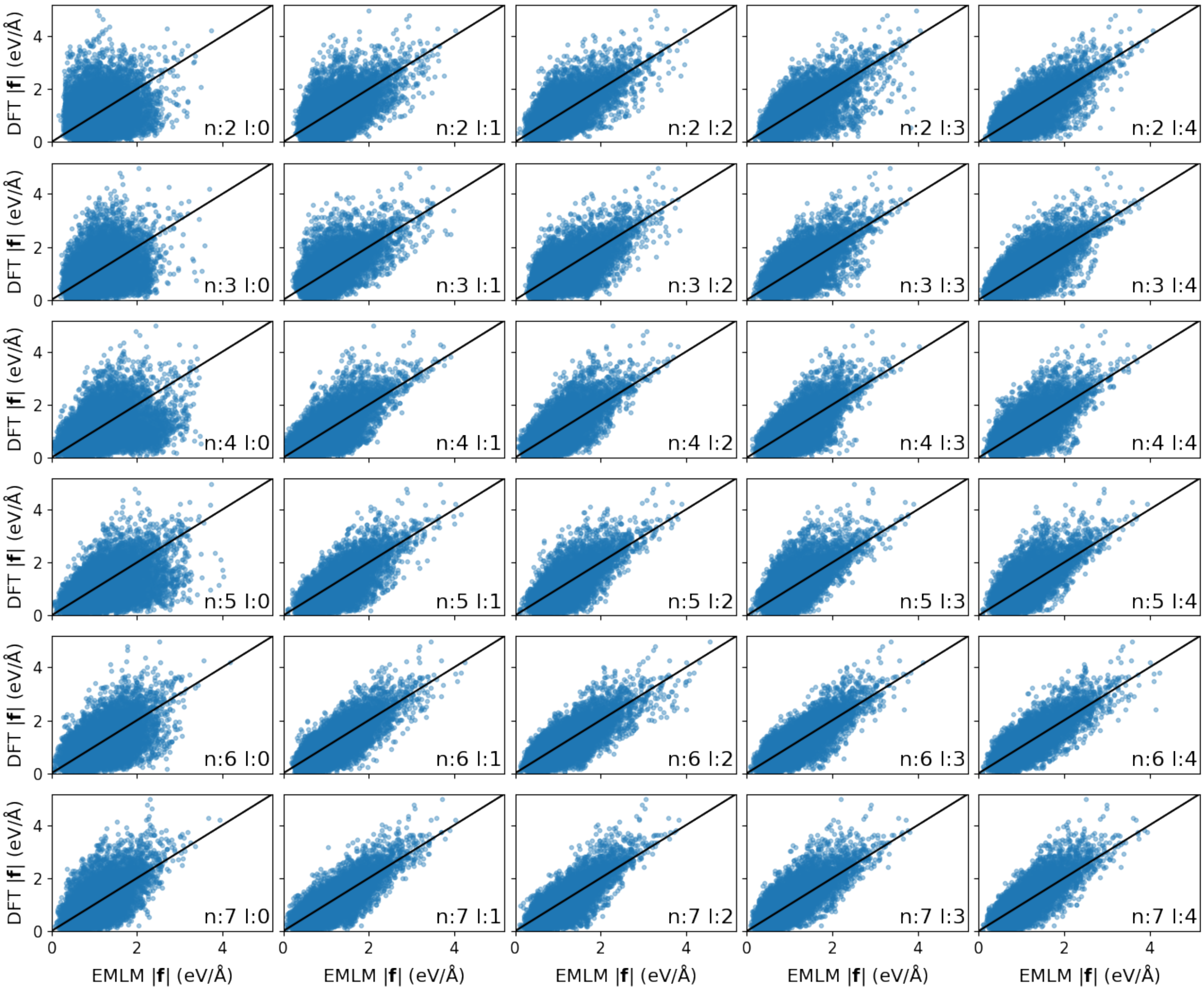


FIG. S19. EMLM force norm predictions are compared to the corresponding DFT values in the case of core gold. The models were trained with T isomer data and tested with Q. For SOAP descriptions cut-off radius $r_{cut} = 4.0$ Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

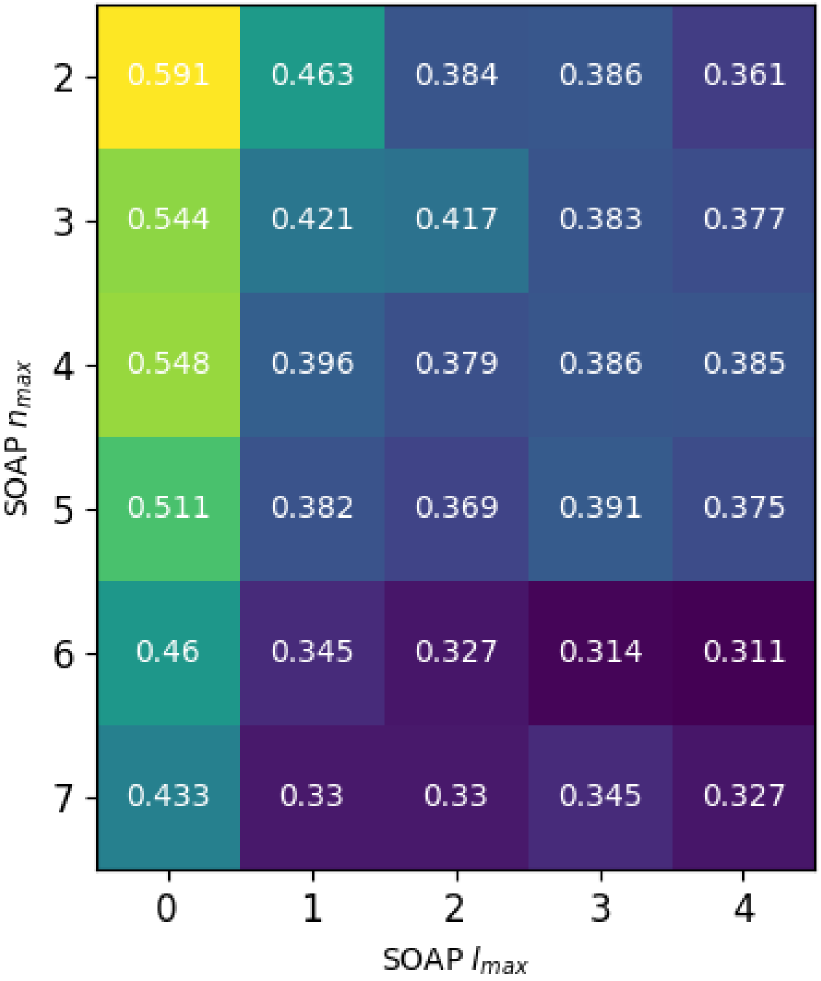


FIG. S20. RMSE values for the testing of EMLM models in FIG. S19. The models were trained with T isomer data and tested with Q. For SOAP descriptions cut-off radius r_{cut} = 4.0 Å and Gaussian broadening $\sigma_{SOAP} = 0.25$ Å.

II. SOAP PARAMETER TESTING VIA FORCE DIRECTIONS

Based on the norm prediction test with EMLM we restricted the parameters to $(n_{\max}, l_{\max}) \in \{(6,4), (7,3), (7,4)\}$, $r_{cut} \in \{4.0, 5.0\}$ Å and $\sigma_{SOAP} = 0.25$ Å. Tests were run similarly as before. First 2500 points were selected from both isomers using RS-maximin. The OAMLM models were trained with training data from a single isomer only and the performance was tested with all data from another isomer. The performance was measured with weighted average of the angles between predicted directions and DFT force vectors. The squared norms of the DFT forces were used as weights. As mentioned in the main article, the used OAMLM scheme has two options for the loss functions: numeric loss function

$$\min_{\hat{\mathbf{v}}_i \in \mathbb{R}^3} J_1(\hat{\mathbf{v}}_i) = -\sum_{j=1}^{K} \exp\left(-\left(\frac{d_{c,j} - (\mathbf{R}_{i,j}\hat{\mathbf{t}}_j) \cdot \hat{\mathbf{v}}_i}{\sigma_1} \right)^2 - \left(\frac{g_{i,j}^{(\prime)}}{\sigma_2} \right)^2 \right), \tag{1}$$

and analytic loss function

$$\min_{\hat{\mathbf{v}}_i \in \mathbb{R}^3} J_2(\hat{\mathbf{v}}_i) = \frac{1}{2} \sum_{j=1}^{K} \omega_{i,j} \left[\hat{\mathbf{v}}_i \cdot (\mathbf{R}_{i,j}\hat{\mathbf{t}}_j) - d_{c,j} \right]^2 \tag{2}$$

where

$$\omega_{i,j} = \exp\left(-\left(\frac{g_{i,j}^{(\prime)}}{\sigma_2} \right)^2 \right). \tag{3}$$

The model was tested with both loss functions with $\sigma_1 = 0.25$ and $\sigma_2 = 0.5$. The results are shown in FIG. S21 for numeric loss function and FIG. S22 for the analytic. The analytic loss function was shown to be better than the numeric one, therefore the effect of σ_2 parameter was tested with it. The tested values were 0.25 and 0.75 in addition to the previously used 0.5. These test are shown in FIG. S23 and S24. For more details look the main article with full explanation.

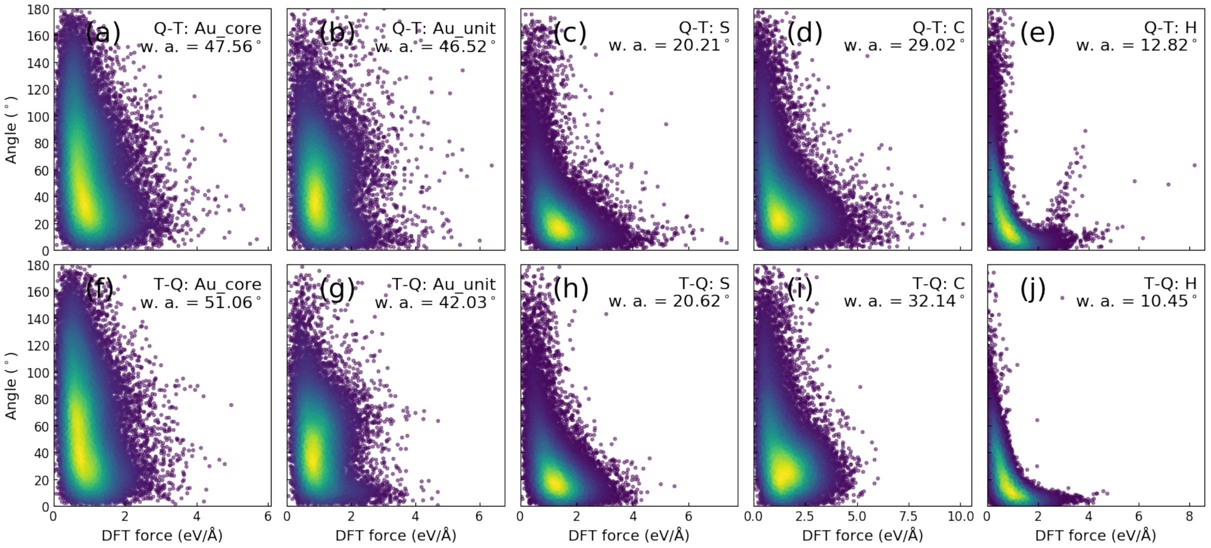


FIG. S21. The performance of the OAMLM with numeric loss function. Vertical axises are the angle between the predicted direction and the DFT force vectors. The horizontal axises show corresponding DFT force norms. In panels (a)-(e) the models were trained with Q isomer and tested with T isomer and (f)-(j) it's vice versa. The tested element is written to the corner of every graph. For hydrogen only third of the data points are plotted. In the graphs "w. a." stands for weighted average angle of the predictions. The colors visualize the density of points: yellow means dense region and purple sparse. SOAP parameters are $n_{\mathrm{max}} = 7$, $l_{\mathrm{max}} = 4$ and $r_{\mathrm{cut}} = 4.0$ Å and loss function parameters $\sigma_1 = 0.25$ and $\sigma_2 = 0.5$.

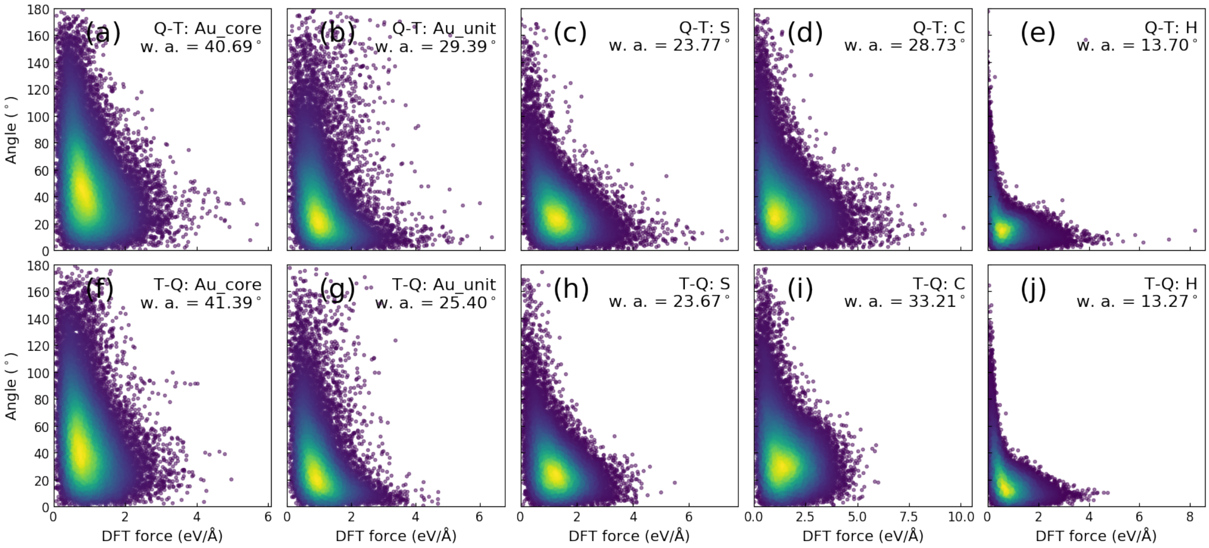


FIG. S22. The performance of the OAMLM with analytic loss function. Vertical axises are the angle between the predicted direction and the DFT force vectors. The horizontal axises show corresponding DFT force norms. In panels (a)-(e) the models were trained with Q isomer and tested with T isomer and (f)-(j) it's vice versa. The tested element is written to the corner of every graph. For hydrogen only third of the data points are plotted. In the graphs "w. a." stands for weighted average angle of the predictions. The colors visualize the density of points: yellow means dense region and purple sparse. SOAP parameters are $n_{\max} = 7$, $l_{\max} = 4$ and $r_{\text{cut}} = 4.0$ Å and loss function parameter $\sigma_2 = 0.5$.

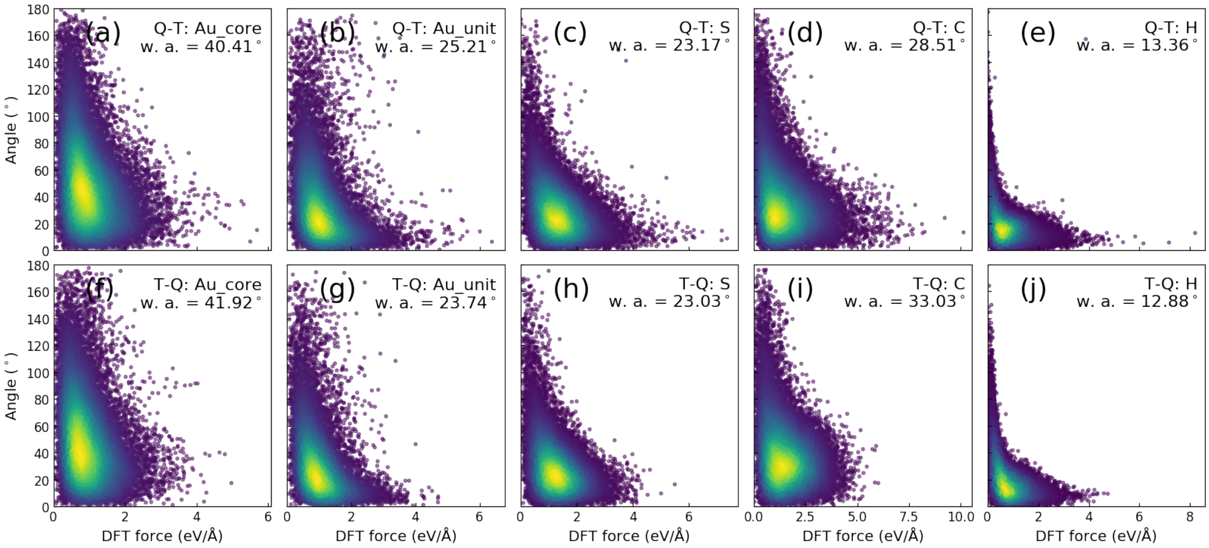


FIG. S23. The performance of the OAMLM with analytic loss function. Vertical axises are the angle between the predicted direction and the DFT force vectors. The horizontal axises show corresponding DFT force norms. In panels (a)-(e) the models were trained with Q isomer and tested with T isomer and (f)-(j) it's vice versa. The tested element is written to the corner of every graph. For hydrogen only third of the data points are plotted. In the graphs "w. a." stands for weighted average angle of the predictions. The colors visualize the density of points: yellow means dense region and purple sparse. SOAP parameters are $n_{\mathrm{max}} = 7$, $l_{\mathrm{max}} = 4$ and $r_{\mathrm{cut}} = 4.0$ Å and loss function parameter $\sigma_2 = 0.25$.

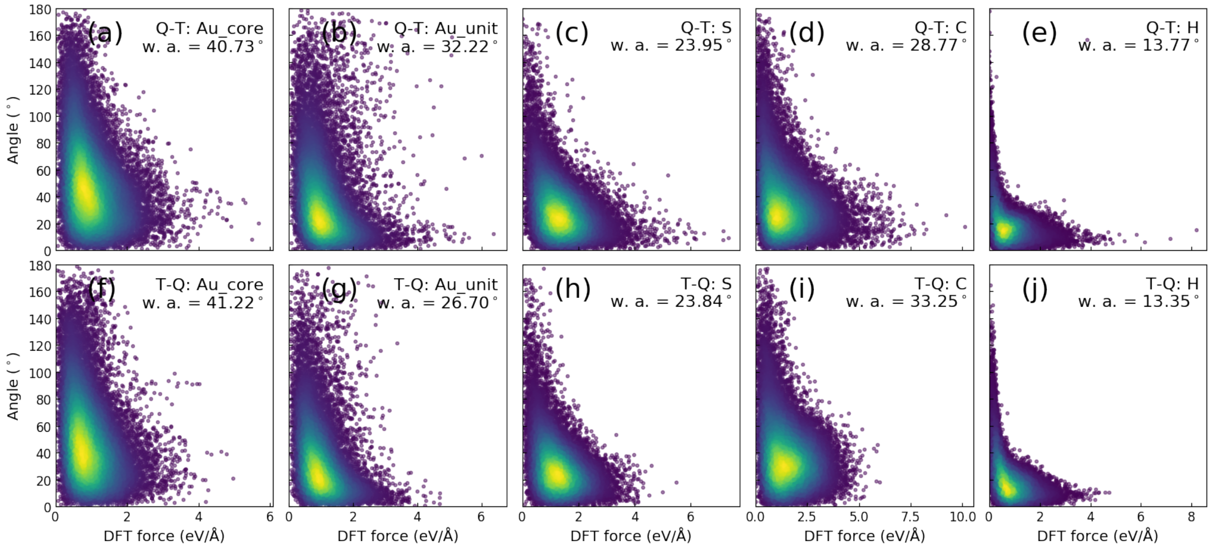


FIG. S24. The performance of the OAMLM with analytic loss function. Vertical axises are the angle between the predicted direction and the DFT force vectors. The horizontal axises show corresponding DFT force norms. In panels (a)-(e) the models were trained with Q isomer and tested with T isomer and (f)-(j) it's vice versa. The tested element is written to the corner of every graph. For hydrogen only third of the data points are plotted. In the graphs "w. a." stands for weighted average angle of the predictions. The colors visualize the density of points: yellow means dense region and purple sparse. SOAP parameters are $n_{\max} = 7$, $l_{\max} = 4$ and $r_{\text{cut}} = 4.0$ Å and loss function parameter $\sigma_2 = 0.75$.

[1] J. Hämäläinen, A. S. C. Alencar, T. Kärkkäinen, C. L. C. Mattos, A. H. Souza Júnior, and J. P. P. Gomes, Minimal learning machine: Theoretical results and clustering-based reference point selection, J. Mach. Learn. Res. **21**, 1 (2020).

[2] H. Qian, W. T. Eckenhoff, Y. Zhu, T. Pintauer, and R. Jin, Total structure determination of thiolate-protected au38 nanoparticles, J. Am. Chem. Soc. **132**, 8280 (2010).

[3] S. Tian, Y.-Z. Li, M.-B. Li, J. Yuan, J. Yang, Z. Wu, and R. Jin, Structural isomerism in gold nanoparticles revealed by x-ray crystallography, Nat. Commun. **6**, 8667 (2015).