THE JOURNAL OF
PHYSICAL
CHEMISTRY A

Article

Monte Carlo Simulations of Au(SCH) Nanocluster Using Distance-Based Machine Learning Methods

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

J. Phys. Chem. A, **Just Accepted Manuscript** • DOI: 10.1021/acs.jpca.0c01512 • Publication Date (Web): 15 May 2020

Downloaded from pubs.acs.org on May 17, 2020

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

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

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 characterised 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 generalisation 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 synthesised via wet chemistry by reducing metal salts in 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 = para mercapto benzoic 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 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 analyse 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 utilised to generate realistic cluster structures for further investigations of thermal-dependent electronic and optical properties of $Au_{38}(SR)_{24}$.

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 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] 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 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] Afterwards 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

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

In equation (1) summations are going through atoms with atomic (element) number 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 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 $g_2(i, j) = R^{-1}_{i,j}$. 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,cut-off}. 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 cut-off. Only the values of the equation (1), which are greater than the cut-off, are used in summation for every value of x. This affects the sensitivity of the descriptor and also the speed of computations. A small cut-off 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 cut-off would allow smaller changes in the structure to be visible in the description than a large cut-off. Using small cut-off 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 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 $f : X \to Y$ between input and output spaces would be used to make 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 $M = \{\mathbf{m}_k\}_{k=1}^{K}$ 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

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

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

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

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 $\mathbf{x}_i$ an $\mathbf{y}_i$. 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

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

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

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

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

$$f(\tilde{\mathbf{x}}) = \mathbf{W}\mathbf{h}(\tilde{\mathbf{x}}), \tag{12}$$

where $\mathbf{h}$ is the same vector valued kernel function as before. With input vector $\tilde{\mathbf{x}}$ it yields $K \times 1$ 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 $Au_{38}(SCH_3)_{24}$ 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 Figures 1C and 1D. 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

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

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.

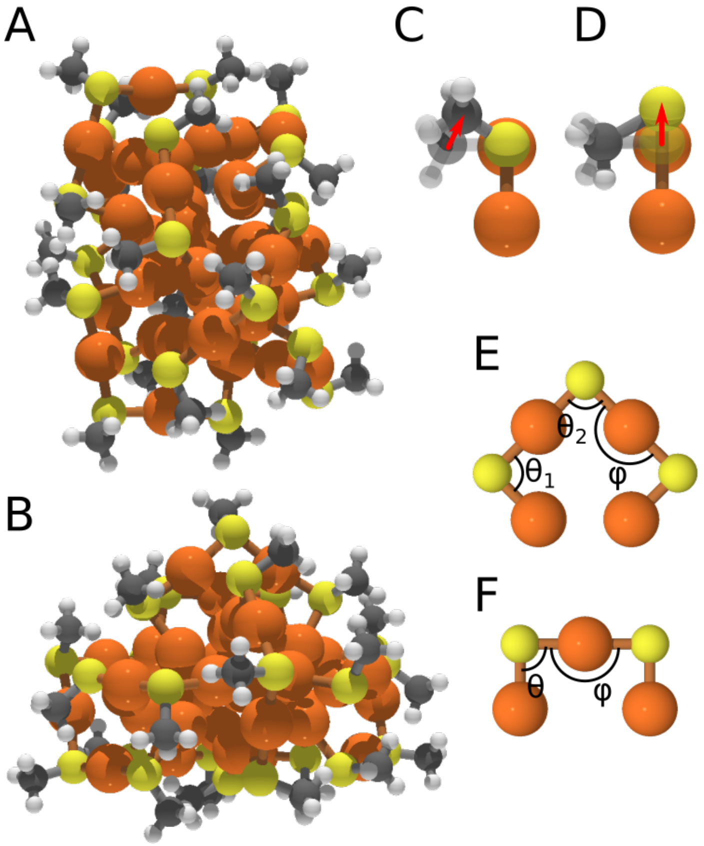


Figure 1: The initial structures of $Au_{38}(SCH_3)_{24}$ are visualized for Q and T isomers in (A) and (B) respectively. While moving sulfur atoms and methyls the orientation of the S-C bond has to be preserved. (C) shows how alignment is preserved if methyl is moved. (D) show the same when sulfur atom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

Results and Discussion

Generating training data and training the models

The training data from the $Au_{38}(SCH_3)_{24}$ clusters was generated using density functional theory (DFT) run with GPAW code.[62,63] The major training data was 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,cut-off}. 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 discussion on choosing the parameters, see 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 non-physical 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 behaviour, 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 Q and 2124 for 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 datapoints are arranged in the feature space.

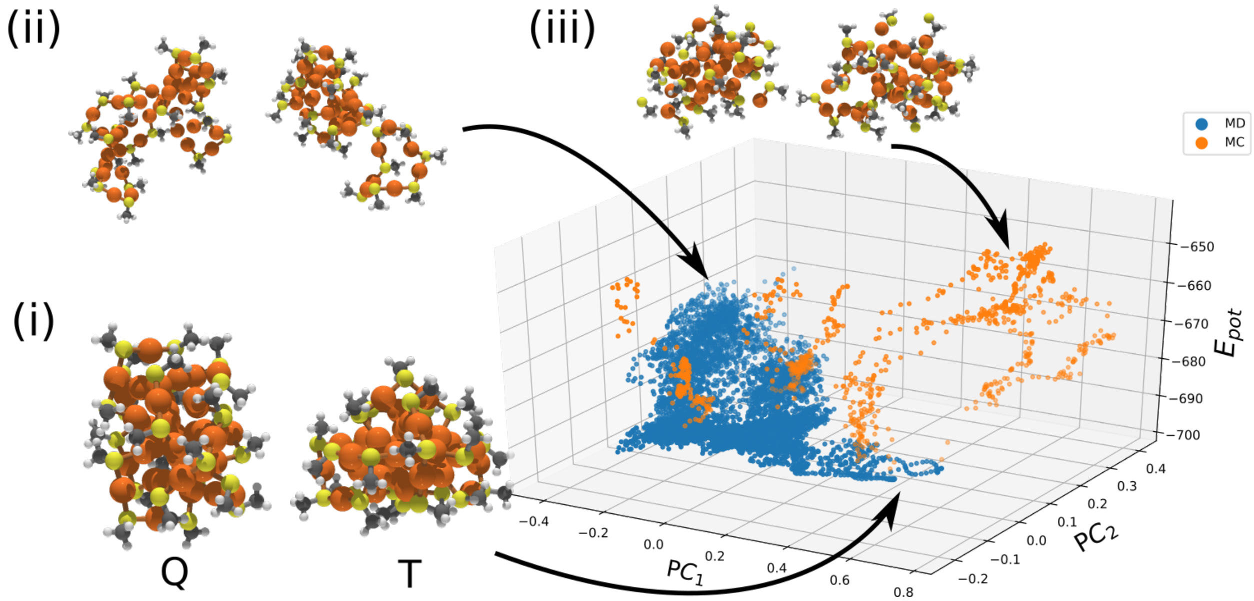


Figure 2: PCA visualisation of MBTR descriptors of the training data. For the sake of clarity only 25% of the points are present in the graph. (i) the initial structures and (ii) high-temperature structures of the original MD simulations[38] (iii) snapshots from Monte Carlo simulations, where S-Au bonds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

Validation: potential energy MLM/EMLM vs. DFT-MD

For validation, we created new independent DFT MD reference data sets both for 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-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.

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 K, 250 K, 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 5(A) indicating that the MC dynamics of both isomers is 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 analysed the statistics of selected bond distances and bond angles for both isomers

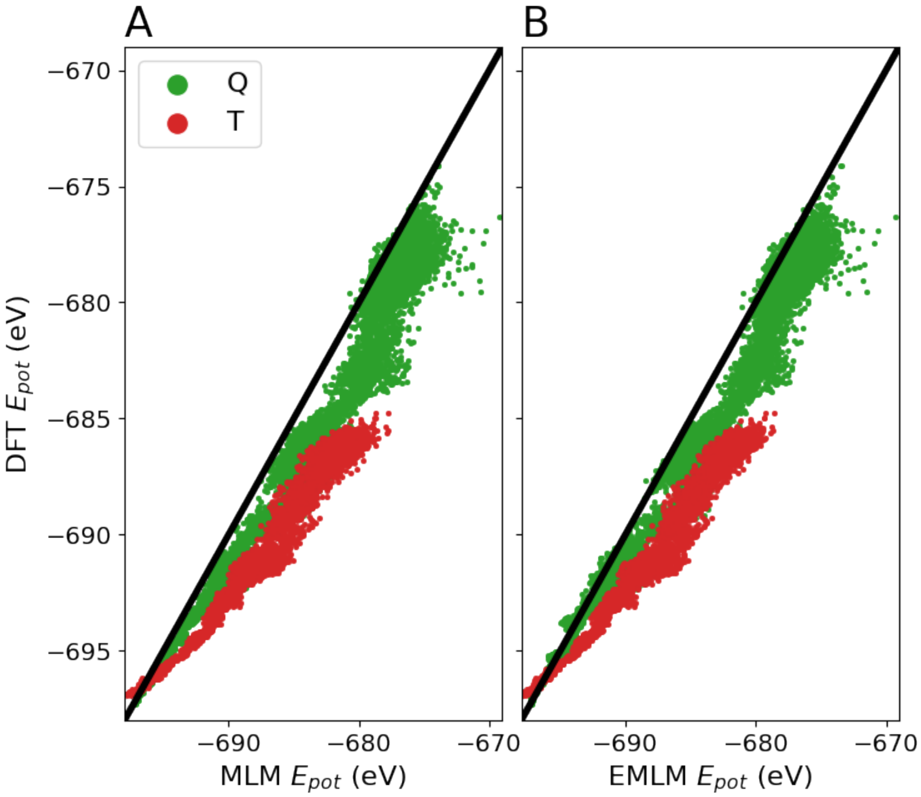


Figure 3: Correlation between the predicted potential energy from (A) MLM and (B) EMLM to the DFT energy from the MD calculations for Q and T isomers.

from the MC runs at 200 K, 250 K, and 300 K. Last 500 MC steps from each simulations were used for the analysis. Figure 6 shows the statistics for the nearest neighbour 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 K and 474 K for Q isomer and 272 K 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

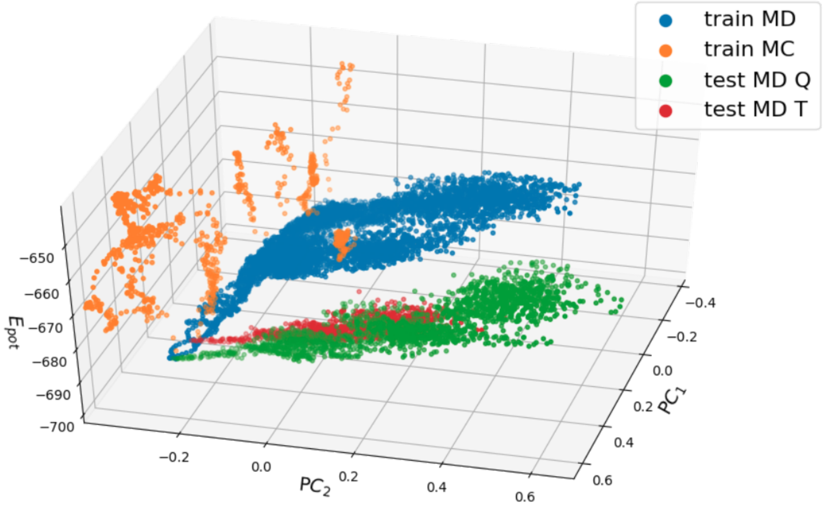


Figure 4: Visualization for PCA from training data and test MD data. Potential energies on z axis are computed with DFT. The graph is rotated with respect to Figure 2. In order to keep visualization clear, only 25% of the points are included

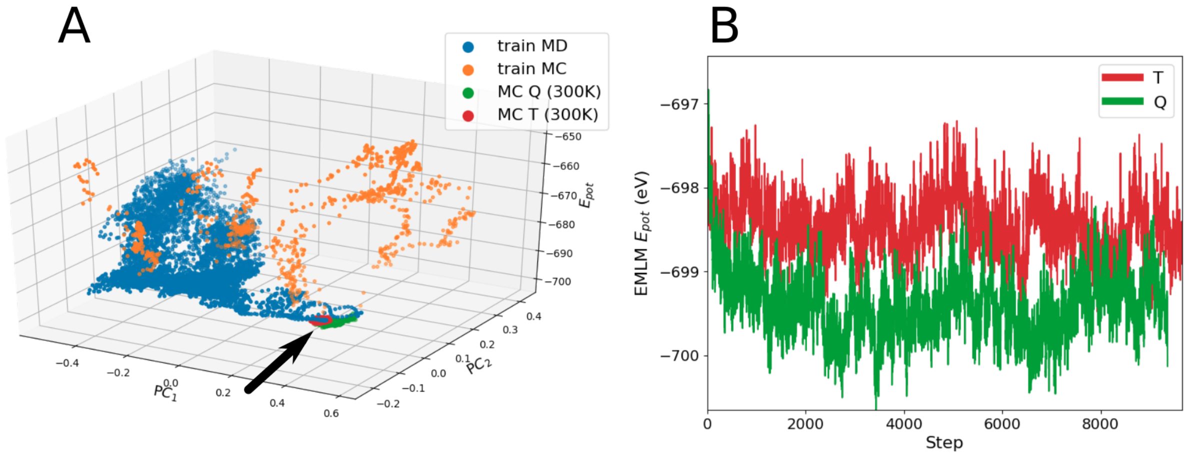


Figure 5: (A) Same as Figure 2, but including also the PCA analysis of EMLM MC runs at 300 K for isomers Q and T. The arrow highlights the region of the MC data. The analysis indicates that both of the isomers are vibrating close to their minima. Only 25% of the points are included into the Figure and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

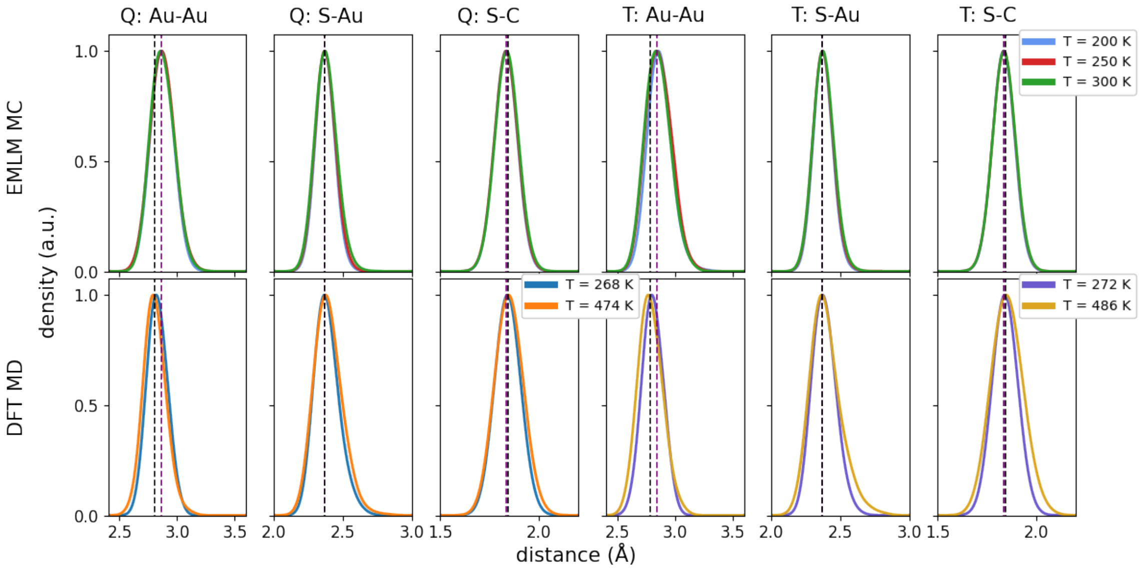


Figure 6: Top row: bond distance distributions from EMLM MC simulations at the indicted temperatures. Bottom row: the same data from DFT MD simulations at indicated temperatures. Labels on the top indicate the isomer and bond type. The vertical dashed lines indicate the average peak positions for every 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 is summed from gaussian-smoothened ($\sigma = 0.05$ Å) data points.

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.

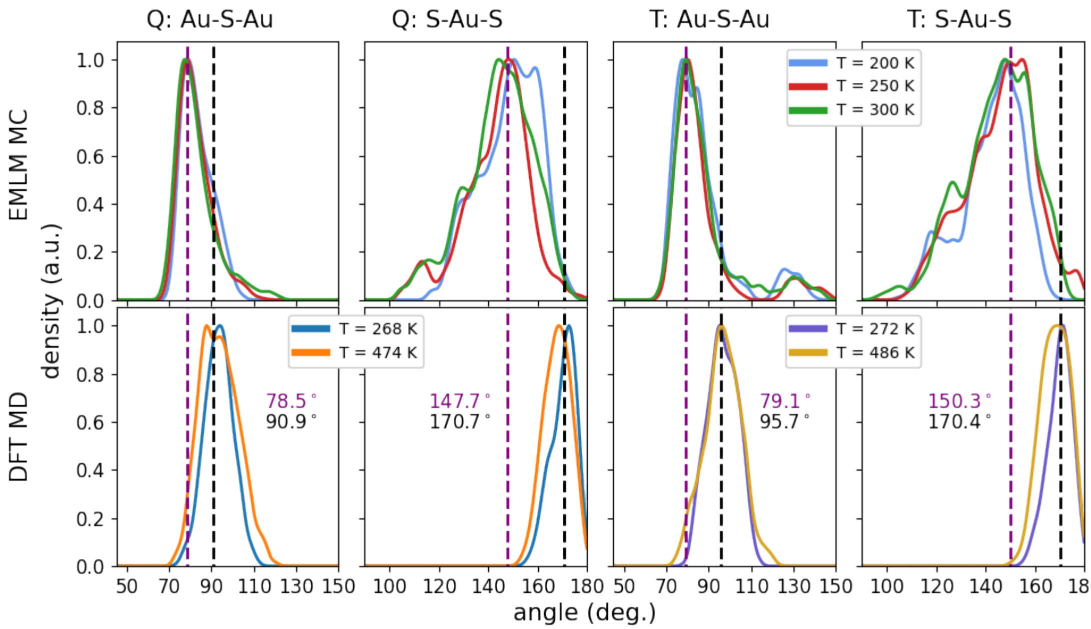


Figure 7: Top row: Selected bond angles distributions from EMLM MC simulations at the indicted temperatures. Bottom row: the same data from DFT MD simulations at indicated temperatures. Labels on the top indicate the isomer and type of the angle. The vertical dashed lines indicate the average peak positions 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 is summed from gaussian-smoothened ($\sigma = 1.75°$) data points.

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 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 @ 2.50GHz with 8GB memory. Computing MBTR k2 with our parameters took about 0.07 seconds for one atomic structure. Prediction of the potential energy using MBTR k2 took about 0.05 seconds with EMLM and 0.56 seconds 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 SI, we could improve the description of the angles of protecting $\mathrm{RS(AuSR)}_{n=1,2}$ 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 showed 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

Additional discussion on testing of the MBTR parameters (SI text and Figures S1, S2). Performance of the MLM with the initial MD training data (Figure S3). The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/

Author Information

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

Notes

The authors declare no competing financial interest.

Acknowledgement

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. MC simulations were done at the FGCI node in the University of Jyväskylä (persistent identifier: urn:nbn:fi:research-infras-2016072533) and the DFT MD simulations were run as part of a PRACE project 2018194723 at the Barcelona Supercomputing Centre.

References

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

(2) McPartlin, M.; Mason, R.; Malatesta, L. Novel cluster complexes of gold(0)–gold(I). *J. Chem. Soc. D* **1969**, 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**, 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 Å resolution. *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. *PNAS* **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* **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. *Comp. Mater. Sci.* **2015**, *110*, 20–28.

(32) Zeni, C.; Rossi, K.; Glielmo, A.; Fekete, Á.; Gaston, N.; Baletto, F.; Vita, A. D. 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, Massachusetts, 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:1704.06439v3 [physics.chem-ph]`. **2017**,

(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* **2013**, *87*.

(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.; Canova, F. F.; Himanen, L.; Foster, A. S. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Comput. Mater.* **2018**, *4*.

(51) Navidi, W.; Jr., W. S. M.; 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.* **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; pp 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; pp 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.

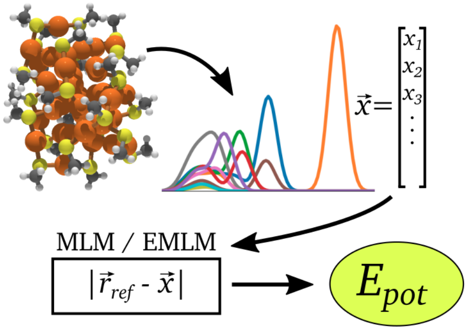
(61) Hämäläinen, J.; Alencar, A. S. C.; Kärkkäinen, T.; Mattos, C. L. C.; Souza Júnior, A. H.; Gomes, J. P. P. Minimal Learning Machine: Theoretical results and clustering-based reference point selection, `arXiv:1909.09978v1 [cs.LG]`. **2019**,

(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* **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.


$\vec{x}=\begin{bmatrix}x_1\\x_2\\x_3\\\vdots\end{bmatrix}$
MLM / EMLM
$|\vec{r}_{ref} - \vec{x}|$
E_{pot}

TOC Graphic