# Isometry Invariants of Crystal Structures Based on Voronoi Domains and Interatomic Distances

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

**Marco Michele Mosca**

October 2022

# Contents

# Abstract

The need for comparison methods between crystal structures led the research to look for proper descriptors that could encode the chemical properties of different materials. Many crystal structures exist in theory, but only some of them may be **synthesized in a laboratory** and used in the real world for **practical applications**. To discover new materials, Crystal Structure Prediction is vital in predicting various crystal structure forms or generating new ones by building blocks or molecules. Usually, a structure prediction computes chemical features that are not **correct properties** because they do not consider the entire 3-dimensional structure of a crystal. However, they rely on rules considering only the type of particles involved. For example, some chemical properties are used frequently to select a good candidate for synthesis because, in theory, they could tell if a crystal may exist in its solid form under some environmental conditions.

This thesis project aims to design and develop new geometric tools or properties that can properly distinguish 3-dimensional structures starting from the raw atom coordinates. The Geometric features developed in this document are fast properties or numerical characteristics that map crystal structure to a **different space** for a more reliable and efficient comparison.

Firstly, we solved the **comparison problem** between crystal lattices by designing a property that maps a crystal lattice to the space of polyhedra and a metric that can distinguish them.

Secondly, we designed a new and faster geometric **property** that relies on vectors of interatomic distances that are proved to change continuously under atom perturbations. Finally, the **chemical property prediction** was addressed in our last work, where we attempted to predict the chemical properties of crystals by using our geometric features.

# Chapter 1

# Introduction to crystal structures

Crystals are solid crystalline materials that can be formed by many substances if cooled sufficiently to reach a solid phase. In this phase, atoms pack together until they arrange in a repeating array. A simple example of a crystal could be the ice structure, where water molecules are arranged in different positions and adjusting their orientation accordingly to chemical forces between atoms. In the following paragraphs, we will deal with atoms and forces, starting from basic definitions and building up the concept of a crystal structure step by step.

## 1.1 Atomic particles

### 1.1.1 Atoms

Before going to some definitions, let us take sodium and chlorine. Sodium ($S$) is a metal that consists of sodium atoms, and chlorine is a green gas made up of two chlorine ($Cl$) atoms paired up in $Cl_2$ molecule. Both are toxic to humans and highly reactive. Nevertheless, when they react together, they form a joint compound called sodium chloride ($NaCl$) or rock salt, the common salt that we can find daily in our kitchen to make our food tastier and full of flavour. If we consider sodium and chlorine singularly, they are called **elements** because they are made up of the same type of atoms and eventually paired up as molecules. On the other hand, sodium chloride ($NaCl$) is called **compound** as different types of atoms form its structure. Atoms could be arranged together in molecules or network structures. When we refer to an atom, we think of a small particle with a **nucleus** in the centre and **electrons** spinning around it. The idea of this small entity is called a model of an atom. According to this model of an atom, the nucleus is the dense and positively charged part that consists of positively charged **protons** and uncharged **neutrons** (except for hydrogen), which keep the nucleus tight by a robust nuclear force. This force is stronger than the repulsion between protons. Electrons are the particles involved in chemical reactions, and they are arranged in circular shells around the nucleus. The outer shell contains the reactive electrons called **valence electrons** on which the number of possible reactions depends. Each element differs from the other based on the number of protons, given by the symbol $Z$ and called **atomic number**.

### 1.1.2 Ions

Under normal conditions, every atom is electrically neutral as the number of protons and electrons are equal. Nevertheless, it is possible to move electrons away from the nucleus during some types of reactions when charge neutrality is broken. For example, when sodium and chlorine react together, energy is released to build bonds between them, and electrons are transferred from sodium $Na$ close to the chlorine nucleus $Cl$. In this case, protons would be non-neutralized in $Na$ forming a positive ion or **cation** $Na^+$ and a negative force will overcome the neutral charged nucleus in $Cl$ forming a negative ion or **anion** $Cl^-$:

$$Na \rightarrow Na^+ + e^- \tag{1.1}$$

$$Cl + e^- \rightarrow Cl^- \tag{1.2}$$

with $Na^+$ the correspondent cation from sodium, $e^-$ the transferred electron, and $Cl^-$ the anion from chlorine. The required energy to move the farthest electron away from the nucleus is called **ionization energy** which is defined as the energy change for the process in equation (1.1) and measured in $\frac{kJ}{mol}$. Each element has its own set of ionization energies depending on the number of shells where electrons lie and their distance from the nucleus. Therefore, the type of bond that two atoms may form depends on it. For example, in sodium chloride, both atoms have different first ionization energies $I_1$ such as 496 kJ/mol for sodium and 1251 kJ/mol for chlorine. The high difference between them suggests that they will form an ionic bond where electrons move closer to one atom with respect to the other forming ions on both sides. The higher first ionization energy $I_1$ of chlorine results from the fact that chlorine needs more energy to release an electron from the outer shell because the high positive charge of protons keeps electrons tighter to the nucleus with a lower distance. Thus, it needs more effort to release the first electron from the outer shell. Moreover, ionization energies can be provided to remove electrons one by one from an atom, and they should always be greater than the previous one. In fact, the electron is removed from a more positively charged ion and could lie in tighter shells, closer to the nucleus and subject to a stronger force. For example, the second ionization energy $I_2$ of sodium is 4562 $\frac{kJ}{mol}$ and will follow the equation:

$$Na^{+} \rightarrow Na^{2+} + e^{-} \tag{1.3}$$

## 1.2 Atomic structure and Interactions

Atoms are small entities that form the matter surrounding us, starting from the most straightforward example, such as the water in our seas or oceans, and ending with the carbon structure of a diamond, which occurs very rarely in nature and makes it the most expensive crystal. Although we can not spot or see atoms, why would we be able to see the matter they form under the light? The answer lies in light waves. Light behaves as a wave, and we can see everything that is larger than the light wavelength (400 nm - 700 nm). The first quantum model of an atom was proposed by **Bohr** in 1913, who suggested that electrons do move around a nucleus at specific distances from it and, indeed, their ionization energy changes at each level. It means that the electronic energy in an atom is quantized. He stated that the electron moves around the nucleus in a fixed orbit in hydrogen, which can be thought of as a spherical shell. Shells are indexed by an integer number $n \in \mathbb{N}$ called **quantum number**, which starts from 1, representing the closest orbit to the nucleus. The electronic energy changes concerning the orbit (given by $n$) on which the electron lies as shown below:

$$E = -\frac{k}{n^2} \tag{1.4}$$

where $n$ is the quantum number and $k$ is a constant which considers the mass and charge of the electron. Bohr's model of an atom became outdated later but the concept of quantized energy and quantum number remained as solid basics for further theories.

### 1.2.1 Wave mechanics

Properties of matter are related to those small particles called electrons that spin around the nucleus. Spectroscopy is a method used to study the atomic structure, and uses the strength of electromagnetic radiations (such as X-rays) that are absorbed or emitted by atoms. **Electromagnetic radiation** is a form of energy that consists of an oscillating electric-magnetic field holding physical properties such as the **wavelength** $\lambda$, **amplitude** and **frequency** $v$. The wavelength $\lambda$ (measured in $m$ or $nm$) is the distance between two wave crests, and the frequency $v$ (measured in hertz $Hz$) refers to the number of waves crests that pass over the origin every second. According to the electromagnetic spectrum, which contains all types of radiation, we can see only a small part of it with our eyes depending on the waves that all matter emits. These emissions allow us to see different colours that lie in the electromagnetic spectrum, provided that a wavelength ranges from 390 nm (purple) to 740 nm (red) with intermediate values that correspond to all remaining colours. Every type of wave before and after this range cannot be seen by our eyes, but they exist as radio waves, microwaves, infrared, ultraviolet, x-rays and $\gamma$-rays. In 1924, **De Broglie** proposed that electrons can behave as waves, so that they have a wavelength property which is inversely proportional to the mass($m$) and the velocity ($v$) as shown below:

$$\lambda\ (wavelength) = \frac{h}{mv} \tag{1.5}$$

where $h$ is the Planck constant. This idea of an electron could seem strange since we cannot see the wave behaviour of the surrounding matter. Only in 1925, **Davison**, **Germer** and **Thomson** carried out an experiment to obtain a diffraction pattern given by the matter. They shot a beam of electrons to a crystal of nickel, matching the electron beam wavelength with the distance between nickel atoms, and modifying the electrons' velocity $v$ according to the equation (1.5). The diffraction pattern was originated by the electrons, which are charged particles (unlike neutrons and X-rays). Therefore, the beam of electrons interacted with nickel atoms, particularly with protons and electrons travelling at a specific wavelength. Indeed, diffraction occurs because the particle's wavelength is the same as the interatomic distance and the crystal acts as a diffraction grid. To study the electron trajectory, we should shoot electromagnetic radiation at high energy that will cause the trajectory deviation and the change in velocity. Therefore, it will not be possible to obtain the perfect position and calculate the exact velocity at which an electron spins around. The electron uncertainty theory was first discussed by **Heisenberg** in the 1920s. Although it is not possible to spot a precise location, we can take into consideration the probability of an electron occupying a region of the space, which is called **atomic orbital**. Finally, after the old Bohr model that belongs to the quantum mechanics theory, the concepts of uncertainty, probability and atomic orbitals led chemistry to a modern theory called **wave mechanics**. Before 1927, electrons were described in terms of position, velocity and energy until **Schrodinger** developed an equation called **wavefunction** $\psi$. This function led to the wave mechanics theories where undetermined orbitals replaced Bohr's determined orbits. Consequently, the wave function is not used to determine the exact position but only to detect the probability that an electron has to occupy a particular region of the space around the nucleus (atomic orbital). The study of electron energy involves computations of partial derivatives meaning that we can study the change in energy associated with $\psi$ along one axis. In contrast, the others are constant or fixed.

The wave function is not directly measurable, but contains only information about the electron behaviour. The German physicist **Born** suggested considering the square of the wave function $\psi^2$ and thinking of it as a measurable property proportional to the probability of finding the electron within a small space $d\tau$. In particular, where $\psi^2$ is large, the probability of finding an electron is high. Finally, the location of an electron cannot be precisely computed, and indeed, we should take the probability that an electron has to occupy a specific space. When we consider a small volume of space $d\tau$, it is suitable to refer to the *probability per unit volume*, defined as the **electron density**.

### 1.2.2 Atomic orbitals

Unlike Bohr's orbit, the atomic orbital does not have precise borders and represents the space around the nucleus where the probability of finding an electron is very high. According to Bohr's model, the hydrogen atom has a positive nucleus and an electron spinning around it on an orbit of radius 0.53 Å. On the other hand, according to wave mechanics, the hydrogen atom consists of a positive nucleus located at the centre of the orbital (centre of a sphere with radius 0.53 Å), which circumscribes the space where the maximum probability of finding an electron lies.

Schrodinger succeeded in solving the total energy equation related to the wave function $\psi$ in 1927, showing that the following equation gave energy levels:

$$E_n = \frac{h\ R}{n^2} \tag{1.6}$$

where $R$ is a constant that involves the mass and charge of the electron, $h$ is the Plank constant and the vacuum permittivity and $n$ is the quantum number. Since there are different quantum numbers resulting from Schrodinger equation, $\boldsymbol{n}$ is called **principal quantum number** and can be assigned to values from 1 to $\infty$ (just the first 7 are chemically significant). The principal quantum number represents the electron's energy level that is almost similar to Bohr's energy rings, but with the difference that there is no fixed distance nucleus-electron (wave mechanics uncertainty). The size of the orbital depends on $n$ together with the maximum number of electrons that it can contain.

Each energy level $n$ may have several types of orbitals that have different shapes. The type is referred to as the **secondary quantum number** or orbital quantum number and it is represented by the symbol $\boldsymbol{l}$. The symbol $l$ can take values from 0 to $n-1$, though only values from 0 to 3 are chemically significant. These sublevels referred by $l$ were found to be strictly related to each level $n$ and indeed each energetic level $n$ may have at maximum $n$ sublevels. As well as representing them by integers from 0 to 3, the following lower case letters were used to name the most significant ones (first four): $\boldsymbol{s}$, $\boldsymbol{p}$, $\boldsymbol{d}$, $\boldsymbol{f}$ (ordered according to the increasing energy) and each sublevel has a different geometric shape: spherical for *s*, a filled 3D infinity symbol shape for *p* and a more complex symmetric shape for *d* and *f*. Take as example the oxygen atom ($O$): in its ground state, electrons occupy two energy levels $n = 2$: the first one contains a spherical sublevel *s* with 2 electrons (exponent), the second has a spheric sublevel *s* plus a *p*-type sublevel with 6 electrons in total.

O

$1s^2 2s^2 2p^4$

Finally, each orbital type $l$ may have different orientations in the 3D space and the set of all possible orientations forms the entire type. For example, the *p* type orbital ($l = 1$) may be located in the 3D space along the three axis thanks to its shape and therefore, it may have three different orientations $m_1 \in [-1, 0, +1]$ that is referred to as the **magnetic quantum number** (symbol $\boldsymbol{m_l}$). The magnetic quantum number $m_l$ depends on $l$ and can be assigned to values from -$l$ to $l$ with $2l+1$ values in total. In particular, the sublevel s ($l = 0$) contains only one orbital type that does not have any preferred orientation (spherical symmetry) and corresponds to $m_0 = 0$. Regarding the sublevel p ($l = 1$), three values are allowed $m_1 \in [-1, 0, +1]$ which define three orbitals along x, y and z axis

$(p_x,\ p_y,\ p_z)$, perpendicular to each other and oriented differently that form the entire orbital $p$.

### 1.2.3 Components of the wave function $\psi$

The wave function $\psi$ determines the behaviour of an electron spinning around a nucleus. According to Born, the square $\psi^2$ can be related to the probability of finding it within a small region of the space $d\tau$. This behaviour can be explained in a 3D space on a Cartesian system with the origin at the point $(0, 0, 0)$ and any other point at coordinates $(x, y, z)$. Therefore the wave function will be a function that takes in input three coordinates $\psi(x, y, z)$. Nevertheless, an atom can be modelled as a sphere where the nucleus lies at the origin coordinates $(0, 0, 0)$ and all electrons spin around it. Since an atom is related to a sphere, it is more suitable to identify points in the space through another system called the spherical system, which is similar to the Cartesian system except for the values used as coordinates. Indeed, the first coordinate will be the distance from the origin to a specific point $x = r$ where $r$ is the distance between the electron and the nucleus, followed by two angles $y = \theta$ and $z = \phi$ that describe the orientation of the electron concerning the nucleus.

$$\psi(x, y, z) = \psi(r, \theta, \phi) = R(r)\ Y(\theta, \phi) \tag{1.7}$$

A spherical system allows us to split the wave function into two components: the **radial wave function** that depends on the distance between the nucleus and the electron $R(r)$, and the **angular wave function** $Y(\theta, \phi)$ that depends on the angles describing the shape of the orbital, as shown in (1.7). Since nuclear forces applied to electrons follow a spherical behaviour, it is reasonable to consider the radial wave function, which depends only on the distance from the nucleus. Given the Born interpretation of $\psi^2$, it represents the probability of finding an electron within a small volume $d\tau$. Consequently, the function $R^2(r)\ d\tau$ is the probability of finding an electron within a small volume $d\tau$ at a distance $r$ from the nucleus, which is equal to the electron density at that point. Finally, the probability of finding the electron anywhere in the whole space must be equal to 1 and it is expressed as follows

$$\int_{r=0}^{r=\infty} \psi^2 d\tau = 1 \tag{1.8}$$

and called single-electron wave function.

Although we need to consider a region of the space to calculate the probability, it is more beneficial to restrict the probability computation to a spherical region around the nucleus with a specific region thickness. Indeed, we may want to compute the most probable distance of finding an electron. It can be achieved by plotting the radial distribution function against $r$. The **radial distribution function** is related to the probability of finding an electron in a spherical shell of radius $r$ and thickness $dr$:

$$RDF = 4\pi r^2 \; R^2(r) \; dr \tag{1.9}$$

### 1.2.4 Octet rule

How can elements join together to form compounds of different types and with various physical properties? It is rare that, in nature, elements exist as isolated atoms. Indeed most of the atoms combine with the same or different types to create molecules of elements or molecules of compounds. To understand how chemical reactions cause the making and breaking of bonds, we need to understand how electrons interact with each other to create a bond and which energy levels are involved.

The bonding in many molecules can be achieved by considering their electronic configuration. Atoms approach a stable state where the outer orbital has the maximum number of electrons. For example, the hydrogen atom is formed by an atomic orbital in the first energy level $n = 1$ with only the spherical type $s$ ($l = 0$). It has 1 electron, which lies in this orbital type $1s^1$, identified by the exponent of $s$. Since the orbital type $s$ can have a maximum of 2 electrons, it needs another electron to reach a **stable state** and match the electronic configuration of helium (He), a noble gas (stable element with the maximum number of electrons in the outer shell that belongs to the VIII group of the periodic table). All elements are listed in the periodic table and belong to a specific group that determines the number of electrons they have in the outer shell called **valence electrons**. For instance, carbon C belongs to the IV group meaning that the outer shell has 4 valence electrons involved in reactions. Some elements tend to complete their outer shell by acquiring electrons, like chlorine *Cl*, from other atoms. Others remove a few remaining electrons of the outer shell to reach stability. This rule is called **octet rule** since for most atoms eight electrons in the outer shell corresponds to the filled $s$ and $p$ orbital types $ns^2np^6$ where $n$ is the energy level, $s$ and $p$ are orbital types of $n$ and the exponents are the number of electrons in the atomic orbital ($2 + 6$ in total).

In 1916, **Lewis** was the first person to develop the concept of electron sharing in molecules and recognized that atoms in a molecule could obey the octet rule by sharing electrons of the outer shell. Bonds and electron configuration can be represented by dots and lines to show lone pairs of electrons in the outer atomic orbital (shell) and highlight those that make a physical bond. For example, an oxygen molecule is formed by two oxygen atoms that belong to the VI group of the periodic table, and it means that they need other 2 electrons to complete and match with the corresponding noble gas Neon ($Ne$). Each atom can be drawn in the **Lewis diagram** with four dots that represent lone pairs and two lines meaning that they share 2 electrons (double bond) to both complete the octet as shown in diagram 1.10. Indeed, the outer shell $n = 2$ is made of $2 + 4$ valence electrons, and 2 of them will be shared with the other to complete the outer atomic orbital $n = 2$ and reach 8.

$$\underset{1\mathrm{s}^2 2\mathrm{s}^2 2\mathrm{p}^4}{:\ddot{\mathrm{O}}} \quad = \quad \underset{1\mathrm{s}^2 2\mathrm{s}^2 2\mathrm{p}^4}{\ddot{\mathrm{O}}:} \tag{1.10}$$

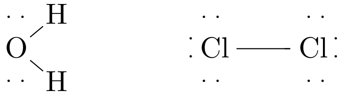
### 1.2.5 Bond types

In Section 1.1.2, we dealt with a structure called sodium chloride $NaCl$ that has **ionic bonding**, a strong and non-directional bond formed by a cation ($Na^+$) and an anion ($Cl^-$). This type of bond forms between two oppositely charged ions thanks to the electrostatic attraction. Electrons are transferred from one atom to another and the attractive force, $F$, that keeps them tight is given by Coulomb's Law: $F = \frac{q_1 q_2}{r^2}$ where $q_1$, $q_2$ are the ions charges and $r$ is the distance between them. Indeed, sodium belongs to the II group of the periodic table. Therefore it is unstable since it needs other 7 electrons to complete the outer shell $n = 3$ according to its electronic configuration $1s^2 2s^2 2p^6 3s^1$. This type of atom tends to get rid of electrons instead of sharing them to become stable, as shown in diagram 1.11.

$$\underset{1\mathrm{s}^2 2\mathrm{s}^2 2\mathrm{p}^6 3\mathrm{s}^1}{.\mathrm{Na}} \quad + \quad \underset{1\mathrm{s}^2 2\mathrm{s}^2 2\mathrm{p}^6 3\mathrm{s}^2 3\mathrm{p}^5}{:\underset{..}{\ddot{\mathrm{Cl}}}\,\dot{}} \quad \longrightarrow \quad \underset{1\mathrm{s}^2 2\mathrm{s}^2 2\mathrm{p}^6}{\mathrm{Na}^+} \quad + \quad \underset{1\mathrm{s}^2 2\mathrm{s}^2 2\mathrm{p}^6 3\mathrm{s}^2 3\mathrm{p}^6}{:\underset{..}{\ddot{\mathrm{Cl}}}:\ ^-} \tag{1.11}$$

After the ionic bond has occurred, sodium ion reaches stability with 8 electrons in the outer shell $n = 2$ after giving the remaining one to chlorine, which becomes a negatively charged

ion. Ionic forces are effective over large distances where ions pack together in repeating arrays in ionic crystals to maximize the Coulomb attraction and minimize repulsion.
Sodium *Na* does not occur in nature, but it can be found in different compounds from which it can be prepared to pack sodium cations and make a crystalline structure. The type of bond that keeps the sodium ions tight is called **metallic bond** which is a characteristic of all metals in the first three groups, including the transition metals such as iron *Fe*. This type of bond is genuine and generates regular arrays of metal cations surrounded by a 'sea' of electrons. Electrons are delocalised because they occupy the space between the cations and can move around, giving these materials the property of *electric conductivity*. Indeed, these atoms, which belongs to the first groups, tend to lose electrons easily since their first ionization energies are lower.
In order to react, all atoms in a molecule try to attract electrons as much as they can. The power of electrons' attraction is called **electronegativity**, and the shared electrons are pulled towards the more electronegative atom. This property depends on how hungry atoms of electrons are, and it is directly related to the number of electrons they need to complete the outer shell. According to the periodic table of elements, electronegativity increases by rows and decreases down a column (or group). For example, a water molecule $H_2O$ consists of 1 oxygen (electronegativity at 3.44) and two hydrogen atoms (electronegativity at 2.2). Because of its higher electronegativity, all electrons in the hydrogen orbital type $s$ are pulled closer to the oxygen one. Instead, let us take a diatomic molecule made of the same elements as in the green gas chlorine $Cl_2$. Electrons are not moved towards the other atom since *Cl* atoms have the same electronegativity (3.16). This type of bonding, in which electrons are shared between atoms, is called **covalent bond**. It is strong and directional, and may involve reactions between different atoms where each of them may be more or equally hungry of electrons, see diagram 1.12.



(1.12)

The presence of a covalent bond can affect the structure of a molecule where electronegative elements such as oxygen attract the bonding electrons considerably. Therefore, electronegative atoms in a molecule, such as water $H_2O$, cause the bond to acquire a partial negative charge $\delta^-$ on one end and a partial positive charge $\delta^+$ on the other end. The separation between these two charges generates an electric dipole, and the molecule is said to be a **polar molecule**. Given their modified electronegative structure, polar molecules in water can interact with each other since the negative part $O^{\delta^-}$ of a molecule $A$ can interact through a weak bond with the positive part $H^{\delta^+}$ of molecule $B$. This type of interaction is called **dipole-dipole interaction**, and it is 100 times weaker than ionic interactions and falls off quickly with the distance $r$ following the function $\frac{1}{r^3}$. Specifically, water molecules have a particular case of dipole-dipole interaction that, in general, concerns highly electronegative atoms connected to hydrogen atoms such as fluorine $F$, nitrogen $N$ and oxygen $O$. In this case, attractions between dipoles are powerful so as to be called **hydrogen bonds**. Hydrogen atoms create proper "bridges" between a molecule $A$ and the most electronegative atom of molecule $B$ thanks to their high positive charge and align all molecules in a crystalline network. This type of bond does not affect the chemical properties of a compound, but it influences the physical properties such as the boiling point, density and solubility.

Polar molecules consist of dipole moments that divide them into two parts. Even in apolar compounds, such as iodine $I_2$ crystal, which are not subject to a high difference in the dipole charges (symmetric distribution), various attractive forces may exist. Assuming electrons move around continuously, their distribution may be altered, causing a small dipole moment that will alter the electronic distribution of molecules in the neighbourhood (this effect is called **induced dipole**). This type of weak connection between diatomic molecules belongs to the non-bonded interaction class referred to as **van der Waals forces** that are dispersion forces which drop off rapidly with the distance $r$, following mostly a leading term close to $\frac{1}{r^6}$.

$$ {}^{\delta^-}\mathrm{I} \text{——} \mathrm{I}^{\delta^+} \quad - - - - \quad {}^{\delta^-}\mathrm{I} \text{——} \mathrm{I}^{\delta^+} \tag{1.13}$$

### 1.2.6 Atomic radii

We cannot measure the correct size of atoms and find the exact position of an electron that behaves like a wave. Although electrons are very difficult to spot, atoms' nuclei in compounds or elements are found at proper distances from each other and are characterized by a specific atomic radius, where they apply their nuclear force. When we deal with covalent bonds, reaction forces are stronger and keep atoms at a certain distance depending on the type of atoms involved in the interaction. According to this bond type, a **covalent radius $r_c$** is assigned to the atoms, and it is defined as half the inter-atomic distance between two singly bonded atoms (or pure elements). For diatomic molecules such as iodine $I_2$, there is no problem in computing the nuclei's distance since they form a molecule with a single bond. However, for elements that do not have diatomic molecule forms, such as carbon $C$–$C$ (e.g. they exist only in larger compounds), an average value is calculated from a range of compounds containing it.

Ionic solids are formed by ionic bonding, where atoms are arranged in a crystalline network and retain different inter-nuclear distances from the usual atomic radii that are called **ionic radii**. According to ionic bond rules, atoms modify their usual atomic radius by increasing or reducing the inter-atomic distance. Indeed, when an atom as $Na$ forms a cation $Na^+$ in sodium chloride $NaCl$, it gives an electron to chlorine $Cl$ reducing its ionic radius (because an electron moved away). As a consequence, chlorine accepts the electron and increases its atomic radius. Ions regularly pack together in crystals. Therefore, their inter-atomic distance can be measured accurately with an average from many crystal structures using X-rays Crystallography and electron density maps.

Furthermore, unlike the covalent radius, **Van der Waals radius** is defined as a non-bonded distance of the closest approach, and it is calculated from the smallest inter-atomic distances in crystal structures that are not bonded to one other. These values are computed as averages compiled from many crystal structures. For example, when bonded, the iodine atom $I$ has a covalent radius of 1.33 Å. On the other end, if not bonded, it has a van der Waals radius of 1.98 Å. It means that, in theory, atoms within 1.98 Å from iodine that are not bonded, should be considered in computing chemical features. Therefore, knowing these experimental values of **bond lengths**, it is easier to treat or identify bonds computationally.

## 1.3 Solids

According to standard conditions (e.g. room temperature and pressure), the matter may be found in different states such as solid, liquid or gas. In the liquid phase, molecules are close together and arranged randomly, occupying a fixed volume but not a fixed shape, such as the water $H_2O$ adapting to different containers. Gaseous substances are in a fluid form as well, where molecules are far apart from each other and spread around by filling all the available space and adapting to the shape of a container, such as oxygen $O_2$ or chlorine $Cl_2$ gases. A **solid** is a rigid form that can arrange in repeating arrays or layers and extend to infinity. In this document we will focus on these perfectly ordered structures, although it is also possible to have non repeating patterns such as in the glass structure. Atoms, ions or molecules in a solid phase of the matter vibrate around their positions and cannot easily move due to the strong interactions between them given by the perfect arrangement with specific distances between atoms. Nevertheless, the temperature is a good condition to use to provide the system with energy by increasing it and pushing the system to instability. Indeed, atoms, ions or molecules vibrate in their position, and these vibrations become intense as the temperature rises, causing a **phase change** which is the transition phase that brings matter from a particular state to another. For example, if we provide a block of ice with heat (just considering a room temperature of 25°), hydrogen bonds become weaker and weaker, allowing the block to melt down, moving from a solid phase to a liquid phase. Moreover, given a sodium crystal $Na$ mixed with some water, sodium melts down at standard conditions. It is a highly reactive metal and changes its phase from solid to liquid, making bonds with hydrogen and oxygen atoms. On the other hand, many substances may reach a solid phase making tighter bonds when environmental conditions change, such as the same water which freezes to **ice** below 0° C after fixing and aligning hydrogen bonds at specific distances.

Furthermore, many simple crystalline solids exist or can be prepared from compounds at room temperature and pressure such as iodine $I_2$, sodium *Na*, lithium *Li* or iron *Fe*. *Iodine* $I_2$, as the ice, is a **molecular solid** which means that covalent bonds connect iodine atoms, but iodine molecules are linked together by weaker non-covalent forces such as van der Waals and hydrogen bond forces. Due to the weaker forces, iodine easily evaporates, passing to gaseous phase and releasing iodine vapour if a small amount of heat is provided to the system.

*Sodium*, *lithium* and *iron* are metals formed by metallic bonds, which may arrange in an infinite **metallic network structure**. It is held up by electromagnetic forces between cations of the same type and surrounded by a sea of electrons that provide them with the electricity conduction property. This property occurs for the metal groups of the periodic table (I, II, III and transition metals) because all valence electrons of the atoms are delocalized and can move freely around. Most of these metals exist in a solid state except for mercury, which is liquid at room temperature.

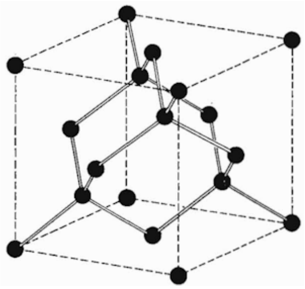


Figure 1.1: [61] Fig. 10. Crystal structure of diamond.

In addition, **ionic solids** usually tend to form between elements on the far left of the periodic table with groups I or II, such as *Li* and *Na* that prefer to remove electrons and form cations. They pair up with elements on the far right with groups VI or VII that form anions and need to reach a stable configuration as the closest noble gas. Elements in the other first groups (III or IV) may also be used, but the reaction becomes harder because they need more potent ionization energies as the number of electrons to be removed increases for making subsequent ions.

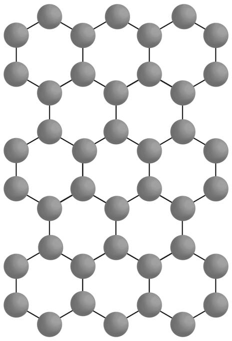


Figure 1.2: [61] Fig. 10. Crystal structure of graphite.

Ionic bonds are non-directional, meaning that their strength decreases with the increasing separation of the ions that pack together to maximize the Coulomb attraction and minimize repulsions. An example of an ionic solid is the sodium chloride $NaCl$ made of an ion of group I ($Na^+$) and an ion of group VII ($Cl^-$).

Many types of structures are found in different non-metal elements and their compounds (right part of the periodic table), including carbon $C$. They are called **covalent network structures** and are formed by covalent bonds, which are stronger than ionic forces. An example is the **diamond** in Figure 1.1, a carbon form structure very rare in nature. Each carbon in the diamond is equivalent and connected to precisely four atoms at a distance of 1.54 Å because four is the maximum number of bonds that a carbon atom can make. Moreover, it forms a big molecule with all carbons connected, and these bonds make it the hardest substance known. Another form of carbon may be described with the **graphite** in Figure 1.2, a soft grey solid that consists of 2-dimensional layers of atoms joined together, where each carbon atom $C$ makes only three bonds with the carbons in the surroundings within a layer, fixing at a distance of 1.42 Å in a hexagonal shape. It is a soft metal because layers are kept closer by weak van der Waals forces with a distance of 3.4 Å from each other. Carbon can form more than one structure, indeed diamond and graphite are called **allotropes** because they are different structures of the same element. All these examples of crystalline materials have a regular structure consisting of a small repeating unit or particles that can build up the entire crystal. The underlying periodic structure where a unit is repeating is called a **lattice**.

## 1.4 Energy changes and disorder

### 1.4.1 Energy changes

Energy is used in everyday life, and indeed, each action we perform requires energy to be completed. During the energy release, work is done, and it involves motion occurring against a force. Even in elements reactions, work is performed by taking in or releasing energy and consequently, the environment is affected by being cooled down or heated up. When a reaction produces energy and heats the surroundings, it is called **exothermic** where the product resulting from it gains less energy than the reactants. On the other hand, if a reaction needs energy to be performed, it takes in energy by cooling down the surroundings, and it is called **endothermic** reaction. It is difficult to measure the energy transferred between a reaction and the surroundings. However, we can only measure the change in energy at constant pressure called **enthalpy change** $\Delta_r H$ where enthalpy $H$ is the heat transferred and $r$ is the reaction type among different processes (e.g. crystal formation identified by $f$).

$$\Delta_r H = H_{products} - H_{reactants} \tag{1.14}$$

Exothermic reactions have a negative value of the enthalpy change addressing the fact that energy is lost to the surroundings. On the contrary, the value of $\Delta_r H$ is positive for endothermic changes because the system gains energy from the surroundings. Usually, standard conditions and temperature in kelvin at which they occur are specified respectively by the superscript symbol $^{\ominus}$ and a numerical subscript as in $\Delta_r H^{\ominus}_{298}$. For example, for the reaction of sodium $Na$ and chlorine $Cl$ we can use the following **thermochemical equation** to describe the beginning of the process, which links the enthalpy change to the molar amount of elements:

$$Na(s) + Cl_2(g) \rightarrow NaCl(s)$$

In the thermochemical equation, we can notice that, in nature, chlorine gas $Cl_2$ exists as an element which is paired up into a molecule of two chlorine atoms. During a reaction, masses should be preserved as the **mass conservation law** states, and therefore the mole of atoms of reactants should be the same in the product. Indeed, in the product of the equation, only one atom of chlorine is counted, and consequently, the thermochemical equation is said to be unbalanced. To find the **correct balance**, we need to match the mole of atoms on both sides of the equation as follows:

$$Na(s) + \frac{1}{2}\, Cl_2(g) \rightarrow NaCl(s) \qquad \Delta_f H^{\ominus}_{298} = -411.12\, \frac{kJ}{mol} \tag{1.15}$$

where 1 mole of sodium solid $Na$ reacts with $\frac{1}{2}$ the mole of $Cl_2$ gas to form 1 mole of sodium chloride solid $NaCl$, and throughout the process 411.12 kJ are transferred to heat the surroundings for crystal formation ($f$).

After a reaction, chemical bonds are broken, atoms rearranged and new bonds are made in the final product releasing or taking in energy measured in $\frac{kJ}{mol}$. When bonds break, an input of energy is taken in to separate the atoms or ions and disrupt the attractive electrostatic forces. So the process is called endothermic because it cools down the environment absorbing energy from it. Conversely, when a bond is formed, energy is released by heating the surroundings. Finally, the overall change in enthalpy is determined to be endothermic or exothermic by the difference in energy between the bond-breaking and bond-forming processes as defined in equation 1.14.

During the phase of bond breaking, the system absorbs energy allowing the atoms to divide depending on their bond strength. To measure the bond strength, the **bond dissociation enthalpy $\boldsymbol{D}$** or $\boldsymbol{\Delta_{diss}H}$ can be used. It is defined as the standard enthalpy change for the reaction in which a bond is broken. For some homonuclear diatomic molecules, we have different dissociation enthalpies that are summarized below:

$$\begin{aligned}
H_2(g) \rightarrow 2H(g) \qquad & \Delta_{diss}H = D_{(H-H)} = +435.8\, \frac{kJ}{mol} \\
O_2(g) \rightarrow 2O(g) \qquad & \Delta_{diss}H = D_{(O-O)} = +498.5\, \frac{kJ}{mol} \\
I_2(g) \rightarrow 2I(g) \qquad & \Delta_{diss}H = D_{(I-I)} = +152.3\, \frac{kJ}{mol} \\
Cl_2(g) \rightarrow 2Cl(g) \qquad & \Delta_{diss}H = D_{(Cl-Cl)} = +242.4\, \frac{kJ}{mol}
\end{aligned}$$

where values of dissociation energy change are positive because bond breaking requires to take in energy in the system (molecule) from the environment.

When a reaction process is ongoing, the reactants may need an amount of energy that allow them to break bonds and consequently form new ones. It could be given by simply providing the system with more heat. We say that during a reaction, reactants need to pass the energy barrier or **activation enthalpy** to start breaking bonds and approach to the formation of the product when the enthalpy of the reacting system gradually decreases. Under normal conditions such as room temperature and standard pressure, some reactions can occur without providing energy because they have small energy barriers. For instance, sodium $Na$ and chlorine $Cl$ cannot react without providing heat to the system, which is performed by adding some water and let the sodium mix to get hotter. Therefore, the chlorine gas molecules $Cl_2$ can react with hot sodium to form sodium chloride $NaCl$ or table salt, unleashing a very bright red light that confirms the exothermic reaction and, if no precautions are taken, the reaction force may break the bottle where the elements are contained.

In order to form a crystal, bonds should be broken and then formed to lead to the product. As the **Hess's law** states, the total enthalpy change for a chemical reaction is independent of the path by which the reaction occurs, provided the starting and ending states are the same for each reaction path. In particular, it does not matter what is the path to form a crystal; the energy released will not change. For instance, given the experimental standard enthalpy change of the sodium chloride formation $\Delta_f H^{\ominus}(NaCl)$ in 1.15, whatever are the states of the reactants, the enthalpy energy of formation will not change [9].

### 1.4.2 Entropy

It is not entirely true that a solid have always an infinite repeating pattern. The repeating structure can be interrupted at some point breaking the periodicity of a crystal structure. This is the case of disordered solids such as the **glass** that we find in our windows or bottles. It is formed from heating up a viscous liquid made by sand and other compounds, and then cooling it down. A glass material is characterized by brittleness, transparency, heat-resistance, and belongs to the class of **amorphous** solids, which may retain in parts the properties of a periodic solid and the molecular randomness of a liquid. These type of solids forms from a disordered system that did not succeed in building the final structure in which all atoms converge into an ordered pattern. One way of looking at the disorder is to consider what is happening to the organisation of molecules in the system and how the energy is exchanged. The factor that affect the disorder of a system is called **entropy** and is given by the symbol $S$. Entropy is a state function and therefore its change is given by the difference between the final and initial states of a system

$$\Delta S = S_{products} - S_{reagents} \tag{1.16}$$

The higher is the number of microstates the more disordered is a macroscopic system. These microstates are the set of different molecules arrangements that form the system. The entropy is a measure of randomness or disorder in a system and can be represented by the Boltzmann formula below:

$$S = k_B \ln W \tag{1.17}$$

where $W$ is the of ways of arranging molecules and their energies and $k_B = 1.381 \times 10^{-23} \frac{J}{K}$ is the Boltzmann constant. Another measure of the entropy change is to considered the energy exchange under a certain temperature at equilibrium (when the reaction has reached a "stable" state).

$$\Delta S = \frac{q}{T} \tag{1.18}$$

where $q = \Delta H$ is the enthalpy change at constant pressure and $T$ is the temperature. Every isolated system tends spontaneously to increase its disorder, reaching the most probable state. An increasing value in $\Delta S$ refers to a transition from an ordered state to a disordered one (from solid to gaseous). Vice versa, a decreasing value in $\Delta S$ defines the inverse process.

Since our experiments are based on perfectly ordered structures, we will refer only to periodic crystal structures.

## 1.5 Atom packing

A lattice highlights the perfect arrangement of particles in a crystal structure that can be aligned on different layers. Indeed, a unit of particles should be repeated in all directions to build the entire structure, and consequently, atoms, ions or molecules may be arranged in layers that superimpose on each other. Many crystals adopt structures based on layer superimposition. How can we model atoms to reach the best superimposition? According to the wave mechanics model, atoms are formed by a nucleus in the centre and electrons spin around different types of orbital. The more an electron is far away from the centre, the less ionization energy is required to remove it. Therefore the atom structure resembles a sphere where the nuclear force fades out concerning the distance from the nucleus (radius). Once we set a theoretical atom structure, we can start building layers of hard spheres and see how they pack together, keeping in mind the principal aim of minimizing the empty space. This type of arrangement on different layers is defined as **close packing**. Sphere packing starts from the first horizontal layer A where spheres are aligned and placed to minimize the empty space by letting each touch six neighbours. Consequently, a second layer B is located on top of it, where each sphere occupies a depression between the spheres of the first layer. When the first two layers are located on top of each other, two different types of hole open in between, occupied by a third layer. Regarding the third layer, since there are two types of depressions formed by superimposing AB layers, we can place the third one in two ways: the first one (ABC), could be found by placing hard spheres of a layer C above the spaces of the first layer A (**cubic close packing** or ccp), and the second one (ABA) consists of another layer A directly above the first one (**hexagonal close packing** or hcp). There are many different types of layer arrangements such as ABCAB or ABCB, but ccp and hcp are the most common used by the majority of metals and noble gases: copper *Cu* and aluminium *Al* have a cubic close-packed structure, magnesium *Mg* and zinc *Zn* adopt the hexagonal close packing (Figure 1.3). After the third layer is added, we can count a total number of neighbour spheres equal to 12 for each one, where 12 is called **coordination number**. Ccp and Hcp structures have coordination number equal to 12 because each sphere of layer B touches: 6 spheres on the same layer B, 3 spheres on layer A and 3 spheres on layer C (or A). The structures of close packing represent the most efficient way to pack spheres or atoms, indeed, spheres occupy 74% of the space, and this percentage value is said **packing efficiency**.

There are different types of cubic structures, and all of them differ from each other by a minor detail. When dealing with a crystal structure, we refer to an object with a particular shape containing all particles and defining the smallest possible set of repeating units. This object is called **unit cell**. In cubic unit cells (Figure 1.4), if atoms, ions or molecules are placed on the vertices and in the centre of each face, we define a **face-centred cubic** unit cell or fcc, which is adopted by metals like calcium *Ca* or aluminium *Al*. Particles can also lie at the centre of the unit cell with no atom in the faces' centres, and in this case, it will be a **body-centred cubic** or bcc structure used by lithium *Li*, sodium *Na* and iron *Fe*. Moreover, the simplest of the cubic types is the **primitive cubic** where all particles are located on the only vertices leaving the empty space inside. Therefore, the face-centred type occupies all possible space of a unit cell with coordination number 12 and could also

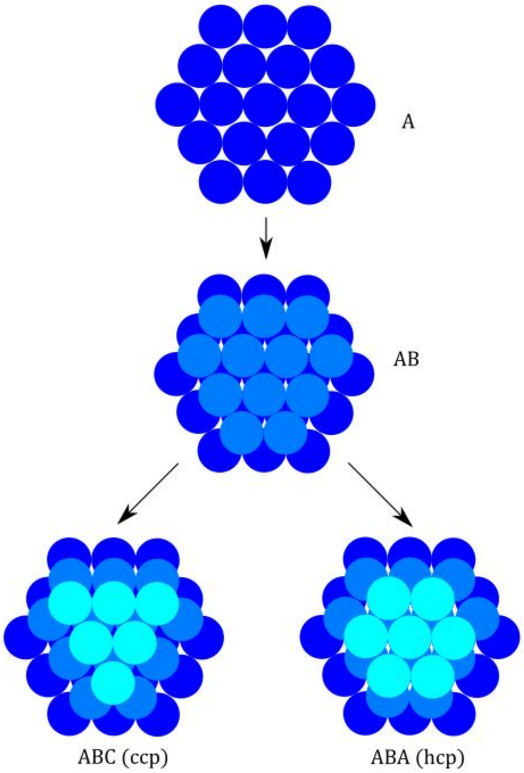


Figure 1.3: Cubic and hexagonal close packing.

be called a close-packed structure. In contrast, body-centred and primitive types contain bigger empty spaces and their packing efficiency drops from 74% of the face-centred to 68% and 52.4% with coordination numbers 8 and 6 respectively [50].

## 1.6 Crystal Structure

Crystals are solid crystalline materials formed by atoms, ions or molecules. They consist of an underlying periodic structure called a **lattice $\mathbf{\Lambda}$** and a set of particles (atoms, ions or molecules) called a **motif M** which is repeated at each lattice point (Figure 1.5).

**Definition 1.1.** (**lattice $\mathbf{\Lambda}$**). *More formally, given a linear basis of n vectors* $\vec{v}_1, \vec{v}_2, ...\vec{v}_n$,

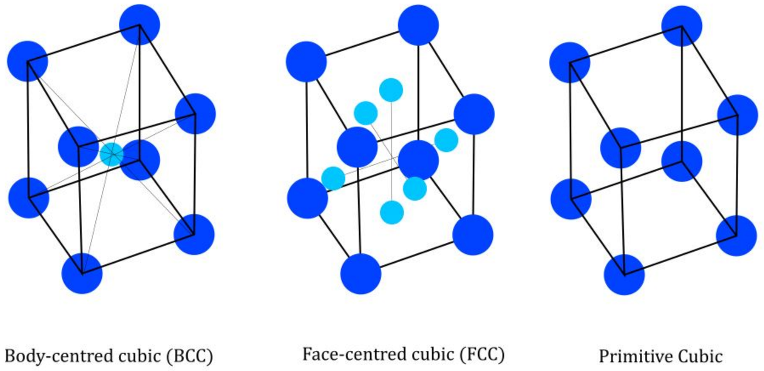


Figure 1.4: The unit cells of body-centred, face-centred and primitive cubic lattices.

*a lattice in* $\mathbb{R}^n$ *is the discrete set of their linear combinations with integer coefficients* $t_i$*:*

$$\Lambda = \{\sum_{i=1}^{n} t_i \vec{v}_i \in \mathbb{R}^n \mid t_i \in \mathbb{Z}\} \tag{1.19}$$

It is an infinite arrangement where all points are equivalent in the sense that each of them can be found by translation along the three axes. Moreover, they define a place in the space where the repeating units lie. In metallic bonded structures such as sodium or iron, each *lattice* point is occupied by a set of metal atoms or ions. Their unit cell follows a body-centred cubic structure with two lattice points in total because 1 point is shared between 8 unit cells ($\frac{1}{8}$) and altogether account for 1, plus 1 point in the centre of the cube. Therefore, the repeating unit is formed by two atoms or ions. The set of lattice points is

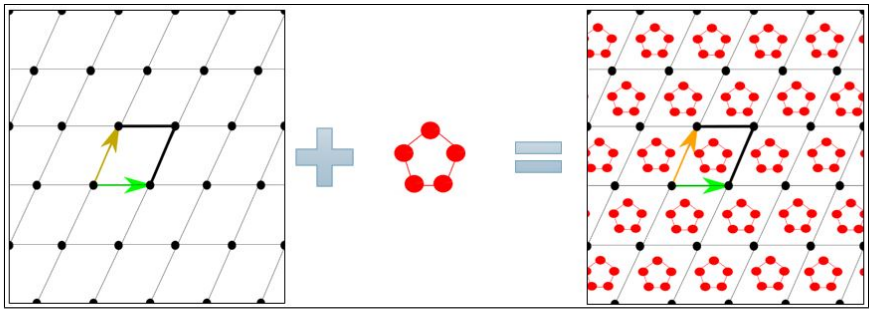


Figure 1.5: A crystal structure made of a lattice plus a motif.

used to simplify the periodic pattern of a structure (underlying periodic structure). Indeed they do not retain any information about chemical bonds. To add chemical information, we need to consider atoms and their positions which repeat at each lattice point.

**Definition 1.2.** (**unit cell $\boldsymbol{U}$, motif $\boldsymbol{M}$**). *A unit cell $U$ is a parallelepiped spanned by the linear basis $\vec{v}_1, \vec{v}_2, ...\vec{v}_n$, and containing a finite set of points $\vec{p}_1, \vec{p}_2, ...\vec{p}_m \in M$. All points of a motif can be addressed by the linear combinations with real coefficients in the interval $r_i \in [0,1]$:*

$$U = \{\sum_{i=1}^{n} r_i\vec{v}_i \in \mathbb{R}^n \mid r_i \in [0,1]\} \tag{1.20}$$

The unit cell is periodically repeated by translations along the unit cell vectors $\vec{v}_i$ to span the infinite crystal structure.

**Definition 1.3.** (**periodic point set $\boldsymbol{S}$**). *Generally, a* periodic point set $S \subset \mathbb{R}^n$ *is the* Minkowski sum $S = \Lambda + M = \{\vec{u} + \vec{v} : u \in \Lambda, v \in M\}$*, so $S$ is a finite union of translates of the lattice $\Lambda$.*

A unit cell $U$ is *primitive* when it contains the smallest part of a crystal structure. Since every smallest part (molecules, atoms or ions) can be repeated at each lattice point, no any other lattice point must be included in the unit cell to be primitive (see Figure 1.4 for the primitive cubic unit cell). An example of periodic set is a lattice $\Lambda$ plus a 1-point motif $M = \{p\}$. The motif point $p$ (smallest part) repeats at each lattice point overlapping it. Each point can be found by translating the other, and so $p$ can be any point in a unit cell $U$. Examples of periodic sets are given in Fig. 1.6a-b that shows equivalent (or isometric) square lattices. Whereas the periodic sets in the bottom part of Fig. 1.6c-d represent isometric hexagonal lattices, because every black point has exactly six nearest neighbours that form a regular hexagon. A lattice $\Lambda$ of a periodic set $S = M + \Lambda \subset \mathbb{R}^n$ is not unique, in fact, $S$ can be generated by different unit cell vectors that contains a motif larger than $M$. This results in a non-primitive unit cell that is bigger. For instance, imagine to double the vectors in Figure 1.5, and consequently, four repeating units can be clearly identified inside the new bigger parallelogram, together with other lattice points included.

In the previous paragraphs we dealt with cubic unit cell types, but, usually, crystals could be also described by non-cubic unit cells where linear bases have angles different from 90° degrees and different bases lengths. Angles and vector lengths are called **unit cell**

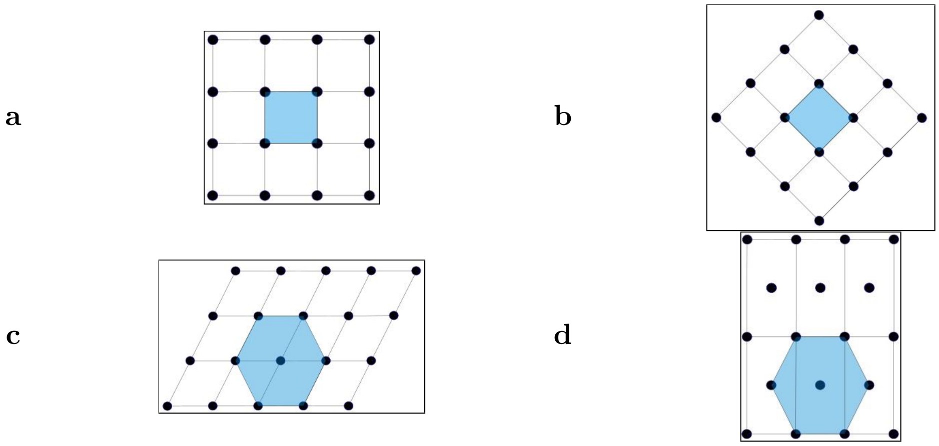


Figure 1.6: Equivalent square lattices (a-b) where b has been transformed by a rigid motion with respect to a. Equivalent hexagonal lattices (c-d) where d has different bases with respect to c.

**parameters** which form the parallelepiped containing the set of particles, and are stored in the **crystallographic information file (CIF)** together with other structural details and chemical information.

## 1.7 Crystal systems

In the previous sections cubic unit cells were mentioned to explain that most of the metallic network structures adopt this type to form a crystal structure of ions which allow the electricity conduction. Cubic unit cells are said to belong to the **7 crystal systems** together with other 6 classes. The unit cell is defined by a parallelepiped and it is stored in the CIF file with 3 lengths $a$, $b$, $c$ and 3 angles $\alpha$, $\beta$, $\gamma$ (see Section 1.10). When $a = b = c$ and $\alpha = \beta = \gamma = 90°$ it belongs to the *cubic* crystal system and all the other possible values for lengths and angles characterize the remaining classes as it is summarized in table 1.1. For example, the hexagonal crystal system, which follows an atom packing of ABA layers, is adopted by metallic networks such as zinc *Zn*, magnesium *Mg* and titanium *Ti*, holding angles $\alpha = \beta = 90°$ and $\gamma = 120°$.

Each crystal system can have at most four types of unit cells. In Section 1.5, we dealt with atom packing and three types of unit cells of the cubic crystal system: face-centred, body-centred and primitive cubic unit cells. Crystal systems may also be formed by a

| System | Lengths | Angles |
|---|---|---|
| Cubic | $a = b = c$ | $\alpha = \beta = \gamma = 90°$ |
| Tetragonal | $a = b \neq c$ | $\alpha = \beta = \gamma = 90°$ |
| Orthorombic | $a \neq b \neq c$ | $\alpha = \beta = \gamma = 90°$ |
| Hexagonal | $a = b \neq c$ | $\alpha = \beta = 90°$, $\gamma = 120°$ |
| Trigonal | $a = b = c$ | $\alpha = \beta = \gamma \neq 90°$ |
| Monoclinic | $a \neq b \neq c$ | $\alpha = \gamma = 90°$, $\beta \neq 90°$ |
| Triclinic | $a \neq b \neq c$ | $\alpha \neq \beta \neq \gamma \neq 90°$ |

Table 1.1: Features of all 7 crystal systems.

fourth type that belongs to the face-centred type but holds lattice points only in two of the face centres. It gains its name from symbols A, B and C depending on the two faces where lattice points lie. For example, a B-type face-centred unit cell has a lattice point at each corner and two lattice points in two parallel faces' centres perpendicular to vector $\vec{b}$. All 4 types are shown in Table 1.2. Crystal systems and unit cell types form a total of 14 lattices called **Bravais lattices**.

| Type | Symbol | Lattice point positions |
|---|---|---|
| Primitive | **P** | Each corner |
| Body-centred | **I** | Each corner and parallelepiped's centre |
| Face-centred | **F** | Each corner and each faces' centres |
| Face-centred | **A, B, C** | Each corner and two parallel faces' centres |

Table 1.2: Features of all 4 unit cell types.

## 1.8 Crystal Packing and Density

### 1.8.1 Packing efficiency

A crystal structure of a metal tends to form a metallic network where atoms of the same type pack together lying at specific distances from each other to support electrical conductivity. The Close-packed structure is the most efficient way to pack spheres in space. The property that addresses how good atoms are packed is called **packing efficiency**, defined as follows:

$$pe = \frac{\sum_{i=1}^{n} V(a_i)}{V(U)} 100 \tag{1.21}$$

where $n$ is the number of atoms in the unit cell $U$, *V()* is the function that outputs the volume, $a_i$ is the *i-th* atom in the unit cell $U$, and the result is multiplied by 100 to get the percentage value. Since the volume of the unit cell *V(U)* usually is higher than the sum of all atoms' volumes the result belongs to the range [0, 100].

To compute the numerator of the packing efficiency, we need to check what atoms we should consider in the unit cell. Indeed, some atoms could belong to different repeated unit cells in the crystal structure. Therefore we need to assign a weight for each of them as follows: $\frac{1}{x}$ where $x$ is the number of unit cells that share that particular atom or sphere. For example, in the close-packed cubic (or face-centred cubic) structure of calcium crystal *Ca* there are 8 atoms on the vertices plus the other 6 at the face centres. In the repeated unit cell domain, each atom on the vertices is shared with 8 unit cells and therefore, it should be accounted with weight $\frac{1}{8}$, whereas each atom on the face centres is shared with 2 unit cells and should be accounted with weight $\frac{1}{2}$. Once we have identified the count values, we can find the **effective number of atoms** in the unit cell:

$$n_{fcc}(number\ of\ atoms\ in\ fcc) = \frac{1}{8}\,8 + \frac{1}{2}\,6 = 4$$

Regarding the other cubic structures, the number of atoms changes reporting a different number of shared atoms as shown below:

$$n_{bcc}(number\ of\ atoms\ in\ bcc) = \frac{1}{8}\,8 + 1 = 2$$
$$n_{p}(number\ of\ atoms\ in\ primitive) = \frac{1}{8}\,8 = 1$$

where *bcc* has 1 atom inside that belongs only to 1 unit cell and the *primitive* structure has atoms only at each lattice point (or cube vertex). Given the effective number of atoms in a unit cell, we need to compute the **volume of an atom** which is modelled as a sphere with radius $r$ and volume $\frac{4}{3}\pi r^3$. Once we can compute the numerator, we need to find the **volume of a cubic close-packed unit cell (fcc)** to solve the packing efficiency for metallic network structures. Suppose we do not know the side length $l$ of the cube. In that case, we can find it in function of the sphere (atom) radius by using the Pythagora's theorem and knowing that each fcc unit cell's face has a diagonal of length $d = 4r$ where each sphere in the face centre is touching other 4 neighbours. Therefore, the side length $l$ can be found as follows:

$$l^2 + l^2 = d^2 \iff l^2 + l^2 = (4r)^2 \iff l = 8^{1/2}r$$

from which the volume of the cubic unit cell is equal to:

$$V(U_{fcc}) = l^3 = (8^{1/2}r)^3 = 8^{3/2}r^3$$

The packing efficiency of a close-packed (fcc) unit cell is shown below:

$$pe_{fcc} = \frac{\sum_{i=1}^{n_{fcc}} V(a_i)}{V(U_{fcc})} 100 = \frac{n_{fcc}\,\frac{4}{3}\pi r^3}{8^{3/2}r^3} 100 = \frac{\frac{16}{3}\pi r^3}{8^{3/2}r^3} 100 = 74.0\%$$

where $\sum_{i=1}^{n_{fcc}} V(a_i)$ is the sum of all atom volumes.
When the unit cell is not represented by a cubic shape, it could be easier to calculate the packing efficiency by relying on the linear bases of the unit cell by the following formula:

$$pe = \frac{\sum_{i=1}^{n_{fcc}} V(a_i)}{V(U)} 100$$

where the volume of the non-cubic unit cell $V(U)$ may be computed with the absolute value of the cross product and the scalar product between unit cell vectors:

$$V(U) = |(v_1 \times v_2) \cdot v_3| \tag{1.22}$$

### 1.8.2 Relative atomic mass and Crystal Density

To compare atomic masses, a relative scale is used. The standard scale is a single atom of carbon-12 ($^{12}C$) of which the **relative atomic mass** ($A_r$) is defined as a value equal to 12. All the other atom masses are considered with respect to this reference value, which has no unit. Starting from the notion that exactly 12 grams of $^{12}C$ contain 6.022 x $10^{23}$ atoms, every other element or compound that has the same weight contains the same number of atoms (**mole of atoms**). The number of atoms per mole is referred to as the **Avogadro constant,** $N_A$. Practically, the unit of the amount of substance is the mole (mol) that always contains 6.022 x $10^{23}$ entities ($N_A$). The relative atomic mass is used to compare masses of atoms. However, if we want to consider a comparison between compounds we need to compute the **relative formula mass** given by the symbol $M_r$ (if discrete molecules are involved such as the benzene molecule $C_6H_6$ the relative formula mass is called **relative molecular mass**). Take the previous example of the common salt sodium chloride $NaCl$ that is formed by two ions arranged in a network structure. Given the relative atomic mass of sodium $A_r(Na) = 22.989$ and chlorine $A_r(Cl) = 35.453$, the relative formula mass is:

$$M_r(NaCl) = (1 \cdot 22.989 + 1 \cdot 35.453) = 58.442 \tag{1.23}$$

When dealing with compounds chemists want to know the amounts of entities that they contain (could be atoms, molecules or ions) since reactions occur between them. Therefore, chemical amounts are based on the concept that the mass of one mole (the **molar mass**, $M$) is equal to the relative atomic mass, relative formula mass or relative molecular mass in grams per mole ($\frac{g}{mol}$). For example, given $^{12}C$ and its relative atomic mass $A_r(^{12}C) = 12.011$, its molar mass is exactly $M(^{12}C) = 12\frac{g}{mol}$. Moreover, if we consider the relative formula mass of the sodium chloride at (1.23), its molar mass will be $M(NaCl) = 58.442\frac{g}{mol}$.

Given 12 grams of common salt or sodium chloride, can we estimate the number of ions contained? The answer is 'yes, we can', indeed the amount of substance (in mol) is computed by the following equation:

$$amount\ (in\ mol) = \frac{mass\ (in\ g)}{molar\ mass\ (in\ \frac{g}{mol})} = \frac{12g}{58.442\frac{g}{mol}} = 0.205\ mol \tag{1.24}$$

It means that there are at least $\frac{1}{5}$ entities (ions in our case) of a mole contained in 12 grams of sodium chloride. The result makes sense because the sum of sodium and chlorine relative atomic masses is 5 times bigger than the reference $^{12}C$ mass, therefore in 12 grams they will contain a less number of entities or ions for 1 mole, which is indeed equal to the number of ions ($N_A$) in 12g of $^{12}C$.

Most of the time, when dealing with crystal structures, we already know the content of the unit cell, such as atom positions and their types. Since atoms in compounds are various and have different atomic radii, it is hard to find the side length in terms of the atomic radii and compute the packing efficiency. Therefore, we can use another measure called **density** ($\boldsymbol{\rho}$) that is defined as the mass per unit volume. It explains how dense atoms or molecules are packed together, and it is measured in $\frac{g}{cm^3}$ where values close to 0 mean that there are many empty spaces:

$$\rho = \frac{m}{V(U)} \tag{1.25}$$

where $V$ is the volume of the unit cell and $m$ is the sum of all molar masses (or relative atomic masses) described in Section 1.8.2, which should be divided by the Avogadro constant $N_A$ to find the mass of the atoms out of 1 mole. For instance, calcium crystal is formed by a close-packed or fcc cubic structure, and this means that we count 4 atoms in total that belong to the unit cell, therefore $m$ is given by:

$$m = \frac{4 \times A_r(Ca)}{N_A}$$

Once we have the mass $m$ and the volume of the fcc unit cell $V(U_{fcc})$, the density is computed as follows:

$$\rho = \frac{m}{V(U_{fcc})} = \frac{4 \times A_r(Ca)}{N_A \, 8^{3/2} r^3} = \frac{4 \times 40.078 \; g \; mol^{-1}}{(6.022 \times 10^{23} \; mol^{-1}) \times 8^{3/2} \times (1.97 \times 10^{-10} m)^3} = 5.972 \frac{g}{cm^3}$$

where 40.078 $\frac{g}{mol}$ is the relative atomic mass of calcium atom and 1.97 Å is its atomic radius. If we consider a different crystal structure characterized by a non-cubic unit cell, the volume can be computed with equation 1.22.

Density is a valuable property that can be applied to find the level of porosity in materials and acts as a threshold measure when a range of structures with a specific porosity level should be explored for further stability assessments (see chapter 3.2).

## 1.9 Lattice Energy

### 1.9.1 Experimental lattice energy

To form a solid, bond breaking and formation processes are involved. Therefore, an exchange in energy occurs between the system and the surroundings. To measure the energy used for lattice formation, we refer to the **lattice energy $\boldsymbol{\Delta_{latt} U}$** that is the negative of the internal energy change of the system where all potential and kinetic energies are summed up:

$$\Delta U = \Delta_{latt} H - p \Delta V \tag{1.26}$$

$$\Delta_{latt}U = -\Delta U \tag{1.27}$$

where $p$ is the pressure constant, $\Delta V$ is the change in volume and $\Delta_{latt}H$ is the lattice enthalpy. The **lattice enthalpy $\boldsymbol{\Delta_{latt}H}$** is the enthalpy change used for the conversion of 1 mole of the ionic solid into the gaseous ions:

$$NaCl(s) \rightarrow Na^{+}(g) + Cl^{-}(g) \qquad \Delta_{latt}H^{\ominus}(NaCl) \tag{1.28}$$

To compute the *lattice enthalpy* $\Delta_{latt}H$, we need to build a formation cycle called **Born-Haber cycle** that describes the enthalpies needed to lead the reactants to their gaseous state ions. The *Born-Haber cycle* for sodium chloride *NaCl*, shown in Figure 1.7, occurs following different steps starting from the balanced thermochemical equation:

$$Na(s) + \frac{1}{2}\,Cl_2(g) \rightarrow NaCl(s) \tag{1.29}$$

1. Sodium should be converted to its gaseous form and chlorine bond has to be cleaved to release one atom of chlorine.

   $$\begin{aligned} Na(s) \rightarrow Na(g) \qquad & \Delta_a H^{\ominus}(Na) = +108\,\frac{kJ}{mol} \\ \frac{1}{2}Cl_2(g) \rightarrow Cl(g) \qquad & \Delta_a H \ominus (Cl) = +121\,\frac{kJ}{mol} \end{aligned}$$

   where $a$ is the *atomization process* to get gaseous atoms.

2. Gaseous atoms should be then ionized to form a sodium cation and a chlorine anion.

   $$\begin{aligned} Na(g) \rightarrow Na^{+}(g) + e^{-} \qquad & \Delta_i H(1)^{\ominus}(Na) = +496\,\frac{kJ}{mol} \\ Cl(g) + e^{-} \rightarrow Cl^{-}(g) \qquad & \Delta_{eg} H(1)^{\ominus}(Cl) = -349\,\frac{kJ}{mol} \end{aligned}$$

   where $i$ is the *ionization process*, $\Delta_i H(1)^{\ominus}(Na)$ is the first (1) ionization energy, *eg* is the *electron gain process* and $\Delta_{eg}H(1)^{\ominus}(Cl)$ is the energy needed to acquire the first (1) electron.

3. Finally, gaseous ions are combined to form the solid state structure.

$$Na^{+}(g) + Cl^{-}(g) \rightarrow NaCl(s) \qquad -\Delta_{latt}H^{\ominus}(NaCl) \tag{1.30}$$

where $-\Delta_{latt}H^{\ominus}(NaCl)$ is the negative of the lattice enthalpy that highlights the inverse process to form a crystalline structure from gaseous phase ions.

As Hess's law states, the final enthalpy of formation involved in a reaction does not change if reactants are found in a different phase state. The Born-Haber cycle is based on this law, and indeed, the entire sequence of phase changes leads to solid formation. Therefore, to find the *lattice enthalpy* of sodium chloride, we need to sum up all enthalpies of the phase change sequence with the following equation:

$$\begin{aligned}\Delta_f H^{\ominus}(NaCl) = \Delta_a H^{\ominus}(Na) &+ \Delta_a H \ominus (Cl) + \\ &+ \Delta_i H(1)^{\ominus}(Na) + \Delta_{eg} H(1)^{\ominus}(Cl) - \Delta_{latt} H^{\ominus}(NaCl)\end{aligned}$$

Knowing the enthalpy of formation we can rearrange the equation to find, finally, the lattice enthalpy $\Delta_{latt}H^{\ominus}(NaCl)$ as arranged below:

$$\begin{aligned}\Delta_{latt} H^{\ominus}(NaCl) = \Delta_a H^{\ominus}(Na) &+ \Delta_a H \ominus (Cl) + \\ &+ \Delta_i H(1)^{\ominus}(Na) + \Delta_{eg} H(1)^{\ominus}(Cl) - \Delta_f H^{\ominus}(NaCl)\end{aligned}$$

$$\begin{aligned}\Delta_{latt} H^{\ominus}(NaCl) = +108 \, \frac{kJ}{mol} &+ 121 \, \frac{kJ}{mol} + \\ &+ 496 \, \frac{kJ}{mol} - 349 \, \frac{kJ}{mol} + 411 \, \frac{kJ}{mol} = +787 \, \frac{kJ}{mol}\end{aligned}$$

Experimentally, lattice energy values are computed with a negative sign to describe the stability of a crystal structure where the lower is the energy the more stable is the structure. Therefore, the inverse process from gaseous ions to solid sodium chloride is taken into account for the heat transferred [9]. Supposing the gaseous ions behave ideally, the following equation holds:

$$p\Delta V = \Delta n_{gas} RT$$

where $\Delta n_{gas} = -2 \; mol$ is the difference of gaseous moles between products and reactants, $T$ is the constant temperature ($298 \; K$) and $R = 8.314 \frac{J}{K \, mol}$. Therefore, the lattice energy $\Delta_{latt}U$ is computed as follows:

$$\Delta_{latt} U = -\Delta_{latt} H + \Delta n_{gas} RT = -787 \, \frac{kJ}{mol} - 5 \, \frac{kJ}{mol} = -792 \, \frac{kJ}{mol}$$

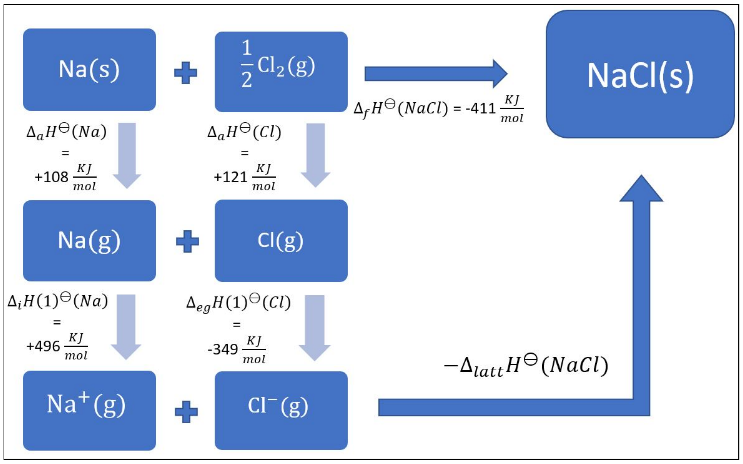

Na(s)
+
$\frac{1}{2}Cl_2(g)$
$\Delta_f H^{\ominus}(NaCl)$ = -411 $\frac{KJ}{mol}$
NaCl(s)
$\Delta_a H^{\ominus}(Na)$ = +108 $\frac{KJ}{mol}$
$\Delta_a H^{\ominus}(Cl)$ = +121 $\frac{KJ}{mol}$
Na(g)
+
Cl(g)
$\Delta_i H(1)^{\ominus}(Na)$ = +496 $\frac{KJ}{mol}$
$\Delta_{eg} H(1)^{\ominus}(Cl)$ = -349 $\frac{KJ}{mol}$
$-\Delta_{latt} H^{\ominus}(NaCl)$
$Na^+(g)$
+
$Cl^-(g)$

Figure 1.7: The Born-Haber cycle of the sodium chloride ($NaCl$) formation.

### 1.9.2 Approximated lattice energy

When we want to calculate the energy and we do not have enough thermodynamic data regarding the enthalpies of the *Born-Haber cycle*, we may use a theoretical method to approximate the energy. For first, ions can be considered as point charges where a distance $r$ lies between them. For ionic solids, like sodium chloride, it is easier because their solid structure consists of only electrostatic interactions between ions of sodium and chlorine and therefore, they are subject to the **Coulomb's Law**:

$$\Delta_C U = -\frac{Z_+ \, Z_- \, e^2}{4\pi\epsilon_0 r}$$

where $Z_+$ and $Z_-$ are the ion charges (cation and anion's), $e = 1.6022 \times 10^{-19}$ $C$ is the charge on the electron, $\epsilon_0$ is the vacuum permittivity constant and $r$ is the distance between them. In ionic solids, ions interact to each other and these interactions should be considered pairwise. Forces that divide cations and anions are attractive, but repulsive forces push away ions of the same charge. Coulomb interactions are calculated by summing up all these ion interactions by producing an infinite sequence, and only the neighbourhood of the ions is considered. The sum of all interaction that involves the geometric structure is called **Madelung constant A** that contributes to the electrostatic potential of the Coulomb law:

$$\Delta_C U = -\frac{A \; N_A \; Z_+ \; Z_- \; e^2}{4\pi\epsilon_0 r} \tag{1.31}$$

where $N_A$ is the *Avogadro constant* which is added as a factor to take into consideration a molar amount of the energy value. The *Madelung constant* $A$ is computed by summing internal energies given by specific pairwise electrostatic interactions between an atom in the centre and its first neighbours $\Delta U_1$, second neighbours $\Delta U_2$ and so on until all ions of the unit cell are involved. For example, for sodium chloride *NaCl*, which has a close-packed cubic (fcc) structure, a sodium ion in the centre is chosen and the potential energies given by the interaction of sodium with its first neighbour at distance $d_1$, second neighbour at distance $d_2$ and third neighbour at distance $d_3$ are computed in the following way:

$$\begin{aligned}
A &= \Delta U_1 + \Delta U_2 + \Delta U_3 \\
\Delta U_1 &= 6 \times -\frac{(1\times 1)e^2}{4\pi\epsilon_0 d_1} = -6 \times \frac{e^2}{4\pi\epsilon_0 d_1} \\
\Delta U_2 &= 12 \times +\frac{(1\times 1)e^2}{4\pi\epsilon_0 d_2} = 12 \times \frac{e^2}{4\pi\epsilon_0 d_2} \\
\Delta U_3 &= 8 \times -\frac{(1\times 1)e^2}{4\pi\epsilon_0 d_3} = -8 \times \frac{e^2}{4\pi\epsilon_0 d_3}
\end{aligned}$$

where $\Delta U_1$ is the total potential or electrostatic energy between the central sodium and the 6 closest chlorine ions (first neighbours) with negative values due to the attractive force and distance $d_1 = r$, $\Delta U_2$ is the potential energy between the central sodium and the 12 sodium ions (second-closest neighbours) with positive values due to the repulsive force, and distance $d_2 = \sqrt{2}\,r$, and $\Delta U_3$ is the potential energy between the central sodium and the 8 third-closest chlorine ions with distance $d_3 = \sqrt{3}\,r$.

Moreover, repulsive forces should be added to equation 1.31. Because ions are not point-charges, their electron cloud should be considered when approaching each other at small distances. Therefore, Max **Born** suggested that the repulsive force potential could be expressed by an additional term for the Coulomb potential:

$$\Delta_B U = \frac{B}{r^n} \tag{1.32}$$

where $B$ is a constant and $n$ is a constant large number called *Born exponent*. Adding this term to the Coulomb potential energy gives the **Born-Landè equation** of the lattice energy with attractive and repulsive forces:

$$\Delta_{latt}U = \Delta_C U + \Delta_B U = -\frac{A\ N_A\ Z_+\ Z_-\ e^2}{4\pi\epsilon_0 r} + \frac{B}{r^n}$$

that can be simplified considering that lattice energy will be a minimum at equilibrium when $r = r_0$, the distance between ions at the equilibrium state. Minimizing the equation will give:

$$\Delta_{latt}U = -\frac{A\ N_A\ Z_+\ Z_-\ e^2}{4\pi\epsilon_0 r_0}\ (1 - \frac{1}{n}) \tag{1.33}$$

where the *Born exponent* $n$ can be found experimentally or approximated according to the electronic configurations of the ions by averaging pre-calculated values for each ion type in the solid structure. For example, in sodium chloride, values of $n$ for sodium and chlorine are respectively 7 and 9. The Born exponent for sodium chloride will be $n = \frac{7+9}{2}$ and the lattice energy $\Delta_{latt}U = -754.7\frac{kJ}{mol}$ which is close to its experimental value found in subsection 1.9.1 [9].

### 1.9.3 Lattice energies from Force Fields

When we consider *molecular crystals*, and we want to compute their approximated lattice energy, the process becomes more complex because a proper cut-off of the molecular structure is needed to compute it on a finite set of molecules. This choice arises because molecules are located in various positions, and all of them may have different orientations. In this case, also distant molecules may participate in assessing the approximated energy since several forces such as Van Der Waals, electrostatic or hydrogen bondings, may affect the lattice energy on a specific part. Therefore, after choosing a cut-off threshold, the sum of all possible interaction forces (intra and inter molecules) should be taken in consideration to better approximate the energy. To achieve this goal, **force fields** could be used to describe a molecular system through a functional part that addresses the geometry of the molecules, and a second part that includes parameters depending on the atomic elements involved. To gather all the essential parameters, computations are usually conducted in experiments among different types of compounds to record and fit the results into the force field function. Many force fields have been developed to reveal the nature of a system through atomistic simulations. They can depends on simpler and general functions such as in the **UFF** [39] force field that was made for any combination of elements. Nevertheless, more complex combinations may raise when taking into consideration conformational properties, energies of formation and atomic vibrations. These parameters are included in force fields such as **MMFF** [20] that is trained on a huge amount of data and used for mostly biological molecules, or the **PCFF** [46] that is trained on organic compounds but does not have suitable parameters for describing the molecular dynamics under fixed environmental conditions such as the temperature. A later approach, based on PCFF, was developed to address the condensed-phase applications. The **COMPASS** [45] force field was re-parametrised to accept various properties of the condensed-phase (when the driving force of transformation is triggered) together with the usual empirical data on isolated molecules.

## 1.10 Crystallographic information file (CIF)

Structure resolution occurs utilizing X-rays crystallography, where X-rays are ”focused” on the crystal until the Bragg equation (see subsection 3.1.1) is fulfilled. After the resolution, each type of information gained is stored in the **crystallographic information file** ending with extension CIF. The main structure of the file contains **attributes**, **list of attributes** (loop) and their **values**. There are following summarized some of the most important attributes:

- **_atom_sites**
  Data items with the _atom_sites attribute store a details' list about crystallographic cell and cell transformations of all atom positions. Usually, atom coordinates are expressed in fractional coordinates which represent a fraction of the unit cell vectors and are assigned to the attributes **_atom_sites_fract_x**, **_atom_sites_fract_y** and **_atom_sites_fract_z**. Every atom is assigned to a name with the attribute **_atom_sites_label** and its type is saved in **_atom_sites_type_symbol**.

- **_atom_type**
  The radius of intramolecular bondings can be found in the attribute **_atom_type_radius_bond** measured in Angstroms Å together with the intermolecular bond length in **_atom_type_radius_contact** which are stored as a list or loop for each atom type **_atom_type_symbol**.

- **_cell**
  This part of the file is formed by different attributes that describe the geometry of the unit cell. Unit cell parameters are stored in **_cell_length_a**, **_cell_length_b**, **_cell_length_c**, **_cell_angle_alpha**, **_cell_angle_beta** and **_cell_angle_gamma** which are respectively the length of unit cell vector $\vec{a}$, $\vec{b}$, $\vec{c}$, the angle $\alpha$ on the plane $b$-$c$, the angle $\beta$ on the plane $a$-$c$ and the angle $\gamma$ on the plane $a$-$b$.

- **_chemical_formula**
  These items specify the composition and chemical properties of the compound. All discrete bonded residues or ions are stored singularly in **_chemical_formula_moiety**, the chemical structure is explicitly expressed in **_chemical_formula_structural** with parenthesis, the total number of moieties is typed in the attribute **_chemical_formula_sum** and the molecular weight of the formula saved in sum is registered in the attribute **_chemical_formula_weight**.

- **_exptl_crystal**
  Data items in the _exptl_crystal category store details about experimental measurements on the crystal such as shape, size or density. The category starts with a descriptive name of the target crystal in **_exptl_crystal_description** followed by values representing the density of a crystal (see subsection 1.8.2) measured from the crystal cell and content in the attribute **_exptl_crystal_density_diffrn** and size of the crystal used for the x-rays diffraction method measured in millimetres in the attributes for the length **_exptl_crystal_size_length**, maximum, medial and minimum dimensions **_exptl_crystal_size_max**, **_exptl_crystal_size_mid**, **_exptl_crystal_size_min**.

- **_geom_angle**
  It is common to add details about bond angles as calculated from the atoms data. The category _geom_angle is a useful loop to check on angles formed by triplets of atoms. There is one line for each triplet with the atom labels that are assigned to three attributes **_geom_angle_atom_site_label** and the angle that they form in **_geom_angle** attribute.

- **_geom_bond**, **_geom_contact** and **_geom_hbond**
  These categories store information about intramolecular, intermolecular and hydrogen bonds respectively. Each category has a loop formed by one line for each bond. Main values of the categories consist of two labels of the atoms involved and the distance between them (**_geom_bond_atom_site_label_1**, **_geom_bond_atom_site_label_2** and **_geom_bond_distance**) in angstroms Å .

## 1.11 Objectives and Thesis Outline

After a wide introduction to crystal structures and their chemical properties, we will dig into a new topic that brings the 3D-dimensional structure of a crystal and its geometric features to be the protagonists of this thesis.

In chapter 2 we start discussing the **equivalence** relation between structures by defining which type of transformations should be considered in $\mathbb{R}^3$ to find a proper solution for crystal structure classification.

Chapter 3 highlights the **chemistry context** in which we are applying our study. It starts by explaining how a crystal structure is resolved (brought from the real world to a CIF file) and what is the recent state-of-the-art method used to address the comparison problem of big datasets of crystal structures. Our new contribution will be strictly related to improving these methods.

The first geometric feature (Voronoi Domain of a lattice), taken into consideration, is shown in chapter 4 and it is introduced together with the definition of **Voronoi** Diagram. This property defines a specific geometric structure in the space by holding interesting characteristics of a point set. Our emphasis was to understand the similarity of the closely related underlying periodic structures of crystal datasets.

The second geometric property, Averaged Minimum Distance (AMD), is explained in chapter 5. This is the fastest developed characteristic of point sets that bases its speed and effectiveness on mapping crystal structures to **distance vectors** (definition 5.1), considering the atoms inside a unit cell. The key advantage of these objects is that they are easily and quickly comparable by usual metrics, such as the Euclidean distance.

Finally, in the last chapter 6, we make use of AMD features from chapter 5 to perform **predictions on the chemical data** retained by a crystal structure.

# Chapter 2

# The Equivalence and Metric Problems of Crystal Structures

## 2.1 Equivalence

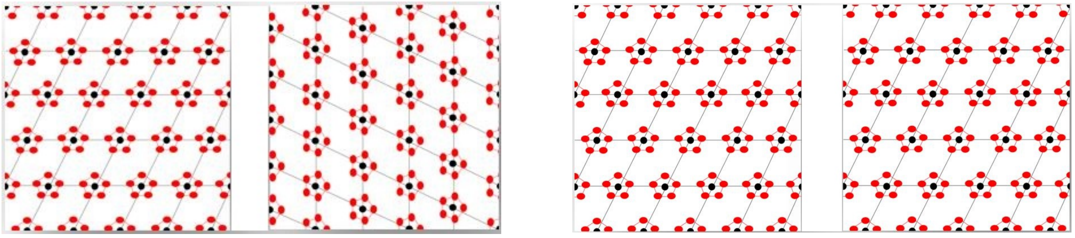
An ideal crystal is formed by a periodic point cloud that extends 3-dimensionally in all directions. Definitions of both a lattice and a motif lead to the crystal periodicity, where each point belongs to a periodic net that builds up the entire crystal structure. Since infinitely many bases can generate the same lattice, it can have many different unit cells that define the repeating unit. A periodic crystal is defined by a lattice $\Lambda$ and a motif M, which is a collection of molecules (for molecular crystals), atoms or ions (in the case of an ionic crystal such as NaCl). The motif is periodically translated in the directions along the 3 vectors that define a unit cell of $\Lambda$. The infinite set of linear bases of a lattice leads to a strong ambiguity. It becomes hard to decide if crystal structures (or lattices) are equivalent, when represented by **infinitely many different unit cells**. This information (which may be ambiguous) about crystals are contained in the Crystallographic Information Files (CIFs), such as edge lengths and angles of a unit cell $U$. All atoms are represented by fractional coordinates concerning the vector of $U$, i.e. as numbers within the interval [0,1]. These coordinates are often given for an asymmetric unit that generates a full motif in $U$ by applying symmetry operations specified in the CIF file (see Section 1.10).

Imagine holding and moving a diamond in your hands. Distances between its corners do not change, and every part of the most precious stone remains intact as before. Crystals are **rigid bodies**, and every transformation, such as a rotation or translation, applied to them does not change the distances between their points. Their structure remains unchanged. Crystals are solid materials; hence no **rigid motion**, which is a composition of rotations and translations in $\mathbb{R}^3$, can affect them. Therefore, any comparison of crystals should consider infinitely many positions (of a crystal or its lattice) related by rigid motions in $\mathbb{R}^3$. Crystal structures (or lattices) are called **equivalent** (or **isometric**) in $\mathbb{R}^3$ if they can be obtained from each other by a rigid motion, which **preserves distances** between any points in $\mathbb{R}^3$. In addition, reflections can be applied to them leading to the definition of **isometry** (rigid motions + reflections). An **isometry** is a linear map that preserves pairwise Euclidean distances between all points. For example, an isometry, like a rotation, maps points of a set to different coordinates but preserves the inter-atomic distances (see Figure 2.1).

**Definition 2.1.** (isometry). *An* isometry *of* $\mathbb{R}^n$ *is a map* $f : \mathbb{R}^n \to \mathbb{R}^n$ *that preserves the Euclidean distance, so* $|p - q| = |f(p) - f(q)|$ *for any points* $p, q \in \mathbb{R}^n$*. The map* $f$ *can also preserve the* orientation *if the matrix whose columns are images under* $f$ *of the standard vectors* $\vec{e}_1, \ldots, \vec{e}_n$ *has a positive determinant. In this case* $f$ *can be specifically called a* rigid motion*.*

Any isometry is a bijective function that can be inverted and decomposed. A composition of isometries is also an isometry and it is defined as an operation in the group $Iso(\mathbb{R}^n)$. A subset $Iso^+(\mathbb{R}^n) \subset Iso(\mathbb{R}^n)$ consists of all those isometries that can also preserve the orientation of an object. For example, a translation by a vector or a rotation over a line are called rigid motions and are a smaller subgroup of those orientation-preserving isometries. When a transformation (or matrix) preserves the orientation of a point cloud, the matrix belongs to the group $SO(\mathbb{R}^3)$. All orientation-preserving isometries in $\mathbb{R}^n$ can be decomposed into translations and rotations $R \in SO(\mathbb{R}^n)$ around the origin in $\mathbb{R}^n$.

Equivalence relations between transformations, applied to point clouds, are crucial for studying crystals as solid materials. Indeed, if we slightly perturb a crystal at the atomic level, its structure will change, but the small perturbation is not easy to detect, though it occurred, and it should be quantified. Since crystal can be equivalent to each other, they can be related by rigid motions (or isometries). An **isometry class** of crystals is a set of structures that may have different representations (e.g. different bases and motif coordinates), but they remain equivalent under isometries. If such crystals are found, we can say that they belong to the same isometry class. So, the space of isometry classes of crystals under isometries is infinite because many structures exist with infinitely many unit



(a) C' on the left, C" on the right

(b) C' on the left, C" on the right

Figure 2.1: Crystal C' and Crystal C" are equivalent because if we rotate C" counter-clockwise by 90° we obtain C'

cells, and continuous because a crystal is a rigid body that can be transformed into another by even small perturbations. Therefore, quantifying similarity between perturbed crystals is an important problem that helps detect any perturbation. Consider crystal lattices, for instance. The Bravais classification puts them into a much smaller number of classes (only 14 types in dimension 3), even if geometrically identical lattices may be classified as different. Hence we need other tools to check if given lattices are not equivalent. These tools are called invariants.

## 2.2 Isometry classification problem of crystal structures

In this section, we define the isometry invariant as a mapping function. Defining a mapping function that transforms crystals into simpler objects may help for a further comparison. Crystals are rigid bodies, and they should be considered under rigid motions (compositions of translations and rotations) or isometries (including reflections).

For a given equivalence relation, any objects can be distinguished only by an invariant. However, many descriptors of crystals include non-invariant values, for example, parameters of a unit cell, which are ambiguous since infinitely many unit cells exist for a given lattice.

**Definition 2.2.** (isometry invariant). *An* ***isometry invariant I*** *on a crystal is a property or numerical characteristic that does not change under isometries. To describe the fact that two crystals are isometric or equivalent (whatever is their orientation), the invariant should map both of them to the same number or property (Figure 2.2).*

**Definition 2.3.** (isometry invariant properties). *To correctly classify crystals, an* *isometry invariant $I$ should satisfy the following conditions:*

- **2.3(a).** ***Invariance****: if any periodic sets $S, Q$ are isometric, then $I(S) = I(Q)$;*
- **2.3(b).** ***Continuity****: $I(S)$ continuously changes under perturbations of points;*
- **2.3(c).** ***Computability****: a distance between values of $I$ is computable fast;*

*In addition, an invariant $I$ is* **complete** *when it can uniquely identify a crystal, so the condition below is satisfied:*

- **2.3(d).** ***Completeness****: if $I(S) = I(Q)$, then the periodic sets $S, Q$ are isometric;*

Condition 2.3(a) is needed for any reliable comparison of crystals. Indeed, it states that if two objects are equivalent or isometric, an invariant should map them to the same value. Many crystal descriptors include cell parameters or fractional coordinates, neither of which are isometry invariants. If a non-invariant takes different values on two crystals, these crystals can still be isometric, hence they can not justifiably distinguish crystals or predict crystal properties. In condition 2.3(b), our invariant should be sensitive to small perturbations of a point cloud and correctly detect those small changes in order to distinguish very similar structures and strongly discern very different crystals. Computability 2.3(c) states that an invariant should be quickly computable, for example in a polynomial time in the number of points in a motif $M$ of a periodic set $S$. Finally, it is possible that some structural information can be lost when applying an invariant to a crystal. The condition 2.3(d) of completeness allows us to uniquely identify any crystal $S$ by its complete invariant since we can easily and uniquely reconstruct the original crystal from the mapped space [2].

**Problem 2.4.** (Isometry classification problem). *Find numeric characteristics or properties (isometry invariants) of crystals or their lattice that can properly allow a further classification and distinguish classes among the entire continuous crystal space.*

An **isometry invariant of crystals** up to a particular relation (e.g. a rigid motion) is a function that should take the same value or property on all isometric crystals.
When crystal structures are not equivalent, we may want to quantify their differences, and therefore, our invariant should be continuous in order to detect and distinguish properly any tiny perturbation, which structures can be subject to. Indeed, the similarity

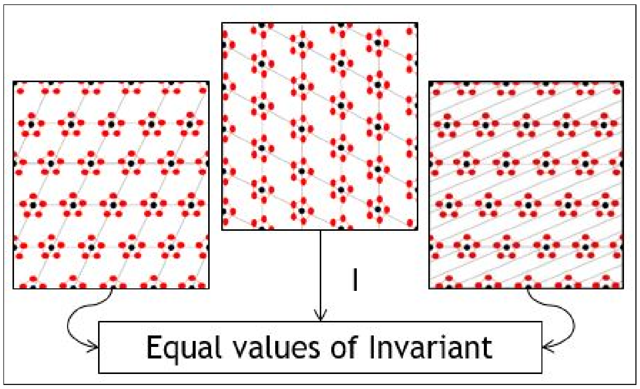


Figure 2.2: The invariant I applied to equivalent crystal structures maps to the same property.

of crystal structures can be assessed by isometry invariants that preserve their continuity in the mapped space under small perturbations. Without the continuity property we can only state whether they are equivalent or not, and we are not able to quantify those perturbations. Then, similar crystal structures, which are distinguished by a small perturbation in their motif, should be mapped by our invariant to close values (fig. 2.3). Let us consider an example of two 1-D periodic point sets (fig. 2.4), for example

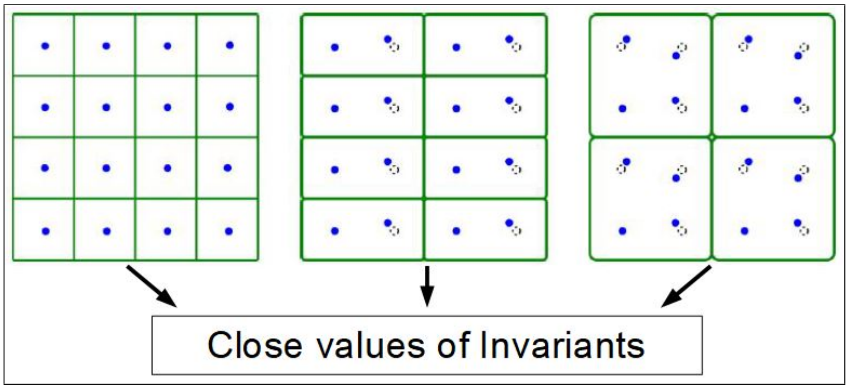


Figure 2.3: [43] Fig. 2. Similar crystals with 1-point, 2-point and 4-point motif should be mapped to close values of invariant.

$S = \{0, 1, 2, 3\} + 8\mathbb{Z}$ *and* $Q = \{0, 3, 4, 5\} + 8\mathbb{Z}$ which are repeated by a period of 8 towards all directions. If we define an invariant as the following:

$$I_{list}(C) = List\ of\ ordered\ set\ of\ pairwise\ distances$$

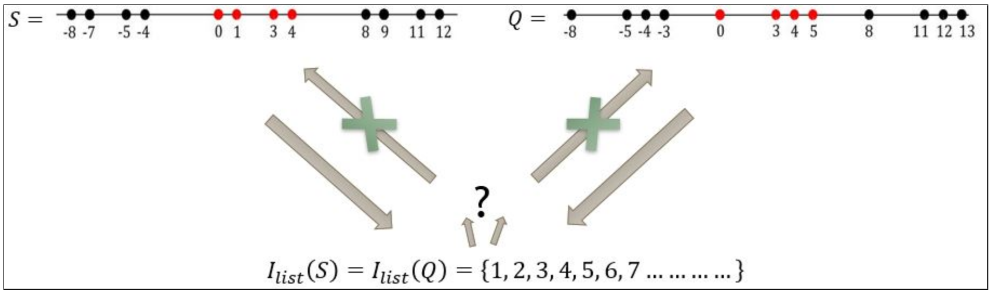


Figure 2.4: Not-complete invariants cannot reconstruct the original point set.

and then apply it to both S and Q:

$$I_{list}(S) = I_{list}(Q) = \{1, 2, 3, 4, 5, 6, 7...\}$$

we can notice that both sets are mapped to the same list of all ordered pairwise distances. Therefore, we cannot reconstruct the original point set from this invariant type. The invariant is called not-injective or **not-complete** in the sense that non-equivalent point clouds may map to the same invariant $I_{list}$ (Figure 2.4). We aim to define new isometry invariants that can distinguish crystal structures properly in a continuous space. Chapters 4 and 5 will deal with two isometry invariants such as the Voronoi Domain of a lattice and the Averaged Minimum Distances (AMDs) of a motif. Moreover, new metrics will be defined to solve the distance problem between lattices through their Voronoi domain, and usual vector metrics (such as Euclidean distance) will be used for AMD comparison of motifs.

## 2.3 Metric problem for crystal lattices comparison

Erroneously, edge lengths and angles (unit cell parameters) were considered before for lattice comparison [35]. However, they are not invariants because infinitely many primitive cells with different unit cell parameters may define the same lattice. The equivalence problem for lattices is to design a metric that accepts two arbitrary lattices and decides whether they are equivalent or not. Theoretically, such an algorithm can be based on Niggli's reduced cell [35] in subsection 2.4.1, and its instability under perturbations [1] leads to the metric problem (Problem 2.6).

Suppose a metric function between crystal lattices satisfies the metric axioms below.

In that case, the crystallography will be open to rigorous methods of metric geometry that will measure what portions of a crystal space are explored and what regions require more sampling in computer simulations.

**Definition 2.5.** (metric). *Let $\mathbb{R}^+$ denote the set of all non-negative real numbers. Let $S$ be any set. A metric on $S$ is a function $d : S \times S \to \mathbb{R}^+$, such that the following conditions (or metric axioms) hold:*

- **2.5(a).** ***Coincidence:*** *For any $\Lambda, \Lambda' \in S$, the metric function $d(\Lambda, \Lambda') = 0$ if and only if $\Lambda, \Lambda'$ are equivalent;*
- **2.5(b).** ***Symmetry:*** *$d(\Lambda, \Lambda') = d(\Lambda', \Lambda)$ for any $\Lambda, \Lambda' \in S$;*
- **2.5(c).** ***Triangle inequality:*** *$d(\Lambda, \Lambda') + d(\Lambda', \Lambda'') \geq d(\Lambda, \Lambda'')$ for any $\Lambda, \Lambda', \Lambda'' \in S$.*

For $S = \mathbb{R}^n$, one will use the Euclidean distance $d(p, q) = \sqrt{(p_1 - q_1)^2 + ... + (p_n - q_n)^2}$ between points $p = (p_1, ..., p_n)$ and $q = (q_1, ..., q_n)$, which satisfies the axioms above. For a set $S$ of crystal structures or arbitrary lattices, it is a hard problem to define a metric function $d$ satisfying the axioms above, because $d$ should not depend on crystal lattices representation, hence should be independent of many unit cells. Axiom 2.5(a) highlights the fact that when equivalency is satisfied a metric should be $d(\Lambda, \Lambda') = 0$, and for any non-equivalent crystal lattices $\Lambda \neq \Lambda'$ the value should increase. Axiom 2.5(b) says that a metric function remains the same if arguments are swapped. Axiom 2.5(c) is motivated by the assumption that a shortest path from Λ to *Λ”* should not be longer than a combination of shortest paths from Λ to *Λ’* and then from *Λ’* to *Λ”*. Any approach should justify that any non-equivalent crystal lattices Λ, *Λ’* have different representations. Else the metric function between numerical properties of non-equivalent crystal lattices Λ, *Λ’* is 0 and axiom 2.5(a) fails. The metric problem for crystal lattices requires a metric function that satisfies metric axioms 2.5(a), 2.5(b), 2.5(c) and also the continuity and scaling conditions below:

- **2.5(d).** ***Continuity:*** *The metric function $d(\Lambda, \Lambda')$ continuously changes under perturbations of crystal lattices, e.g., if cell parameters are noisy; in particular, the range of $d$ should be a continuous interval, possibly $[0, \infty)$, but not only a finite collection of discrete values.*
- **2.5(e).** ***Uniform Scaling:*** *The metric function $d(\Lambda, \Lambda')$ should remain unchanged*

*if both sets $\Lambda, \Lambda' \subset \mathbb{R}^n$ are scaled by the same factor $s > 0$, i.e., $d(\Lambda, \Lambda') = d(s \times \Lambda, s \times \Lambda')$, where $s \times \Lambda = \{s \times p \in \mathbb{R}^n : \text{for any point } p \in \Lambda\}$.*

Equivalence under scaling could be in theory useful to distinguish lattices that are scaled by a different factor. We did not dive into this topic in our experiments, but we provide a possible future application below. Let us take the example in subsection 3.2.3 where triptycene-based molecules are formed (mostly) by a various number of carbon rings. If we imagine adding more rings to all the molecules' arms, we scale up the molecules and eventually the lattice. Since molecules are repeated at each lattice point and may exist in differently scaled forms, the scaled lattices can be grouped for a further assessment on the structure of their motif (which may be related by some moieties addition).

From the above definition of metric and its axioms, the problem below on a metric of lattices follows:

**Problem 2.6.** (metric problem). *The metric problem for crystal lattices is to find a function that satisfies the metric axioms 2.5(a), 2.5(b), 2.5(c) and 2.5(d) to correctly distinguish them in a continuous space when small perturbations may be applied.*

## 2.4 Past methods to quantify differences between crystal lattices and structures

Most solid minerals in nature and several important synthetic materials are periodic crystals at the atomic scale.

A periodic point set, even without extra links (bonds), is a fundamental model for any solid crystalline material. Indeed, atomic centres (nuclei) have a less ambiguous physical meaning than inter-atomic bonds, which can be challenging to classify and require bespoke and empirical definitions. For example, at what distance does a hydrogen bond become a bond? However, for any zero-size point representing an atom, one can add a label such as a chemical element or a radius or another physical property.

Typically, a periodic crystal is stored as a Crystallographic Information File (CIF), including parameters of a unit cell (three edge-lengths and three angles in $\mathbb{R}^3$) and all atoms with chemical labels and coordinates of atomic centres in the unit cell. The input size of a crystal is best measured as the number of atoms in its unit cells or the number of points in a motif of a periodic set. Unit cells may have a high number of atoms and so a near-linear time algorithm should be preferred to find proper isometry invariants. Many methods have been developed in the past years trying to address and classify the vast space of crystal structures, although some of them neglect the continuity property as described below.

### 2.4.1 Niggli's reduced cell

Crystallographers used **Niggli's reduced cell** [35] as a canonical cell for a lattice. This reduced cell is unique but at the price of continuity, and retains the interesting properties below:

- Angles close to 90°
- Uniqueness
- Can be found regardless of the starting point

Following the general idea of the reduction process, every unit cell vector is reduced with respect to the others. The reduction continues until every condition is satisfied. For instance, given the bidimensional unit cell with the two vectors $\vec{a}$ and $\vec{b}$, the vector $\vec{b}$ is considered reduced with respect to $\vec{a}$ when the following condition holds: $|b \ \cos\gamma| \leq \frac{1}{2}a$ with $b = ||\vec{b}||$ and $a = ||\vec{a}||$. It means that, by fixing $\vec{a}$, the projection of $\vec{b}$ on $\vec{a}$ is lower than half of $\vec{a}$.

In Figure 2.5, for example, it is shown a skewed unit cell with vector $\vec{b_1}$ and $\vec{a}$. Since the projection of $\vec{b_1}$ on $\vec{a}$ is bigger than half of $a$, $\vec{b_1}$ needs to be reduced to $\vec{b_2}$ where $\vec{b_2} = \vec{b_1} - \vec{a}$. With the last step, the condition holds, and $\vec{b_2}$ is considered reduced. The same procedure is performed on $\vec{a}$ by following the condition: $|a \ \cos\gamma| \leq \frac{1}{2}b$. It means that $\vec{b}$ is fixed and $\vec{a}$ should be reduced. The projection of $\vec{a}$ on $\vec{b_2}$ is lower than half of $b_2 = ||\vec{b_2}||$ and the reduction is complete. Regarding the three-dimensional unit cell, the number of conditions rises up to 6: two conditions for every pair of the unit cell vectors $\vec{a}$, $\vec{b}$ and $\vec{c}$. In this case, the unit cell is in its reduced form if it satisfies all conditions:

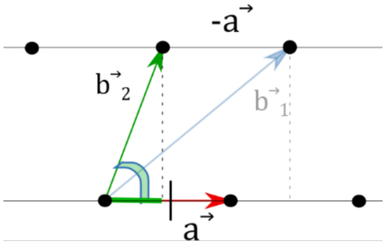


Figure 2.5: Vector reduction: Fix $\vec{a}$ and reduce $\vec{b_1}$ to $\vec{b_2}$.

$$\begin{aligned}
|a\ \cos\gamma| &\le \frac{1}{2}b & \qquad |b\ \cos\gamma| &\le \frac{1}{2}a \\
|b\ \cos\alpha| &\le \frac{1}{2}c & \qquad |c\ \cos\alpha| &\le \frac{1}{2}b \\
|c\ \cos\beta| &\le \frac{1}{2}a & \qquad |a\ \cos\beta| &\le \frac{1}{2}c
\end{aligned}$$

where $a = ||\vec{a}||$, $b = ||\vec{b}||$, $c = ||\vec{c}||$, $\gamma$ is the angle between $\vec{a}$ and $\vec{b}$, $\alpha$ is the angle between $\vec{b}$ and $\vec{c}$, $\beta$ is the angle between $\vec{a}$ and $\vec{c}$.

Despite the uniqueness property [35], the Niggli's reduced cell cannot be used directly to study the similarity between crystal structures because of its instability [1]. Indeed, there is also the possibility that similar crystals may have a very different reduced cells. It means that a reduced cell of a slightly perturbed lattice is not close to a reduced cell of the non-perturbed lattice, as shown in Figure 2.6. Even a small perturbation may increase the number of reduction steps and, consequently, change the resulted cell parameters such as the angle between unit cell vectors drastically. Examples of discontinuous Niggli's cells were known since 1980 [1]. Discontinuity under small atomic perturbations is the major weakness of all discrete invariants, including symmetry groups. In practice, the nearly identical periodic sets in the last two pictures of Fig. 2.3 should be recognisable as very similar.

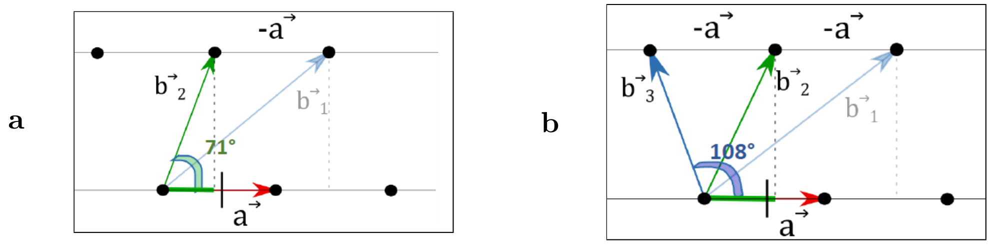


Figure 2.6: (a) Non-perturbed lattice with 1 reduction step from $b_1$ to $b_2$. (b) Slightly perturbed lattice with 2 reduction steps which lead to a very different angle between $\vec{a}$ and $\vec{b}_3$.

### 2.4.2 The COMPAK algorithm for the Cambridge Structural Database (CSD)

Though there was no justified distance that satisfies all metric axioms for any periodic crystals, the **COMPACK** algorithm [10] is widely used for pairwise comparison of crystals. Within given tolerances (20° for angles and 20% for distances), up to a given number (15 by default) of molecules from two crystals are matched by a rigid motion that minimizes the **Root Mean Square deviation** of $N$ matched atoms $RMS = \sqrt{\frac{1}{N}\sum\limits_{i=1}^{N} |p_i - q_i|^2}$.

| $m$ matched molecules | 5 of 5 | 9 of 10 | 12 of 15 | 16 of 20 | 21 of 25 | 26 of 30 | 28 of 35 |
|---|---|---|---|---|---|---|---|
| RMS, $1\mathring{A} = \frac{1}{10}nm$ | 0.603 | 0.708 | 0.874 | 0.969 | 1.080 | 1.040 | 1.044 |
| running time, seconds | 0.168 | 0.422 | 2.026 | 14.61 | 63.51 | 151.4 | 759.3 |

Table 2.1: The Root Mean Square (RMS) deviation between the experimental T2-$\delta$ crystal [37] and its closest simulated version with ID 14. The irregular dependence of RMS on $m$ makes this comparison unreliable. The running time substantially grows in the number of molecules.

Informally, the RMS is restricted to a finite subset of atoms or molecules whose choice may depend on extra parameters. If a match between these finite subsets is extended, $RMS$ can also increase, potentially to infinity, for example, between cubic lattices of sizes 1 and 1.1. Table 2.1 shows how RMS depends on the maximum number $m$ of attempted molecules to match by rigid motion.

### 2.4.3 Powder X-rays Diffraction Pattern similarity (PXRD)

When single-crystal diffraction data are missing, PXRD could be used to characterise a solid material. Since crystallites may be oriented in different ways, the diffraction is performed in a range of angles guided by the mechanical motion of the source and detector together. Crystals are assessed in a different orientation; therefore, this procedure helps to reveal crystalline impurities or phase mixtures in products [48]. The procedure involves the comparison of diffraction profiles of two crystal structures. The squared difference method is used to overlap, in theory, both patterns and check similarities among their intensity peaks.

### 2.4.4 The COMPSTRU algorithm of Bilbao Crystallographic Server (BCS)

Similarly to COMPACK, the recent COMPSTRU algorithm [15] measures similarity between a given reference structure $S$ and crystal structures whose lattice parameters should be close to those of $S$ (by default 0.5 Å for distances and 5° for angles). This comparison is restricted to crystal structures that have the same space-group type. A small perturbation of atomic positions of the reference $S$ will produce a nearly identical crystal that is not comparable to the reference $S$, breaking the continuity condition 2.3(b).

### 2.4.5 Pair Distribution Function (PDF)

More complex materials as disordered crystals require the PDF analysis. It gives a deep quantitative insight of a material since these type of crystals may have a very noisy diffraction pattern where peaks of intensities lose information due to a nanoscale structure [5]. Therefore, the idea is to count the number of atoms on a shell of radius $r$, starting from a specific origin. The probability of meeting an atom at distance $r$ starting from the reference atom (origin) is the Pair Distribution Function. The computation extends for each type of atoms in the crystal. However, it is not a continuous function and it is very sensitive to small changes since they can result in large signal shifts.

### 2.4.6 Radial Distribution Function (RDF)

The Radial Distribution Function (RDF), described partially in subsection 1.2.3 for electrons, is based on the density of atoms in a shell of radius $r$ and width $dr$ centred around an atom of a specific type [49]. Since atom types are essentially used, the RDF can be used for comparing crystals that are composed of the same atom types. Due to averaging across atoms of a specific type within a unit cell, the RDF is independent of a cell choice. A similar distance-based fingerprint was introduced earlier by Valle and Oganov [52]. As in the PDF, small changes in the internal structure may cause large shifts of the distribution since atoms may not be found in the same shell of radius $r$ and width $dr$ after small perturbations.

## 2.5 Recent progress on Isometry Invariants

Our goal is to develop and implement new geometric features that can unveil the structural composition of crystal structures and allow an easy and fast comparison of their characteristics. By doing so, we will be able to bring the materials' discovery to the next level: comparing properly more crystal structures means having more possibility to **discover new structures** with different functions applicable in the real world.

Our group of Data Science and Theory Applications made different steps forward in developing new invariants of crystal structures that are summarised below.

More recently, for any periodic point set $S \subset R^n$ with a motif $M$ in a unit cell $U$, Edelsbrunner et al. [12] introduced the density functions $\psi_k(t)$ for any integer $k \geq 1$. The *k-th* **Density Function** $\psi_k(t)$ is the total volume of the regions within the unit cell $U$ covered by exactly $k$ balls $B(p;t)$ with a radius $t \geq 0$ and centres at points $p \in M$, divided by the unit cell volume $Vol[U]$. In practice, $k$ balls grow around points passing through various states of intersection between each other. When an intersection occurs, a new density function is released, indicating the volume of the new regions generated when their intersection began. The density function $\psi_k(t)$ was proved to be invariant under isometry, continuous under perturbations, complete for periodic sets in general position in $\mathbb{R}^3$, and computable in time $O(m\,k^3)$, where $m$ is the motif size of $S$. The concept of single-value density $\rho$ is practically extended to density functions which consider the periodic geometric structure in a proper and continuous way. The resulting densigram is provably complete for periodic sets in a general position, but is slow to compute.

In addition, the invariant **Isoset** [2] reduces the isometry classification of all periodic point sets to a finite collection of isometry classes of $\alpha$-clusters around points in a motif at a certain radius $\alpha$ and proved to be continuous under perturbations. Checking if two isosets coincide needs a cubic algorithm, which is not yet implemented. The idea is to find all clusters of points depending on the $\alpha$-radius chosen and check if they could be related by equivalence or similarity under isometries. An isometry $f \in Iso(\mathbb{R}^n)$ between local clusters should match their centres.

Moreover, another new invariant **Root Invariant** has been developed to solve the problem of isometry classification of lattices where a list of specific vectors, called superbases, can be found from any lattice and stored as a list of squared lengths or norms (vonorms). They can be computed in function of their vectorial scalar product (conorms) and vice versa. Conorms are then square rooted and ordered (root form RF) to represent a complete invariant of a lattice [26] [25].

The latest invariant **Pointwise Distance Distribution** (or PDD) [59] of periodic sets, is considered also up to isometry in $\mathbb{R}^3$. It involves a simpler computation of inter-atomic distances that maps a crystal structure to distance distributions. The new representation of a crystal by a numerical matrix is faster and continuous under perturbations. Moreover, each mapped crystal can be reconstructed from the numerical matrix space. A recent computation of PDD ran on the entire CSD dataset consisting of 660K crystals, and 5 crystals have been found to have identical structures but with different atom types.

# Chapter 3

# Introduction to materials discovery

## 3.1 X-rays crystallography

To be studied at the atomic level, solids should be resolved to gather structural information that will help determine the structure and the physicochemical properties. A powerful method for resolving a solid crystalline structure is X-rays crystallography. A crystalline solid acts as a grid when directed by a beam of X-rays. It means that the electron cloud of the atoms scatters the X-rays leading to a **diffraction pattern** formed by a series of spots on an image plate. From the diffraction pattern, the electron density can be rebuilt with computer software allowing to determine the bond lengths and angles. In the following paragraphs we will deal with two types of X-rays diffraction methods, **powder** and **single-crystal diffraction**.

### 3.1.1 X-rays diffraction

The discovery of X-rays is dated to 1895 when **Rontgen** benefited of his discovery for medical diagnosis and treatment. Diffraction techniques can measure bond lengths. The **electron beam** is a powerful tool to resolve structures because electrons can be forced to travel at a specific wavelength to match the size of the object involved. Indeed, when an electron beam wavelength is of the same order of the distance between two atoms, it can generate a diffraction pattern. Since the wavelength depends on the electron velocity, it can be tuned easily by applying a specific potential difference.

Emission of electrons comes from an electrically heated filament of tungsten where electrons are accelerated by a high potential difference of 20-50 kV, allowing to change the wavelength at which they travel accordingly to De Broglie equation 1.5. Moreover, this beam of electrons is directed to a metal target or anode that is water-cooled. It redirects the electron beam out as high-intensity X-rays for the specific X-rays wavelength. Usually, copper and molybdenum are used as target metals in X-ray crystallographic studies because they can redirect the electron beam with high intensities at a wavelength of 1.54 and 0.71 Å respectively. Crystalline solids are formed by periodically arranged arrays of atoms, ions or molecules with inter-atomic spacing of the order of 1 Å . The beam's wavelength must be of the same order of magnitude as the grating spacing, and a crystal can act as a diffraction grid with specific wavelength values. In 1913, W.H. and W.L. **Bragg** started their experiments on using X-rays diffraction to determine the structures of solids. The first structure that they resolved was the sodium chloride $NaCl$ followed by many others. They noticed that crystal diffraction behaves as **reflections** of structure layers on which atoms lie and that only specific orientations of a crystal with respect to the source and detector are suitable to reflect X-rays. When we deal with parallel layers of atoms (each of them identified by a triplet $hkl$), we may want to consider also the inter-layer distance $d_{hkl}$ or inter-planar spacing. When the source shoots X-rays, the beam with parallel X-rays is incident to the planes at an angle $\theta_{hkl}$ which hits atoms on different layers. If an atom is directed and struck by the beam, X-rays are reflected with an angle $2\theta_{hkl}$ travelling towards the detector placed accordingly. The reflected beam should arrive in phase to the detector as a single one to hold a strong intensity, and this property is known as **constructive interference**. For example, two X-rays beams strike two atoms on the first and second layer with incidence $\theta$. Both the redirected rays are reflected by $2\theta$ towards the detector (fig 3.1). To detect both atoms on the same vertical line, the X-rays beams must arrive in phase given that the second one traverses an extra length. Therefore, the first beam and the second bean have a path difference as described below:

$$2d_{hkl}\sin\theta \tag{3.1}$$

To arrive in phase, the difference in equation 3.1 must be equal to an integral number of wavelengths as the following **Bragg equation** states:

$$n\lambda = 2d_{hkl}\sin\theta \tag{3.2}$$

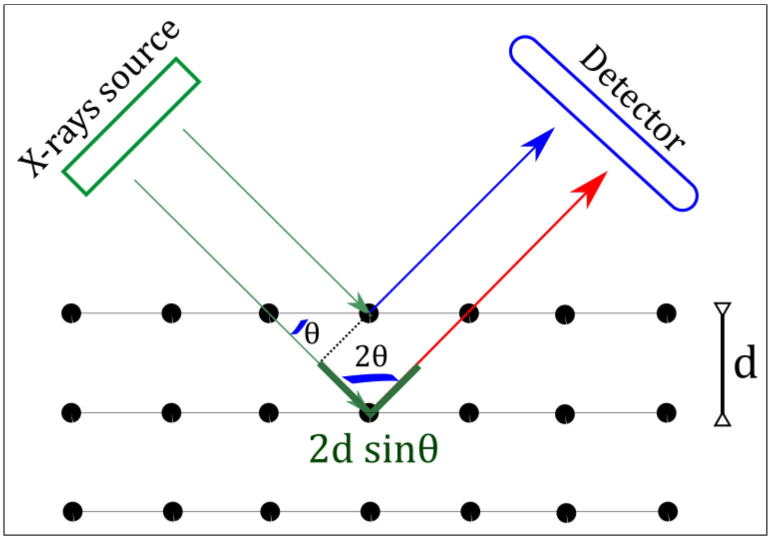


Figure 3.1: An example of X-rays shot by a source and reflected by atoms A, B and C on different planes

which relates the spacing between crystal planes, $d_{hkl}$, to Bragg angles, $\theta$, used to observe them. The property of an atom to scatter X-rays is called **scattering factor $f_0$**, which depends on the atomic number $Z$. Indeed, the higher the number of electrons, the more strongly X-rays will scatter on the Bragg angle $\theta$ and the beam's wavelength. Moreover, the scattering power could decrease when the angle $\theta$ increases. Once X-rays hit an atom and reflect towards the detector, they could miss other electrons around the nucleus as the angle becomes wide. Practically, the Bragg equation serves as a mathematical tool used to increase the focus on the crystal to hit as many atoms as possible on the layers.

### 3.1.2 Powder diffraction

To study high symmetric crystals, a thin ground crystalline powder is used. It contains numerous small crystals known as **crystallites** which are oriented in different ways. When a beam of X-rays strikes the crystallites, it will hit only those oriented in such a way that the Bragg equation gets integral values on the left side, and the diffracted beams make an angle of $2\theta$ with the incident beam. Even if the crystallites can lie in different orientations, they could be still suitable to fulfil the Bragg equation, and, in this case, the diffraction pattern behaves as a set of cones where reflections lie. The collection of powder diffraction patterns is performed automatically by a diffractometer that adjusts the angle of diffraction accordingly to fix the intensity of diffracted beams by storing intensity values, angles and positions of the X-ray tube and detector sequentially.

### 3.1.3 Single crystal X-rays diffraction

It is also possible to work on a single crystal in order to measure the position or intensity of the *hkl* reflections and determine the unit cell dimensions, space groups and atomic positions. Single crystal X-ray diffraction data is collected using an automated diffractometer that measures Bragg angles $\theta$ and the intensity for each reflection *hkl* where all of them can be collected and measured at the same time by a flat-plate detector (CCD). Thanks to this method, we can determine the size and shape of the unit cell by finding all unit cell parameters. Each reflection *hkl* is indexed, and symmetry elements can be determined, allowing to identify the space group to which the crystal belongs. In addition to the reflection collection, determined structures should undergo a refinement process to allow the electron density of an atom to be better refined around the nucleus. To perform the refinement, we need to collect the result of the waves scattered by all atoms in the unit cell that is called **structure factor $\boldsymbol{F_{hkl}}$** and depends on each atom position and its scattering factor:

$$F_{hkl} = 2\sum_{i}^{n} f_i \; cos2\pi \; (hx_i + ky_i + lz_i) \tag{3.3}$$

where $i$ is the atom index, $x_i, y_i, z_i$ are its fractional coordinates, $f_i$ is its scattering factor, and it defines a way of combining the power of reflections with atomic coordinates. The presence of the cosine confirms the periodic behaviour of the waves originated from a reflection.

As mentioned before, the structure factor can be computed knowing about the atom type and position together with its scattering factor. Consequently, we may want to rebuild (or resolve) the atomic structure and store the crystal in a CIF file by detecting intensities of reflections at which atoms occurs. We can do that by reversing the computation and be able to find the atom position starting from the amplitudes of the structure factors. Unfortunately, this reversed procedure can lead us to experience loss of information known as the **phase problem**. In fact, intensities produced from each atom $I_{hkl}$ are proportional to the square of the structure factor $I_{hkl} \propto F^2_{hkl}$. Therefore, if we take the square root of the intensity, we will not know the amplitude sign $|F_{hkl}| \propto \sqrt{I_{hkl}}$ of the waves. Intensities are crucial for detecting atom types especially for heaviest atoms that produce a stronger signal as they have more electrons. One solution to the phase problem is the **Patterson** method that makes use of the so-called Patterson function, named after the scientist that first proposed it. The idea is to consider $N$ atoms and all inter-atomic vectors up to $N^2 - N$ that represent peak positions. Peak heights refer to the product of electron (atomic) numbers involved in specific atom pairs.

Once each atom has been identified with its atomic coordinates, many structure factors are calculated $F_c$ and compared to the observed amplitudes $F_o$ to refine the atom positions. Finally, to assess the quality of the determined structure, they are used in a final measure that describes its correctness and precision, called **R factor**. It is usually applied to estimate errors in a data set and follows the function below:

$$R = \frac{\sum |(|F_o| - |F_c|)|}{\sum |F_o|} \tag{3.4}$$

where $F_o$ is the observed structure factor after data correction (or reduction) and $F_c$ is the calculated structure factor. The lower is the *R factor* the better is the structure determination since the difference between observed and calculated factors decreases [50].

## 3.2 Materials discovery

Materials have been resolved through different crystallographic methods, and two of the most important ones have been described in the previous paragraph. A significant milestone, which helped the progress of materials science, consists of methods that use previous structural information of resolved crystals and generate new structures that could be more or less suitable for a crystalline material to be called a solid and to exist in the real world. Most important, the first step of a scientist is to **design** the composition or structure of a new compound, knowing that there are some sets of physicochemical rules that should be considered. Finding new **physico-chemical properties** of a compound is the primary purpose of materials discovery science that is strictly related to the possibility of designing and synthesising a new crystal or, practically, to the possibility for a chemist to shape materials to their will. Regarding the **synthesis** of a compound, the aim is to look for a structure that, following all set of rules, may be synthesised or, in particular, keep a stable structure that prevents it from changing its solid phase according to different environmental conditions such as the temperature. The foundation of this approach is based on the *lattice energy* (see subsection 1.9.1 and 1.9.2), a measure of structural stability that gathers all possible potential energies of atoms and molecules. It should be minimised to find the most negative value as the number of interactions between particles increases. Therefore, to explore the space of stable conformations, it is a valuable decision to project crystals onto an **energy landscape** that embraces the whole world of known compounds where some of them may not have been synthesised yet but, in theory, are capable of existence. When a structure has the lowest lattice energy value, it is called **thermodynamically stable** and corresponds to the global minimum. The potential energy is taken into consideration and could be written in terms of external thermodynamic variables or considering atomic charges and distances between neighbours (see subsection 1.9.2). Moreover, the remaining local minima of the landscape may exist in theory and are called **metastable structures**. In the case of molecular solid, creativity can give birth to materials with hopefully various properties where the main focus falls on the concept of **polymorphism** which relates to the possibility of having the same compound with a different arrangement of molecules in the 3D space [22]. Although only one molecule can be enough to generate a vast energy landscape of compound polymorphs as in **porous molecular materials** , such as **hydrogen-bonded organic frameworks (HOFs)**, it would be also possible to use more of them. Some of them can serve as **linkers** that connect target molecules and keep them tight by building blocks such as in **extended frameworks** like **covalent organic frameworks (COFs)** or **metallic-organic frameworks (MOFs)**. The difference between the former and the latter is that *porous molecular materials* are porous solids formed by discrete molecules held together by weak interactions such as hydrogen bonds, which may support the solubility of a compound in organic solvents. So they can be produced by crystallisation where no new bonds are formed. Single crystal x-rays diffraction becomes possible to resolve them, whereas *extended organic frameworks* have molecules covalently linked together by strong interactions [11]. These particular examples are open-pored structures, three-dimensional frameworks used for important applications such as gas storage for vehicles' tanks. These molecular solids arrange their building units through covalent or non-covalent interactions in such a way that they form infinite channels passing throughout the structure used to trap gas molecules inside.

### 3.2.1 Functional materials discovery

*Molecular solids*, such as HOFs, do not follow the usual rules of bondings as *extended frameworks* do because weak interaction positions and orientations are not easily predictable.

Therefore, **crystal structure prediction** has been used over the past years as a tool to generate new possible crystal structures that may exist in the landscape. Moreover, the prediction of chemical properties became possible through maps that relate crystal functions to both their structure and lattice energy [37]. The method consists of a first step that involves the choice of a specific experimental molecule already known. Secondly, it would be optimised in its internal energy to make a better structure which has more suitable bond lengths and where atoms occupy the best positions to balance covalent forces within the molecule. It has already been applied for several materials such as allotropes of common elements, organic photovoltaics, and porous solids where a thousand or million crystal structures are predicted with the only molecular structure as input.

Despite the high number of structures produced, the final aim is to predict the properties of hypothetical crystal structures identified as best candidates for synthesis in a laboratory which implies the structure-related knowledge of their stability considering various atomic configurations. *Molecular solids* are characterized by complex **energy landscapes** because weak interactions are less predictable than covalent bonds. Therefore, small changes in the molecular structure may result in big changes in the molecule packing. The **stability** of each structure depends on its predicted lattice energy and so, the projection of a structure on the energy landscape became the first necessary step.

### 3.2.2 Crystal Structure Prediction (CSP)

Challenges in material science involve the prediction of structures at the atomic level, which may depend on the lattice energy or, mostly, on the atom packing. A model should be created to approximate bonding rules because it includes all possible characteristics used to express a compound behaviour about the internal or intermolecular bonds. A **model** and a **prediction** have two different meanings. The former is more related to the fact that an experimental structure could be validated under certain rules by checking the equilibrium state that keeps a stable solid phase and minimises a lattice energy function. Indeed, a model may take as input an experimental structure and apply modifications to the bond lengths according to the rules expressed to find a more stable state. In contrast, the latter has a significantly different approach because we have no empirical information about the atom positions that should be firstly generated. Moreover, models of force fields for lattice energy minimisation could be used by prediction methods to refine the procedure of building a crystal structure by predicting its atomic configuration. There are many methods

used in crystal structure predictions that explore the configuration space and use different procedures to cover the whole landscape. A review of different methods can be found in [62] of which some are dealt with in the current paragraph.
The **simulated annealing** method is based on the physical concept of increasing and lowering the temperature to simulate a disordered and ordered state of atoms in a compound, respectively. Ions in a crystal structure are continuously perturbed at each temperature through the Monte Carlo approach, where a random number of atoms is affected. Metropolis criterion is used to decide whether the Monte Carlo move or random perturbations are accepted or not. Since the disordered status is reached by increasing the temperature, the initial temperature should be high enough to let the compound pass the energy barriers to break bonds physically. The consequent cooling down phase will lead to a local minimum with bond formations, decreasing the probability of jumping among other minima. Indeed, if the annealing process of perturbing and cooling down is done slowly enough, the global minimum could be reached. A modified version of the Monte Carlo move has been called Monte Carlo basin hopping (MCBH [54]) where the current perturbed structure is optimised (or technically relaxed) at once to find first the closest stable phase. Then the Monte Carlo move is assessed for acceptance. The problem with the simulated annealing is that the exploring phase starts from a single point (or single atomic configuration) and could miss different minima wells around the landscape.
**Genetic algorithm** methods are another way to explore the lattice energy landscape that relies on the fact that crystals are gathered in a population of structures following the analogy of species' evolution. The first generation is the starting point that includes many structures that will be "mating" to each other for the first time and that are generated by a random arrangement of atoms. As in the species evolution, crossing over and mutations occur to support diversity in the next population and so, combination or insertion of features are performed during the perturbation phase. Crossing over and mutations are two operators that help the population to evolve to a better one. Therefore, the previous population of structures needs to evolve to approach a better evaluation of the lattice energy function. Thus, not all structures are allowed to mate because some of them, resulting from two parent crystals, may retain bad mutations or insertions that could lead to high energy values. A selection of each population occurs to choose those that can evolve to a lower energy structure. Furthermore, structure relaxation could be used for each resulted structure to better and quicker converge to lower energy packings.
**Molecular packing** approaches act differently because of the high possibility of molecular

crystals forming polymorphs of the same compound. The lattice energy landscape is more complex as the molecules may pack differently. Small changes in their orientation and position may result in high lattice energy changes. Grid-search approaches, such as MOLPAK method [21], find densely packed structures modelling atoms as hard spheres, together with simulated annealing and Monte Carlo methods adapted for predicting polymorphic structures [16].

### 3.2.3 Energy-structure-function maps

In the work mentioned above [37], the energy-structure-function map of molecular organic crystals is used to guide the discovery of new molecular materials with specific porosity and high predicted gas selectivity. Authors studied different molecules used as candidate building blocks for porous solids, but we will deal with three of them for the explanatory purpose of this chapter. The energy-structure related landscape involves the use of density property (see subsection 1.8.2), which can be plotted against lattice energy to check and select a subset of crystal structures having good possibilities to be candidates for synthesis. Target molecules have a tripod shape with three arms and mostly consist of carbon rings followed by a few nitrogen and oxygen atoms. **Triptycene-based molecules T1** and **T2** tend to form intermolecular hydrogen bonding stabilizing in a porous phase, except for **T0** (Fig 3.2a-c). Specifically, T2 structures are made of two carbon rings per arm plus other two in the geometric (or mass) centre, four nitrogen atoms and one oxygen atom in each extreme of the arms. Many simulated crystal structure from this molecule present, in their energy-density landscape (Fig 3.2f), some low density structures indicating unusual stability for their density values such as T2-$\gamma$ found in the spike of 0.4 $\frac{g}{cm^3}$, T2-$\beta$ and T2-$\alpha$ in the 0.8 $\frac{g}{cm^3}$ spike, whereas T2-$\delta$ is the second global minimum after T2-$\epsilon$ lying around 1.3 $\frac{g}{cm^3}$.

The landscapes of T0 and T1 belong to the typical behaviour of organic molecules that have the potential to densely pack, retaining low lattice energy where the energy-density distribution decreases monotonically. (Fig 3.2d-e). T0-$\alpha$ is an example of a non-porous structure which keeps a stable structure with only a high-density value. In the T2 dataset, 5 out of 5679 have been synthesised with the global minimum $T2\epsilon$, despite the low density featuring hydrogen-bonded networks with two-dimensional rings. They form one-dimensional pore channels and lie in the minima spikes of the landscape by keeping the most stabilising electrostatic interactions. Low-density structures such as T2-$\gamma$ showed a high predicted

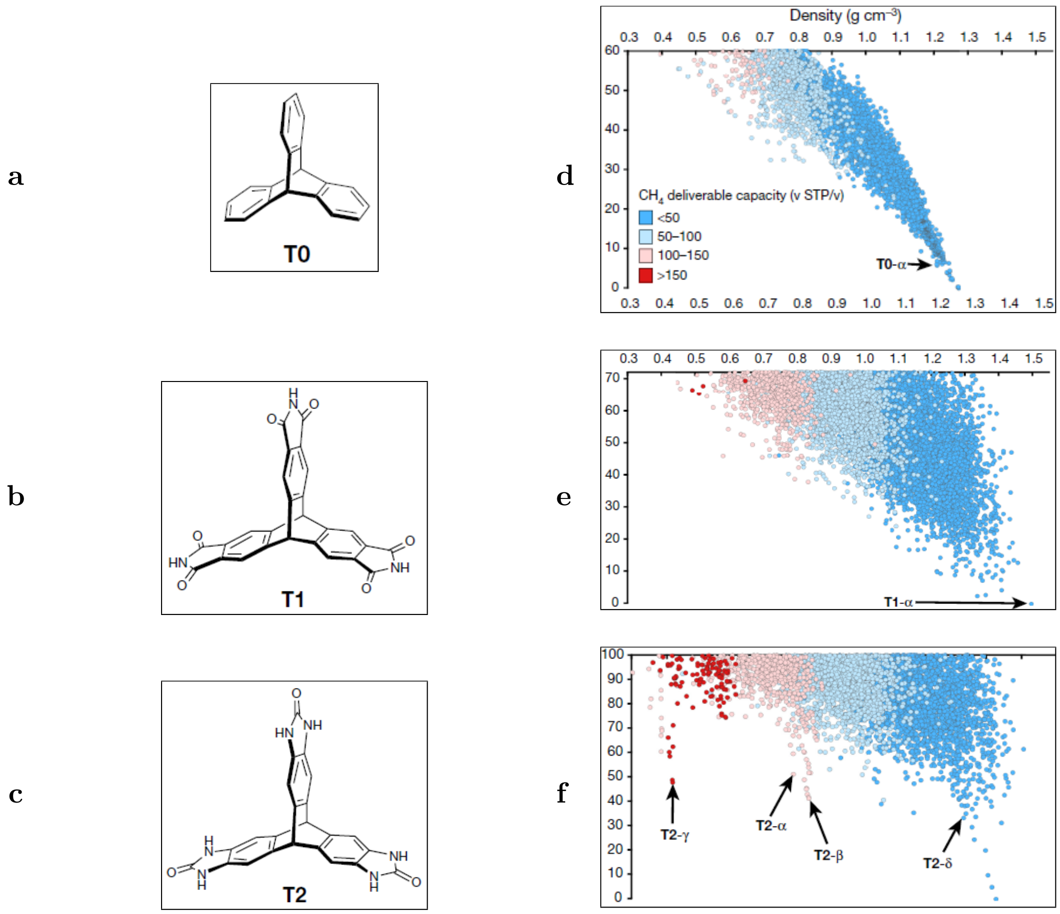


Figure 3.2: [37]. Target molecules (a-c) and their respective energy-density landscapes (d-f) coloured by methane capacity [37]

methane capacity concerning the previous synthesised T2-$\alpha$ (about 32% higher), confirming the more substantial potential for methane storage of the T2 simulated dataset.

Laboratory syntheses were guided by the energy landscapes in Fig 3.2d-e of simulated crystals. Only the density value was taken into account to distinguish crystal structures, which were only nano-porous organic crystals. Indeed, inorganic crystals, which may retain close-packed structures and consequently high density, are not suitable targets for this approach as they may not be separated appropriately. Using the density $\rho$ of an experimental crystal, one can search through multiple simulated crystals within a vertical 'stripe' of the energy-vs-density landscape in Fig. 5.5 over a small density interval to allow for errors.

One might take the simulated crystal with the lowest energy as the most likely structure from this stripe. As such, a result depends on the tolerance error for the density among other factors. A final match is confirmed by the non-invariant RMS deviation between finite portions of crystals, see Table 2.1.

# Chapter 4

# Voronoi-based distances between Crystal Lattices

The current chapter deals with the concept of metric (or distance function) between objects which aims to compare crystal lattices, and solve the metric problem 2.6 that is the key contribution of this first PhD project. It is based on our paper published in the Crystal Research and Technology journal [29]. The C++ code is available on my github account [28]. Before defining the developed metrics on crystal lattices, we need to explain the Voronoi Diagrams and how we can compute them.

## 4.1 Voronoi Diagrams

Nowadays, renewable energy sources are the main topic, which all countries in the world are discussing. The ecological transition will be necessary to keep our earth clean and tidy. To introduce the concept of Voronoi Diagrams, I will make a simple example on how it could be used in order to highlight its properties.

Consider that ecological transition has been found so necessary in a country that they decided to install more solar panels in the countryside to provide more cities with clean energy. To predict whether a new solar panel station will be advantageously located for inhabitants and not overload the central system, we must estimate the number of cities that will use it. In particular, we need to find the area for which each solar panel station is providing electricity. More generally, we have a set of places, called *sites*, that provides clean energy to the population, and we want to know for each site which city receives the electricity resource. To study this question, we need to follow the simple assumptions below:

- All cities have the same number of inhabitants
- All stations are working perfectly

The model above could give a very simple approximation; indeed, cities may have a different number of people living there. The cost may vary depending on cities' policies and solar stations may be damaged or not working at maximum performance. However, for our discussion, we will assure those simple assumptions. Finally, we are interested in **subdividing the territory into subregions** to find all areas close to the sites (or solar panel stations) where cities lie. It means that cities receive the electricity resource from the nearest site. The action area of a solar station consists of all those cities for which that station is closer than any other, as shown in Figure 4.1. The model where every point (or city) is assigned to the nearest site (or station) is called Voronoi assignment model. The subdivision induced by this model is called the **Voronoi Diagram** of the set of sites.

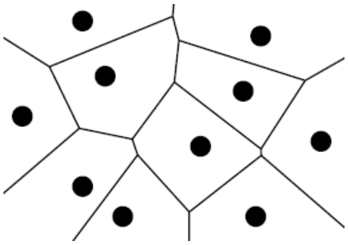


Figure 4.1: A Voronoi diagram of 11 sites or solar panel stations.

We can extract different geometric information about the cities' locations such as: which are those cities that receive electricity from 2 or more stations, or where is the closest city in a specific range or, moreover, which is the best location to install a new solar panel station taking into accounts the system overload and the area provided with electricity. In addition, this beneficial geometric structure finds its applications in different sectors such as physics, biology, robotics and other fields.

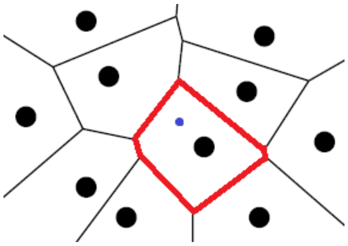


Figure 4.2: A Voronoi domain of a point cloud (red) and a possible city location (blue).

In the next chapter, Voronoi diagrams will be used to extract a Voronoi domain of a lattice which is just one area related to one site (Figure 4.2), however in the current chapter we will discuss how to compute this interesting geometric structure.

Let us start with some notation and definitions. Denote the Euclidean distance between two points $p$ and $q$ in $\mathbb{R}^3$ by $d(p,q)$ which is defined as follows:

$$d(p,q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2 + (p_z - q_z)^2} \tag{4.1}$$

**Definition 4.1.** (voronoi diagram). *Let $S$ be a set of $n$ distinct points $p_i$ in $\mathbb{R}^3$. The **Voronoi diagram VD** of $S$ is the subdivision of the space into $n$ domains $V(p_i)$, one for each point in $S$. A random point $q$ lies in a particular domain $V(p_i)$ if and only if $d(q,p_i) \leq d(q,p_j)$ for each site $p \in S$ with $j \neq i$. [4].*

We use the notation $VD$ to address the set of vertices and edges of the Voronoi structure and the symbol $V(p_i)$ to consider the single Voronoi domain (or the action area of the panel station) of a site $p_i$. The single Voronoi domain is computed by intersections that involve **halfspaces** defined below.

**Definition 4.2.** (halfspaces). *For $p,q \in S$ in $\mathbb{R}^3$, let us draw a segment $\overline{pq}$ between the points and mark the middle point $m$. Let $B(p,q)$ be the bisector plane perpendicular to $\overline{pq}$ and passing through $m$ that divides the space in half. The halfspace hs(p,q) contains all the points lying on one side of $B(p,q)$.*

$$hs(p,q) = \{x \mid d(p,x) \leq d(q,x)\} \tag{4.2}$$

A point $r \in hs(p,q)$ if and only if $d(r,p) \leq d(r,q)$. Therefore, the Voronoi domain of a point $p_i$ is computed by intersecting all $n-1$ halfspaces $hp(p_i,p_j)$:

$$V(p_i) = \bigcap_{1 \leq j \leq n, j \neq i} hs(p_i,p_j) \tag{4.3}$$

It could happen that some bisector planes are not involved in creating the Voronoi domain region because $V(p_i)$ may be generated with only a subset of $n-1$ bisector planes. It occurs when some of them lie beyond the newly formed polyhedron. In a Voronoi diagram, the common boundary of two Voronoi domains of $p_i$ and $p_j$ is called **Voronoi edge** $e$ which is basically part of the bisector plane between the two corresponding points, $e \subset hs(p_i,p_j)$. Let us call the points $p_i$ and $p_j$ in $S$ **Voronoi sites**. If we expand a ball starting from a point on the edge $e$, it will intersect both sites $p_i$ and $p_j$ at the same time. On the other end, if we expand a ball from a point $r$ inside the Voronoi domain of the site $p_i$, it will intersect the point $p_i$ before the others, hence $r \in V(p_i)$. If the ball expanding from point $r$ intersects three or more sites at the same time, then $r$ is a **Voronoi vertex** of the Voronoi diagram.

The intersection of all halfspaces $hs(p_i, p_j)$ is an expensive procedure that is more than quadratic with respect to the number of sites (even with halfplanes in $\mathbb{R}^2$). The next section discusses another method that could be adopted to faster compute the Voronoi diagram by using Delaunay Triangulations.

## 4.2 Triangulations

Before dealing with specific cases, let us find a different problem from the previous one to help the reader understand a triangulation and how we can use it.

We are given a set of points $S$ that represent solar panel stations placed in a territory. We want to install a closed network that connects all stations to keep the entire system online and capable of providing electricity to the population without interruptions. Hence, each station needs to exchange messages related to eventual malfunctions to the other neighbours to be replaced at once. In other words, the entire system should provide cities with a continuous service when some stations could be shut down for maintenance or run into sudden hardware or software failure.

To not connect all the solar stations directly to each other, it is better to install physical cables from each station to only a few neighbours. A simple partition of this space through triangles, called triangulation, may help identify each solar station's proper neighbours and build the network structure. Before defining formally the triangulation, we need to explain what is a convex hull.

**Definition 4.3.** (convex hull). *A point set $S$ is called convex if and only if for any pair of points $p, q \in S$ the line segment $\overline{pq}$ is completely contained in $S$. The convex hull of $S$ is the smallest convex set that contains $S$.*

An example of a convex hull in $\mathbb{R}^2$ is shown in (Figure 4.3a) and it could be imagined as a fence that encloses $S$ by surrounding all its points and everything inside. In $\mathbb{R}^3$, we can refer to a convex hull as the smallest box that would be enough to contain all points of $S$ considering even vertices, sides and edges of the box as spots where the points may lie.

**Definition 4.4.** (triangulation). *A triangulation $T$ of a set $S$ of points $p_i \in \mathbb{R}^3$ is a partition of the convex hull of $S$ into tetrahedra (or triangles in $\mathbb{R}^2$) whose vertices are the* *points of S.*

The partitioning of the convex hull of $S$ results in a graph $G(T_S)$ whose vertices $v$ are the points of $S$ and the physical cables are the edges $e$ connecting the solar stations. A triangulated point set $T_S$ is subdivided into simple regions such as triangles in $\mathbb{R}^2$ (or tetrahedra in $\mathbb{R}^3$) for a point cloud on a plane as shown in Figure 4.3b. Triangles partition the space and are not always incident to all vertices, which, as consequence, are directly connected only to a few neighbours.

## 4.3 Delaunay Triangulations

A triangulation is not unique because more than one type of partitions of a convex hull may exist. Changing the convex hull division leads to a change in the angles of different triangles. Let us denote $A_T$ a vector of angles from all triangles that belong to the triangulation $T$.

**Definition 4.5.** *A Delaunay Triangulation DT for a set of points S is a triangulation T such that no point lies inside the circumdisk passing through any 3 adjacent vertices of DT in* $\mathbb{R}^2$ *(or in a circumscribing ball passing through any 4 adjacent vertices in* $\mathbb{R}^3$*).*

To assure that every circumcircle passing through 3 adjacent vertices does not contain any other point of $S$ inside, each edge that belongs to the graph should be legalized or, better to say, the minimum angle of $T$ must be maximized. It means that if it does not satisfy the condition above, some edges should be replaced and different points may be allowed to form a connection by changing the partition type. Consider the 2D example in Figure 4.4, a partition of a 4-points set $P$ has the edge $\overline{p_ip_j}$ that makes the triangulation illegal because the circumcircle passing through $p_i, p_k, p_j$ contains the point $p_l$.

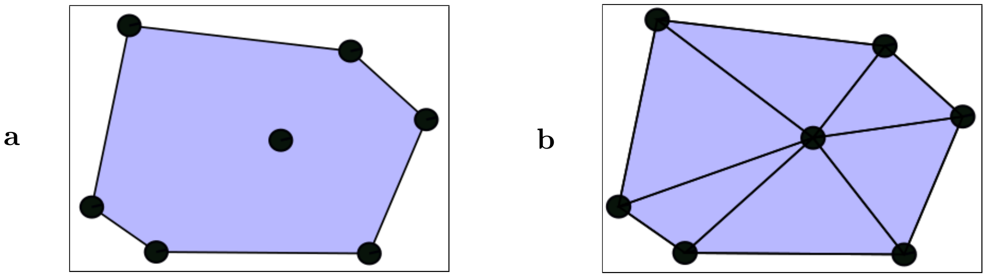


Figure 4.3: (a) The convex hull of S which contains all points inside the formed polygon. (b) Triangulation or partition of the convex hull of S.

The triangle $p_i, p_k, p_j$ is said to be in a conflict zone. In order to legalize an edge, a replacement should occur by flipping the edge $\overline{p_i p_j}$ itself and moving the link between other two vertices $\overline{p_k p_l}$. Given the ordered angle vector of the illegal triangulation $T$ on the left: $A^{il}_T = \{30, 30, 30, 35, 115, 120\}$, the resulting legalization led the minimum angle value (30°) to maximization. Indeed, the minimum angle in the angle vector of the Delaunay Triangulation $DT_P$, $A_{DT_P} = \{50, 55, 60, 65, 65, 65\}$, is 50°.

The Delaunay triangulation can be computed in a subquadratic time complexity $O(n^{(2d-1)/d})$ by using the Watson algorithm [57] that works in an arbitrary dimension $d$ on $n$ points. For simplicity consider the space $\mathbb{R}^2$. The main idea of this algorithm is to start with a triangle that fully contains all points of a set $S$. The algorithm begins with no points inside the triangle. We need to populate the outer triangle with all the points $s_i \in S$ and every time a point is added, new triangles are formed. So, one by one, take a point $s_i \in S$ and locate the triangle where it should be contained. Check all triangles in a conflict zone because of the insertion of $s_i$ and remove their edges. Finally, triangulate the remaining empty region with the new point and its neighbours to satisfy the non-empty circle condition. The number of searched regions that should be fixed depends on the location of the newly inserted point, which can lie inside a triangle or on its edges.

The Delaunay triangulation always exists for points in a generic position (e.g. all of them should not be co-planar or co-linear to each other). An algorithm based on points insertion can detect a non-optimal local case and modify the structure accordingly to satisfy the Delaunay condition of the circumsphere or circumdisk [57] [23].

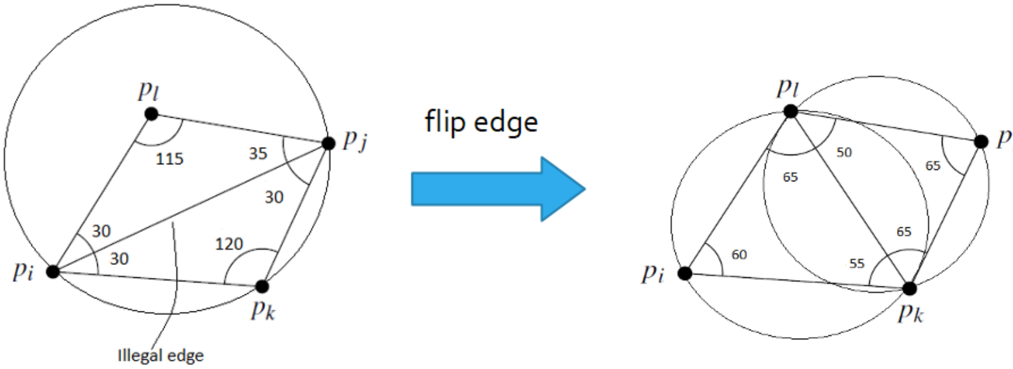


Figure 4.4: Illegal edge $\overline{p_i p_j}$ is legalized and replaced by $\overline{p_l p_k}$ in a 4-points set $P$.

## 4.4 Voronoi Diagram from Delaunay Triangulation

Definition 4.5 is based on the Delaunay empty-ball or empty-disk property, where there exists a circumscribing ball or circle whose interior does not contain any vertex of the triangulation. If this condition is satisfied in a triangulation, the Delaunay triangulation itself could be used as a starting point to generate the Voronoi Diagram of a point set.

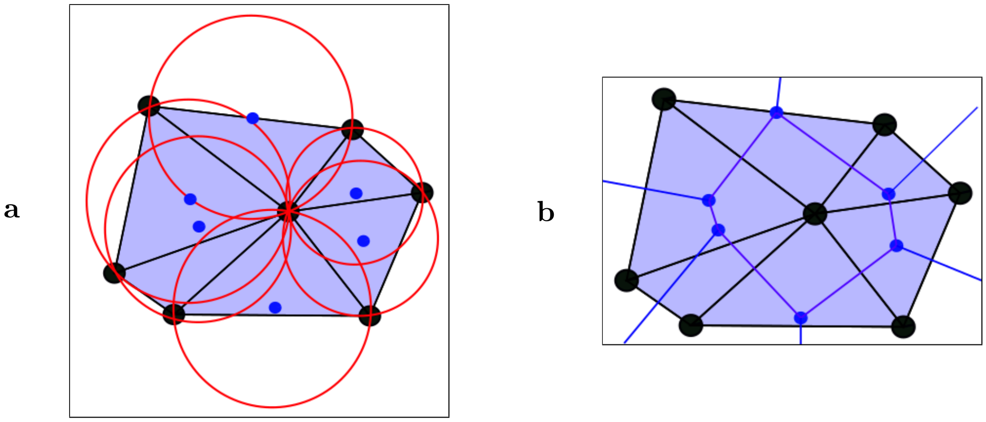


Figure 4.5: (a) The Delaunay triangulation of $DT_S$ (black line) and its circumcircles with centres (blue points). (b) Voronoi diagram (blue line) with Voronoi vertices in blue computed from the Delaunay triangulation $DT_S$.

According to Definition 4.1, a Voronoi Diagram is a subdivision of the space where a circumcircle expanding from each vertex $r$ intersects three or more sites. Hence, the distances of those sites to the Voronoi vertex $r$ are the same. Moreover, definition 4.1 assures that the Voronoi diagram takes into consideration a **balanced subdivision** among all the space, where all Voronoi vertices are **equidistant** from the closest sites. To confirm the balanced subdivision of the space, each circumdisk that intersects three or more sites should not contain any other site. If it was the case, the vertex $r$ would not have an equal distance to all its closest sites and the definition breaks. This is exactly the **key-property** of the Delaunay triangulation which does not have any points inside the circumdisk of any 3 or more adjacent vertices. Therefore, starting from a Delaunay triangulation we can generate the Voronoi Diagram of a point set by finding the circumcentre of each triangle (or tetrahedron). All triangles in Figure 4.5a are not in a conflict zone because every circumdisk does not contain any other black point, hence all edges are legalized. The circumcentre of each triangle is a Voronoi vertex of the Voronoi diagram and it is connected to its Voronoi neighbour only if the triangles inscribed into the circumcircles have a common edge as shown in Figure 4.5b. The resulting graph is called **dual graph** of the Delaunay triangulation.

Starting from the Delaunay triangulation of a point set in $\mathbb{R}^3$, all polyhedra should be iterated to compute the circumcentres and build the Voronoi diagram. The number of polyhedra may vary from $\Theta(n)$ to $\Theta(n^{\lceil \frac{d}{2} \rceil})$ concerning $n$ points [44] and the time complexity of computing the Delaunay triangulation is $O(n^{(2d-1)/d})$ [57] with $d$ number of dimensions.

## 4.5 Voronoi Domain of a point of a lattice

The geometric structure of a Voronoi Diagram is used for many applications. We decided to use this geometric object as a tool to develop two distance functions or metrics by using a Voronoi Domain of a Lattice. The Voronoi Domain of a point $O$ of a lattice $L$ is a convex and centrally symmetric polyhedron, or better to say a convex polyhedron that is symmetric with respect to its centre. It means that a plane passing through its centre may split the polyhedron in two specular parts. It has several faces equal to the number of closest neighbours of $O$. This geometric object allows us to have local information of $O$ neighbourhood encoded in its shape.

**Definition 4.6.** *Given a fixed origin point $O$ in a lattice $L \subset \mathbb{R}^3$, the* **Voronoi Domain** *$V_O(L)$ is the set of all points $p \in \mathbb{R}^3$ that are (not necessarily strictly) closer to $O$ (in the usual Euclidean metric) than to all other points of $L$. [4].*

$$V_O(L) = \{p \in \mathbb{R}^3 \mid d(p, O) \leq d(p, q) \; for \; any \; q \in L \setminus O\} \tag{4.4}$$

Since all points in a lattice are translationally equivalent, the Voronoi Domain of point $O$ will be the same for any other point. Therefore, it can cover all the space by its tessellation (Figure 4.6a). We will use the notation $V(L)$ to address the Voronoi domain of a generic point in the lattice.

We want our isometry invariant to rely on a stable object, continuously changing under lattice perturbations. The Voronoi Domain of any lattice is not combinatorially stable because lattice perturbations can lead to a different underlying graph structure. Hence the number of vertices may change from 4 to 6 in a 2D square example (Figure 4.6b) or from 8 to 12 in a 3D cube example (Figure 4.6c). Nevertheless, it can be defined as a geometrically stable object if we consider its shape. Indeed, if a linear basis is slightly perturbed, it can lead to a small change from a rectangular to a hexagonal shape (Figure 4.6b) or from a cubic to a truncated cubic shape (Figure 4.6c). A stability theorem can be found in [40] which was proved by D. Reem and links Voronoi Domains with their corresponding lattices. The geometric stability of the Voronoi Domain is necessary for comparison because it allows us to assure the stability of our lattice metrics.

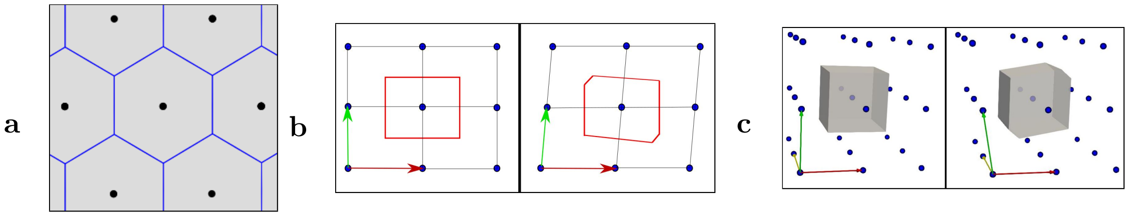


Figure 4.6: (a) The Voronoi Domains of points of a 2D hexagonal lattice. (b) 2D square lattice with a square Voronoi Domain perturbed to a hexagon. (c) Cubic lattice with a cubic Voronoi Domain perturbed to a truncated cube.

## 4.6 Voronoi-based metrics between arbitrary crystal lattices

### 4.6.1 Neighbourhood of a point set

Before going through the definitions, we need to explain some concepts needed to define our metrics. The first concept is the **neighbourhood of a point set**.

**Definition 4.7.** (neighbourhood). *Formally, given any subset $C \subseteq \mathbb{R}^n$, its r-offset neighbourhood N(C; r) is the set of all points $p \in \mathbb{R}^n$ that are at distance at most r from $C$.*

$$N(C;r) = \{p \in \mathbb{R}^n \mid d(p,q) \leq r \;\; for \;\; some \;\; q \in C\} \tag{4.5}$$

If $C$ has only a point, *N(C; r)* will be a closed ball with radius *r*. If we consider a point set that forms a convex polyhedron, *N(C; r)* will contain the area around and the polyhedron itself. In Figure 4.7, a 2D example is shown where the point in $C$ has a neighbourhood defined by a yellow disk of radius *r*, including the point itself, and the hexagon has a neighbourhood defined by the yellow zone around it with the hexagon itself.

### 4.6.2 Hausdorff metric

We defined the first metric based on the already known **Hausdorff metric $d_H$** which measures how far two point sets $A$ and $B$ are from each other. The main idea is to try to include point set $A$ into $B$ and vice versa by considering a specific offset $r$ between them, but knowing that we want to make a minimum effort to cover both. When we say that a

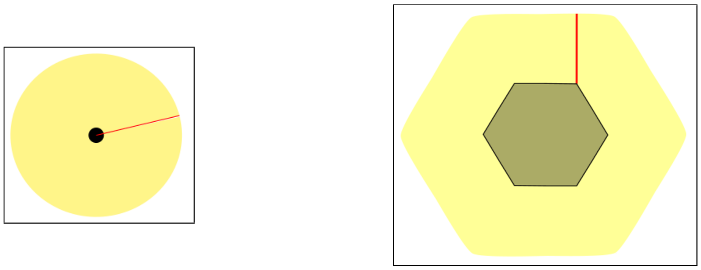


Figure 4.7: Left: The yellow neighbourhood *N(C; r)* where $C$ is a point. Right: The yellow neighbourhood *N(C; r)* where $C$ is a hexagon.

set $A$ covers a set $B$, we mean that the union of all balls with radius $r$, drawn from each point of $A$, includes the whole set $B$. Therefore, we need to find an $r$-offset to compute the Hausdorff metric defined as follows:

$$d_H(A, B) = \min\{r \geq 0 : B \subset N(A; r) \; and \; A \subset N(B; r)\} \tag{4.6}$$

where *N(A; r)* is the neighbourhood of $A$ with offset $r$. The Hausdorff metric outputs a value that quantifies the similarity between fixed point sets where no rigid motion is applied. Whereas, the new metrics defined in the following paragraphs consider both point sets over all rotations.

### 4.6.3 Rotationally-invariant metric $d_R$

The previous *Hausdorff metric* was defined for fixed point sets where no isometry is taken into account. For example, let us consider two isometric hexagons located in different positions with their centres being at distance $r$ from each other. The Hausdorff metric between hexagon $H_1$ and hexagon $H_2$ will be $r$ as well, although both polygons are isometric. Therefore, we defined an extended version of the Hausdorff metric that detects equivalent and similar lattices through their Voronoi polyhedra. Two polyhedra are the input point sets. They need to have the same centre to minimize the $r$-offset and so, they need to be shifted or translated by a vector $\vec{v}$. A translation $T_v$ of a vector set $P$ is the set of vectors $P_{T_v}$ resulted from the sum of $\vec{v}$ applied to any vector of $P$.

$$P_{T_v} = \{\vec{u} + \vec{v} \mid for \; any \; \vec{u} \in P\}$$

**Lemma 4.8.** *For any centrally symmetric polyhedra* $P, P' \subset \mathbb{R}^n$ *and a translation* $T_v$ *by a vector* $\vec{v} \in \mathbb{R}^n$*, the offset parameter* $\min\{r : T_v(P) \subset N(P'; r)\}$ *is minimal when* $T_v$ *moves the centre c(P) of the polyhedron P to the centre c(P') of P'.*

*Proof.* Assume by contradiction that $\min\{r : T_v(P) \subset N(P'; r)\}$ is minimized for a vector that differs from $c(P') - c(P)$. Without loss of generality, one can assume that $v = 0$ and $r = 0$ because we consider $P$ completely included in $P'$ and we want to prove that if $P \subset P'$, then this inclusion is preserved when the centre $c(P)$ is shifted to $c(P')$. Let us consider Figure 4.8 with $P$ and a symmetric copy $S(P)$ of it that is mirrored over the centre $c(P')$. The polyhedron $P'$ remains at the same position and covers the symmetric image $S(P)$ of $P$. Here, it is allowed to talk about symmetric images because Voronoi domains of lattices are centrally symmetric. If we move the centre $c(P)$ by continuous motion to its symmetric image $S(P)$ through the centre $c(P')$, all intermediate images of $P$ remain covered by $P'$ due to the convexity of $P'$. Indeed, any two points belong to $P'$ together with the line segment connecting them. Hence, the polyhedron $P$ shifted by the vector $c(P') - c(P)$ is also covered by $P'$. Symmetrically, if one fits a translational image of $P'$ into a minimal offset of $P$, then an optimal translation should make the centres of $P, P'$ identical. □

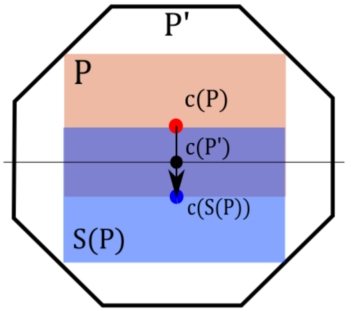


Figure 4.8: A centrally symmetric polyhedron $P'$ covers a centrally symmetric polyhedron $P$. Then the symmetric image of $P$, $S(P)$, is also covered by $P'$

Secondly, consider both polyhedra over all rotations to pick the best one that will quantify their similarity or assure their equivalence. Rotations are defined by square and orthogonal matrices $n \times n$ that belong to the group of rotations $\mathrm{SO}(n)$ with $n$ number of dimensions.

To finally design our metric we need to define the $r$-**offset** distance for polyhedra shifted to the same origin and subject to all rotations.

**Definition 4.9.** ($r$-offset). *Given two crystal lattices $L$ and $L'$ and their respective Voronoi Domains $V(L)$ and $V(L')$ at the origin $O \in \mathbb{R}^3$, we define the* minimum offset $r$ *over all rotations $R$:*

$$\mathit{offset}(L, L') = \min\{r \geq 0 : R(V(L)) \subset N(V(L'); r)\} \tag{4.7}$$

where the minimum is taken over all rotations $R \in SO(3)$ of $V(L)$ about the origin in $\mathbb{R}^3$. In Figure 4.9 an example is shown where we are trying to compute the *offset*$(L_{hex}, L_{rect})$ between two lattices $L_{hex}$ and $L_{rect}$ that have, respectively, a hexagonal and a rectangular Voronoi domain. $L_{rect}$ is fixed and $L_{hex}$ is considered over all rotations. In practice, the offset *offset*$(L_{hex}, L_{rect})$ is the minimum effort needed to cover all vertices of the hexagon inscribed in a neighbourhood of the rectangle defined by $r$-offset.

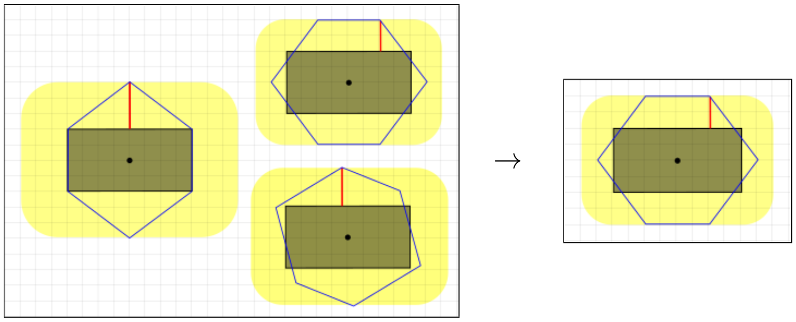


Figure 4.9: Left: Possible rotations of the hexagon inside the respective $r$-offset neighbourhoods of the rectangle. Right: The minimum offset *offset*$(L_{hex}, L_{rect}) = 2$ (red segment) between a hexagonal $V(L_{hex})$ and a rectangular $V(L_{rect})$ Voronoi domains over all rotations.

**Definition 4.10.** (rotationally-invariant metric). *Finally, the **rotationally-invariant metric** between two lattices will be the symmetric maximum between two offsets:*

$$d_R(L, L') = \max\{\ \mathit{offset}\ (L, L'),\ \ \mathit{offset}\ (L', L)\} \tag{4.8}$$

The Voronoi domain is defined in terms of distances to lattice nodes and computed from the unit cell vectors stored in the CIF file of a crystal. Since the unit cell parameters are stored in Angstroms, and the $offsets$ are found through the Euclidean distance between nodes, the computation of $d_R$ will output a value with the Angstrom measure unit. To continue with the previous example of the hexagon and rectangular Voronoi domains, Figure 4.10 shows the final result of the rotationally-invariant metric in Angstroms. The second offset $\mathit{offset}(L', L)$ is already computed on the right where $V(L')$ rotates and $V(L)$ is fixed.

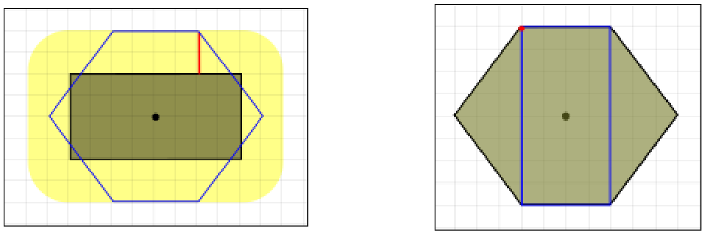


Figure 4.10: The two non-symmetric offsets $\mathit{offset}(L_{hex}, L_{rect}) = 2$ on the left and $\mathit{offset}(L_{rect}, L_{hex}) = 0$ on the right will give a rotationally-invariant metric $d_R(L_{hex}, L_{rect}) = 2$

The *rotationally-invariant metric* can be considered independent of a lattice representation. It means that although two crystal lattices may be represented by different unit cell vectors, the metric will output the same value as their representation changes (i.e. different angles or vector lengths). It satisfies the identity, symmetry and triangle inequality axioms together with the continuity condition as stated and proved in the following theorem.

**Theorem 4.11.** *The rotationally-invariant metric $d_R$ is independent of a lattice representation and satisfies the axioms 2.5(a), 2.5(b), 2.5(c) and condition 2.5(d).*

*Proof.* The rotationally-invariant metric between lattices is based on the Voronoi domains, which are defined in terms of distances to lattice nodes, hence are independent of a linear basis of a lattice. By equation 4.7 $d_R(L, L')$ is always not negative and equals 0 only when there is a rotation $R$ that matches $R(V(L)$ with $V(L')$, hence the Voronoi domains $V(L), V(L')$ become identical under the rotation $R$, so the lattices $L, L'$ are equivalent, which proves axiom 2.5(a). Axiom 2.5(b) follows from equation 4.8 taking the maximum of two offsets when $L, L'$ are swapped. To check axiom 2.5(c), one can assume that maxima in equation 4.8 are attained on first offsets. Let us fix two rotations $R, R'$ so that $d_H(L, L') =$

$\mathit{offset}(R(V(L)), V(L'))$ and $d_H(L', L'') = \mathit{offset}(R'(V(L')), V(L''))$. In practice, consider just 1 rotation for $V(L)$ and 1 for $V(L')$ and no more transformations. Both refer to a minimum inclusion offsets $r$ and $r'$ as follows. By equations 4.6, 4.7, 4.8 the first Hausdorff metric $d_H(L, L')$ above has a minimum value (say, $r$) when

$$R(V(L)) \subset N(V(L'); r) \; \mathit{and} \; V(L') \subset N(R(V(L)); r) \; \mathit{or} \; R^{-1}(V(L')) \subset N(V(L); r) \quad (4.9)$$

Similarly, the second Hausdorff metric $d_H(L', L'')$ has a minimum value (say, $r'$) when

$$R'(V(L')) \subset N(V(L''); r') \; \mathit{and} \; V(L'') \subset N(R'(V(L')); r') \; \mathit{or} \; (R')^{-1}(V(L'')) \subset N(V(L'); r') \quad (4.10)$$

When applying the first rotation, the offset is set to a minimum value $r$ where both Voronoi domains $V(L)$ and $V(L')$ are included in each other's neighbourhood. The Hausdorff metric with the second optimal rotation outputs an $r'$ value that includes $V(L')$ in $V(L'')$' neighbourhood and vice versa. Given the previous rotations, their composition (applied from right to left) rotates the Voronoi domain $V(L)$ to the position $R'(R(V(L)))$. The first inclusions from 4.9, 4.10 above imply the inclusion below:

$$R'(R(V(L))) \subset R'(N(V(L'); r)) = N(R'(V(L')); r) \subset N(N(V(L''); r'); r) = N(V(L''); r+r') \quad (4.11)$$

Similarly, the opposite composition of rotations $R^{-1}(R')^{-1}$ (applied from right to left) rotates the Voronoi domain $V(L'')$ to $R^{-1}((R')^{-1}(V(L'')))$. In equation 4.11, one rotation causes the output of $r$ of $V(L)$ with respect to $V(L')$, but the second rotation results in the offset $r'$ of $V(L')$ with respect to $V(L'')$. The axiom 2.5(c) wants that both rotations are applied to the first domain $V(L)$, to demonstrate that it is included in the neighbourhood of the third domain $V(L'')$. The inclusion is possible since offsets may be added. Indeed, take the example when the second offset $r' = 0$. It means that $V(L')$ is equivalent to $V(L'')$ (no bigger neighbourhood needed), therefore $V(L)$ does not need any other than $r + 0 = r$ to be included in $V(L'')$. The sum increases if the $V(L')$ and $V(L'')$ are different and need a wider neighbourhood. Inclusions 4.9, 4.10 mean that the Hausdorff metric $d_H(L, L'')$ has the upper bound $r + r'$ attained for the rotations $R'R$ and $R^{-1}(R')^{-1}$. The minimum over all rotations can be even smaller, hence the triangle inequality $d_R(L, L'') \leq r + r'$ in axiom 2.5(c) holds. Continuity condition 2.5(d) follows from the stability of Voronoi domains in Section 4.5. □

### 4.6.4 Scale-invariant metric $d_S$

We prove that the rotationally-invariant metric $d_R$ satisfies all metric axioms and is continuous under perturbations of unit cell parameters. In this section we introduce the new notation $s \times P$ used when a point set $P$ is scaled uniformly by a factor $s$. The following scale-invariant metric is additionally invariant under uniform scaling of a lattice. As for the previous metric, the minimum can be obtained by shifting the polyhedra to a common centre as the lemma below states:

**Lemma 4.12.** *For any centrally symmetric polyhedra $P, P' \subset \mathbb{R}^n$ and a translation $T_v$ by a vector $\vec{v} \in \mathbb{R}^n$, the scale factor* $\min\{s > 0 : T_v(P) \subset s \times P'\}$ *is minimal when $T_v$ moves the centre c(P) of the polyhedron P to the centre c(P') of P'.*

*Proof.* Proof of this lemma is similar to the proof of Lemma 4.8 with $r$-offset $N(P; r)$ replaced by scaled polyhedra $s \times P \subset \mathbb{R}^n$, because all other inclusion and convexity arguments remain valid. □

**Definition 4.13.** (scale). *Given two Voronoi Domains of crystal lattices V(L) and $V(L')$, we want to find the minimum scale factor s over all rotations as follows:*

$$scale(L, L') = \min\{s > 0 : R(V(L)) \subset s \times V(L')\} \tag{4.12}$$

where $R \in SO(3)$ and $SO(3)$ is the group of all rotations in $\mathbb{R}^3$. Similarly to the previous *offset*, the *scale* measure is computed over all rotations by fixing one polyhedron $V(L')$ and rotating the other one $V(L)$. Instead of considering a neighbourhood of the fixed domain, in the *scale* function we consider a scaled version of the fixed domain that should include the rotating domain.

Let us take the example in Figure 4.11 where the hexagonal $V(L_{hex})$ and rectangular $V(L_{rect})$ Voronoi domains are reproposed and adapted for the scale measure. Two segments are considered for each rotation to find the $scale(L, L')$. The first segment (red lines) is between the centre of both Voronoi domains $O$ and a vertex $p$ of the rotating domain $d(O, p)$, which intersect a side of the fixed domain at point $q$. The second (green lines) is the segment between the centre $O$ and point $q$, $d(O, q)$. Since the *scale factor s* is a ratio between the two distances in Angstroms, it will result in a dimension-less measure which represents the number of times ($s$) a fixed domain should be enlarged to cover all vertices of the rotating domain. Practically, the scale $scale(L_{hex}, L_{rect})$ is the minimum effort needed to cover all vertices of the hexagon inscribed in the $s$-scaled version of the rectangle.

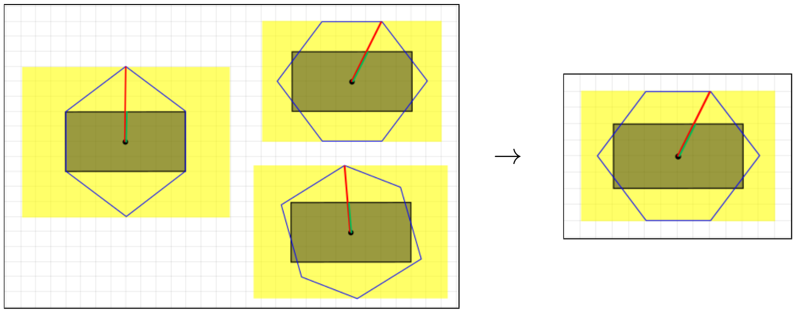


Figure 4.11: Left: Possible rotations of the hexagon inside the respective $s$-scaled versions of the rectangle. Right: The minimum scale $scale(L_{hex}, L_{rect}) = \frac{2\sqrt{5}}{\sqrt{5}} = 2$ (red over green distances) between a hexagonal $V(L_{hex})$ and a rectangular $V(L_{rect})$ Voronoi domains over all rotations.

**Definition 4.14.** (scale-invariant metric). *Finally, the* ***scale-invariant metric*** *is defined as the logarithm of the symmetric maximum between two scale measures:*

$$d_S(L, L') = \ln\{\max\{scale(L, L'), scale(L', L)\}\} \tag{4.13}$$

where the logarithm is in base $e$ which is needed to make the metric additive and map isometric (or equivalent) lattices to 0. Figure 4.12 shows the results of the scale-invariant metric on the comparison between the hexagonal and rectangular Voronoi domains. Moreover, the *scale-invariant metric* can be considered independent of an isometry class and satisfies the identity, symmetry and triangle inequality axioms together with continuity and scaling conditions as stated and proved in the following theorem.

**Theorem 4.15.** *The scale-invariant metric $d_S$ is independent of a lattice representation and satisfies the axioms 2.5(a), 2.5(b), 2.5(c), and both conditions 2.5(d), 2.5(e).*

*Proof.* Similarly to the proof of Theorem 4.11, the scale-invariant metric between lattices

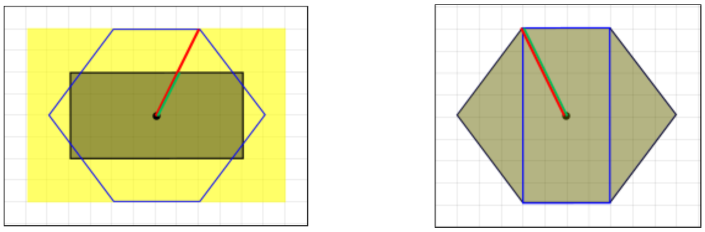


Figure 4.12: The two non-symmetric scales $scale(L_{hex}, L_{rect}) = \frac{2\sqrt{5}}{\sqrt{5}} = 2$ on the left and $scale(L_{rect}, L_{hex}) = \frac{2\sqrt{5}}{2\sqrt{5}} = 1$ on the right will give a scale-invariant metric $d_S(L_{hex}, L_{rect}) = \ln 2$

is based on the Voronoi domains and is independent of a lattice representation. The scaling factor is the main feature of this metric that refers to the ability to recognise scaled versions of specific lattices. To check that $d_S(L, L') \geq 0$, let $R, R'$ be optimal rotations that minimize the factors $s = scale(L, L')$ and $s' = scale(L', L)$, respectively. Equation 4.12 implies that

$$R'(R(V(L))) \subset R'(s \times V(L')) = s \times R'(V(L')) \subset s \times s' \times V(L)$$

The composition of both rotations applied to $V(L)$ does not change the volume of $V(L)$ and $R'R(V(L))$ and $V(L)$ are isometric. Therefore, the inclusion

$$R'(R(V(L))) \subset s \times s' \times V(L)$$

implies that $s \times s' \geq 1$, and $\ln\{\max\{s, s'\}\} \geq 0$. The equality is possible only if $s = s' = 1$. It means that no scaling operation needs to be applied and $V(L), V(L')$ can be found from each other by applying a rotation, hence the lattices $L$, $L'$ are equivalent, so axiom 2.5(a) is proved. Axiom 2.5(b) follows from symmetric equation 4.13. To check axiom 2.5(c), let us fix optimal rotations $R, R'$ such that $scale(L, L')$ and $scale(L', L'')$, and consider minimum values of scaling (say, $s$ and *s’*, respectively) when

$$R(V(L)) \subset s \times V(L')\, and\, R'(V(L')) \subset s' \times V(L'')$$

Then

$$R'(R(V(L))) \subset R'(s \times V(L')) = s \times R'(V(L')) \subset s \times s' \times V(L'')$$

Similarly to the $d_R$, take the example of $s > 1$ where $V(L)$ needs a wider scale of $V(L')$ to be contained, and $s' = 1$ which refers to the fact that $V(L')$ and $V(L'')$ are equivalent. If $V(L)$ is included in $V(L')$, it does not need any other scale factor to be contained in $V(L'')$ because $s \times 1 = s$. Differently, if the second couple outputs a higher scale, the total scale factor will be affected. Hence, $scale(L, L'') \leq s \times s'$, because an optimal rotation from $V(L)$ to $V(L'')$ may have a smaller scale. The symmetric $scale(L'', L)$ has a similar upper bound from optimal rotations or $scale(L'', L)$ and $scale(L', L)$. The triangle inequality follows after taking the logarithm of both sides:

$$\begin{aligned}&\max\{scale(L, L''), scale(L'', L)\} \leq \\ &\qquad \leq \max\{scale(L, L'), scale(L', L)\} \times \max\{scale(L', L''), scale(L'', L')\} \qquad (4.14)\end{aligned}$$

To prove continuity condition 2.5(d), let us fix two Voronoi Domain and define the distance that we require. Let $r(L)$ be the distance from the origin $O \subset L$ to the boundary of the Voronoi domain $V(L)$ of which we consider its neighbourhood. The geometric stability of Voronoi domain, in Section 4.5, confirms that the Voronoi domain $V(L')$ of a perturbed lattice is included in the $r$-offset $N(V(L); r)$ of $V(L)$ for a small $r > 0$. Let us assume that their centres are superimposed and coincide with the origin in $\mathbb{R}^n$. We may want to use the neighbourhood definition to find the correspondent upper bound of a scale factor, indeed by fixing $r$ and $r(L)$. For any $p \in N(V(L); r)$, let $R(O, p)$ be the straight ray emanating from $O$ and passing through $p$. Let $q$ be the intersection of $R(O, p)$ with the boundary of $V(L')$. The total distance from the origin $O$ to the furthest point of $V(L')$ is $d(p, O) = d(p, q) + d(O, q)$, hence the $scale(L', L)$ is equal to

$$scale(L', L) = \frac{d(p, q) + d(O, q)}{d(O, q)}$$

which is upper bounded by the correspondent ratio of $r$ and $r(L)$ as follows

$$scale(L', L) = \frac{d(p, q) + d(O, q)}{d(O, q)} \leq \frac{r + r(L)}{r(L)}$$

Then the Voronoi domain V(L’) is included in the neighbourhood of V(L) and consequently in its respective scaled version

$$V(L') \subset N(V(L); r) \subset \left(1 + \frac{r}{r(L)}\right) V(L)$$

Therefore

$$scale(L', L) \leq \left(1 + \frac{r}{r(L)}\right)$$

Swapping *L and L'* and computing the $scale(L, L')$, we get the upper bound for the scale-invariant metric

$$d_S(L, L') \leq \ln\left(1 + \frac{r}{\min\{r(L), r(L')\}}\right)$$

which means that *L'* remains close to *L*. The scale-invariant in condition 2.5(e) holds by equation 4.12, because the inclusion $R(V(L)) \subset s \times V(L')$ remains unchanged then both lattices *L, L'* are simultaneously scaled by the same factor. □

### 4.6.5 Metric Algorithms

In order to quantify lattice similarities, we need to consider their Voronoi domains over all rotations. In our algorithm we use a **uniform rotation sampling** procedure depending on a parameter *n*. The Voronoi domain of a lattice is a centrally symmetric polyhedron, and thanks to this property, we can limit the number of rotations to a certain amount. Let us consider the unit sphere in Figure 4.13. Each rotation is performed in the 3D space over an **axis of rotation** (which is a unit vector). The axes shown in the figure represent two axes of rotations with a different height from the plane *x-y*. The parameter *n* defines the number of intervals of the *z*-axis where axes fall at the centre in order to find a set of axes of rotations, with specific **height** and **orientation**, around which the polyhedron is permitted to rotate. The symmetry of this type of convex polyhedron allows us to restrict the number of rotations to half the sphere meaning that the generated axes of rotations belong only to the upper hemisphere. Therefore, a polyhedron may rotate around the axis of rotation at about $\theta$ and around the *z*-axis at about $\mu$.

**Theorem 4.16.** *For any polyhedra* $P, P' \in \mathbb{R}^n$ *symmetric with respect to the origin O, offset*$(P, P') = \min\{r \geq 0 : P \subset N(P'; r)$ *and scale*$(P, P') = \min\{s > 0 : P \subset s \times P'\}$ *can be computed in a linear time with respect to the number of vertices and faces of P, P'.*

*Proof.* The minimum distance *r* in the offset $offset(P, P') = \min\{r > 0 : P \subset N(P'; r)\}$

is updated with the minimum value by iterating all vertices $v$ of $P$ as follows. Find the intersection of the line segment $[0, v]$ from the origin $O$ to $v$ with a face $F$ of $P'$. If there is such an intersection, then $r$ increases to the distance $d(v, F)$. Similarly, $\text{scale}(P, P') = \min\{s > 0 : P \subset s \times P'\}$ is computed by finding the minimum scale value according to which each vertex $v$ of $P$ is inside $P'$. Detect the intersection of the ray $R(O, v)$ passing, from the origin $O$, through $v$ with a face $F$ of $P'$, then $s$ increases to $\frac{d(O,v)}{d(O,R(O,v)\cap F)}$. In the worst case, intersecting lines through vertices of $P$ and flat faces of $P'$ requires a loop over all vertices and a loop over all faces. An upper bound for the asymptotic complexity is the product of the numbers of vertices and faces, which is a linear function in each number. Indeed, those numbers are in a fixed range, as explained below. □

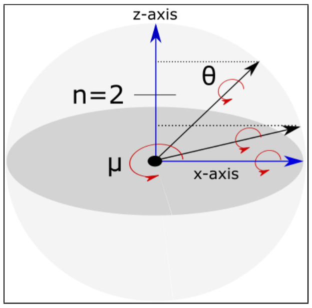


Figure 4.13: Rotation sampling example from a unit sphere with $n = 2$. In addition to $x$-axis and $z$-axis, two rotation axes are generated, around which the Voronoi domain may rotate.

The Voronoi domain of a lattice in $\mathbb{R}^3$ may have at least 8 vertices and 6 faces and at most 24 vertices and 14 faces. It may rotate about $\theta \in [0, 360°)$ around a unit length rotation axis $\vec{v}$ in the upper hemisphere defined as follows:

$$\vec{v} = (\sqrt{1 - z^2} \cos \mu, \sqrt{1 - z^2} \sin \mu, z) \tag{4.15}$$

where $z \in (0, 1)$ is the height parameter which has $n$ samples as decided by the user input. Each of the $n$ rotation axes generated with a specific height is rotated by $\theta$ and $\mu$ accounting for $2\pi n$ samples for each angle. In total, about $4\pi^2 n^3$ rotations will be sampled. In our experiments, we used several samples equal to 1065 given by $n$=3. Each vertex $\vec{u}$ of the Voronoi domain is rotated by the following **Rodriguez formula** around a rotation axis $\vec{v}$

by $\theta$:

$$R(\vec{u}) = \vec{u}\,\cos\theta + (\vec{v}\times\vec{u})\,\sin\theta + \vec{v}\,(\vec{v}\times\vec{u})(1-\cos\theta) \tag{4.16}$$

Let $R_{samples}$ be the set of all rotation samples generated with the previous method, the algorithms used to compute the above-mentioned metrics $d_R$ and $d_S$ are shown below with a pseudo-code.

**Algorithm 1** Rotationally-invariant metric

```
procedure Offset($L$, $L'$)
    $R_{samples} = \{R_i \text{ is a rotation sample s.t. } R_i \in SO(3)\}$
    $offset \leftarrow -1$
    for $R_i \in R$ do
        $dist_{rot} \leftarrow r$                        ▷ with $r \geq 0 : R_i(V(L)) \subset N(V(L'); r)$
        if ($offset$ == -1) then
            $offset \leftarrow dist_{rot}$
        else if ($dist_{rot} < offset$) then
            $offset \leftarrow dist_{rot}$
        end if
    end for
    return $offset$
end procedure
procedure RID $d_R(L, L')$
    offset$_{LL'} \leftarrow$ OFFSET$(L, L')$
    offset$_{L'L} \leftarrow$ OFFSET$(L', L)$
    return max(offset$_{LL'}$,offset$_{L'L}$)
end procedure
```

**Algorithm 2** Scale-invariant metric

```
procedure Scale($L$, $L'$)
    $R_{samples} = \{R_i \text{ is a rotation sample s.t. } R_i \in SO(3)\}$
    $scale \leftarrow -1$
    for $R_i \in R$ do
        $dist_{rot} \leftarrow s$                        ▷ with $s > 0 : R_i(V(L)) \subset s \times V(L')$
        if ($scale$ == -1) then
            $scale \leftarrow dist_{rot}$
        else if ($dist_{rot} < scale$) then
            $scale \leftarrow dist_{rot}$
        end if
    end for
    return $scale$
end procedure
procedure SID $d_S(L, L')$
    scale$_{LL'} \leftarrow$ SCALE$(L, L')$
    scale$_{L'L} \leftarrow$ SCALE$(L', L)$
    return ln(max(scale$_{LL'}$,scale$_{L'L}$))
end procedure
```

## 4.7 Results and conclusions

When comparing crystal lattices, structure similarity is satisfied when $d_{\mathrm{S}}(L, L')$ or $d_{\mathrm{R}}(L, L')$ are close to 0. We performed some experiments on a manually curated version of T2 dataset made of 5688 structures using our metrics to assess the similarity between crystal lattices. There is following a plot with the heatmap of the first 100 lowest-Energy crystals compared by our rotational-invariant metric and scale-invariant metric [Figure 4.14]. A white colour highlights the close similarity. On the other hand, different crystal pairs appear in black.

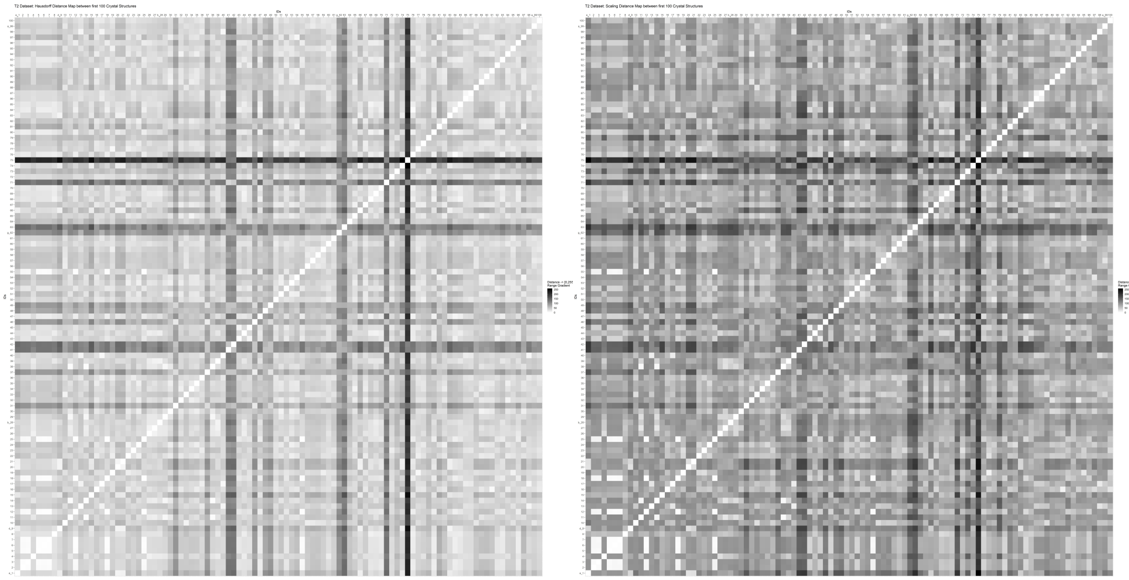


Figure 4.14: [29] Fig. 7. Heatmaps of the Voronoi-based geometric invariants. On the left the rotationally-invariant metric and on the right the scale-invariant metric.

We want to show that our metrics can find geometric differences in crystal lattices that could not be addressed by energy or density values. The aim was to emphasize that the usual approach for studying crystal structures should be based on geometric information and chemical data separately. The following experiment shows that energy or density values cannot be used to imply the geometric similarity. Indeed, we could highlight it only by looking at the differences between crystal lattices.

We applied our metrics to the T2 dataset made of 5688 simulated crystal structures. Among the results, we looked for pairs of crystals very close in energy and density values but very different in the values of our metrics. The search resulted in several pairs of crystals, e.g., with IDs (41,47), (68,71), (63,73), (71,83), (71,93), which have small energy differences within $3\frac{kJ}{mol}$ and also small density differences within $0.01\frac{g}{cm^3}$. However, the lattices of these crystal structures have Rotationally-Invariant metrics $d_R \geq 15$ Angstroms and scale-invariant metrics $d_S \geq 1.1$, i.e., with scale factors more than $e^{1.1} \sim 3$. Crystal lattices 41 and 47 have unit cell angles close to 90°, but different unit cell sides: (53.3, 23.7, 7.3), (15.4, 12.9, 16.5). These differences in their structure were found only through geometric information by using the new metrics. Chemical information such as energy and density were not considered. Finally, to summarize:

- Arbitrary crystal lattices are considered up to rigid motions, and their equivalence classes can be distinguished by the new metrics ($d_S$ and $d_R$). Their computation is independent of a choice of unit cells or coordinates in crystal representations.

- Differences are quantified continuously. This concept is fundamental in producing a continuous hierarchy of crystal structures and showing new patterns among their classes.

- Energy and density are not proper characteristics to distinguish the geometric structure. Indeed experiments on simulated T2 crystal structures show that $d_R$ and $d_S$ better distinguish crystal lattices that have almost identical energy and density.

# Chapter 5

# Average minimum distances of a periodic point set

The following chapter deals with a new isometry invariant of crystals focusing on their motif. It aims to develop a property that can correctly classify the space of crystals, solving the isometry classification problem 2.4. My contribution to this project was to develop and implement an early version of the algorithm in C++ that made use of the proper data structure to compute distances. Later, it was extended by new progresses of the other authors with theorems, proofs, a python version [58] and new plots that I collected in this chapter from our paper published in the MATCH journal [60].
The new isometry invariant extracts a geometric feature that directly considers the inter-atomic distances, and maps the atomic structure to distance vectors. The early version of the software can be found on my github account [27].

## 5.1 Inter-atomic distance distribution

The previous invariant example in chapter 2 Figure 2.4 is not enough to distinguish non-equivalent point clouds since they could map to the same list of ordered pairwise distances. Instead of a full list of distances, consider a matrix of distances: 1 row (list of distances) for each point in the central unit cell. Let a crystal $C$ have $m$ points in a unit cell $p_1, ...., p_m$. For any integer $k \geq 1$, we define a $m \times k$ matrix $D(S; k)$ of distances whose $i$-th row consists of the ordered Euclidean distances $d_{i1} \leq ... \leq d_{ik}$ measured from $p_i$ to its first $k$ nearest neighbours within the infinite crystal $C$. For example, let us consider the 1-D periodic point set $S = \{0, 1, 3, 4\} + 8\mathbb{Z}$ repeated by a period of 8 towards both directions and its matrix of distances $D$ (Figure 5.1). $D$ would consist of four rows (one for each point) of distances between a point in the central unit cell and its $k$ nearest neighbours. In the example, only $k = 3$ neighbours are considered.

## 5.2 Averaged Minimum Distances for Periodic Sets

As stated in [60] and repeated below, the distance matrix $D$ is used to extract an isometry invariant called Average Minimum Distance.

**Definition 5.1.** (Average Minimum Distance AMD$_k$(S)). *Let a periodic point set $S = \Lambda + M \subset \mathbb{R}^n$ have points $p_1, ...., p_m$ in a primitive unit cell. For a fixed integer $k \geq 1$ and $i = 1, ..., m$, the i-th row of the $m \times k$ matrix D(S;k) consists of the ordered Euclidean distances $d_{i1} \leq ... \leq d_{ik}$ measured from the point $p_i$ to its first k nearest neighbours within the infinite set S.*
*The j-th* Average Minimum Distance *of a periodic motif S with m points $p_1, ...., p_m$ in a unit cell U is the average of the j-th column in the matrix $D(S; k)$ of distances to neighbours.*

$$AMD_j(S) = \frac{1}{m} \sum_{i=1}^{m} D_{ij}(S; k) \tag{5.1}$$

*The full $AMD^{(k)}$ vector of k distances of S is the set of $AMD_j$ that can be computed as follows:*

$$AMD^{(k)}(S) = (AMD_1(S), AMD_2(S), ..AMD_k(S)) \tag{5.2}$$

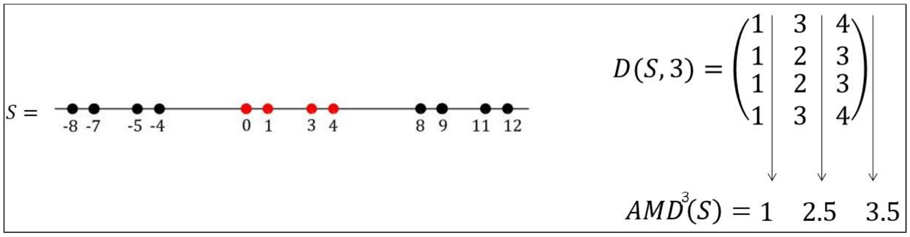

S =
-8 -7 -5 -4 0 1 3 4 8 9 11 12
$D(S,3) = \begin{pmatrix} 1 & 3 & 4 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 3 & 4 \end{pmatrix}$
$AMD^3(S) = 1 \quad 2.5 \quad 3.5$

Figure 5.1: Periodic point Set S and its AMD vector.

This invariant is independent of a unit cell because it does not matter whether a unit cell is primitive or not-primitive; the AMD vector of distances will be the same between two isometric crystals. For example, a periodic point set may be described by two unit cells where one is doubled with respect to the other, as shown in Figure 5.3 for the periodic point sets $S$ and $S_{np}$. In addition, similarity between two crystal structures is correctly assessed thanks to the tendency of the AMD invariant to detect small perturbations within the point cloud. Figure 5.2 presents a toy example where the AMD distance values change slightly after a small point perturbation of the point 1 in 1.5.

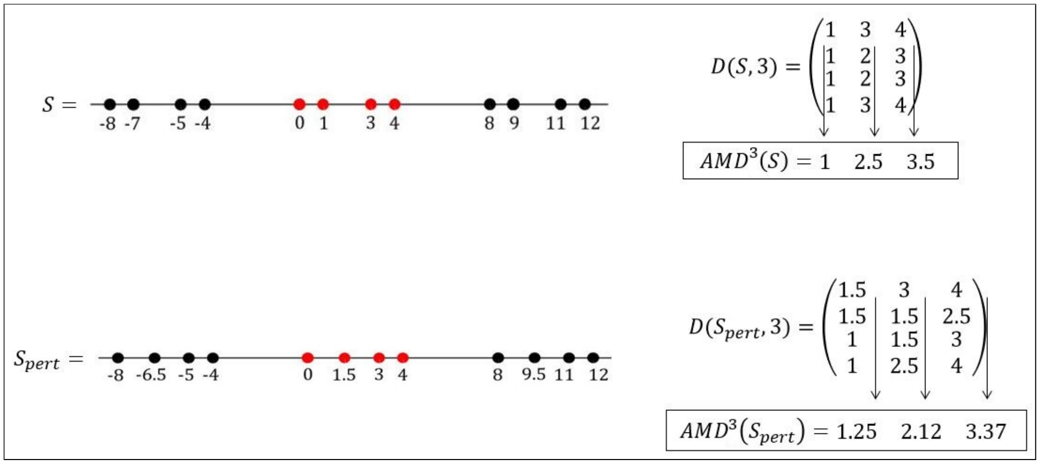


Figure 5.2: Periodic point Set S and its slightly perturbed version $S_{pert}$ with their respective and similar AMD vectors.

Moreover, if two unit cells $U$ and $U'$ have different unit cell parameters, although they belong to the same crystal, AMD can recognise them as the same crystal structure since they map to the same AMD vector. Having different unit cell parameters means that the points have different coordinates, but distances between them are the same. Indeed, regardless the starting point, the set of distances $AMD^{(k)}$ of the crystal generated by both $U$ and $U'$ will be the same after ordering. Therefore, AMD solves the ambiguity problem regarding the infinitely many unit cells that a crystal structure may have. Indeed, ordered lists of distances allow us to retrieve the exact distances in a unique order starting from a central unit cell. Moreover, the number of neighbours $k$ can be increased to involve more distances within the infinite point cloud and add extended information to the AMD invariant about the periodic structure. AMD is an isometry invariant demonstrated by the theorem and proof below.

**Theorem 5.2.** (Theorem 4 in [60]. isometry invariance of AMD). *For any finite*

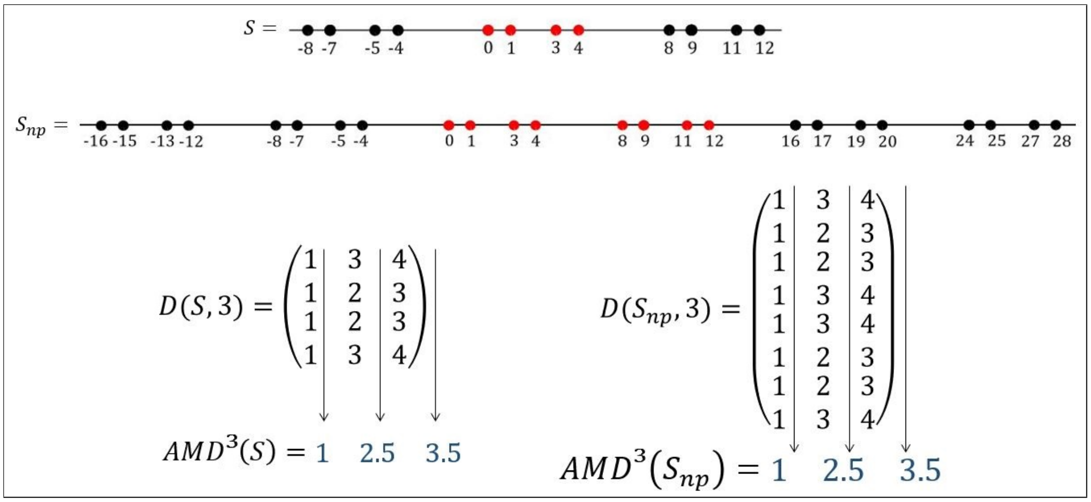


Figure 5.3: Periodic point Set S with a primitive unit cell (red points), periodic point set $S_{np}$ with a non-primitive unit cell (red points) and their respective AMD vectors.

*or periodic point set $S \subset \mathbb{R}^n$, the Average Minimum Distance $AMD_k$(S) is an isometry invariant of S for any $k \geq 1$.*

*Proof.* AMD is an isometry invariant of crystal structures since, firstly, any primitive and non-primitive unit cell of a crystal structure S can be mapped to the same AMD vector. Let us take two primitive unit cells $U$ and $U$' of the same crystal lattice $\Lambda$, consequently with the same number of points. We can establish a bijection between them because points of $U$ can be translated along a unit cell vector $\vec{v} \in \Lambda$ to match points in $U$'. Hence, the distance matrix $D$ will be the same up to rows permutations, where each row is a distance-vector that consists of real value distances between one point and all $k$ neighbours.

Secondly, the distance matrix $D$ is preserved under any isometry $f : S \to Q$ applied to the points in the unit cell with $S$ and $Q$ periodic point sets. Let us consider two primitive unit cells $U_S$ and $U_Q$ that are related to each other by a function map $f$ which maps all points of $U_S$ into points of $U_Q$. Q is preserved by a translation along vector $\vec{v}$, therefore S is preserved by a translation along $f^{-1}(\vec{v})$ having points defined by different integer coefficient of linear combinations. They are primitive unit cells, so they contain the same number of motif points, and all points are bijectively mapped to each other from $S$ to $Q$ and vice versa. Since $f$ preserves distances, every ordered row of the distance matrix contains real values that do not change after applying a certain bijection where each point $p_i$ of $S$ is mapped to its corresponding point in $Q$. If the bijection exists, distance matrices $D(S,k)$ and $D(Q,k)$ are equivalent up to rows' permutations because rows could be ordered in different ways but still exist in both matrices, hence $AMD_k(S) = AMD_k(Q)$. Moreover, being invariant under isometry means that neighbours do not change their distance to points in the surroundings when any isometry is applied to the point cloud. Therefore, to assure invariant values of $AMD$, distances are ordered within each row to gather them in the matrix at specific positions according to the vicinity between points and their neighbours. □

### 5.2.1 Continuity of AMD under perturbations

Crystal structures are rigid bodies subject to perturbations (or atom vibrations) around their position resulting from different environmental conditions, such as the temperature at which their structure is resolved. The simplest way to measure a perturbation is a deviation of their position. The deviation over a full infinite crystal is defined as the Bottleneck Distance.

**Definition 5.3.** (Definition 6 in [60]. bottleneck distance between sets). *For a bijection $g : S \to Q$ between finite or periodic point sets $S, Q \subset R^n$, the maximum deviation is the supremum $\sup\limits_{p \in S} |p - g(p)|$ over $p \in S$. The bottleneck distance is defined as $d_B(S,Q) = \inf\limits_{g:S \to Q} \sup\limits_{p \in S} |p - g(p)|$ is the infimum over bijections $g : S \to Q$ where all bijections are considered from $S$ to $Q$.*

Informally, it is the minimum effort that we need to individually shift and match points in S to points in Q. Given the maximum deviation between point sets, when points in both periodic sets are slightly perturbed, neighbour distances in AMD vectors are subject to minor changes, as explained by the following lemmas and theorems.

**Lemma 5.4.** (Lemma 2 in [12]). *Let periodic point sets $S, Q \subset \mathbb{R}^n$ have the bottleneck distance $d_B(S,Q) < r(Q)$, where $r(Q)$ is the minimum half-distance between points of $Q$. Then $S, Q$ have a common lattice $\Lambda$ with a unit cell $U$ such that $S = \Lambda + (U \cap S), Q = \Lambda + (U \cap Q)$.*

The packing radius $r(Q)$ refers to that distance between particles that confirms the existence of the smallest bond in an infinite crystal motif. When we match points in S with points in $Q$, we may want to shift and match $S$ onto $Q$ by performing a minimum effort. Practically, a small enough distance between two sets may assure that points of $U_S$ remains inside $U_Q$ up to translations. Indeed, consider a point $\vec{s} \in S$ and its infinitely many copies $\vec{s}_n = \vec{s} + n\vec{v}$ towards a direction along vector $\vec{v}$ with $n \in \mathbb{Z}$. For any point $\vec{s} \in S$ we can always find its copy at a lattice point of $\Lambda_Q$ inside a unit cell $U_Q$ translated by a number of times towards that direction, $\vec{s}_n \in \vec{q}_n + U_Q$ where $q_n \in \Lambda_Q$.

**Lemma 5.5.** (Lemma 8 in [60]. perturbed distances). *For some $\epsilon > 0$, let $g : S \to Q$ be a bijection between finite or periodic sets such that $|a - g(a)| \leq \epsilon$ for all $a \in S$. For any $i \geq 1$, let $a_i \in S$ and $b_i \in Q$ be the i-nearest neighbours of points $a \in S$ and $b = g(a) \in Q$, respectively. Then the Euclidean distances from a,b to their i-th neighbours $a_i, b_i$ are $2\epsilon$-close, i.e $||a - a_i| - |b - b_i|| \leq 2\epsilon$.*

Below an explanation of the proof is presented from the related paper. First, we need to fix a value (or upper bound) that will define the interval within which distances may change. The maximum value will be based on the maximum deviation $\epsilon$ allowed for a distance of two points. When a maximum deviation is to be found, it is better to consider both point clouds superimposed. Think of the bijection $g$ as a tool used for the superimposition. Each point $a \in S$ is linked by the function $g$ to a point $b \in Q$ and vice versa (bijection). In theory those points should be superimposed if shifted, but that depends on the rest of the point cloud. Indeed, we may want to translate the full set $Q$ by the vector $\vec{a} - g(\vec{a})$ to shift, and eventually match, $S$ with $Q$. A minimum effort for the superimposition is considered, and therefore $\epsilon$ is found over all the maximum distances (Bottleneck distance). In practice, given the maximum distances over pairs of points, the minimum is computed. Provided that any two points change their distance at most by $\epsilon$, distances between neighbours have at most a $2\epsilon$ offset. Finally, considering the space inside a ball of diameter $2\epsilon$, the new distances between $a$, $b$ and their neighbours will be similar to each other by at most $2\epsilon$. It means that perturbing neighbours around will lead to small deviations which are upper bounded by $2\epsilon$, and if point sets are isometric these distances between neighbours will be equivalent $||a - a_i| - |b - b_i|| = 0$. If distances of $a$, $b$ and their $i - th$ neighbours were completely different, and $a$ kept the maximum deviation allowed $2\epsilon$ with its $i - th$ nearest neighbour $a_i$ (e.g. both located in the closed-ball border), then $||a - a_i| - |b - b_i||$ will tend to $2\epsilon$ as $b$ and $b_i$ get closer. Hence, distance perturbations are bounded by a closed ball of diameter $2\epsilon$.

**Theorem 5.6.** (Theorem 9 in [60]. continuity of AMD under any small perturbations). *Let finite or periodic sets $S, Q \subset \mathbb{R}^n$ satisfy $d_B(S, Q) < r(Q)$, where $r(Q)$ is the packing radius of $Q$. Then $|AMD_k(S) - AMD_k(Q)| \leq 2d_B(S, Q)$ for any $k \geq 1$.*

In an informal discussion we may say that since all atoms vibrate, the simplest way to measure a perturbation is a deviation of their position. However, the deviation can be measured only over a bounded domain. The deviation over full infinite crystals is defined as the Bottleneck Distance $d_B(S,Q)$. From $S$ and $Q$, we compare the matrices $D(S;k)$ and $D(Q;k)$ and the distance between the $k$-th values of both is upper bounded by $2\epsilon = 2d_B(S,Q)$ by the Lemma 5.8. Therefore, the average of each column of both $D(S;k)$ and $D(Q;k)$ will be bounded to the same value.

### 5.2.2 The asymptotic behaviour of AMD

$AMD_k(S)$ approaches $c(S)\sqrt[n]{k}$. The point packing coefficient $c(S)$ is introduced below. The volume of the unit ball in $\mathbb{R}^n$ is $V_n = \dfrac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}$, where $\Gamma$ denotes Euler's Gamma function $\Gamma(m) = (m-1)!$ and $\Gamma(\frac{m}{2}+1) = \sqrt{\pi}(m-\frac{1}{2})(m-\frac{3}{2})\cdots\frac{1}{2}$ for any integer $m \geq 1$.

**Definition 5.7.** (Definition 10 in [60]. $(U,m)$-sets $S$, $AMD_k(S;U)$, the point packing coefficient $c(S)$). *Let $U$ be a unit cell of a lattice $\Lambda \subset \mathbb{R}^n$. For any fixed $m \geq 1$, a set $S \subset \mathbb{R}^n$ is called a* $(U,m)$-set *if $S \cap (U+\vec{v})$ consists of $m$ points for any vector $\vec{v} \in \Lambda$. For any point $p \in S \cap U$, let $d_k(S;p)$ be the distance from $p$ to its $k$-th nearest neighbour in $S$. The* Average Minimum Distance *is $AMD_k(S;U) = \dfrac{1}{m}\sum\limits_{p\in S\cap U} d_k(S;p)$. The* Point Packing Coefficient *is $c(S) = \sqrt[n]{\dfrac{Vol[U]}{mV_n}}$.*

For example, a periodic set $S = \Lambda + M$ is generated by a unit cell $U$ with a lattice $\Lambda$ and a motif $M$ with $m$ points. If a non-periodic perturbation of $S$ is considered, the unit cell above changes its $AMD_k(S;U+\vec{v})$ value, because only one point is perturbed, and it belongs to the original unit cell. Moreover, if the perturbation is periodic (as it is in crystal structures), the same perturbation will affect all copies in the point cloud $S$, resulting in another change of the $AMD_k(S;U+\vec{v})$ value. While the $AMD_k(S;U+\vec{v})$ values depend on those perturbations, the number of motif points and the unit cell volume remain the same. Therefore, $\frac{Vol[U]}{m}$ is independent of a choice of $U$, and it is an isometry invariant also for any $(U,m)$-set $S$. If all points have the weight $V_n$ of the unit ball in $\mathbb{R}^n$, then $(c(S))^n$ is inversely proportional to the density $\rho = \frac{mV_n}{Vol[U]}$ of $S$. The *diameter* of a unit cell $U$ is $d = \sup\limits_{a,b\in U} |a-b|$.

Distances in AMD vector follows a cubic root behaviour as shown in Figure 5.4 for nine T2 experimental crystals. This plot highlights experimental crystals' AMDs when $k = 1000$

neighbours are chosen, and each $k$ on the x-axis is the index in the AMD vector. All gamma structures resolved under different temperatures overlap in the first curve from the top, followed by alpha, both (overlapping) betas, delta and epsilon structures.

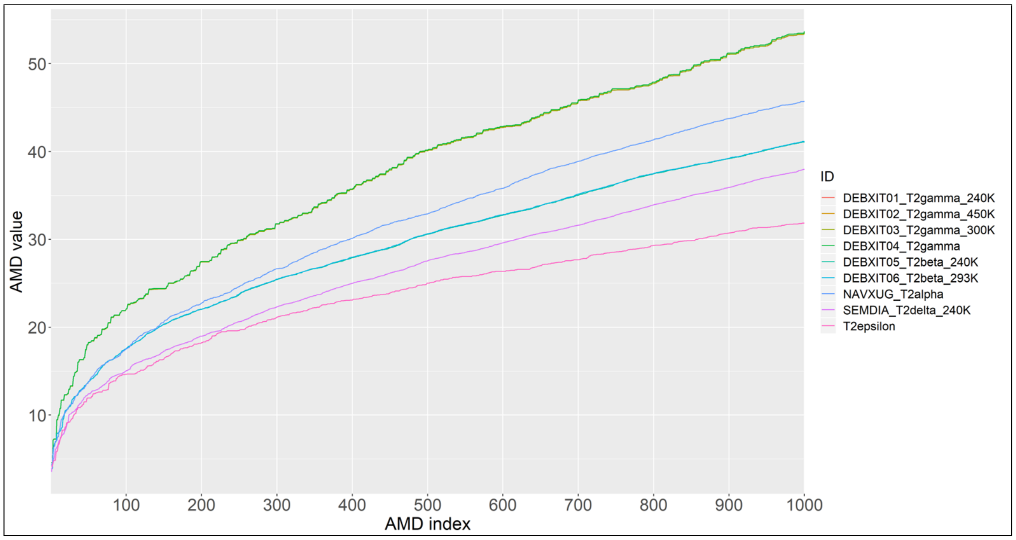


Figure 5.4: AMD values of T2 experimental crystals with k = 1000.

**Lemma 5.8.** (Theorem 12 in [60]. distance bounds). *Let $S \subset \mathbb{R}^n$ be any $(U, m)$-set with a unit cell $U$ of diameter $d$, see Definition 5.7. For any point $p \in S \cap U$, let $d_k(S; p)$ be the distance from $p$ to its $k$-th nearest neighbour in $S$. Then $c(S)\sqrt[n]{k} - d < d_k(S; p) \leq c(S)\sqrt[n]{k} + d$ for any $k \geq 1$.*

The AMD retains the behaviour of $\sqrt[n]{k}$ and each distance within the point cloud bounded by the diameter of a ball follows that trend. This behaviour is demonstrated by the following theorem that defines an upper bound to limit allowed perturbations ($< d$) within the point cloud included in a ball of diameter $d$. AMD is linked to the cubic root trend through the point packing coefficient, which in this case represents the isometry invariant part of a distance distribution owned by each periodic point set.

**Theorem 5.9.** (Theorem 13 in [60]. asymptotic behaviour of AMD). *For any $(U, m)$-set $S \subset \mathbb{R}^n$ from Definition 5.7, we have $|AMD_k(S; U) - c(S)\sqrt[n]{k}| \leq d$ for any $k \geq 1$ and $\lim\limits_{k \to +\infty} \dfrac{AMD_k(S; U)}{\sqrt[n]{k}} = c(S)$.*

## 5.3 A near linear time algorithm for AMD

AMD is a powerful isometry invariant that can encode a large number of information in a vector of distances. The following section describes the algorithm used in Theorem 5.10. The input is strictly related to the 3D-dimensional structure of its crystal, where the entire motif with all $m$ motif points is stored as atom coordinates in the CIF file, following the Cartesian system. Each crystal structure can be mapped to a vector $AMD^{(k)}(S) = (AMD_1, \dots, AMD_k)$ that is independent of a periodic point set $S$. Increasing $k$ adds more components to the vector $AMD^{(k)}$ without changing any previous values. There is the following theorem for near linear time complexity of AMD.

**Theorem 5.10.** (Theorem 14 in [61]. a near linear time algorithm for AMD). *Let a periodic set $S \subset \mathbb{R}^n$ have $m$ points in a unit cell $U$. For a fixed dimension $n$ and $i = 1, \dots, k$, $AMD_i(S)$ can be computed in a time $O(\nu(S;n)km\log(km))$, where $\nu(S;n)$ is independent of $k, m$.*

The algorithm starts by building an $n$-d tree [8] of an extended point cloud of $\mu$ points in time $O(n\mu\log\mu)$. Briefly, all space, in which a point cloud lie, is divided into regions, and each of these regions can be accessed by a point query $p$ that interrogates the $n$-d tree structure. Regions of the space are traversed until the closest point to $p$ is found. For each $p \in M$, all $k$ neighbours of $p$ can be found and ordered by distances to $p$ in time $O(\mu\log\mu)$. By Definition 5.1 we lexicographically sort $m$ lists of ordered distances in time $O(km\log m)$, because a comparison of any two ordered lists of length $k$ takes $O(k)$ time. The ordered lists of distances are the rows of the matrix $D(S;k)$. All $AMD_i(S)$ are found in time $O(km)$. The total time is $O((m+n)\mu\log\mu + km\log m) = O(\nu(S;n)km\log(km))$.

## 5.4 Comparison between crystal structures through AMD

The motivation that led us to develop new tools, which can continuously describe the space of crystals, was based on the fact that many simulated crystal structures coming from CSP methods may be very similar to each other. Indeed, the lattice energy minimisation is performed on supercomputers, which requires many weeks and may gather all close crystals from the same local minima. Indeed, the landscape in Fig. 5.5 required 12 weeks of supercomputer time [37]. This issue has been assessed as the ‘over-prediction’ because of the high amount of generated structures of similar crystals. The lattice energy has no closed expression due to the dependence on infinitely many interactions between all atoms within a periodic crystal. Simulated crystals are often visualised by the energy-vs-density *landscape* where each point represents a crystal with two coordinates (density, energy). This single-value density is insufficient to differentiate crystals because many non-isometric sets can have the same density.

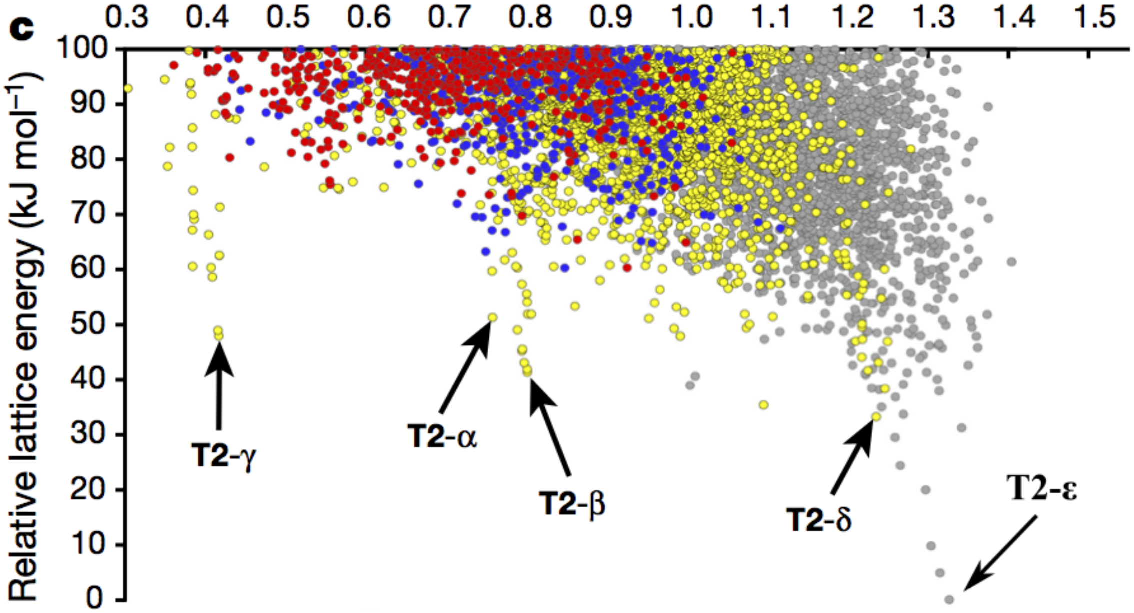


Figure 5.5: [37, Fig. 2d]: energy-vs-density plot shows many of 5679 crystals as nearly identical.

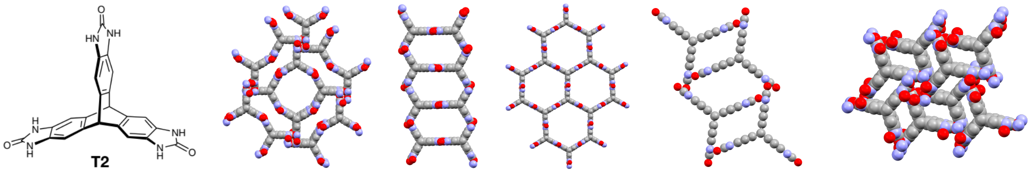


Figure 5.6: [61, Fig. 12]. T2 molecule and the crystals T2-$\alpha$, T2-$\beta$, T2-$\gamma$, T2-$\delta$, T2-$\epsilon$ based on the T2 molecule were synthesized in the laboratory after the Crystal Structure Prediction in Fig. 5.5 reported in [37].

A manual approach was used to detect the candidates for synthesis among the crystal structure, and only 5 of them were chosen and synthesised. However, now, AMD can automatically identify them as close neighbours in the clustering dendrogram in Fig. 5.7.

Here are the numerical IDs in the T2 dataset: 0186 for T2-$\alpha$, 0054 for T2-$\beta$, 0120 for T2-$\gamma$, 0014 for T2-$\delta$, 0001 for T2-$\epsilon$. Fig. 5.7 shows the dendrogram obtained by the $L_\infty$-distance, which is more stable with respect to the length $k$ of AMD vectors than Euclidean $L_2$ due to Theorem 5.9. All expected similar crystals, including the four versions of T2-$\gamma$ that relate to different conditions of synthesis, belong to the same small cluster highlighted by the orange colour in Fig. 5.7. Fig. 5.4 clearly shows five different patterns for the various experimental structures of T2, thus showing that AMD curves can distinguish between polymorphs (that is, different crystal packings of the same molecule).

For molecular crystals, which are the focus here, chemical bonds can range in strength, from strong covalent bonds to weak hydrogen bonds and even weaker inter-atomic dispersion forces. The concept of bonding is qualitative and quasi-continuous in nature, while atomic positions are more unambiguously defined. To identify any crystal, we should avoid ambiguous descriptors that are not preserved under isometries and changes of a basis, such as chemical properties-based descriptors or non-invariant values. Table 2.1 showed comparisons between the experimental crystal T2-$\delta$ and its closest simulated version (crystal 14 in the T2 dataset). The RMS deviations do not provide a single distance but strongly depend on the number of molecules that are matched in two crystals.

## 5.5 Conclusions

A key tool in any classification is an *invariant* that is a function or a property preserved under equivalence (see chapter 2).

In the experiments on T2 crystals, each of 42 atoms in the T2 molecule was represented by one point. Each of 5679 T2 crystals [37] contains up to 16 T2 molecules in a unit cell, with eight molecules on average. $AMD^{(200)}$ was computed in 5 min over 5679 T2 crystals on a modest desktop, which is negligible in comparison with the 12-week supercomputer time for the CSP plot in Fig. 5.5. Computational CSP datasets have the potential to be much larger than this, so a complete classification of periodic crystals and quick computations are the key characteristics to understand structure-property relationships. X-ray diffraction determines crystal structures in a rigid form so that material transformations can be connected in the continuous space of all isometry classes of periodic crystals.

AMD$^{(k)}$ vectors were computed for full periodic structures and much larger $k$ on the modest desktop AMD Ryzen 5 6-core 4.60Ghz, 32GB DDR4. Although the Cambridge Structural Database (CSD) has more various crystal forms in comparison with T2 struc-

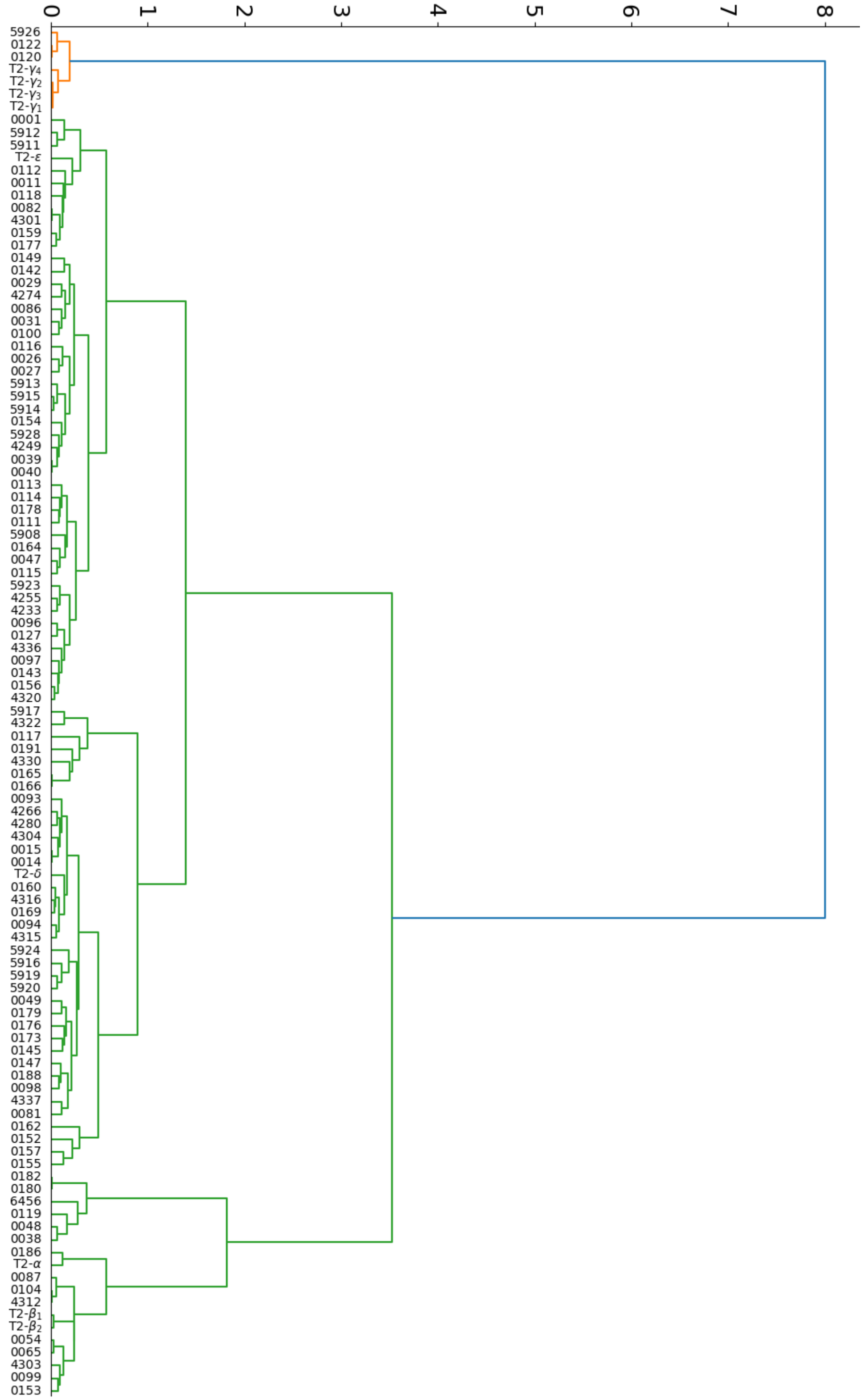


Figure 5.7: [61, Fig. 13]. Complete-linkage clustering by the $L_\infty$-distance on $AMD^{(1000)}$ of T2 crystals: nine experimental and 100 simulated crystals with lowest energies, see [37].

tures based on the same molecule, $\mathrm{AMD}^{(100)}$ required less than 52 min for all 228,994 organic structures. Their relations are showed by the TreeMap in Figure 5.8.

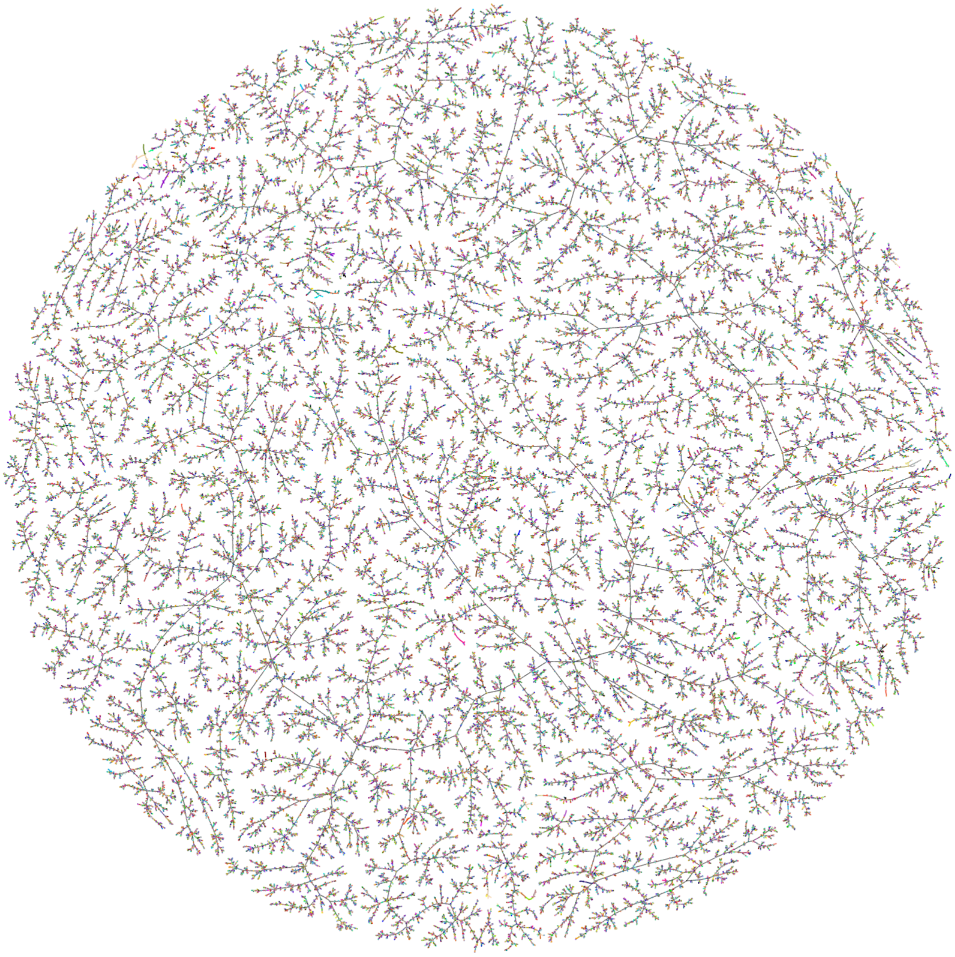


Figure 5.8: [61, Fig. 18]. TreeMap of the Cambridge structural Database with 228,994 molecular organic crystal structures.

# Chapter 6

# Fast prediction of lattice energies by AMD invariant

In this chapter, we will go through a machine learning method used to predict chemical properties that makes use of our previous isometry invariant AMD. It is based on our paper published in 2022 [43] and accompanied by a new abstract for the Congress of the International Union of Crystallography (IUCr) [42]. The goal of this third PhD project is to solve the problem of structure-related property prediction 6.1. T2 dataset (see subsection 3.2.3) is taken into account to perform our experiments. Average Minimum Distance is computed to predict the **lattice energy** by splitting the dataset in a test set made of 1136 AMD vectors and a training set of 4543 AMD input vectors containing ordered distances. The last set trains the machine learning model explained in Section 6.4.

## 6.1 Importance of structural information for crystal properties

**Problem 6.1.** (Structure-related property prediction). *Find geometric characteristics and proper isometry invariants that thoroughly predict desired properties of crystals such as the lattice energy.*

Each crystal retains a specific function that manifests in the real world as physical property such as electrical conductivity, malleability and gas absorption rate. Since physicochemical properties depends on the structure, distances between atoms should be considered for a proper classification and new characteristics' discovery.

The most important property of a crystal is the energy of its crystal structure, which is usually called the *lattice energy* or *potential energy surface* or *energy landscape* [55]. Chemists refer to these chemical properties to assess a structure and its stability or, better to say, to decide its capability to retain a solid phase under standard environmental conditions. If the energy value is acceptable, then such a crystal can be accessible for **synthesis** in a lab and hopefully can remain stable under particular environmental conditions. Since the lattice energy is a steep and high-frequency function with no closed analytic expression, calculations are always approximated, from the *force field (FF)* level [31] to the more exact *density functional theory (DFT)* [19].

The main novelty of our approach to energy predictions is using a fast computable and easily interpretable invariant of crystals, which should follow the conditions highlighted in Section 2.2.

Many widely used isometry invariants, including symmetry groups, split the space of all crystal structures in a discrete set of classes. However, what should be decided when an unknown structure lies on a boundary of two classes? How should it be classified? Therefore, these types of invariants are considered **discontinuous** under perturbations of atoms because they isolate classes of crystals, stating that there is no relation between them. Conversely, there is a huge relation between each type of structure: they belong to a continuous space. Perturbations are then crucial for distinguishing simulated crystals obtained via Crystal Structure Prediction (CSP). Indeed, CSP methods stop at some local minima [33] causing the simulation of many structures to converge to the same local minimum.

The lattice energy is a function of the continuous crystal space whose geometry needs to be unrevealed by using proper invariants that can help distinguish one structure from another.

## 6.2 AMD: a resource for lattice energy prediction

The key problem is to describe **structure-property relations** through suitable isometry invariants that can approximate chemical properties of crystals such as the lattice energy. The Average Minimum Distances (AMD) (see chapter 5) solve the above-mentioned problem of property prediction 6.1. AMD is an infinite sequence of isometry invariants that can change by at most $2\epsilon$ if the atom coordinates are perturbed in their $\epsilon$-neighbourhoods. The continuity of AMD in Theorem 5.6 confirms that small perturbations of crystal structures lead to small changes in AMD values and can be tested for checking **continuity of energy**

under crystal perturbations. Continuity is a local property that can be assessed by finding a boundary within which a specific property may change (in our case, the lattice energy). This boundary is given by the difference in AMD. In practice, we need to find a constant $\lambda$ and $\delta$ such that a small distance $d < \delta$ between AMD vectors can limit a slight change in energy, within $1\frac{kJ}{mol}$ for a good outcome to rise small changes in AMD values followed by small changes in energy. Past invariants, such as density, RMSD (see subsection 2.4.2), PXRD (see subsection 2.4.3) do not satisfy the continuity property (no boundary constant at all), indeed same values from their computations do not guarantee close values of energy. PRDF (see subsection 1.2.3 where refers to electrons) was used in the past to predict the lattice energy. The computation requires distance thresholds $r$ and $dr$, which can affect the output. Schutt et al. confirm in [[49], Table I] that the PRDF consists of non-invariant features such as cell parameters. The mean absolute error (MAE) of energy predictions based on PRDF is $0.68\frac{eV}{atom}$ or $65.6\frac{kJ}{mol}$.
Wigner-Seitz cells (also called Dirichlet or Voronoi domains) of atoms were used by Ward et al. [56] and 271 cell-based geometric and chemical attributes were extracted to reach the MAE of 0.09eV/atom or 8.7kJ/mole to predict the formation enthalpy. A further neural network approach [[47], Fig. 4] improved the mean absolute error (MAE) to $1.8\frac{kcal}{mol} = 7.56\frac{kJ}{mol}$. Egorova et al. [13] predicted the difference between the accurate DFT energy and its force field approximation with MAE less than 2kJ/mole by using GGA DFT (PBE) calculations and symmetry function descriptors [3].

Lattice energies of the T2 dataset come from force field calculations and DFT optimizations. The COMPASS force field (Section 1.9.3) is used to compute energy values of all the simulated crystal structures. It consists of different terms for bonds ($b$), angles ($\theta$), dihedrals ($\phi$), out-of-plane angles ($\chi$), cross-terms, two non-bonded functions, electrostatics ($E_q$) and van Der Waals ($E_{vdW}$) interactions.

$$E_{total} = E_b + E_\theta + E_\phi + E_\chi + E_{b,b'} + E_{b,\theta} + E_{b,\phi} + E_{\theta,\phi} + E_{\theta,\theta'} + E_{\theta,\theta',\phi} + E_q + E_{vdW} \quad (6.1)$$

Especially, we may want to focus on the van der Waals term $E_{vdW}$ that follows a 9-6 Lennard-Jones potential to describe dispersion forces of non-bonded interactions.

$$E_{vdW} = \sum_{ij} \epsilon_{ij}[2(\frac{r^o_{ij}}{r_{ij}})^9 - 3(\frac{r^o_{ij}}{r_{ij}})^6] \quad (6.2)$$

where $r_{ij}$ is the distance between two atoms. The property of the AMD that could relate to these interactions is parametrised in the number of neighbours $k$ that should be considered during the computation of distances. A higher number of neighbours will lead to a higher number of pairwise distances encoded in the AMD vector. Therefore, it can store the geometric information of long-range interactions between particles, averaged by definition.

We achieved a mean absolute error less than $5\frac{kJ}{mol}$ using AMD vectors as descriptors. The fact that we wanted to exclude chemical data arises from our concept of isometry invariant. Geometric-based descriptors should assess chemical properties to relate the structure of a crystal with its function properly. The training process took 10 minutes using a modest desktop on the T2 dataset made of 5679 crystal structures.

## 6.3 Continuity of structure-property relation in terms of AMD

A crystal function depends on its structure, and therefore, it becomes essential to reveal which geometric features can be computed. Machine learning approaches rely on properties from crystal descriptors where not all are invariants up to isometry. It is not demonstrated or may be hard to find that small changes in the input data can release small perturbations in the output.

**Continuity of a structure-property relation** can be mathematically expressed as Lipschitz continuity [32], Section 9.4]:

$$|E(S) - E(Q)| \leq \lambda \, d(S, Q) \tag{6.3}$$

where $\lambda$ is a constant, $E$ is a chemical property such as the lattice energy, $d(S, Q)$ is a distance function. The inequality 6.3 should hold for all crystals $S$, $Q$ with small distances $d(S, Q) < \delta$.

Figure 6.1,6.2,6.3 show past methods of crystal similarity, which are insufficient to assess the continuity of the lattice energy. Crystal were paired up and each pair were represented by a square dot. Differences in the past metric between two crystals link to the x-axis and differences in energies on the y-axis.

Figure 6.1 shows crystal pairs with very close densities and somewhat different lattice energies, which means that the energy varies very differently from the density changes. However, density is still used as a key property of crystal structures in CSP landscapes. It

is an isometry invariant, continuous and constant under perturbations.
Figure 6.2 illustrates the packing similarity computed by the COMPACK algorithm (see subsection 2.4.2 as the Root Mean Square Deviation (RMSD). Atomic positions are matched up to 15 molecules in two crystals, which correspond to the default value. Different thresholds are considered for the computation, such as angle offsets whose values affect the RMSD. For example, when only 1 of 15 molecules is matched, the RMSD equals 0 because each crystal among 5679 consists of the same T2 molecule in Figure 3.2. Finally, the powder X-ray diffraction (PXRD) similarity has the range [0,1] with values close to 1, indicating the similarity of diffraction patterns. Figure 6.3 has $1 - PXRD$. The same conclusion of density and RSMD is also adopted for this structural feature since high energy differences may have shallow differences in PXRD.

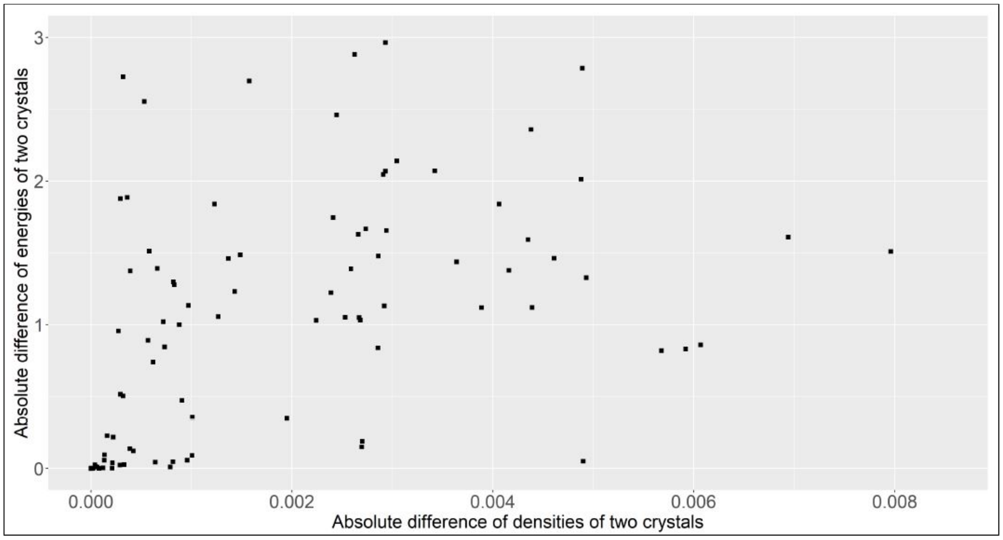


Figure 6.1: [43] Figure 5. 5679 T2 crystals in Figure 3.2 have the density in the range [0.3; 1.4]. Several crystals have differences in densities within 0.003 $\frac{g}{cm^3}$ and differences in the energy up to $3\frac{kJ}{mol}$.

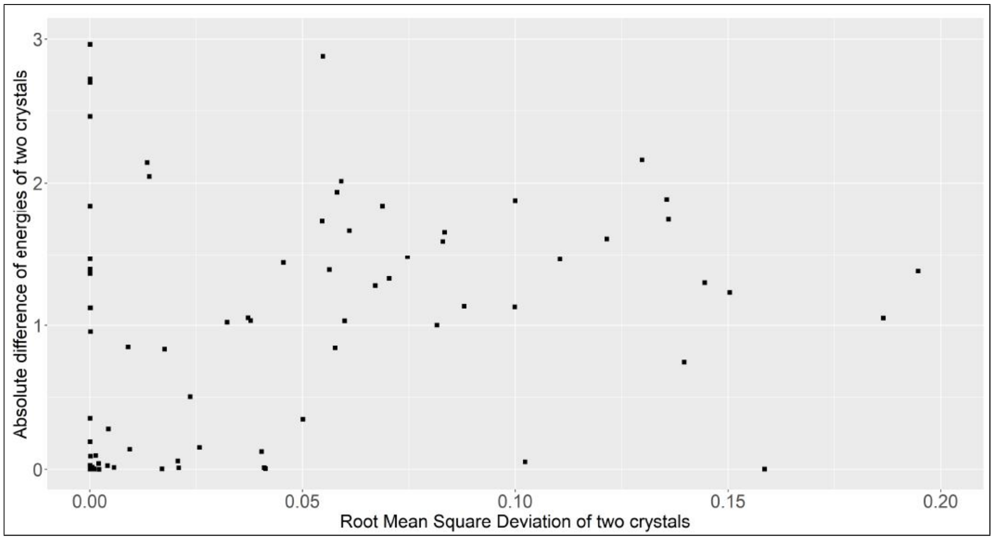


Figure 6.2: [43] Figure 6. Crystal pairs with $RMSD < 0.1\mathring{A}$ have energy differences up to $3\frac{kJ}{mol}$.

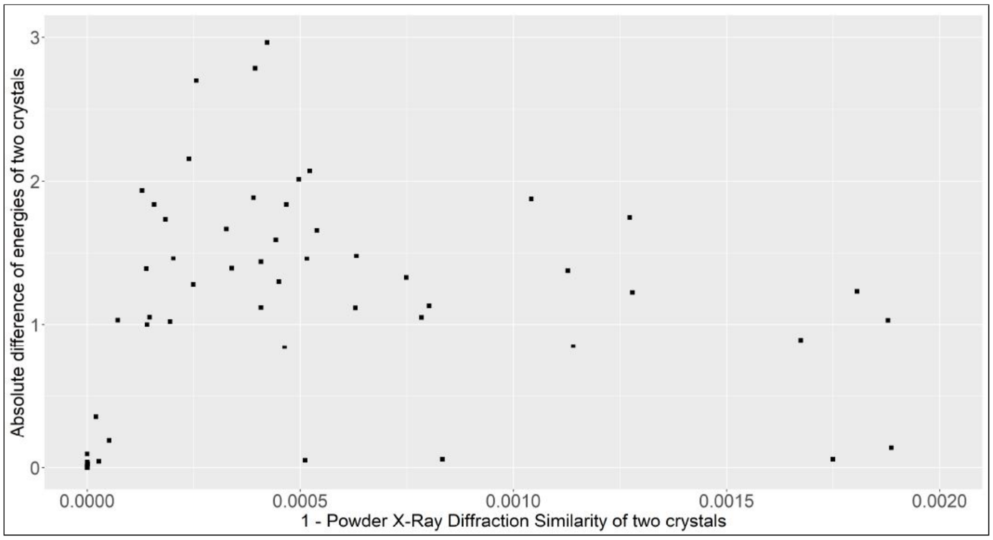


Figure 6.3: [43] Figure 7. Crystal pairs with $1 - PRXD < 0.0005$ may have high energy differences ($\Delta E > 1\frac{kJ}{mol}$), though $1 - PXRD \in [0, 0.0005]$ suggests they are very similar structures.

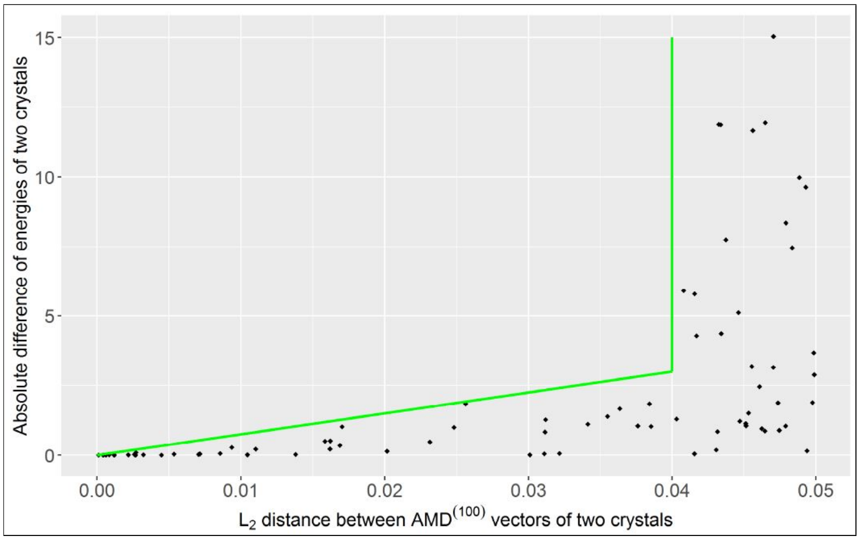


Figure 6.4: [43] Figure 8. The green line $|\Delta E| = 75L_2$ with $L_2 \in [0, 0.04]$ shows that if crystals have a distance $L_2 < 0.04\mathring{A}$ between $AMD^{100}$ vectors, their energy difference $\Delta E$ measured in $\frac{kJ}{mol}$ is at most $75L_2$.

In Figure 6.4, 6.5, 6.6 the continuous AMD invariant of the T2 dataset changes together with the lattice energy. Crystals are paired up and compared through $AMD^{100}$ vectors of length $k = 100$ and represented by a rhombus-shaped dot. The distances between vectors $\vec{p} = (\vec{p}_1, .., \vec{p}_k)$ and $\vec{q} = (\vec{q}_1, .., \vec{q}_k)$ on the horizontal axis are computed by the Euclidean metric

$$d(p,q) = L_2(p,q) = \sqrt{\sum_{i=1}^{k}(|p_i - q_i|^2)}$$

the Chebyshev metric

$$L_\infty(p,q) = \max_{i=1,..,k} |p_i - q_i|$$

and the Manhattan metric

$$L_1(p,q) = \sum_{i=1}^{k} |p_i - q_j|$$

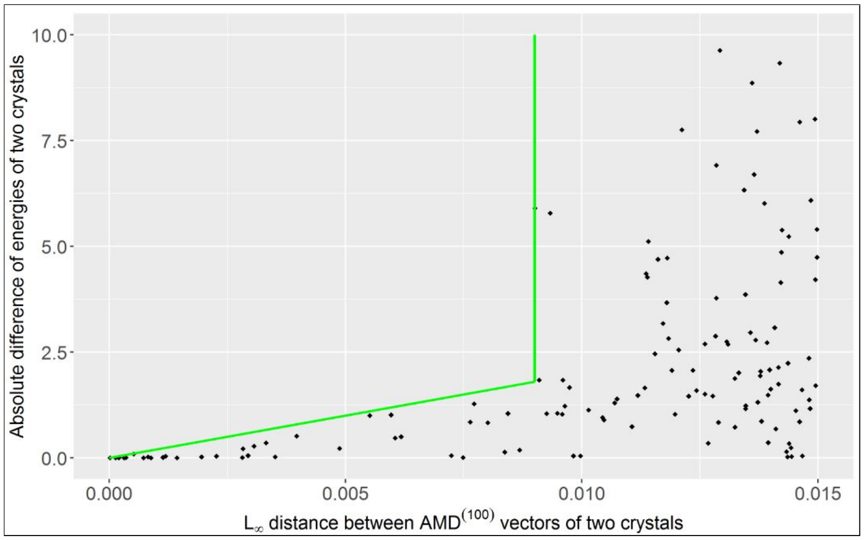


Figure 6.5: [43] Figure 9. The green line $|\Delta E| = 200L_\infty$ with $L_\infty \in [0, 0.009]$ shows that if crystals have a distance $L_\infty < 0.009\mathring{A}$ between $AMD^{100}$ vectors, their energies differ by at most $200L_\infty$.

In Figure 6.4 the Lipschitz's continuity for the energy

$$|\Delta E| = |E(S) - E(Q)| \leq \lambda_2 L_2(AMD^{100}(S), AMD^{100}(Q)) \tag{6.4}$$

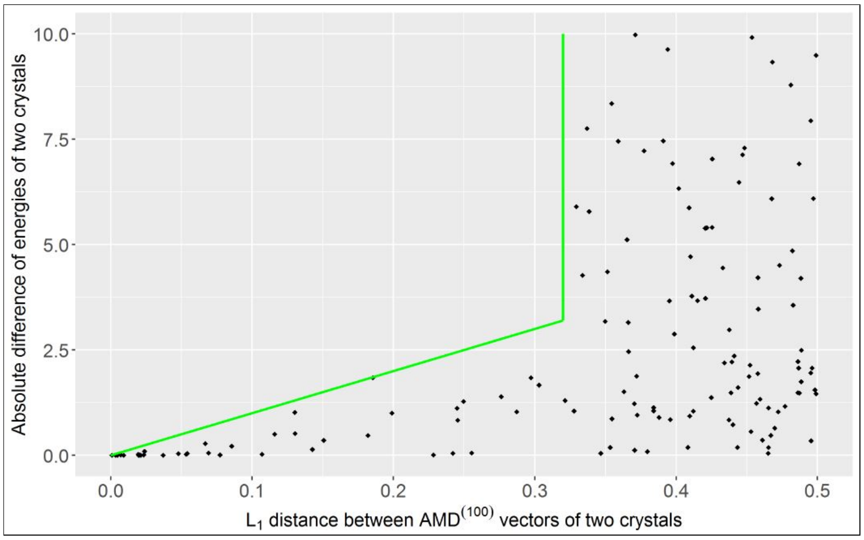


Figure 6.6: [43] Figure 10. The green line $|\Delta E| = 10L_1 \in [0, 0.32]$ shows that if crystals have a distance $L_1 < 0.32\mathring{A}$ between $AMD^{100}$ vectors, their energies differ by at most $10L_1$.

holds for $\lambda_2 = 75$ and all pairs of crystals $S,\ Q$ whose $AMD^{100}$ vectors have the Euclidean distance $L_2 < \delta_2 = 0.04\mathring{A}$. These pairs are below the green line $\Delta E = 75L_2$ up to the distance threshold $\delta_2 = 0.04\mathring{A}$. Figure 6.5 similarly illustrates continuity of the lattice energy with respect to the metric $L_\infty(p, q)$ between $AMD^{100}$ vectors. All pairs of crystals with distances $L_\infty < \delta_\infty$ have energy differences less than $\lambda_\infty L_\infty$ with $\lambda_\infty = 200$, so all dots are below the green line $|\Delta E| = 200L_\infty$. In Figure 6.6 is shown that the lattice energy is continuous for the metric $L_1(p, q)$ between $AMD^{100}$ vectors. All pairs of crystals with distances $L_1 < \delta_1 = 0.32\mathring{A}$ have energy differences less than $\lambda_1 L_1$ with $\lambda_1 = 10$, so all dots are below the green line $|\Delta| = 10L_1$. The thresholds $\delta_1 = 0.32\mathring{A}$ and $\delta_2 = 0.04$ are larger than $\delta_\infty = 0.009\mathring{A}$, because the metrics $L_1, L_2$ sum up all deviations between corresponding coordinates of $AMD^{100}$ vectors. On the other hand, the metric $L_\infty$ measures only the maximum deviation. Lipschitz continuity for the previous descriptors (density, RMSD, PXRD) in Figure 6.1,6.2,6.3 would have had huge slopes or gradients (Lipschitz constants) as the difference in their values increases.

## 6.4 Mean Absolute Error of energy prediction

This section will deal with the important contribution of inferring the lattice energy from a crystal structure by using a training dataset of ground truth energies. The discontinuity of the previous descriptors in Figure 6.1,6.2,6.3 does not allow resolving the energy prediction problem. Indeed, if we input a perturbed crystal, we expect a close energy value in the output.

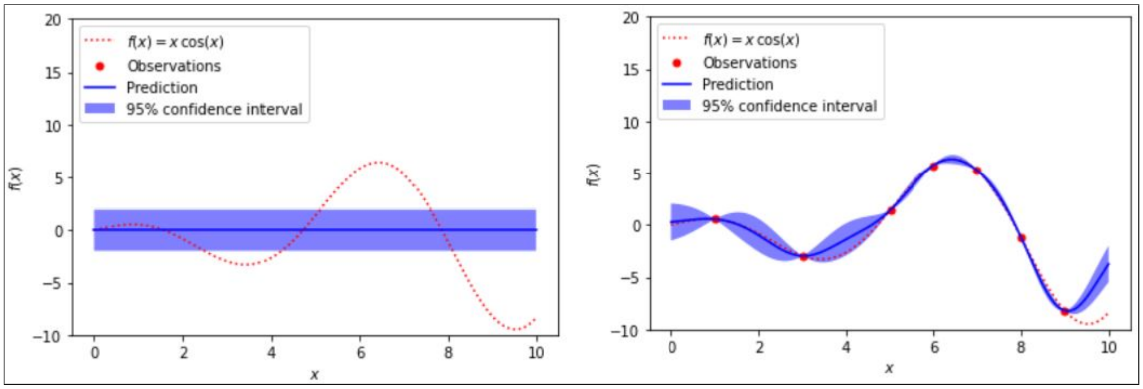


Figure 6.7: [43] Figure 11. Example of Gaussian Process that predicts values of $f(x) = x \cos x$ by training on observed data points. Left: the starting prediction is 0 for any $x$. Right: After training on six data points the prediction improves.

The training and prediction procedures have been performed by the Gaussian Process Regression [24] as implemented in SciKit Learn [36] (see Figure 6.7 for a general example). This procedure achieved the highest score on the T2 dataset of 5679 crystals. Let us move briefly to the explanation of this type of regression.

This machine learning method belongs to the **supervised learning** class. Like all methods in this class, its procedure consists of a training set full of empirical data (or labelled data). In our case, the model is trained through AMD values (or inter-atomic distances), and it aims to link lattice energy values with distances' information. Given an unknown crystal structure represented by its AMD vector, the prediction output will be the lattice energy. The computation is performed by a regression since the output falls in a continuous range of values. The model start with a prior distribution or, better to say, a poor knowledge of how our data behaves. In Figure 6.7a, the prior knowledge states that all values are 0 (mean is 0), and no initial noise is considered. When the first set of observed data populates the prior knowledge, the model changes its state and a posterior distribution (or knowledge) is generated given the observations(Figure 6.7b). The state's change implies the alteration of the mean either in the observed data points that perfectly match it or in the confidence intervals where, most probably, it may lie.

Each crystal structure S is mapped to a vector and represented by its $AMD^k(S)$ vector. The length $k$ is in the interval $[50, 500]$. The distance $d$ chosen between $AMD^k$ vectors was $L_\infty$ due to the smallest Lipschitz constant $\lambda = 2$ in the continuity property $|AMD_k(S) - AMD_k(Q)| \leq \lambda d_B(S; Q)$. For the metrics $L_1, L_2$, the Lipschitz constants would be $2k, 2\sqrt{k}$. Machine learning methods need to run on mapping functions (kernels) used to map input features in a space where they can be compared and their parameters updated. These functions are called Kernels, and for any pair of crystals S and Q, we consider the Rational Quadratic Kernel defined as follows:

$$K(S, Q) = (1 + \frac{d^2(S, Q)}{2\alpha l^2})^{-\alpha} \tag{6.5}$$

where $\alpha, l$ are scale parameters optimized by training. A training set (80%, 4543 crystals) was randomly separated from the T2 dataset and the remaining 20% was used as test subset of m = 1136 crystals. Table 6.1 shows three averages of errors, over 10 runs:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m} |E_{true}(S_i) - E_{pred}(S_i)|^2} \tag{6.6}$$

is the root-mean-square error in the lattice energy averaged over m crystals $S_1, .., S_m$ from the test subset, then

$$MAE = \frac{1}{m} \max_{i=1,..,m} |E_{true}(S_i) - E_{pred}(S_i)| \tag{6.7}$$

is the mean absolute error and

$$MAPE = \frac{1}{m} \max_{i=1,..,m} \frac{|E_{true}(S_i) - E_{pred}(S_i)|}{E_{true}(S_i)} \tag{6.8}$$

is the mean absolute percentage error.

All errors *RMSE, MAE, MAPE* are consistent across different values of $k$ as shown in Table 6.1, where each row corresponds to 10 runs for a specific AMD length $k$, with the

| k | RMSE $\pm std$ | MAE $\pm std$ | MAPE $\pm std$ | training time, sec | full test time, ms |
|---|---|---|---|---|---|
| 50 | 6.503 ± 0.123 | 4.900 ± 0.86 | 3.509 ± 0.059 | 627 ± 85 | 15961 ± 183 |
| 100 | 6.344 ± 0.152 | 4.801 ± 0.103 | 3.439 ± 0.070 | 349 ± 47 | 7979 ± 564 |
| 150 | 6.607 ± 0.119 | 4.977 ± 0.077 | 3.559 ± 0.053 | 400 ± 23 | 12789 ± 203 |
| 200 | 6.617 ± 0.147 | 4.966 ± 0.114 | 3.554 ± 0.079 | 506 ± 40 | 15943 ± 46 |
| 250 | 6.517 ± 0.109 | 4.914 ± 0.082 | 3.514 ± 0.055 | 574 ± 91 | 16464 ± 193 |
| 300 | 6.632 ± 0.139 | 5.003 ± 0.092 | 3.577 ± 0.062 | 545 ± 15 | 16431 ± 52 |
| 350 | 6.615 ± 0.077 | 4.990 ± 0.077 | 3.581 ± 0.053 | 500 ± 22 | 12395 ± 44 |
| 400 | 6.611 ± 0.149 | 4.984 ± 0.080 | 3.569 ± 0.053 | 585 ± 25 | 17906 ± 201 |
| 450 | 6.559 ± 0.179 | 4.954 ± 0.127 | 3.545 ± 0.085 | 512 ± 21 | 12927 ± 67 |
| 500 | 6.622 ± 0.116 | 5.004 ± 0.092 | 3.581 ± 0.068 | 598 ± 24 | 18429 ± 219 |

Table 6.1: [43] Table 1. The Gaussian Process with the Rational Quadratic Kernel predicts the energy reported in [37] and discussed in subsection 3.2.3 with the mean absolute error (MAE) of less than $5\frac{kJ}{mol}$ on $m = 1136$ crystals. The training is performed on the isometry invariants $AMD^k$ of 4543 crystals for several $k$.

empirical standard deviation $\pm std$ computed from all runs. Speed is the key advantage over past methods and reports 10 min time for training 4543 vectors AMD(k). The last column collects the full test time on $m = 1136$ crystals. Referring to the highest $k$ chosen ($k = 500$), we achieved a mean absolute error of $5.004 \pm 0.092$ that is the maximum among all the runs for each value of $k$.

We tried to run the Gaussian Process Regression on the density functions descriptors $\psi_k(t)$ [12], which are continuous isometry invariants extending the single-value density for a variable radius $t \geq 0$. Smaller values of AMD-based predictions were reported with respect to density functions $\psi_k$, which are slower to compute than AMD (cubic time in $k$). Two other machine learning methods were trained on AMD and density functions such as the Random Forest [30] and Dense Neural Network [18] that performed slightly worse than the Gaussian Process, although the training and test times were much faster (seconds instead of minutes) [41].

# Chapter 7

# Conclusions

This PhD project found its importance in developing new geometric tools called isometry invariants of crystal structures. This section will highlight my main contribution to the projects.

My first project on crystals started in 2018 when I studied the concept of a Voronoi domain. We needed a proper metric to compare crystal lattices without considering discrete classifications such as unit cell parameters-based methods (e.g. Niggli's reduced cell). I contributed to the development and implementation of the algorithm to compute and compare Voronoi Domains of lattices (chapter 4). Most importantly, we addressed the problem of finding a continuous metric between crystal lattices (problem 2.6). It was our first paper published about periodic geometry where the definition of a metric was highlighted. The key contribution includes the development of two **metrics on lattices** (rotationally-invariant 4.10, and scale-invariant 4.14) defined in terms of Voronoi domains that satisfy the metric axioms (theorems 4.11 and 4.15). My software computes the Voronoi Domain for each given lattice in a dataset. All crystal files must be in CIF format. The algorithm proceeds by triangulating the domain of lattice points and computing their Voronoi Diagram. Moreover, the above-mentioned metrics can be calculated on all pairs of lattices during the software run by specifying the related command line options. The metric computation runs in multithreading where the number of threads can be chosen by the user and different rotation samples are processed in parallel for each pair of lattices (see [28] for more details). This software can be used to output information on lattices that can be mapped to Voronoi domains and saved in different formats for both visualization and geometric information storage. In addition, the metrics values are stored in comma separated values (csv) files and can be used to plot heatmaps or dendrograms. So, the end-user can analyse clusters of crystal lattices that are grouped by similarity from these files. I implemented this algorithm in **C++** and used different software libraries such as **CGAL** for computational geometry algorithms [51], **Gemmi** for handling the unit cell data [17], **VTK** for visualizing Voronoi Domains [53] and **Eigen** to cope with linear algebra data [14]. Heatmaps of the results were generated with **R** statistics programming language [38].

After dealing with lattices, we moved our focus on inter-atomic distances inside the unit cell. For this second project, I contributed to the development and implementation of the early version of the algorithm to **compute inter-atomic distances** by proposing and using the kd-tree data structure to efficiently query points in a neighbourhood. This algorithm was then proved and supported by the other authors with the theorems and proofs in Chapter 5. Average Minimum Distance (definition 5.1) has been proved to be an isometry invariant (theorem 5.2) that led us to solve the problem of **isometry classification** (problem 2.4) because any two equivalent crystal structures map to the same AMD vector. Since small perturbations affect the rigid form of a crystal, the **continuity** property must be taken into consideration to correctly detect these small changes (theorem 5.6). My AMD software takes in input a dataset of CIF files and for each crystal builds the $k$-d tree to speed up the neighbours retrieval. Supported by the $k$d-tree, it computes the pairwise distances between the particles inside the unit cell and their neighbours. At the end, all distances are averaged by the number of particles. AMD computations run in multithreading and therefore more crystals are processed in parallel (see [27] for more details). This software can be used to output a set of isometry invariants ($AMD_j$) that can identify a crystal structure, and be used by further studies that need to consider continuous geometric information of a crystal. Finally, I implemented the software in **C++** using the following libraries: **Gemmi** for handling the unit cell data [17], **Eigen** to cope with linear algebra vectors [14], **Boost** and **OpenBabel** for building molecular graphs [7] [34], and **Nanoflann** for building the kd-tree and for quick distance computations [6].

Capable of retaining vital information about the geometric structure of a crystal, our invariants can be used for predicting chemical properties. My third contribution includes the visualization of plots related to the change of energy values versus different similarity measures. In addition, I checked and proposed various methods to undergraduate students to **predict the lattice energy** by using AMD as a source (problem 6.1). We found that Gaussian Processes performed better than other regression algorithms in our experiments (Section 6.4). Although the prediction was not perfect and the structure-property problem

(6.1) is not solved, the key point was to highlight the continuity of the lattice energy with respect to the AMD invariant and other similarity measures. I used **R** statistic programming language [38] to plot and visualize how the **energy continuously changes** when proper invariants, such as AMD, are considered (Section 6.3).

# Acknowledgements

I would like to acknowledge and give my thanks to my supervisors Dr. Vitaliy Kurlin and Prof. Andrew I. Cooper, who helped and supported me with their advice through all the last 4 years of my PhD. Their guidance led me to complete a project that included different type of topics as well as to learn more and more about crystal structures and computational geometry. I also want to thank my IPAP members Linjiang Chen and Vladimir Gusev for their advice during my yearly assessments about how to improve each topic and my skills.

My special thanks go to my family, who continuously sustained me during all these years of studies and their understanding when undertaking my research and writing my project.

Finally, I would like to thank the Electrical Engineering, Electronics and Computer Science School of Liverpool for giving me the opportunity to have a bursary to fund my studies through my period of stay in Liverpool.

# List of Figures

# List of Tables

# Bibliography

[1] L. C. Andrews, H. J. Bernstein, and G. A. Pelletier. "A perturbation stable cell comparison technique". In: *Acta Crystallographica Section A* 36.1 (1980), pp. 248–252.

[2] O. Anosova and V. Kurlin. "An Isometry Classification of Periodic Point Sets". In: *Discrete Geometry and Mathematical Morphology.* Ed. by Joakim Lindblad, Filip Malmberg, and Nataša Sladoje. Cham: Springer International Publishing, 2021, pp. 229–241. ISBN: 978-3-030-76657-3.

[3] J. Behler. "Atom-centered symmetry functions for constructing high-dimensional neural network potentials." In: *The Journal of chemical physics* 134.7 (2011).

[4] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars. *Computational Geometry: Algorithms and Applications.* Springer, 2010.

[5] S. J. L. Billinge. "The rise of the X-ray atomic pair distribution function method: a series of fortunate events." In: *Phil.Trans. R. Soc. A377: 20180413* (2019).

[6] J. L. Blanco and P. K. Rai. *nanoflann: a C++ header-only fork of FLANN, a library for Nearest Neighbor (NN) with KD-trees.* https://github.com/jlblancoc/nanoflann. 2014.

[7] *Boost.* https://www.boost.org/.

[8] R. A. Brown. "Building a Balanced $k$-d Tree in $O(kn \log n)$ Time". In: *J. Computer Graphics Techniques* 4.1 (2015), pp. 50–68.

[9] A. Burrows, J. Holman, A. Parsons, G. Pilling, and G. Price. *Chemistry 3: Introducing inorganic, organic and physical chemistry.* 2017.

[10] J. Chisholm and S. Motherwell. "COMPACK: a program for identifying crystal structure similarity using distances". In: *J. Applied Crystallography* 38.1 (2005), pp. 228–231.

[11] A. I. Cooper. "Porous Molecular Solids and Liquids". In: *ACS Cent. Sci.* 3 (2017), pp. 544–553. DOI: 10.1021/acscentsci.7b00146.

[12] H. Edelsbrunner, T. Heiss, V. Kurlin, P. Smith, and M. Wintraecken. "The Density Fingerprint of a Periodic Point Set". In: *Proceedings of Symposium on Computational Geometry*. 2021. DOI: 10.4230/LIPIcs.SoCG.2021.32.

[13] O. Egorova, R. Ha zi, D.C. Woods, and G.M. Day. "Multifidelity statistical machine learning for molecular crystal structure prediction." In: *The Journal of Physical Chemistry A 124* 124.39 (2020), pp. 8065–8078.

[14] *Eigen library*. https://gitlab.com/libeigen/eigen.

[15] G. de la Flor, D. Orobengoa, E. Tasci, J. M. Perez-Mato, and M. I. Aroyo. "Comparison of structures applying the tools available at the Bilbao Crystallographic Server". In: *J. Appl. Cryst.* 49 (2016), pp. 653–664. DOI: https://doi.org/10.1107/S1600576716002569.

[16] R. J. Gdanitz. "Prediction of molecular crystal structures by Monte Carlo simulated annealing without reference to diffraction data". In: *Chem. Phys. Lett.* 190 (1992), pp. 391–396.

[17] *Gemmi library*. https://gemmi.readthedocs.io/en/latest/.

[18] I. Goodfellow, Y. Bengio, and A. Courville. "Deep learning." In: *MIT Press* 1 (2016).

[19] E. Gross and R.: Dreizler. "Density functional theory." In: *Springer Science & Business Media* 3 (2013).

[20] T. A. Halgren. "Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94". In: *Journal of Computational Chemistry* 17.5-6 (1996), pp. 490–519. DOI: https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.

[21] J. R. Holden, Z. Y. Du, and H. L. Ammon. "Prediction of possible crystal structures for C-, H-, N-, O- and F-containing organic compounds". In: *J. Comput. Chem.* 14 (1993), pp. 422–437.

[22] M. Jansen and C. Schon. “”Design” in Chemical Synthesis - An Illusion?” In: *Angew. Chem. Int. Ed.* 45 (2006), pp. 3406–3412. DOI: 10.1002/anie.200504510.

[23] B. Joe. “Construction of three-dimensional Delaunay triangulations using local transformations”. In: *Computer Aided Geometric Design* 8 (1991).

[24] C. KI Williams. “Gaussian processes for machine learning”. In: *Taylor & Francis* (2006).

[25] V. Kurlin. “A complete isometry classification of 3-dimensional lattices”. In: *arXiv* (2022). URL: https://arxiv.org/abs/2201.10543.

[26] V. A. Kurlin. “Mathematics of 2-dimensional lattices”. In: *Foundations of Computational Mathematics* (2023).

[27] M. M. Mosca. *C++ AMD github.* URL: https://github.com/mmmosca/AMD.

[28] M. M. Mosca. *VoronoiLatticeDistance github.* URL: https://github.com/mmmosca/VoronoiLatticeDistances.

[29] M. M. Mosca and V. Kurlin. “Voronoi-based similarity distances between arbitrary crystal lattices”. In: *Cryst. Res. Technol.* 55 (2020). DOI: 10.1002/crat.201900197.

[30] A.J. Myles, R.N. Feudale, Y. Liu, N.A. Woody, and S.D. Brown. “An introduction to decision tree modeling.” In: *Journal of Chemometrics* 18.6 (2004), pp. 275–285.

[31] S.R. Niketic and K. Rasmussen. “The consistent force field: a documentation.” In: *Springer Science & Business Media* 3 (2012).

[32] M. O’Searcoid. “Metric spaces”. In: *Springer Science & Business Media* (2006).

[33] A. Oganov. “Modern methods of crystal structure prediction”. In: *Wiley & Son* (2011).

[34] *Openbabel.* http://openbabel.org/wiki/Main_Page.

[35] Niggli P. In: *Handbuch der Experimentalphysik, Leipzig: Akademische Verlagsgesellschaft* 7 (1928), p. 750.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and et al. “Scikit-learn: Machine learning in python.” In: *Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[37] A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. A. Little, S. Y. Chong, B. J. Slater, D. P. McMahon, B. Bonillo, C. J. Stackhouse, A. Stephenson, C. M. Kane, R. Clowes, T. Hasell, A. I. Cooper, and G. M. Day. "Functional materials discovery using energy-structure-function maps". In: *Nature* 543 (2017), pp. 657–664. DOI: `10.1038/nature21419`.

[38] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria, 2020. URL: `https://www.R-project.org/`.

[39] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard III, and W. M. Skiff. "UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations". In: *J. Am. Chem. Soc.* (1992).

[40] D. Reem. "The geometric stability of Voronoi diagrams with respect to small changes of the sites". In: (2011). DOI: `10.48550/ARXIV.1103.4125`. URL: `https://arxiv.org/abs/1103.4125`.

[41] J. Ropers. *Applying machine learning to geometric invariants of crystals to predict crystal energy.* URL: `https://github.com/JRopes/CrystalEnergyPrediction`.

[42] J. Ropers, M. M. Mosca, O. Anosova, and V. Kurlin. "Introduction to invariant-based machine learning for periodic crystals." In: *Acta Cryst. A77, C671.* (2021).

[43] J. Ropers, M. M. Mosca, O. Anosova, V. Kurlin, and A. I. Cooper. "Fast predictions of lattice energies by continuous isometry invariants of crystal structures". In: *International Conference on Data Analytics and Management in Data Intensive Domains.* 2022, pp. 178–192.

[44] B. L. Rotschild and E. G. Straus. "On triangulations of the convex hull of n points". In: *Combinatorica* 5 (1985). URL: `https://doi.org/10.1007/BF02579380`.

[45] Huai S. "COMPASS: An ab Initio Force-Field Optimized for Condensed-Phase ApplicationssOverview with Details on Alkane and Benzene Compounds". In: *J. Phys. Chem. B* (1998).

[46] Huai S., Stephen J. M., Jon R. M., and Arnold T. H. "An ab Initio CFF93 All-Atom Force Field for Polycarbonates". In: *J. Am. Chem. Soc.* (1994).

[47] P. Sacchi, M. Lusi, A.J. Cruz-Cabeza, E. Nauha, and J. Bernstein. "An extensible neural network potential with dft accuracy at force field computational cost." In: *Chem. Science* 8 (2017), pp. 3139–3203.

[48] P. Sacchi, M. Lusi, A.J. Cruz-Cabeza, E. Nauha, and J. Bernstein. "Same or different that is the question: identification of crystal forms from crystal structure data." In: *CrystEngComm* 22.43 (2020), pp. 7170–7185.

[49] K. Schutt, H. Glawe, F. Brockherde, A. Sanna, K.R. Muller, and E. Gross. "How to represent crystal structures for machine learning: Towards fast prediction of electronic properties." In: *Physical Review B* 89.20 (2014).

[50] L. E. Smart and E. A. Moore. *Solid State Chemistry: An Introduction.* 2005.

[51] The CGAL Project. *CGAL User and Reference Manual.* 4.14.3. CGAL Editorial Board. URL: `https://doc.cgal.org/4.14.3/Manual/index.html`.

[52] M. Valle and A. R. Oganov. "Crystal fingerprint space-a novel paradigm for studying crystal structure sets." In: *Acta Crystallographica Section A: Foundations of Crystalloraphy* 66.5 (2010), pp. 507–517.

[53] *Visualization ToolKit library.* `https://vtk.org/`.

[54] D. J. Wales and J. P. K. Doyle. "Global optimization of clusters, crystals and biomolecules". In: *Science* 285 (1999), pp. 1368–1372.

[55] D.J. Wales. "Exploring energy landscapes." In: *Annual review of physical chemistry* 69 (2018), pp. 401–425.

[56] L. Ward, R. Liu, A. Krishna, V. Hegde, A. Agrawal, A. Choudhary, and C. Wolverton. "Including crystal structure attributes in machine learning models of formation energies via voronoi tessellations". In: *Physical Review B* 96.2 (2017).

[57] D. F. Watson. "Computing the $n$-dimensional Delaunay tassellation with application to Voronoi polytopes". In: *The Computer Journal* 24.2 (1981).

[58] D. Widdowson. *Python AMD github.* URL: `https://github.com/dwiddo/average-minimum-distance`.

[59] D. Widdowson and V. Kurlin. "Resolving the data ambiguity for periodic crystals". In: *Advances in Neural Information Processing Systems (Proceedings of NeurIPS 2022)* 35 (2022).

[60] D. Widdowson, M. M. Mosca, A. Pulido, V. Kurlin, and A. I. Cooper. “Average Minimum Distances of periodic point sets - foundational invariants for mapping periodic crystals”. In: *MATCH Communications in Mathematical and in Computer Chemistry* 87.3 (2022), pp. 529–559. DOI: 10.46793/match.87-3.529W.

[61] D. Widdowson, M. M. Mosca, A. Pulido, V. Kurlin, and A. I. Cooper. “The asymptotic behaviour and a near linear time algorithm for isometry invariants of periodic sets”. In: *arXiv* (2021).

[62] S. M. Woodley and R. Catlow. “Crystal structure prediction from first principles”. In: *Nature materials* 7 (2008), pp. 937–946.