CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

CLUSTER GEOMETRY OPTIMIZATION USING MECHANISTIC RANDOM GENERATIONS TO IMPROVE GENETIC ALGORITHMS

WAGNER BERNARDES QUINTAO ROMERO

Orientador: BRENO RODRIGUES LAMAGHERE GALVAO
Centro Federal de Educação Tecnológica

Coorientador: JOSE LUIZ ACEBAL FERNANDES
Centro Federal de Educação Tecnológica

BELO HORIZONTE
DECEMBER, 2019

WAGNER BERNARDES QUINTAO ROMERO

CLUSTER GEOMETRY OPTIMIZATION USING MECHANISTIC RANDOM GENERATIONS TO IMPROVE GENETIC ALGORITHMS

Dissertação apresentado ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional

Linha de pesquisa: Métodos Matemáticos Aplicados

Orientador: BRENO RODRIGUES LAMAGHERE GALVAO
Centro Federal de Educação Tecnológica

Coorientador: JOSE LUIZ ACEBAL FERNANDES
Centro Federal de Educação Tecnológica

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL
BELO HORIZONTE
DECEMBER, 2019

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
COORDENAÇÃO DO PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

"CLUSTER GEOMETRY OPTIMIZATION USING MECHANISTIC RANDOM GENERATIONS TO IMPROVE GENETIC ALGORITHMS"

Dissertação de Mestrado apresentada por **Wagner Bernardes Quintão Romero**, em 20 de fevereiro de 2020, ao Programa de Pós-Graduação em Modelagem Matemática e Computacional do CEFET-MG, e aprovada pela banca examinadora constituída pelos professores:

Prof. Dr. Breno Rodrigues Lamaghere Galvão
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. José Luiz Acebal Fernandes
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Mateus Xavier Silva
Universidade Federal de Minas Gerais

Prof. Dr. Rodrigo Tomás Nogueira Cardoso
Centro Federal de Educação Tecnológica de Minas Gerais

Visto e permitida a impressão.

Prof. Dr. Thiago de Souza Rodrigues
Coordenador do Programa de Pós-Graduação em
Modelagem Matemática e Computacional

Romero, Wagner Bernardes Quintão
R763c Cluster geometry optimization using mechanistic random generations to improve genetic algorithms / Wagner Bernardes Quintão Romero. – 2020.
xi, 36 f.

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional.
Orientador: Breno Rodrigues Lamaghere Galvão.
Coorientador: José Luiz Acebal Fernandes.
Dissertação (mestrado) – Centro Federal de Educação Tecnológica de Minas Gerais.

1. Análise de aglomeração – Teses. 2. Otimização matemática – Teses. 3. Algoritmos genéticos – Teses. 4. Química quântica – Teses. I. Galvão, Breno Rodrigues Lamaghere. II. Fernandes, José Luiz Acebal. III. Centro Federal de Educação Tecnológica de Minas Gerais. IV. Título.

CDD 511.6

Elaboração da ficha catalográfica pela Biblioteca-Campus II / CEFET-MG

Resumo

Nanopartículas são pequenos aglomerados (clusters) com poucos nanômetros de dimensão podendo conter até cerca de um milhão de átomos. O comportamento físico-químico destes clusters está fortemente ligado à sua geometria, número de partículas e composição. Devido à complexidade do problema, a configuração destes clusters não pode ser compreendida puramente por dados experimentais. Portanto, são necessários modelos ou hipóteses para dar sentido aos resultados destes experimentos. Computadores modernos possibilitam o avanço de abordagens *ab initio* que são capazes de prever a estrutura de menor energia de um cluster mesmo sem nenhum dado experimental, embora sejam eficientes apenas para pequenos clusters demandam alto custo computacional. Fenômenos quânticos contribuem para a dificuldade de se prever a geometria ótima de clusters em casos que se necessitam de grande exatidão. Quanto maior a exatidão esperada para os resultados, mais fenômenos quânticos devem ser levados em consideração, o que aumenta a complexidade do cálculos da energia do cluster. Métodos de busca por soluções, especialmente Algoritmos Genéticos (GA), tem sido usados em uma variedade de estudos apresentando boas soluções em uma maneira altamente paralelizável. Estes GAs se beneficiam de metáforas elaboradas e têm sido constantemente aprimorados ao longo dos anos. Neste trabalho, um método mecanicista para a geração aleatória de clusters é usado a fim de melhorar a eficiência de GAs usados para optimização de clusters.

Palavras-chave: Otimização de clusters. Algoritmos Genéticos. Química quântica.

Abstract

Nanoparticles are small particle agglomerates (clusters) of few nanometers ranging from a dozen up to a million atoms. The cluster behavior is strongly linked to its geometry, the number of particles and composition. The cluster configuration cannot be determined by experimental data solely and further hypothesis are necessary for giving meaning to the experimental results. The improvement of the current computer programs makes possible the advance of *ab initio* approaches which are able to determine the structure of the lowest energy without the need for any experimental data, but only for small clusters and at expenses of high computational cost. Quantum effects involved in clusters increase the difficulty of valuing geometries, especially when high accuracy is required. The higher the expected accuracy is, the more of those quantum aspects should be taken under consideration, increasing the complexity of the required algorithms. Search for solutions methods, especially using Genetic Algorithms (GA), have been used in a variety of cluster studies bringing light to good solutions in a highly parallel fashion assessing marginally the geometry domain. GA methods benefit from elegant metaphors and have been constantly improved throughout the years. In the present work, a mechanistic method for random generation of clusters is studied to improve the GA learning in cluster geometry optimization methods.

Keywords: Cluster optimization. Genetic Algorithm. Quantum Chemistry.

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

SE	Schrödinger equation.
PES	Potential Energy Surface.
GA	Genetic Algorithms.
DFT	Density Functional Theory.
IEG	Inhomogeneous Electron Gas.
L-BFGS	Broyden-Fletcher-Goldfarb-Shanno (limited memory).
BFGS	Broyden-Fletcher-Goldfarb-Shanno (full memory).
GUI	Graphical user interface.
GAMESS	General Atomic and Molecular Electronic Structure System.

Contents

Chapter 1

Introduction

Every known industry sector is intimately constrained to a specific set of materials. Curtarolo et al. (2013) stated that the rare event of changing those materials in a well-established technology is to be considered a revolution. Particularly the growth of industrial demands in specialized sectors such as electronics or biotechnology highlights the necessity for the comprehension of materials in the nanoscopic scale for business development.

Nanoparticles are small particle agglomerates (clusters) of few nanometers ranging from a dozen to a million atoms. Despite having the same chemical composition, the physical-chemical behavior of substances varies significantly when comparing cluster state to bulk solid state. Johnston (2003) states that is particularly relevant for both research and development in fields like catalysis, fullerenes or nanotechnology in general.

Clusters can be composed of different atom types or even a collection of molecules, being observed in various environments like molecular beams, vapor phase, colloidal suspensions, isolated in inert matrices or over surfaces (JOHNSTON, 2003). The behavior of clusters is strongly linked to its geometry, the number of particles and composition so that relevant physical properties can be calculated in terms of those features.

As stated by Marks (1994) and Silva, Galvão & Belchior (2014), the cluster configuration cannot be understood by experimental data solely meaning that mathematical models and further hypotheses are necessary to give meaning to the experimental results.

It is necessary to elaborate a suitable approximate energy function to represent the cluster, considering the quantum and electromagnetic phenomena involved. The problem is then to minimize this energy function to find the target cluster geometry that corresponds to the most probable atomic arrangement represented as the global minimum of the total energy of the system.

The *ab initio* approaches model the cluster behavior by basic physical principles instead of relying on any empirical data for the energy calculation. That constitutes an unbiased approach considering quantum aspects of the clusters and hence going beyond the problem of minimizing electronic attraction and repulsion or any empirical force field models.

The higher the expected accuracy is, the more of those quantum aspects should be taken under consideration. Fortunately, the underlying physical laws requested for those are completely known as refereed by Dirac (1929) almost a hundred years ago. However, the celebrated resolution of the complex physical models has the drawback of a considerable computational cost.

The quantum effects involved in cluster energy calculations contribute to the unpredictability of the output geometries, especially when high accuracy is required. Therefore every possible geometry is a reasonable prospect solution to the problem. The so-called Potential Energy Surface (PES) to be surveyed, is a hypersurface of potential energy values, and it is used to represent the domain of the N-dimensional space of cluster geometries.

The use of Genetic Algorithms (GA) focuses the core of the computational effort to the search for minima among the PES in a highly parallel fashion recombining randomly generated building blocks of candidate solutions (genes) to yield the putative global minimum while assessing only marginally the PES.

According to Mitchell (1998), the application of a GA involves a considerable amount of choices, and since search methods, especially GA, benefit from elegant metaphors. There is an inherent need for proper encoding, and there is no ideal choice for the coding of mutation, crossover, or even for the random generation of the initial population. Some approaches to implement GA in cluster optimization apply empirical chemical properties as heuristics for the initial population (JOHNSTON, 2003; HEILES et al., 2012; KAZAKOVA; WU; RAHMAN, 2013).

The mechanisms in which such heuristics affect the GA can be either positive or detrimental since they can create an unwanted bias, preventing a proper exploration of the PES. Many of those approaches use references from the average atomic displacement of the bulk solid state to limit the space in which the initial prospect solutions are generated.

That creates a demand for knowledge about the sparsity of the clusters' atoms before geometry optimization. Furthermore, the function to be optimized is unconstrained, and the wrong estimate of those limits can affect the efficiency of the GA.

A proposal for different use of the atomic distance is presented. The proposed approach is used for the generation of the first population and does not have any constrain to the overall packing factor or the cluster's limiting shape.

Chapter 2

Objectives

2.1 General objectives

The objective of this study is **to evaluate a new strategy for the generation of the initial population of clusters to be used in Genetic Algorithms (GA)**.

2.2 Specific objectives

1. *Elaborate a mechanistic algorithm for a random cluster generation in a 3-dimensional space subjected to constraints such as bond length and normal distribution parameters.*
2. *Improve the strategy used for first GA population using optimal bond distances or bulk average distance as a reference.*
3. *Evaluate the proposed generation algorithm in comparison to reference one.*

Chapter 3

Theoretical Background

The theoretical background of this work is divided into two main subjects: I. Calculation of the cluster energy (section 3.1), and II. Implementation of Genetic Algorithms (section 3.2 and section 3.3).

3.1 Quantum aspects of clusters

> The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. (DIRAC, 1929).

The methods that only tackle the Schrödinger equation (SE) for giving information without the need for any experimental data are called *ab initio* methods. The *ab initio* assessment for complex systems is challenging without proper approximations and some computational heuristics even considering current computers. Although, as suggested by Dirac (1929), the whole of chemistry lies in the solution of SE equation.

Some of those approximations not only make possible the *ab initio* assessment in theoretical chemistry but also enables inherent notions of the chemical practice like molecular shape, orbitals, and atom types.

3.1.1 Born-Oppenheimer approximation

Since the SE is related to the Hamiltonian operator for the total energy of the system, one can take much advantage from the mass difference between the electron and the proton. The proton weights roughly 1800 times more than the electron, motivating the approximation that considers the nucleus — a collection of protons and neutrons — fixed in regard to the motion of the electrons.

As explained by Koch & Holthausen (2015), in regard to the much-celebrated work of Born & Oppenheimer (2000), the kinetic energy of a nucleus can be disregarded, thus, the potential energy of the nucleus is constant. In this case, the Hamiltonian operator shrinks to an electronic Hamiltonian:

$$H_{elec} = T + V_{Ne} + V_{ee} \ , \tag{1}$$

where the electronic Hamiltonian (1) is influenced by the electronic kinetic energy, T, the potential energy between nucleus-electron, V_{Ne} and between electron-electron, V_{ee}.

This allows the electronic part of SE to be solved as a function of nuclear positions, resulting in the Potential Energy Surface (PES), a hypersurface in N-dimensional space. An energy value on the PES would be an eigenvalue of the electronic Hamiltonian according to the eigenvalue equation (2).

$$H_{elec}\Psi = V_{PES}\Psi \ , \tag{2}$$

where Ψ is the eletronic wave-function, or the eigenfunction associated to the eigenvalue V_{PES}.

The cluster geometry study benefits from the Born-Oppenheimer approximation for being able to separate nuclear and electronic motion with considerable reliability, that enables the assessment of V_{PES} as a function of the nuclear coordinates. The study of this PES constitutes the basis of the cluster geometry optimization.

3.1.2 Cluster Energy Calculations

Two approaches for cluster energy calculations are presented here. Empirical potentials are compared to an *ab initio* approach. Both approaches can be used to elaborate the objective function to be optimized for obtaining the most adequate cluster geometry.

3.1.2.1 Empirical potentials

Experimental information can be used to model atomic interactions and even to neglect quantum aspects of nuclear motion in an attempt to overcome the difficulties imposed by the SE. In this context, the motion of the atoms is considered as subjected to Newton's second law and the whole cluster is modeled as interacting spheres. Such approach is called force field model. At that approach, the problem of calculating optimal energy would be approximated by a search in a PES generated by a simpler objective function.

As referred by Jensen (2017), the metaphor used in the empirical potential models is mostly described as "ball and spring" model with atoms being the balls with different sizes and "softness" and the spring representing bonds having different lengths and "stiffness".

That metaphor apart, empirical potential models are mostly responsible for the notion of rigid directional bonds between atoms, the concept of atom types and the whole organic chemistry. There are many force field models. Each of them, most applicable to specific systems, being responsible for several branches of chemistry that wouldn't exist without such approximations.

From the perspective of force field methods, the clusters are usually seen as a collection of different atoms, or molecules, subjected to atomic or molecular bonds. The bond here is as an abstraction of a distinct interaction between every pair of atoms regarded as a spring.

3.1.2.2 Many-body Gupta potentials

As described by Johnston (2003), hetero-elemental metallic clusters or nanoalloys can be modeled with reasonable accuracy by the many-body Gupta potential which approximates the cluster's potential energy by a sum of attractive energy, V_{Ne}^{i}, and repulsive energy, V_{ee}^{i} for each atom i, where the overall energy of the cluster is given as:

$$V_{PES} = \sum_{i}^{N} \left(V_{ee}^{i} - V_{Ne}^{i} \right) . \tag{3}$$

The repulsive and attractive terms can be calculated separately concerning the interaction of each atomic pair i-j:

$$V_{ee}^{i} = \sum_{j \neq i}^{N} A(a,b) e^{-p(a,b)\left(\frac{r_{ij}}{r_0(a,b)} - 1\right)} , \tag{4}$$

and

$$V_{Ne}^{i} = \left(\sum_{j \neq i}^{N} \zeta^2(a,b) e^{-2q(a,b)\left(\frac{r_{ij}}{r_0(a,b)} - 1\right)} \right)^{\frac{1}{2}} . \tag{5}$$

In the equations (5) and (4), the quantities: A, r_0, ζ, p and q, are empirical quantities fitted experimentally by the values of cohesive energy, lattice parameters, and independent elastic constants of the materials in the bulk form, given in respect to the atom types a and b for the reference crystal structure at 0 K. Table 1 shows the values for the parameters used following the construction of Paiva & Galvão (2015).

Table 1 – Parameters for the Gupta potential for the Al-Mg system.

	Al-Al	Mg-Mg	Al-Mg
r_0 (Å)	2.8637	3.21	3.03685
A (eV)	0.1221	0.198722	0.160411
ζ	1.316	0.349712	0.832856
p	8.612	15.977780	12.29849
q	2.516	1.684590	2.100295

Figure 1 – Diatomic energy optimization for the bond Al-Mg.

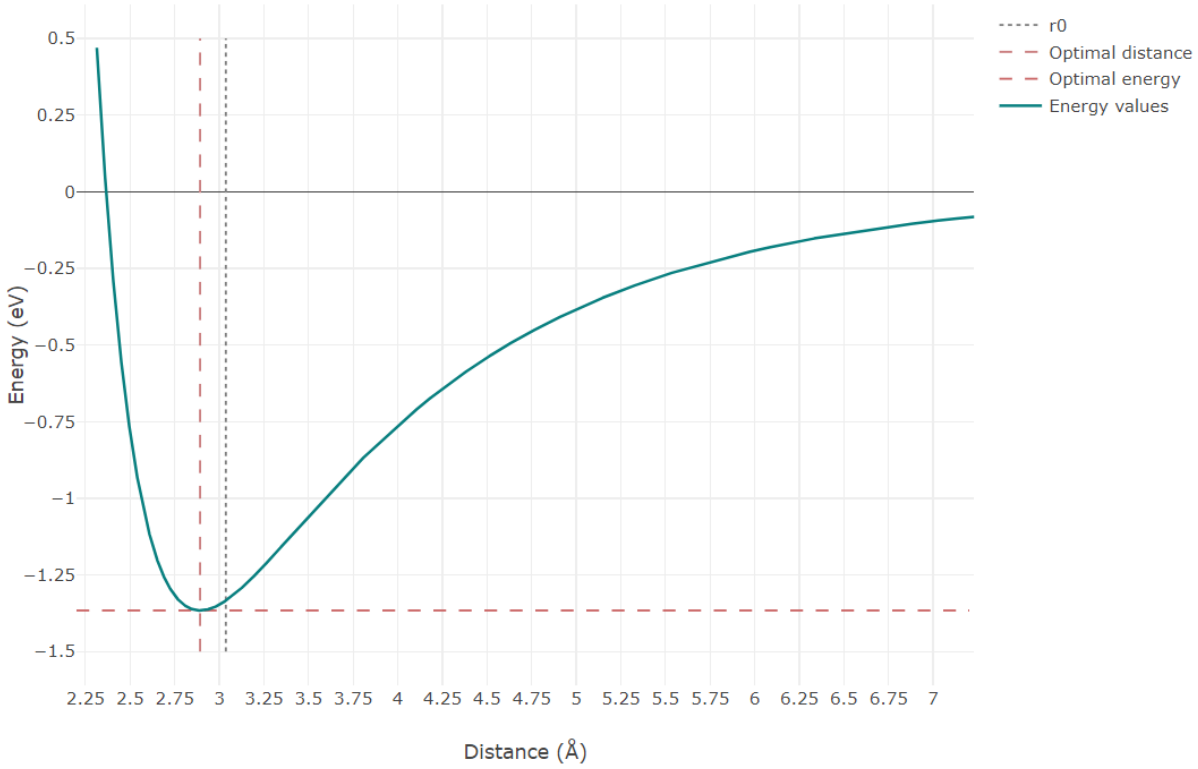


Figure 1 illustrates the Gupta energy as a function of the distance between the elements (r_{12}) for the diatomic system Al-Mg. In this case, the L-BFGS minimization can quickly obtain the global minimum of the Gupta function. In such a simplification, the distance r_{12} represents the cluster's geometry, and the optimization result is the diatomic optimal distance between the two specimens. This function depicts the diatomic optimization model in which the asymmetrical curve shows a single minimum of -1.366 eV, at 2.89246 Å relatively close to the empiric fitted minimal distance r_0 of the atoms at the bulk form (shown at table Table 1).

The curve in Figure 1 captures some of the Gupta model characteristics. The starting point is the null energy of the system with the elements apart, the energy value decreases smoothly, before reaching the optimal distance. Equation (4) asymptotically dominates the exponential growth in energy for short distances reaching even undesired positive values for $r_{12} < 2.366$ Å.

Notice that the proximity between r_0 and the diatomic optimal value is not providential since both exponents in equations (5) and (4) are affected by the distance ratio:

$$\frac{r_{ij}}{r_0(a,b)}\,. \tag{6}$$

Many works use the $r0$ distance as a reference for the packing factor or a first trial distance of the atoms in a random cluster generation (JOHNSTON, 2003; KAZAKOVA; WU; RAHMAN, 2013).

3.1.2.3 Density functional theory

The Density Functional Theory (DFT) is a method that simplifies calculations considering quantum aspects involved in clusters based on the Inhomogeneous Electron Gas (IEG) approximation as presented by Hohenberg & Kohn (1964). DFT extends the concept of the electron density to the idea of a pair density that contains all information about electron-electron correlation explained in detail in Koch & Holthausen (2015). The DFT provides an approximate function for obtaining the eigenvalue V_{PES} shown in the electronic Schrödinger equation (2).

3.2 Genetic Algorithm

> The traditional theory of GAs (first formulated in Holland 1975) assumes that, at a very general level of description, GAs work by discovering, emphasizing, and recombining good “building blocks" of solutions in a highly parallel fashion. (MITCHELL, 1998).

3.2.1 Overview of evolutionary search methods

Many problems require solutions that are too complex to program by hand. This is the case of cluster geometry optimization. The most suitable atomic arrangement is the one of the lowest energy *i.e.* the global minimum of the PES.

Assessing the energy of every possible solution of the PES means solving the SE for all possible geometries. An effort that could take centuries of computational effort requiring the algorithm to be adaptive for being able to assess only marginally the PES and yet being able to propose a reasonable solution for the putative global minimum as an alternative for a simple try-an-error procedure among an enormous number of possible solutions.

Genetic Algorithm (GA) is an example of a method for searching for solutions, where the prospect solutions are evolved to better ones by means of selection. There are other general methods like Hill Climbing, Simulated Annealing, and Tabu Search. Mitchell (1998) states that all “Search for Solutions” methods are based on the following steps:

- Initially generate candidate solutions (or a set of);
- Evaluate the candidate solutions;
- Decide which candidates will be kept and which will be discarded;
- Produce variants by using some kind of operators on the surviving candidates.

Johnston (2003) highlights that the use of GA in geometry optimization can benefit from more sophisticated approaches. Such approaches would be implemented in order to reduce the space of variables, increase accuracy and convergence.

3.2.2 Overview of Genetic Algorithms

Both the complex systems in which the GA is applied and biological evolution in nature represent cases in which there is an enormous set of possible solutions. Those solutions are expressed as genetic sequences and the best solutions are represented by the sequences that would result in highly adapted organisms to a given environment following the natural selection theory of evolution. Many adaptive computer programs use the phenomena involved in evolution as a metaphor. Although the evolution, in nature, depends on several other factors like competition and cooperation between the individuals, and even the fitness criteria itself can change in time.

The GA evolves a group of candidate solutions expressed by “chromosomes” which are a collection of “genes”. In essence, each chromosome can be coded as a string or a list of values which characterizes the state of the individuals, and each value at that list is regarded as a gene. Preferably, any change in a gene of the chromosome would produce a different individual, a different state, and a different prospected solution as well.

A feature of interest in the GA application is that instead of evolving a single solution, it evolves a subset of solutions, called population. The variability of the initial population can result in a broader search and the stability of the final population indicates convergence.

The GA implementation together with its operators are based on the specifications of the approached problem. The simplest form of GA application would require at least three types of operators: selection, crossover, and mutation.

Selection: Takes place to distinguish solutions. The operator will be taken over the population to select well fitted chromosomes that would be most likely reproduced.

Crossover: Such operator is inspired by reproduction and provides the exchange of information between two different chromosomes to create an offspring.

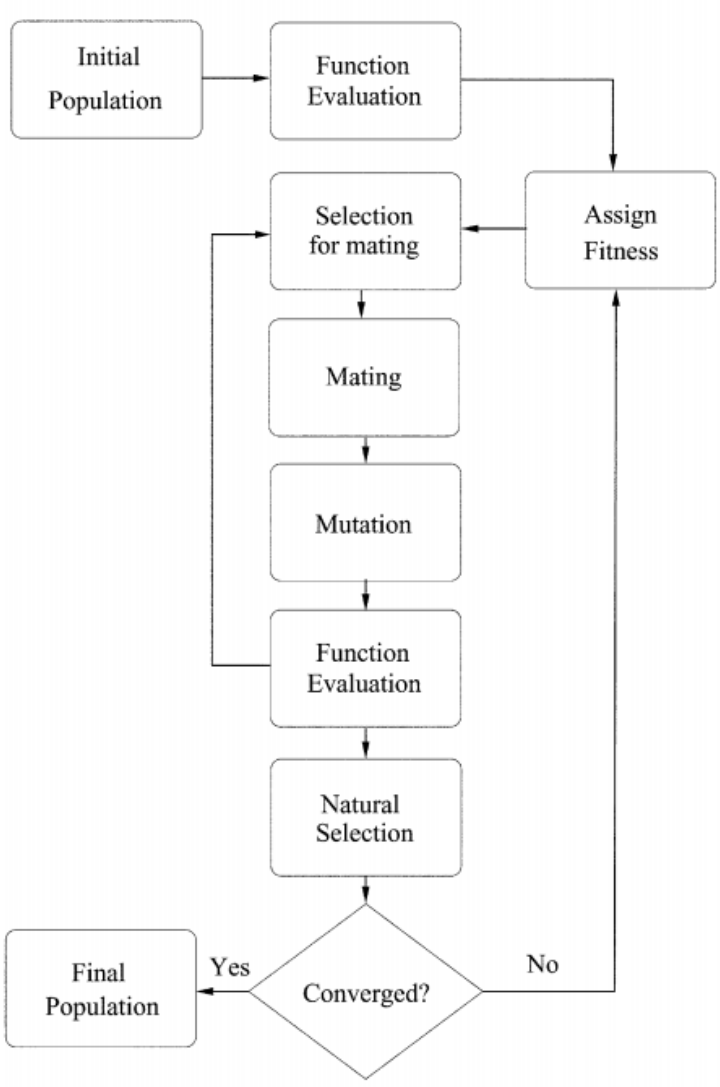
Mutation: Such operation occurs with a relatively rare frequency. Variations of the mutation operator are used to prevent stagnation of the population since objective functions could have basins of attraction that would work as traps, and thus enabling a better exploration of the space of solutions.

3.3 GAs for cluster geometry optimization

GAs have been used in chemistry since the 1990s. The early works in cluster geometry optimization were done independently by Hartke (1993) and Xiao & Williams (1993) using GA for minimizing the empiric equations of molecular clusters.

Figure 2 illustrates the flow of a standard GA substituting the usual operator crossover by the so-called mating operator (subsection 3.3.4).

Figure 2 – Example of flow chart for a GA



Source: Johnston (2003)

3.3.1 Encoding cluster optimization for the Genetic Algorithms

The application of a GA involves a considerable amount of choices. There is an inherent need for proper encoding. The proposed data structure must capture every aspect of the candidate solution affecting the calculation of the objective function, and it governs both the resolution and the computation cost of the operators used throughout the search. The data structure is a strong pillar of an efficient algorithm, but proper encoding also constrains the interpretation of the results. A particular encoding could enable further insights, metaphors, or any heuristics, that could benefit both the efficiency and the accuracy of the GA.

The early works of cluster geometric optimization used a binary encoded GA for the geometric optimization small silicon clusters (up to Si_4) (HARTKE, 1993), and benzene (C_4H_6), naphthalene ($C_{10}H_8$) and anthracene ($C_{14}H_{10}$) (XIAO; WILLIAMS, 1993). The binary encoding involves the discretization of the spatial coordinates converted into binary values, an approach that enables the GA operators to work in a bitwise fashion. In binary encoded candidate solutions, essential GA operations, like Crossover and Mutation, work simply by trimming and flipping bits in a string.

Zeiri (1995) introduced a GA using real-valued Cartesian coordinates opposing the bitwise encoding. The new encoding used continuous variables in a vector space to represent each atom position in the cluster. Using the vector representation, the GA operates over a schema made of the three variables that represent each atom positions in the cluster, rather than a bit locus in a binary transcription of the whole candidate solution. Each atomic coordinate can be then associated with a specific atom type. The correlation between each atomic coordinate is a relevant parent characteristic the offspring preserves. The use of real-valued Cartesian coordinates thus adds accuracy to the GA algorithm while also enabling a higher resolution for each candidate solution.

The use of real-valued Cartesian coordinates thus adds accuracy to the GA algorithm while also enabling a higher resolution for each candidate solution. Deaven & Ho (1995) stated that standard bitwise Crossover operation does not preserve sufficiently the characteristics of the parents.

With a suitable encoding for the candidate solutions, the next logical step is to create a method for the generation of a random individual and then the first population. Several strategies have been presented throughout the years with the aim to improve the search domain exploration, speeding up convergence towards the global minimum without stalling or biasing the algorithm. The evolution of such strategies is presented here in this chapter.

3.3.2 The initial population

Since GA is a stochastic algorithm, the starting point has a significant effect on the convergence. There is a chance that the algorithm would get stuck in a basin showing early convergence or would not have a budget to explore thoroughly a particular configuration that could reach a lower energy level.

There is a set of GA parameters that could prevent early convergence on-the-fly leveling the selective pressure by acting upon fitness scaling or just by the result of mutation. Both resources are explained later in this work, although there is an intrinsic limitation of mutation and fitness scaling in the context of GA. Considering that the essence of GA, as explained by Holland (1992), is to recombine solutions to create better ones, the first population not only sets a starting point for the algorithm but it also feeds the search with the building blocks that will be recombined and mutated to form better solutions.

The generation approach is intimately linked to how the space of solutions is explored for the energy minimization problem. These structures can be more compact or sparse depending on the strategy used, and even an unbiased generation would have a chance of finding preferably certain types of structures.

A cluster shape cannot be modeled solely by labeling the apparent two-dimensional (2D) envelope shape. This type of oversimplification might force one's intuition to disregard the contribution of the atoms at the core of the cluster. Noticeably, a shape evaluation throughout the cluster atoms is hardly conspicuous from any 3-dimensional (3D) visualization since the N atoms in the cluster would have at least $3N - 6$ degrees of freedom. A brief description of cluster shape descriptors is shown in appendix A.

A suitable generation approach would lead to:

1. Better exploration of the space of solutions by increasing initial population diversity;
2. Reduced population of conceptually unfit structures;
3. Increased proportion of structures with lower raw energy value.

3.3.3 Local minimization

Deaven & Ho (1995) performed a gradient driven local optimization of the cluster energy for the C_{60} molecule. In their approach, every individual was subjected to a gradient driven local or minimization molecular dynamics quenching to perform what they called relaxation. In such a method, every cluster geometry is relaxed to the nearest a local minimum in the PES.

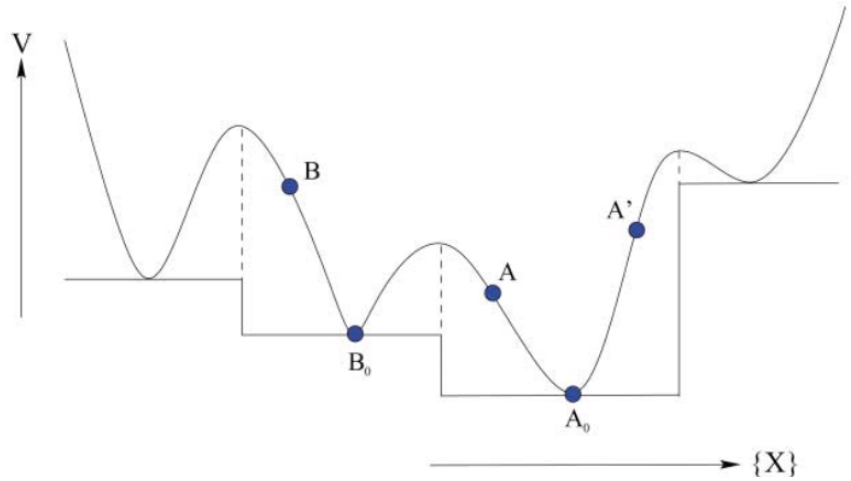
Pereira et al. (2007) highlighted the use of the Broyden-Fletcher-Goldfarb-Shanno limited memory quasi-Newton method (L-BFGS) as a powerful optimization technique combining features of modest computational costs and the super-linear convergence exhibited by the full memory quasi-Newton method (BFGS).

The L-BFGS algorithm is a quasi-Newton algorithm used for local minimization of strictly convex functions, continuously differentiable in unconstrained systems ((NOCEDAL, 1980)). L-BFGS works well with strictly convex functions, continuously differentiable for unconstrained systems. In such method, the user specifies the number of corrections (k) of an initial Hessian matrix (H_0) to be stored in memory. These corrections are used to create a sparse approximation of the Hessian matrix H_k used for finding a minimum in a multivariable objective function as described by Liu & Nocedal (1989).

As illustrated in Figure 3 the relaxation occur moving individuals from a continuous energy surface (illustrated by the continuous curve) to the nearest point of attraction *i.e.* the local minimum (illustrated by the stepped lines bellow the continuous curve). In the figure, the point B were relaxed into point B_0 and both points A and A' were relaxed into A_0.

Johnston (2003) presents the local minimization as a Lamarckian approach in the GA context. Most GA applications perform the relaxation to the nearest minimum at every step. As a result, the offspring's genetic characteristic is the one obtained after the local minimization rather than the one it inherited directly from the parents.

Figure 3 – Illustration of the potential energy surface



Source: Johnston (2003)

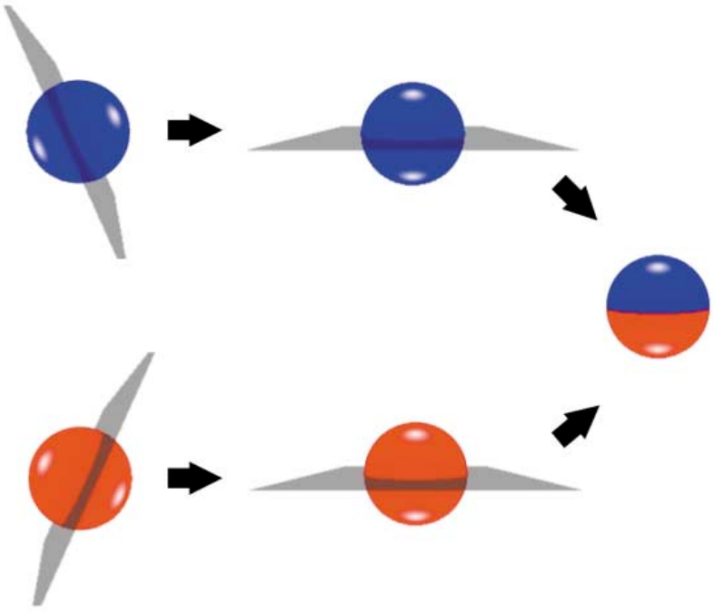
3.3.4 Cut and splice crossover operator (mating)

Deaven & Ho (1995) showed a significant improvement in the energy optimization by using a cut and splice crossover operator or mating operator as presented by the authors. The operation takes over two parent geometries G and G' to yield the child $G"$:

$$P(G, G') \rightarrow G" \ . \tag{7}$$

The single-cut method, as implemented by Johnston (2003), selects a random plane passing through the center of mass of each parent to yield a single offspring. The plane virtually cuts both parents in two collections of atoms, one above and other below the plane as shown in the Figure 4. The atoms above the plane belonging to one parent are then spliced to the atoms below the plane of the other parent to create the child $G"$. If the offspring does not yield the correct number of atoms, both parents have to move the same distance in opposite directions perpendicular to the plane. The spliced offspring would also undergo relaxation to the nearest local minimum so that the algorithm and its operations would only input and output local minima geometries as presented by Deaven & Ho (1995).

Figure 4 – Schematic representation of the single cut mating operator as proposed by Deaven & Ho (1995)



Source: Johnston (2003)

Johnston (2003) stated this mating operator has been employed in most subsequent cluster GA work among a significant number of GA programs used in cluster geometry optimization. In Johnston's implementation of the mating operator, the random plane would be chosen by a random rotation of the cluster in two directions and the cut itself could be done through the middle of the cluster or in a weighted fashion taking more atoms from the fittest cluster. The

same study refers to improvements done by other authors highlighting the implementation of new pseudo-genetic operators or new ways of handling populations in GA operations like crossover and mutation. Pereira & Marques (2010) presented a variant for the mating operator taking only the n-nearest neighbors of a randomly chosen atom of a parent, adding a few from the second parent and even generating random new atoms if required.

3.3.5 Mutation schemes

Mitchell (1998) emphasizes that the role of mutation is to prevent the loss of diversity. Many studies — including Deaven & Ho (1995) and Johnston (2003) — revealed the notion that even good mating and selection operators wouldn't prevent stagnation. They noticed that some ecology seemed to be trapped in an attractive watershed increasing the number of iterations necessary to find the fittest minimum in comparison to the algorithms with mutation. Johnston (2003) presented different schemes that would be used for mutation, depending on the type of cluster being studied:

Atom displacement: Since the cluster is represented as a set of atom types and their coordinates a simple mutation would just replace a subset of those coordinates by randomly generated values.

Twisting: The first steps of the Deaven and Ho mating operator are applied to simply rotate the upper half of the cluster's atoms in the axis perpendicular to the plane.

Cluster replacement: An entire cluster is replaced by a randomly generated one.

Atom permutation: Hetero-elemental clusters would have the position of pairs of different atom types switched without altering the clusters' geometry.

The local minimization operator can also affect the output of the mutation complementarily.

3.3.6 Dynamic fitness scaling

In most applications of GA for cluster geometry optimization the cluster energy is the matter of concern for the selection. The energy value itself is often used alone for differentiating clusters and maintaining diversity in the ecology even though many geometries could have similar energy values.

To work around the similarities in the energy value a dynamic fitness scaling was implemented by Johnston (2003). A dimensionless normalized value for energy was implemented in his work to compare the clusters in the population. The assigned value ρ_i of the energy V_i of the current cluster i was taken in regard to the worst individual of the current population (*i.e.* the one with the biggest energy V_{max}) and the current fittest individual with the lowest energy

V_{min} as in (8):

$$\rho_i = \frac{V_i - V_{min}}{V_{max} - V_{min}} \,, \tag{8}$$

this normalized energy value is then assigned to a fitness function to control how rapidly fitness falls off with an increasing cluster energy.

3.3.7 The solution "building blocks"

As stated by Mitchell (1998) in an attempt to describe GA effectiveness: good solutions tend to be made up of good building blocks. The statement was concerning the theoretical background of GA first explained by Holland (1992). In Holland's theory, GA would work combining randomly generated genes but also learning from the combinations in the way. The notion of building blocks was formalized as schemata. Every GA operation could create a bias towards a larger group of individuals emphasizing the fittest and learning about the search environment on the fly. Johnston (2003) considers good schemata to be regions of the parent clusters, subgroups of atomic arrangement, that contribute to the overall energy minimization. This learning mechanism is arguably translated in the Deaven & Ho (1995) mating operator that recombines a collection of such subgroups by trimming geometrical hemispheres rather than randomly sorting atom positions regardless of the surroundings.

Chapter 4

Related Work

4.1 Random cluster generation

Random cluster generation algorithms aim to yield atom positions to be associated with atom types for the realization of the initial population. Such coordinates usually are real-valued numbers following the insights of Zeiri (1995).

4.1.1 Scaling limitations of standard generations

If the algorithm generates each coordinate value randomly and regardless of the previous choices, the range (R_{max}) in which the algorithm chooses each distance constitutes an enclosing cube for the whole cluster where each coordinate is generated according to:

$$x_i \leq R_{max} \ . \tag{9}$$

The fitness of such an algorithm is intrinsically dependent on the choice for the magnitude of the displacement (R_{max}). Hence, creating a scaling problem relative to many cluster properties such as the number of atoms and each atom's physical-chemical properties. If the same range is carelessly used to enclose different clusters with different atom numbers or different atom types, some clusters would be generated with an unwanted bias towards a higher or lower packing factor.

Johnston (2003) generated real-valued Cartesian coordinates randomly following a uniform distribution $[0, \sqrt[3]{N}]$ so that the volume of the box would scale proportionally to the number of atoms (N). Kazakova, Wu & Rahman (2013) implemented a GA for the geometry optimization of nanoalloys of Cu–Au using a force field model for the cluster energy. They have followed the same strategy of Zeiri (1995), but also using the parameter (r_0) as a reference for the cube scale targeting better Gupta potential values. The edge of the cube thus grew proportionally to:

$$x_i \leq r_0 \sqrt[3]{N} \ . \tag{10}$$

Heiles et al. (2012) implemented a similar approach for studying nanoalloys of Au–Ag clusters, combining GA with DFT calculations generating the initial population randomly in a sphere whose radius, ρ_{max} is given as:

$$\rho_{max} = 1.1\ r_0 \sqrt[3]{N} \tag{11}$$

The target quantity r_0 in equations (10) and (11) can be both the bonding distance in the pure solid or arithmetic mean of the different references for atomic distances in hetero-elemental clusters taken as an average effect for the whole cluster.

4.1.2 Other strategies

Cai et al. (2002) have implemented a Fast Annealing Evolutionary Algorithm in which it followed one of two possible strategies for the initial population: (i) a completely random generation in a shpere or (ii) growing a seed taken from previous results.

For the random generation in a sphere, the radius scaled with (12):

$$\rho_{max} = r_e \left(\frac{1}{2} + \sqrt[3]{\frac{3N}{4\pi\sqrt{2}}} \right), \tag{12}$$

where r_e is called pair equilibrium separation and is set to 1. This quantity is analogous to the target for the average atomic distance (r_0). While growing a seed of optimizated results, clusters with less atoms are used to start the new cluster generation. In this approach the optimized cluster with $N - m$ atoms is used as the core of a N atoms cluster where the m missing atoms were generated as increment atoms at the ρ_{max} radius with random spherical displacements $\Delta\theta$ and $\Delta\phi$.

While growing seed from previous results, Cai et al. (2002) uses optimized smaller clusters as a starting point or the core of a bigger cluster subjected to the optimization. The proposed solution then needs then to randomize only the position of the missing atoms. Naturally, such an approach would require data from previous generations and a careful mixture with other approaches since the seed used as a core could create an undesired bias to the search.

The approaches presented in this chapter aim to circumvent the inherent problem of enclosing a cluster to a particular geometry at the starting point of the geometry optimization. The downfall is that even with the knowledge of empirical properties such as a reference for the target atomic distance (r_0) for a given atom type, one can not accurately learn the overall packing factor or its external shape before the optimization. Chapter 5 presents a proposal for better use of the distance r_0, which does not have any constrain to the overall packing factor or the cluster's limiting shape.

Chapter 5

Metodology

In this work, each individual of the initial population is generated as an iterative atomic addition over previously positioned atoms. Such an approach prevent the generation of clusters with stray atoms or extreme atomic repulsion.

5.1 The cluster representation

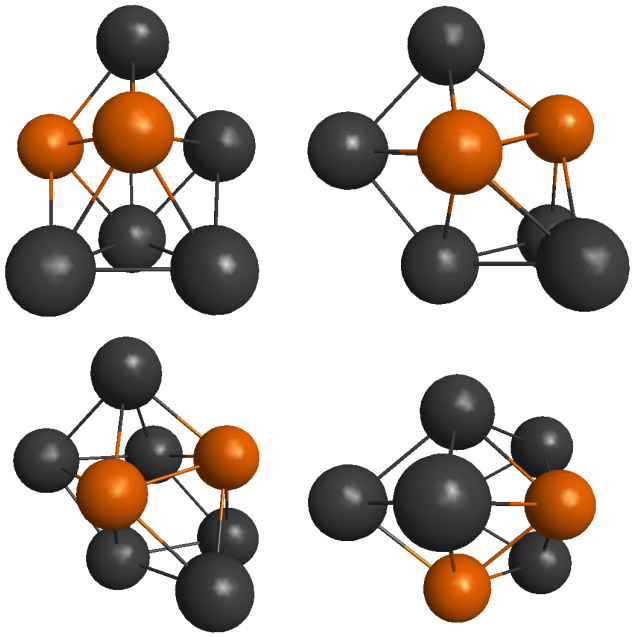
The cluster is modeled as a list of Cartesian coordinates and the associated atom types as explained in subsection 3.3.1. There are many redundancies not approached at the generation being that the same cluster can be equally represented by different lists of coordinates. Figure 5 shows different views (with different Cartesian coordinates) of the expected global minimum of the Al_6Cu_2 cluster.

Any **permutation of the coordinates** belonging to atoms of the same type in the list of coordinates would generate the same cluster since there is no expected label or locality between atoms of the same type. Any **rigid body rotation** or **translation** of the geometry would alter the coordinates of the cluster while representing the same geometric structure *i. e.* same individual.

The presence of these redundancies means that the cluster is represented by a $3N$ variable structure being N the number of atoms in the cluster. Naturally none of the transformations mentioned before would change the energy value and the energy value alone is enough to discern effectively between the clusters.

The cluster visualization is done with the help of an open source graphical user interface (GUI): $WxMacMolPlt$. This GUI is used for the General Atomic and Molecular Electronic Structure System (GAMESS) as presented by Bode & Gordon (1998) and provides 3D visualization with rotation possibilities helping intuition about the subject of optimization.

Figure 5 – Different views of the expected global minimum of Al_6Cu_2 cluster



Source: Prints taken from the program images.

5.2 The increment atoms approach

For the random generation of clusters an incremental approach was used. In such approach a string containing all atom types is generated and then sorted to work as a random queuing list for atom types. The first atom is placed at the origin. While there are still atom types in the queue, a random one is chosen as a pivot. A random increment vector is then generated in spherical coordinates and added to the pivot atom coordinates. The new coordinate is attributed to the sorted atom type and the procedure repeats until the queue is empty.

The increment approach randomizes the distance between the pivot atom and the appended one at every iteration to increase the size of the cluster without the need to guess the packing factor or the cluster volume.

The cluster generation is described in the algorithm 1. The elaboration of the increment vector is taken in a spherical coordinates system with random uniform θ and ϕ ranging from $[0, \pi]$ and $[0, 2\pi]$ respectively and the ρ value taken from a normal distribution centered in the expected bulk lattice parameter r_0 or a diatomic optimized distance. The normal distribution standard deviation is usually set as 10% of r_0.

The method does not constraint the volume of the cluster. The cluster is evaluated with respect to the distance between the increment atom and its nearest neighbor. This distance should not be closer than 80% of r_0 (*i.e.* the proximity range) for preventing extreme atomic repulsion, and it should not be larger than 250% of r_0 (*i.e.* maximum distance) for preventing the inclusion of any stray atoms in the cluster. If an increment atom triggers a failure in either proximity or maximum distance check it is disregarded and generated again as shown in algorithm 1.

Algorithm 1: The increment generation algorithm

Input: List of atom types and their quantities (*e.g.* Al:10, Mg:10, ...)
Output: Representation of the created cluster
create a shuffled list of atom types, $QUEUE$.
remove an $ATOM\_TYPE$ from $QUEUE$ and place it at $(0, 0, 0)$ in the $CLUSTER$.
while *there are still items in* $QUEUE$ **do**
    choose a random $BASE\_ATOM$ from $CLUSTER$
    create an increment vector in polar coordinates, INC_VECTOR with:
    $\rho \leftarrow$ selected from a random normal distribution $N(r_0, \sigma)$;
    $\theta \leftarrow$ selected from a random uniform distribution $U(0, 360)$;
    $\phi \leftarrow$ selected from a random uniform distribution $U(0, 180)$.
    convert INC_VECTOR into the rectangular coordinates vector, $RECT\_INC$.
    $NEW\_ATOM.coordinates \leftarrow (BASE\_ATOM.coordinates + RECT\_INC)$.
    $NEW\_ATOM.type \leftarrow$ next $ATOM\_TYPE$ in $QUEUE$.
    evaluate NEW_ATOM in respect to other atoms in the $CLUSTER$.
    if NEW_ATOM *pass all requirements in evaluation* **then**
        append NEW_ATOM at cluster;
        remove atom type from the $QUEUE$.
    end
end

5.2.1 Bulk and optimal diatomic distances

For *ab initio* approaches, no reference of the experimentally determined lattice parameter is presented, but the cluster generation method can be roughly the same while calculating an optimal diatomic distance as a substitute reference for r_0. A previous step is then necessary for calculating r_0 as the optimal diatomic distance of all the species involved considering the objective function being optimized (in this case, DFT).

The relaxation to the nearest local minimum was done using the python's optimization function from the scipy library that follows L-BFGS approach explained by Liu & Nocedal (1989). And all generated clusters were also subjected to relaxation after the generation process.

5.3 The reference atoms approach

This reference method can randomly generate the coordinate of the cluster's atoms using a uniform distribution and yet resembling the increment approach in terms of the cluster evaluation for a fair comparison and is shown in the algorithm 2.

As mentioned before, the evaluation of the cluster is done following the same parameters used in the increment atoms approach for proximity range and maximum distance.

The calculation of ρ_{max} shown in algorithm 2 follows the Kazakova, Wu & Rahman (2013) scalability insight where the volume of the limiter sphere scales proportionally to the number of atoms (N) and is also dependent on the Gupta parameter r_0. The calculated limiter sphere radius is:

$$\rho_{max} = r_0 \sqrt[3]{N} \,. \tag{13}$$

Algorithm 2: The reference generation algorithm

Input: List of atom types and their quantities (*e.g.* Al:10, Mg:10, ...)
Output: Representation of the created cluster
calculate the spherical radius limiter, ρ_{max}.
create a shuffled list of atom types, $QUEUE$.
while *there are still items in* $QUEUE$ **do**
    create a position vector in polar coordinates, $POSITION$ with:
    $\rho \leftarrow$ selected from a random uniform distribution $U(0, \rho_{max})$;
    $\theta \leftarrow$ selected from a random uniform distribution $U(0, 360)$;
    $\phi \leftarrow$ selected from a random uniform distribution $U(0, 180)$.
    convert $POSITION$ into the rectangular coordinates vector.
    $NEW\_ATOM.coordinates \leftarrow RECT\_POSITION$.
    $NEW\_ATOM.type \leftarrow$ next $ATOM\_TYPE$ in $QUEUE$.
    evaluate NEW_ATOM in respect to other atoms in the $CLUSTER$.
    if NEW_ATOM *pass all requirements in evaluation* **then**
        append NEW_ATOM at cluster;
        remove atom type from the $QUEUE$.
    end
end

Chapter 6

Results

6.1 Energy distribution of randomly generated clusters

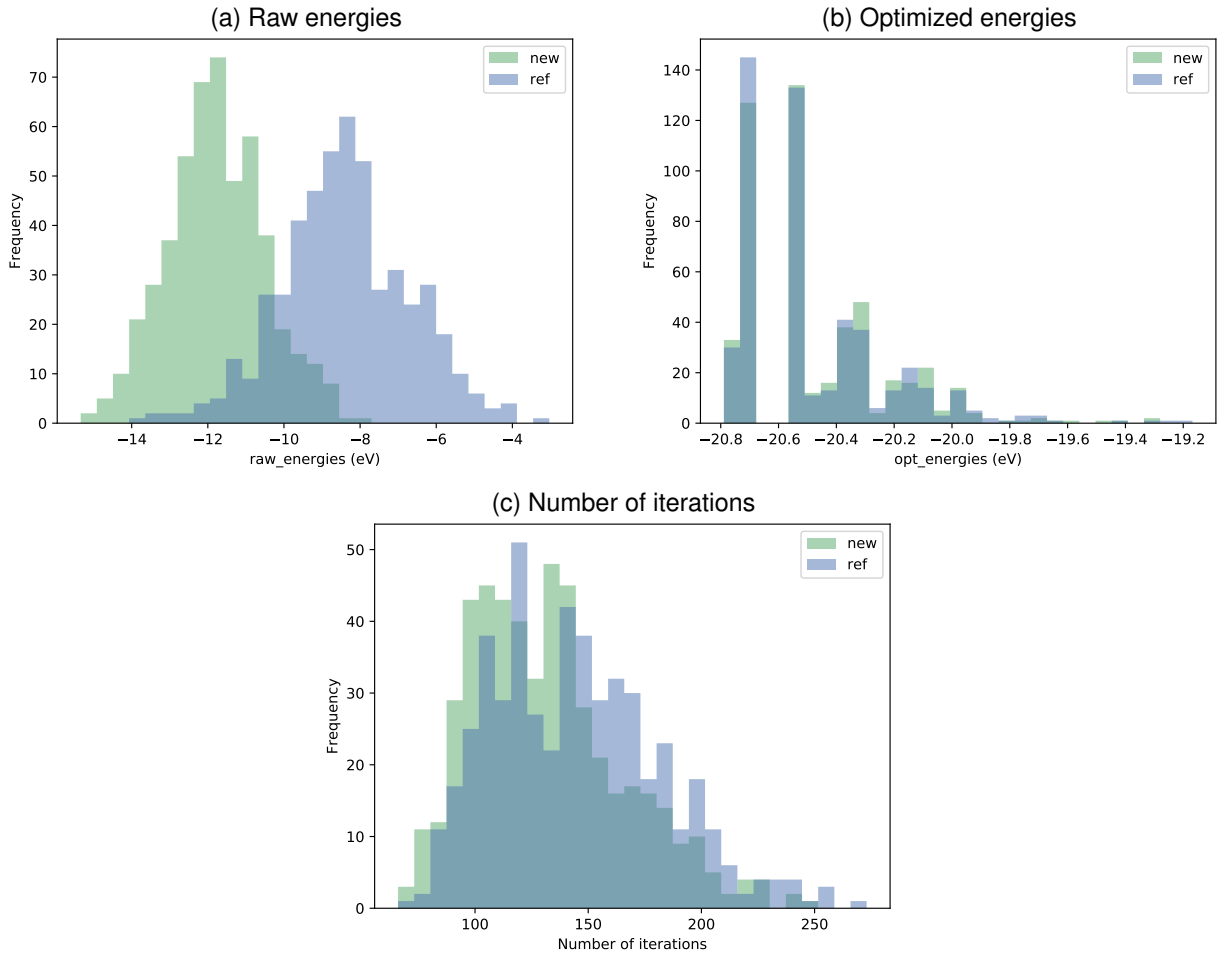
The new incremental strategy (hereafter denoted as ***new***) was compared to the spherical random reference generation (termed ***ref***). In the reference procedure the atoms are simply sorted in a sphere whose volume grows proportionally to the number of atoms. The main difference is that in the reference strategy the inclusion of an atom in the cluster was roughly independent of the atoms previously included although respecting the same validation parameters for proximity and stray atoms.

For validating the proposed strategy, 500 generated clusters of the Al_xMg_x stoichiometry with 12, 16 and 20 atoms (Al_6Mg_6, Al_8Mg_8 and $Al_{10}Mg_{10}$) were analysed with respect to 3 indicators: *The cluster energy before any optimization step (**raw energy**)*, *the cluster energy after L-BFGS optimization (**optimized energy**)* and *the **number of iterations** performed by the local optimization process (L-BFGS)*. The average values are shown in Table 2 and their distributions are pictured in Figure 6.

Table 2 – Comparison between new and ref average results.

		Average number of iterations	**Raw energy (eV)**	**Optimized energy (eV)**
Al_6Mg_6	new	131.944	-11.774245	-20.467526
	ref	144.422	-8.389084	-20.477823
Al_8Mg_8	new	160.512	-16.153776	-28.310082
	ref	165.752	-11.485943	-28.324609
$Al_{10}Mg_{10}$	new	174.744	-20.791567	-36.28177
	ref	181.404	-14.639601	-36.25815

Figure 6 – Distribution of values for the 500 Al_6Mg_6 clusters



Both incremental (proposed) and the reference approach were tested. The raw energy value of larger randomly generated clusters of the same stoichiometry with 40 and 60 atoms ($Al_{20}Mg_{20}$ and $Al_{30}Mg_{30}$) were also analyzed for giving scalability insights for both strategies.

6.1.1 The raw energy.

The Raw energy values resembled a normal distribution with similar standard deviation, but with a shifted average as shown in the Table 3. The distributions of energy values were closely studied for Al_xMg_x structures with 12 atoms shown in Figure 6 comparing the indicators of the incremental strategy (new) and reference strategy (ref).

A hypothesis test was performed for evaluating the hypothesis of both distributions (raw and new) being samples of the same population for each indicator. $\alpha < 0.05$ was considered a reasonable threshold for the rejection of that hypothesis. The tests considered the two samples to be independent of one another and using a scipy module for testing the null hypothesis that the 2 distributions could have identical average (expected) values.

Table 3 – Statistical tests result for ref and new raw energy values.

		Average raw energies (eV)	Standard deviation (eV)	p-value
Al_6Mg_6	new	-11.774245	1.276208	$< 10^{-5}$
	ref	-8.389084	1.696443	
Al_8Mg_8	new	-16.153776	1.664292	$< 10^{-5}$
	ref	-11.485943	2.056415	
$Al_{10}Mg_{10}$	new	-20.791567	1.917559	$< 10^{-5}$
	ref	-14.639601	2.274293	

The shift in the raw energies' distribution at Figure 6a (also visible in Table 3) indicates a larger chance of generating a better structure before optimization using the proposed strategy (new). This gap is enforced by the hypothesis test result (p-value) shown in Table 3. The significantly low p-value gives support to the conclusion that each approach has its distribution of raw energies (*i. e.* a different average and standard deviation). It is also believed that systems with more atoms tend to generate more isolated distributions.

This result encourages the search for further improvements in the generation strategy for being able to make GA runs operating over raw clusters and without frequent local relaxation. Such an improvement could reduce the computational cost of the algorithm and provide better consistency for the GA operators and their learning process.

6.1.2 The optimized energy.

The distributions for the optimized energy values are shown in Figure 6b and in Table 4. The histograms indicate that the optimized energies of both generation strategies follow the same distribution function instead of two dissociated distributions. This result is supported by the hypothesis test (p-value) shown in Table 4. The considerable value (greater than 0.05) indicates that both approaches reach the optimal energy values with similar probability.

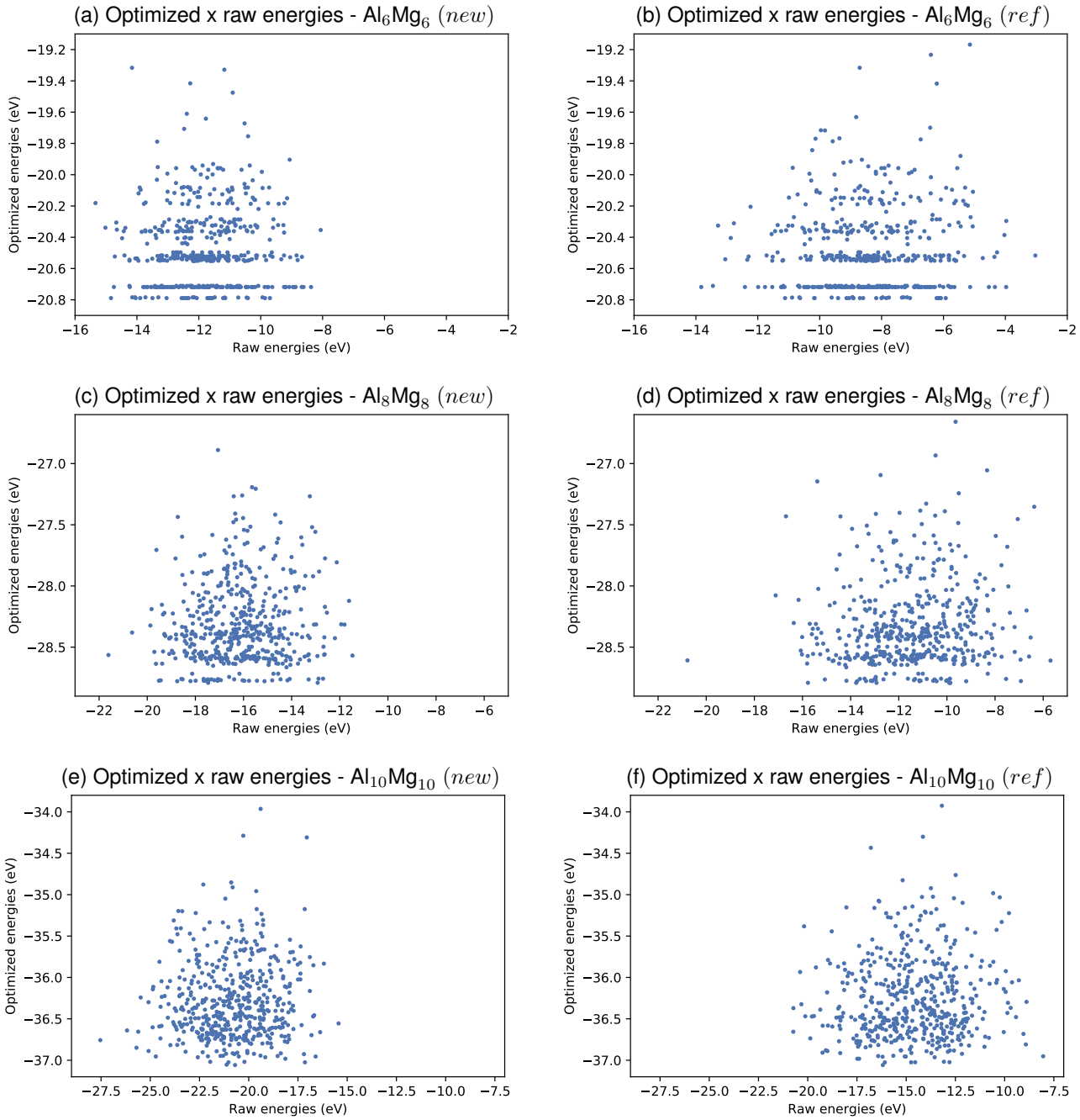
Table 4 – Comparison between new and ref optimal energy values.

		Average optimal energies (eV)	Standard deviation (eV)	p-value
Al_6Mg_6	new	-20.467526	0.259007	0.5352
	ref	-20.477823	0.265292	
Al_8Mg_8	new	-28.310082	0.331634	0.4992
	ref	-28.324609	0.347130	
$Al_{10}Mg_{10}$	new	-36.281770	0.467075	0.4380
	ref	-36.258150	0.494354	

Figure 7 presents a set of scatter plots for each individual matching the raw energy with the optimized energy showing that the frequency of achieving the global minimum is roughly the same between the proposed (new) and the reference strategy (ref) despite the proposed strategy scatter plots showing more clusters randomly generated with lower energy values.

It is noticeable in Figure 7 that the clusters with the lowest raw energies do not necessarily generate the clusters with the lowest optimized energy. No correlation was found between raw and optimized energies for the clusters studied.

Figure 7 – Raw energy against optimized energy for the 500 Al_xMg_x clusters



Another point of interest in this comparison is that the lowest optimized energies of Al_6Mg_6, shown in Figure 7a (new) and Figure 7b (ref), fall in disconnected plateaus indicating a level of discretization of the optimized structures, coherent with the results of Doye, Miller & Wales (1999) that showed the exponential growth of the number of minima with the size of the cluster.

6.1.3 The number of iterations.

The number of iterations taken by the L-BFGS method to relax the generated structure to a local minimum is relevant for a highly time-consuming objective function. In this case, the computational cost will be directly proportional to this number, and if one of the generating methods requires fewer iterations, it will be the most cost-effective. This values are shown in Table 5.

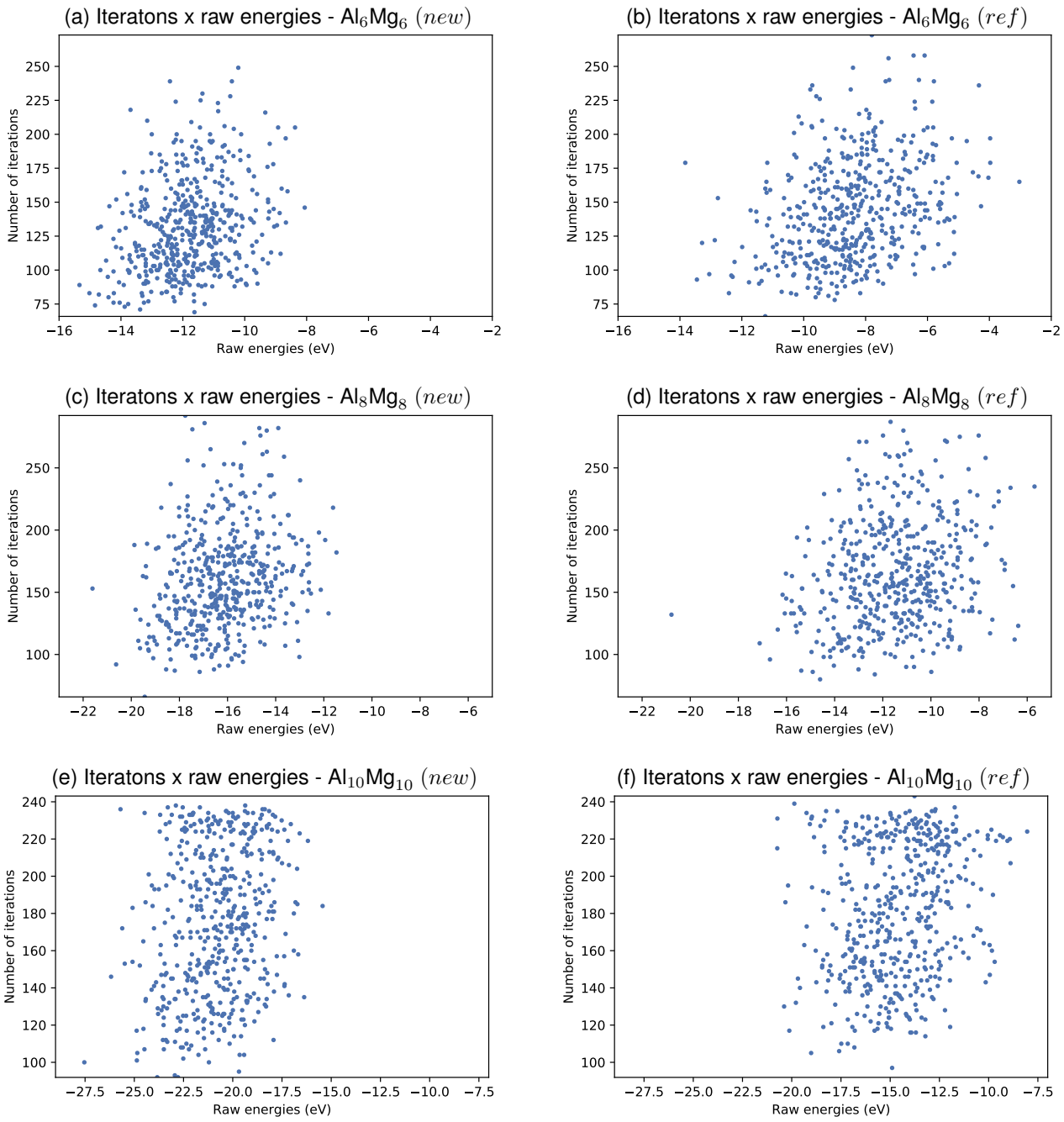
The results showed little difference between the two methods, approximately, from 65 to 300 iterations in all distributions as shown in Figure 6c. The hypothesis test result shown by the p-value in Table 5 indicated that the two approaches generated different average iterations being the new approach consistently (although moderately) more efficient.

Table 5 – Comparison between new and ref number of iterations.

		Average number of iterations	**Standard deviation (eV)**	**p-value**
Al_6Mg_6	new	131.944	34.006247	$< 10^{-5}$
	ref	144.422	37.238474	
Al_8Mg_8	new	160.512	40.024366	0.0456
	ref	165.752	42.621573	
$Al_{10}Mg_{10}$	new	174.744	38.944043	0.0049
	ref	181.404	35.542155	

No correlation was found between the raw energy value and the number of iterations as show in Figure 8. In many cases, the cluster with worse raw energy required fewer iterations for the optimization. These results are credited to the complexity of the Hessian matrix, considering that *e. g.* $Al_{10}Mg_{10}$ systems have 54 degrees of freedom.

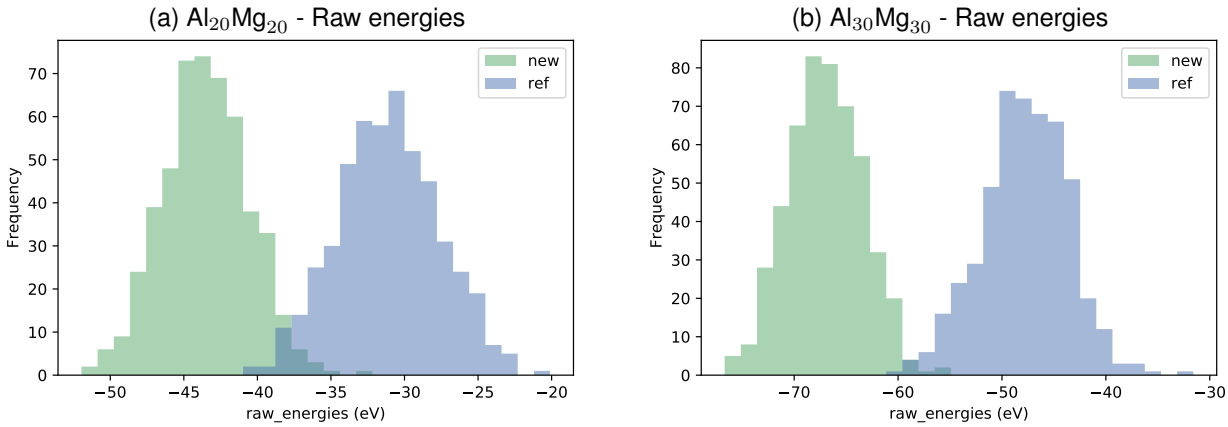
Figure 8 – Raw energy against number of iterations for the 500 Al_xMg_x clusters



6.2 Results of larger clusters

The larger structures showed, in average, a similar behavior. The mean and standard deviation ($\mu \pm \sigma$) for the raw energy distribution was of -44 ± 3 (new) against -31 ± 3 (ref) for $Al_{20}Mg_{20}$ structures, and -67 ± 4 (new) against -48 ± 4 (ref) for $Al_{30}Mg_{30}$. This behavior indicates consistency among the different distributions of each cluster generation strategy.

Figure 9 – Distribution of raw energy values for the 500 instances of larger Al_xMg_x clusters



The number of possible structures increase significantly with the number of atoms creating an immensely diverse set of expected structures for the random generation following the results of Doye, Miller & Wales (1999). Figure 9 shows two separate distributions with a significant difference in the raw energy mean values and slightly different standard deviations.

These fairly symmetrical bell-shaped distributions were also tested in regard to the normality of the distribution. A **normal probability plot** was constructed for a visual test of normality of the raw energy values of the 500 clusters with 20, 40 and 60 atoms. Each probability plot compares the normalized 500 clusters sample with a benchmark Montecarlo cumulative normal distribution in which the expected result would contract to a straight line centered on the origin if a normal distribution is found. In the normality test presented by Filliben (1975), the correlation coefficient (R) can be used to quantify the normal distribution hypothesis in the normal probability plot.

Another test was performed complementarily following the approach of d'Agostino (1971) and showed that all these distributions could be considered normally distributed. In this **normality test** both Skewness and kurtosis are considered in a two-sided hypotesis test in which the acceptance criteria is a p-value greater than 0.05 in which the clusters with 22, 40 and 60 atoms scored 0.3795. 0.5808 and 0.6188 respectively.

The normality test aims to verify if the mechanistic route for realizing the first population creates any undesirable bias against the exploration of the PES. Since all verified realizations resulted in normally distributed raw energies, **the method is credited sufficiently unbiased**.

Chapter 7

Conclusion

The proposed method showed better results for the raw energy values scoring a gap of 6, 13, 19 eV when compared to the reference method for the Al_xMg_x clusters with 20, 40 and 60 atoms respectively. This approximate 40% increase in the negative energy did not impose a significant bias, since the proposed method yielded a normally distributed random function with lower raw energies

Even though the proposed method required, on average, slightly fewer ($< 5\%$) iterations for reaching the same optimized energies in the smaller clusters, all the ecologies from both methods are believed to reach the same optimized structures with approximately the same computational cost. Hence the proposed method was considered suitable for creating the first population in a GA.

The indication of a mechanistic generation improving the raw energy distribution raises the question about other inprovements in the GA inspired by such heuristics. The randomized test showed no correlation between the distributions of raw and optimized energies, indicating that lower raw energies alone would not link to lower optimized energies. On the other hand, the construction of a more suitable and unbiased first population could lead to a more coherent learning in a GA, reducing the frequency needed for the BFGS optimization. Specially in regard to the mating operator (DEAVEN; HO, 1995).

Another remaining question is whether or not the implemented heuristics improve the GA in the case of *ab initio* approachs. For that, the whole generation procedure is expected to be done using DFT as the objective function which is more time-consuming.

Implementation of other GA operators should be tested in different objective functions as well. The goal is the evaluation of the benefit of such approaches to the GA and understand how they behave in a less predictable environment such as DFT.

Bibliography

ARTECA, G. A. Molecular shape descriptors. **Reviews in computational chemistry**, Wiley Online Library, p. 191–253, 1996. Cited 2 times in pages 35 and 36.

BODE, B. M.; GORDON, M. S. Macmolplt: a graphical user interface for gamess. **Journal of Molecular Graphics and Modelling**, Elsevier, n. 3, p. 133–138, 1998. Cited in page 20.

BORN, M.; OPPENHEIMER, R. On the quantum theory of molecules. In: **Quantum Chemistry: Classic Scientific Papers**. [S.l.]: World Scientific, 2000. p. 1–24. Cited in page 6.

CAI, W. et al. Optimization of lennard-jones atomic clusters. **Journal of Molecular Structure: THEOCHEM**, Elsevier, v. 579, n. 1-3, p. 229–234, 2002. Cited in page 19.

CURTAROLO, S. et al. The high-throughput highway to computational materials design. **Nature materials**, Nature Publishing Group, v. 12, n. 3, p. 191, 2013. Cited in page 1.

D'AGOSTINO, R. B. An omnibus test of normality for moderate and large size samples. **Biometrika**, Oxford University Press, v. 58, n. 2, p. 341–348, 1971. Cited in page 30.

DEAVEN, D. M.; HO, K.-M. Molecular geometry optimization with a genetic algorithm. **Physical review letters**, APS, v. 75, n. 2, p. 288, 1995. Cited 7 times in pages vi, 12, 13, 15, 16, 17, and 31.

DIRAC, P. Series a, containing papers of a mathematical and physical character. **Proceedings of the Royal Society of London**, v. 123, n. 792, 1929. Cited 2 times in pages 2 and 5.

DOYE, J. P.; MILLER, M. A.; WALES, D. J. Evolution of the potential energy surface with size for lennard-jones clusters. **The Journal of Chemical Physics**, AIP, v. 111, n. 18, p. 8417–8428, 1999. Cited 2 times in pages 28 and 30.

FILLIBEN, J. J. The probability plot correlation coefficient test for normality. **Technometrics**, Taylor & Francis Group, v. 17, n. 1, p. 111–117, 1975. Cited in page 30.

HARTKE, B. Global geometry optimization of clusters using genetic algorithms. **The Journal of Physical Chemistry**, ACS Publications, v. 97, n. 39, p. 9973–9976, 1993. Cited 2 times in pages 10 and 12.

HEILES, S. et al. Dopant-induced 2d–3d transition in small au-containing clusters: Dft-global optimisation of 8-atom au–ag nanoalloys. **Nanoscale**, Royal Society of Chemistry, v. 4, n. 4, p. 1109–1115, 2012. Cited 2 times in pages 2 and 19.

HOHENBERG, P.; KOHN, W. Inhomogeneous electron gas. **Physical review**, APS, v. 136, n. 3B, p. B864, 1964. Cited in page 9.

HOLLAND, J. H. **Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence**. [S.l.]: MIT press, 1992. Cited 2 times in pages 13 and 17.

JENSEN, F. **Introduction to computational chemistry**. [S.l.]: John wiley & sons, 2017. Cited in page 6.

JOHNSTON, R. L. Evolving better nanoparticles: Genetic algorithms for optimising cluster geometries. **Dalton Transactions**, Royal Society of Chemistry, n. 22, p. 4193–4207, 2003. Cited 11 times in pages 1, 2, 7, 8, 9, 11, 14, 15, 16, 17, and 18.

KAZAKOVA, V. A.; WU, A. S.; RAHMAN, T. S. Cluster energy optimizing genetic algorithm. In: ACM. **Proceedings of the 15th annual conference on Genetic and evolutionary computation**. [S.l.], 2013. p. 1317–1324. Cited 4 times in pages 2, 8, 18, and 23.

KOCH, W.; HOLTHAUSEN, M. C. **A chemist's guide to density functional theory**. [S.l.]: John Wiley & Sons, 2015. Cited 2 times in pages 6 and 9.

LIU, D. C.; NOCEDAL, J. On the limited memory bfgs method for large scale optimization. **Mathematical programming**, Springer, v. 45, n. 1-3, p. 503–528, 1989. Cited 2 times in pages 14 and 22.

MARKS, L. Experimental studies of small particle structures. **Reports on Progress in Physics**, IOP Publishing, v. 57, n. 6, p. 603, 1994. Cited in page 1.

MITCHELL, M. **An introduction to genetic algorithms**. [S.l.]: MIT press, 1998. Cited 4 times in pages 2, 9, 16, and 17.

NOCEDAL, J. Updating quasi-newton matrices with limited storage. **Mathematics of computation**, v. 35, n. 151, p. 773–782, 1980. Cited in page 14.

PAIVA, M. A. M. de; GALVÃO, B. R. Analysis of aluminium-magnesium alloy clusters through genetic algorithm. **Revista Processos Químicos**, v. 9, n. 18, p. 291–294, 2015. Cited in page 7.

PEREIRA, F. B. et al. Designing efficient evolutionary algorithms for cluster optimization: A study on locality. In: **Advances in Metaheuristics for Hard Optimization**. [S.l.]: Springer, 2007. p. 223–250. Cited in page 14.

PEREIRA, F. B.; MARQUES, J. M. Towards an effective evolutionary approach for binary lennard-jones clusters. In: IEEE. **Evolutionary Computation (CEC), 2010 IEEE Congress on**. [S.l.], 2010. p. 1–7. Cited in page 16.

SILVA, M. X.; GALVÃO, B. R.; BELCHIOR, J. C. Theoretical study of small sodium–potassium alloy clusters through genetic algorithm and quantum chemical calculations. **Physical Chemistry Chemical Physics**, Royal Society of Chemistry, v. 16, n. 19, p. 8895–8904, 2014. Cited in page 1.

XIAO, Y.; WILLIAMS, D. E. Genetic algorithm: a new approach to the prediction of the structure of molecular clusters. **Chemical Physics Letters**, Elsevier, v. 215, n. 1-3, p. 17–24, 1993. Cited 2 times in pages 10 and 12.

ZEIRI, Y. Prediction of the lowest energy structure of clusters using a genetic algorithm. **Physical Review E**, APS, v. 51, n. 4, p. R2769, 1995. Cited 2 times in pages 12 and 18.

Appendix

APPENDIX A – Shape descriptors

Shape descriptors have to be sensitive to highlight system properties as they are perceived by the objective function and at the level of detail used for the energy calculations. Arteca (1996) stated that it is essential to select an appropriate descriptor that addresses the problem at hand sustaining reasonable discriminating power — since the use of a descriptor carries a loss of information with respect to the original system.

Although small scale characteristics will influence the large scale, it will not be possible to tell small scale characteristics from large scale since the latter relies on averaged contributions from small scale. Extensively averaging local interactions could hide important information creating a evaluation problem. Therefore, it is necessary to adapt the descriptors to the size of each system, considering that the chosen descriptor could show important features of clusters at the chosen scale, but also inebriate the analysis in a different scale.

In this present work, estimators were used as shape descriptors for qualifying the cluster structure to bring light to the subtle characteristics of each generation method.

A.1 Characterizing the nuclear geometry

The simplest descriptor of a closed surface is the surface volume. The volume indicates size but can also bring (*e. g.*) insights about relative atomic spread. The natural instinct is to model the nuclear structure by a sphere that encloses all the atoms involved. Different structures can be, thus, visualized as fluctuations of the cluster spread.

A.1.1 Measurements of structure size

For the scope of this work, the cluster is centered on the global origin. A decision that enables the reduction of three position variables when this origin is chosen to be one of the atoms in the cluster. Any distance taken from the center of mass ($\vec{c}_m$), can also be represented by the subtraction:

$$d = ||\vec{c}_m - \vec{r}_i|| \tag{14}$$

The radius of the smallest sphere that encloses the atoms of the cluster, R_{min}, is calculated as half of **the furthest distance**, R_{max}, between two atoms in the cluster:

$$R_{max} = max\left(||\vec{r}_i - \vec{r}_j||\right) = 2R_{min} \tag{15}$$

A.1.2 Matrix of inertia, I

The matrix of inertia:

$$I = \begin{bmatrix} \sum m_i(y_i^2 + z_i^2) & -\sum m_i x_i y_i & -\sum m_i x_i z_i \\ -\sum m_i y_i x_i & \sum m_i(x_i^2 + z_i^2) & -\sum m_i y_i z_i \\ -\sum m_i z_i x_i & -\sum m_i z_i y_i & \sum m_i(x_i^2 + y_i^2) \end{bmatrix} \tag{16}$$

Where the values (x, y, z) are distances taken form the center of mass:

$$\vec{c}_m - \vec{r}_i = (x_i, y_i, z_i) \tag{17}$$

Can also be shown in terms of the its eigenvalues:

$$I = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \tag{18}$$

A.1.3 Molecular asphericity, ω

Arteca (1996) introduced the molecular asphericity, ω, as a shape descriptor based on the three moments of inertia, λ_i:

$$\omega = \frac{1}{2} \left\{ \sum_{i=1}^{2} \sum_{j=i+1}^{3} (\lambda_i - \lambda_j)^2 \right\} \left\{ \sum_{i=1}^{3} \lambda_i \right\}^{-2} \tag{19}$$

The closer the value gets to 0 the closer the structure shape is to a sphere. Since molecular dynamics are not considered here, the spherical interpretation of the static might seem vague. Although, even while considering only rigid body rotation, highly symmetrical would still be observed as close to spheres.

A.1.4 Mean atomic distance

This descriptor is presented as a reference for the atomic separation. Naturally, a more sparse cluster would have a larger mean distance between atoms. To the scope of this work, special attention was given to the standard deviation of the atomic distances, based on the assumption that different distributions indicate different structures. Higher fluctuations of the atomic distances would also indicate a deviation from a cubic or spherical shell structure towards a more linear structure.