Development and application of Machine Learning approaches to investigate the configurational space of binary alloys

Marios Gkanas

Departement of Physics

University of Crete

June 2024

Abstract

In this thesis, a novel methodology has been developed that allows for the identification and analysis of configurational patterns in alloys. This methodology explicitly incorporates the symmetry properties of the parent lattice and is based on unsupervised machine learning approaches. More specifically, unit cells of various sizes and symmetries are used to describe the configurations. These configurations are represented by vectors with lengths equal to the number of atoms enclosed by the unit cells.

To search for patterns and dominant configurations within the thousands of vectors, a self-consistent clustering algorithm has been developed. By applying this approach, cluster centers are constructed, and the representation vectors are assigned to these cluster centers. Moreover, a degree of order parameter is defined, allowing the assignment of the highest symmetry cluster center to each lattice site.

The aforementioned methodology is applied to pseudobinary InGaN alloys. As input, large alloy structures consisting of more than 10^5 atoms, produced by Monte Carlo calculations, are used. The results confirm the tendency of In atoms to align as second nearest neighbors in InGaN, leading to $\sqrt{3} \times \sqrt{3}$ translational symmetry. The outcome of the developed methodology, i.e., cluster centers, can be directly used in density functional theory calculations or to produce special quasirandom structures with modified probabilities for the occupation of the lattice sites.

Acknowledgements

I would like to express my deepest appreciation to L. Lymperakis, my thesis supervisor, whose guidance, support, and encouragement were invaluable throughout this research. He provided insightful feedback and constructive criticism that significantly contributed to shaping this thesis.

I am also grateful to N. Christakis, supervisor, for his expertise and assistance in the data analysis and machine learning area. His dedication and willingness to share knowledge were instrumental in overcoming various challenges encountered during the research process. Special thanks are due to E. Iliopoulos for his invaluable assistance.

To my friends and family, I owe an immense debt of gratitude. Their unwavering support, patience, and belief in me have been a constant source of strength throughout this journey. Whether through words of encouragement, understanding during stressful times, or simply being there to listen, their presence has made all the difference.

Lastly, I extend my appreciation to the staff and resources at the University of Crete for providing the necessary facilities and environment for conducting this research.

This thesis would not have been possible without the contributions of each individual mentioned above. Thank you all for your support.

ChatGPT has been used for English language polishing.

Contents

List of Figures

Chapter 1

Introduction

A necessary prerequisite for the design, growth, and synthesis of alloys is to gather insights into the thermodynamics of these materials and specifically derive and understand the corresponding phase diagrams. These diagrams provide crucial information regarding constraints and growth conditions such as temperature and partial pressures.

One of the most crucial factors underlying the construction of phase diagrams is the atomistic configurations present at the nanoscale, which simultaneously dictate the properties of alloys at the macroscopic level. These configurations can significantly impact the optoelectronic properties of semiconducting materials or the mechanical properties of metallic structural materials.

In Computational Materials Science, a workhorse in the study of these properties is methodologies that combine the accuracy of first principles calculations with techniques such as Cluster Expansion (CE) calculations and Monte Carlo simulations. In this approach, a CE Hamiltonian is trained and validated against Density Functional Theory (DFT) calculations. The CE Hamiltonian enables the efficient and accurate description of the energetics of numerous alloy systems comprising thousands or even millions of atoms. With an efficient computational Hamiltonian at hand to describe the energetics, Monte Carlo calculations are executed. These calculations, conducted in various ensembles such as canonical, grandcanonical,

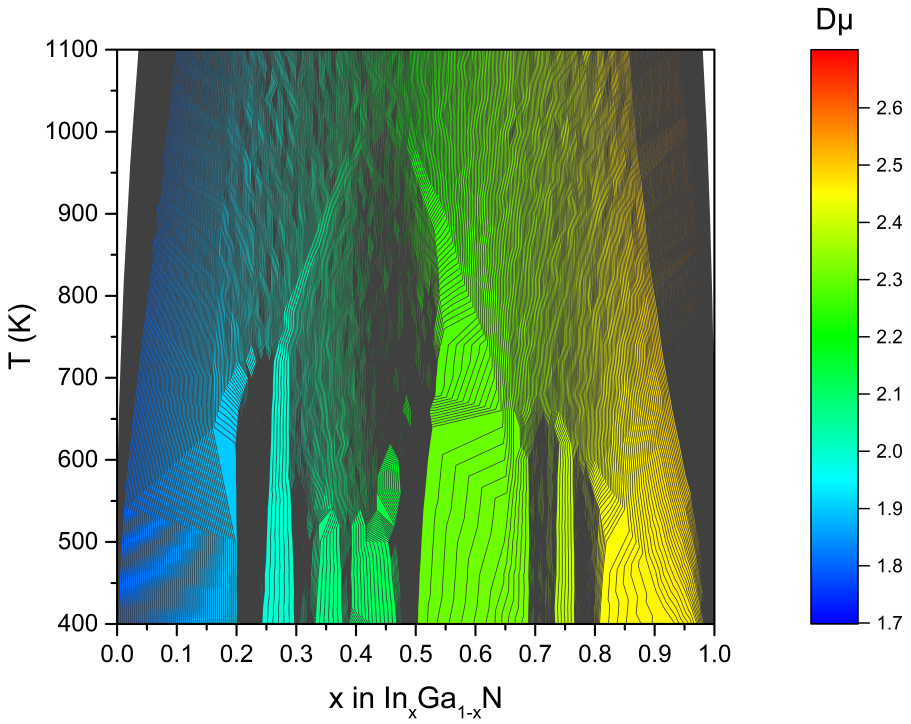

Figure 1.1: Phase diagram of InGaN pseudobinary alloys. The color code is the difference in the In and Ga chemical potentials: $\Delta\mu = \mu_{\mathrm{In}} - \mu_{\mathrm{Ga}}$ in eV. The phase diagram has been calculated my employing canonical MC calculations in a $40 \times 40 \times 40$ cell consisting of 128000 atoms [16]

etc., yield valuable information including chemical potentials, total energies, and heat capacities as functions of temperature. This wealth of data can be leveraged to derive phase diagrams and discern order/disorder transitions within the system (see Fig. 1.1).

A significant outcome of these calculations is the generation of atomic configurations at different temperatures and alloy compositions. These configurations provide invaluable insights into the structural evolution of the material under varying thermodynamic conditions, shedding light on the intricate interplay between composition, temperature, and order/disorder phenomena.

III-Nitride alloys hold significant importance in the realm of materials science and technology due to their diverse range of applications and unique properties. InGaN alloys are extensively utilized in **optoelectronic** devices such as light-emitting di-

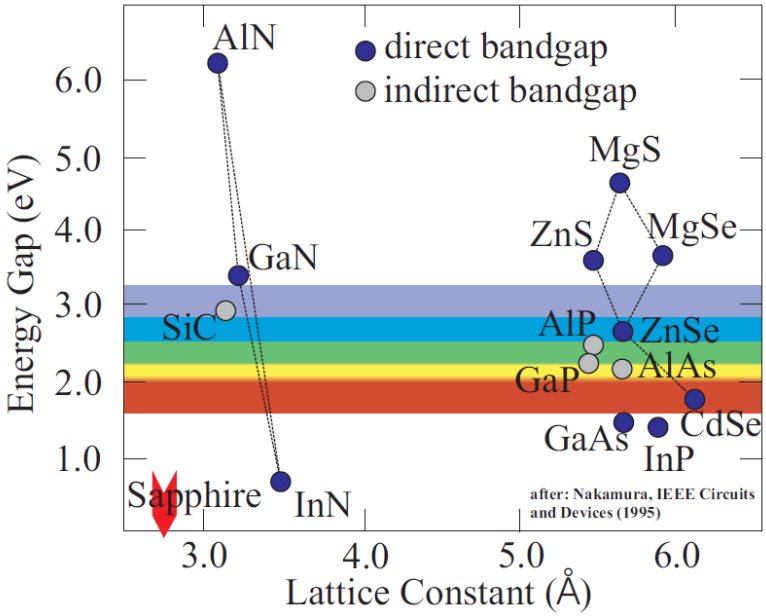

Figure 1.2: Bandgap energy as a function of the lattice constant for various technologically important semiconductors [18]. The value of the InN band gap has been corrected to 0.7 eV (instead of the value of 2 eV used in the original figure).

odes (LEDs), laser diodes, and photovoltaic cells. Their tunable bandgap spanning the visible spectrum makes them versatile materials for producing efficient light emission across a wide range of wavelengths (see Fig. 1.2). Understanding the **thermodynamics of InGaN alloys**, specifically phase separation and order-disorder transitions, is crucial for optimizing device performance and enhancing their efficiency.

As can be deduced from Fig. 1.2, to access the green region of the spectrum, InGaN films with In content as high as ≈30% are required. However, the growth of high-quality and high In content InGaN films is challenging: In and Ga atoms have very different atomic radii (136 pm for GaN vs 156 pm for InN [3]). Moreover, the bond strength In-N and Ga-N bonds are very different: The cohesive energy of InN has been calculated by DFT-GGA calculations to be equal to 7.695 eV and that of GaN 9.265 eV [21]. These disparity in the properties of the end constituents, i.e., InN and GaN, has been suggested to result in spinodal decomposition and phase separation (see Ref. [19] and references therein).

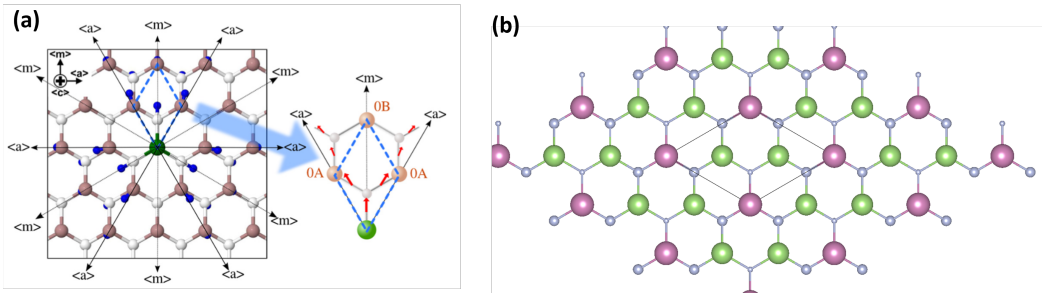

Figure 1.3: (a) Left: Displacement of Ga (denoted by brown balls) and N (denoted by white balls) atoms in the (0001) plane of GaN, around an In atom ((denoted by green balls). Small blue spheres indicate the relaxed position of Ga and N atoms after. Right: Schematic representation of the displacement channeling mechanism that allows for efficient strain accommodation. The figure has been adopted from Ref. [[15]]. (b) Ball and stick model the $\sqrt{3} \times \sqrt{3}$ structure of the $In_{1/3}Ga_{2/3}N$. Big brown and green balls denote In and Ga atoms, respectively. Small gray balls are N atoms.

Spinodal decomposition is a phenomenon governed by bulk diffusion limitations, commonly observed in binary or pseudobinary alloys. This occurs when the Gibbs free energy associated with mixing, exhibits partial convexity, presenting dual minima. In regions where the second derivative of the Gibbs free energy concerning composition is negative, the uniform alloy structure becomes susceptible to composition fluctuations, thus resulting in phase segregation. Various theoretical investigations have investigated the bulk thermodynamics of InGaN alloys, unveiling a substantial miscibility gap [2, 8].

All the above mentioned calculations considered incoherent growth for InGaN. This entails an assumption that, for any given composition, the InGaN alloy is fully rexaled from the strain induced by the lattice mismatch between the substrate and the epilayer. However, it has demonstrated that when considering coherent growth, wherein InGaN is biaxially strained to GaN, the width of the miscibility gap diminishes significantly and/or shifts towards the In-rich region of the phase diagram [5, 12, 20].

1.1 Phase diagram of InGaN alloys

More recently it has been shown that at $In_{1/3}Ga_{2/3}N$ alloys ordering can be induced at the surface during N-rich Molecular Beam Epitaxy (MBE) growth [16]. By combining DFT calculate surface calculations, MBE and High Resolution Transmission Electron Microsopy (HR-TEM) experiments it was demonstrated that a $\sqrt{3} \times \sqrt{3}$ reconstruction (see Fig. 1.3(b)) for In content 33% is energetically favorable at the surface and thermodynamically stable at temperatures as high as 950 K. In this ordered InGaN structurethe In atoms are spatially distributed as 2nd nearest neighbors aligned along the $\langle 1\bar{1}00 \rangle$ direction. The origin of this structure is the interplay between two mechanisms: (a) Efficient strain accommodation (see Fig. 1.3(a)) [15] and (b) a novel reconstruction mechanism, elastically frustrated rehybridization.

In bulk InGaN alloys the surface reconstruction mechanism is not present. However, efficient strain accomodation is relevant. Indeed it has been shown that the above mentioned ordered structure is favorable [15]. Nevertheless, MC calculations revealed that in bulk and at $x_{In} = \frac{1}{3}$ an order-disorder transition will occur at a $\approx$200 K lower temperature (see Fig. 1.1). The phase diagram of bulk InGaN biaxially strained to GaN in Fig. 1.1 reveals that phase separation towards the end constituents is suppressed. However, it also reveals the presence of miscibility gaps (areas where equipotential trajectories are not present) at low temperatures.

1.2 Motivation

Monte Carlo simulations offer a vast array of insights into alloys, enabling thorough exploration, comprehension, and even design of their properties. A prime illustration is the phase diagram of InGaN, which elucidates the thermodynamic intricacies of such alloys. Additionally, these simulations yield invaluable insights into atomic arrangements, crucial for understanding the electronic properties of compound semiconductors. Notably, the spatial distribution of alloy constituents

profoundly impacts these properties, as evidenced by the redshift of the bandgap observed due to ordering in AlGaN alloys [1, 11].

Atomic geometries obtained through Monte Carlo (MC) calculations offer a means to derive the Warren–Cowley ordering parameters [4]. While these parameters can offer insights into alloy ordering and can aid in exploring order-disorder transitions, they do not directly quantify the ratio of ordered structures, their spatial distribution, or their translational and rotational symmetry. Such details are essential prerequisites for density functional theory (DFT) calculations, particularly when utilizing codes with periodic boundary conditions.

To address the above mentioned, in the present thesis a methodology based on Machine Learning clustering algorithms is implemented to investigate the structural properties of binary and pseudobinary alloys obtained by MC calculations.

1.3 Crystal Structure of InGaN

In the realm of solid-state physics and materials science, the understanding of crystal structures and their properties plays a fundamental role. The Bravais lattice, named after Auguste Bravais (1850), is a fundamental concept in crystallography. A Bravais lattice is a mathematical concept used in crystallography to describe the periodic arrangement of points (atoms, ions, or molecules) in a crystal structure. The Bravais lattice describes the periodicity and symmetry of the crystal lattice. There are 14 Bravais lattices in three dimensions which are grouped in seven lattice systems. These are the following:

- Triclinic.
- Monoclinic.
- Orthorhombic.
- Tetragonal.
- Rhombohedral.
- Hexagonal.
- Cubic.

Essential concepts in describing the Bravais lattices are the primitive and unit cells. Unit cells constitute the basic building blocks from which larger cells or even the entire crystal structure through translational symmetry operations. Primitive cells are the smallest possible unit cells. Primitive cells are parallelepiped and are defined by three vectors called primitive vectors which are also translational vectors of the crystal.

In topic of the present thesis is the development and application of ML approaches to investigate the configurational space of the pseudobinary InGaN alloys. As

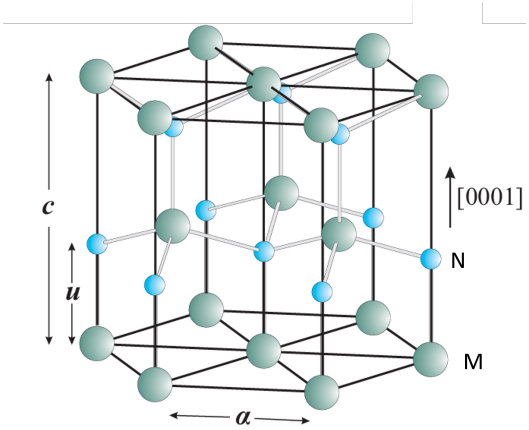

Figure 1.4: Schematic representation in ball and stick model of the wz structure. N denotes that Nitrogen atoms and M the group III atoms.

has already been mentioned in the introduction, these alloys belong to the family of group III-Nitrides which include InN, GaN, and AlN as well as their alloys. The thermodynamically most favorable structure of these alloys is the hexagonal wurtzite structure (Space group: P63mc). This structure can be described by two sublattices, the metal and the Nitrogen sublattice. The primitive vectors that describe the translational symmetry are the following:

$$\begin{aligned} \boldsymbol{a}_1 &= \left[\tfrac{1}{2}a, -\tfrac{\sqrt{3}}{2}a, 0\right] \\ \boldsymbol{a}_2 &= \left[\tfrac{1}{2}a, \tfrac{\sqrt{3}}{2}a, 0\right] \\ \boldsymbol{a}_3 &= [0, 0, c] \end{aligned} \tag{1.1}$$

a and c are the lattice constants. In the ideal wurtzite crystal $c = \sqrt{\frac{8}{3}}a$.

Crystal points that belong to the metal sublattice are at positions (in direct or reduced coordinates):

$$\begin{aligned} \boldsymbol{b}_1 &= [0, 0, 0] \\ \boldsymbol{b}_2 &= \left[\tfrac{2}{3}, \tfrac{1}{3}, \tfrac{1}{2}\right] \end{aligned} \tag{1.2}$$

and the Nitrogen atoms at:

$$\begin{aligned} \boldsymbol{b}_3 &= [0, 0, u] \\ \boldsymbol{b}_4 &= \left[\tfrac{2}{3}, \tfrac{1}{2}, u + 0.5\right] \end{aligned} \tag{1.3}$$

Here, u is an internal lattice parameter which describes the shift of the one sublattice with respect to the other. In the ideal wurtzite structure $u = \frac{3}{8}$. Vectors $\boldsymbol{b}_i$ are called basis vectors. The wurtzite structure is shown in Fig. 1.4.

Antisite point defects denoted as M_N and N_M, (M stands for In, Ga, or Al) are defects where metal atoms sit at N sites and vice-versa and have high formation energies. Therefore, their concentrations are typically very small. Hence, in the present study, we treat the InGaN system as a pseudobinary In_xGa_{1-x} alloy with hexagonal symmetry, i.e., the primitive vectors are the same but we use only the $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ basis vectors.

As it becomes clear from the discussion above, the primitive cell of our system is described by the primitive vectors in Eq. 1.1, and the two basis vectors in Eq. 1.2. The primitive vectors can be used to define larger translational vectors of the crystal:

$$\boldsymbol{R} = n_1\boldsymbol{a}_1 + n_2\boldsymbol{a}_2 + n_3\boldsymbol{a}_3 \tag{1.4}$$

where $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, and $\boldsymbol{a}_3$ are the primitive vectors.

Moreover, the volume of the primitive cell is given by the following equation:

$$V = |\boldsymbol{a}_1 \cdot (\boldsymbol{a}_2 \times \boldsymbol{a}_3)| \tag{1.5}$$

We should also note that the primitive cell contains exactly the two atoms defined by the two basis vectors. However, due to the crystal translational symmetry, lattice points exist at the eight vertices of the unit cell. These points are considered to be shared with neighboring cells. Hence each cell is considered to contain $\frac{1}{8}$ of each site at the eight vertices. This argument will be used extensively in our code in order to achieve the minimum representation of our structures.

At this point, it is important to make a distinction between the usage of the terms *"unit"* and *"primitive"* cells in the present thesis. As primitive cell, we refer to the cell described by Eqs. (1.1) and (1.5) above. In our study, the different structures will be described by cells that are equal or larger than the primitive and may contain more than two basis atoms. We will refer to these larger cells as unit cells and keep the term "primitive cell" for the primitive cell of the wurtzite crystal, which represents the smallest repeating unit within a crystal lattice.

In the present methodology we employ unit cells to invesigate the configurational space of the pseudobinary alloys. More specifically we will search for and apply clustering to configurations that are contained in unit cells. More details are given in the next Section (Sec. 2).

Chapter 2

Datasets

To address the challenge of identifying highly symmetric ordered structures within an alloy, we developed a specialized approach. This approach efficiently shifts through millions of diverse configurations to pinpoint structures of particular interest. Central to our efforts is the necessity for a robust method capable of identifying patterns amidst this extensive array of configurations, with a primary focus on configurations exhibiting high symmetry.

In our developed approach, we use unit cells as fundamental descriptors. These unit cells are defined by translation vectors, establishing translational symmetry, as well as basis vectors representing lattice points within the unit cell and their occupation, such as Ga or In atoms. This systematic approach facilitates the categorization and analysis of the multitude of configurations present in the alloy. Efficiency is a critical factor underlying our methodology. With millions of configurations to sort through, computational efficiency is non-negotiable. Our algorithm is meticulously designed to optimize computational resources.

Another challenge we face is that our approach needs to be robust, i.e., the ability to uniquely identify structures of interest. In order to achieve this, the underlying rotational and translational symmetries have to be explicitly incorporated and utilized in our methodology and algorithm. It is paramount to ensure that symmetry equivalent configurations are not erroneously treated as different structures.

2.1 Input Structures

The atomic structures provided as input are the results of Cluster Expansion Monte Carlo calculations for $In_xGa_{1-x}N$ pseudobinary alloys, conducted across temperatures ranging from 400 K to 2000 K. These structures are represented by $40 \times 40 \times 40$ cells, each containing $40 \times 40 \times 40 \times 2$ atoms. Ga and In atoms are represented by 0 and 1, respectively.

Our approach involves exploring the configurations in the aforementioned cells. These cells are defined by three translational vectors. These are a linear combinations of the wurtzite primitive cell's translational vectors (the primitive cell being the smallest group of atoms with the overall symmetry of a crystal, as explained in Section 1.3). We also prioritize unit cells that contain structures with low mixing enthalpy, as determined by a DFT calculations as well as by CE Hamiltonian (see Introduction). The mixing enthalpies of these structures are at most 5/;meV per cation above the convex hull line (see Fig. 2.1). Eventually we include 65 symmetry inequivalent unit cells.

Before we discuss the main part of the code, let us describe how we define and construct the different structures and how we represent them with vectors. As has already been mentioned, the translation and basis vectors define the shape and size of the unit cell as well as the lattice points contained by the unit cell. The lattice points within and at the boundaries of this cell are occupied by In or Ga atoms.

In our approach we go through all lattice points and for each of the aforementioned 65 unitcells we construct a vector. This vector contains the occupations Ga (0) or In (1) of the lattice points of the unit cell. The dimension of the vector is equal to the number of lattice points. There are numerous ways to map the lattice points to the vector components/indices. The mapping between the lattice points and the vector components is done using the coordinates of the lattice points: The lattice points are sorted using their direct coordinate with respect to the local

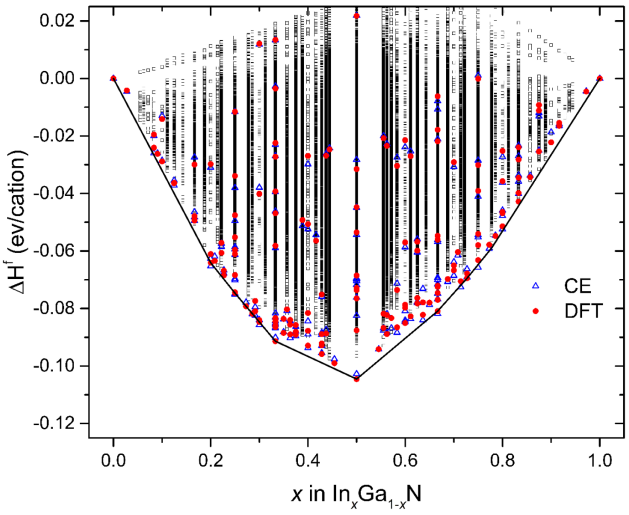

Figure 2.1: Calculated mixing enthalpies of various configurations of InxGa$_{1-x}$N alloys. Blue triangles and red dots indicate enthalpuies cal;culated with a CE Hamiltonian and DFT calculations. Taken from Ref. [].

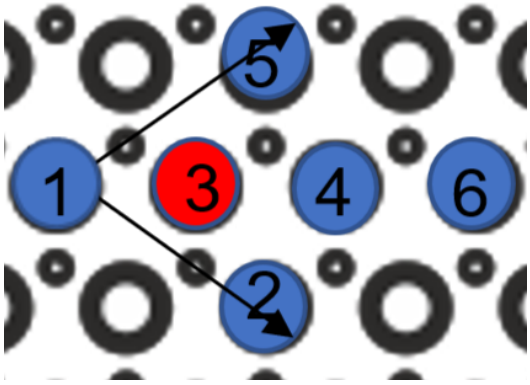

Figure 2.2: Schematic representation of a 2D unit cell with a $\sqrt{3} \times \sqrt{3}$ translational symmetry. Blue and red points denote Ga and In atoms, respectively. The vector representation of this structure is $[0, 0, 1, 0, 0, 0]$.

frame of reference: First by their z-coordinate, then by their y- and finally by their x-coordinate. This ordering of the components allows for a global and unique description of the structures with vectors.

2.2 Vector description of the configurations.

Figure 2.2 depicts a selected representative 2D cell wherein lattice points are occupied by Ga (0) or In (1) atoms. The corresponding vector serves as a unique descriptor of the structure: The first component of the vector corresponds to the lattice point located at the origin of the local frame of reference defined by the translational vectors. This lattice site, occupied by a Ga atom, is designated with a value of 0. The second component of the vector corresponds to the lattice point at the lower edge of the cell, assigned a value of 1, indicating occupancy by an In atom.

2.3 Rotational Symmetry

To fully incorporate the rotational symmetry of the parent crystal structure, at each lattice point, we construct representation vectors. We do this by rotating the unit cell 0, 120, and 240 degrees. A schematic representation of this is shown in Fig. 2.3. This process yields three representatiion vectors for each lattice point in our input structure. As a result, we obtain a total of $128,000 \times 3$ representation vectors.

2.4 Pseudo-periodic unit cells.

As has already been mentioned, in our approach, we employ unit cells to construct the structures and the corresponding vector represention. Nevertheless, the unit cells in computational materials exhibit translational symmetry, implying periodicity and hence order within. However, it's important to note that not all generated unit cells are inherently periodic. This occurs when atoms positioned at the edges or opposite sides of the cell differ. To address this discrepancy, we adopt a strategy to enforce periodicity. For atoms within the unit cell connected by translational

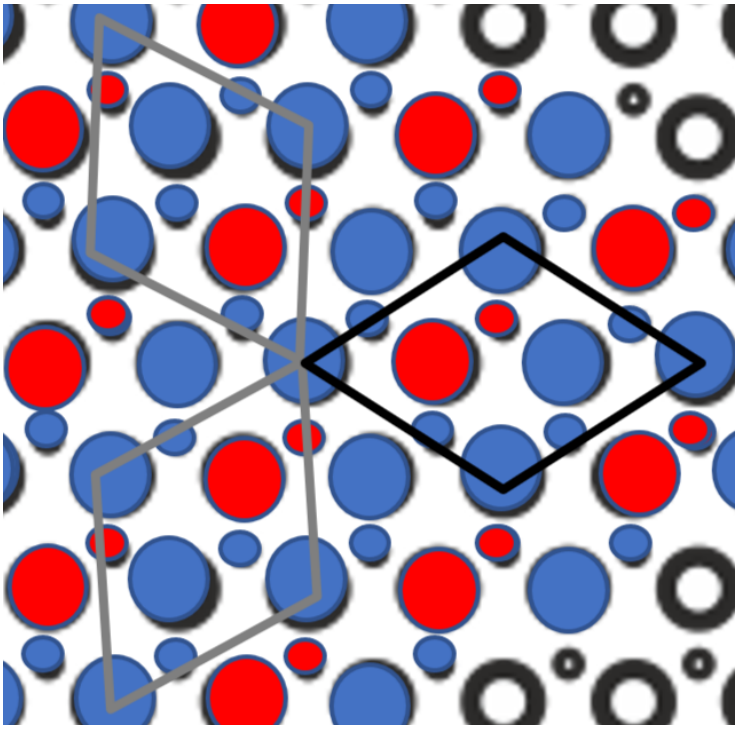

Figure 2.3: Schematic representation of a 2D unit cell with a $\sqrt{3} \times \sqrt{3}$ translational symmetry and its vector representation. The color code is as in Fig. 2.2. The three diamonds indicate the theree unit cells produced by 0 (black) and 120 and 240° (gray) rotation.

vectors, i.e. they are translationally symmetry equivalent, we assign an average occupation. For instance, at the 8 edges where 6 atoms are Indium (0) and 2 are Ga (1), we assign a value of 0.25 to these sites. This approach renders nonperiodic cells pseudo-periodic.

Pseudo-periodic unit cells enable the application of a minimal representation of the structure. Instead of necessitating indices for all atoms within the cell, including those at boundaries and edges, we only consider the indices of atoms incapable of being connected via translational vectors. This approach enhances the efficiency of our approach significantly. Moreover, the concept of partial occupation within these cells serves as a key indicator of disorder. Drawing inspiration from entropy in statistical physics, we proceed to define a descriptor of order within the cell (see below). By characterizing the extent of partial occupation and the arrangement of atoms within the unit cell, we can quantify the level of disorder present, providing

valuable insights into the overall structural properties of the alloy system.

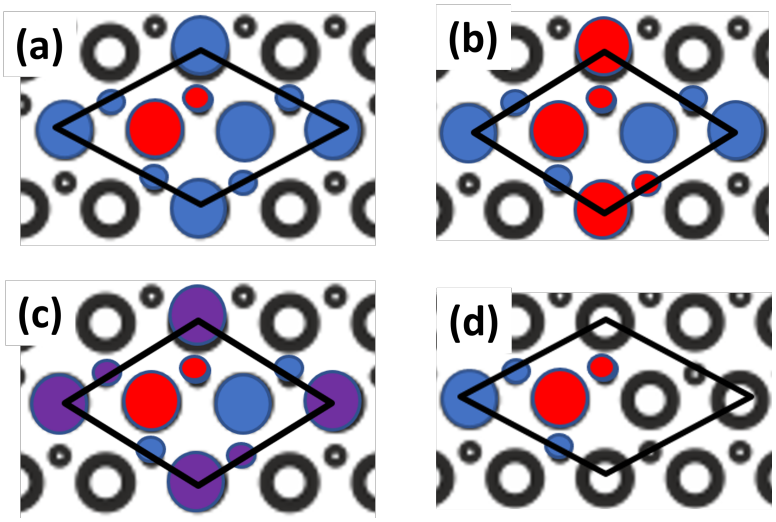

Figure 2.4: (a) Schematic representation of a unit cell that obeys periodic boundary conditions, i.e., translational. All translational symmetry equivalent lattice sites are occupied by the same species (In or Ga). (b) The same unit cell but the four symmetry equivalent sites at the four edges of the cell are occupied by both In and Ga atoms. This cell disobeys periodic boundary conditions. (c) Pseudoperiodic unit cell constructed from the cell in (b). A partial average occupation of 0.5 is assignes to lattice sites at the corners of the cell. (d) The unit cell in (a) in the minimum representation. Since periodic boundary conditions are justified, all corner/edge atoms but one, are not included in the representation.

2.5 Downsampling

For each input unit cell, i.e., for each temperature and content, datasets consisting of 128000×3 representation vectors are constructed. This renders the application of clustering methodologies computationally cumbersome. In order to address this, we apply downsampling to our datasets. This entails condensing a smaller dataset's size while retaining its critical features. In this context we select from our datasets atoms that are homogeneously distributed in the volume of the input cells. The aim is to construct a representative set of atoms that preserve the essential characteristics of the whole input lattice while significantly reducing the dataset's size.

To execute downsampling we select atoms with a step equal to 5 along each dimen-

sion. We have carefully checked if this approach provides a consistent description of the properties of the whole system. Our check indicate that (to be updated at the end)... By downsampling with a step equal to 5 along each dimension, the resulting downscaled dataset has dimensions of $\frac{128000\times3}{125}$. This downscaled dataset serves as a condensed yet representative sample of the original lattice, facilitating more efficient data processing and analysis. The procedure of downsampling is schematically shown in Fig. 2.5

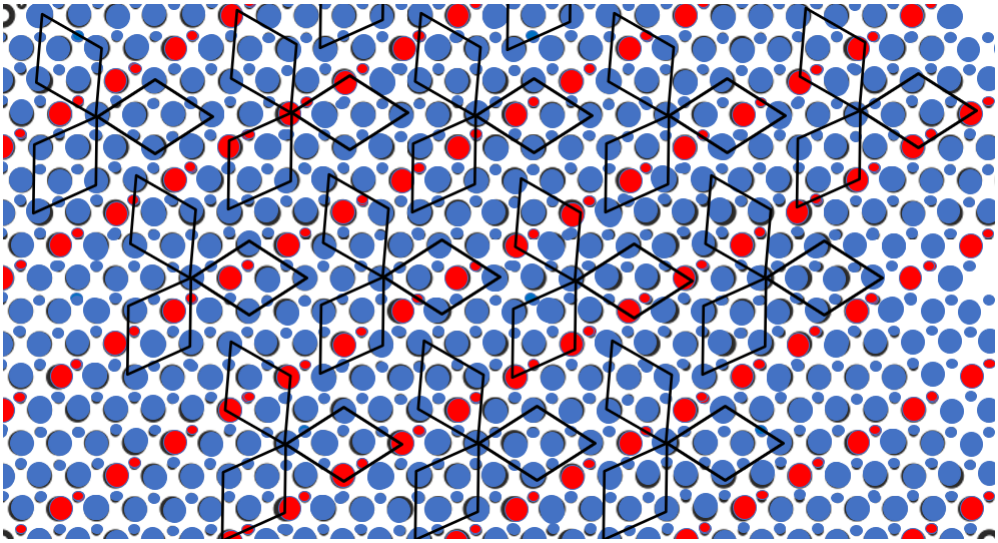

Figure 2.5: Schematic representation of an ordered structure. Blue and red balls indicate Ga and In atoms, respectively. The representation vectors are taken at the sparse/downsampled matrix. These points are located at the origin of the unit cells represented by the diamonds. The rotated unit cells are also shown. As can be seen, despite the downsampling of the input data, the sparse matrix includes the majority of lattice points and hence can provide a good representation of all different configurations.

2.6 Translational Symmetry

Within the set of representation vectors, there's a possibility of encountering vectors that could be erroneously perceived as symmetrically inequivalent, indicated by the norm of their differences not being equal to zero. This discrepancy can arise when translational symmetry-equivalent cells exhibit varying arrangements of indices. In essence, despite the cells being translationally equivalent, differences

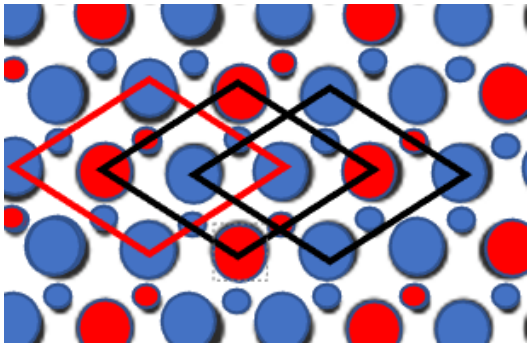

Figure 2.6: Schematic representation of an ordered configuration and a $\sqrt{3} \times \sqrt{3}$ unit cell placed at three different origins.

in the atomic arrangement can lead to variations in the representation vectors, potentially resulting in misinterpretations of symmetry.

To make this clear, in Fig. 2.6 we show an ordered configuration and a $\sqrt{3}\times\sqrt{3}$ unit cells placed at three different origins. Due to the long range ordering the structures contained by the three unit cells shiuld be the same. The representation vectors of these are: $v_1 = [1,0,1,1,0,1]$, $v_2 = [0,1,1,0,1,1]$, and $v_3 = [1,1,0,1,1,0]$. However, the distance between any of these two is $|v_i - v_j| = 2$, $i \neq j$. Therefore, these vector representations will erroneously be treated as different configuration in a clustering algorithm.

Failure to address this issue could lead to complications in the clustering algorithm, potentially resulting in the erroneous assignment of vectors to incorrect cluster centers. Moreover, it might generate cluster centers that do not accurately represent the underlying data. Inaccuracies stemming from the misinterpretation of symmetrically inequivalent vectors could distort the clustering process, undermining the algorithm's ability to effectively identify meaningful patterns and structures within the dataset. Therefore, addressing this concern is crucial to ensure the reliability and validity of the clustering results.

To address this issue, we implement a procedure where we iterate through all representation vectors. At each vector, we apply translational symmetry operations, corresponding to rearrangements of the vectors components. This process allows us

to identify all symmetry-equivalent representations at each lattice point. Among these equivalents, we selectively retain only the representation vector that matches an existing representation vector at another lattice point in the dataset. If none of the equivalents match any existing representation vector, we retain one of them arbitrarily. By doing so, we ensure that all symmetry-equivalent vectors share the same representation vector. This method effectively prevents discrepancies that could adversely affect the clustering algorithm's.

2.7 Summary

In this chapter the preperation of the our datasets have been described in details. Our input datasets are alloys configurations consisting of 124000 atoms. Employing 65 different unit cells, considering rotation and translational symmetries, and downsampling our data, we end with 1024×3 representation vectors for each of the 65 unit cells. The aim to (1) identify representative configurations for each unit cell and (2) for each lattice point to asign a certain unit cell and configuration. To achieve this we develop and apply a ML clustering approach which is described in the nect Chapter.

Chapter 3

Unsupervised Learning

Unsupervised learning (UL) stands out in the realm of machine learning (ML) as an approach that effectively eliminates human bias from the analysis, as acknowledged in numerous studies [10] [13] [9] [1,2,3]. Going a step further, it can be argued that UL not only addresses human bias but also plays a pivotal role in mitigating computational bias and reducing uncertainty associated with numerical algorithms and processes. In UL, computational groups are determined by the algorithm itself rather than predefined by the researcher, making it particularly well-suited for tackling problems of higher complexity when identifying groups.

Furthermore, UL proves advantageous in scenarios where obtaining unlabelled data—whether experimental, computational, or derived from field measurements—is more feasible than acquiring labelled data requiring user intervention. The absence of pre-defined labels in UL, though posing challenges, makes it a valuable approach for navigating problems that lack readily available labels. UL represents a foundational machine learning approach that has endured over time and remains crucial in a number of diverse research and application domains, including image processing, sleep stages classification, and mechanical damage detection.

3.1 Existing Challenges

UL algorithms, while powerful and versatile, do face significant challenges when it comes to particle clustering tasks:

- **Ambiguous cluster boundaries:** Defining clear boundaries for complex spatial distributions or overlapping clusters poses challenges, impacting the accurate separation and identification of distinct particle clusters.
- **Sensitivity to hyperparameters:** Selecting optimal hyperparameter values, such as the number of clusters, can be difficult and may lead to varied results, requiring careful parameter tuning.
- **Dimensionality and feature selection:** Handling high-dimensional particle data and selecting meaningful features are challenges, with poor choices impacting clustering performance.
- **Cluster shape and size variability:** Assumptions of predefined cluster shapes may not align with irregular shapes and varying sizes in particle clusters, affecting accurate representation.
- **Robustness to noise:** Sensitivity to noise can result in spurious clusters or inaccuracies in clustering results due to measurement errors or outliers.
- **Limited supervision:** Unsupervised algorithms lack external guidance, potentially missing valuable information that could enhance clustering accuracy.
- **Scalability:** Some algorithms become computationally expensive with larger datasets, posing challenges in efficiently handling large-scale particle data due to memory and processing limitations.

To address the above mentioned optimization problems we apply the RUN-ICON algorithm which is based around the K-Means clustering method.

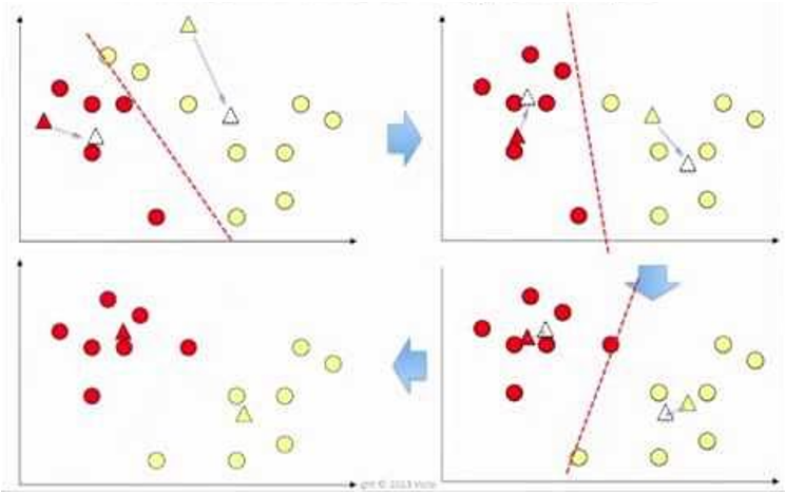

Figure 3.1: Schematic representation of K-Means clustering algorithm. Taken from https://www.e2matrix.com/blog/2018/01/01/kmeans-clustering-with-example/.

3.2 Clustering

Clustering is a method used to group similar items together. It works by organizing data into clusters, where each cluster contains items that are more alike to each other than to those in other clusters. In this project, clustering was utilized to analyze the 3×1024 representation vectors from the simulation. This allows us to identify groups of vectors, and hence of configurations, that share common characteristics, revealing patterns in the data.

K-Means is a widely used algorithm for clustering data into groups. The goal is to divide a set of data points into a specific number, k, of clusters. The process begins by randomly selecting k initial points from the dataset, which will serve as the centers or centroids of the clusters.

Once the centroids are chosen, each data point in the dataset is assigned to the nearest centroid. This step forms k clusters based on the proximity of the data points to the centroids. After all the points are assigned, the algorithm calculates new centroids by finding the average position of all the points in each cluster.

These steps of assigning data points to the nearest centroid and updating the centroids are repeated until the centroids no longer change significantly, or the

algorithm reaches a maximum number of iterations. The final result is k clusters with data points grouped around each centroid. This shows how the data naturally forms groups.

K-Means is a simple and efficient method for clustering, making it a popular choice for many applications. It helps to uncover patterns and structures in data, which can be useful for further analysis and decision-making. In my research, I use K-Means to analyze the $3\times$ 1024 representation vectors from the simulation, helping to identify meaningful patterns within the configurational space.

3.3 The RUN ICON algorithm

In this project, the RUN-ICON UL algorithm was employed. The primary objective of RUN-ICON [17] is to overcome various challenges associated with cluster selection, ensuring a high level of confidence and low uncertainty. Unlike traditional methods that rely on intuitive criteria for determining the optimal number of clusters, RUN-ICON employs a unique approach. It identifies the optimal number of clusters by consistently identifying dominant centers across multiple repetitions of the K-means algorithm. Instead of relying on the Sum of Squared Errors, the algorithm introduces innovative metrics such as the Clustering Dominance Index (CDI) and Uncertainty.

The CDI is associated with the frequency of a specific clustering configuration occurring when splitting the dataset into a certain number of clusters. It can be interpreted as the probability of that particular configuration occurring. On the other hand, uncertainty represents the relative difference between upper and lower CDI bounds for a clustering configuration, indicating the maximum variance from the mean for that configuration.

The RUN-ICON algorithm enhances the traditional K-Means clustering method by repeating the process multiple times with different numbers of clusters. For each iteration, it performs K-Means clustering, varying the number of clusters

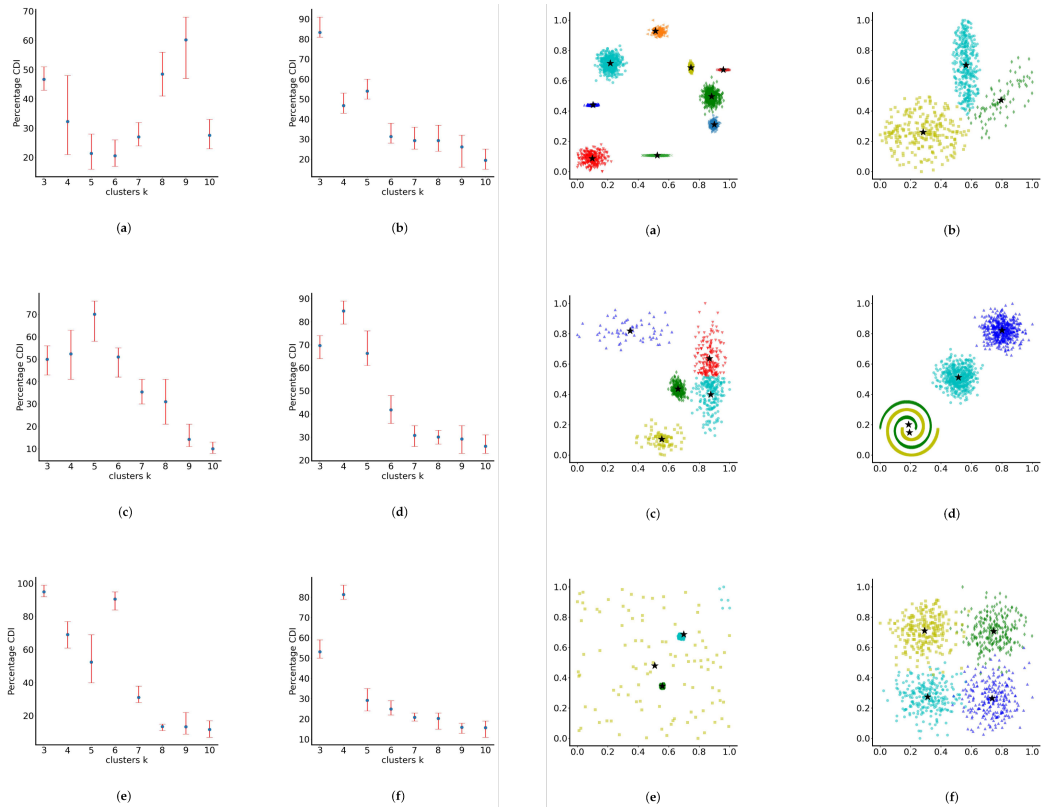

Figure 3.2: Left Panel: Averaged percentage Cluster Dominance Index (i.e., frequency of occurrence of a specific clustering configuration when requesting a particular number of clusters) for different clustering requirements ranging from 3 to 10 clusters. Right Panel: Optimal clusters generated by the RUN-ICON algorithm for the data sets tested in the left panel. Different colours correspond to different clusters. The black stars correspond to the cluster centres. Ref [17]. Right:

to explore different clustering configurations. This repetition helps in identifying the most dominant cluster centers across different runs, reducing uncertainty and increasing the confidence in the clustering results.

By analyzing the outcomes of these repeated runs, the RUN-ICON algorithm calculates the probability of occurrence for each cluster center. This means it determines how often certain cluster centers appear as the dominant ones in various clustering configurations. The result is a more robust and reliable clustering, as it highlights the most consistent and significant patterns in the data, providing a clearer understanding of the underlying structure. Clustering results from the RUN-ICON algorithm are presented in Fig. 3.2, where the ability of the algorithm to correctly

identify dominant clusters is demonstrated through the use of a number of test datasets.

The efficiency of RUN-ICON has been demonstrated in recent studies, showcasing its superiority over various UL algorithms, such as Repeat K-means [7], which iteratively refines cluster centers for improved stability in the presence of noisy data, Bayesian K-means [14], incorporating Bayesian inference principles for a more flexible and probabilistic clustering approach, and DBSCAN [6], a density-based algorithm capable of identifying arbitrarily shaped clusters and noise points.

3.4 Self Consistent Clustering

As has already been mentioned, clustering is a fundamental unsupervised learning technique used to identify inherent structures within a dataset by grouping similar data points together. In the context of atomic configurations represented as vectors, clustering serves to organize atoms into coherent groups based on their spatial arrangements. However, traditional clustering algorithms may face challenges when dealing with complex systems exhibiting translational symmetry.

The essence of clustering lies in partitioning the dataset into subsets, or clusters, such that data points within the same cluster share similarities, while those in different clusters are dissimilar. This is typically achieved by optimizing an objective function that quantifies the compactness of clusters and the separation between them.

In clustering, we optimize the norm of the vectors. However, here we face the same problem as we had with the construction of the representation vectors and the effect of translational symmetry (see Section 2.6). More specifically, due to translational symmetry, some vectors may erroneously be assigned to more distant cluster centers. To correct for this issue, we developed and applied an iterative self-consistent approach. After determining the optimal number of clusters using RUN-ICON algorithm, the clustering process is refined iteratively. This involves applying

translational symmetry operators to the representation vectors and assessing their similarity to existing cluster centers. If a transformed configuration is closer to a cluster center other than the original, the vector is updated by applying this translational and/or rotational operators. Once all vectors have been tested and, where neccesary, updated, RUN-ICON algorith is employed again. This iterative refinement continues until no further updates can be made. This ensures that each vector is assigned to the cluster that best represents its configuration.

3.5 Order descriptor.

At the end of the self-consistent clustering process, each lattice point is assigned to 65 different cluster centers (one for each unit cell). In order to uniquely assighn a single cluster center to each lattice point we define an order descriptor, termed "entropy" . The equation for the “entropy” is as follows:

$$S_{CC} = \sum_{i=1}^{n} \left(x_i^{-x_i} \right) \cdot \left((1 - x_i)^{-(1-x_i)} \right) \tag{3.1}$$

where the index i runs over all the sites of cluster center and x_i is the occupation of the site i.

The cluster with the smallest ”entropy” S_{CC} is chosen, i.e., “more ordered” and higher symmetry cluster centers are chosen. In cases where more than one cluster center exhibit the same value of S_{CC}, priority is given to the center with the smaller volume. Ultimately, for each lattice point, one cluster center is assigned.

Chapter 4

Results and Discussion

In this chapter, we present the results of our investigation into the behavior of cluster centers across various temperatures Tand compositions x. Applying the methodology described in the previous Chapter we investigated the different alloy configurations for $x = 0.2$, 0.25 , 0.30, 0.35, and 0.40 and T in the range from 400 to 1000 K. This methodology provided for each lattice site the most representative cluster center.

4.1 Order-Disorder

In a first step we estimate the degree of order at different temperatures and compositions. To achieve this for each T and x we define the a degree of diosder $S(T, x)$ as:

$$S(T, x) = \frac{1}{N} \sum_{i=1}^{N} S_i(T, x) \ , \tag{4.1}$$

where $S_i(T, x)$ is the degree of order of the cluster assigned to the lattice site i at T and x (see Eq. 3.1). N is the number of lattice sites. Smaller/larger values of $S(T, x)$ indicate higher/lower degree of ordering in our system.

In Fig. 4.1 the degree of order is plotted against temperature for different selected compositions. As expected the general trend is that the degree of order is reduced as the temperature increases, since configurational entropic contributions dominate

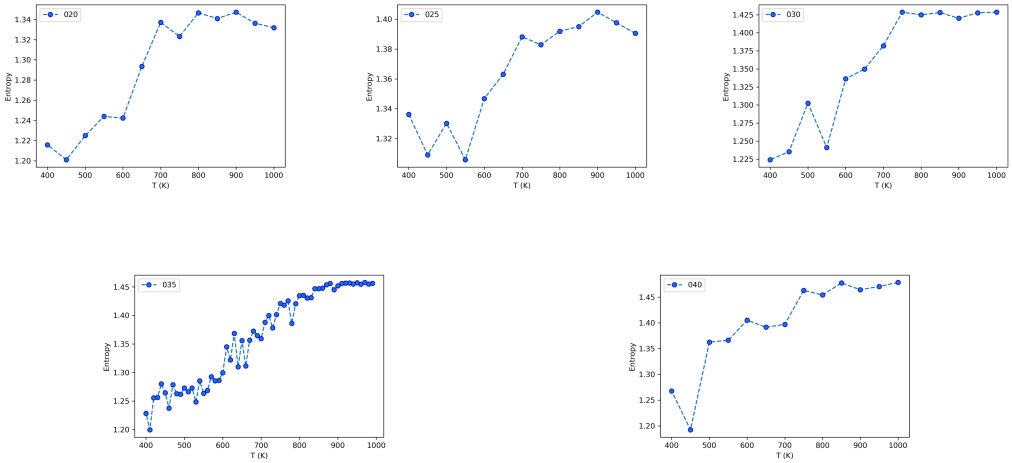

Figure 4.1: Degree of order plotted against temperature for selected compositions.

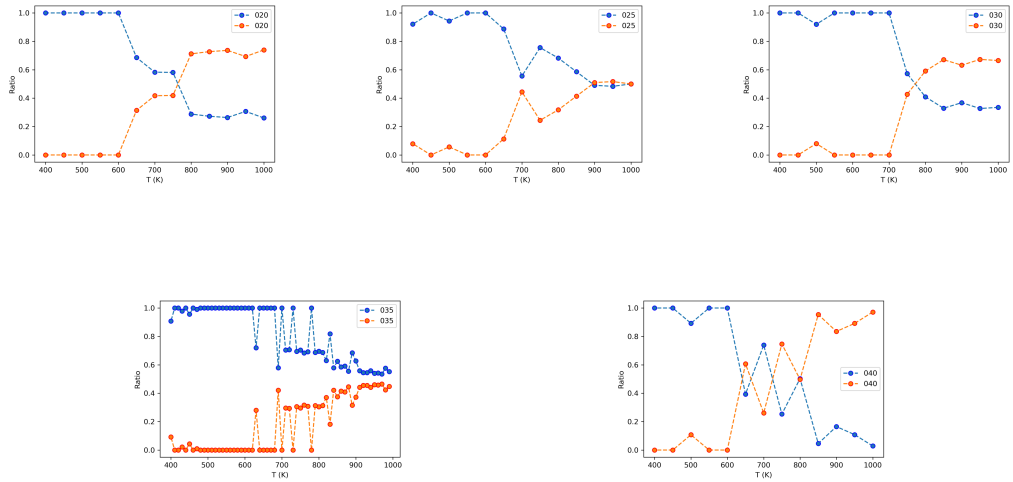

Figure 4.2: Ratio of lattice sites at ordered (blue) and disordered (red) configurations with respect to the total number of sites as function of temperature for slected compositions.

at elevated temperatures. Nevertheless, fluctuations in this parameter are present, i.e., the degree of order seems not to be monotonous with temperature. This is attributed to the following two reasons: (i) At regions of the phase diagram where order/disorder transitions occur, longer MC runs are neccesary in order to accurately describe the thermodynamics of the system. However, this is not the case for the MC claulcations from which the input structures have been derived. (ii) Due to available computational power restrictions we had to downsample our data and use a sparse representative set of representation vectors.

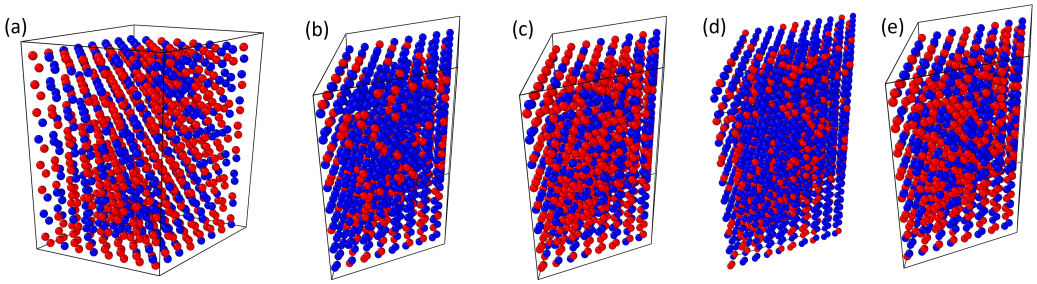

Figure 4.3: Distribution of ordered (blue) and disordered (red) configurations at T=800 K at (a)20%, (b) 25%, (c) 30%, (d) 35%, and (d) 40% In content.

In the next step we identify the ratio of lattice sites that belong to ordered and disordered configurations. To achieve this we define a critical degree of order parameter S_{crit}, which is a function of composition x:

$$S_{\text{crit}} = x^{-x} \tag{4.2}$$

This critical value serves as a threshold, distinguishing between ordered and disordered cluster centers within the alloy. Specifically, if a cluster center surpasses this critical value, it is considered disordered. Conversely, if it falls below S_{crit}, the cluster center is considered to demonstrate an ordered configuration. Although this definition is arbitrary, it allows to group the different onfigurations into two different classes and at least qualitatevely describe the degree of ordering.

The ratio of ordered and disordered configurations as function of temperature are plotted in Fig. 4.2. A notable revelation from the findings is the presence of phase transitions from ordered to disordered phases for each content. This transition signifies a critical shift in the alloy's configuration, indicative of a change in its state. As with the dependence of the degree of order on temperature, the aforementioned dependence is not monotonous and exhibit some fluctuations. This is attrubuted to the same two sources, i.e., not sufficiently long MC runs and sparse representation of the configurational space.

T=800 K is a temperature that is typically applied at the growth of InGaN alloys. Therefore, in order to investigate the distribution of ordered and disordered structures in the bulk of the alloy, in Fig. 4.3 the distribution of ordered and disordered

configurations is schematically shown for $x = 0.2 - 0.4$. As can be clearly seen, there is no separation between the ordered and disordered configurations for all contents.

4.2 Representative configurations

The investigation into the results of the problem revealed a consistent pattern across all content analyses conducted within the scope of this study. Notably, each structure analysis consistently yielded unit cell, i.e. translational symmetrie, that exhibited a high frequency of appearances, particularly under conditions of low temperature.

In the following we examine and discuss the most probable to occur cluster centers. VASP format is adopted to describe the crystal structure *:

- The first line includes the cluster center (in vecto format). In fully ordered structure all indeces would be 0 (In) or 1 (Ga). A partial occupation x, i.e., an index between 0 and 1, should be treated as a probability of find a Ga (x) or an In ($1 - x$) atom.
- The second line includes a rescale factor applied to the primitive vectors of the lattice. In all cases we use 1.0
- The next three lines include the three primitive vectors of the unit cell.
- The next three lines are species (here we denote them as X, Y, Z), the number of atoms of each type, and the type of coordinates (direct or cartesian).
- The atom positions are listed next in ascending order with respect to occupation, i.e., sites of lowest occupation are listed first.

*VASP is on of the most widely used DFT codes.

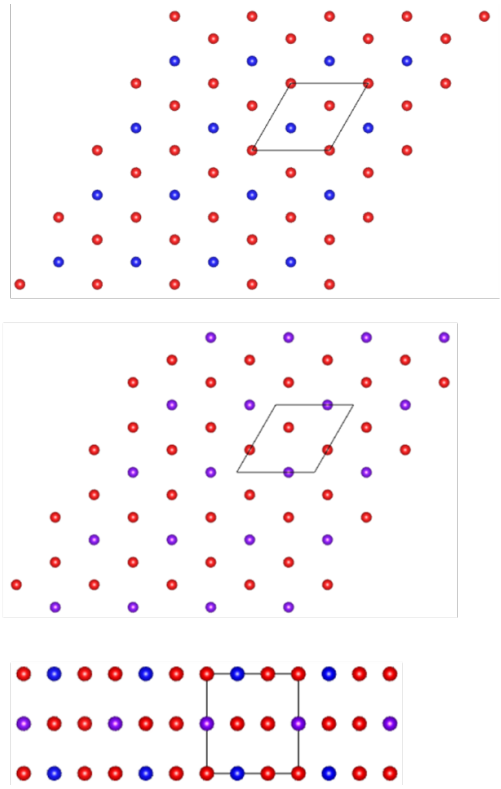

Figure 4.4: Schematic representation of the atoms at the two basal planes (top two figures) and in side view (bottom figure) of the Structure 1, at x=0.20. Blue balls represent occupation less than 0.2, red occupation above 0.8, and purple denotes intermediate occupations.

4.2.1 x=0.20

Structure 1

30% of lattice sites are in T=300 K belong to this luster center. The atomic geometry of this structure is shown in Fig. 4.4. This structure has the $\sqrt{3} \times \sqrt{3}$ unit cell. One c-plane has a content of $x = \frac{1}{3}$ and integer occupancies of all three lattice sites. The second basal lattice plane contains 2 Ga atoms, and in the third lattice site, we have a 50% probability of finding an In atom. The content of this plane is $x = \frac{1}{6}$. This gives a total In content of 25%. The In atoms are arranged as second nearest neighbors in both planes. The detailed information of this structure in vasp format follows:

```
#[1.  0.  1.  0.5 1.  1. ]
1.0
1.5000000 0.8660254 0.0000000
0.0000000 1.7320508 0.0000000
0.0000000 0.0000000 1.6329932
X Y Z
1 1 4
Direct
0.3333333 0.3333333 0.0000000
0.6666667 0.0000000 0.5000000
0.0000000 0.0000000 0.0000000
0.6666667 0.6666667 0.0000000
0.0000000 0.3333333 0.5000000
0.3333333 0.6666667 0.5000000
```

Structure 2

39% of lattice sites are in T=700 K belong to this luster center. The atomic geometry of this structure is shown in Fig. 4.5. The structure has the $\sqrt{3} \times \sqrt{3}$ unit cell. The first lattice plane contains 1 Ga atoms and the other two lattice sites 60% and 70% probability of being occupied by a Ga atom. The second lattice plane contains 2 Ga atoms, and the third site has 20% probability of finding an In atom. This gives a total In content of 15%. The detailed information of this structure in vasp format follows:

```
#[0.6 1.  0.7 0.8 1.  1. ]
1.0
1.5000000 0.8660254 0.0000000
0.0000000 1.7320508 0.0000000
0.0000000 0.0000000 1.6329932
```

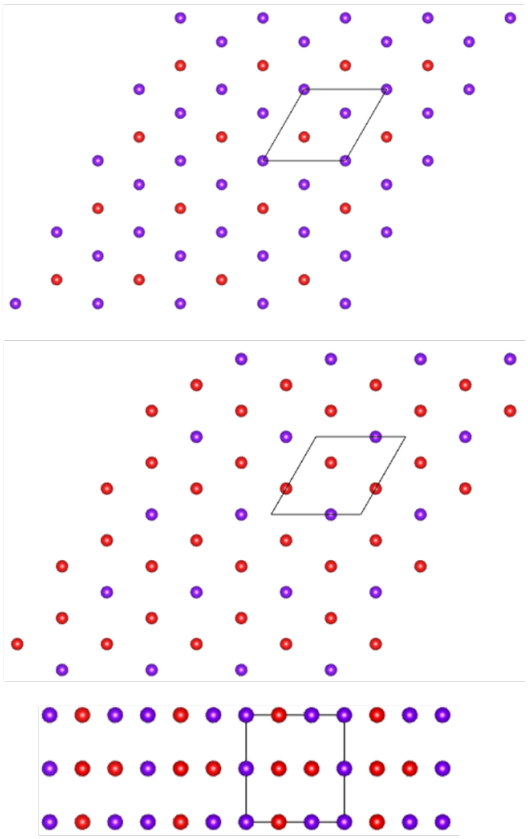

Figure 4.5: Schematic representation of the atoms at the two basal planes (top two figures) and in side view (bottom figure) of the Structure 2, at x=0.20. Blue balls represent occupation less than 0.2, red occupation above 0.8, and purple denotes intermediate occupations

```
X Y

3 3

Direct

0.0000000 0.0000000 0.0000000

0.6666667 0.6666667 0.0000000

0.6666667 0.0000000 0.5000000

0.3333333 0.3333333 0.0000000

0.0000000 0.3333333 0.5000000

0.3333333 0.6666667 0.5000000
```

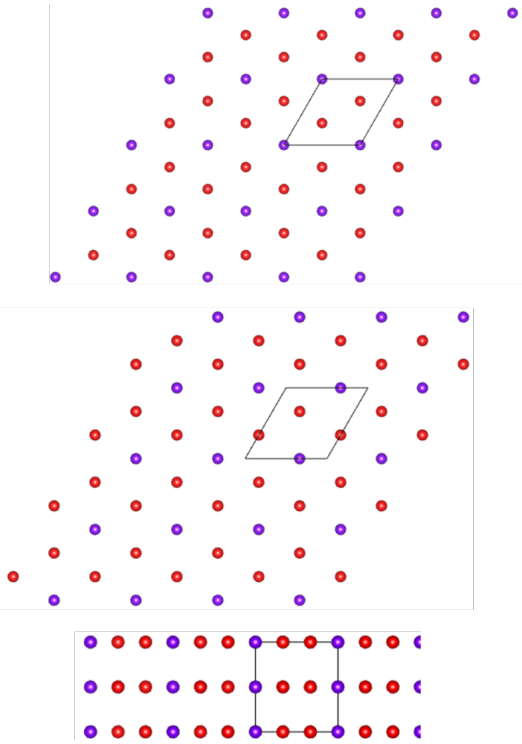

Figure 4.6: Schematic representation of the atoms at the two basal planes (top two figures) and in side view (bottom figure) of the Structure 1, at x=0.25. Blue balls represent occupation less than 0.2, red occupation above 0.8, and purple denotes intermediate occupations

4.2.2 x=0.25

Structure 1

54% of lattice sites at T=700 K belong to this cluster center. The atomic geometry of this structure is shown in Fig. 4.6. The structure has the $\sqrt{3} \times \sqrt{3}$ unit cell. The first lattice plane contains 2 Ga atoms and the other lattice site has 0.5 occupancy, i.e., 50% probability of being occupied by an In atom. The second lattice plane contains 2 Ga atoms, and the third site has 70% probability of finding an In atom. This gives a total In content of 20%. The detailed information of this structure in vasp format follows:

```
#[0.5 1.  1.  0.3 1.  1. ]
1.0
```

```
1.5000000 0.8660254 0.0000000
0.0000000 1.7320508 0.0000000
0.0000000 0.0000000 1.6329932
X Y
2 4
Direct
0.6666667 0.0000000 0.5000000
0.0000000 0.0000000 0.0000000
0.3333333 0.3333333 0.0000000
0.6666667 0.6666667 0.0000000
0.0000000 0.3333333 0.5000000
0.3333333 0.6666667 0.5000000
```

4.2.3 x=0.30

Structure 1

52% of lattice sites at T=650 K belong to this cluster center. The atomic geometry of this structure is shown in Fig. **??**. The structure has the $\sqrt{3} \times \sqrt{3}$ unit cell. This cluster center is rather disordered compared to the other structures we have examined so far. The first lattice plane contains one Ga atom and the other lattice sites have 0.5 and 0.8 occupancy, i.e., 50% and 80% probability of being occupied by an Ga atom. The second lattice plane contains 2 Ga atoms, and the third site has 80% probability of finding an In atom. This gives a total In content of 25%. The detailed information of this structure in vasp format follows:

```
#[0.8 0.5 1.  0.2 1.  1. ]
1.0
1.5000000 0.8660254 0.0000000
0.0000000 1.7320508 0.0000000
0.0000000 0.0000000 1.6329932
```

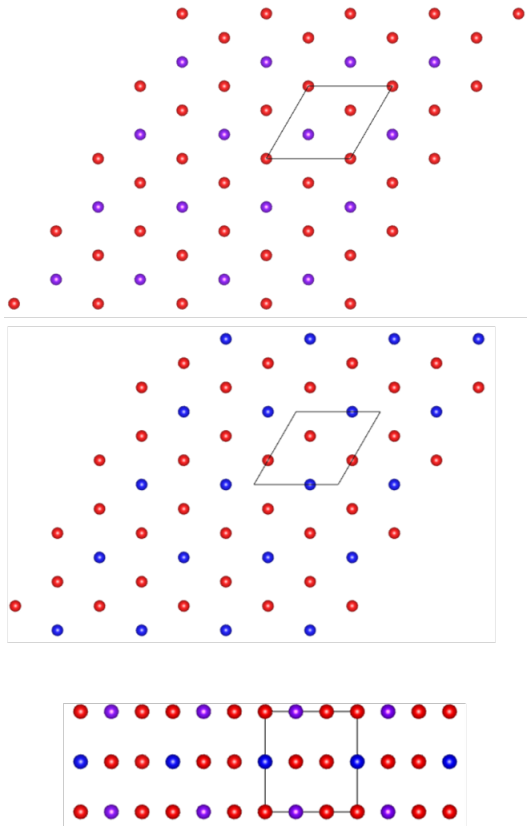

Figure 4.7: Schematic representation of the atoms at the two basal planes (top two figures) and in side view (bottom figure) of the Structure 1, at x=0.30. Blue balls represent occupation less than 0.2, red occupation above 0.8, and purple denotes intermediate occupations

```
X Y Z

1 1 4

Direct

0.6666667 0.0000000 0.5000000

0.3333333 0.3333333 0.0000000

0.0000000 0.0000000 0.0000000

0.6666667 0.6666667 0.0000000

0.0000000 0.3333333 0.5000000

0.3333333 0.6666667 0.5000000
```

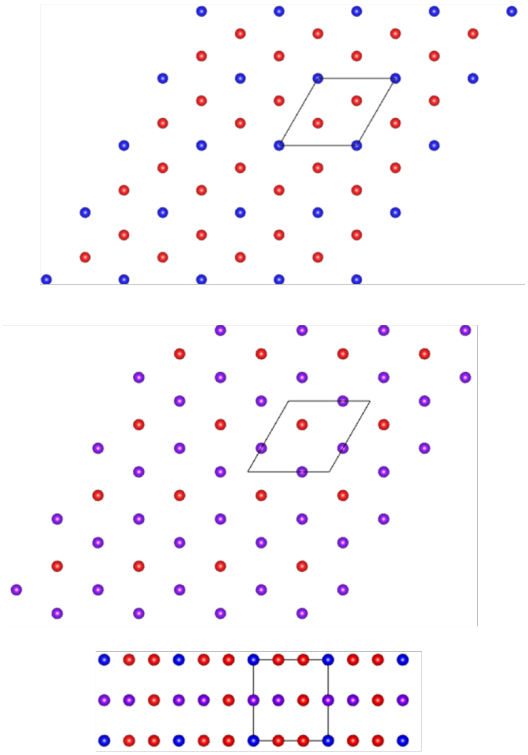

Figure 4.8: Schematic representation of the atoms at the two basal planes (top two figures) and in side view (bottom figure) of the Structure 2, at x=0.30. Blue balls represent occupation less than 0.2, red occupation above 0.8, and purple denotes intermediate occupations.

Structure 2

60% of lattice sites at T=600 K belong to this cluster center. The atomic geometry of this structure is shown in Fig. 4.8. The structure has the $\sqrt{3} \times \sqrt{3}$ unit cell. The first lattice plane contains two Ga atoms and the other lattice site has 10% probability of being occupied by an Ga atom. The second lattice plane contain one Ga atom, and the other lattice sites have 0.6 and 0.7 occupancty. This gives a total In content of 26%. The detailed information of this structure in vasp format follows:

```
#[0.1 1.  1.  0.6 0.7 1. ]
1.0
1.5000000 0.8660254 0.0000000
0.0000000 1.7320508 0.0000000
```

```
0.0000000 0.0000000 1.6329932
X Y Z
1 2 3
Direct
0.0000000 0.0000000 0.0000000
0.6666667 0.0000000 0.5000000
0.0000000 0.3333333 0.5000000
0.3333333 0.3333333 0.0000000
0.6666667 0.6666667 0.0000000
0.3333333 0.6666667 0.5000000
```

4.2.4 x=0.35

Structure 1

49% of lattice sites at T=670 K belong to this cluster center. The atomic geometry of this structure is shown in Fig. 4.9. The structure has the $\sqrt{3} \times \sqrt{3}$ unit cell. The first lattice plane contains two Ga atoms and the other lattice site has 80% probability of being occupied by an In atom. The second lattice plane contain one Ga atom with inenger occupancy, and the other lattice sites have 0.2 and 0.8 occupancy. This gives a total In content of 30%. The detailed information of this structure in vasp format follows:

```
#[0.2 1.  1.  0.2 0.8 1. ]
1.0
1.5000000 0.8660254 0.0000000
0.0000000 1.7320508 0.0000000
0.0000000 0.0000000 1.6329932
X Y Z
2 1 3
Direct
```

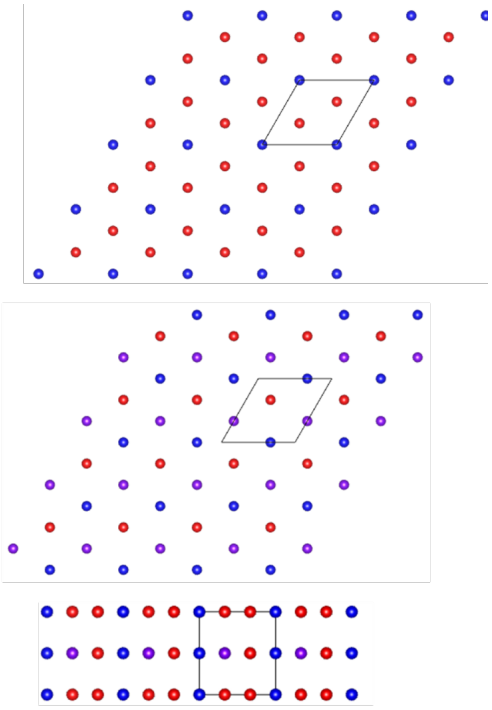

Figure 4.9: Schematic representation of the atoms at the two basal planes (top two figures) and in side view (bottom figure) of the Structure 1, at x=0.35. Blue balls represent occupation less than 0.2, red occupation above 0.8, and purple denotes intermediate occupations.

```
0.0000000 0.0000000 0.0000000
0.6666667 0.0000000 0.5000000
0.0000000 0.3333333 0.5000000
0.3333333 0.3333333 0.0000000
0.6666667 0.6666667 0.0000000
0.3333333 0.6666667 0.5000000
```

Structure 2

50% of lattice sites at T=650 K belong to this cluster center. The atomic geometry of this structure is shown in Fig. 4.10. The structure has the $\sqrt{3} \times \sqrt{3}$ unit cell. The first lattice plane contains one Ga atom and the other lattice sites have 30% and 80% probability of being occupied by an In atom. The second lattice plane contain one Ga atom with intenger occupancy, and the other lattice sites have 0.1.

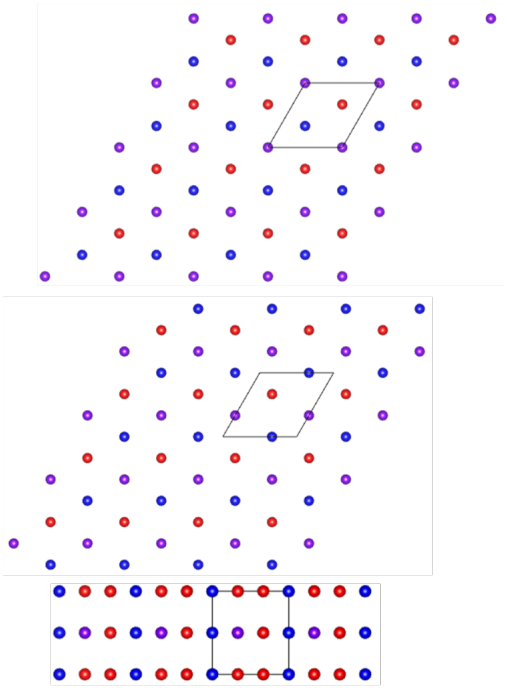

Figure 4.10: Schematic representation of the atoms at the two basal planes (top two figures) and in side view (bottom figure) of the Structure 2, at x=0.35. Blue balls represent occupation less than 0.2, red occupation above 0.8, and purple denotes intermediate occupations

This gives a total In content of 33%. The detailed information of this structure in vasp format follows:

```
#[0.7 0.2 1.  0.1 1.  1. ]
1.0
1.5000000 0.8660254 0.0000000
0.0000000 1.7320508 0.0000000
0.0000000 0.0000000 1.6329932
X Y Z
2 1 3
Direct
0.6666667 0.0000000 0.5000000
0.3333333 0.3333333 0.0000000
0.0000000 0.0000000 0.0000000
```

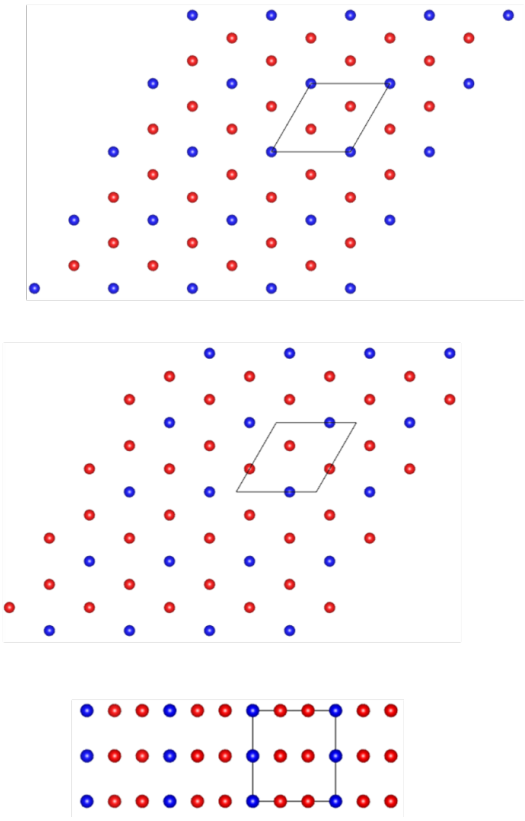

Figure 4.11: Schematic representation of the atoms at the two basal planes (top two figures) and in side view (bottom figure) of the Structure 1, at x=0.40. Blue balls represent occupation less than 0.2, red occupation above 0.8, and purple denotes intermediate occupations.

```
0.6666667 0.6666667 0.0000000
0.0000000 0.3333333 0.5000000
0.3333333 0.6666667 0.5000000
```

4.2.5 x=0.40

Structure 1

59% of lattice sites at T=500 K belong to this cluster center. The atomic geometry of this structure is shown in Fig. 4.11. The structure has the $\sqrt{3}\times\sqrt{3}$ unit cell. The lattice sites at the first lattice plane has no intenger occupancies. Nevertheless, the probability of find a Ga at each of these sites is 20%, 80%, and 80%. The second

basal plane contains two Ga atoms and one In atom. This gives a total In content of 36%. The detailed information of this structure in vasp format follows:

```
#[0.2 0.8 0.8 0.  1.  1. ]
1.0
1.5000000 0.8660254 0.0000000
0.0000000 1.7320508 0.0000000
0.0000000 0.0000000 1.6329932
X Y
2 4
Direct
0.6666667 0.0000000 0.5000000
0.0000000 0.0000000 0.0000000
0.3333333 0.3333333 0.0000000
0.6666667 0.6666667 0.0000000
0.0000000 0.3333333 0.5000000
0.3333333 0.6666667 0.5000000
```

Structure 2

The percentage of the cluster center in addition with others is about 42% at 400K 42% of lattice sites at T=400 K belong to this cluster center. The atomic geometry of this structure is shown in Fig. 4.12. The structure has the $\sqrt{3} \times \sqrt{3}$ unit cell. The first *c*-plane includes one Ga and one In atom. The third latticde site has 60% probabillity to be occupied by a Gatom. The second basal plane has one In atoms. The other two lattice sites have 80% and 90% probability to be occupied by a Ga tom. This gives a total In content of 40%. The detailed information of this structure in vasp format follows:

```
#[0.6 1.  0.  0.  0.8 0.9]
1.0
```

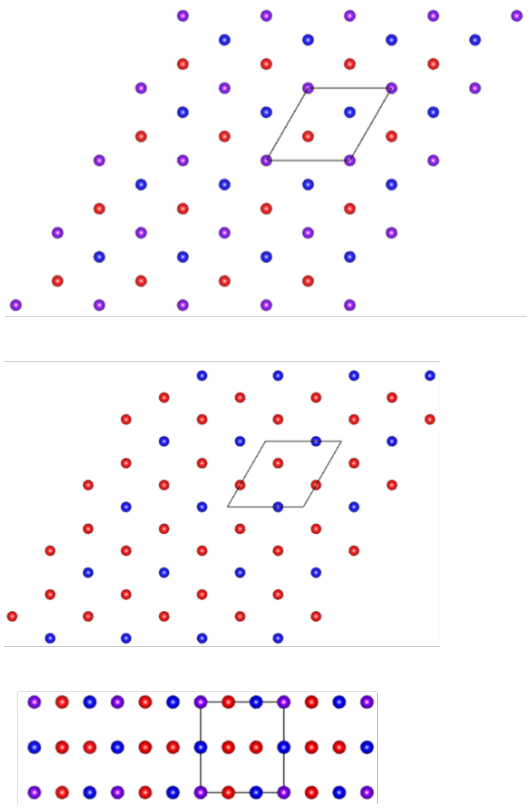

Figure 4.12: Schematic representation of the atoms at the two basal planes (top two figures) and in side view (bottom figure) of the Structure 2, at x=0.40. Blue balls represent occupation less than 0.2, red occupation above 0.8, and purple denotes intermediate occupations.

```
1.5000000 0.8660254 0.0000000

0.0000000 1.7320508 0.0000000

0.0000000 0.0000000 1.6329932

X Y Z

2 1 3

Direct

0.6666667 0.6666667 0.0000000

0.6666667 0.0000000 0.5000000

0.0000000 0.0000000 0.0000000

0.0000000 0.3333333 0.5000000

0.3333333 0.6666667 0.5000000

0.3333333 0.3333333 0.0000000
```

4.3 General trends

Based on the aformentioned analysis we see that the most probable configuration for InGaN alloys is described by a $\sqrt{3} \times \sqrt{3}$ unit cell. As has already described in the Introduction (see Fig. 1.3) at 33% In content, In atoms preferencially allign as second nearest neighbohrs. This is due to efficient strain relaxation. This allignement results in a $\sqrt{3} \times \sqrt{3}$ unit cell. The methodology developed and applied in the present thesus indicate that the same mechanism is active at other compositions as well. Another key outcome of these calculations is that unit cells with $\sqrt{3} \times \sqrt{3}$ symmetry can be used in DFT calculations either to investigate ordered structures of random structures. For the later, special quasirandom structures, where the occupacion probabilities are adjusted according the occupancies of the cluster center, can be applied.

Chapter 5

Conclusion

In this thesis, we have developed and implemented a novel methodology to investigate the configurational space of binary alloys, specifically focusing on InGaN alloys. This methodology integrates symmetry properties of the parent lattice and employs unsupervised machine learning approaches to identify and analyze configurational patterns.

We introduced a self-consistent clustering algorithm that uses vectors representing unit cells of varying sizes and symmetries to search for patterns within thousands of configuration vectors. This algorithm effectively identifies cluster centers and assigns representation vectors to these centers, allowing for the construction of a degree of order parameter and the assignment of high symmetry cluster centers to lattice sites.

Our methodology was applied to pseudobinary InGaN alloys, utilizing large alloy structures produced by Monte Carlo calculations. The results confirmed the tendency of indium atoms to align as second nearest neighbors, leading to a $\sqrt{3} \times \sqrt{3}$ translational symmetry. This key structural motif is prevalent across different compositions of InGaN alloys due to efficient strain relaxation, particularly at 33% indium content.

The cluster centers derived from our methodology can be directly used in Density Functional Theory (DFT) calculations or to produce special quasirandom struc-

tures (SQS) with modified probabilities for the occupation of lattice sites. This demonstrates the robustness and versatility of our approach in modeling both ordered and random alloy behaviors while maintaining computational efficiency. In conclusion, the novel methodology developed in this thesis provides a powerful tool for investigating and understanding the configurational

Bibliography

[1] M. Albrecht, L. Lymperakis, J. Neugebauer, J. E. Northrup, L. Kirste, M. Leroux, I. Grzegory, S. Porowski, and H. P. Strunk, Phys. Rev. B **71** (2005), 035314.

[2] C. Caetano, L. K. Teles, M. Marques, A. Dal Pino, and L. G. Ferreira, Phys. Rev. B **74** (2006), 045215.

[3] E. Clementi, D. L. Raimondi, and W. P. Reinhardt, The Journal of Chemical Physics **47** (1967), no. 4, 1300–1307.

[4] J. M. Cowley, Phys. Rev. **77** (1950), 669–675.

[5] Andrew I. Duff, Liverios Lymperakis, and Jörg Neugebauer, physica status solidi (b) **252** (2015), no. 5, 855–865.

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al., kdd, vol. 96, 1996, pp. 226–231.

[7] Pasi Fränti and Sami Sieranoja, Pattern Recognition **93** (2019), 95–112.

[8] C. K. Gan, Y. P. Feng, and D. J. Srolovitz, Phys. Rev. B **73** (2006), 235214.

[9] Raia Hadsell, Sumit Chopra, and Yann LeCun, *Dimensionality reduction by learning an invariant mapping*, 2006 IEEE computer society conference on

computer vision and pattern recognition (CVPR'06), vol. 2, IEEE, 2006, pp. 1735–1742.

[10] Dayan P Hinton GE, Frey BJ, and Neal RM., Science (1995).

[11] E. Iliopoulos, Jr. Ludwig, K. F., T. D. Moustakas, and S. N. G. Chu, Applied Physics Letters **78** (2001), no. 4, 463–465.

[12] S. Yu. Karpov, MRS Internet Journal of Nitride Semiconductor Research **3** (1998), 16.

[13] Hopfield JJ. Krotov D, Natl Acad Sci U S A. (2019).

[14] B Kulis and MI Jordan, *A new k-means algorithms via bayesian nonparametrics*, Proceedings of the 29th International Conference on Machine Learning, 2012.

[15] Sangheon Lee, Christoph Freysoldt, and Jörg Neugebauer, Phys. Rev. B **90** (2014), 245301.

[16] L. Lymperakis, T. Schulz, C. Freysoldt, M. Anikeeva, Z. Chen, X. Zheng, B. Shen, C. Chèze, M. Siekacz, X. Q. Wang, M. Albrecht, and J. Neugebauer, Phys. Rev. Mater. **2** (2018), 011601.

[17] D. Drikakis N. Christakis, Mathematics (2023).

[18] S. Nakamura and G. Faso, *The blue laser diode*, © Springer-Verlag, Berlin Heidelberg, 1997.

[19] G.B. Stringfellow, Journal of Crystal Growth **312** (2010), no. 6, 735–749.

[20] A. Tabata, L. K. Teles, L. M. R. Scolfaro, J. R. Leite, A. Kharchenko, T. Frey, D. J. As, D. Schikora, K. Lischka, J. Furthmüller, and F. Bechstedt, Applied Physics Letters **80** (2002), no. 5, 769–771.

[21] Agostino Zoroddu, Fabio Bernardini, Paolo Ruggerone, and Vincenzo Fiorentini, Phys. Rev. B **64** (2001), 045208.