Bachelor Mathematics

*Bachelor thesis*

# Topological Data Analysis in Crystallography

by

Mark Lapidus

July 1, 2024

Supervisor: Magnus Bakke Botnan, Senja Barthel

Department of Mathematics

Faculty of Science

VU VRIJE UNIVERSITEIT AMSTERDAM

# Abstract

This thesis explores the application of topological data analysis to crystallography with a specific focus on classifying crystal systems through the use of persistent homology. The methodology involves constructing Rips complexes from crystallographic data, computing persistent homology to capture the multi-scale topological features, and employing machine learning techniques to classify crystal systems based on these features. The results demonstrate that persistent homology can effectively distinguish between crystal systems, offering a new dimension of insight into their classification.



Title: Topological Data Analysis in Crystallography
Author: Mark Lapidus, m.lapidus@student.vu.nl, 2734866
Supervisor: Magnus Bakke Botnan, Senja Barthel
Date: July 1, 2024

Department of Mathematics
Vrije Universiteit Amsterdam
de Boelelaan 1081, 1081 HV Amsterdam
https://www.math.vu.nl/

# Contents


**1. Introduction** **5**

**2. Crystallography** **7**

2.1. Crystal Structures . . . . . 8

2.2. Crystal Systems . . . . . 11

2.2.1. Point Groups . . . . . 11

2.2.2. Space Groups . . . . . 14

2.2.3. Crystal Families . . . . . 14

**3. Topological Data Analysis** **16**

3.1. Complexes . . . . . 16

3.1.1. Simplices . . . . . 16

3.1.2. Simplicial Complexes . . . . . 17

3.1.3. Rips Complex . . . . . 18

3.2. Homological Algebra . . . . . 18

3.2.1. Euler Characteristic . . . . . 18

3.2.2. n-Chains . . . . . 19

3.2.3. Boundary Operator . . . . . 19

3.2.4. Chain Complexes . . . . . 20

3.2.5. Betti Numbers . . . . . 21

3.2.6. Simplicial Homology . . . . . 22

3.3. Persistent Homology . . . . . 23

3.3.1. Chain Maps . . . . . 24

3.3.2. Persistence Barcodes and Diagrams . . . . . 25

**4. Analysis** **26**

4.1. Constructing Crystals . . . . . 26

4.1.1. Fractional Coordinate System . . . . . 26

4.1.2. Symmetry Application . . . . . 27

4.1.3. Transformation to Cartesian Coordinate System . . . . . 28

4.1.4. Unit Cell Normalisation . . . . . 28

4.1.5. Periodic Boundary Conditions . . . . . 29
4.2. Building Rips Complexes . . . . . 30
4.3. Random Crystals . . . . . 31
4.3.1. Group Order . . . . . 31
4.4. Classifying Space Groups . . . . . 32

**5. Conclusion** **37**

**A. CIF File** **39**

# 1. Introduction

In this project we explore two, seemingly unrelated fields: crystallography and topological data analysis.

Crystallography is the study of crystal structures and their properties. A fundamental aspect of crystallography is the classification of crystal structures into crystal systems, based on their symmetry properties, which are described by point and space groups. Early crystallographers, such as René-Just Haüy, recognised that crystals could be described by repeating units, while a thorough mathematical foundation of crystal symmetry was only developed in the 19-th century, culminating in the derivation of the 230 unique space groups by Evgraf Fedorov and Arthur Schoenflies independently in 1891 [13].

Topological data analysis, on the other hand, stems from two fundamental questions: how can high-dimensional structures be inferred from low-dimensional representations, and how can discrete points be assembled into a global structure? Ghrist and Edelsbrunner [11, 10] address these questions by converting sets of data points into families of simplicial complexes, indexed by a proximity parameter. These complexes are then analyzed using algebraic topology, particularly persistent homology, which encodes the homological features of a point cloud in the form of a barcode, providing a visual summary of the data's topological properties.

So what do these topics have in common? Carlsson [7] demonstrated the capability to differentiate between two distinct planar crystalline structures using topological methods. In their work, one structure corresponded to a rectangular lattice, while the other corresponded to a hexagonal lattice. The differences between the two structures were due to how the Rips complex detected homologies. The rectangular lattice produced rectangular features that were identifiable, whereas the hexagonal lattice produced triangular features that were not. This work highlighted the utility of persistent homology and topological data analysis in distinguishing between different crystalline structures, providing a powerful tool for their analysis.

We, however, are interested in a more general setting where we aim to capture the differences in all crystal systems using homological features. To achieve this, we will first describe how crystal structures are categorized based on their symmetries, eventually deriving the seven crystal systems. Next, we will study simplicial complexes, which will allow us to convert a point cloud into a topological space. Additionally, we will delve into homological algebra, from which we will learn how to count $n$-dimensional holes in a simplicial complex. Our work culminates in the Analysis chapter, where we combine our newly acquired knowledge to study the actual crystal symmetries, leading to some remarkable results.

# 2. Crystallography

First, let us understand what a crystal structure is comprised of.

**Definition 2.0.1** (Lattice)**.** A **lattice** is a discrete array of points in a vector space.

In crystallography, however, one encounters lattices which display translational symmetry. In $\mathbb{R}^2$, for instance, a translation of the lattice by the vector

$$\vec{T} = m\vec{a}_1 + n\vec{a}_2$$

where $m, n \in \mathbb{Z}$, has the property that all the points in the lattice can be accessed and have an identical environment. Vectors $\vec{a}_1$ and $\vec{a}_2$ here are referred to as **primitive translation vectors**, and the lattices that can be characterised in this way are called Bravais lattices.

**Definition 2.0.2** (Bravais Lattice)**.** A **Bravais lattice** is a lattice that is generated by the primitive translation vectors.

Note, however, that the choice of the primitive translation vectors is not unique for a given Bravais lattice. In other words, any two linearly independent vectors $\vec{a}_1$ and $\vec{a}_2$ can be used to generate a Bravais lattice in $\mathbb{R}^2$.

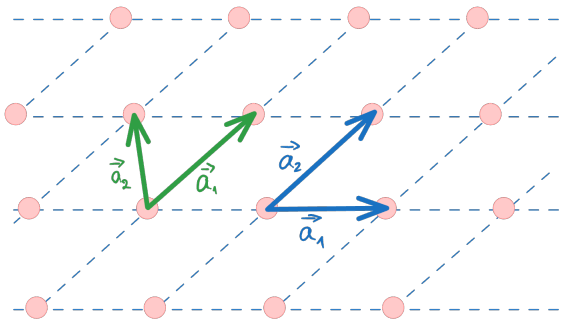



Figure 2.1.: Bravais lattice generated by two primitive translation vectors $\vec{a}_1$ and $\vec{a}_2$.

This definition can be extended to $\mathbb{R}^3$ by adding a third linearly independent vector $\vec{a}_3$. One could, however, replace the lattice points [1] by more complex objects, such as a group

[1] More generally - lattice sites.

of atoms or a molecule, which, in turn, would form its basis. This leads us to the definition of a crystal structure.

**Definition 2.0.3** (Crystal Structure)**.** A **crystal structure** is a Bravais lattice with a basis added to each lattice site.

It is worth noting that Bravais lattices assume that the basis is reduced to a single point. As a result, Bravais lattices correspond to crystal structures when the basis of a crystal is in form of atoms.

## 2.1. Crystal Structures

Since crystals can, in theory, extend infinitely, we need to define a section that can be repeated - its unit cell.

**Definition 2.1.1** (Unit Cell)**.** The **(conventional) unit cell** is the smallest repeating unit of atoms having the full symmetry of a crystal.

In other words, the unit cell completely reflects the symmetry and structure of the entire crystal, which is built up by repetitive translation of the unit cell along its primitive translation vectors. Note, however, that for a given crystal there may be different possible unit cells, which are varied by their inner symmetries.

### 2D Bravais Lattices

In the example below, the cell (a) is a valid unit cell spanned by two linearly independent translation vectors. However, it does not show the full symmetry of the lattice, whereas cell (b) clearly shows the two axes of rotation, and thus is a more conventional choice. The cell (c) has the least number of symmetries among them.

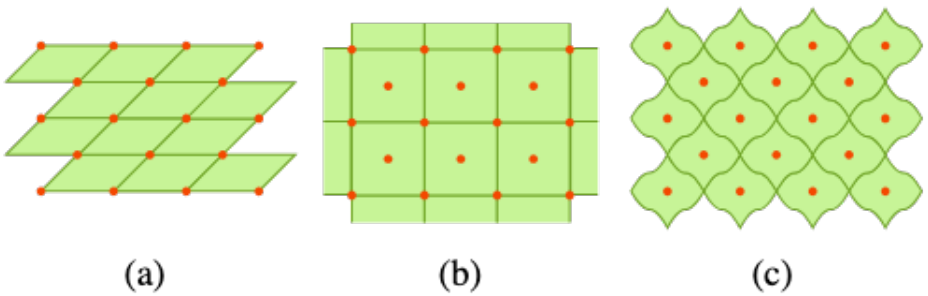



Figure 2.2.: Examples of different unit cells for a two-dimensional Bravais lattice.

And so the general approach is to choose the unit cell with the highest level of inner symmetry. Notably, in $\mathbb{R}^2$, there are 5 possible lattice systems summarised in the table below.

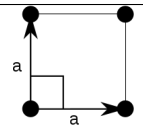
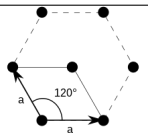
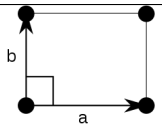
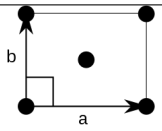
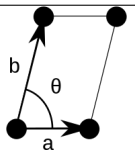
| **Lattice System** | **Axial Distances** | **Axial Angle** | **Unit Cell** |
|---|---|---|---|
| Square | $a = b$ | $\theta = 90°$ |   |
| Hexagonal | $a = b$ | $\theta = 120°$ |   |
| Rectangular | $a \neq b$ | $\theta = 90°$ |   |
| Centered Rectangular | $a \neq b$ | $\theta = 90°$ |   |
| Rhomboidal | $a \neq b$ | $\theta \neq 90°$ |   |

Table 2.1.1.: The 5 Bravais lattice systems in $\mathbb{R}^2$.

Consequently, any other pattern could be reduced to one of them.

**Example 2.1.1: Why is there no centered square lattice?**

If a point was added to the center of a square lattice, the lattice would still be a square but with smaller axial distances.

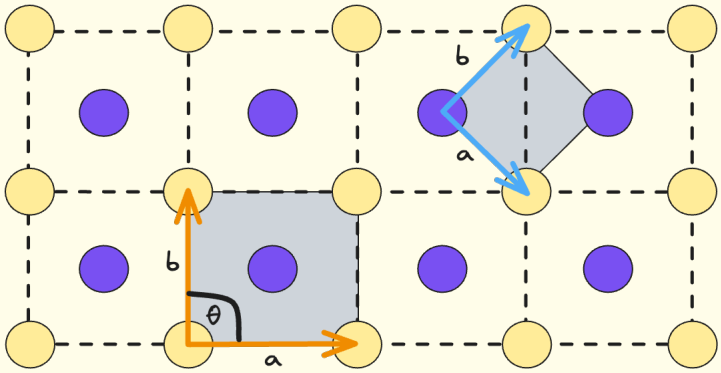



Figure 2.3.: Centered Square Lattice

The parameters of the unit cell are specified by the lattice constants.

**Definition 2.1.2** (Lattice Constants)**. Lattice constants** are the side lengths of the unit cell and the angles between them.

### 3D Bravais Lattices

In $\mathbb{R}^3$, the side lengths are commonly displayed as $a$, $b$ and $c$, and the angles are denoted by $\alpha$, $\beta$ and $\gamma$. Conventionally, $\alpha$ is the angle between $b$ and $c$, $\beta$ is the angle between $a$ and $c$, and $\gamma$ is the angle between $a$ and $b$, respectively.

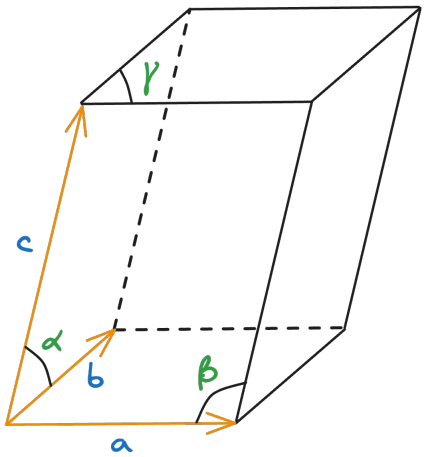



Figure 2.4.: A 3-dimensional unit cell specified by the principal axes and the angles between them.

Analogously to 5 Bravais lattice systems in $\mathbb{R}^2$, there can only be 14 Bravais lattices in $\mathbb{R}^3$ [14]. By removing the centering operations, these, in turn, can be grouped into 7 lattice systems [2], each characterised by the lattice constants. For the representations of their unit cells see the appendix.

| **Lattice System** | **Axial Distances** | **Axial Angles** |
|---|---|---|
| Triclinic | $a \neq b \neq c$ | $\alpha \neq \beta \neq \gamma$ |
| Monoclinic | $a \neq b \neq c$ | $\alpha = \gamma = 90° \neq \beta$ |
| Orthorhombic | $a \neq b \neq c$ | $\alpha = \beta = \gamma = 90°$ |
| Tetragonal | $a = b \neq c$ | $\alpha = \beta = \gamma = 90°$ |
| Rhombohedral | $a = b = c$ | $\alpha = \beta = \gamma < 120°, \neq 90°$ |
| Hexagonal | $a = b \neq c$ | $\alpha = \beta = 90°$, $\gamma = 120°$ |
| Cubic | $a = b = c$ | $\alpha = \beta = \gamma = 90°$ |

Table 2.1.2.: The 7 Bravais lattice systems in $\mathbb{R}^3$.

## 2.2. Crystal Systems

Besides the translational symmetry from which we derived 14 Bravais lattices, it is also worth looking at the point symmetries - the group of symmetry operations, such as reflections, rotations, inversions and their combinations, that leaves at least one point fixed. Note that any operation including lattice translation is discarded.

**Definition 2.2.1** (Point Group)**.** A **(crystallographic) point group** is a set of rigid transformations, excluding translation, that leave at least one point fixed and map a crystal onto itself.

### 2.2.1. Point Groups

According to the crystallographic restriction theorem, the rotational symmetries may only contain 1-fold, 2-fold, 3-fold, 4-fold and 6-fold axes of rotation [3] [9]. As a result, the number of crystallographic point groups is limited to 32 in $\mathbb{R}^3$ [4], compared to an infinite number of general point groups.

[2] One can assign each Bravais lattice system to a crystal system to get 7 crystal systems.

[3] In polar coordinates, an $n$-fold rotation axis of a point $(R, \theta, \varphi)$ perpendicular to $\theta$ can be written as $(R, \theta + \frac{2\pi}{n}, \varphi)$.

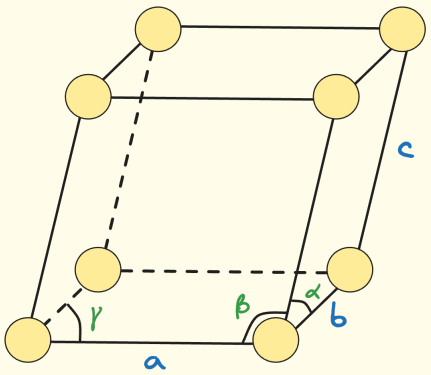
Point groups form the foundation for the classification of crystals: one could, for instance, count the number of axes of rotation and their respective multiplicities in order to compare different crystals in regard to their symmetry. This is the basis for defining the crystal systems.

**Definition 2.2.2** (Crystal System)**.** A **crystal system** is a classification of crystals for point groups defined by rotational axes.

The seven crystal systems are: triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal and cubic. But let us reconstruct the point groups from some of them.

**Example 2.2.1: Triclinic Point Groups**

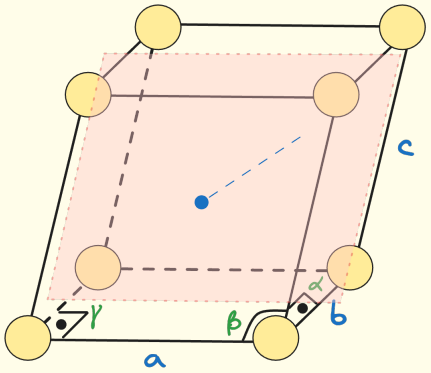
The triclinic crystal system has the lowest symmetry since its sides and angles are distinct. Besides the trivial 360° rotation, it has also an inversion symmetry. So there are only two point groups.




When some of the angles or edges are equivalent, more symmetries are formed. This leads us to the crystal with the second least number of symmetries - monoclinic.

**Example 2.2.2: Monoclinic Point Groups**

The monoclinic crystal system has a 2-fold axis of rotation, as well as the mirror plane reflection. Along with a 2-fold axis of rotation perpendicular to a mirror plane, this gives 3 point groups in total.




Similarly, the other crystal systems can be treated in the same manner, giving in total 32 point groups.

| **Crystal System** | **Number of Point Groups** | **Axial Distances** | **Axial Angles** |
|---|---|---|---|
| Triclinic | 2 | $a \neq b \neq c$ | $\alpha \neq \beta \neq \gamma$ |
| Monoclinic | 3 | $a \neq b \neq c$ | $\alpha = \gamma = 90° \neq \beta$ |
| Orthorhombic | 3 | $a \neq b \neq c$ | $\alpha = \beta = \gamma = 90°$ |
| Tetragonal | 7 | $a = b \neq c$ | $\alpha = \beta = \gamma = 90°$ |
| Trigonal | 5 | $a = b = c$ | $\alpha = \beta = \gamma < 120°, \neq 90°$ |
| Hexagonal | 7 | $a = b \neq c$ | $\alpha = \beta = 90°$, $\gamma = 120°$ |
| Cubic | 5 | $a = b = c$ | $\alpha = \beta = \gamma = 90°$ |

Table 2.2.1.: The 7 crystal systems and their point groups.

It is important to remember, however, that lattice constants are simply a consequence of the underlying symmetries, reflecting the intrinsic spatial arrangement of atoms within the crystal.

### 2.2.2. Space Groups

The space groups in three dimensions are formed by combining the 32 crystallographic point groups, each associated with one of the 7 crystal systems, with the 14 Bravais lattices. This means that the action of any element of a space group can be represented as the action of an element from the corresponding point group, optionally followed by a translation. A space group is, therefore, a combination of the translational symmetry of a unit cell, the point group symmetry operations of reflection, rotation, and inversion, as well as the screw axis [4] and glide plane [5] symmetry operations. Together, these symmetry operations result in a total of 230 distinct space groups, each corresponding to one of the 7 crystal systems, which are summarised in the table below.

| **Crystal System** | **Space Groups** |
|---|---|
| Triclinic | $1-2$ |
| Monoclinic | $3-15$ |
| Orthorhombic | $16-74$ |
| Tetragonal | $75-142$ |
| Trigonal | $143-167$ |
| Hexagonal | $168-194$ |
| Cubic | $195-230$ |

Table 2.2.2.: Crystal systems & Space groups.

### 2.2.3. Crystal Families

We can also unify the Bravais lattice systems and crystal systems into 6 crystal families by combining the rhombohedral lattice system and trigonal crystal system into the existing hexagonal crystal family. The reason being is that both trigonal and hexagonal crystal systems exhibit 3-fold rotational symmetry [6]. And since the crystal systems are a classification for point groups, while the lattice systems are a classification for Bravais lattices, crystal families can be regarded as a classification for space groups.

In summary, we depict the relationship between lattice and crystal systems in the diagram.

[4] A symmetry operation which combines translation & rotation.
[5] A symmetry operation which combines translation & reflection.
[6] Note that a 6-fold axis of rotation also includes 1-fold, 2-fold and 3-fold axes of rotation.

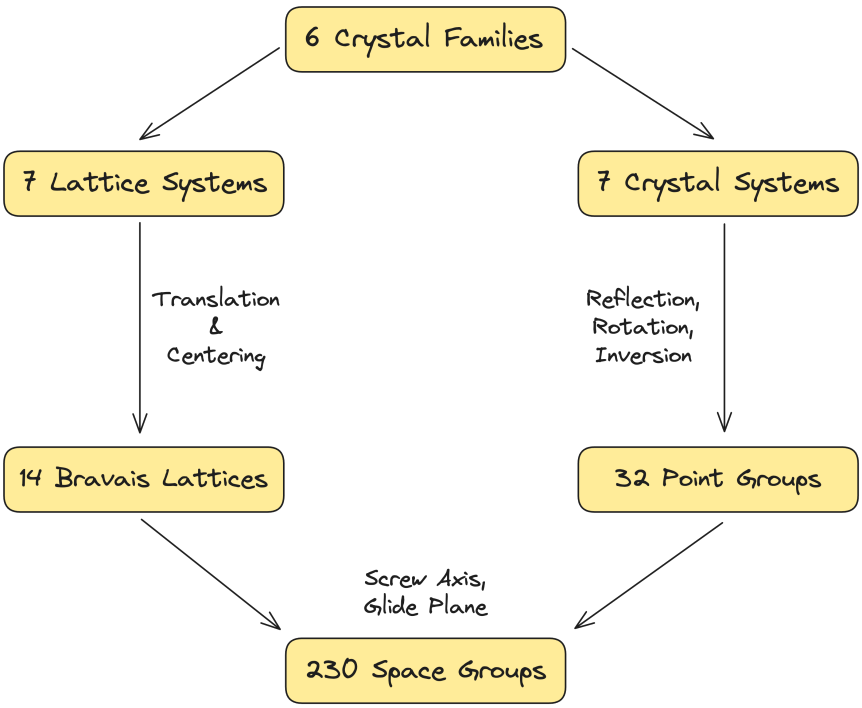



Figure 2.5.: Symmetry relationship between crystal & lattice systems.

# 3. Topological Data Analysis

## 3.1. Complexes

The concept of topological data analysis involves assigning topological invariants to data, which is usually presented as a discrete set, whose topology is rather trivial. To address this, we need to convert the data into a *continuous* object that is expected to be topologically similar to the underlying geometric shape from which the data was sampled. This transformation process is done by means of simplicial complexes.

### 3.1.1. Simplices

**Definition 3.1.1** (n-Simplex)**.** Let $\{p_0, \cdots, p_n\}$ be an affinely independent set of points in $\mathbb{R}^d$. The **n-simplex** $\sigma$ spanned by the points $p_i$ is the set of convex combinations $x \in \mathbb{R}^d$ of the form

$$x = \sum_{i=0}^{n} \lambda_i p_i$$

where $\sum_{i=0}^{n} \lambda_i = 1$ and $\lambda_i \geq 0$ for all $i$.

The dimension of a simplex $\sigma$ is given by $\dim \sigma = |\sigma| - 1$, so an $n$-simplex is of dimension $n$. Note that there can be no $n$-simplex in $\mathbb{R}^d$ for $n > d$.

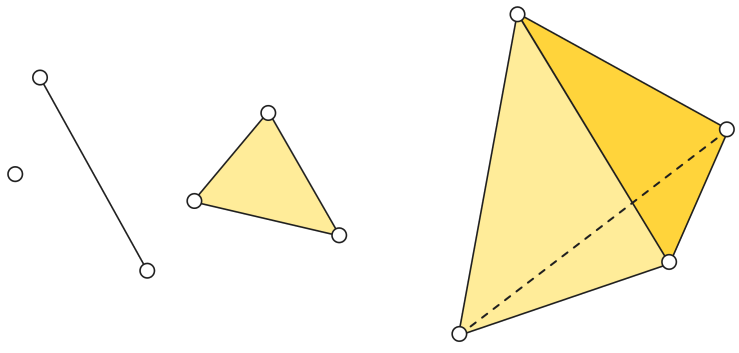


Figure 3.1.: Examples of simplices of dimensions: 0 - point, 1 - line segment, 2 - triangle and 3 - tetrahedron.

If $\tau \subseteq \sigma$ is spanned by a non-empty subset of the points defining $\sigma$, then $\tau$ is a **face**

[1] of $\sigma$. The **boundary** $\operatorname{Bd}\sigma$ of $\sigma$ is the union of all proper faces, and the **interior** $\operatorname{Int}\sigma = \sigma - \operatorname{Bd}\sigma$ is everything else. Hence, every point $x \in \sigma$ belongs to $\operatorname{Int}\sigma$ of exactly one face $\sigma$, specifically the one spanned by the points $p_i$.

### 3.1.2. Simplicial Complexes

By selecting sets of simplices that are closed under the operation of taking faces and ensuring there are no improper intersections, we form simplicial complexes.

**Definition 3.1.2** (Simplicial Complex)**.** A **simplicial complex** $K \in \mathbb{R}^d$ is a finite collection of simplices $\sigma$ such that:

1. Every face $\tau \subseteq \sigma$ is in $K$.
2. For any two simplices $\sigma_1, \sigma_2 \in K$, the non-empty intersection $\sigma_1 \cap \sigma_2$ is a face of both.

The **dimension** of a simplicial complex $K$ is the maximum dimension of its simplices.

**Example 3.1.1: Graph**

A graph is a 1-dimensional simplicial complex which consists of 0-simplices (vertices) and 1-simplices (edges).

The **underlying space** $|K|$ of a simplicial complex $K$ is the union of its simplices together with the subspace topology inherited from the Euclidean space in which the simplices belong.

In order to avoid the need for embedding a simplicial complex in a specific geometric space, those can be defined purely in terms of sets and their relationships.

**Definition 3.1.3** (Abstract Simplicial Complex)**.** An **abstract simplicial complex** is a finite collection of non-empty sets $A$ such that for each simplex $\sigma \in A$, all subsets of $\sigma$ are also in $A$.

Any simplicial complex $K$ can be defined as an abstract simplicial complex $A$ by replacing all simplices of $K$ with their corresponding sets of vertices, thus making $A$ a **vertex scheme** of $K$. As has been mentioned before, by constructing abstract simplicial complexes, we can work with sets instead of geometric simplices.

[1] We say that $\tau$ is a proper face of $\sigma$ if $\tau \subset \sigma$.

**Example 3.1.2: Vertex scheme of a triangle**

The 2-simplex connecting $p_0$, $p_1$ and $p_2$ has a vertex scheme $\{\{p_0\}, \{p_1\}, \{p_2\}, \{p_0, p_1\}, \{p_0, p_2\}, \{p_1, p_2\}, \{p_0, p_1, p_2\}\}$.

Similarly, an abstract simplicial complex, which is combinatorial in nature, can be **geometrically realised** into a simplicial complex. Specifically, every abstract simplicial complex of dimension $d$ has a geometric realisation in $\mathbb{R}^{2d+1}$ [10].

### 3.1.3. Rips Complex

Now that we have established the theoretical underpinnings of simplicial complexes, we can discuss the computational methods for constructing them from a collection of points in a metric space. The most widely used methods include Čech [12] and Rips complexes. Our choice falls on the Rips complex as it is computationally less expensive, since it can be stored as a graph, unlike the Čech complex, which requires storing the entire boundary operator [11]. Moreover, there is an efficient implementation of the Ripser complex in Python through the Ripser package [15, 3].

**Definition 3.1.4** (Rips Complex)**.** The **Rips complex** of a finite metric space $(X, d)$ at threshold $\varepsilon > 0$ is the simplicial complex

$$\mathcal{R}_\varepsilon = \{\sigma \subseteq X : d(x_i, x_j) \leq \varepsilon \quad \forall x_i, x_j \in \sigma\}.$$

Intuitively, given a set of points, a radius is chosen, and an edge is formed between any two points whose distance is less than or equal to this radius. Higher-dimensional simplices such as triangles and tetrahedra are then added for any collection of points that are pairwise connected.

## 3.2. Homological Algebra

Deciding if two simplicial complexes are homeomorphic is an undecidable problem, but it is decidable to show if two spaces are not homeomorphic. Therefore, we need to find topological properties which remain invariant under homeomorphisms.

### 3.2.1. Euler Characteristic

One of the simplest topological invariants is the Euler characteristic.

**Definition 3.2.1** (Euler Characteristic [6])**.** Let $K$ be a simplicial complex and let $K_i$ denote the number of $i$-dimensional simplices in $K$. The **Euler characteristic** of $K$ is the integer

$$\chi(K) = \sum_{i=0}^{\infty} (-1)^i K_i.$$

Even though Euler characteristic outputs a single quantity, which only provides a global summary of the space's topology, it already helps in classifying surfaces and spaces. That is, if two spaces do not have the same Euler characteristic, it means that they are topologically distinct and cannot be homeomorphic to each other. We, however, aim at providing a more detailed breakdown of the topological features in each dimension. For that, maps between simplices within a simplicial complex need to be studied.

### 3.2.2. n-Chains

**Definition 3.2.2** (Vector Space of $n$-chains)**.** The **vector space of $n$-chains** $C_n(K)$ in a simplicial complex $K$ is a formal sum of $n$-simplices of $K$ with coefficient in $\mathbb{Z}_2$:

$$C_n(K) = \Big\{ \sum_{i=0}^{n} a_i \sigma_i \mid a_i \in \mathbb{Z}_2, \sigma_i \in K \Big\}$$

*Remark.* For simplicity, we define the $n$-chains $C_n(K)$ as the vector space over $\mathbb{Z}_2 = \{0, 1\}$, so that the orientations can be ignored.

Note that the $n$-simplices of $K$ form a basis of $C_n(K)$, so $\dim C_n(K)$ equals their count.

### 3.2.3. Boundary Operator

We can now define the map $\partial_n : C_n(X) \to C_{n-1}(X)$ that is sending a linear combination of $n$-simplices to a linear combination of their faces. Observe that by linearity of $\partial_n$ we have that

$$\partial_n \sum_{i=0}^{n} a_i \sigma_i = \sum_{i=0}^{n} a_i \partial_n(\sigma_i)$$

so it is sufficient to specify the action $\partial_n$ on a single simplex.

**Definition 3.2.3** (Boundary Operator)**.** Given a simplicial complex $K$, the **boundary operator** $\partial_n : C_n(K) \to C_{n-1}(K)$ is the linear transformation that assigns each simplex $\sigma = \{p_0, \ldots, p_n\} \in K$ to its boundary:

$$\partial_n(\{p_0, \ldots, p_n\}) = \sum_{i=0}^{n} \{p_0, \ldots, \hat{p_i}, \ldots, p_n\}.$$

*Remark.* In other words, the boundary of an $n$-simplex is the formal sum of its $(n-1)$-dimensional faces that do not contain the $i$-th vertex.

**Example 3.2.1: Boundary of a triangle**

Consider an example of a triangle scheme. The boundary is $\partial_2(\{p_0, p_1, p_2\}) = \{p_0, p_1\} - \{p_0, p_2\} + \{p_1, p_2\}$

In $C_n(K)$, the elements of $\operatorname{Im} \partial_{n+1}$ are called **boundaries**, and the elements of $\operatorname{Ker} \partial_n$ are called **cycles**. A cycle is thus an $n$-chain whose boundary is zero.

**Lemma 3.2.1.** *The composition $\partial_n \circ \partial_{n+1} = 0$ for all $n \geq 0$.*

*Proof.*

$$\begin{aligned}
\partial_n \circ \partial_{n+1}(\sigma) &= \partial_n \Big( \sum_{i=0}^{n+1} \{p_0, \ldots, \hat{p}_i, \ldots p_{n+1}\} \Big) \\
&= \sum_{i=0}^{n+1} \partial_n(\{p_0, \ldots, \hat{p}_i, \ldots p_{n+1}\}) \\
&= \sum_{i=0}^{n+1} \sum_{j=0, j\neq i}^{n+1} \{p_0, \ldots, \hat{p}_j, \ldots, \hat{p}_i, \ldots p_{n+1}\} \\
&= 0.
\end{aligned}$$

And since for $i \neq j$ the $(n-1)$-simplex $\{p_0, \ldots, \hat{p}_j, \ldots, \hat{p}_i, \ldots p_{n+1}\}$ appears twice in the sum, the result follows as we are working in $\mathbb{Z}_2$. ■

From this result it follows that every boundary is a cycle, meaning that the image of the boundary operator $\partial_{n+1}$ lies within the kernel of the boundary operator $\partial_n$. We then denote the vector space of $n$-cycles by

$$Z_n(K) = \operatorname{Ker} \partial_n = \{c \in C_n(K) : \partial_n(c) = 0\}$$

and the vector space of $n$-boundaries by

$$B_n(K) = \operatorname{Im} \partial_{n+1} = \{\partial_{n+1}(d) : d \in C_{n+1}(K)\}.$$

### 3.2.4. Chain Complexes

Observe that the collection of $n$-chains are connected by the boundary operators into a chain complex $C_\bullet$ of the form:

$$\cdots \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \cdots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0.$$

Whenever we have a chain complex, we can examine the extent to which $\operatorname{Im} \partial_{n+1}$ equals $\operatorname{Ker} \partial_n$. If the sequence is exact at $C_n(K)$, then $\operatorname{Ker} \partial_n = \operatorname{Im} \partial_{n+1}$, meaning $\mathcal{H}_n(K) = 0$ as we will see further. This would indicate that there are no $n$-dimensional holes in the simplicial complex at that dimension. Figure below illustrates the subspace relation between the cycles and boundaries of the corresponding $n$-chains in a chain complex. It then follows form the rank-nullity theorem [2] that the dimension of an $n$-chain $C_n(K)$ can be decomposed into the dimension of the cycle space $Z_n(K)$ and the dimension of the boundary space $B_{n-1}(K)$:

$$K_n = \dim C_n(K) = \dim Z_n(K) + \dim B_{n-1}(K). \tag{3.2.1}$$

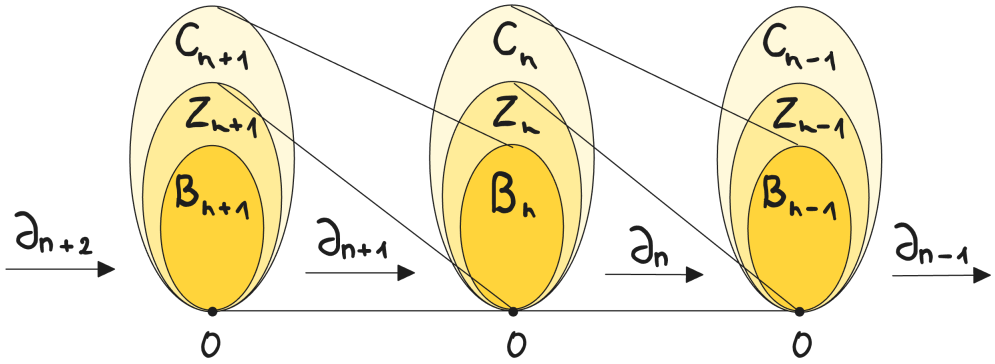



Figure 3.2.: A chain complex consisting of $n$-chains, cycles and boundaries.

### 3.2.5. Betti Numbers

We can now define Betti numbers as follows.

**Definition 3.2.4** (Betti Number)**.** The **$n$-th Betti number** of a simplicial complex $K$ is $\beta_n(K) = \dim Z_n(K) - \dim B_n(K)$.

Compared to the Euler characteristic, Betti numbers are a sequence of numbers that provide more detailed information about the topological features of a space at different dimensions. Each Betti number counts the number of $n$-dimensional holes in a space where, in general,

- $\beta_0(K)$: number of connected components.
- $\beta_1(K)$: number of loops.

- $\beta_2(K)$: number of voids.

The connection between the Euler characteristic and Betti numbers is captured in the Euler-Poincare formula.

**Theorem 3.2.1** (Euler-Poincare Formula)**.** *For a simplicial complex $K$,*

$$\chi(K) = \sum_{i=0}^{\infty}(-1)^i K_i = \sum_{i=0}^{\infty}(-1)^i \beta_i(K).$$

*Proof.* As we saw from equation 3.2.1,

$$K_i = \dim C_i(K) = \dim Z_i(K) + \dim B_{i-1}(K).$$

By definition of the Euler characteristic it the follows that:

$$\begin{aligned}
\chi(K) = \sum_{i=0}^{\infty}(-1)^i K_i &= \sum_{i=0}^{\infty}(-1)^i \Big(\dim Z_i(K) + \dim B_{i-1}(K)\Big) \\
&= \sum_{i=0}^{\infty}(-1)^i \dim Z_i(K) + \sum_{i=0}^{\infty}(-1)^i \dim B_{i-1}(K) \\
&= \sum_{i=0}^{\infty}(-1)^i \dim Z_i(K) + \sum_{i=-1}^{\infty}(-1)^{i+1} \dim B_i(K) \\
&= \sum_{i=0}^{\infty}(-1)^i \dim Z_i(K) + \sum_{i=0}^{\infty}(-1)^{i+1} \dim B_i(K) \\
&= \sum_{i=0}^{\infty}(-1)^i \dim Z_i(K) - \sum_{i=0}^{\infty}(-1)^i \dim B_i(K) \\
&= \sum_{i=0}^{\infty}(-1)^i \Big(\dim Z_i(K) - \dim B_i(K)\Big) \\
&= \sum_{i=0}^{\infty}(-1)^i \beta_i(K).
\end{aligned}$$

Note that there are no $-1$-dimensional simplices, hence $B_{-1}(K) = 0$. ■

## 3.2.6. Simplicial Homology

Not all chain complexes are exact. To understand how an arbitrary chain complex deviates from exactness, we use homology, which quantifies these deviations by measuring the presence of cycles that are not boundaries. This process helps in identifying the underlying topological features of the space, such as holes and voids, providing insight into

its structure.

**Definition 3.2.5** ($n$-th Simplicial Homology)**.** The **$n$-th simplicial homology vector space** of a simplicial complex $K$ is the quotient space

$$\mathcal{H}_n(K) = Z_n/B_n = \operatorname{Ker} \partial_n / \operatorname{Im} \partial_{n+1}.$$

It then follows that $\beta_n(K) = \dim \mathcal{H}_n(K)$. Analogously to Betti numbers then, the dimension of $\mathcal{H}_n(K)$ can be interpreted as the number of $n$-dimensional holes in the simplicial complex $K$.

**Example 3.2.2: Computing Simplicial Homology**

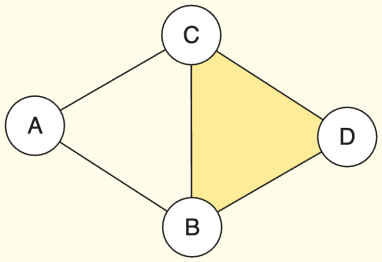
Consider an abstract simplicial complex:

$$X = \{\{A\}, \{B\}, \{C\}, \{D\}, \{AB\}, \{AC\}, \{BC\}, \{BD\}, \{CD\}, \{BCD\}\}$$

depicted below. The associated chain complex is of the form:

$$0 \xrightarrow{\partial_3} \mathbb{R} \xrightarrow{\partial_2} \mathbb{R}^5 \xrightarrow{\partial_1} \mathbb{R}^4 \xrightarrow{\partial_0} 0.$$

Since $\{ABC\}$ is not a 2-simplex, the 1-cycle $\{BC\} - \{AC\} + \{AB\}$ is not the boundary of a higher dimensional simplex, so it is a hole. Hence, $\dim \mathcal{H}_1 = 1$.




## 3.3. Persistent Homology

Recall that transforming a point cloud data set into a Rips complex necessitates selecting a parameter $\varepsilon$. But is there an optimal $\varepsilon$ that best captures the topology of the data set? While the homology of a simplicial at a particular value of $\varepsilon$ can be computed and provides valuable insights, it is not sufficient to determine the optimal value of $\varepsilon$. Simply counting the number and types of holes that appear at each parameter value $\varepsilon$ is also inadequate. Indeed, the more interesting object to consider is the family of Rips complexes, which allows us to make comparisons between different complexes within the same family and draw meaningful conclusions.

### 3.3.1. Chain Maps

For that, we need to discuss not only the properties of a simplicial complex but also the properties of maps between them.

**Definition 3.3.1** (Chain Map)**.** A **chain map** $f_\bullet : C_\bullet(X) \to C_\bullet(Y)$ is a sequence of linear transformations between two chain complexes that commute with their respective boundary operators, such that $f_n \circ \partial_{n+1} = \partial'_{n+1} \circ f_{n+1}$.

Chain maps form a commutative diagram:

$$\begin{array}{ccccccccc} \cdots & \xrightarrow{\partial} & C_{n+1}(X) & \xrightarrow{\partial} & C_n(X) & \xrightarrow{\partial} & C_{n-1}(X) & \xrightarrow{\partial} & \cdots \\ & & \downarrow{\scriptstyle f_\bullet} & & \downarrow{\scriptstyle f_\bullet} & & \downarrow{\scriptstyle f_\bullet} & & \\ \cdots & \xrightarrow{\partial'} & C_{n+1}(Y) & \xrightarrow{\partial'} & C_n(Y) & \xrightarrow{\partial'} & C_{n-1}(Y) & \xrightarrow{\partial'} & \cdots \end{array}$$

and are important in preserving the algebraic structure as well as allowing comparisons between different topological spaces.

In the case of simplicial complexes from the same family, chain maps are used to link the chain complexes of the subcomplexes in a filtration, enabling the study of how topological features evolve as the filtration parameter changes.

**Definition 3.3.2** (Filtration)**.** A **filtration** of a simplicial complex $K$ is a nested sequence of subcomplexes $\{K_i \subseteq K\}_{i=0}^m$ such that

$$K_0 \subseteq K_1 \subseteq \cdots \subseteq K_m = K.$$

*Remark.* The length of the filtration is determined by the number of terms in it.

The key idea of persistent homology is then to track the changes in homologies over *time*. In this setting, homologies that persist over a substantial parameter range are regarded as signal, while short-lived features are considered noise. This implies that one can examine not only the homology of an individual complex but also the homology of a map between complexes.

**Definition 3.3.3** (Persistent Homology)**.** Given a filtration of $K$, the $(i, j)$**-persistent homology** of $K$, denoted $\mathcal{H}_n^{i \to j}(K)$, is the image of the induced map $f_{\mathcal{H}}^{i,j} : \mathcal{H}_n(K_i) \to \mathcal{H}_n(K_j)$ for all $i \leq j$.

The associated $n$-th persistent Betti numbers are then given by $\beta_n^{i,j} = \dim \operatorname{Im} f_{\mathcal{H}}^{i,j}$.

**Example 3.3.1: The Rips Filtration**

Assume that $(\mathcal{R}_i)_0^m$ is a filtered sequence of Rips complexes associated to a discrete point cloud for an increasing scale $(\varepsilon_i)_0^m$. Since Rips complexes increase with $\varepsilon$, the chain maps $f_\bullet$ are naturally identified with inclusions

$$\mathcal{R}_0 \xhookrightarrow{f_{\mathcal{H}}} \mathcal{R}_1 \xhookrightarrow{f_{\mathcal{H}}} \cdots \xhookrightarrow{f_{\mathcal{H}}} \mathcal{R}_{m-1} \xhookrightarrow{f_{\mathcal{H}}} \mathcal{R}_m$$

where $f_{\mathcal{H}}^{i,j} : \mathcal{H}_n(\mathcal{R}_i) \to \mathcal{H}_n(\mathcal{R}_j)$ are the induced maps in homology for all $i \leq j$. In other words, instead of examining the homology of the individual terms $\mathcal{R}_i$, one focuses on the homology of the iterated inclusions $f_{\mathcal{H}}$.

### 3.3.2. Persistence Barcodes and Diagrams

In order to visually represent the changes in topological features of a filtration across different scales, we introduce two similar methods: persistence barcodes and persistence diagrams.

Given a filtration $\{K_i\}$ of a simplicial complex $K$, the **$k$-th persistence barcode** is a collection of intervals $[b, d)$, where each interval represents a $k$-dimensional homological feature that appears at filtration step $b$ (birth) and disappears at step $d$ (death). The length of each bar indicates the persistence of the feature across the filtration - the shorter the barcode the more noisier the feature and vice versa. [11]

In a **persistence diagram**, however, each homological feature is plotted as a point in the plane, where the $x$-coordinate represents the birth time and the $y$-coordinate represents the death time of the feature. Points close to the diagonal represent short-lived features, while points far from the diagonal represent long-lived, significant features.

# 4. Analysis

In previous sections, we have seen how different crystal systems are classified based on their symmetries. We have also described a method to analyse a point cloud by first constructing a Rips complex and then using the tools of persistent homology. Let us now combine our knowledge to study point clouds from actual crystals and their symmetries. The question we aim to answer is: can persistent homology distinguish between two different space groups? Recall that every space group corresponds to one of the seven crystal systems. Thus, distinguishing space groups ultimately allows us to differentiate between crystal systems.

## 4.1. Constructing Crystals

The information about the crystal can be retrieved from the CIF file. The Crystallographic Information File (CIF) format is a standardised text file format used for storing and exchanging crystallographic data. It contains detailed descriptions of crystal structures, including unit cell parameters, symmetry information and atomic coordinates. Therefore, reading a CIF file is the first step in our pipeline of analysing the crystals. For an example of a CIF file see the Appendix A.

### 4.1.1. Fractional Coordinate System

The atomic coordinates within a unit cell of a crystal are usually expressed as fractional coordinates. Instead of using absolute distances, fractional coordinates express atomic positions as fractions of the normalised unit cell dimensions. In this coordinate system, the basis vectors are chosen to be primitive translation vectors. The fractional coordinates of a point in space $\rho = (\rho_{x_1}, \rho_{x_1}, \ldots, \rho_{x_d})$ in terms of the primitive translation vectors can then be defined as

$$\rho = \rho_{x_1}\vec{a}_1 + \rho_{x_2}\vec{a}_2 + \cdots + \rho_{x_d}\vec{a}_d.$$

*Remark.* Observe that the fractional coordinates are in the range $[0, 1)$.

By reading the initial fractional coordinates from the CIF file, we can proceed by applying symmetries to them.

### 4.1.2. Symmetry Application

Symmetry application involves generating all equivalent positions of the atoms within the unit cell by applying the defined symmetry operations of the space group to the initial atomic coordinates using the Seitz matrix.
A Seitz matrix $S$ is represented as

$$\mathbf{S} = \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix}$$

where

- $\mathbf{R}$ is a $3 \times 3$ rotation matrix.
- $\mathbf{t}$ is a $3 \times 1$ vector representing a translation.
- $\mathbf{0}$ is a $1 \times 3$ vector of zeros.

To find the new coordinate $\mathbf{x}'$ after applying the Seitz matrix, multiply $\mathbf{S}$ by $\mathbf{x}_h$ [1]:

$$\mathbf{x}'_h = \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}$$

Finally, the new fractional coordinate $\mathbf{x}'$, ignoring the homogeneous component, is:

$$\mathbf{x}' = \begin{pmatrix} R_{11}x + R_{12}y + R_{13}z + t_x \\ R_{21}x + R_{22}y + R_{23}z + t_y \\ R_{31}x + R_{32}y + R_{33}z + t_z \end{pmatrix}$$

*Remark.* Points outside of the unit cell can be wrapped to $[0, 1)$ through standardisation:

$$x_{\text{wrap}} \equiv ((x \mod 1) + 1) \mod 1. \tag{4.1.1}$$

This expanded set of fractional coordinates can already be used for comparing the crystal structures. Note, however, that its unit cell is normalised with sides of length 1, effectively shrinking the atoms into a cubic shape. Consequently, we would not be able, for example,

[1] To apply the Seitz matrix, we extend the fractional coordinate into a 4-dimensional vector to incorporate the homogeneous coordinate component.

to distinguish a tetragonal system from a cubic system using fractional coordinates alone. Therefore, we proceed to orthogonalise these coordinates into a Cartesian coordinate system.

### 4.1.3. Transformation to Cartesian Coordinate System

The relationship between fractional and Cartesian coordinates can be described by the matrix transformation $\mathbf{r} = \mathbf{A}\rho$:

$$\begin{pmatrix} r_x \\ r_y \\ r_z \end{pmatrix} = \begin{pmatrix} a\sin(\beta)\sqrt{1-(\cot(\alpha)\cot(\beta)-\csc(\alpha)\csc(\beta)\cos(\gamma))^2} & 0 & 0 \\ a\csc(\alpha)\cos(\gamma)-a\cot(\alpha)\cos(\beta) & b\sin(\alpha) & 0 \\ a\cos(\beta) & b\cos(\alpha) & c \end{pmatrix} \begin{pmatrix} \rho_x \\ \rho_y \\ \rho_z \end{pmatrix}$$

giving us a set of positional coordinates $\vec{r} = (r_x, r_y, r_z)$, which can be used to build a simplicial complex from. We can visualize these coordinates to reveal the symmetric patterns of the crystals.

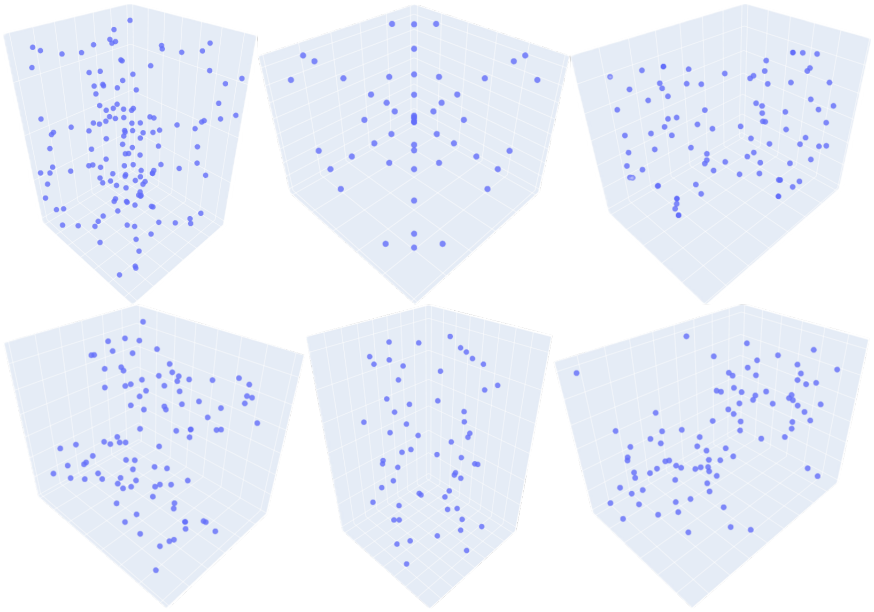


Figure 4.1.: Sample of 6 crystals plotted from positional coordinates.

### 4.1.4. Unit Cell Normalisation

At this point, however, we have an issue that the sides of the unit cell might differ. Therefore, we remove the effects of varying unit cell sizes by scaling them to a common reference frame.

The volume of the unit cell can be calculated by the formula

$$V = a \cdot b \cdot c \cdot \sqrt{1 + 2\cos(\alpha)\cos(\beta)\cos(\gamma) - \cos^2(\alpha) - \cos^2(\beta) - \cos^2(\gamma)},$$

where $a, b, c, \alpha, \beta, \gamma$ are the lattice constants. We can then use it to calculate the normalising constant:

$$k = \left(\frac{1}{V}\right)^{1/3}. \tag{4.1.2}$$

And after multiplying the positional coordinates in our unit cell by this quantity we get the normalised positional coordinates.

*Remark.* Note that normalising positional coordinates in a unit cell does not affect the calculation of persistent homology, as it is invariant under geometric transformations such as scaling.

### 4.1.5. Periodic Boundary Conditions

Another issue we need to address is that the unit cell is currently treated as a finite system, which can include surface and edge effects. However, we want to ensure that atoms at one edge of the unit cell have equivalent atoms at the opposite edge, creating a seamless transition and accurately reflecting the infinite periodic nature of the crystal lattice. This can be achieved by applying periodic boundary conditions.

Recall that the distance between two positional coordinates $r_i$ and $r_j$ in an Euclidean space can be defined by:

$$d_{r_i r_j} = \sqrt{(d_{x_i x_j})^2 + (d_{y_i y_j})^2 + (d_{z_i z_j})^2} \tag{4.1.3}$$

where coordinate-wise distances are given by $d_{x_i x_j} = |x_i - x_j|$ for $x$-coordinate, for instance. Now if $d_{x_i x_j} > \frac{1}{2} \cdot a$, we replace $d_{x_i x_j}$ with a new distance $d_{x_i x_j} = a - d_{x_i x_j}$ in the equation 4.1.3. With a similar reasoning, we replace $d_{y_i y_j}$ with $d_{y_i y_j} = b - d_{y_i y_j}$ if $d_{y_i y_j} > \frac{1}{2} \cdot b$, and $d_{z_i z_j}$ with $d_{z_i z_j} = c - d_{z_i z_j}$ if $d_{z_i z_j} > \frac{1}{2} \cdot c$, respectively. In the end, we get an $n \times n$ distance matrix, where $n$ is the number of positional coordinates in the unit cell.

The whole pipeline of generating a crystal and preparing it for the subsequent analysis can be summarised in the figure below.

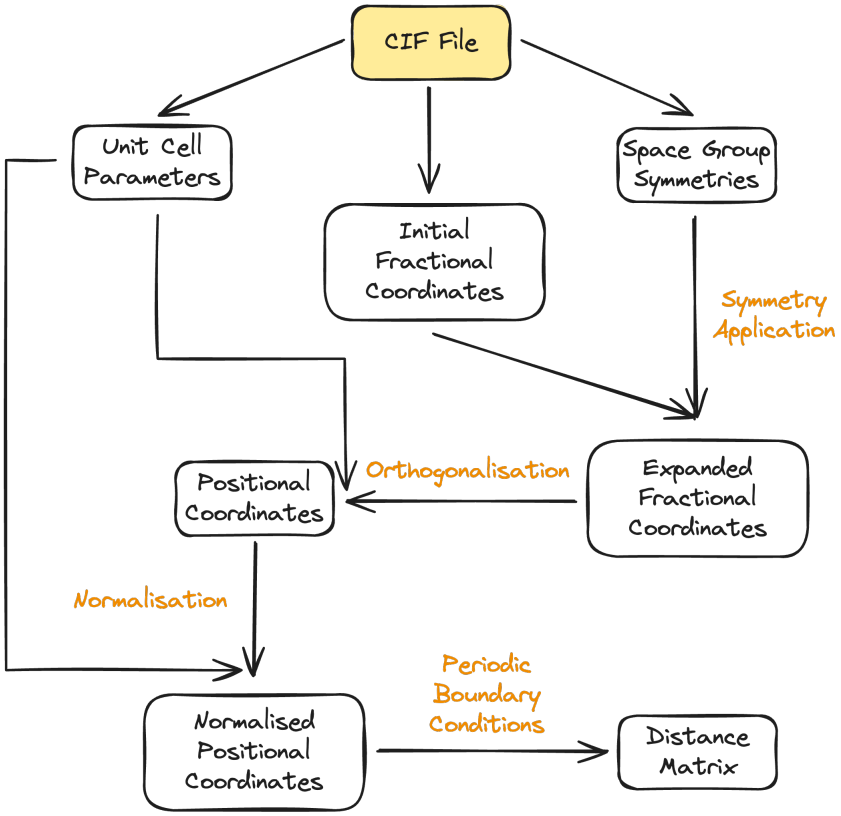



Figure 4.2.: Pipeline for constructing a crystal.

## 4.2. Building Rips Complexes

We can build the Rips complex directly from the (normalised) positional coordinates or using the distance matrix we calculated after applying the periodic boundary conditions. The figure below illustrates the construction of a sequence of Rips complex from the normalised positional coordinates for one of the crystals.

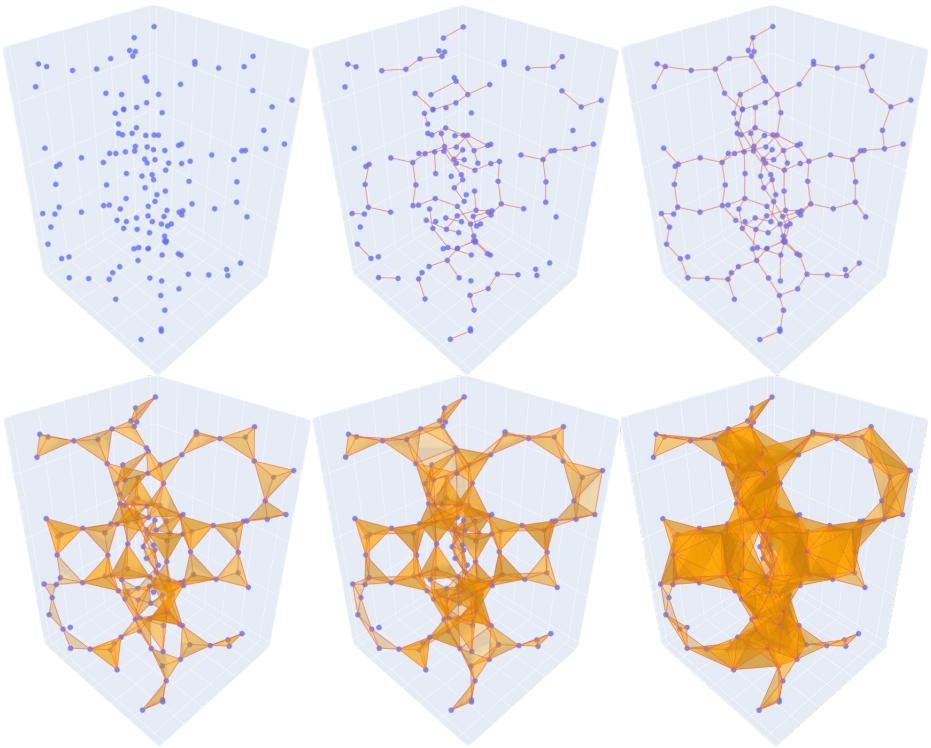


Figure 4.3.: Build-up of a Rips complex from positional coordinates for AEI crystal.

## 4.3. Random Crystals

Unfortunately, we have a limited amount of crystallographic data available. As a result, for certain crystal systems, such as triclinic, we have an insufficient sample size to conduct a thorough analysis.

The solution involves constructing a crystal from randomly generated fractional coordinates and applying the same procedure described in Figure 4.2. The obvious question is: how many random fractional coordinates should be generated for each space group?

### 4.3.1. Group Order

To answer this question, we need to know how many equivalent points are generated from a single point by the symmetry operations of a space group. This number is determined by the group order - count of distinct symmetry operations that map the space onto itself.

**Example 4.3.1: Space Group 229**

The space group 229 is of order 96, meaning that a single random point is mapped to 96 distinct equivalent points through the group's symmetry operations.

As can be imagined, space groups may have different group orders. However, it is essential that the number of randomly generated fractional coordinates be the same across all space groups we compare. One obvious reason is that we want to observe the differences in the homologies of the symmetries, rather than differences in the number of points. Hence, to make sure that the number of points is the same between two or more space groups, we can take the least common multiple of group orders and divide it by the group order of each space group. Below is an example of two groups of Rips complexes generated from random fractional coordinates.

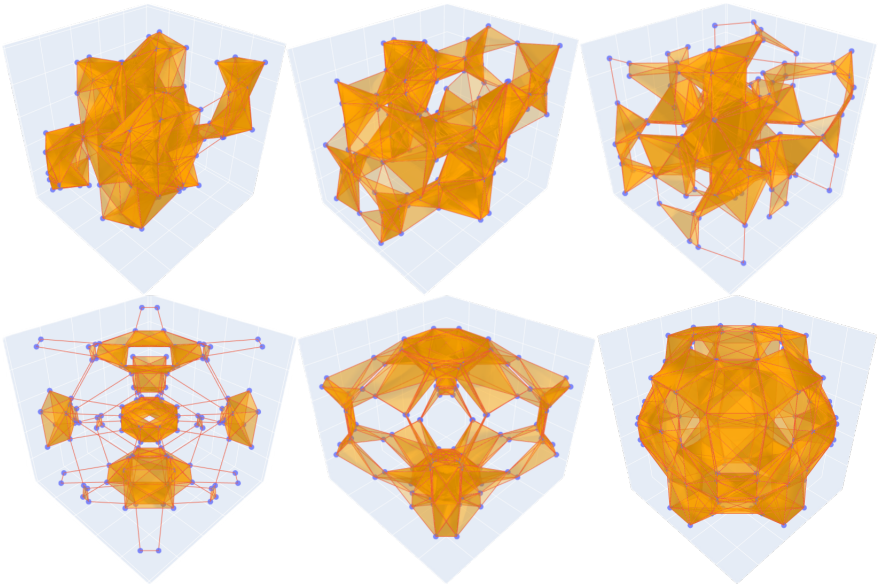


Figure 4.4.: Rips complexes constructed from randomly generated point clouds. Top row - using symmetries from space group 63. Bottom row - using symmetries from space group 139.

## 4.4. Classifying Space Groups

Now that we know how to generate random point clouds from different space group symmetries and create Rips complexes from them, we can proceed with the analysis. Our goal is to classify two space groups corresponding to two different crystal systems. Since a single crystal system may correspond to multiple space groups, we randomly choose space groups that correspond to each particular system. We generate a sample of 50 crystals for each group with 100 in total. For each of those crystals we compute persistent homology which is visualised in the form of persistence diagrams as can be seen in figures 4.5 and 4.6.

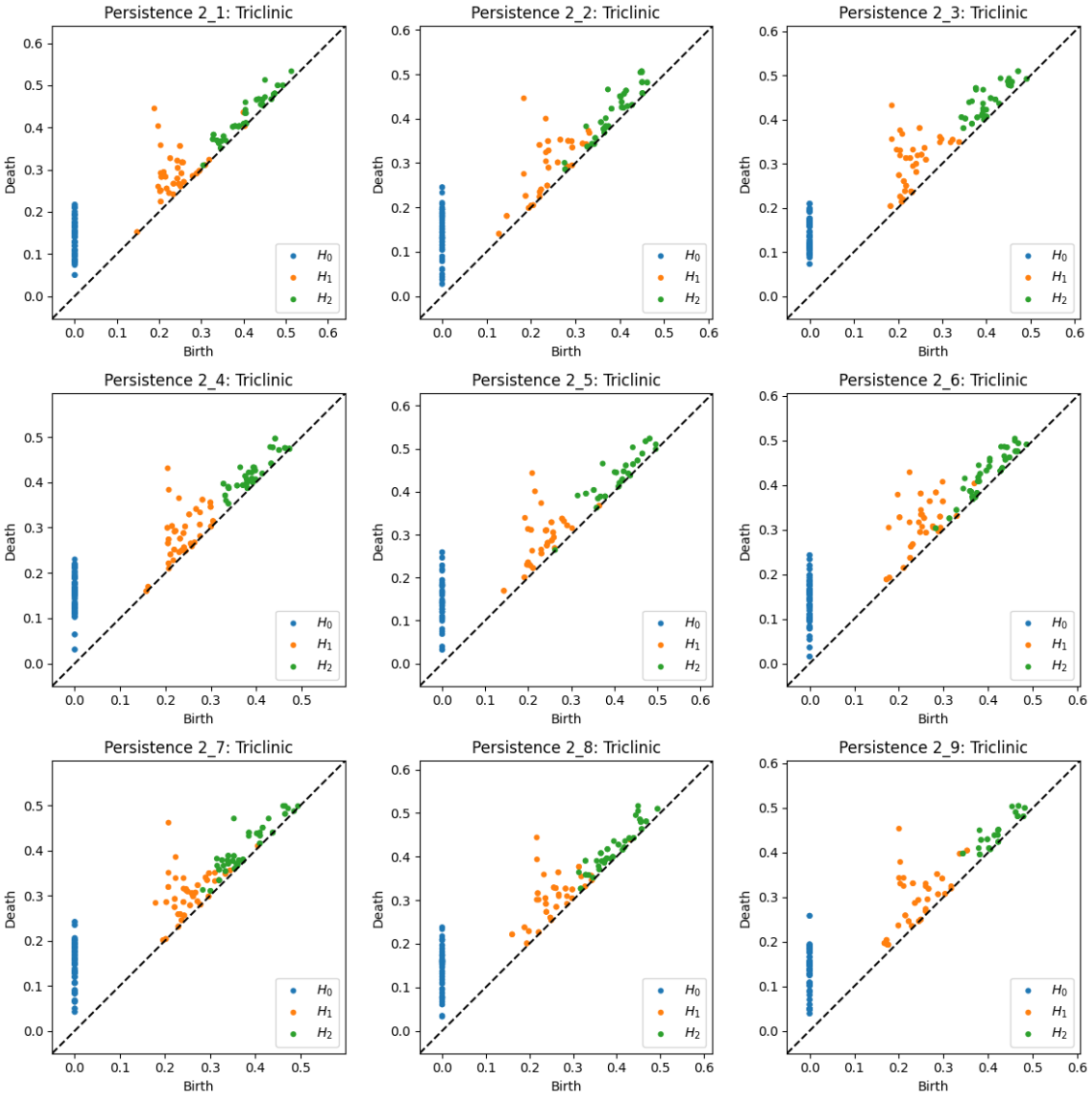



Figure 4.5.: The first 9 persistence diagrams for a space group 2 corresponding to the triclinic crystal system.

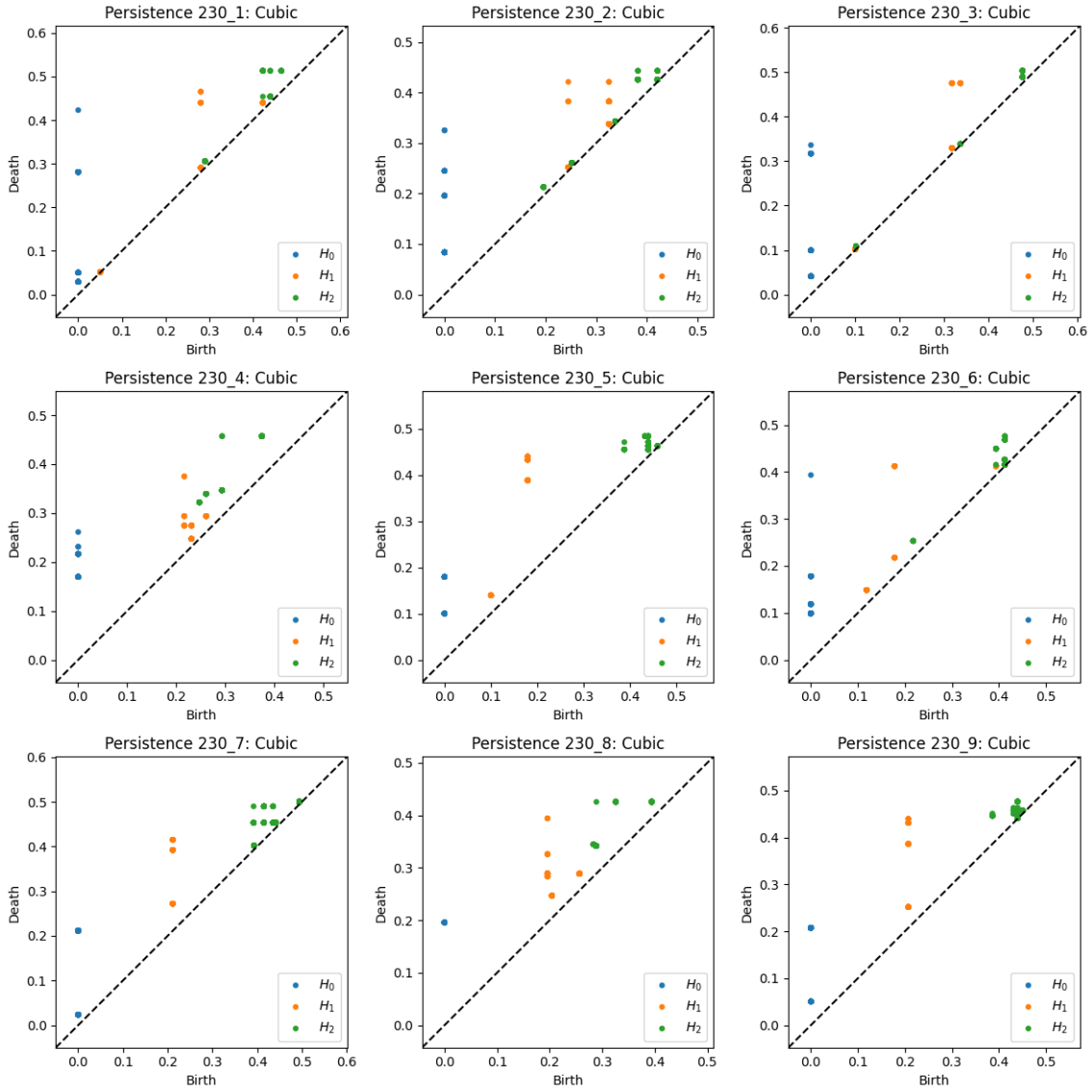



Figure 4.6.: The first 9 persistence diagrams for a space group 230 corresponding to the cubic crystal system.

Even purely visually we can see that the two groups have completely different homologies which can potentially be distinguished. In order to compare two persistence diagrams we utilise the Wasserstein distance, which measures the similarity between two diagrams by calculating the minimal value achieved by a perfect matching of the two diagrams using the sum of all edges lengths. Wasserstein distance is calculated for each dimension, so as an output we get three $100 \times 100$ matrices corresponding to distances in dimensions 0, 1 and 2. We then take an element-wise maximum between the three of them to get a single distance matrix. In order to make the distance matrix usable for traditional machine learning tools, we transform it into a lower-dimensional space while preserving the pairwise distances between points using the Multidimensional scaling [5]. In the analysis,

a 5-dimensional embedding was used. In the plot below, the first two dimensions are visualised.

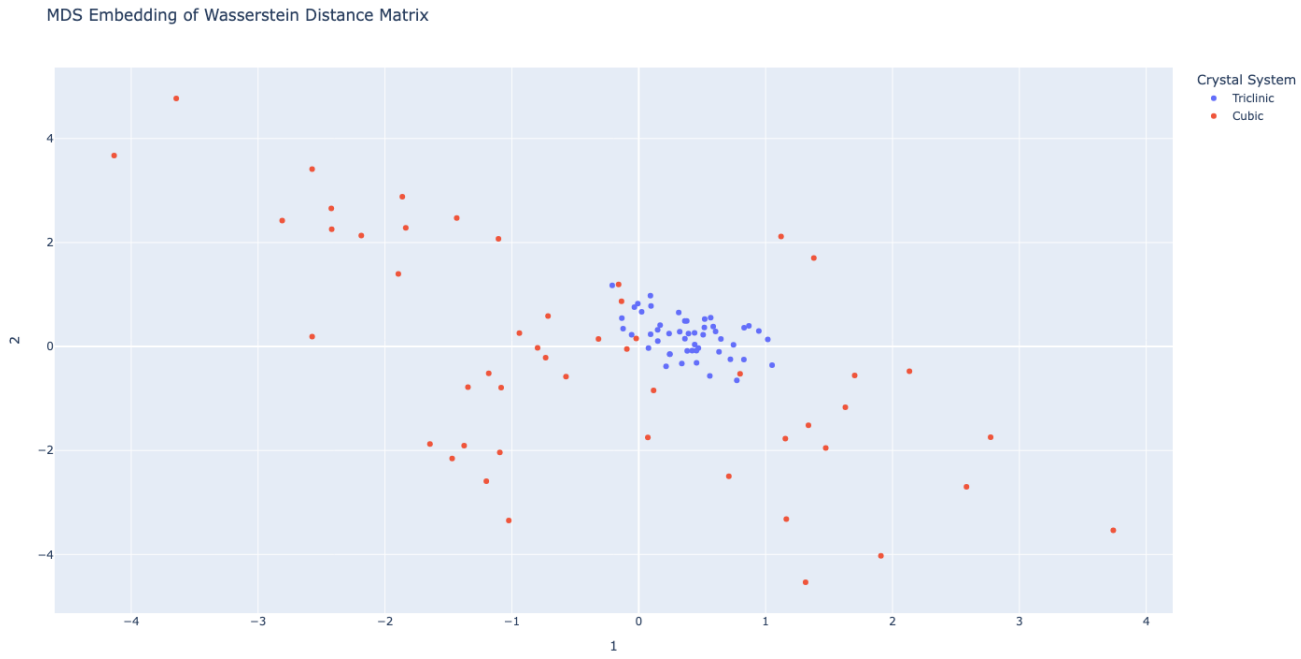



Figure 4.7.: Embedding of a Wassertstein distance into $\mathbb{R}^2$ using Multidimensional scaling.

As for the classification algorithm itself, a Support Vector Machine classifier was used due to its robustness to overfitting [8]. The algorithm was able to categorise the test samples with an accuracy of 100% in this particular example as can be seen from the classification report below.

```
Accuracy: 1.00
Classification Report:
              precision    recall  f1-score   support

   Triclinic       1.00      1.00      1.00        13
       Cubic       1.00      1.00      1.00        17

    accuracy                           1.00        30
   macro avg       1.00      1.00      1.00        30
weighted avg       1.00      1.00      1.00        30
```

A similar procedure was performed for all the other crystal systems with results summarised in a matrix:

| Triclinic | Monoclinic | Orthorhombic | Tetragonal | Trigonal | Hexagonal | Cubic | |
|---|---|---|---|---|---|---|---|
| $*$ | 0.93 | 0.97 | 0.90 | 1.0 | 1.0 | 1.0 | Triclinic |
| 0.93 | $*$ | 0.90 | 0.93 | 0.97 | 0.97 | 1.0 | Monoclinic |
| 0.97 | 0.90 | $*$ | 0.87 | 1.0 | 0.97 | 0.93 | Orthorhombic |
| 0.90 | 0.93 | 0.87 | $*$ | 0.93 | 0.83 | 0.97 | Tetragonal |
| 1.0 | 0.97 | 1.0 | 0.93 | $*$ | 0.90 | 1.0 | Trigonal |
| 1.0 | 0.97 | 0.97 | 0.83 | 0.90 | $*$ | 0.87 | Hexagonal |
| 1.0 | 1.0 | 0.93 | 0.97 | 1.0 | 0.87 | $*$ | Cubic |

All calculations as well as the results can be verified in the repository. We see that in all of the examples, we were able to distinguish pairs of crystal systems with a sufficiently high accuracy, indicating that homological features do have a signal when comparing two symmetries.

# 5. Conclusion

This project has demonstrated the efficacy of employing persistent homology, a powerful tool within topological data analysis, to distinguish between crystal symmetries. By leveraging homological features from multiple dimensions, we have solidified previous results which go beyond traditional geometric and algebraic methods. Through the construction of Rips complexes from crystallographic data and the computation of persistence diagrams, we identified unique topological signatures that are indicative of different crystal systems. Our results consistently showed that persistent homology could effectively capture and utilise the intricate topological characteristics inherent in crystal structures. After all, the ability to identify and classify crystal systems with high accuracy has profound implications for material science, particularly in the automated discovery and categorisation of novel crystal structures. Future work could expand on this foundation by exploring other methods of analysing persistence diagrams, including vectorising through persistence images [1] or clustering techniques such as K-Means.

# Bibliography


[1] Henry Adams. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017.

[2] Sheldon Axler. *Linear Algebra Done Right*. Springer, 4 edition, 2024.

[3] Ulrich Bauer. Ripser: efficient computation of vietoris-rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3):391–423, 2021.

[4] Walter Borchardt-Ott. *Crystallography*. Springer-Verlag Berlin, 3 edition, 2012.

[5] Ingwer Borg and Patrick Groenen. *Modern Multidimensional Scaling*. Springer Series in Statistics, 1 edition, 1997.

[6] Magnus Bakke Botnan. *Lecture Notes on Topological Data Analysis*. 2022.

[7] Gunnar Carlsson. *Topological Data Analysis with Applications*. Cambridge University Press, 2021.

[8] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2013.

[9] B. D. Cullity. *Elements of X-ray Diffraction*. Prentice Hall, 3 edition, 2001.

[10] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. Departments of Computer Science and Mathematics Duke University, 2008.

[11] Robert Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[12] Robert Ghrist. *Elementary Applied Topology*. Createspace, 2014.

[13] Maureen Julian. *Foundations of Crystallography with Computer Applications*. CRC Press, 2 edition, 2015.

[14] Charles Kittel. *Introduction to Solid State Physics*. Wiley, 8 edition, 2004.

[15] Christopher Tralie, Nathaniel Saul, and Rann Bar-On. Ripser.py: A lean persistent homology library for python. *The Journal of Open Source Software*, 3(29):925, 2018.

# A. CIF File

```
data_AEI

_cell_length_a                  13.6770(0)
_cell_length_b                  12.6070(0)
_cell_length_c                  18.4970(0)
_cell_angle_alpha               90.0000(0)
_cell_angle_beta                90.0000(0)
_cell_angle_gamma               90.0000(0)

_symmetry_space_group_name_H-M     'C m c m'
_symmetry_Int_Tables_number         63
_symmetry_cell_setting             orthorhombic

loop_
_symmetry_equiv_pos_as_xyz
'+x,+y,+z'
'1/2+x,1/2+y,+z'
'-x,+y,+z'
'1/2-x,1/2+y,+z'
'+x,-y,1/2+z'
'1/2+x,1/2-y,1/2+z'
'-x,-y,1/2+z'
'1/2-x,1/2-y,1/2+z'
'-x,-y,-z'
'1/2-x,1/2-y,-z'
'+x,-y,-z'
'1/2+x,1/2-y,-z'
'-x,+y,1/2-z'
'1/2-x,1/2+y,1/2-z'
'+x,+y,1/2-z'
'1/2+x,1/2+y,1/2-z'

loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
    O1    O     0.0000    0.0004    0.1614
    O2    O     0.1449    0.0438    0.2500
    O3    O     0.1252    0.1515    0.1293
    O4    O     0.1804    0.9517    0.1253
    O5    O     0.1465    0.8322    0.0116
    O6    O     0.0000    0.7364    0.9468
    O7    O     0.1794    0.6672    0.9281
    O8    O     0.7404    0.0000    0.0000
    T1    Si    0.1126    0.0369    0.1664
    T2    Si    0.1128    0.7711    0.9394
    T3    Si    0.7733    0.9042    0.0521
```