

REPORT DOCUMENTATION PAGE			Form Approved OMB NO. 0704-0188		
<p>The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p> <p>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</p>					
1. REPORT DATE (DD-MM-YYYY) 01-11-2010		2. REPORT TYPE Final Report		3. DATES COVERED (From - To) 1-Aug-2007 - 31-Jul-2010	
4. TITLE AND SUBTITLE Development of ab-initio multibody energy expansions for the design of metallic materials with extremal properties			5a. CONTRACT NUMBER W911NF-07-1-0519		
			5b. GRANT NUMBER		
			5c. PROGRAM ELEMENT NUMBER 611102		
6. AUTHORS Nicholas Zabaras			5d. PROJECT NUMBER		
			5e. TASK NUMBER		
			5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAMES AND ADDRESSES Cornell University Office of Sponsored Programs Cornell University Ithaca, NY 14853 -2801			8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211			10. SPONSOR/MONITOR'S ACRONYM(S) ARO		
			11. SPONSOR/MONITOR'S REPORT NUMBER(S) 52605-MS.1		
12. DISTRIBUTION AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited					
13. SUPPLEMENTARY NOTES The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.					
14. ABSTRACT Computational design of new materials relies on accurate descriptions of interatomic potentials. Such potentials can be realized within the Multi-Body Expansion (MBE) framework, where the expansions constructed using ab-initio calculations offer a generalized potential that can be used to describe energetics, since energies can be conceived as summations of the small cluster contributions. Furthermore, MBE technique focus on positional degrees of freedom, thus, it would eliminate a significant amount of expensive and time consuming energy minimization					
15. SUBJECT TERMS ab initio, potential energy surfaces, configuration space, interpolation, reduced-order model, multi body energy expansion, interatomic potentials					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Nicholas Zabaras
a. REPORT UU	b. ABSTRACT UU	c. THIS PAGE UU			19b. TELEPHONE NUMBER 607-255-9104

Report Title

Development of ab-initio multibody energy expansions for the design of metallic materials with extremal properties

ABSTRACT

Computational design of new materials relies on accurate descriptions of interatomic potentials. Such potentials can be realized within the Multi-Body Expansion (MBE) framework, where the expansions constructed using ab-initio calculations offer a generalized potential that can be used to describe energetics, since energies can be conceived as summations of the small cluster contributions. Furthermore, MBE technique focus on positional degrees of freedom, thus, it would eliminate a significant amount of expensive and time consuming energy minimization required to search for stable phase structures. However, in practice, obtaining the N-body ($N > 2$) potentials is quite a challenging problem and this has been the focus of our work.

List of papers submitted or published that acknowledge ARO support during this reporting period. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

V. Sundararaghavan and N. Zabaras, "Many-body expansions for computing stable structures of multi-atom systems", Physical Review B, Vol. 77 (6), pp. 064101-1--064101-10, 2008.

V. Sundararaghavan and N. Zabaras, "A multilength scale continuum sensitivity analysis for the control of texture-dependent properties in deformation processing", International Journal of Plasticity, Vol. 24, pp. 1581-1605, 2008

B. Kouchmeshky and N. Zabaras, "Modeling the response of HCP polycrystals deforming by slip and twinning using a finite element representation of the orientation space", Computational Materials Science, Vol. 45, Issue 4, pp. 1043--1051, 2009

V. Sundararaghavan and N. Zabaras, "A statistical learning approach for the design of polycrystalline materials", Statistical Analysis and Data Mining, Vol. 1, Issue 5, pp. 306--321, 2009 (invited paper for the special issue on 'Materials Informatics: Data-Driven Discovery in Materials Science', Krishna Rajan and Patricio Mendez, eds.)

B. Kouchmeshky and N. Zabaras, "The effect of multiple sources of uncertainty on the convex hull of material properties", Computational Materials Science, Volume 47, Issue 2, pp. 342--352, 2009

B. Kouchmeshky and N. Zabaras, "Microstructure model reduction and uncertainty quantification in multiscale deformation processes", Computational Materias Science, Vol. 48, Issue 2, pp. 213--227, 2010

Zheng Li, Bin Wen and N. Zabaras, "Computing mechanical response variability of polycrystalline microstructures through dimensionality reduction techniques", Computational Materials Science, 49 (2010) 568-581

Number of Papers published in peer-reviewed journals: 7.00

(b) Papers published in non-peer-reviewed journals or in conference proceedings (N/A for none)

Number of Papers published in non peer-reviewed journals: 0.00

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts): 0

Peer-Reviewed Conference Proceeding publications (other than abstracts):

B. Ganapathysubramanian and N. Zabaras, "Multibody expansions: An ab initio based transferable potential for computational thermodynamics", presented at the 'Computational thermodynamics and kinetics' symposium in the 2008 TMS Annual Meeting & Exhibition, (Y. Wang, L.-Q. Chen, J. J. Hoyt, Y. U. Wang, organizers), New Orleans, Louisiana, March 9-13, 2008

Baskar Ganapathysubramanian and Nicholas Zabaras, "Characterizing adsorption on metallic surfaces: effect of composition", presented at the 'Computational Thermodynamics and Kinetics' symposium (Long Qing Chen, Yunzhi Wang, Pascal Bellon, Yongmei Jin, organizers) at the 2009 TMS Annual Meeting & Exhibition, San Francisco, CA, February 15-19. 2009

B. Kouchmeshky and N. Zabaras, "A simple non-hardening rate-independent constitutive model for HCP polycrystals deforming by slip and twinning", presented at the 'Deformation Twinning: Formation Mechanisms and Effects on Material Plasticity: Experiments and Modeling' in the 2008 TMS Annual Meeting & Exhibition, (G. T. Gray, S. Mahajan, E. K. Cerreta, organizers), New Orleans, Louisiana, March 9-13, 2008

Wei Li and N. Zabaras, "Finite element modeling of the deformation of 3D polycrystals including the effect of grain size distribution", presented at the '3-Dimensional Materials Science' symposium in the 2008 TMS Annual Meeting & Exhibition, (M. D. Uchic, E. M. Taleff, A. C. Lewis, J. P. Simmons, M. J. DeGraef, organizers), New Orleans, Louisiana, March 9-13, 2008

B. Kouchmeshky and N. Zabaras, "A microstructure-sensitive design approach for controlling properties of HCP materials", presented at the '9th Global Innovations Symposium: Trends in Integrated Computational Materials Engineering for Materials Processing and Manufacturing' symposium in the 2008 TMS Annual Meeting & Exhibition, (C. C. Battaile, A. Misra, J.A. Hines, J. W. Sears, organizers), New Orleans, Louisiana, March 9-13, 2008

B. Kouchmeshky and N. Zabaras, "Advances on multiscale design of deformation processes for the control of material properties", presented at the 'Materials Processing Fundamentals' symposium (Prince N. Anyalebechi, organizer) at the 2009 TMS Annual Meeting & Exhibition, San Francisco, CA, February 15-19. 2009

N. Zabaras and B. Kouchmeshky, "Modeling uncertainty propagation in deformation processes", presented at the 'General Abstracts: Materials Processing and Manufacturing Division' symposium (Thomas R. Bieler, Neville R. Moody, organizers) at the 2009 TMS Annual Meeting & Exhibition, San Francisco, CA, February 15-19. 2009

Babak Kouchmeshky and Nicholas Zabaras, "Uncertainty quantification in multiscale deformation processes", presented at the 'Stochastic material models' symposium (Sharif Rahman et al., organizers) at the 10th U.S. National Congress in Computational Mechanics, Columbus, OH, July 16-19, 2009

Zheng Li and Nicholas Zabaras, "Low-dimensional models for microstructure representation: A data-driven approach", presented at the 'Stochastic material models' symposium (Sharif Rahman et al., organizers) at the 10th U.S. National Congress in Computational Mechanics, Columbus, OH, July 16-19, 2009

Bin Wen and Nicholas Zabaras, "Grain-size effect in 3D polycrystalline microstructure including texture evolution", presented at the 'Modeling issues and computational methodologies of virtual polycrystals' symposium (P. Dawson et al., organizers) at the 10th U.S. National Congress in Computational Mechanics, Columbus, OH, July 16-19, 2009

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts): 10

(d) Manuscripts

Bin Wen, Zheng Li and N. Zabaras, "Thermal response variability of random polycrystalline microstructures", Communications in Computational Physics, submitted.

Ilias Bilionis and N. Zabaras, "Multi Body Expansion Using Sparse, Permutation Invariant and Bayesian Estimation of Potential Energy Surfaces", Physical Reviews B, submitted.

Number of Manuscripts: 2.00

Patents Submitted

Patents Awarded

Awards

Graduate Students

<u>NAME</u>	<u>PERCENT SUPPORTED</u>
Veera Sundararaghavan	1.00
Peng Chen	1.00
Ilias Bilonis	1.00
Baskar Ganapathysubramanian	1.00
Zhoulong Huang	1.00
Babak Kouchmeshky	1.00
FTE Equivalent:	6.00
Total Number:	6

Names of Post Doctorates

<u>NAME</u>	<u>PERCENT SUPPORTED</u>
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

<u>NAME</u>	<u>PERCENT SUPPORTED</u>	National Academy Member
Nicholas Zabaraz	0.10	No
FTE Equivalent:	0.10	
Total Number:	1	

Names of Under Graduate students supported

<u>NAME</u>	<u>PERCENT SUPPORTED</u>
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

- The number of undergraduates funded by this agreement who graduated during this period: 0.00
- The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 0.00
- The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 1.00
- Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 1.00
- Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00
- The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00
- The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: 1.00

Names of Personnel receiving masters degrees

<u>NAME</u>
Total Number:

Names of personnel receiving PHDs

<u>NAME</u>
Veera Sundararaghavan
Baskar Ganapathysubramanian
Babak Kouchmeshky
Total Number: 3

Names of other research staff

<u>NAME</u>	<u>PERCENT SUPPORTED</u>
FTE Equivalent:	
Total Number:	

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

Technology Transfer

Development of ab-initio multibody energy expansions for the design of metallic materials with extremal properties

ARO GRANT NUMBER W911NF-07-1-0519

Final Report: October 31, 2010

Prof. Nicholas Zabaras
Materials Process Design and Control Laboratory
Sibley School of Mechanical and Aerospace Engineering
101 Frank H.T.Rhodes Hall
Cornell University, Ithaca, NY-14853-3801

zabaras@cornell.edu

<http://mpdc.mae.cornell.edu/>

Table of Contents

Abstract	Page 2
Introduction and Background	Page 2
Methods for constructing the potential	Page 3
MBE expansions for computing stable structures of multi-atom systems	Pages 1-10
Towards the construction of fully transferable multi-atom potentials	Pages 1-47

Abstract

Computational design of new materials relies on accurate descriptions of interatomic potentials. Such potentials can be realized within the Multi-Body Expansion (MBE) framework, where the expansions constructed using ab-initio calculations offer a generalized potential that can be used to describe energetics, since energies can be conceived as summations of the small cluster contributions. Furthermore, MBE technique focus on positional degrees of freedom, thus, it would eliminate a significant amount of expensive and time consuming energy minimization required to search for stable phase structures. However, in practice, obtaining the N -body ($N>2$) potentials is quite a challenging problem and this has been the focus of our work.

Introduction and Background

Advances in Molecular Dynamics (MD) and Monte Carlo (MC) techniques have made possible in the recent years the systematic probing of material properties (phase transitions, thermophysical properties) using computer experiments. The coupling of these techniques with advanced statistical methods would enable us to systematically scan for materials with extreme properties. One can imagine a scenario in which the desired properties are first specified and then an extensive computational search is performed to discover a particular material that realizes those (Materials by Design). Subsequently, targeted experiments are performed to actually create this material in the lab. Such a procedure would increase the rate of new discoveries having a profound effect on the development of new technologies, saving at the same time billions of dollars. However, the magic ingredient that connects the MD and MC methods with the real world is a physically accurate description of the energetics of the materials. Such a description is provided through quantum mechanical calculations, albeit at a tremendous computational cost. Thus, the development of efficient ways to tabulate the results of quantum mechanical calculations is a necessary requirement towards the realization of Materials by Design.

We first review the atomistic aspects of this work using the MBE (multibody energy expansion). We want to predict extremal properties from first principles. To do this intelligent alloying, we need a method that allows structure and property prediction of multi-atom systems that are not necessarily on a Bravais lattice (like FCC, BCC, etc.). We need to be able to examine configurations of atoms that are placed anywhere in space (this makes this method very different from the cluster expansion method where the atoms are in a given fixed lattice and the only thing you change is what atom let us say A or B you place in each lattice location). We expand the energy in two body, three-body,

etc. interactions (Figure 1). This is not an obvious trivial expansion as the curse of dimensionality hits you fast and the recursive calculation of these many body potentials is very difficult.

Approach

Require first-principles based calculations. However, currently infeasible to analyze large ($\sim 10^4$) systems

Multi-body expansions: Energy is represented using a hierarchy of structure-independent, transferable many-body potentials



Construct these potentials from first-principle calculations (VASP, Quantum Espresso and DFT++).

These potentials describe energies of arbitrary atomic structures ($\sim 10^4$) as a function of atomic positions very efficiently

Figure 1: The multibody energy expansion (MBE).

Methods for Constructing the Potentials

In the initial execution of this project, N -body potentials were generated by tessellating the hyper-surface and approximating the energy using the finite element method. Convergence characteristics were significantly improved by weighting the energies obtained from various truncation of the many body expansion. However, the finite element based tessellation of the hyper-space places extensive restrictions on the accuracy of the approximation. Moreover, the N -th order potential lies in $3N-6$ dimensional space and finite element tessellation (and subsequent searching and interpolation) of spaces beyond 6 dimensions becomes computationally ineffective. In the second year, we incorporated the newly developed adaptive sparse grid collocation (ASGC) method based on Smolyak algorithm into sampling the topology of the clusters to construct these N -body potentials. Unfortunately, we were unable to interpolate energies of larger than $N=5$ clusters with this method because the increasing dimensionality of the configuration space required a computationally forbidding number of electronic energy calculations.

During the third year, it became apparent we had to move to a grid-less interpolation scheme. We studied Distance Geometry techniques used in the Protein Folding literature to mathematically describe the configuration space and developed a new that enabled us to sample it efficiently. The recently developed Multinomial Expansion Method (MEM) - used in the computational Chemistry literature for the construction of Potential Energy Surface (PES) to study chemical reactions – was chosen as the most promising candidate interpolation scheme. It is the first interpolation scheme that effectively incorporates all invariance principles of a potential energy surface in a single functional form. The most important such invariance principle is the permutation invariance with respect to atoms of

the same type. The effect of this is a drastic reduction on the number of required ab initio calculations, thus making possible the construction of MBE of higher orders. We subsequently improved the fitting capabilities (MEM) by putting it in a Bayesian framework. The most important new contribution is the use of the Bayesian variance to quantify the informational content of each point in the configuration space that lead us to an efficient adaptive scheme that minimizes the required number of electronic calculations even further. We demonstrated that this new technique considerably improves the quality of the samples and outperforms the random selection of data points. We were able to construct the ab initio PES of Platinum clusters of up to 6 atoms with only a few thousand electronic calculations. The ab initio PES were used to find the stable structures of small Pt clusters using Simulated Annealing. The results were found to be in very good agreement with those found in the literature. The constructed Pt PES's were also used to fit the interatomic potentials up to order 6. It was shown that those become less and less important as their order increases, albeit slowly in low energy regions. We used the potentials to investigate the performance Multi-Body Expansions of various orders for Pt clusters of up to 10 atoms. It was demonstrated that interactions of at least 5 atoms are required to qualitatively describe Pt clusters. Finally, we observed that the error introduced during the fitting procedure of the interatomic potentials propagates in a complicated manner through the Multi-Body Expansion formula making its naive application to big clusters questionable. It is the object of our current research to investigate the propagation of this error through the MBE formula and design effective techniques to filter it out. We believe that such filtering schemes have to be case specific (different for each material) and should utilize further physical information. This problem constitutes the final obstacle towards the construction of fully transferable potential energy surfaces using the MBE framework.

With the PES surfaces in place, exploration of the energy landscape in the high dimensional configuration space becomes an easier task that could potentially revolutionize the search for materials for extremal properties. In our immediate plans we are working towards integration of PES meta models with MD and MC techniques to allow us computing materials with desired mechanical and thermophysical properties, phase transition characteristics, etc. Many applications to surface design (e.g. of Pt clusters to maximize H-adsorption) are also anticipated.

In the remaining of this final report, we first briefly discuss the developments of FEM like tessellation techniques of the configuration space for PES construction. For brevity of the report we do not discuss the activities on the sparse grid interpolation approach since the number of ab initio simulations needed with this method was prohibitory high. We finally conclude with the Bayesian framework for interpolating potentials using invariant polynomials in the high-dimensional configuration space. It is this framework that we believe provides the best available option for PES surrogate construction using the minimum number of ab initio runs for properly chosen realizations in the configuration space.

Weighted multi-body expansions for computing stable structures of multi-atom systems

Veera Sundararaghavan, Baskar Ganapathysubramanian, Xiang Ma, Peng Chen and Nicholas Zabaras*
*Materials Process Design and Control Laboratory,
 Sibley School of Mechanical and Aerospace Engineering,
 188 Frank H. T. Rhodes Hall,
 Cornell University,
 Ithaca, NY 14853-3801, USA*

(Dated: November 1, 2010)

The effect of structural relaxations in alloys is described using a multi-body energy expansion formalism. N -body potentials in the multi-body expansion are computed from energies of isolated clusters, which in turn, are calculated from empirical potentials or self-consistent quantum mechanical calculations. Convergence characteristics of multi-body expansions (MBE) are improved by weighting energies obtained from various truncations of many-body expansion in a new method called weighted MBE (wMBE). It is shown that multi-body expansions of many-atom systems can be efficiently constructed using interpolation of isolated cluster energies from databases. In contrast to the method of cluster expansion, wMBE focuses on positional degrees of freedom and hence, explicitly handles structural relaxations during computations of stable atom clusters and periodic or amorphous phase structures.

PACS numbers: 64.60.Cn, 65.40.-b, 61.66.Dk, 05.50.+q

I. INTRODUCTION

Calculation of stable structures of alloys, clusters, surfaces and molecules from first-principles is an important step towards design of materials with exceptional properties. Identification of stable alloy phases aid in construction of phase diagrams from first-principles. Because of the immense variety of phase structures, identification of stable structures at different combinations of the alloying elements is a non-trivial problem. While a first-principles approach based on density functional theory (DFT) provide a rigorous way for calculating formation energies of phase structures, the computational complexity of performing fully-relaxed calculations over the entire set of possible phase structures makes this method prohibitive. Techniques such as cluster expansion¹⁻⁷ and more recently, data mining techniques^{8,9} allow one to accelerate the search for stable phase structures.

In cluster expansion methods (CEM)¹⁻⁴, the relaxed energy of an atomic structure is represented as a linear combination of characteristic energies of clusters of atoms over a fixed lattice. The coefficients in the cluster expansion are computed using relaxed DFT energy calculations of few prototype structures¹. This method includes only ordering degrees of freedom as provided by different possible arrangements of atom types on a fixed parent lattice. Consequently, CEM fails in cases where the alloy phases have complex structures that are different from the superstructures of the underlying parent lattice (for example, FCC or BCC lattices) and exhibits convergence issues in cases where structural relaxation effects are dominant^{5,6} (for example, in alloys involving constituents with large size differences).

In another technique called multi-body expansion

(MBE), N -body potentials (or otherwise, cluster potentials¹⁰) constructed from ab-initio calculations are used to describe energies of arbitrary atomic structures as a function of atom positions. The total energy is represented as a summation over potentials of underlying isolated atom clusters in the structure, with series terms involving pair, three-body, four-body,..., N -body potentials. Up to third-order truncations of multi-body expansions have been previously used in related empirically derived potentials, namely the Gupta¹¹ and Murrell-Mottram (MM)¹²⁻¹⁵ potentials. Multi-body potentials focus on positional degrees of freedom and hence, explicitly handles structural relaxations during computations of stable phase structures. Structural relaxation effects can be treated in a cluster expansion approach by combining it with position-dependent potentials in the form of a hybrid cluster expansion⁶. Another method combining CEM and multi-body potentials was proposed recently for introducing positional degrees of freedom in a more generalized cluster expansion^{16,17}. However, building the N -body potentials from atomistic calculations is quite a challenging problem. Firstly, the number of clusters (and thus, the number of cluster energies that need to be calculated) in an N -atom system increases geometrically with the order of expansion (Fig. 1). Secondly, although the approach provides good convergence for rare-gas crystals, convergence of MBE is not smooth for metallic crystals^{18,21}. Absence of smooth convergence does not allow establishment of a hard cut-off for the series terms. Consequently, there has been no published reports of a multi-body expansion constructed directly from first-principle calculations.

This paper addresses these drawbacks by proposing a multi-body expansion with weighted terms. Using this

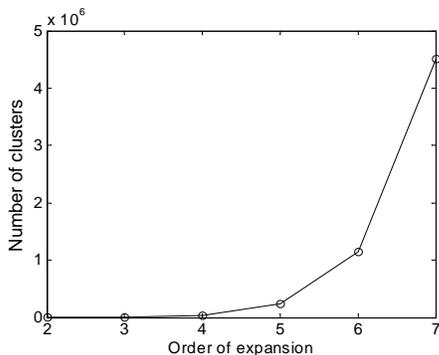


FIG. 1: Increase in the total number of clusters involved in a multi-body expansion for a 32-atom system with the order of expansion.

technique, it becomes possible to accurately compute energies of N -atom systems from knowledge of small cluster energies computed from first principles. The efficiency of this new method called weighted multi-body expansion (wMBE) is emphasized through examples in this work. Another contribution of this paper is a formal technique to rapidly calculate multi-body expansions using linear interpolation over tessellated cluster configurational spaces. MBE using interpolated energies is several orders of magnitude faster than using DFT calculations, since cluster energies are computed beforehand and are directly sampled from the database when computing the multi-body expansion.

A. Multi-body expansion methodology

Consider, for instance, a configuration of M -atoms (possibly all different), whose energy we intend to compute. We denote the total energy of this M -particle system using $E_P = E_P(X_1, X_2, \dots, X_M)$, where P is the order of the expansion used and the position \mathbf{R}_n of atom n is grouped together with the species of atom n denoted by an integer σ_n , $X_n = (\mathbf{R}_n, \sigma_n)$. As the order of labelling the M atoms is arbitrary, the form of $E_P(X_1, \dots, X_i, \dots, X_j, \dots, X_M)$ must be symmetric with respect to interchange of X_i and X_j .

From here on, we denote M as the total number of atoms in the system, $N = 1, 2, \dots, P$ denotes a N -atom cluster within the M -atom system. Further, $L = 1, 2, \dots, N$ denotes an arbitrary L -atom cluster within an N -atom cluster. The energy E_P of an M -particle system is represented as summation over a series of N -body

interaction potentials $V^{(N)}$ via

$$E_P(X_1, X_2, \dots, X_M) = \sum_{N=1}^P E^{(N)}(X_1, X_2, \dots, X_M),$$

$$E^{(N)} = \sum_{m_1=1}^M \sum_{m_2=m_1+1}^M \dots \sum_{m_N=m_{N-1}+1}^M$$

$$V^{(N)}(X_{m_1}, X_{m_2}, \dots, X_{m_N}). \quad (1)$$

The potentials can be inverted via the Mobius inversion approach from number theory. Mobius inversion has been used previously for extraction of potentials from energy data by Chen^{19,20} although in a different context. In the case of multi-body potentials $V^{(N)}$, the Mobius inversion is given as¹⁶:

$$V^{(N)}(X_1, X_2, \dots, X_N) = \sum_{L=1}^N (-1)^{N-L} \sum_{m_1=1}^N$$

$$\sum_{m_2=m_1+1}^N \dots \sum_{m_L=m_{L-1}+1}^N E^*(X_{m_1}, X_{m_2}, \dots, X_{m_L}) \quad (2)$$

Here, we denote the energies of L -atom clusters within the N -atom cluster as E^* . The above equation constitutes a unique definition of N -body potentials $V^{(N)}$ which are structure-independent because this equation does not carry any information about the environment of the atom clusters¹⁶. $V^{(2)}(X_i, X_j)$ can be understood as the excess energy attributed to pair interactions in an isolated atom pair i, j , i.e., $V^{(2)}(X_i, X_j) = E^*(X_i, X_j) - E^*(X_i) - E^*(X_j)$. Similarly, $V^{(3)}(X_i, X_j, X_k)$ can be understood as the excess energy attributed to three-body interactions in a isolated trimer (i, j, k) :

$$V^{(3)}(X_i, X_j, X_k) = E^*(X_i, X_j, X_k)$$

$$- (V^{(2)}(X_i, X_j) + V^{(2)}(X_j, X_k) + V^{(2)}(X_i, X_k))$$

$$- (E^*(X_i) + E^*(X_j) + E^*(X_k)). \quad (3)$$

Once the potentials $V^{(N)}$ have been constructed, they can be used to calculate the energy $E_P(X_1, X_2, \dots, X_M)$ for a M -atom system using Eq. (1). The first critical requirement of the technique is the knowledge of complete energy surface (cluster energies versus atom positions and types) of small isolated clusters of atoms (E^* , $L = 1, \dots, 5$) for building the potentials in Eq. (2). Secondly, it is essential that the expansion converges within a small-order of expansion (i.e. $P \leq 5$) for computational efficiency. These two aspects are addressed in the next two sections. Complete energy surface for small isolated clusters is created by mathematically defining the configurational space of clusters, tessellation of the space, computation of cluster energies on nodal points, followed by interpolation of cluster energies as described in the next section. Computational efficiency is improved through the use of weighted multi-body expansions as explained in section C. Efficiency can be further improved

by performing computations in parallel by distributing the M atoms involved in the loop index m_1 in Eq. (1) to different processors.

B. Construction of cluster energy surfaces

The basic idea of the approach to rapidly compute multi-body expansions of arbitrary systems is to build an interpolation function for the isolated cluster energies E^* from the pre-computed database. Given a set of n m -atom clusters represented as $\Theta = \{\xi_d^i\}_{i=1}^n$ in the d -dimensional configurational space, we try to build a smooth function that maps clusters to ab-initio energies, $f: \mathbb{R}^d \rightarrow \mathbb{R}$. In particular, we use an interpolant $\mathcal{I}f$ such that $\mathcal{I}f(\xi_d^i) = f(\xi_d^i)$, $\forall i = 1, \dots, n$.

The first step in this procedure will be to define the d -dimensional configurational space of an m -atom cluster. The positions of the atoms in the cluster are represented by the distance between atoms, $R_{ij} > 0$. For two-atom clusters ($m = 2$), the configurational space is one-dimensional, with each point x in the space representing a two-atom cluster with inter-atomic distance of $R_{12} = x$. As the number of atoms, m , in the cluster increases, the number of distances, R_{ij} necessary to completely and uniquely describe the cluster increases rapidly. Up to $m \leq 4$, clusters are uniquely represented by $\frac{1}{2}m(m-1)$ independent variables.

For example, the space of all possible three-atom clusters is three-dimensional as shown in Fig. 2(a). This space is a convex hull with 9 planes (symmetries not included) due to a linear set of constraints arising from three triangle inequalities of the form $R_{ij} + R_{jk} \geq R_{ik}$ that constrain the location of atoms in the three-atom cluster and the upper and lower cutoff used for possible cluster sizes in the database: $R_{ij} > l$ and $R_{ij} < u$ with $i, j, k = 1 \dots 3$. Cluster symmetries can be used to further reduce the space and consequently, reduce the number of energy calculations required. Figure 2(b) shows the reduced space accounting for symmetries ($R_{12} \leq R_{23} \leq R_{13}$) in the case where all 3 atoms are of the same type. Also shown in Fig. 2(a) is the tessellation of the configurational space of clusters. The energy of a cluster corresponding to each nodal point in the space is calculated and stored in the database. The plot of energy versus interatomic distance for a two-atom Pt cluster is shown in Fig. 3 with location of nodal points for two-atom clusters. Higher-dimensional spaces are adaptively tessellated as shown in Fig. 2(a) with a finer discretization of regions involving small clusters. The tessellation of the configurational space is carried out using n -dimensional Delaunay triangulation, as implemented in the qhull²² program. Tessellation generates elements (known as a simplex) over which local linear interpolation is carried out to find the energy of any other three-atom cluster within the space. The discretization and interpolation techniques are, in essence, same as those used in the popular finite element techniques for PDEs. Energy (E) of an arbitrary cluster

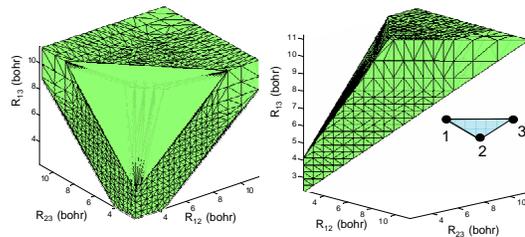


FIG. 2: (left) shows the space of all possible three-atom clusters within an upper and lower cutoff cluster size. This space represents a convex hull in 3D. (right) Use of symmetries (in the case where all three atoms are of one type, e.g. Pt-Pt-Pt clusters) can further reduce the space. The simplices used to perform local linear interpolation of energies are also shown. In 3D, the simplex is a tetrahedron.

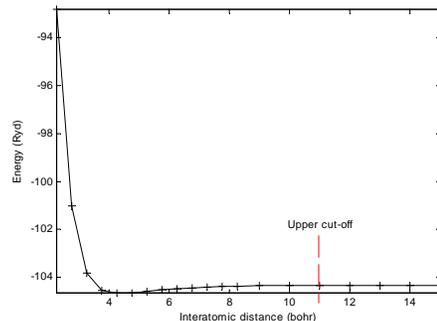


FIG. 3: The plot of energy versus interatomic distance for a two-atom Pt cluster. The location of nodal points in this one-dimensional configurational space and the upper cutoff used for calculations are indicated.

with cluster specifier $\xi_d^* = [\xi_1^*, \xi_2^*, \dots, \xi_d^*]$ in a tessellated d -dimensional configurational space is given as:

$$E = \alpha^T \mathbf{E}^e, \quad (4)$$

where \mathbf{E}^e is the vector containing energies at the nodes of the simplex within which ξ_d^* is located. α is obtained as $\alpha = \mathbf{A}^{-1}\mathbf{b}$, where $\mathbf{A} = [\mathbf{1}, \xi_1^e, \xi_2^e, \dots, \xi_d^e]^T$ and $\mathbf{b} = [1, \xi_1^*, \xi_2^*, \dots, \xi_d^*]^T$. Here, ξ_i^e denotes a vector containing the i^{th} coordinate value of all nodes in an element e . The element e is located by calculating α for every element in sequence and selecting the element e where all elements of $\alpha > 0$. This step becomes more time-consuming as the dimensionality of configurational space (hence, the number of elements) increases. Further, the geometry of the configurational space becomes more complex as the dimensionality of the configurational space increases. For example, the configurational space of a fourth-order cluster (excluding symmetries) involves 24 linear constraints and a quartic constraint.

The number of independent variables specifying a $m > 4$ atom cluster is given by $d = 3m - 6$ although $\frac{1}{2}m(m-1)$ variables are needed to uniquely define a cluster²³. An example of how cluster specifiers are determined for a 5-atom cluster is illustrated in Fig. 4. In this example,

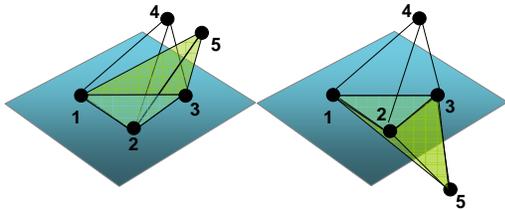


FIG. 4: In the case of a five-atom cluster, the locations of the fourth and fifth atoms can be fixed with respect to the plane formed by atoms 1 – 2 – 3 using the cluster specifiers $[R_{12}, R_{23}, R_{13}, R_{14}, R_{24}, R_{34}, R_{15}, R_{25}, R_{35}]$. However, these specifiers do not completely represent the cluster. The dependent variable in this case is R_{45} which can take one of possible two values based on the location of atom 5 either above or below the plane formed by atoms 1 – 2 – 3.

there are 9 independent variables and 1 dependent variable (R_{45}) that can take one of two values based on the location of the fifth atom. Thus, $m > 4$ cases present special difficulties associated with dependent variables. We address this issue by creating different configuration spaces corresponding to the values that each dependent variable takes. For the case of a 5-atom cluster, this means that two potentials need to be created, one for the case where atom 5 is above the plane formed by atoms 1 – 2 – 3 and another when it is below that plane.

For a binary AB system, all possible cluster configurational spaces are created for a given cluster size, e.g. for L -atom clusters, $L + 1$ energy databases (e.g. for $L = 2, 3$ databases containing $E^*(X_A, X_A), E^*(X_A, X_B), E^*(X_B, X_B)$) need to be generated. The upper and lower cutoff were selected by carefully analyzing the energies of two-atom clusters over a large range of R_{12} to locate an upper cut-off beyond which the interaction between atoms were not significant and a lower cutoff where the interaction energy is positive.

For Platinum with lattice parameter of $a = 7.5$ bohr, the lower cutoff of atom spacing in a cluster within the database was fixed as $R_{ij} > 0.3a$ and upper cutoff was fixed as $R_{ij} < 1.5a$. The plot of energy versus interatomic distance for a two-atom Pt cluster, from which the cut-offs were identified, is shown in Fig. 3. The cut-offs signify that clusters with $R_{ij} < 0.3a$ and $R_{ij} > 1.5a$ are not available in the database. During MBE calculations, energies of clusters containing such interatomic distances are approximated using the following means. For $R_{ij} < 0.3a$, cluster energies were given an artificial high value to signify that such configurations are not energetically feasible. For N -atom clusters with $R_{ij} > 1.5a$, the excess energy attributed to N -body interactions is assumed to be zero (i.e. $V^{(N)} \approx 0$, for N -atom clusters with an $R_{ij} > 1.5a$). This is mathematically equivalent to approximating the energies of large clusters using energies of smaller sub-clusters. For example, Fig. 5(a) shows the energy surface of three-atom Platinum clusters up to an

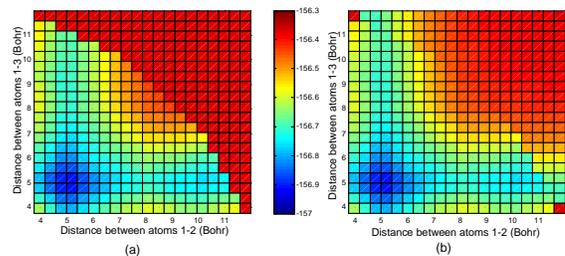


FIG. 5: Energy surface ($E^*(X_1, X_2, X_3)$) for 3-atom Pt clusters whose atoms are positioned at the vertices of a right angled triangle with line joining atoms X_2 and X_3 forming the hypotenuse. Figure (a) shows computed Platinum three-atom cluster energies, while (b) shows extension of energies beyond the cutoff using energies of smaller clusters ($E^*(X_i, X_j)$ and $E^*(X_i)$).

upper cutoff size of $1.5a$ and Fig. 5(b) shows the complete energy surface when the energies beyond the upper cutoff are approximated using two- and one-atom energies.

C. Weighted multi-body expansion

Multi-body expansion has been shown to work very well for rare-gases where the expansion is dominated by pair interactions making higher terms in the expansion negligible. Total energy of metallic systems, however, has significant contributions from higher-order interactions and the expansion has non-smooth convergence behavior¹⁸. Figure 6 shows the behavior of multi-body expansion for an eight atom (2 unit cell) FCC Platinum cluster that requires at least a 7th order expansion to capture the true energy. It is observed here that energies computed by including successively higher-orders of interaction, in fact, oscillate around the true energy. An ad-hoc numerical approach for estimating the true energies for the case in Fig. 6 will be to appropriately weight the energies obtained at different orders of multi-body expansion, which is akin to smoothing (or filtering) the energy oscillations in Fig. 6. Numerical experiments presented in the next section indicate that weighted MBE (wMBE) calculations lead to dramatic improvement to the convergence behavior of the multi-body expansion. In the wMBE approach, the energies up to a cut-off order of expansion P are weighted so that we reach as close to the true energy (E_M) of an M -atom system as follows:

$$E_M(X_1, X_2, \dots, X_M) = \alpha_1 E_1(X_1, X_2, \dots, X_M) + \alpha_2 E_2(X_1, \dots, X_M) + \dots + \alpha_P E_P(X_1, \dots, X_M) \quad (5)$$

The coefficients $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_P]^T$ are computed by solving the equation:

$$\alpha = C^+ E, \quad (6)$$

where, E are the true energies of q M -atom clusters ($X^i, i = 1, \dots, q$) computed with self-consistent DFT

calculations. Each row of \mathbf{C} contains the energies $[E_1, E_2, \dots, E_P]^{(i)}$ obtained from multi-body expansion of each of these clusters (where $E_p^{(i)} = E_p(\mathbf{X}^i)$). \mathbf{C}^+ is the pseudo-inverse of matrix \mathbf{C} . The technique to obtain coefficients α is thus, similar to the method of Connolly and Williams¹ used for cluster expansions. In their technique, truncation of the expansion is based on which clusters are important, for example, in FCC crystals where only clusters containing nearest neighbors are important, the series is truncated at fourth-order. In the case of the wMBE, however, the cutoff of the order of interactions needs to be identified through numerical experiments as will be demonstrated in the next section.

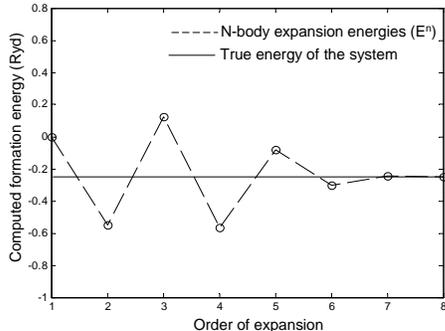


FIG. 6: Convergence of the many-body energy expansion of an eight-atom FCC Platinum cluster requires at least a 7th order expansion to reasonably capture the true energy.

Before proceeding to examples, we summarize the steps involved in the overall algorithm as follows:

1. *Offline calculations (steps 1-4): Constructing a database.*

Generate coordinates ($\{\xi_d^i\}_{i=1}^n$) for sampling the configurational space of all cluster sizes involved. For example, each three-atom cluster (1-2-3) corresponds to the coordinate $\xi_3^i = (R_{12}, R_{23}, R_{13})$ in the configurational space where e.g. R_{12} is the interatomic distance between atoms 1 and 2. During this step, various constraints based on geometry or symmetry are used to reduce the number of nodes in the configurational space.

2. For an L -atom cluster of a binary system, coordinates for $L + 1$ configurational spaces need to be created during step 1. Each configurational space corresponds to a different atom type list, e.g. for $L = 2$ atom clusters of a binary alloy $A - B$, configurational spaces for clusters of types $A - A$, $A - B$ and $B - B$ need to be generated.

3. Perform tessellation of coordinates in all configurational spaces and store nodal coordinates and element-node lists in the database.

4. Generate input files and perform self-consistent DFT calculations to compute energies ($E^*(\xi_d^i)$) at

nodal locations of all configurational spaces. Energies from the DFT calculation are read and stored in databases, one corresponding to each configurational space.

5. *Calculation of MBE coefficients.*

Compute self-consistent ab-initio calculations to compute energies (to obtain \mathbf{E} in Eq. (6)) of a few (three or four) different N -atom configurations.

6. Compute energies using MBE with increasing orders of expansion and obtain $[E_1, E_2, \dots, E_P]^{(i)}$ for each N -atom configuration used in step 5. During multi-body expansion, potentials ($V^{(N)}$) are created using Eq. (2) on the fly, using cluster energies E^* obtained by interpolating from the database constructed in steps (1-4). The steps involved to compute cluster energy, E^* , of an arbitrary cluster are:

(a) Locate the cluster in the corresponding configurational space. For example, a three-atom cluster of type $A - A - B$, is located at the coordinate $\xi_i^3 = (R_{12}, R_{23}, R_{13})$ in the three dimensional configurational space of $A - A - B$ type.

(b) Identify the element in which the cluster is located and perform linear interpolation using known energies at nodal values in that element using Eq. (4).

(c) Energies of clusters that are not available in the database are approximated using the methods detailed at the end of Section B.

7. Compute the coefficients α of weighted MBE using Eq. (6). Perform tests for convergence by comparing energies predicted by wMBE with ab-initio calculation for few other configurations of N -atoms.

8. *MBE calculations for arbitrary N -atom configurations.*

The converged weighted expansion can be now be employed for computing energies of other N -atom systems using Eq. (5) and Eq. (1). During calculations, cluster energies E^* are again interpolated from the database as in step 6.

D. Results for metallic systems

1. *Extrapolatory performance of wMBE approach:* In the first test case, energies predicted by multi-body expansion are compared with true energies obtained using the embedded atom potential of Sutton and Chen²⁴ for Platinum atom clusters. Convergence of the expansion is tested using exact cluster energies (without performing interpolation). Atom configurations used in these cases correspond to $n_x \times n_y \times n_z$ clusters with n_i unit cells located in the i^{th} direction.

Figure 7 shows the energies obtained for an isolated $4 \times 1 \times 1$ (16-atom) cluster of Platinum computed using

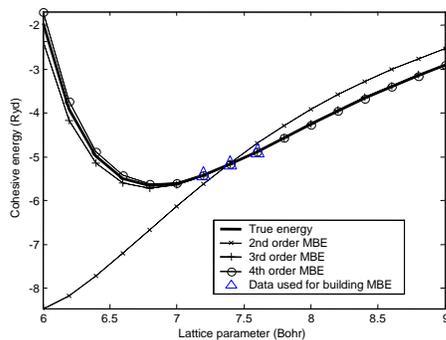


FIG. 7: Comparison of true energies of a $4 \times 1 \times 1$ (16-atom) FCC Platinum cluster with that predicted by multi-body expansion. Weights in the multi-body expansion were computed using 3 energies at lattice parameters of 7.2, 7.4 and 7.6 bohr.

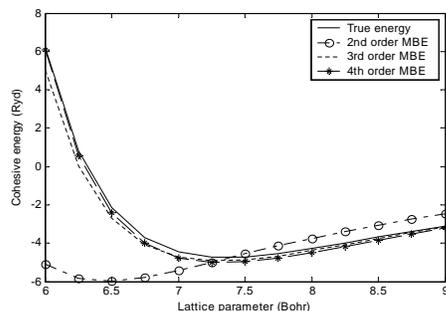


FIG. 8: Comparison of true energies of a $4 \times 1 \times 1$ (16-atom) FCC Platinum cluster with MBE expansion results for an extrapolatory case of a FCC lattice where the face centered atom in the $x - y$ plane and $y - z$ plane in the FCC basis are translated by $(-0.1, -0.1, 0)$ and $(0, 0.1, 0.1)$, respectively in crystal coordinates.

2^{nd} , 3^{rd} and 4^{th} order wMBE and the true energies. In all cases, the parameters α were computed using $4 \times 1 \times 1$ Pt clusters using just 3 energies at lattice parameters of 7.2, 7.4 and 7.6 bohr as indicated in Fig. 7. Energies increase linearly within this range of lattice parameters. In spite of this, predicted energies from the 3^{rd} and 4^{th} order multi-body expansions exactly capture the parabolic nature of the true energy profile. As a test of the extrapolatory performance of wMBE, we perturb the face centered atoms in the $x - y$ plane and $y - z$ plane of the FCC basis by $(-0.1, -0.1, 0)$ and $(0, 0.1, 0.1)$, respectively in crystal coordinates. In spite of the large changes in energy resulting from this perturbation, the expansion built previously for a FCC cluster is able to reproduce the energy profile of this distorted cluster accurately (Fig. 8).

2. *Convergence of wMBE in extrapolatory cases:* Figure 9 shows the energies predicted at various lattice parameters using 2^{nd} and 3^{rd} order wMBE for an isolated $2 \times 2 \times 1$ FCC Platinum cluster. In this case, the parameters α were originally computed using $2 \times 2 \times 1$ FCC clusters of Pt using 11 lattice parameters between 6 bohr to 8 bohr in the increments of 0.2 bohr. Although Fig. 9

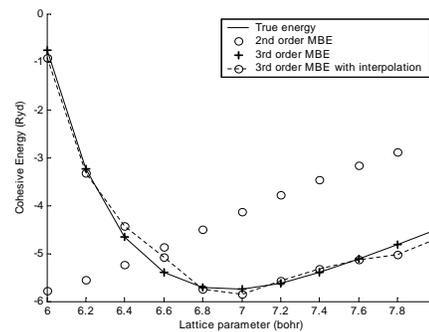


FIG. 9: Comparison of true energies obtained for a $2 \times 2 \times 1$ (16-atom) FCC Platinum cluster with energies computed using 2^{nd} and 3^{rd} order wMBE.

shows that the third-order expansion is adequate to capture the true energy profile, use of higher-orders of expansion improve the performance in extrapolatory cases. Figure 10 depicts the performance of 3^{rd} , 4^{th} and 5^{th} order MBE in an extrapolatory case where the face centered atom in the $x - y$ plane of the FCC basis is translated by $(-0.1, -0.1, 0)$ in crystal coordinates. Figure 11 shows the decrease in the l^2 norm error in energies predicted with increasing order of multi-body expansion. Several other numerical experiments of this kind indicate that the wMBE approach captures the energy profile for any random configurations of N -atom Pt clusters, and thus, has potential applications in NVE or NVT atomistic simulations. The weighting procedure aims to average out the extraneous energy contributions (eg. surface energies) arising due to lack of environment in isolated clusters. The limitation in the procedure is that a change in number of atoms (N) simulated necessitates re-calibration of MBE coefficients. Figure 9 shows the energy variation with lattice parameter obtained from an MBE expansion calculated using cluster energies interpolated from a database. For interpolation, the second-order configurational space is discretized into 20 linear elements (21 nodes) and the third-order configurational space (including symmetries) was approximated using 16374 tetrahedral elements on which energies were calculated at 3191 nodal locations. Although discretization and linear interpolation introduce errors in calculation of energies, it is seen that the technique still reasonably captures the energy profile and the energy minima. The advantage of interpolation approach is that it is order of magnitude faster since cluster energies are computed beforehand and are directly sampled from the database during simulations.

3. *wMBE using interpolated energies from ab-initio calculations:* Figure 12 shows structure optimization to find the lattice constants for FCC Platinum system using *interpolated* energies of clusters computed from first principles DFT calculations. Since MBE inherently uses non-periodic configurations, energies of periodic structures are computed by considering supercells

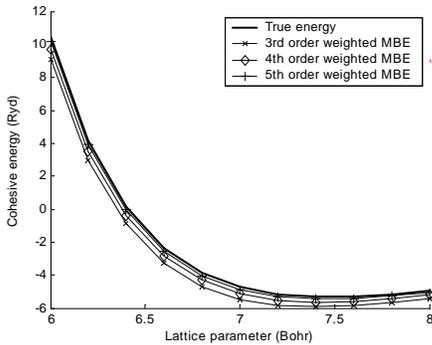


FIG. 10: Comparison of true energies obtained for a $2 \times 2 \times 1$ (16-atom) FCC Platinum cluster with energies computed using 3^{rd} , 4^{th} and 5^{th} order wMBE for an extrapolatory case of a FCC lattice with the face centered atom in the $x-y$ plane of the FCC basis is translated by $(-0.1, -0.1, 0)$ in crystal coordinates.

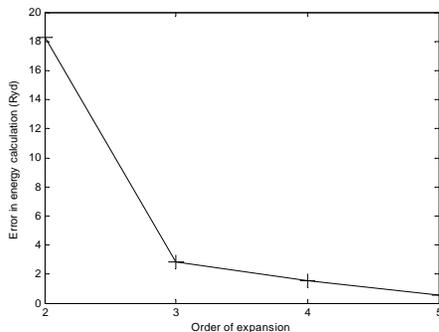


FIG. 11: Decrease in the l^2 norm error in energies with increasing order of multi-body expansion for the extrapolatory case in Fig. 10.

with true energies E used for fitting Eq. (5) obtained from self-consistent DFT calculations of a periodic unit cell. In this example, a $5 \times 5 \times 5$ (500-atom) FCC cluster is considered. The variation of cohesive energy ($E_c(X_1, \dots, X_m) = E^*(X_1, \dots, X_m) - \sum_{i=1}^m E^*(X_i)$) of 3-atom clusters with interatomic distances (R_{12}, R_{23}, R_{13}) is shown on the configurational space (accounting for symmetry) in Fig. 12. The configurational space is discretized into 4609 tetrahedral elements on which linear interpolation is carried out. Ab-initio energy data were computed on 1027 nodal locations. Figure 13 shows comparison of the energies computed using 3^{rd} and 4^{th} order wMBE with the true energies. Coefficients in the multi-body expansion were generated using three ab-initio energy calculations of a periodic FCC Platinum lattice with lattice parameters of 6.5, 8.5 and 9.0 bohr. It is seen from Fig. 13 that the energy profile is well captured using the technique within expected error bounds as discussed later in this section. The significant advantage of using interpolated energies is that it does not utilize any significant computational resource. This is due to the fact that all heavy ab-initio calculations are performed beforehand

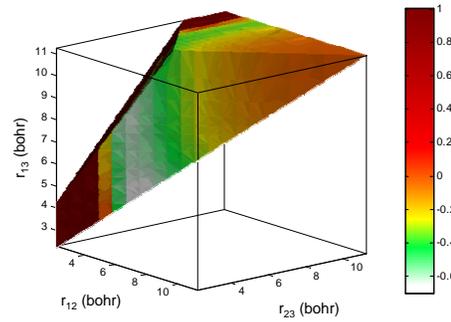


FIG. 12: The three-body cluster configurational space for Platinum colored by the cohesive energies ($E_c(X_1, X_2, X_3) = E^*(X_1, X_2, X_3) - \sum_{i=1}^3 E^*(X_i)$) computed from ab-initio simulations.

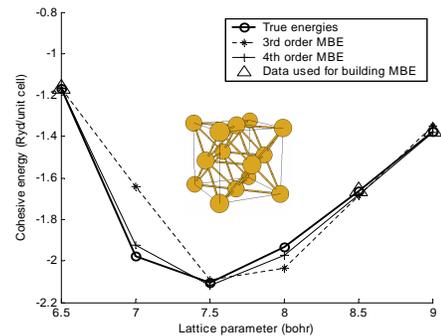


FIG. 13: Comparison of variation of energies with lattice parameter for a periodic FCC Pt lattice with wMBE calculations involving cluster energy interpolation.

and the data is stored for interpolation.

4. Analysis of accuracy of wMBE with interpolated ab-initio energies

The main sources of error in the wMBE procedure are the errors involved in the interpolation of energies from the database, fitting weighting coefficients and convergence accuracy for the ab-initio energy data. The maximum interpolation error over any element (including higher dimensional elements) is tightly bounded by $c_t r_{mc}^2$, where the absolute curvature of the true ab-initio energy surface is bounded in each element t by a constant $2c_t$ and r_{mc} is the minimum containment radius of an element. For the 2-atom Pt cluster energy surface, the maximum interpolation error was 0.03 mRyd. Although this error cannot be completely eliminated, we use smaller element sizes in the regions where large energy variations are expected in order to reduce the interpolation error for larger clusters. The convergence accuracy of self-consistent DFT calculations of small clusters was within 0.01 mRyd in all cases. In order to study the error in fitting MBE weights, we carried out a leave-one-out cross-validation (CV) procedure. Here, the error in reproduction of energies is studied by fitting the energy with $N - 1$ clusters and computing the error in reproduction of energy of the left-out cluster (E_i). The process is

repeated with every single cluster used once as a left-out cluster. The CV error is computed as the mean error $\frac{1}{N} \sum_1^N |E_{true} - E_i|$. Compared to statistical estimates such as variance, CV error provides a more reliable estimate of future performance of wMBE when energies of new clusters need to be predicted.

Ab-initio energies of $N = 300$ randomly generated 24-atom Pt clusters were used for testing the accuracy of the wMBE procedure. The 24 atoms were randomly placed at grid points spaced 7 bohr apart in each direction over a cube of 105 bohr length and ab-initio energies (E) for use in Eq. (6) are computed. The mean CV error for third-order MBE was found to be 0.381 Ryd (15.9 mRyd per atom). The mean CV error during cross validation for fourth-order expansion reduces to 0.121 Ryd (5.04 mRyd per atom). The average cohesive energy per atom for the complete data set was 312.4 mRyd. This demonstrates convergence towards ab-initio energies, although the error may still be significant for modeling phenomena such as phase transformations where accuracy in the order of mRyd may be required.

5. *Convergence of wMBE for a binary system (α -alumina Al_2O_3):* A multi-body expansion is constructed for α -Alumina (Al_2O_3) system using cluster energies computed using the Streitz-Mintmire (SM) model²⁵. Streitz-Mintmire potential is a many-body functional that merges electrostatic potential with an embedded-atom potential to describe metal-oxide energies. α -Alumina (Al_2O_3) has a rhombohedral primitive unit cell and is described in space group $R\bar{3}c$ (no.167) with two lattice parameters a, b . The lattice parameter a is varied while b is fixed at 0.4856 bohr. Figure 14 plots the variation of energies, computed using wMBE, as a function of lattice parameter a for a $2 \times 2 \times 1$ cluster of α -Alumina. The true energies as computed by the SM model at each lattice parameter are also shown. Four energies at lattice parameters $a = 7.0, 7.2, 7.4$ and 7.6 bohr were used to compute the MBE coefficients. Within this range of lattice parameters, energies increase linearly as indicated in Fig. 14. In spite of this, a fourth-order expansion is able to represent the curvature of the α -Alumina energy profile predicted by the Streitz-Mintmire (SM) model. In contrast to the FCC Pt case in Fig. 7, predicted energies from the 3^{rd} order multi-body expansion is not able to predict the energy minima, while the fourth-order predicts the lattice parameter $a = 6.6$ bohr accurately. Instead of the SM model, ab-initio calculations of isolated cluster energies (E^*) could have been used. Since we compute energies of isolated clusters by approximating a periodic lattice, care must be taken to avoid the influence of lattice Coulomb potential on the ionic Al-O cluster (due to finite size effects) by using a large enough unit cell. DFT calculations were avoided in this example due to the computational complexity of handling a large number of plane waves because of the sharply peaked valence states in oxygen and requirement of a large unit cell. wMBE of a binary metallic system that uses ab-initio calculations is reported in the next example.

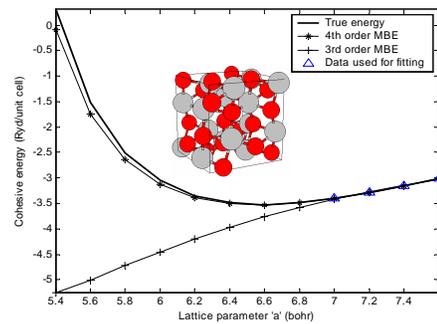


FIG. 14: Comparison of variation of energies with lattice parameter for a $2 \times 2 \times 1$ supercell of α -alumina (space group $R\bar{3}c$) using 3^{rd} and 4^{th} order wMBE. The true energies and the energies used for computing MBE coefficients are indicated.

6. *wMBE of a binary ($Au-Cu$) system using interpolated ab-initio energies:* This example demonstrates structure optimization to find the lattice constants for FCC $CuAu_3$ system (space group $Pm\bar{3}m$, no. 221) using interpolated energies of clusters computed from first principles DFT calculations. As in the case of FCC Pt, the energies E used for fitting Eq. (5) are obtained from self-consistent DFT calculations of a periodic unit cell. A $6 \times 6 \times 6$ (864-atom) FCC cluster is considered to approximate the periodic lattice, and MBE expansion is constructed using energies interpolated from the tessellated configurational space. As an example, the cohesive energy (E_c) variation with cluster specifiers (R_{12}, R_{23}, R_{13}) in the configurational space for 3-atom $Cu - Cu - Au$ and $Cu - Au - Au$ clusters is shown in Fig. 15(a) and (b), respectively. Apart from the 9 constraints discussed in section B, the inequalities $R_{23} < R_{13}$ and $R_{12} < R_{13}$, respectively, are additionally used to account for cluster symmetries in the space shown in Fig. 15(a) and (b). The lower and upper cutoffs used for constructing these spaces were 2.19 bohr and 10.95 bohr, respectively. For single atom-type clusters of copper or gold, the upper and lower cutoffs were fixed at 0.3 and 1.5 times the lattice parameters of pure FCC Cu and Au lattices.

Figure 16 shows comparison of the energies computed using 3^{rd} and 4^{th} order wMBE with the true energies. Coefficients in the multi-body expansion were generated using three ab-initio energy calculations of a periodic FCC $CuAu_3$ lattice with lattice parameters of 8.6, 8.7 and 8.8 bohr. Similar to the Al_2O_3 case, the 3^{rd} order multi-body expansion is not able to capture the energy profile of FCC $CuAu_3$ whereas a fourth-order expansion provides a reasonable approximation of the energy profile. wMBE approach allows computation of the energy of large systems with accuracy subject to the errors discussed previously. Cross-validation accuracy for this system using a similar procedure as described before was also carried out. We employed 300 random clusters of 24-atom $CuAu_3$ clusters for testing the accuracy of the wMBE procedure. The 24 atoms were

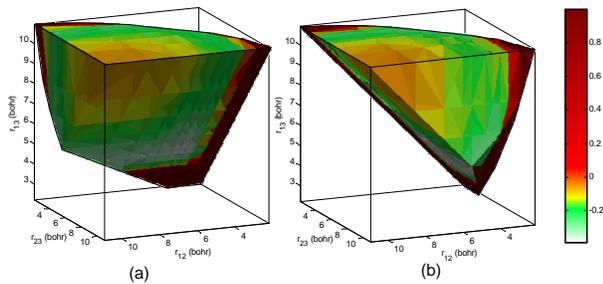


FIG. 15: The three-body cluster configurational space for (a) $Cu - Cu - Au$ and (b) $Cu - Au - Au$ colored by the cohesive energies (as defined in Fig. 12) computed from ab-initio simulations. The two spaces have different geometries due to different underlying symmetries.

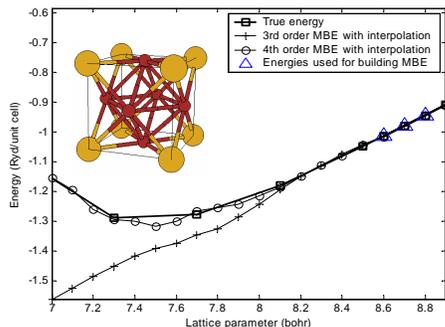


FIG. 16: Comparison of variation of energies with lattice parameter for a periodic FCC $CuAu_3$ lattice with wMBE calculations involving cluster energy interpolation.

randomly placed at grid points spaced 7.5 bohr apart in each direction over a cube of 112.5 bohr length. A cross validation error of 0.187 Ryd (7.8 mRyd per atom) was achieved when a fourth-order expansion was used. The average cohesive energy per atom for the complete data set was 207.1 mRyd. This error may be significant for modeling phenomena such as phase transformations, but wMBE is a good replacement for empirical potentials in several other multiscale modeling applications where reasonable accuracy is required. Although use of higher (5+) body interactions is expected to improve the fit, it greatly increases computational overhead in tessellation and data generation. We are currently working on the use of data-adaptive hierarchical interpolation to address this issue.

E. Conclusions

Developments presented here advance the existing state of the art in multi-body expansion technique for representation of energies of alloy systems through the following new contributions:

- The convergence characteristics of multi-body ex-

pansions (MBE) were improved by weighting energies obtained from various orders of atom interactions in a new method called weighted multi-body expansion (wMBE).

- In contrast to methods such as cluster expansion that involve the ordering degrees of freedom, wMBE focuses on the positional degrees of freedom. This allows one to explicitly model structural relaxations.
- Database interpolation techniques are demonstrated for accelerating computation of energies using multi-body expansions. For the first time in literature, multi-body expansions were computed directly from ab-initio energies of small clusters to model energies of Platinum and a binary alloy (Au-Cu) system. The quality of the expansion was quantified using leave-one-out cross-validation technique.
- The technique involves considerably lesser computational cost, with no requirement of periodicity, and hence, could be used to perform more accurate NVE or NVT molecular simulations of metallic clusters and complex phase structures compared to other commonly used position-dependent potential approximations. We are currently working on data-adaptive hierarchical interpolation which would allow us to build higher (5+) order potentials that would lead to improved accuracy.

APPENDIX A: AB-INITIO CALCULATIONS

Ab initio electronic-structure calculations were carried out using density functional theory in the local density approximation, as implemented in the PWscf package, using Perdew-Zunger parameterization of the exchange correlation energy and Rabe-Rappe-Kaxiras-Joannopoulos²⁶ (ultrasoft) pseudopotential. Kohn-Sham orbitals were expanded in a plane wave basis up to an energy cutoff calculated to ensure convergence. Brillouin zone integrations were carried out using single k -point calculation and Methfessel-Paxton first-order spreading²⁷. The cell size is taken to be sufficiently large to effectively simulate an isolated cluster. For Platinum, the energy cutoff was 244.8 eV, the cell size was taken as four times the maximum size of the cluster.

ACKNOWLEDGMENTS

The authors acknowledge the support of the Materials Science Division of the Army Research Office and of the Computational Mathematics Program of the Air Force Office of Scientific Research.

-
- * Electronic address: njz1@cornell.edu
- ¹ J.W.D. Connolly and A.R. Williams, *Phys. Rev. B* **27**, 5169–5172 (1983).
 - ² J.M. Sanchez, F. Ducastelle and D. Gratias, *Physica A* **128**, 334–350 (1984).
 - ³ A. Zunger, NATO Advanced Study Institute on Statics and Dynamics of Alloy Phase Transformations (ed. P. Turchi and A. Gonis), New York: Plenum, 1994.
 - ⁴ D. de Fontaine, in *Solid State Physics*, edited by H. Ehrenreich and D. Turnbull, Academic Press, New York, 1994, Vol. 47, p. 33.
 - ⁵ Z.W. Lu, S.H. Wei, A. Zunger, S. Frota-Pessoa and L.G. Ferreira, *Phys. Rev. B* **44**, 512–544 (1991).
 - ⁶ H.Y. Geng, M.H.F. Sluiter and N.X. Chen, *Phys. Rev. B* **73**, 012202(1–4) (2006).
 - ⁷ A. van de Walle and G. Ceder, *J. of Phase Equil.* **23**(4), 348–359 (2002).
 - ⁸ C.C. Fischer, K.J. Tibbetts, D. Morgan and G. Ceder, *Nature materials* **5**, 641–646 (2006).
 - ⁹ S. Curtarolo, D. Morgan, K. Persson, J. Rodgers and G. Ceder, *Phys. Rev. Lett.* **91**, 135503 (1–4) (2003).
 - ¹⁰ A.E. Carlsson, *Solid State Physics* (vol. 43) (eds. H. Ehrenreich and D. Turnbull), Boston, MA: Academic (1990).
 - ¹¹ R. P. Gupta, *Phys. Rev. B* **23**, 6265–6270 (1981).
 - ¹² Y. Li, E. Blaisten-Barojas and D.A. Papaconstantopoulos, *Phys. Rev. B* **57**, 15519–15532 (1998).
 - ¹³ J.N. Murrell and R.E. Mottram, *Mol. Phys.* **69**, 571–585 (1990).
 - ¹⁴ J.-Y. Fang, R. L. Johnston and J. N. Murrell, *Mol. Phys.* **78**(6), 1405–1422 (1993).
 - ¹⁵ H. Cox, R.L. Johnston and J.M. Murrell, *J. Solid Stat. Chem.* **145**(2), 517–540 (1999).
 - ¹⁶ R. Drautz, M. Fahnle and J.M. Sanchez, *J. Phys.: Condens. Matter* **16**, 3843–3852 (2004).
 - ¹⁷ M. Fahnle, R. Drautz, F. Lechermann, R. Singer, A. Diaz-Ortiz and H. Dosch, *Phys. Status Solidi B* **242**, 1159–1173 (2005).
 - ¹⁸ B. Paulus, K. Rosciszewski, N. Gaston, P. Schwerdtfeger and H. Stoll, *Phys. Rev. B* **70**, 165106–165115 (2004).
 - ¹⁹ N.X. Chen, *Phys. Rev. Lett.* **64**, 1193–1195 (1990).
 - ²⁰ N.X. Chen and G.B. Ren, *Phys. Rev. B* **45**, 8177–8180 (1992).
 - ²¹ B. Paulus, *Physics Reports* **428**, 1–52 (2006).
 - ²² Qhull program (<http://www.qhull.org/>).
 - ²³ J.W. Martin, *J Phys C. Solid state Phys.* **8**, 2837–2857 (1975).
 - ²⁴ A.P. Sutton and J. Chen, *Philos. Mag. Lett.* **61**(3), 139–146 (1990).
 - ²⁵ F.H. Streitz and J.W. Mintmire, *Phys. Rev. B* **50**, 11996–12003(1994).
 - ²⁶ A.M. Rappe, K.M. Rabe, E. Kaxiras, and J.D. Joannopoulos, *Phys. Rev. B* **41**, 1227–1230 (1990).
 - ²⁷ M. Methfessel and A.T. Paxton, *Phys. Rev. B* **40**, 3616–3621 (1989).

Towards the construction of fully transferable multi-atom potentials

Ilias Bilonis^{a,b}, Nicholas Zabaras^{a,b}

^a*Material Process Design and Control Laboratory, Sibley School of Mechanical and Aerospace Engineering, 101 Frank H.T. Rhodes Hall, Cornell University, Ithaca, NY 14853-3801, USA*

^b*Center for Applied Mathematics 657 Frank H.T. Rhodes Hall Cornell University Ithaca, NY 14853, USA*

Abstract

A Bayesian scheme to fit Potential Energy Surface of clusters of N atoms is proposed using a permutationally invariant polynomial basis. The evidence approximation is employed to fit the missing prior parameters and identify the length scale. Distance geometry techniques are introduced to efficiently sample the configuration space. The Bayesian variance is used to quantify the informational content of each point in the configuration space leading to an efficient adaptive scheme that minimizes the required number of expensive ab initio calculations. Objective stopping criteria are provided.

Keywords: PES interpolation, model selection, invariant polynomial basis

1. Introduction

The potential energy surface (PES) plays a central role in the computational simulation of all types of atomic interactions of interest. Once the PES is constructed, Molecular Dynamics (MD) or Monte Carlo (MC) methods can be employed to investigate the system's dynamical behavior. Of course, the accuracy of such simulations depends crucially on the accuracy of the PES used. The ideal PES is the so-called Born-Oppenheimer PES obtained by solving the Schrödinger equation using the adiabatic approximation [1]. Di-

Email addresses: ib227@cornell.edu (Ilias Bilonis), zabaras@cornell.edu (Nicholas Zabaras)

rectly using the ab initio PES in simulation is computationally infeasible for all but extremely simple systems.

As the number of atoms in the system increases, simple functional forms based on physical models are a necessary choice. Their parameters are fitted using a small set of experimental data: bond energies, bond distances and angles, elastic moduli, vibrational frequencies etc. Such PES give qualitative descriptions of the system and their applicability depends closely on what range of experimental data was used to fit their parameters, i.e. they are not transferable.

Recently it has become possible to obtain the PES directly from ab initio calculations for relatively small and not too complex systems. The basic problems addressed in the literature consist of 1) determining the important areas of the configuration space, 2) finding the minimum number of the configuration space points required to obtain an accurate PES and 3) fitting the electronic energies to an analytic model. The first problem is usually addressed by employing classical trajectories whose initial points are selected to match the distribution of these variables under the conditions of the experiments being investigated [2]. Ab initio calculations are performed on a set of system configurations along these trajectories. Alternatively, an approximate empirical PES can be used to initiate the trajectory sampling of the configuration space [3]. The second problem, testing the convergence of the PES, is performed by computing various dynamical properties of the system and examining their invariancy with respect to the database size [4]. Finally several methods have been proposed to accurately fit ab initio databases. Many methods assume parametrized analytical forms for the surface. Such are the many-body expansion method [5] and the recently successful multinomial expansion method [6, 7, 8]. Other, basis free, methods are a) moving Shepard interpolation techniques [4, 9, 10, 11, 12, 13], b) reproducing kernel Hilbert spaces [14], c) interpolating moving least squares [15, 16, 17, 18, 19, 20, 21] and d) Neural Networks methods [22, 23, 24, 25].

Despite the appeal and usefulness of the above mentioned techniques they all suffer from the curse of dimensionality: it is practically impossible to construct a PES for a multi-atom system. To overcome this barrier we employ the Multi-Body Expansion (MBE) technique [26]. MBE provides a systematic framework in which the total energy of a multi-atom system is represented as a summation over potentials of isolated clusters, with series terms involving pair, three-body, four-body, . . . , N -body potentials. This results in structure independent, fully-transferable many-body potentials [27].

The N -body potentials can be constructed using the Möbius transformation from the K -body PES, where $K = 1, 2, \dots, N$. However, building the N -body potentials is not an easy problem: 1) The order of the expansion is a priori unknown and although it is small for rare-gases, it is of relatively high order in metals [28]. 2) The accuracy of the N -body potential depends in a complicated manner on the accuracy of the K -body PES for $K = 1, \dots, N$. Both these problems pose daunting restrictions on the applicability of the method since they require a very accurate PES fitting scheme that utilizes as little as possible electronic structure calculations. Furthermore, for the constructed potentials to be of any use, the analytical form of the PES needs to be able to capture a wide range of the configuration space. Consequently, there have been no published reports of a multibody expansion constructed explicitly from first-principle calculations. The main goal of this work is to address exactly these issues.

To this end, we propose a Bayesian variant of the multinomial expansion method for the construction of the K -body PES. The multinomial expansion method provides an analytical functional form for the PES satisfying permutation invariance with respect to like atoms. Most schemes can account for permutation invariance only by explicitly replicating all possible permutations of each data point. As a result, permutation invariance of like atoms is learned from the data despite the fact that it constitutes a well-known property of any PES - a well-established prior knowledge. The accuracy of the fitting procedure is further enhanced by introducing a Linear Bayesian Regression scheme and the evidence approximation to optimize the scale parameter of the Morse-variables. This avoids overfitting and increases the predictive capabilities of the PES. An additional benefit of the Bayesian framework is that it provides us with a way to quantify the informational content of each point of the configuration space. Our lack of knowledge about the energy value of a particular point of the configuration space is associated with the variance of the Bayesian prediction. This information is then used to adaptively select new data points to be included into the fitting procedure until objective stopping criteria are met. Distance Geometry techniques ([29, 30, 31]) are introduced to facilitate the sampling of the configuration space. In our numerical examples, we demonstrate that this improved multinomial expansion greatly increases the accuracy of the obtained PES and reduces the number of required electronic calculations.

In Section 2.1 we introduce the distance matrix as our variables of choice to describe the configuration space and we discuss why it is necessary to

map them to the Morse variables. In Section 2.2 we discuss the multinomial expansion method and give a simple illustration on how the Monomial Symmetrization Approach (MSA) [8] can be used to construct a basis of permutationally invariant polynomials. In Section 2.3 we mathematically define the configuration space as the set of distance matrices satisfying certain bounds and introduce Distance Geometry techniques that allow us to sample from it. Section 2.4 describes the Linear Bayesian Regression scheme and Section 2.5 employs the evidence approximation in order to select all the missing parameters of the model. In Section 2.6 we propose objective measures to test the goodness of fit, and in Section 2.7 we show how one can use the Bayesian variance to adaptively select new configuration points. Finally in Section 2.8 we introduce the MBE methodology and show how our framework can be applied to the construction of the N -body potentials from the K -body PES.

2. Methodology

Consider a cluster of N atoms not necessarily of the same type. Atom i can be described by its cartesian coordinates $\mathbf{r}_i \in \mathbb{R}^3$ and its type σ_i . In order to simplify the notation we will refer only to the cartesian coordinates $\mathbf{R} = (\mathbf{r}_1, \dots, \mathbf{r}_N)$ but keeping always in mind that these are associated with particular atomic types.

2.1. Choice of variables

We choose to describe the configuration space in terms of interatomic distances. This is not the optimal choice since we use $\frac{1}{2}N(N-1)$ variables instead of the $3N-6$ independent degrees of freedom¹. However, interatomic distances provide a nice way to sample the configuration space using Distance Geometry techniques (see Section 2.3) and guarantee translation and rotation invariance.

Let d_{ij} be the distance between atoms i and j , i.e.

$$d_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|_2, \quad (1)$$

where $\|\cdot\|_2$ is the Euclidean norm of \mathbb{R}^3 . The symmetric matrix

$$\mathbf{D} = (d_{ij}), \quad (2)$$

¹ $3N - 6 = 3N - 3$ translations $- 3$ rotations

is called the *distance matrix* corresponding to the Cartesian coordinates $\mathbf{R} = (\mathbf{r}_1, \dots, \mathbf{r}_N)$. Since \mathbf{D} has a zero diagonal, it contains exactly $\frac{1}{2}N(N-1)$ elements.

Our plan is to expand the energy surface in a polynomial basis (see Section 2.2). Unfortunately using d_{ij} as the variables of the polynomials will be an unphysical choice since as atoms i and j move far apart ($d_{ij} \rightarrow \infty$), the energy would diverge. For this reason, we introduce one more coordinate change, the so called Morse variables $\exp(-\lambda d_{ij})$. To simplify the notation let $K = \frac{1}{2}N(N-1)$ and define the K -dimensional vector $\mathbf{z} = (z_1, \dots, z_K)$ by

$$z_1 = e^{-\lambda d_{12}}, z_2 = e^{-\lambda d_{13}}, \dots, z_K = e^{-\lambda d_{K-1,K}}. \quad (3)$$

The newly introduced parameter λ is related to the scale of the problem. We consider λ as an unknown parameter to be inferred from the data.

2.2. Choice of basis functions

Following the recent advances in the field, we introduce a polynomial basis on the \mathbf{z} variables that is invariant with respect to permutation of atoms of the same type. In [7] computational algebra techniques are used to find a basis on the space of invariant polynomials using commercial computational algebra software like MAGMA [32]. In this work, we follow the Monomial Symmetrization Approach (MSA) of [8] to construct the polynomials.

To introduce MSA consider a cluster of 4 atoms of the same type. In this case the variable describing the system, \mathbf{z} , is 6-dimensional. The energy is approximated by a sum of monomials up to degree k

$$\hat{E}(\mathbf{z}) = \sum_{a+b+c+d+e+f=0}^k C_{a,b,c,d,e,f} z_1^a z_2^b z_3^c z_4^d z_5^e z_6^f, \quad (4)$$

where a, b, c, d, e and f are all non-negative integers and $C_{a,b,c,d,e,f}$ is the coefficient of the monomial $z_1^a z_2^b z_3^c z_4^d z_5^e z_6^f$. This expression is clearly not invariant with respect to permutations of the atoms for arbitrary choices of coefficients. However, if we demand to have permutation invariance, it turns out that many of these coefficients must be the same. To give a concrete example, suppose we permute the atoms as

$$(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4) \rightarrow (\mathbf{r}_4, \mathbf{r}_2, \mathbf{r}_1, \mathbf{r}_3).$$

The corresponding permutation in the distance matrix elements is

$$(y_{12}, y_{13}, y_{14}, y_{23}, y_{24}, y_{34}) \rightarrow (y_{24}, y_{14}, y_{34}, y_{12}, y_{23}, y_{13})$$

and so the \mathbf{z} variables permute as

$$(z_1, z_2, z_3, z_4, z_5, z_6) \rightarrow (z_5, z_3, z_6, z_1, z_4, z_2).$$

We want $\hat{E}(\mathbf{z})$ to be invariant under such a permutation, i.e.

$$E(z_1, z_2, z_3, z_4, z_5, z_6) = E(z_5, z_3, z_6, z_1, z_4, z_2),$$

which can happen only the monomials $z_1^a z_2^b z_3^c z_4^d z_5^e z_6^f$ and $z_5^a z_3^b z_6^c z_1^d z_4^e z_2^f$ have the same coefficient, i.e.

$$C_{a,b,c,d,e,f} = C_{d,f,b,e,a,c}.$$

Listing all permutations will give us exactly which coefficients should be equal. The corresponding monomials will sum to form a single basis function invariant with respect to the permutation group. We call this procedure symmetrization and we denote the symmetrized sum of monomials by $\mathcal{S}[z_1^a z_2^b z_3^c z_4^d z_5^e z_6^f]$ and their common coefficient by $D_{a,b,c,d,e,f}$. Under this notation the energy is written as

$$\hat{E}(\mathbf{z}) = \sum_{a+b+c+d+e+f=0}^k D_{a,b,c,d,e,f} \mathcal{S}[z_1^a z_2^b z_3^c z_4^d z_5^e z_6^f], \quad (5)$$

where the summation is over exponents a, b, c, d, e and f that give a unique $\mathcal{S}[z_1^a z_2^b z_3^c z_4^d z_5^e z_6^f]$. To avoid unnecessary complications we denote these basis functions by $\phi_i(\mathbf{z}), i = 1 \dots M$ and write the energy as

$$\hat{E}(\mathbf{z}; \mathbf{w}) = \sum_{i=1}^M w_i \phi_i(\mathbf{z}), \quad (6)$$

where $w_i, i = 1, \dots, M$ are the corresponding coefficients. These polynomials

up to degree 3 are:

$$\begin{aligned}
\phi_1(\mathbf{z}) &= 1, \\
\phi_2(\mathbf{z}) &= z_1 + z_2 + z_3 + z_4 + z_5 + z_6, \\
\phi_3(\mathbf{z}) &= z_3z_4 + z_2z_5 + z_1z_6, \\
\phi_4(\mathbf{z}) &= (z_3 + z_4)(z_2 + z_5) + z_1(z_2 + z_3 + z_4 + z_5) + (z_2 + z_3 + z_4 + z_5)z_6, \\
\phi_5(\mathbf{z}) &= z_1^2 + z_2^2 + z_3^2 + z_4^2 + z_5^2 + z_6^2, \\
\phi_6(\mathbf{z}) &= z_3z_4(z_5 + z_6) + z_1(z_3z_4 + z_2z_5 + (z_2 + z_3 + z_4 + z_5)z_6) \\
&\quad + z_2(z_3(z_4 + z_5) + z_5(z_4 + z_6)), \\
\phi_7(\mathbf{z}) &= z_1z_2z_4 + z_1z_3z_5 + z_2z_3z_6 + z_4z_5z_6, \\
\phi_8(\mathbf{z}) &= z_1z_2z_3 + z_1z_4z_5 + z_2z_4z_6 + z_3z_5z_6, \\
\phi_9(\mathbf{z}) &= z_3z_4(z_3 + z_4) + z_2z_5(z_2 + z_5) + z_1z_6(z_1 + z_6), \\
\phi_{10}(\mathbf{z}) &= z_1^2(z_2 + z_3 + z_4 + z_5) + z_5(z_3^2 + z_4^2 + (z_3 + z_4)z_5) \\
&\quad + z_1(z_2^2 + z_3^2 + z_4^2 + z_5^2) + (z_3^2 + z_4^2 + z_5^2)z_6 + (z_3 + z_4 \\
&\quad + z_5)z_6^2 + z_2^2(z_3 + z_4 + z_6) + z_2(z_3^2 + z_4^2 + z_6^2), \\
\phi_{11}(\mathbf{z}) &= z_1^3 + z_2^3 + z_3^3 + z_4^3 + z_5^3 + z_6^3.
\end{aligned}$$

The whole procedure can be carried out using Zhen Xie's code which can be found at [33]. One may notice from the basis functions given above, that many monomials of the z_i variables appear again and again. For example z_1^2 appears in ϕ_5 and in ϕ_{10} . This fact is more pronounced in polynomials of higher degree. Calculating each basis function from scratch would result in repetitive calculations of the same monomials and hence it would be highly impractical. One of the extremely useful features of the above mentioned program is that it uses the algorithms described in [8] to break down the computation of the basis functions in reusable parts. This significantly reduces the computational time needed for their calculation. We have developed a Python script that uses the output of this program to produce C++ and Matlab code.

2.3. Sampling the configuration space

We manage to sample the configuration space using Distance Geometry techniques which have been applied successfully to nuclear magnetic resonance (NMR) structural determination problems [34]. For a review of the topic and further information on the algorithms described here see [29].

Algorithm 1 USAMPLE(\mathbf{L}, \mathbf{U}): Samples a distance matrix from the configuration space $\mathcal{C}(\mathbf{L}, \mathbf{U})$.

Require: \mathbf{L}, \mathbf{U} and $\text{unif}()$ (a random number generator).
 {Construct a matrix $\tilde{\mathbf{D}}$ satisfying the bounds and the triangle inequality}
for $i = 1$ to $N - 1$ **do**
 for $j = i + 1$ to N **do**
 {Sample an element i, j uniformly within the bounds.}
 $d_{ij} \leftarrow l_{ij} + \text{unif}() * (u_{ij} - l_{ij})$
 {Update bounds}
 $u_{ij} = \tilde{d}_{ij}$
 $l_{ij} = \tilde{d}_{ij}$
 {Enforce triangle inequality}
 FLOYD(\mathbf{L}, \mathbf{U})
 end for
end for
 {Project $\tilde{\mathbf{D}}$ to the closest distance matrix \mathbf{D} }
 $\mathbf{D} \leftarrow \text{MME}(\tilde{\mathbf{D}})$
if \mathbf{D} is not within the initial bounds **then**
 Go to the beginning of the algorithm.
end if
return \mathbf{D}

We start by defining the configuration space. Not every possible configuration is of physical interest to us - e.g. situations where two atoms are very close together or very far apart. Furthermore, we may wish to restrict our attention to regions of the configuration space of particular interest, like average bond lengths or angles measured in experiments. A very natural way to impose such constraints to the configuration space is provided via the Lower Cut-Off Matrix:

$$\mathbf{L} = (l_{ij}) \tag{7}$$

and the Upper Cut-Off Matrix:

$$\mathbf{U} = (u_{ij}). \tag{8}$$

We define the *configuration space* to be the collection of distance matrices that lie between these two bounds, i.e.

$$\mathcal{C}(\mathbf{L}, \mathbf{U}) := \{\mathbf{D} : \mathbf{D} = (d_{ij}) \text{ is a distance matrix s.t. } l_{ij} \leq d_{ij} \leq u_{ij}\}. \tag{9}$$

Algorithm 2 FLOYD(L, U): Update lower and upper bounds of the distance matrix by enforcing the triangle inequality.

Require: L, U

{Loop over all pairs of atoms N times.}

```
for  $k = 1$  to  $N$  do
  for  $i = 1$  to  $N - 1$  do
    for  $j = i + 1$  to  $N$  do
      if  $u_{ij} > u_{ik} + u_{kj}$  then
         $u_{ij} \leftarrow u_{ik} + u_{kj}$ 
      end if
      if  $l_{ij} < l_{ik} - u_{kj}$  then
         $l_{ij} \leftarrow l_{ik} - u_{kj}$ 
      end if
      if  $l_{ij} < l_{jk} - u_{ki}$  then
         $l_{ij} = l_{jk} - u_{ki}$ 
      end if
      if  $l_{ij} > u_{ij}$  then
        print "Bad Bounds"
      end if
    end for
  end for
end for
return  $L, U$ 
```

This is the space we wish to sample. Figure 2.3 shows how the three dimensional configuration space of a three atom cluster looks like.

Sampling a distance matrix within specific bounds is not a trivial task. A distance matrix corresponding to a cluster of N -atoms needs to satisfy N different types of inequalities: the triangle inequality, and corresponding inequalities involving four distances, five distances and so on.

For this reason we use a type of proposal-rejection sampling technique. We first sample a matrix satisfying the bounds and the triangle inequality and we project it to the nearest distance matrix in a least squares sense. If the resulting matrix satisfies the bounds we stop, otherwise we repeat the procedure. This is outlined in Algorithm 1. To sample the matrix satisfying the bounds and the triangle inequality we iterate over each (i, j) pair of atoms. We sample uniformly a distance d_{ij} between l_{ij} and u_{ij} . We set

Algorithm 3 MME($\tilde{\mathbf{D}}, \mathbf{U}$): It projects $\tilde{\mathbf{D}}$ to the closest distance matrix \mathbf{D} in a least square sense. For details see [citation needed].

Require: $\tilde{\mathbf{D}} = (\tilde{d}_{ij})$

Calculate the N -dimensional vector \mathbf{d}_{cm} .

for $i = 1$ to N **do**

$$\tilde{d}_{\text{cm},i} \leftarrow \frac{1}{N} \sum_{j=1}^N \tilde{d}_{ij}^2 - \frac{1}{N^2} \sum_{j=1}^N \sum_{k=j+1}^N \tilde{d}_{jk}^2$$

end for

If \mathbf{D} is a true distance matrix, \mathbf{d}_{cm} is the distance of each atom from the center of mass.

Calculate the $N \times N$ matrix $\mathbf{W} = (w_{ij})$.

for $i = 1$ to N **do**

for $j = 1$ to N **do**

$$\tilde{w}_{ij} \leftarrow \frac{1}{2}(\tilde{d}_{\text{cm},i} + \tilde{d}_{\text{cm},j} - \tilde{d}_{ij})$$

end for

end for

If \mathbf{D} is a true distance matrix, then \mathbf{W} would be the metric matrix of the cluster i.e. $\mathbf{W} = (\mathbf{r}_i \cdot \mathbf{r}_j)$.

Compute the three greatest eigenvalues of \mathbf{W} , $\lambda_1 > \lambda_2 > \lambda_3$ and the corresponding eigenvectors $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$.

Define a matrix \mathbf{X} of size $N \times 3$.

Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ (Column view).

for $i = 1$ to 3 **do**

$$\mathbf{x}_i \leftarrow \sqrt{\lambda_i} \mathbf{w}_i$$

end for

Now each row of $\mathbf{X} = (\mathbf{r}_1 \mathbf{r}_2 \dots \mathbf{r}_N)^T$ (row view) contains a cartesian representation of the closest metric matrix to \mathbf{W} .

Calculate the distance matrix \mathbf{D} associated with $\mathbf{r}_1, \dots, \mathbf{r}_N$.

return \mathbf{D}

the bounds corresponding to this distance equal to d_{ij} ($l_{ij} = u_{ij} = d_{ij}$) and finally refine all other upper and lower bounds so that the triangle inequality is not violated using the FLOYD algorithm (see Algorithm 2). Finally the resulting matrix is projected to the nearest distance matrix by employing the Metric Matrix Embedding algorithm (see Algorithm 3). Figure 2.3 shows the histograms of the interatomic distances obtained using USAMPLE. Notice that the sampling is not uniform, but all regions of the configuration space are covered.

2.4. Fitting the data

Suppose we have a total of S data points $(\mathbf{z}^{(i)}, E^{(i)})_{i=1}^S$. To avoid lengthy notation we will refer to these data collectively as

$$\mathcal{D} = (\mathbf{z}^{(i)}, E^{(i)})_{i=1}^S. \quad (10)$$

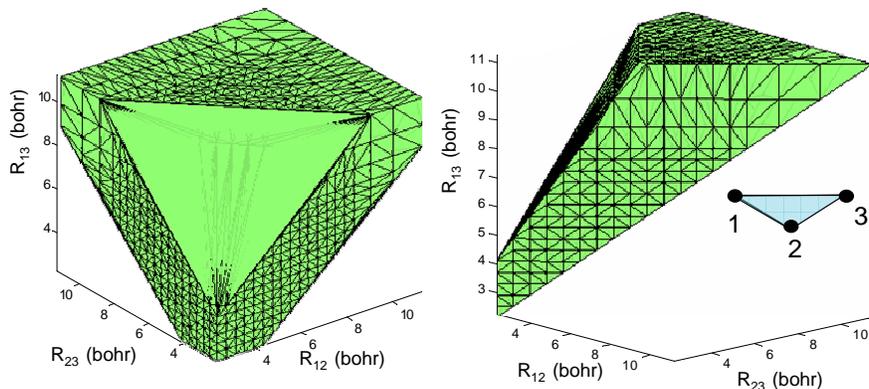


Figure 1: A view of the configuration space for a cluster of 3 atoms.

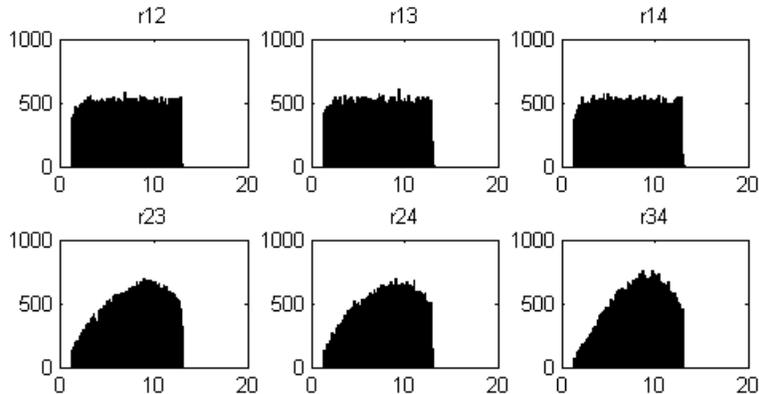


Figure 2: Histograms of the interatomic distances obtained using the USAMPLE algorithm. The minimum distance is set to 1.2 Å and the maximum distance to 13 Å. It is obvious that the sampling is not uniform but any point of the configuration space has the probability of occurring.

In this section we will outline how we fit \mathcal{D} to the model described in Section 2.2. The usual approach (see [6]) is to fix the scale parameter λ (usually to 2 – 3 Bohr) and minimize the least squares error. This procedure is straightforward to implement but has several drawbacks:

1. it can lead in severe overfitting if many basis functions are used.
2. it does not provide a quantification of the uncertainty when doing predictions.
3. it does not give a systematic way to select the scale parameter λ or test alternative coordinate transformations than the Morse variables (e.x. $y_{ij} = 1/d_{ij}$ as in [10]).

We propose the use a Bayesian Linear Regression scheme in order to cope with precisely these problems. We use the generalized linear model given in Eq. (6) with the additional assumption of an additive noise

$$E(\mathbf{z}) = \hat{E}(\mathbf{z}; \mathbf{w}) + \sigma Z, \quad (11)$$

where $Z \sim \mathcal{N}(0, 1)$ and σ^2 is the variance of the data. Of course, our ab initio calculations are deterministic and hence have no noise. However, in case the basis functions are not adequate to describe the energy surface the model will fit data with noise. Not being able to interpolate between data points of the surface, it will assume that their variability is due to a big σ . That is, the value of σ^2 will be an indicator of how good the fit really is. Under this assumption the likelihood of the real energy $E(\mathbf{z})$ is

$$p(E|\mathbf{z}, \mathbf{w}, \sigma^2) = \mathcal{N}(\hat{E}(\mathbf{z}; \mathbf{w}), \sigma^2). \quad (12)$$

Finally, we pose a Gaussian prior distribution over the weights

$$p(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \quad (13)$$

where $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ is the multivariate normal distribution with mean $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$. It turns out [35] that the posterior distribution of the weights given the data is also Gaussian

$$p(\mathbf{w}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (14)$$

with mean and inverse covariance matrix given by

$$\boldsymbol{\mu} = \boldsymbol{\Sigma} (\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \sigma^{-2} \boldsymbol{\Phi}^T \mathbf{E}), \quad (15)$$

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_0^{-1} + \sigma^{-2} \boldsymbol{\Phi}^T \boldsymbol{\Phi}, \quad (16)$$

where Φ is the so called $S \times M$ design matrix:

$$\Phi = \begin{pmatrix} \phi(\mathbf{z}^{(1)})^T \\ \vdots \\ \phi(\mathbf{z}^{(S)})^T \end{pmatrix},$$

with

$$\phi(\mathbf{z}) = (\phi_1(\mathbf{z}), \dots, \phi_M(\mathbf{z}))^T$$

and the observed energy vector

$$\mathbf{E} = (E^{(1)}, \dots, E^{(S)})^T.$$

Selection of the Prior. We propose starting with an isotropic Gaussian prior

$$\boldsymbol{\mu}_0 = \mathbf{0}, \tag{17}$$

$$\boldsymbol{\Sigma}_0 = \alpha \mathbf{I}, \tag{18}$$

where α is an additional unknown parameter which will be inferred from the data and \mathbf{I} is the $M \times M$ unit matrix. Notice that the prior becomes uninformative as $\alpha \rightarrow 0$ allowing for a very flexible choice of weights.

Predictive Distribution. The predictive distribution at a new point \mathbf{z} can be calculated by integrating out the weights using their posterior distribution. Not surprisingly this is a Gaussian also:

$$p(E|\mathbf{z}, \mathcal{D}, \alpha, \sigma^2) = \int p(E|\mathbf{z}, \mathbf{w}, \sigma^2) p(\mathbf{w}|\mathcal{D}, \alpha) d\mathbf{w} = \mathcal{N}(\boldsymbol{\mu}^T \phi(\mathbf{z}), \sigma^2(\mathbf{z})), \tag{19}$$

where the variance $\sigma^2(\mathbf{z})$ is given by

$$\sigma^2(\mathbf{z}) = \sigma^{-2} + \phi(\mathbf{z}) \boldsymbol{\Sigma} \phi(\mathbf{z}). \tag{20}$$

A point estimate of $E(\mathbf{z})$ can be given by the mean $\boldsymbol{\mu}^T \phi(\mathbf{z})$. Notice that we automatically get a quantification of the uncertainty of our prediction through $\sigma^2(\mathbf{z})$. This will be used in Section 2.7 in the selection of new data points.

2.5. Bayesian Model Selection

In the previous section we have shown how the model fits the data for a fixed choice of α and σ^2 . The careful reader would have noticed that the prediction is also implicitly dependent on the choice of the scale parameter λ , i.e.

$$\mathbf{z} = \mathbf{z}_\lambda(\mathbf{D}),$$

where \mathbf{D} is the distance matrix. In a fully Bayesian scheme we would have to define prior distributions on all those three parameters and then integrate over them to get the fully Bayesian predictive distribution. This is not an easy task to perform. In this section we will motivate the so called *evidence approximation* [36] which, under special assumptions, can provide an alternative way to solve the problem.

Suppose we have introduced some prior distribution $p(\alpha, \sigma^2, \lambda)$ on $(\alpha, \sigma^2, \lambda)$. By Bayes Theorem the posterior distribution is

$$p(\alpha, \beta, \lambda | \mathcal{D}) = p(\mathcal{D} | \alpha, \sigma^2, \lambda) p(\alpha, \sigma^2, \lambda), \quad (21)$$

and the fully Bayesian predictive distribution is

$$p(E | \mathbf{D}, \mathcal{D}) = \int \int \int \int p(E | \mathbf{z}_\lambda(\mathbf{D}), \mathbf{w}, \sigma^2) p(\mathbf{w} | \mathcal{D}, \alpha) p(\alpha, \beta, \lambda | \mathcal{D}) d\mathbf{w} d\alpha d\sigma^2 d\lambda.$$

If the posterior is sharply peaked around $(\hat{\alpha}, \hat{\sigma}^2, \hat{\lambda})$ it can be treated as a δ function and hence the above integral may be approximated by

$$p(E | \mathbf{D}, \mathcal{D}) \approx \int p(E | \mathbf{z}_{\hat{\lambda}}(\mathbf{D}), \mathbf{w}, \hat{\sigma}^2) p(\mathbf{w} | \mathcal{D}, \hat{\alpha}) d\mathbf{w}.$$

Intuitively, this would be the case if the basis functions we are using can describe the true energy surface and if we have sufficient data at our disposal. In this case predictions can be made using Eq. (19) at $(\hat{\alpha}, \hat{\sigma}^2, \hat{\lambda})$.

Now the problem becomes to maximize the posterior of the hyperparameters $(\alpha, \sigma^2, \lambda)$ given by Eq. (21). This is not a trivial task either. However, if the prior $p(\alpha, \sigma^2, \lambda)$ is relatively flat (which would be the case if we have no specific preference about these parameters), then this maximum will be effectively the maximum of the likelihood function (see Eq. (21) again):

$$\mathcal{L}(\alpha, \sigma^2, \lambda) = p(\mathcal{D} | \alpha, \sigma^2, \lambda).$$

This function is called the *marginal likelihood* or the *evidence function*. Our model selection problem becomes:

$$(\hat{\alpha}, \hat{\sigma}^2, \hat{\lambda}) = \arg \max_{(\alpha, \sigma^2, \lambda)} \mathcal{L}(\alpha, \sigma^2, \lambda). \quad (22)$$

In principle it can be solved by conjugate gradient methods. Unfortunately this would require the derivative of the basis functions with respect to λ which is a very involved calculation. In our numerical experiments, we exploit the fact that for fixed λ there exist a robust algorithm (see Chapter 3.5 of [35]) to solve

$$(\hat{\alpha}, \hat{\sigma}^2) = \arg \max_{(\alpha, \sigma^2)} \mathcal{L}(\alpha, \sigma^2, \lambda),$$

in order to pose the problem as

$$\hat{\lambda} = \arg \max_{\lambda} \left(\max_{(\alpha, \sigma^2)} \mathcal{L}(\alpha, \sigma^2, \lambda) \right). \quad (23)$$

The optimization over λ can be carried out using Brent's method (see [37]).

2.6. Error Evaluation

A concrete way to account for the error of the approximation is to leave out of the fitting procedure some samples and compare the prediction of their energy with the true value. The measure we propose to use is the mean square error of the energy per atom

$$\text{MSE}(\mathcal{D}_{\text{test}}) = \frac{\sqrt{\sum_{i=1}^{S_{\text{test}}} \left(E_{\text{test}}^{(i)} - \hat{E}(\mathbf{z}_{\text{test}}^{(i)}) \right)^2}}{NS_{\text{test}}}, \quad (24)$$

where we have left out of the fitting procedure S_{test} data points:

$$\mathcal{D}_{\text{test}} = (E_{\text{test}}^{(i)}, \mathbf{z}_{\text{test}}^{(i)})_{i=1}^{S_{\text{test}}}.$$

An alternative measure we will also use is the maximum absolute error of the energy per atom

$$\text{MABSE}(\mathcal{D}_{\text{test}}) = \max_{i \leq 1 \leq S_{\text{test}}} \left\{ \left| E_{\text{test}}^{(i)} - \hat{E}(\mathbf{z}_{\text{test}}^{(i)}) \right| \right\}.$$

Both of them are objective measures of the predictive ability of the fit we have achieved. Ideally, one would like to see $\text{MABSE}(\mathcal{D}_{\text{test}})$ becoming approximately the same as $\text{MSE}(\mathcal{D}_{\text{test}})$. This would mean that there are no

outlying energies and that our fitting is uniform. In Section 2.7 we will use $\text{MSE}(\mathcal{D}_{\text{test}})$ to define a stopping criterion of our scheme. It will provide us with a definite way to judge if we need more data points or we have reached the maximum predictive ability of the chosen basis.

2.7. Adaptive selection of data points

Ab initio data is expensive so we cannot afford to waste any. Inspired by the seminal paper of Sacks [38], we propose a simple experimental design

Algorithm 4 BFED: Fit the PES of a given cluster.

Require: 1) Cluster type X_nY_m , 2) bounds of the configuration space (\mathbf{L}, \mathbf{U}) , 3) maximum polynomial degree d_{max} 4) initial number of data points S_{init} , 5) number of test samples S_{test} , 6) number of MC samples whose variance is tested in every cycle S_{MC} , 7) number of samples that are added to the fitting procedure after its cycle S_{add} .

Generate using $\text{USAMPLE}(\mathbf{L}, \mathbf{U})$ S_{test} samples and calculate their energies. We denote them with

$$\mathcal{D}_{\text{test}} = (E_{\text{test}}^{(i)}, \mathbf{z}_{\text{test}}^{(i)})_{i=1}^{S_{\text{test}}}$$

Generate using $\text{USAMPLE}(\mathbf{L}, \mathbf{U})$ S_{init} samples and calculate their energies. We denote them with

$$\mathcal{D} = (E^{(i)}, \mathbf{z}^{(i)})_{i=1}^{S_{\text{init}}}$$

loop

Fit \mathcal{D} as described in Section 2.4 and Section 2.5.

if $\text{MSE } \mathcal{D}_{\text{test}}$ doesn't change significantly from its previous value **then**
return

else

Generate S_{MC} configuration point samples using $\text{USAMPLE}(\mathbf{L}, \mathbf{U})$.

Calculate their variance using Eq. (20).

Find the S_{add} of the S_{MC} samples with the maximum variance.

Calculate their energy and add them to \mathcal{D} .

end if

end loop

scheme. The goal is to extract information from existing data that would help us select new points of the configuration points that will improve our fit by keeping the ab initio calculations to a minimum.

We have already mentioned that the Bayesian scheme described in Section 2.4, provides a natural way to quantify our uncertainty at any point of the configuration space \mathbf{z} through the predictive variance $\sigma^2(\mathbf{z})$ defined in Eq. (20). This variance represents our *lack of knowledge* about the PES and not the real error. It corresponds to the real error (qualitatively) only to the extent that the selected basis can actually describe the true energy surface. We postulate that the inclusion of new configuration points with high Bayesian variance will improve the fitting achieved as measured by the $\text{MSE}(\mathcal{D}_{\text{test}})$ more than the merely adding random points. This is valid to the extent that the basis functions we use can actually describe the PES under consideration correctly. Our numerical experiments have shown that this is the case for the permutationally invariant polynomial basis alongside with the Morse variables, if a sufficient number of basis functions are used. We have found that the number of require energy calculations is reduced significantly.

Associating the value of $\sigma(\mathbf{z})$ with the informational content of the point \mathbf{z} , a natural strategy is to add to the scheme the \mathbf{z} 's that maximize it. We use a plain Monte Carlo procedure to identify the important points of the configuration space. The proposed algorithm, called Brute Force Experimental Design (BFED), is extremely simple to implement since all it requires is the ability to sample the configuration space and calculate the Bayesian variance. One starts with some random configuration points, calculates the corresponding energy and fits them using the scheme described in Section 2.4 and Section 2.5. The MSE is evaluated on some random energies left out of the fitting procedure as described in Section 2.6. Then we generate many configuration points within the specified bounds using the USAMPLE algorithm and calculate the predictive variance of each one of them using the fitted surface. This is computationally negligible compared to the ab initio calculations. Finally we select the ones with the maximum Bayesian variance, calculate their energies, add them to the scheme and refit ². If the MSE

²The refitting procedure doesn't have to start from scratch. One can initialize the α, β and λ parameters at the previously obtained optimum values resulting in improve convergence the optimization schemes. However, the cost of the fitting procedure is so small compared to the ab initio calculations that making it faster would have no observable

of the new fit stops changing, we stop otherwise we repeat the procedure. This is summarized in Algorithm 4.

2.8. Multi-Body Expansion

One can easily imagine that the energy of a cluster of N atoms with Cartesian coordinates $\mathbf{R} = (\mathbf{r}_1, \dots, \mathbf{r}_N) \in \mathbf{R}^{3N}$ can be written as an infinite series

$$E(\mathbf{r}_1, \dots, \mathbf{r}_N) = \sum_{i=1}^N E(\mathbf{r}_i) + \sum_{i<j} V^{(2)}(\mathbf{r}_i, \mathbf{r}_j) + \sum_{i<j<k} V^{(3)}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) + \dots \quad (25)$$

The $V^{(K)}$'s are functions of K atoms and are called multibody potentials. One can think of $\sum_{i_1 < \dots < i_K} V^{(K)}(\mathbf{r}^{i_1}, \dots, \mathbf{r}^{i_K})$ as the energy added to the system due to interactions between K atoms. It seems reasonable that after a certain index P , interactions between $P+1$ atoms are unimportant, so that

$$V^{(P+1)} \approx 0. \quad (26)$$

If this is the case, then we call this expansion a P -order expansion. We denote the total energy of an N -atom system using a P -order expansion by $E_P = E_P(\mathbf{r}_1, \dots, \mathbf{r}_N)$. We write

$$E_P(\mathbf{r}_1, \dots, \mathbf{r}_N) = \sum_{K=1}^P E^{(K)}(\mathbf{r}_1, \dots, \mathbf{r}_N), \quad (27)$$

where

$$E^{(K)}(\mathbf{r}_1, \dots, \mathbf{r}_N) := \sum_{i_1 < \dots < i_K} V^{(K)}(\mathbf{r}_{i_1}, \dots, \mathbf{r}_{i_k}). \quad (28)$$

Finally, the $V^{(K)}$'s can be readily found by using the Möbius inversion approach from number theory [27, 39]. Möbius inversion has been used previously for the extraction of potentials from energy data in [40, 41]. We have

$$V^{(K)}(\mathbf{r}_1, \dots, \mathbf{r}_N) = \sum_{L=1}^K (-1)^{K-L} \sum_{i_1 < \dots < i_K} E(\mathbf{r}_{i_1}, \dots, \mathbf{r}_{i_K}), \quad (29)$$

where with E we denote the true energy function. Once the multibody potentials $V^{(K)}$ have been constructed, the P -order energy E_P of an N -atom system can be calculated by using Eq. (27).

benefit.

3. Numerical Examples

3.1. Pt clusters using EAM energy and GULP

In the first set of examples we wish to test the convergence properties of the suggested scheme. Due to the computational burden of ab initio calculations, exhaustive tests can only be performed using an empirical potential. In what follows we consider solely clusters of Pt with Embedded-Atom Method (EAM) [42] potential energy as implemented in GULP [43]. The minimum distance between Pt atoms is taken to be 2 Å and the maximum distance 13 Å. All energies are in eV.

First, we investigate the properties of the adaptive addition scheme of Section 2.7 and demonstrate that it greatly improves the quality of the sampled energies. Then, we test the convergence of fitting scheme with respect to the polynomial degree of the basis and finally we compute the EAM Pt7 PES to assess the applicability of our work to relatively big clusters.

Adaptive selection of data points. Our first goal is to demonstrate that the adaptive addition of data points (described in Section 2.7) has a positive effect on the accuracy of the fit. To test this claim, we choose a particular Pt cluster and:

1. Generate a certain number S_{test} of random test points and calculate their energies. These are left out of the fitting procedure and are used only for the evaluation of the error.
2. We generate S_{init} initial random points and calculate their energies.
3. Using the same S_{init} initial random points, we run the BFED Algorithm for different adaptive strategies (varying the number S_{MC} of samples among which we select the points that should be added to the dataset).
4. For each strategy, we trace how the two error estimates MSE and AMSE (described in Section 2.6) vary as a function of the number of data points used to fit the PES.

We perform this test on Pt3 and Pt4 clusters. In both cases we add $S_{\text{add}} = 100$ data points at each step of the BFED Algorithm, set the basis to polynomials of up to degree 10 and we test the same six different strategies of adding new data points:

1. No Selection. S_{add} totally random points are added at each cycle, i.e. $S_{\text{MC}} = S_{\text{add}} = 100$.
2. Choose 100 among 200, i.e. $S_{\text{MC}} = 200$.

3. Choose 100 among 500, i.e $S_{MC} = 500$.
4. Choose 100 among 1000, i.e $S_{MC} = 1000$.
5. Choose 100 among 10000, i.e $S_{MC} = 10000$.
6. Choose 100 among 20000, i.e $S_{MC} = 20000$.

For Pt3, the number of test points is set to $S_{test} = 1000$ and the number of initial random points to $S_{init} = 200$. For Pt4, the number of test points is set to $S_{test} = 3000$ and the number of initial random points to $S_{init} = 500$. In Figure 3 and Figure 4 we plot the evolution MSE and AMSE respectively for Pt3. Figure 5 and Figure 6 depicts the same quantities for Pt4.

These figures show clearly that the naive random selection technique is inferior to the suggested scheme. One can notice two underlying properties of the choice of S_{MC} :

1. The error drops faster as a function of the number of data points, when S_{MC} is initially increased but
2. there is a natural limit to this effect. In both tests, setting S_{MS} to a value greater than 1000 does not refine the errors any more.

The former property provides good evidence that the scheme indeed adds informationally rich data points. The latter, puts a barrier on this improved performance.

Finally, in Figure 7 we plot the histograms of 1000 sampled Pt4 energies that resulted from two of the strategies discussed above: the random addition strategy $S_{MC} = 100$ shown in Figure 7(a) and the ‘‘Choose 100 among 10000’’ strategy $S_{MC} = 10000$ shown in Figure 7(b). Notice that the adaptive strategy yields a considerably broader distribution of energies, covering low energy (stable) configuration points as well as high energy (unstable) ones. This effect, i.e. the more uniform sampling of the configuration space, is the reason why the proposed adaptive scheme actually has this effect to the observed errors. As the S_{MC} initially increases, provides better sampling of the configuration space as seen by the distribution of the energies. However, after a critical value of S_{MC} the ‘‘optimum’’ energy distribution has been reached and increasing it further has no effect.

Convergence with respect to polynomial degree. An important question that we want to pose is to what extent does the choice of the polynomial degree effects the accuracy of the final PES. There are basically two reasons why one should care about this:

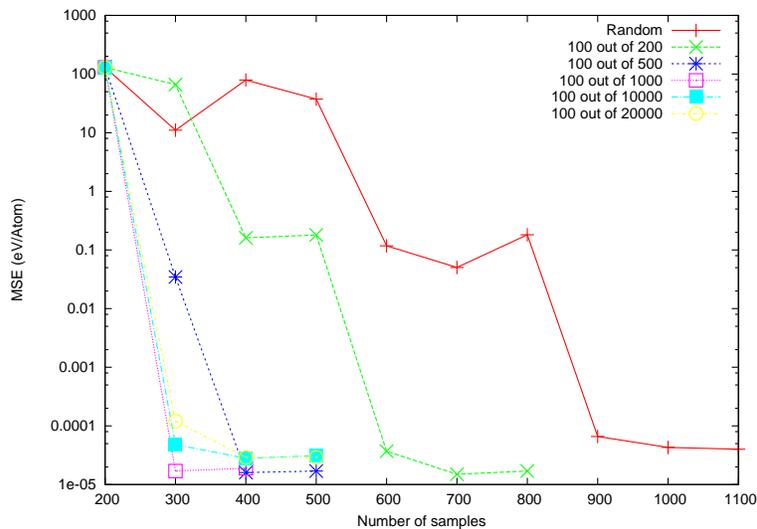


Figure 3: Pt3: evolution of $MSE(\mathcal{D}_{\text{test}})$ as the number of samples increases.

1. If the polynomial degree is low, the basis has low expressivity and it might not be able to capture the real PES and
2. increasing the polynomial degree arbitrarily might result in overfitting which would yield a PES with limited predictive capabilities.

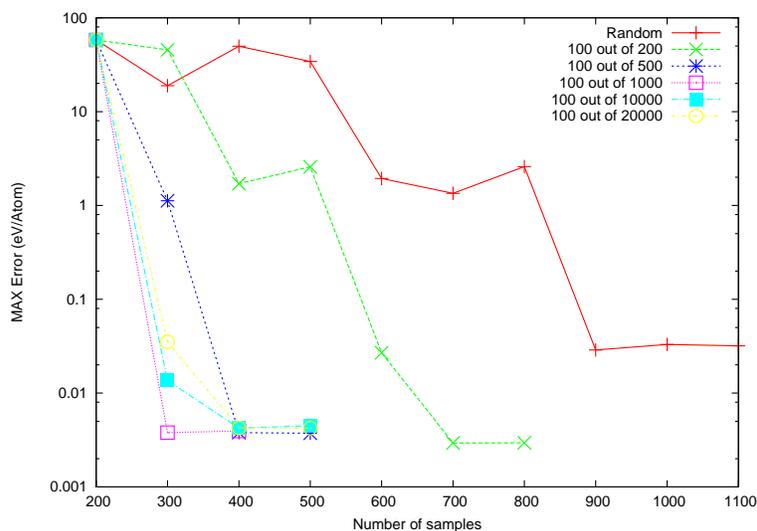


Figure 4: Pt3: evolution of $MABSE(\mathcal{D}_{\text{test}})$ as the number of samples increases.

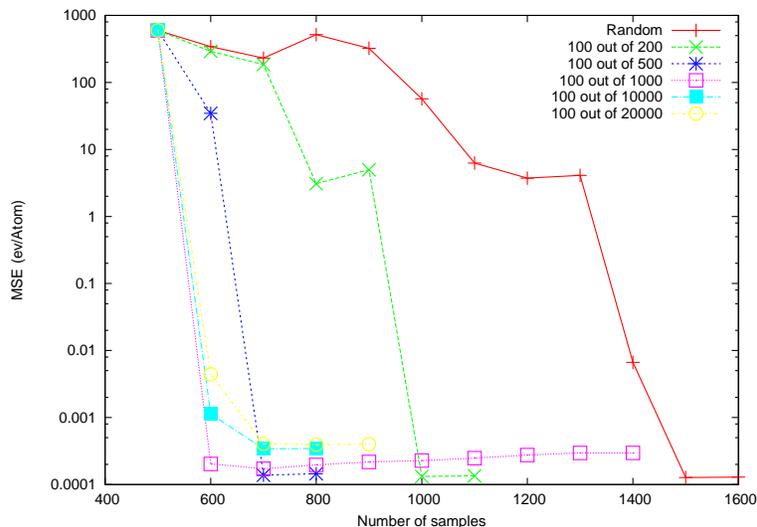


Figure 5: Pt4: evolution of $MSE(\mathcal{D}_{\text{test}})$ as the number of samples increases.

To resolve the first problem, the natural strategy is to use higher order polynomials. In the case study that follows it is clearly demonstrated that the second problem (overfitting) is missing from our scheme. This is due to its Bayesian nature. We consider a Pt4 cluster and fit its PES using polynomials

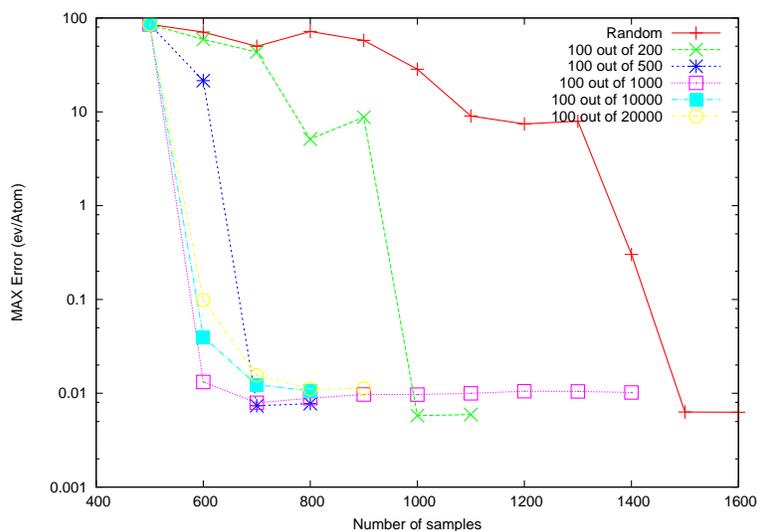


Figure 6: Pt4: evolution of $MABSE(\mathcal{D}_{\text{test}})$ as the number of samples increases.

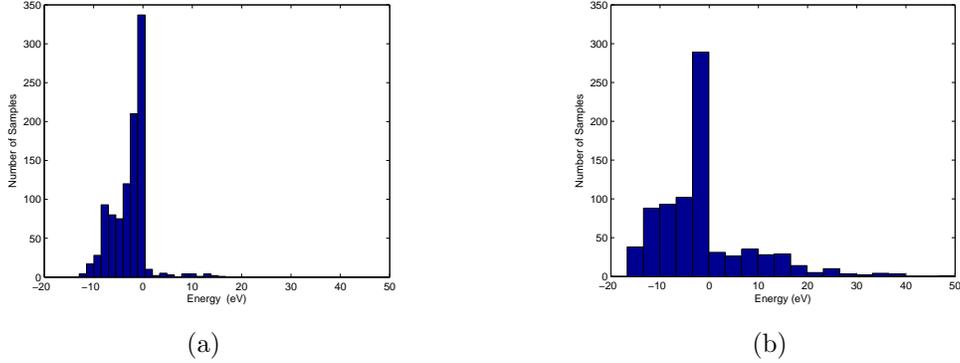


Figure 7: Pt4: The histograms shows 1000 data points used in the fitting procedure. 7(a) shows data points selected completely at random while 7(b) shows the end result of the proposed scheme when selecting 100 out of 10000 data points.

of 4th, 6th, 7th, 8th and 10th degree. For each degree we use exactly the same $S_{\text{test}} = 100$ and $S_{\text{init}} = 1000$ test and initial data points, respectively. The adaptive addition strategy is “choose 100 among 10000” ($S_{\text{add}} = 100$ and $S_{\text{MC}} = 10000$). To give a pictorial representation of the convergence, we fix atoms 2, 3 and 4 to

$$\begin{aligned}\mathbf{r}_2 &= (1.1, 1.1, 0), \\ \mathbf{r}_3 &= (-1.1, 1.1, 0), \\ \mathbf{r}_4 &= (1.1, -1.1, 0),\end{aligned}$$

and we plot the energy as a function of the position of the 1st atom allowing it to move on the $z = 0$ plane, i.e.

$$\mathbf{r}_1 = (x, y, 2.1).$$

We choose this particular set up because for $\mathbf{r}_1 = (0, 0, 2.1)$, we obtain close to the tetrahedral stable configuration of Pt4, thus the test is performed in a region of the configuration space of a high variability in energy. Figure 8 depicts the corresponding cut of the PES for each polynomial degree. The EAM energy is also shown for reference. One can clearly notice that after degree 7 the picture stabilizes. Figure 9 shows the absolute error the 6, 7, 8 and 10th degree PES. This is evaluated by comparing the predicted energy at each point with the true EAM value.

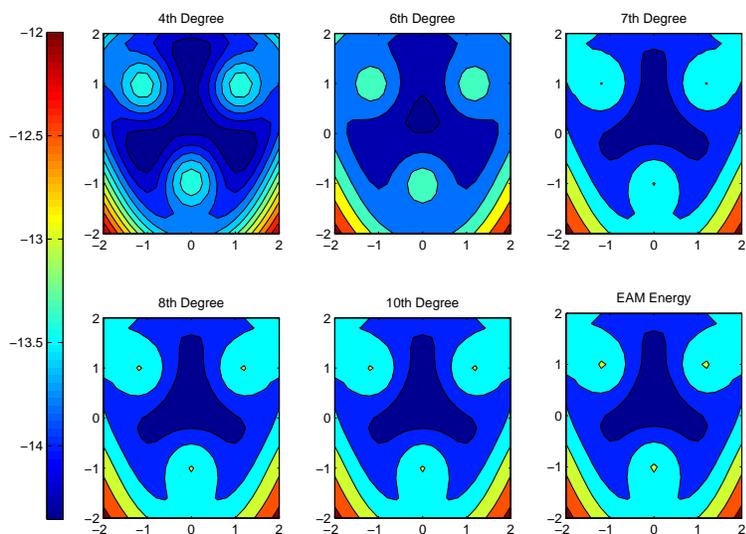


Figure 8: Showing convergence of the fitted PES for a Pt4 cluster with respect to the polynomial degree used.

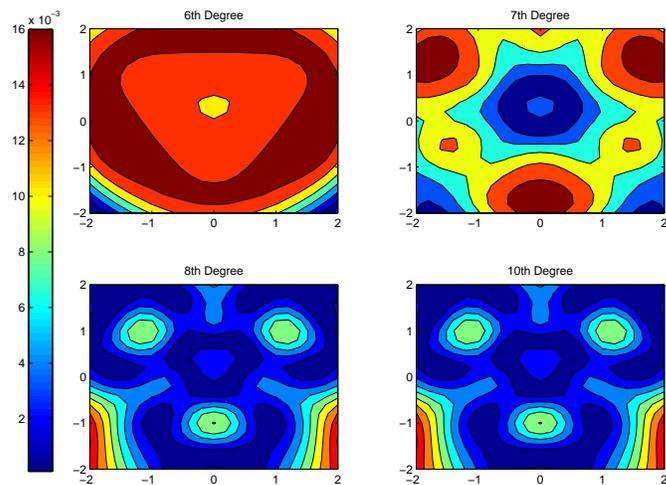


Figure 9: Showing the absolute error of the fitted PES for Pt4 for different polynomial degrees.

Demonstrating the scheme’s accuracy for big clusters. As yet another case study, we test the accuracy of the scheme used for relatively big clusters. We fit the EAM PES for Pt7 using 5th degree polynomials, $S_{\text{test}} = 100$, $S_{\text{add}} = 100$ and $S_{\text{MC}} = 1000$. We add data points until subsequent values MSE do not differ more than 10^{-3} . The total number of “electronic” calculations is 2000. In Figure 10, we compare the fitted energy to the true EAM energy as a function of the x-y position of the first atom

$$\mathbf{r}_1 = (x, y, 0),$$

while keeping the rest fixed to

$$\begin{aligned} \mathbf{r}_2 &= (2, 2, 2.1), \\ \mathbf{r}_3 &= (2, -2, 2.1), \\ \mathbf{r}_4 &= (-2, 2, 2.1), \\ \mathbf{r}_5 &= (-2, -2, 2.1), \\ \mathbf{r}_6 &= (0, 0, 2.1), \\ \mathbf{r}_7 &= (0, 0, -2.1). \end{aligned}$$

Again, this is a region of high variability in energy. We observe that the PES remains fairly accurate even for this relatively low polynomial degree.

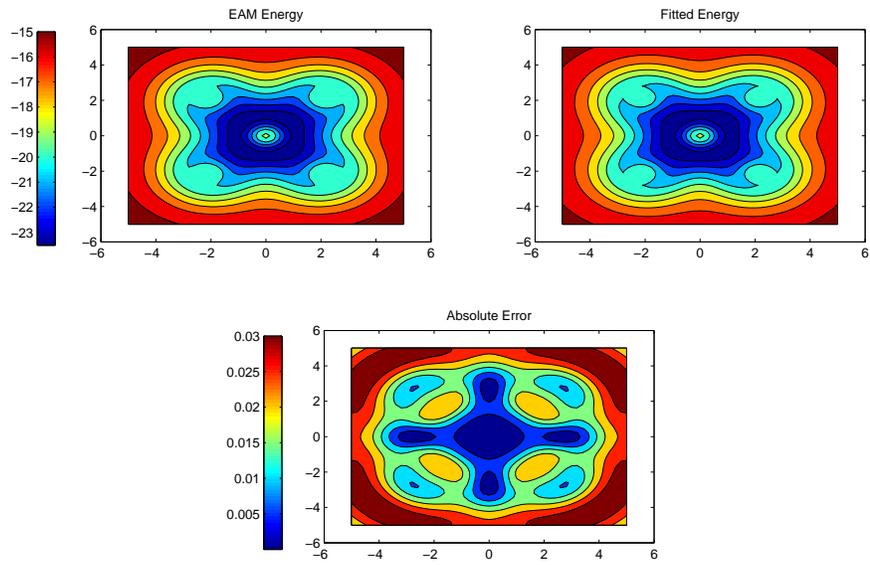


Figure 10: Pt7: Plots of the true EAM energy, the prediction of our scheme and the absolute error in energy.

3.2. Pt clusters using DFT calculations

In what follows, ab initio calculations were performed using density functional theory (DFT) in the local density approximation (LDA) as implemented in the PWSF software package [44], using the Perdew-Zunger parametrization of the exchange correlation energy and Rabe-Rappe-Kaxiras-Joannopoulos [45] (ultrasoft) pseudopotential. The cell size was taken to be sufficiently large (4 times the maximum interatomic distance, a.k.a. 0.05 to 0.2 nano-meters) to effectively simulate an isolated cluster. The energy cut-off was 22 Ry (≈ 300 eV). One k -point was used for the plane wave basis. The above parameters were chosen so that the accuracy of the DFT calculations was within 0.01eV/atom. All distances are in Bohr and energies in Ryd unless otherwise specified.

We computed the DFT PES for Pt2, Pt3, Pt4, Pt5 and Pt6 clusters. The data collection procedure was parallelized using MPI [46] so that each CPU core computed a single data point. In all computations we used $S_{MC} = 10000$ and $S_{add} = 256$. The minimum cut-off distance was set to 3.7 Bohr for all clusters. The maximum cut-off was set to 14 Bohr for Pt2 and 10 Bohr for all others cases. We stopped the BFED Algorithm either when MSE per atom was less than $1e - 03$ Ryd, or when its change between two consecutive iterations was less than $1e - 05$ Ryd. Table 1 shows the detailed parameters of each run along with the computational resources that were required. The vast majority of the computational time was spent in ab initio calculations. For example, the Pt5 PES required approximately 39 hours using 256 CPU cores from which only 6 minutes was spent to actually fit the data. The computational effort put in our fitting scheme can safely be neglected in comparison to the cost of the DFT calculations. In the Pt6 calculation we were forced to use a smaller polynomial basis because of the computational burden involved. This resulted in a considerably less accurate PES. In several cases - for example when the random clusters were in highly unstable configurations - the DFT self-consistent calculations failed to converge within 100 iterations. The configurations for which this happened were removed from the data set.

After assessing the accuracy of the fitted surfaces, we use them to compute stable structures of small Pt clusters and compare the results with the literature. We continue by investigating the decomposition of the PES in K -body potentials V_K . Finally, we put these potentials together to explore the predictive capabilities of raw MBE.

Accuracy of the fitting procedure. Table 2 shows the errors of each of the fitted PES as measured using the test points left out of the scheme. We show both the Mean Square Error as well the Maximum Absolute Error (Section 2.6). Notice that the Maximum Absolute Error is of the order of $1e-03$ Ryd (≈ 0.01 eV) per atom which is exactly the accuracy of our DFT calculations. Figure 11 gives a pictorial representation of the goodness of fit. For a given cluster we plot the true DFT energy versus the energy prediction for each test data point. The horizontal axis of the plots corresponds to the DFT calculated energy and the vertical axis to the prediction of the energy using the fitted PES. The straight line is the $y = x$ line. The deviation of each red point from this line represents the error of the corresponding test point. It is apparent that the predictions become more noisy as the cluster size increases. However, the important feature that one should notice is that the error is *uniformly bounded* over the energy range.

Using the fitted PES to predict stable structures. As an elementary application, we use the fitted PES to predict stable Pt structures of up to 6 atoms. The optimization is performed using the Simulated Annealing technique [47] as implemented in the GSL library [48]. The quantities we report are the average bond length, the symmetry of the structure and the binding energy. The binding energy E_B is defined by

$$E_B = E_{\text{atom}} - E_{\text{cluster}}/N, \quad (30)$$

where $E_{\text{atom}} = E_1 = -52.157\text{Ryd}$ is the one atom energy (see next paragraph for details on how we get calculate this), E_{cluster} is the energy of the stable cluster and N the number of atoms. Table 3 summarizes our results. For easy comparison with the results in the literature lengths are reported in Å and energies in eV. Overall, we obtain the same stable structures as the ones found

Cluster	Degree	No. Basis	S_{init}	S_{test}	S_{final}	CPU Cores	Time
Pt2	10	11	512	128	512	128	03:05:16
Pt3	10	67	1024	256	1024	128	09:31:53
Pt4	8	195	1024	256	1279	256	17:11:06
Pt5	8	580	1024	256	3071	256	39:30:57
Pt6	5	86	1024	256	1280	256	17:12:31

Table 1: Computational details of the fitting procedure of Pt clusters.

Cluster	MSE($\mathcal{D}_{\text{test}}$)	MSE($\mathcal{D}_{\text{test}}$)/Atom	MABSE($\mathcal{D}_{\text{test}}$)	MABSE($\mathcal{D}_{\text{test}}$)/Atom
Pt2	2.2e-03	1.1e-03	4.9e-04	2.4e-04
Pt3	1.2e-02	3.9e-03	5.1e-03	1.7e-03
Pt4	2.9e-02	7.3e-03	8.0e-03	2.0e-03
Pt5	4.9e-02	9.7e-03	1.6e-02	3.1e-03
Pt6	1.4e-01	2.3e-02	2.9e-02	4.8e-03

Table 2: The definition of the errors

Notation	Pt3	Pt4	Pt5	Pt6-1	Pt6-2
Structure					
Bond Len. (Å)	2.43	2.54	2.50-2.59	2.50-2.7	2.43-2.57
Bind En. (eV/atom)	3.25	3.63	3.94	4.19	4.23
Symmetry	$C_{2\nu}$	C_s	D_{3h}	$C_{2\nu}$	$C_{2\nu}$

Table 3: Stable structures of Pt clusters predicted using the fitted PES.

in [49]. The bond lengths predicted are slightly smaller than the ones found in the literature while the binding energies greater. These differences were expected, since in [49] they employ the generalized gradient approximation (GGA) instead of the LDA we use in the present work. However, the behavior of both the binding energy as well as the average bond length as a function of the cluster size is consistent with the [49] results. A comparison is shown in Figure 12.

Decomposition of the PES in potentials. The values of the potentials V_K can be approximated through the Möbius transformation Eq. (29) by using the fitted PES in place of the real one. We use the values thus obtained to fit the potentials using exactly the same regression scheme as the one used to fit the PES. The polynomial basis of the K -order potential V_K is chosen to be the same as the one used to fit to PtK PES. The one-atom energy $E_1(\mathbf{r})$ is equal to half the limiting value of the fitted energy $E_2(\mathbf{r})$ of two atoms as

the interatomic distance goes to infinity, i.e.

$$E_1 = \frac{1}{2} \lim_{r \rightarrow \infty} E_2(r) \approx -52.157 \text{ Ryd.}$$

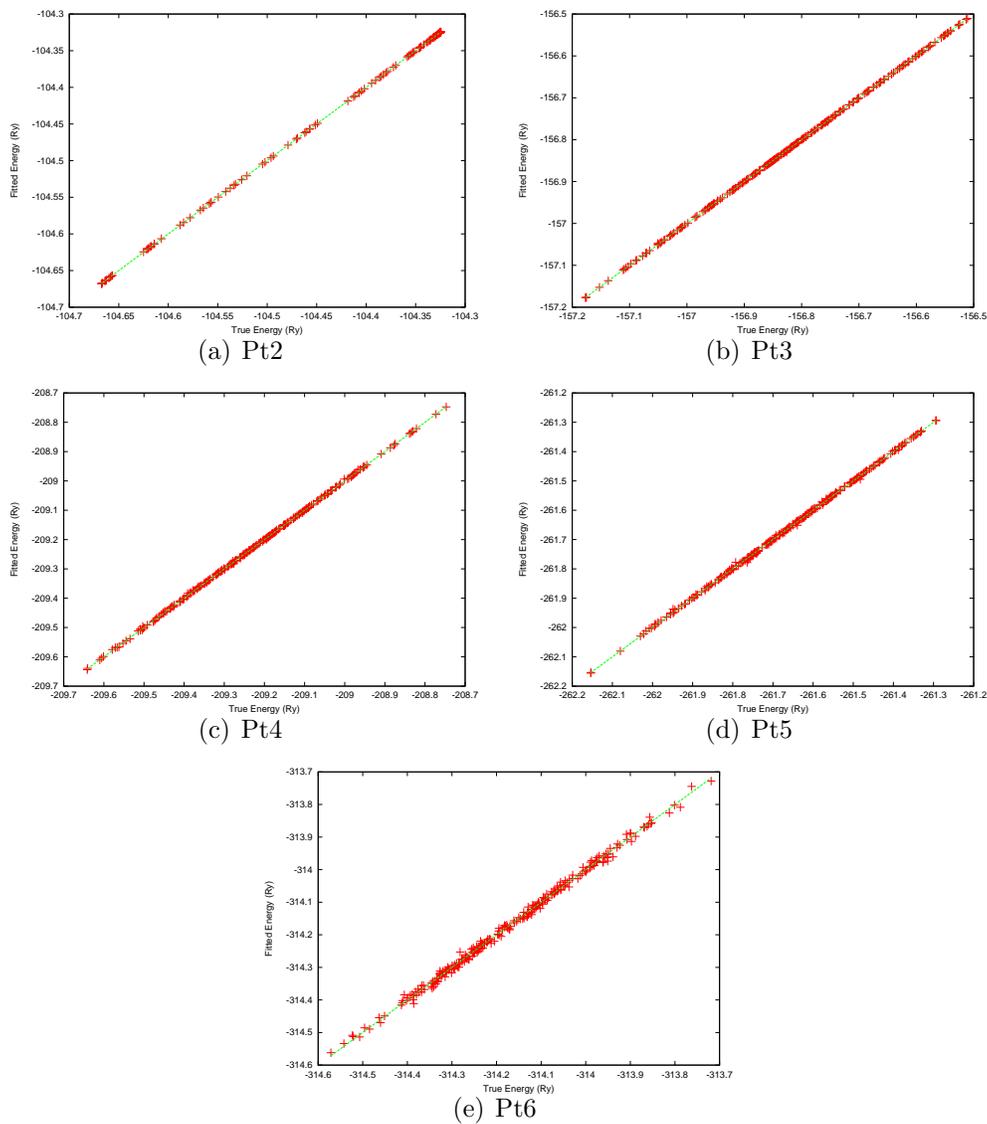


Figure 11: Testing the goodness of fit using full DFT calculations of random clusters. For Pt2 we use 128 energies and 256 for Pt3, Pt4, Pt5 and Pt6.

This is half the amount of energy required to disassociate the two atom system. Figure 13 shows the original $E_2(r)$ and the derived potential $V_2(r)$. Visualizing the higher order potentials is best achieved by showing a two-dimensional piece of the configuration space. In Figure 14 we consider a Pt3 cluster and plot E_3 , V_2 and V_3 as a function of the x-y position of the first atom

$$\mathbf{r}_1 = (x, y, 3.7),$$

while keeping the rest fixed to

$$\mathbf{r}_2 = (-2.2, 0, 0),$$

$$\mathbf{r}_3 = (2.2, 0, 0).$$

Figure 15 shows E_4 , V_2, V_3 and V_4 as a function of the $x - y$ position of the first atom

$$\mathbf{r}_1 = (x, y, 3.7),$$

while keeping the rest fixed to

$$\mathbf{r}_2 = (-3.7, 0, 0),$$

$$\mathbf{r}_3 = (3.7, 0, 0),$$

$$\mathbf{r}_4 = (0, 3.2, 0).$$

Figure 16 shows E_5 , V_2, V_3 , V_4 and V_5 as a function of the $x - y$ position of the first atom

$$\mathbf{r}_1 = (x, y, 3.7),$$

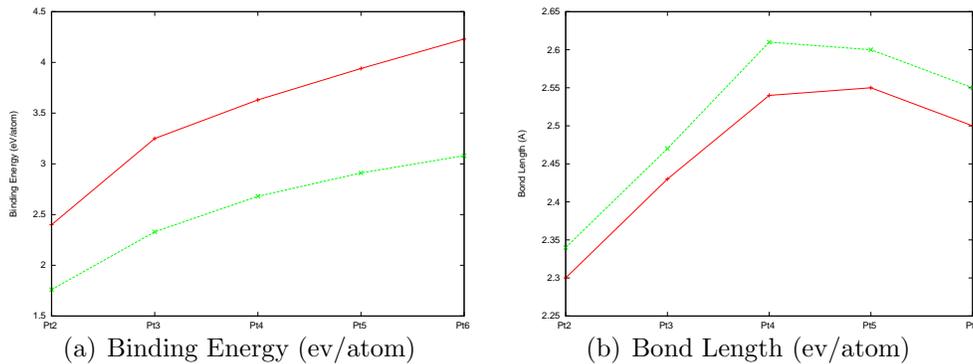


Figure 12: Comparison of binding energy and average bond lengths obtained using our fitted PES for Pt2, Pt3, Pt4, Pt5 and Pt6 with the results found in the literature [49].

while keeping the rest fixed to

$$\begin{aligned}\mathbf{r}_2 &= (-3.7, 0, 0), \\ \mathbf{r}_3 &= (3.7, 0, 0), \\ \mathbf{r}_4 &= (0, 3.2, 0), \\ \mathbf{r}_5 &= (0, -3.2, 0).\end{aligned}$$

Finally, Figure 16 shows E_5 , V_2 , V_3 , V_4 , V_5 and V_6 as a function of the $x - y$ position of the first atom

$$\mathbf{r}_1 = (x, y, 3.7),$$

while keeping the rest fixed to

$$\begin{aligned}\mathbf{r}_2 &= (0, 0, 0), \\ \mathbf{r}_3 &= (-6.4, 0, 0), \\ \mathbf{r}_4 &= (6.4, 0, 0), \\ \mathbf{r}_5 &= (3.2, 6.4, 0), \\ \mathbf{r}_6 &= (-3.2, 6.4, 0).\end{aligned}$$

The key observations one can make out of these figures are:

1. The sign of the interatomic potential is systematically reversed from one order to the other, i.e. V_2 and V_4 are negative, while V_3 and V_5 are positive.
2. V_K is indeed becoming less and less important as K increases.
3. The importance of each V_K becomes greater in regions of the configuration space of low energy.

MBE approximation to ab initio energies. In what follows, we denote with MBE- K the Multi-Body Expansion of order K given by Eq. (27) and Eq. (28) of Section 2.8. We wish to investigate to what extent MBE- K can accurately predict electronic energies of Pt clusters with number of atoms $N > K$. Figure 17 shows the predictions of MBE-2, MBE-3 and MBE-4 on 256 Pt4 test points. Figure 18 shows how the surface of Figure 15 can be approximated by MBE of order 2, 3 and 4. Figure 19 shows the predictions of MBE-3, MBE-4 and MBE-5 on 256 Pt5 test points. Figure 20 shows how the surface of Figure 16 can be approximated by MBE of order 3, 4 and 5. Figure 21 shows the predictions of MBE-4, MBE-5 and MBE-6 on

256 Pt5 test points. Figure 22 shows how the surface of Figure ?? can be approximated by MBE of order 4, 5 and 6. Based on these results one can conclude that:

1. MBE-3 can give a qualitatively consistent picture of the Pt4 PES but for ab initio accuracy MBE-4 is needed (Figure 18).
2. MBE-3 completely fails to give a good Pt5 PES. MBE-4 gives a qualitatively consistent PES but MBE-5 is needed to achieve ab initio accuracy (Figure 20).
3. MBE-4 captures many important features of the Pt6 PES, MBE-5 performs much better but without ab initio accuracy (Figure 22).
4. In Figure 22 we notice that MBE-6 does not reproduce the exact Pt6 PES.

It is evident that MBE-4 or higher is required to get a qualitatively correct PES for Pt clusters. However, the applicability of relatively low order MBE to bigger clusters requires further investigation. The final point, is a demonstration of another difficulty that needs to be addressed. The reason why MBE-6 fails to reproduce the exact Pt6 PES is the small errors introduced to the potentials V_K during their fitting procedure. This error is of the same order as the accuracy of the ab initio calculation but it propagates in a complicated manner through the Multi-Body Expansion Eq. (27) possibly amplifying itself. Even if the error of each V_K is reduced, application of raw MBE to sufficiently large clusters could potentially amplify it. The nature of this propagation is largely unknown and further research is required to address it properly.

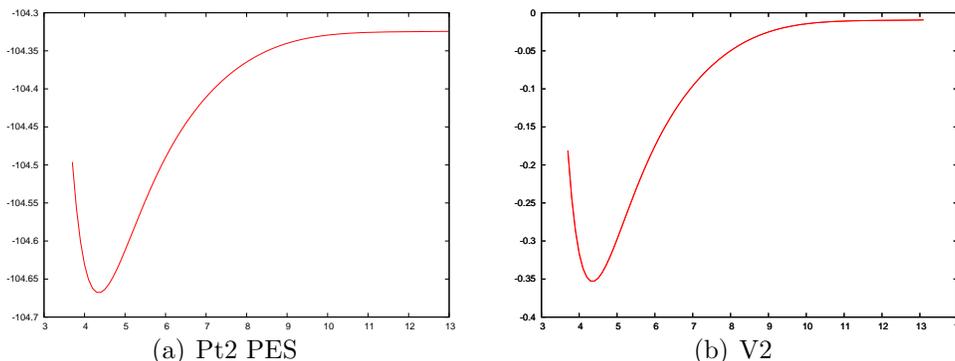


Figure 13: Pt2: Extracting the V2 potential from the PES.

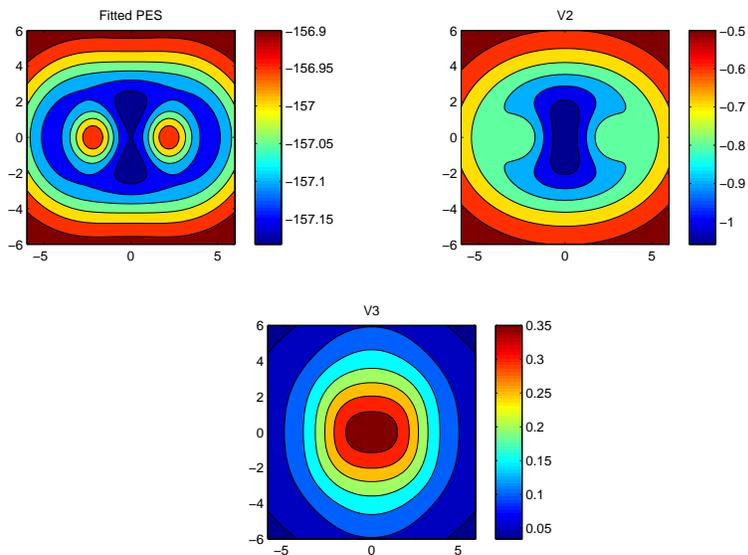


Figure 14: Pt3: Decomposition of the energy of a Pt3 cluster.

Using MBE to predict the energies of larger clusters. In this very last example, we sample some Pt clusters with number of atoms $N = 7, 8$ and 10 , calculate their ab initio energies and compare them to MBE-K $K = 2, 3, 4, 5, 6$. Figure 23 plots jointly all the results. Notice that the main contribution to the expansion comes from the one energy terms and that potentials add corrections of the order of $1 - 2$ Ryd. The accuracy after MBE-3 for Pt7 and Pt8 of the order of 0.05 Ryd (0.6 eV) per atom which is below our goal of 0.01 eV per atom. Furthermore, notice that despite the fact that consecutive MBE approximations fluctuate about the correct value of the energy, they don't seem to converge (up to order 6). As a matter of fact, the oscillations are more pronounced for the largest Pt10 cluster. We believe that this is a clear demonstration of the unaccounted propagation of the error in the potentials we first observed in the results of the previous paragraph.

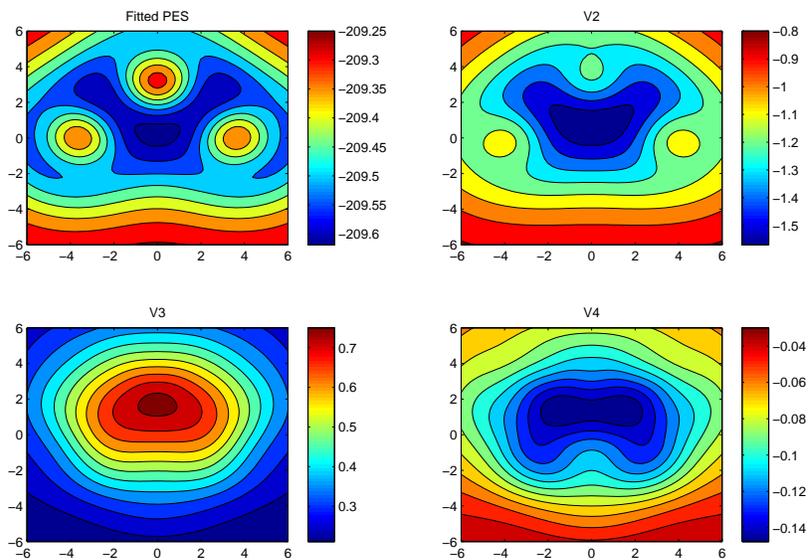


Figure 15: Pt4: Decomposition of the energy of a Pt4 cluster.

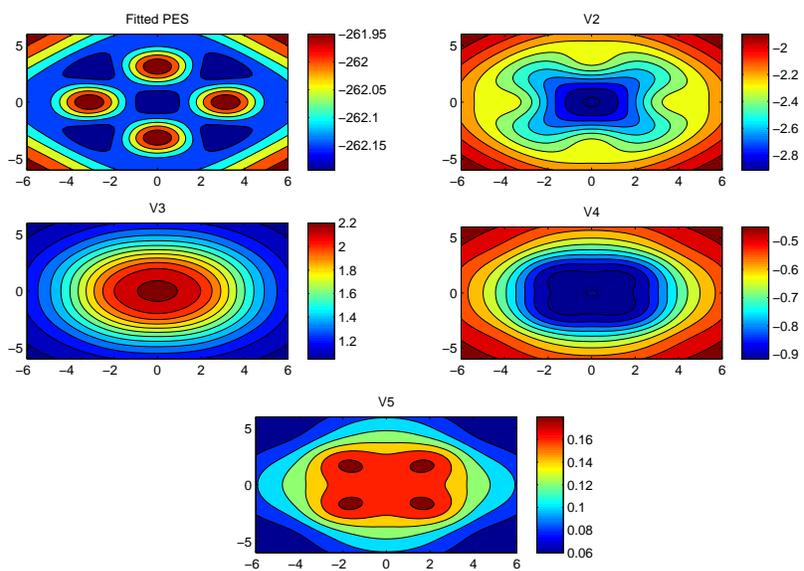


Figure 16: Pt5: Decomposition of the energy of a Pt5 cluster.

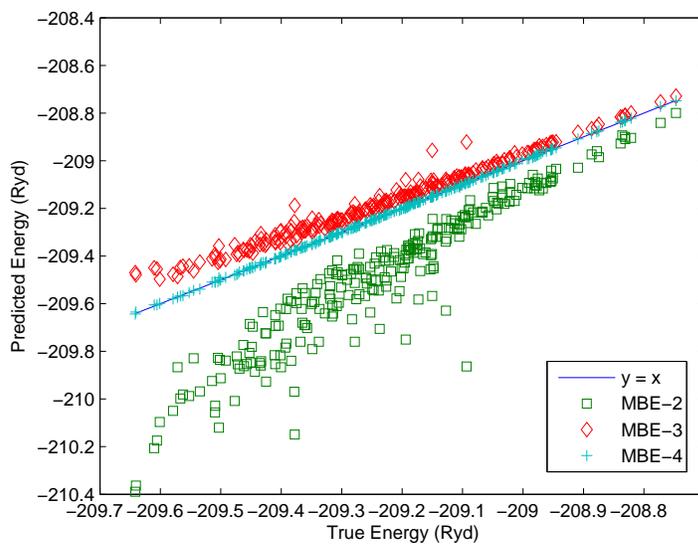


Figure 17: Pt4: Comparing the prediction of successive MBE approximations on 256 test points.

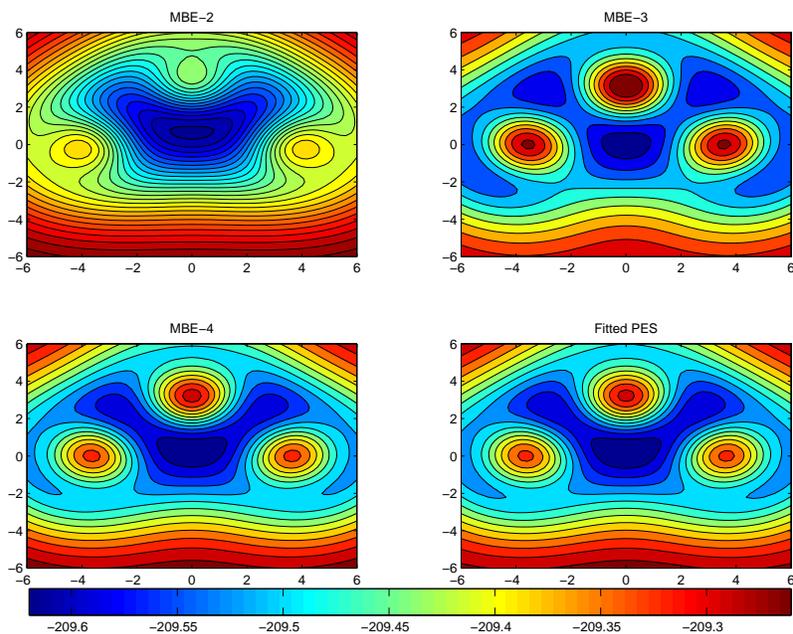


Figure 18: Pt4: Prediction of successive MBE approximations.

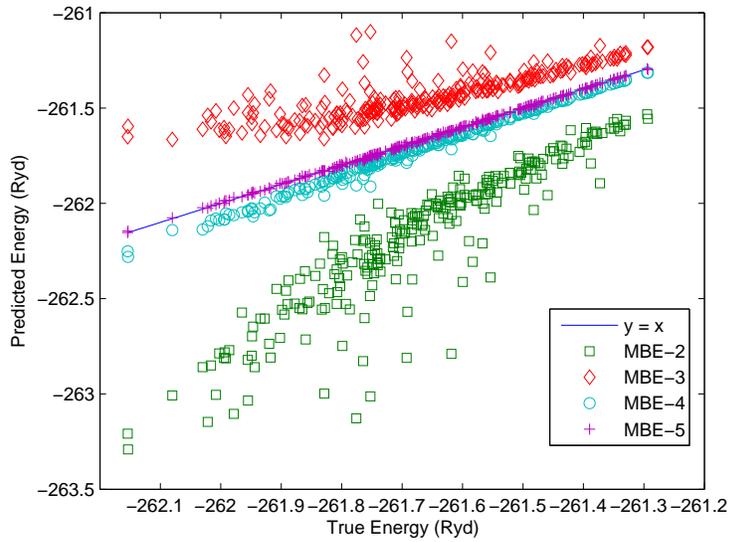


Figure 19: Pt5: Comparing the prediction of successive MBE approximations on 256 test points.

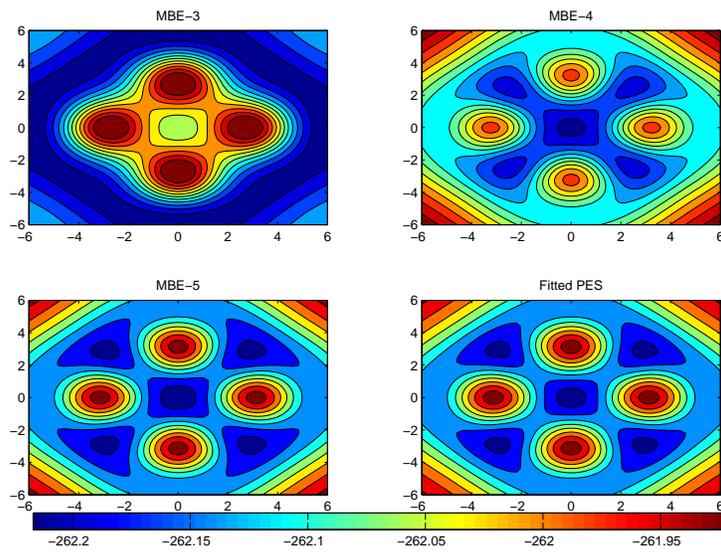


Figure 20: Pt5: Prediction of successive MBE approximations

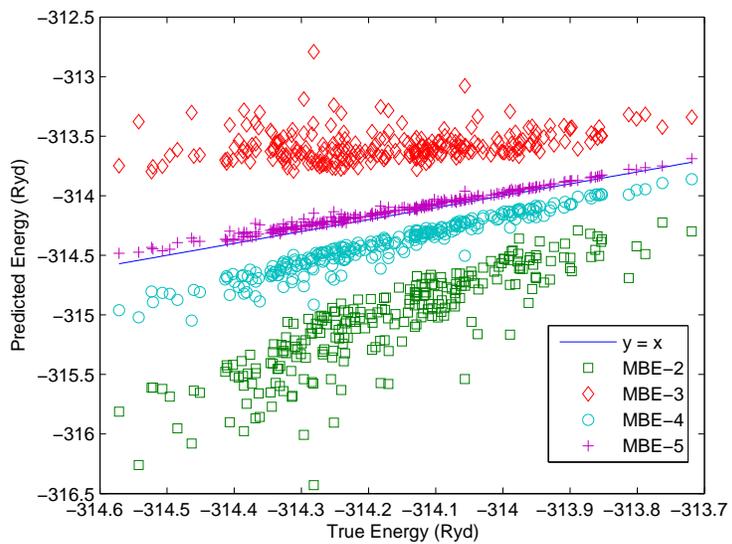


Figure 21: Pt6: Comparing the prediction of successive MBE approximations on 256 test points.

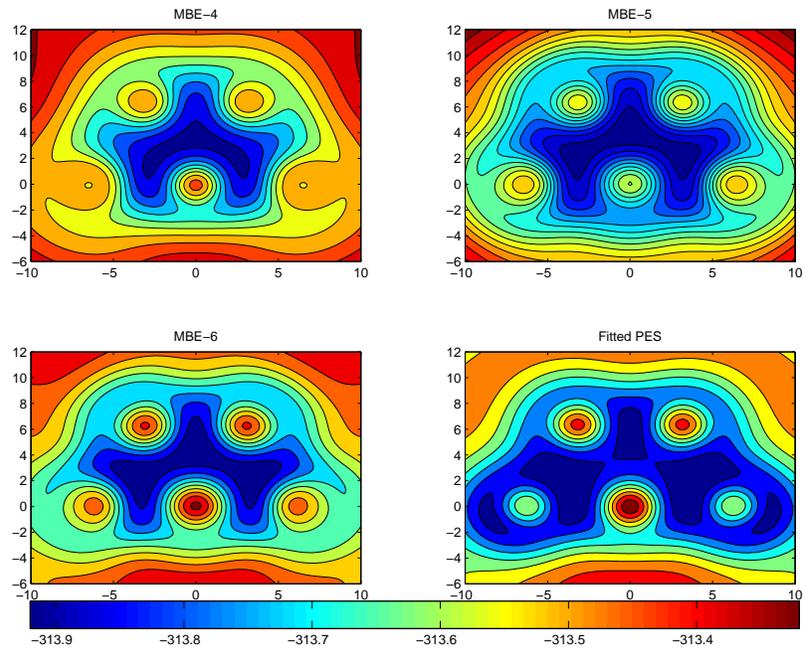


Figure 22: Pt6: Prediction of successive MBE approximations

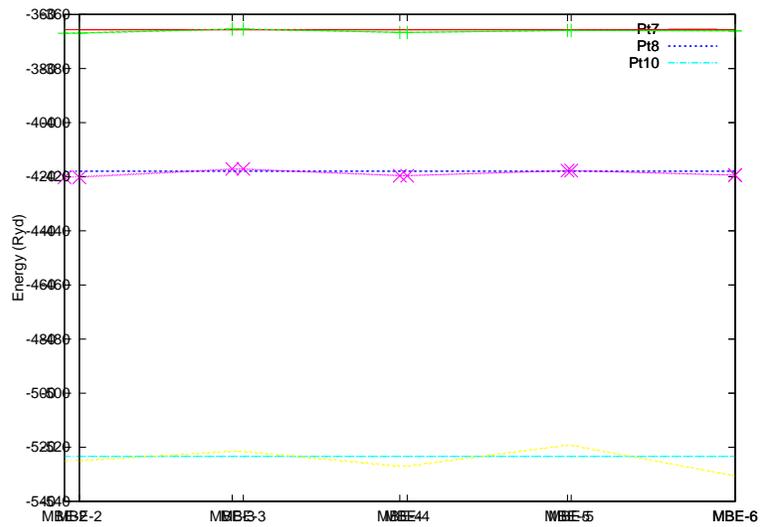


Figure 23: MBE predictions of 3 random clusters Pt7, Pt8 and Pt10.

4. Conclusions

The construction of ab initio accurate Potential Energy Surfaces constitutes a problem of extreme complexity, albeit of vital importance towards the search for materials of extremal properties. The computational cost of ab initio calculations makes mandatory the search for sophisticated techniques to tabulate ab initio data. The Multi-Body Expansion provides a mathematically rigorous framework that allows one to break down the PES of an arbitrarily sized cluster to interatomic potentials of relatively small order. These potentials are connected to Potential Energy Surfaces of small order through the Möbius transformation. In this work, we presented a set of computational techniques to efficiently solve the PES interpolation problem. The potentials can be interpolated using the exact same scheme.

We defined mathematically the configuration space and provided algorithms derived from the Distance Geometry literature to efficiently sample it. The recently developed multinomial expansion method was called to provide a polynomial basis invariant with respect to permutations of like-atoms. This provided us with candidate functions that satisfied all invariance principles of an energy surface. We introduced a Bayesian Linear Regression scheme to fit the weights of the polynomial basis as well as the evidence approximation to optimize the scale parameter. The variance of the prediction was used to devise a simple, yet effective, way to adaptively add data points. We demonstrated that this new technique considerably improves the quality of the samples obtained and outperforms the random selection of data points. Furthermore, it minimizes the number of ab init calculations required enabling us to construct the ab init PES of Pt clusters of up to 6 atoms using a reasonable amount of computational resources. The ab initio PES was used to find the stable structures of small Pt clusters with the aid of Simulated Annealing. The results were found to be in good agreement with the literature. The constructed Pt PES was also used to fit the interatomic potentials up to order 6. It was shown that those become less and less important as their order increases, albeit slowly in low energy regions. We used the potentials to investigate the performance Multi-Body Expansions of various order for Pt clusters of up to 10 atoms. It was demonstrated that interactions of at least 5 atoms are required to qualitatively describe Pt clusters. Finally, we observed that the error introduced during the fitting procedure of the interatomic potentials propagates in a complicated manner through the Multi-Body Expansion formula making its naive application to big clus-

ters questionable. It is the object of our current research to investigate the propagation of this error through the MBE formula and design effective techniques to filter it out. We believe that such filtering schemes have to be case specific (different for each material) and should utilize further physical information. This problem constitutes the final obstacle towards the construction of fully transferable potential energy surfaces using the MBE framework. The impact of such reduced-order fully transferable PES is expected to be significant in the search for new materials, exploring materials and materials surface design, optimizing mechanical and thermophysical properties, etc.

References

- [1] N. W. Ashcroft, N. D. Mermin, Solid State Physics, Academic Press Inc., 1976.
- [2] M. A. Collins, Molecular Potential Energy Surfaces for Chemical Reaction Dynamics, *Theor. Chem. Acc.* 108 (6) (2002) 313–324. doi:10.1007/s00214-002-0383-5. URL <http://www.springerlink.com/index/NVXYCK37WJ75CLN4.pdf>
- [3] L. M. Raff, M. Malshe, M. Hagan, D. I. Doughan, Ab Initio Potential-Energy Surfaces for Complex, Multichannel Systems Using Modified Novelty Sampling and Feedforward Neural Networks, *J. Chem. Phys.* URL <http://link.aip.org/link/?jcp/122/84104>
- [4] J. Ischtwan, M. A. Collins, Molecular Potential Energy Surfaces by Interpolation, *J. Chem. Phys.* 100 (11) (1994) 8080–8088. doi:10.1063/1.466801. URL <http://link.aip.org/link/?JCPSA6/100/8080/1>
- [5] A. J. C. Varandas, Extrapolation to the Complete-Basis-Set Limit and the Implications of Avoided Crossings: The $X^1\Sigma_g^+$, $B^1\Delta_g$, and $B'^1\Sigma_g^+$ states of C_2 , *J. Chem. Phys.* 129 (23) (2008) 234103. doi:10.1063/1.3036115. URL <http://www.ncbi.nlm.nih.gov/pubmed/19102522>
- [6] J. M. Bowman, B. J. Braams, S. Carter, C. Chen, G. Czako, B. Fu, X. Huang, E. Kamarchik, A. R. Sharma, B. C. Shepler, Y. Wang, Z. Xie, Ab-Initio-Based Potential Energy Surfaces for Complex Molecules and

- Molecular Complexes, *J. Phys. Chem. Lett.* 1 (12) (2010) 1866–1874.
doi:10.1021/jz100626h.
URL <http://pubs.acs.org/doi/abs/10.1021/jz100626h>
- [7] B. J. Braams, J. M. Bowman, Permutationally Invariant Potential Energy Surfaces in High Dimensionality, *Int. Rev. Phys. Chem.* 28 (4) (2009) 577–606. doi:10.1080/01442350903234923.
- [8] Z. Xie, J. M. Bowman, Permutationally Invariant Polynomial Basis for Molecular Energy Surface Fitting via Monomial Symmetrization, *J. Chem. Theory Comp.* 6 (1) (2010) 26–34. doi:10.1021/ct9004917.
URL <http://pubs.acs.org/doi/abs/10.1021/ct9004917>
- [9] T. Ishida, G. C. Schatz, Automatic Potential Energy Surface Generation Directly from Ab Initio Calculations Using Shepard Interpolation: A Test Calculation for the $H_2 + H$ System, *J. Chem. Phys.* 107 (9) (1997) 3558–3568. doi:10.1063/1.474695.
URL <http://link.aip.org/link/?JCP/107/3558/1>
- [10] R. P. A. Bettens, M. A. Collins, Learning to Interpolate Molecular Potential Energy Surfaces with Confidence: A Bayesian Approach, *J. Chem. Phys.* 111 (3) (1999) 816. doi:10.1063/1.479368.
URL <http://link.aip.org/link/?JCP/111/816/1>
- [11] R. P. A. Bettens, T. A. Hansen, M. A. Collins, Interpolated Potential Energy Surface and Reaction Dynamics for $O(^3P) + H_3^+(^1A'_1)$ and $OH^+(^3\Sigma^-) + H_2(^1\Sigma_g^+)$, *J. Chem. Phys.* 111 (14) (1999) 6322. doi:10.1063/1.479937.
URL <http://link.aip.org/link/?JCP/111/6322/1>
- [12] R. P. A. Bettens, M. A. Collins, Potential Energy Surfaces and Dynamics for the Reactions between $C(^3P)$ and $H_3^+(^1A'_1)$, *J. Chem. Phys.* 108 (6) (1998) 2424. doi:10.1063/1.475655.
URL <http://link.aip.org/link/?JCP/108/2424/1>
- [13] R. P. A. Bettens, M. A. Collins, Interpolated Potential Energy Surface and Dynamics for the Reactions Between $N(^4S)$ and $H_3^+(^1A'_1)$, *J. Chem. Phys.* 109 (22) (1998) 9728. doi:10.1063/1.477643.
URL <http://link.aip.org/link/?JCP/109/9728/1>

- [14] T. Ho, T. Hollebeek, H. Rabitz, L. B. Harding, G. C. Schatz, A Global H_2O Potential Energy Surface for the Reaction $O(^1D) + H_2 \rightarrow OH + H$, J. Chem. Phys. 105 (23) (1996) 10472–10486. doi:10.1063/1.472977.
URL <http://link.aip.org/link/?JCP/105/10472/1>
- [15] A. Kawano, Y. Guo, D. L. Thompson, A. F. Wagner, Improving the Accuracy of Interpolated Potential Energy Surfaces by Using an Analytical Zeroth-Order Potential Function, J. Chem. Phys. 120 (14) (2004) 6414–6422. doi:10.1063/1.1667458.
URL <http://link.aip.org/link/?JCPSA6/120/6414/1>
- [16] R. Dawes, D. L. Thompson, A. F. Wagner, M. Minkoff, Interpolating Moving Least-Squares Methods for Fitting Potential Energy Surfaces: A Strategy for Efficient Automatic Data Point Placement in High Dimensions, J. Chem. Phys. 128 (8) (2008) 84107. doi:10.1063/1.2831790.
URL <http://link.aip.org/link/?JCP/128/084107/1>
- [17] G. G. Maisuradze, D. L. Thompson, A. F. Wagner, Interpolating Moving Least-Squares Methods for Fitting Potential Energy Surfaces: Detailed Analysis of One-Dimensional Applications, J. Chem. Phys. 119 (19) (2003) 10002–10014. doi:10.1063/1.1617271.
URL <http://link.aip.org/link/?JCPSA6/119/10002/1>
- [18] Y. Guo, A. Kawano, D. L. Thompson, A. F. Wagner, M. Minkoff, Interpolating Moving Least-Squares Methods for Fitting Potential Energy Surfaces: Applications to Classical Dynamics Calculations, J. Chem. Phys. 121 (11) (2004) 5091. doi:10.1063/1.1777572.
URL <http://link.aip.org/link/?JCPSA6/121/5091/1>
- [19] Y. Guo, L. B. Harding, A. F. Wagner, M. Minkoff, D. L. Thompson, Interpolating Moving Least-Squares Methods for Fitting Potential Energy Surfaces: An Application to the H_2CN Unimolecular Reaction, J. Chem. Phys. 126 (10) (2007) 104105. doi:10.1063/1.2698393.
URL <http://link.aip.org/link/?JCP/126/104105/1>
- [20] R. Dawes, D. L. Thompson, Y. Guo, A. F. Wagner, M. Minkoff, Interpolating Moving Least-Squares Methods for Fitting Potential Energy Surfaces: Computing High-Density Potential Energy Surface Data from Low-Density Ab Initio Data Points, J. Chem. Phys. 126 (18) (2007)

184108. doi:10.1063/1.2730798.
URL <http://link.aip.org/link/?JCPA6/126/184108/1>
- [21] Y. Guo, I. Tokmakov, D. L. Thompson, A. F. Wagner, M. Minkoff, Interpolating Moving Least-Squares Methods for Fitting Potential Energy Surfaces: Improving Efficiency via Local Approximants, *J. Chem. Phys.* 127 (21) (2007) 214106. doi:10.1063/1.2805084.
URL <http://link.aip.org/link/?JCP/127/214106/1>
- [22] D. F. R. Brown, M. N. Gibbs, D. C. Clary, Combining Ab Initio Computations, Neural Networks, and Diffusion Monte Carlo: An Efficient Method to treat Weakly Bound Molecules, *J. Chem. Phys.* 105 (17) (1996) 7597. doi:10.1063/1.472596.
URL <http://link.aip.org/link/?JCPA6/105/7597/1>
- [23] S. Manzhos, T. Carrington, Using Neural Networks to Represent Potential Surfaces as Sums of Products, *J. Chem. Phys.* 125 (19) (2006) 194105. doi:10.1063/1.2387950.
URL <http://adsabs.harvard.edu/abs/2006JChPh.125s4105M>
- [24] S. Manzhos, T. Carrington, A Random-Sampling High Dimensional Model Representation Neural Network for Building Potential Energy Surfaces, *J. Chem. Phys.* 125 (8) (2006) 84109. doi:10.1063/1.2336223.
URL <http://link.aip.org/link/?JCPA6/125/084109/1>
- [25] H. M. Le, L. M. Raff, Cis \rightarrow trans trans \rightarrow Cis isomerizations and NO bond dissociation of nitrous acid (HONO) on an ab initio potential surface obtained by novelty sampling and feed-forward neural network, *J. Chem. Phys.* 128 (19) (2008) 194310. doi:10.1063/1.2918503.
URL <http://link.aip.org/link/?JCPA6/128/194310/1>
- [26] V. Sundararaghavan, N. Zabaras, Weighted Multibody Expansions for Computing Stable Structures of Multiatom Systems, *Phys. Rev. B* 77 (6) (2008) 1–10. doi:10.1103/PhysRevB.77.064101.
URL <http://link.aps.org/doi/10.1103/PhysRevB.77.064101>
- [27] R. Drautz, M. Fähnle, J. M. Sanchez, General Relations Between Many-Body Potentials and Cluster Expansions in Multicomponent Systems, *J. Phys.: Condens. Matter* 16 (23) (2004) 3843. doi:10.1088/0953-8984/16/23/005.
URL <http://iopscience.iop.org/0953-8984/16/23/005>

- [28] B. Paulus, K. Rosciszewski, N. Gaston, P. Schwerdtfeger, Convergence of the Ab Initio Many-Body Expansion for the Cohesive Energy of Solid Mercury, *Phys. Rev. B* 70 (16) (2004) 165106.
URL <http://link.aps.org/doi/10.1103/PhysRevB.70.165106>
- [29] T. F. Havel, Distance Geometry: Theory, Algorithms, and Chemical Applications, *Enc. Comp. Chem.* (1998) 1–20.
URL <http://www.ti.inf.ethz.ch/ew/courses/GCMB07/material/lecture13/havel-dist>
- [30] T. F. Havel, I. D. Kuntz, G. M. Crippen, The Theory and Practice of Distance Geometry, *Bull. Math. Biol.* 45 (5) (1983) 665–720.
URL <http://www.springerlink.com/index/162704705886080m.pdf>
- [31] T. F. Havel, Metric Matrix Embedding in Protein Structure Calculations, NMR Spectra Analysis, and Relaxation Theory, *Magn. Reson. Chem.* 41 (S1) (2003) S37–S50. doi:10.1002/mrc.1242.
URL <http://onlinelibrary.wiley.com/doi/10.1002/mrc.1242/pdf>
- [32] W. Bosma, J. J. Cannon, C. Playoust, The Magma Algebra System I: The User Language, *J. Symb. Comput.* 24 (3/4) (1997) 235–265.
URL <http://www.math.ru.nl/~bosma/pubs/JSC1997Magma.pdf>
- [33] Z. Xie, Effective Monomial Symmetrization Approach (EMSA) Program.
URL <http://www.mcs.anl.gov/research/projects/msa/>
- [34] T. F. Havel, K. Wüthrich, An Evaluation of the Combined Use of Nuclear Magnetic Resonance and Distance Geometry for the Determination of Protein Conformations in Solution, *J. Mol. Biol.* 182 (2) (1985) 281–294.
URL <http://linkinghub.elsevier.com/retrieve/pii/0022283685903468>
- [35] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer New York, 2006.
- [36] C. P. Robert, *The Bayesian Choice*, Springer New York, 2001.
- [37] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes: the Art of Scientific Computing*, Cambridge Univ. Press, 2007.

- [38] J. Sacks, W. J. Welch, T. J. Mitchell, H. P. Wynn, Design and Analysis of Computer Experiments, Stat. Science.
URL <http://www.jstor.org/stable/2245858>
- [39] K. F. Ireland, M. I. Rosen, A Classical Introduction to Modern Number Theory.
- [40] N. Chen, Modified Möbius Inverse Formula and its Applications in Physics, Phys. Rev. Lett. 64 (1990) 1193–1195.
doi:10.1103/PhysRevLett.64.1193.
URL <http://link.aps.org/doi/10.1103/PhysRevLett.64.1193>
- [41] N. Chen, G. Ren, Carlsson-Gelatt-Ehrenreich Technique and the Möbius Inversion Theorem, Phys. Rev. B 45 (14) (1992) 8177–8180.
doi:10.1103/PhysRevB.45.8177.
URL <http://link.aps.org/doi/10.1103/PhysRevB.45.8177>
- [42] A. P. Sutton, J. Chen, Long-Range FinnisSinclair Potentials, Phil. Mag. Lett.
- [43] J. D. Gale, A. L. Rohl, The General Utility Lattice Program (GULP), Mol. Sim. 29 (5) (2003) 291–341.
URL <http://www.informaworld.com/index/713812272.pdf>
- [44] P. Giannozzi, S. Baroni, N. Bonini, QUANTUM ESPRESSO: A Modular and Open-Source Software Project for Quantum Simulations of Materials, J. Phys.: Condens. Matter.
- [45] A. M. Rappe, K. M. Rabe, E. Kaxiras, J. D. Joannopoulos, Optimized Pseudopotentials, Phys. Rev. B 41 (2) (1990) 1227.
URL <http://link.aps.org/doi/10.1103/PhysRevB.41.1227>
- [46] M. Snir, MPI: the Complete Reference, MIT Press, 1998.
- [47] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, Optimization by simulated annealing, Science.
- [48] M. Galassi, GNU Scientific Library Reference Manual, 2009.
URL <http://portal.acm.org/citation.cfm?id=1538674>

- [49] L. Xiao, L. Wang, Structures of Platinum Clusters: Planar or Spherical?,
J. Phys. Chem. A.
URL <http://pubs.acs.org/doi/abs/10.1021/jp0485035>