University of Southampton

Faculty of Chemistry
School of Southampton

The Topological and Geometric Analysis of Organic Crystal Systems

by

Jack Robert Doyle

ORCiD: 0000-0002-6534-7318

A thesis for the degree of
Doctor of Philosophy

20th April 2024

University of Southampton

Abstract

Faculty of Chemistry
School of Southampton

Doctor of Philosophy

The Topological and Geometric Analysis of Organic Crystal Systems

by Jack Robert Doyle

The aim of this study was to use topological descriptors to gain insight into the crystal packing of organic compounds and generate crystal structure landscapes that are representative of the packing motif that might be identified by a crystallographer. These descriptors are applied to both sets of experimental compounds, as might be found in the Cambridge Structural Database, for example, or to the large sets of compounds that might be generated as the output of crystal structure prediction calculations.

The crystal structures of fluorinated benzylideneanilines, polyaromatic hydrocarbons, azapentacenes and the nictotinamide:benzoic acid co-crystal were studied through the lens of a novel topological descriptor. This descriptor is constructed from the persistent homology of a set of molecular centroids and orientation vectors extracted from the crystal structure, the homology being computed on a six dimensional space. We were able to generate crystal structure landscapes that completely separated all known packing classes of fluorinated benzylideneaniline as identified by a subject matter expert. We were also able to completely separate the structures of two classes of nictotinamide:benzoic acid co-crystals that were identified to belong to two funnels on the potential energy landscape corresponding to its known polymorphs. While the azapentacens and polyaromatic hydrocarbons proved more resistant to a full description with persistent homology, we were able to produce landscapes that preserve some trends which are consistent with their canonical packing motifs. We also showcase how crystal structure landscapes can be constructed using supervised dimensionality reduction in the context of some existing high fidelity data with known packing motifs in order to obtain landscape that extenuate these chemically relevant features.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signed:.. Date:..................

Chapter 1

Introduction

The ubiquity of easily accessible data has secured the place of data science and machine learning at the forefront of science such that data science is rapidly becoming ”the fourth paradigm” of modern science Hey et al. (2009). The applications of these data intensive techniques in chemistry are extensive in both depth and breadth Bartók et al. (2024); Jung et al. (2020); Tetko and Engkvist (2020). These applications extend into the world of materials and crystallography Damewood et al. (2023); Li et al. (2023a); Isayev et al. (2015) not least because of the massive amount crystallographic data available to us via the Cambridge Structural Database Taylor and Wood (2019) and the output of crystal structure prediction algorithms Day (2011); Bowskill et al. (2021).

In order to make use of and understand the vast amounts of data pertaining to the solid state that is available to us, it is imperative that we have stable representations of these systems. These representations are called descriptors Damewood et al. (2023); Jablonka et al. (2020); Musil et al. (2021). In order for these descriptors to be useful in real world applications these descriptors must be invariant to translation and rotation of the whole system, permutation of the atoms within the system and expansion of the system to its supercell if defined on a smaller unit Jablonka et al. (2020). There are other useful criteria that these descriptors should meet such as uniqueness, continuity and ease of calculation, see the three reviews Damewood et al. (2023); Jablonka et al. (2020); Musil et al. (2021) for more details. These descriptors can also in general be subdivided into two categories: there are *local* descriptors that pertain to the local environment of a given atom - these are useful for building models of potential energy surfaces for example Behler (2011); Anstine and Isayev (2023); Brezina et al. (2023); Lee et al. (2021); then there are *global* descriptors which describe the entire structure at once - these are useful for directly comparing different systems and making predictions Musil et al. (2018); Jiang et al. (2021); Li et al. (2023a); Bartók et al. (2013). In this work we mostly focus on the latter class of descriptors.

These descriptors have two key applications in chemical informatics. Firstly they can be used to predict properties such as energy or electronic conductivity or mechanical properties Cao et al. (2019); Jiang et al. (2021); Lee et al. (2021). Secondly they can be used qualitatively in order to aid our understanding of larger regions of chemical space. This is the application that we are focused on in this work.

There are two main types of chemical data that we focus on for which a descriptor can help qualitatively understand the wider dataset. There are smaller sets of *experimental* crystal structures and there are larger sets of *predicted* crystal structures.

Describing the packing motif of a crystal structure is a qualitative process and is often done heuristically by subject matter experts, often by eye Dodd (2020); Desiraju and Gavezzotti (1989b); Dance (2003); Taylor and Macrae (2001); Florence et al. (2006). There already exist computational techniques such as COMPACK Chisholm and Motherwell (2005) and XPac Grossel et al. (2005) which, while they provide valuable information about crystal similarity and common motifs between different structures, they do not attempt to *predict* what the packing of a given crystal structure is and they are not necessarily amenable to generating plots that visualise the set of crystal structures in a way that naturally respects the packing motif that might be assigned by a human and are amenable for subsequent analysis with the traditional tool kit of statistical learning. Crystal descriptors provide a means of producing more effective visualisations of the dataset that can aid in the understanding of the physical and chemical properties of sets of related compounds. They also offer the attractive proposition of automatically labelling large sets of unknown crystal structures without human input.

The process of crystal structure prediction often generates a large (on the order of 1000s) set of potential crystal structures for a given chemical compound Bowskill et al. (2021); Day (2011). These are often visualised as a crystal structure landscape Ceriotti et al. (2011); Desiraju (2017). The most common form of crystal structure landscape takes the form of a plot of energy vs. density. This is convenient computationally but does not provide much useful chemical information and nor does it provide information about the packing motifs of the different predicted compounds. Chemical descriptors can be used to generate a plot (often by reducing the dimension of the resultant descriptor space) such that crystal structures with similar properties are close together and crystal structures which are very different are far away. This can help gain rapid insight into the nature of the crystal structure landscape and possibly aid future property prediction. Crystal structure landscapes also have significant applications in crystal engineering Aakeröy (1997) and the prediction and categorisation of polymorphic structures Price (2008, 2009); Ismail et al. (2013). This process is outlined in figure 1.1 which we have reproduced from the work of Musil *et al.* Musil et al. (2018).

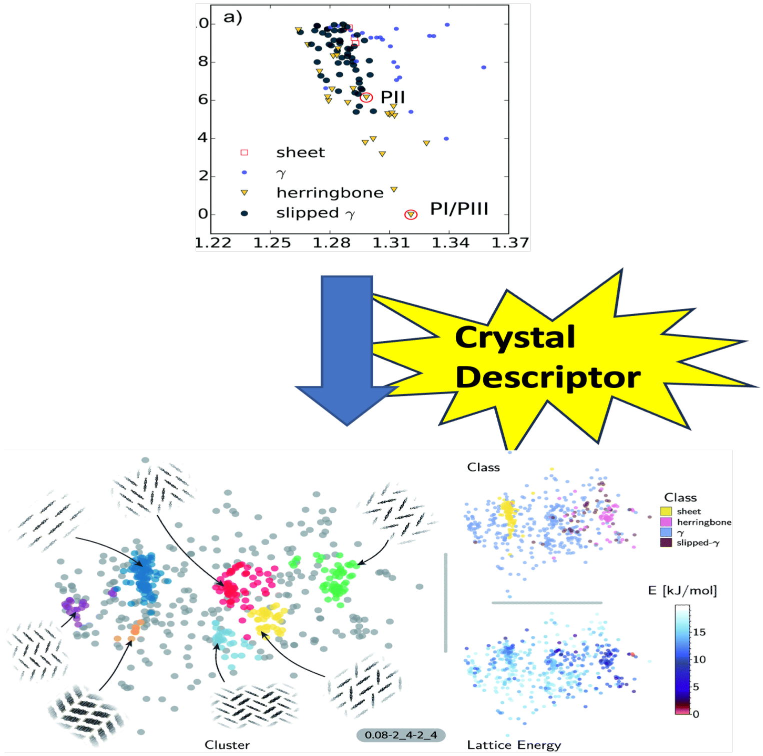


FIGURE 1.1: Graphic demonstrating how a crystal structure descriptor can be used to generate crystal structure landscapes which provide more insight to the packing geometry (and hence chemical and physical properties) of the predicted crystal structures. This figure was reproduced from Musil *et al.* Musil et al. (2018). It shows the crystal structure landscape of a set of azapentacenes before and after a crystal structure descriptor was used to generate a more physically informative landscape.

In this work we generate crystal structure descriptors using the techniques of topological data analysis (specifically persistent homology) - a technique that has found multiple applications in chemistry already Steinberg (2019); Xia et al. (2015); Minamitani et al. (2023) and easily satisfies the invariance properties of a descriptor as we shall see. We apply this to both the kinds of datasets which have been described above - that is, both small and large - in order, primarily, to build physically meaningful crystal structure landscapes that extenuate the differences between crystal structures which have different identifiable packing motifs.

Chapter 2

Topological Data Analysis

The motivation of topological data analysis is to use ideas from topology in order to study the ″shape″ of data in a way that is robust to noise and the chosen metric, that is how we measure the ″distance″ between the constituent points of our data set. Topological data analysis has found a great many applications in fields as diverse as financial networks Gidea (2017); political voting patterns Dlotko et al. (2019); Feng and Porter (2021); cosmology Christian et al. (2022); Heydenreich et al. (2021); the analysis of mobile phone networks Bajardi et al. (2015) and cancer diagnosis Vipond et al. (2021); Lawson et al. (2019). Applications in chemistry include solubility prediction Steinberg et al. (2019); Pirashvili et al. (2018); the study of amorphous materials Nakamura et al. (2015); Hiraoka et al. (2016); zeolites and metal-organic frameworks Lee et al. (2018) and developing molecular descriptors for both organic Li et al. (2022) and biochemistry Xia et al. (2015).

Most of our discussion will centre around persistent homology Edelsbrunner et al. (2002); Zomorodian and Carlsson (2005), a technique for computing the ″holes″ in a dataset at different spatial resolutions. The precise meaning of a ″hole″ in this context will be explained later but the upshot is that these holes capture the general connectivity of the set of data points in such a way that if we suppose that these points are sampling a topological space, then these ″holes″ can uniquely define this space.

The other main technique used in topological data analysis is the mapper algorithm Singh et al. (2007) which generates a graph which summaries the topological features of a dataset - this technique mostly finds applications in exploratory data analysis, is very sensitive to the model parameters and is better suited to very large datasets. As such there are no applications of this technique in our work.

In order to give an adequate background of persistent homology a basic overview of the subject of homology in algebraic topology is required. These ideas will be built on to give an the overview of the mathematical tools that have been used in this work.

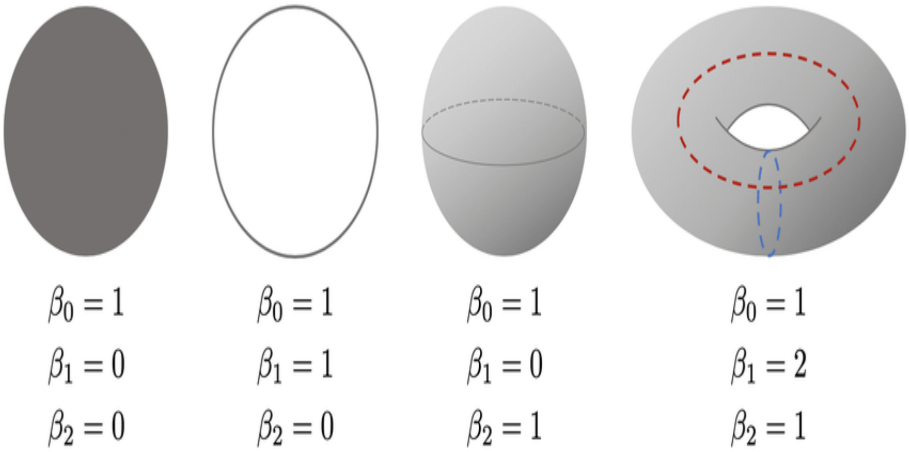


FIGURE 2.1: Some examples of some spaces with their first three Betti numbers Bobrowski and Skraba (2020)

2.1 Homology

The idea of homology is to find a set of Abelian (commutative) groups, corresponding to features in different dimensions, which encode the topological properties of the underlying space. In algebraic topology, spaces are classified according to the behaviour of closed loops that can be defined in the space - if the loops can be continuously (i.e. without cutting or gluing) deformed into one another then they are considered the same object, topologically speaking. In the case of a simply connected space (such as n-dimensional space $\mathbb{R}^n$, the surface of a sphere S^2, or a disk in 2D B^2), any loop that can be defined on the space (think of drawing any closed loop on the plane, for example) are topologically equivalent as they can all be deformed into a single point. A space in which this is *not* the case is the (space defining a) circle, S^1, as any loop covers the circle any number of times but cannot be deformed to single point. In general we find that any space in which the loops pass over some ”hole” cannot be contracted to a single point; in homology we count holes by systematical counting those cycles which cannot be contracted into themselves.

The homology groups are basically equivalence classes of such cycles in each dimension, considering higher order ”loops” for dimensions greater than one - the zero dimensional ”loops” are essentially points and count the number of connected components of the space while the two dimensional ”loops” encode trapped volumes or ”voids”. The holes in each dimension can be counted by considering the ranks of the homology groups or the Betti numbers, β_i. Some examples of different spaces and their Betti numbers are shown in figure 2.1.

In order to find these cycles in practice we need a way of representing the skeleton of the underlying space in different dimensions. This skeleton is called a (abstract) simplicial complex which is in turn composed of simplices. An N-simplex is a

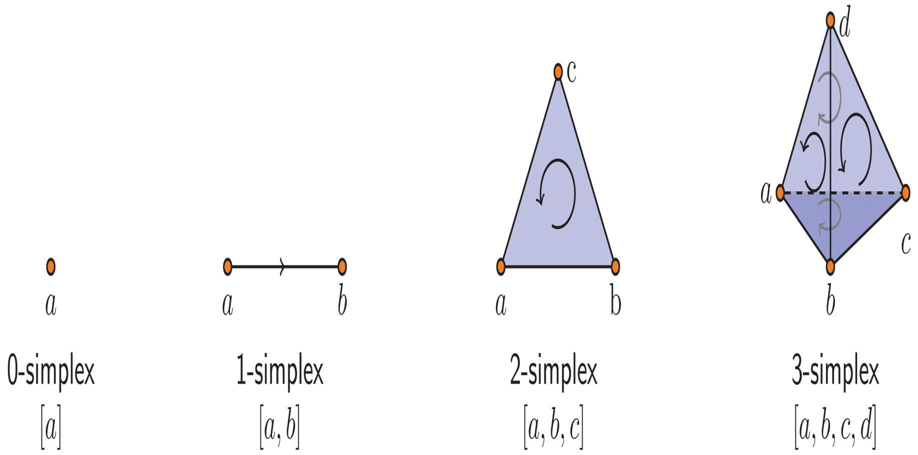


FIGURE 2.2: Some example simplices Topaz et al. (2014)

generalisation of a triangle in N dimensions, so, for example, a 0-simplex is a point, a 1-simplex a line segment, a 2 simplex a triangle and a 3 simplex is a tetrahedron. Obviously the simplices for which $n > 3$ do not have an intuitive geometric interpretation. Some example simplices are shown in figure 2.2.

A simplicial complex,$\mathcal{K}$, is then defined as a set of simplices that obey the following two conditions:

- Every face of a simplex in $\mathcal{K}$ is also in $\mathcal{K}$
- Any non-empty intersection of two simplices in $\mathcal{K}$ is a face of both simplices

The faces of a simplex are those simplices obtained by building a simplex from any subset of the $N + 1$ points that define the simplex. So that, for example, the faces of a tetrahedron contain all the triangles of its geometric faces as well as its edges and vertices.

For the purposes of finding homology we also define each simplex as having a given orientation. For any N-simplex with vertices labelled $(v_0, v_1, ..., v_N)$: if the vertices of two simplices are an even permutation of one another they have the same orientation; if the vertices of the simplices are an odd permutation of one another they have opposite orientations thus if the order of two vertices of a given simplex is swapped the orientation changes. Formally we say (for example) that $(v_0, v_1) = -(v_1, v_0)$.

If we take the vector space [1] of all N-simplices as C_N over the field $\mathbb{Z}_2 = \{0, 1\}$, such that each element of C_N is a linear combination of N-simplices - called a N-chain -

[1] In general, homology theory is defined using groups as opposed to vector spaces. We elect to define everything in terms of vector spaces as it makes the ensuing analysis simpler. The resulting theory is entirely rigorous enough for our purpose.

with coefficient either 0 or 1 (i.e. it is either in the sum or is not), we can define the boundary of a given simplex, $\sigma = (v_0, v_1, ..., v_N)$,as follows.

$$\partial_N(\sigma) = \sum_{i=0}^{N} (-1)^i (v_0, v_1, ..., \hat{v}_i, ..., v_N) \tag{2.1}$$

Where the simplex, $(v_0, v_1, ..., \hat{v}_i, ..., v_N)$, refers to the original simplex with the vertex v_i removed. The boundary is a map from C_N to C_{N-1} and coincides with our intuition of what a boundary should be: the boundary of an edge comprises the points at the beginning and end of the edge while the boundary of a triangle comprises its edges. We can then define a cycle as any N-chain which has zero boundary or as a member of the kernel of ∂_N. The space of cycles is $Z_N = \ker(\partial_N) = \{c \in C_N : \partial_N(C) = 0\}$. We can also define the space of boundaries, B_N, as the image of ∂_{N+1}:
$B_N = \text{im}(\partial_{N+1}) = \{c \in C_N : \exists c' \in C_{N+1}, \partial_{N+1}(c') = c\}$. That is, all elements of B_N, are the boundary of some higher dimensional object. As discussed earlier the homology groups (in our case vector spaces) contain all those cycles which are non-contractible - in practice this means that the cycle is not the boundary of some higher dimensional object (i.e. it is not filled in: a disk is simply connected while the circle is not). We can then define the homology space as the quotient space:

$$H_N = Z_N / B_N \tag{2.2}$$

This is equivalent to "factoring out" all cycles that are the boundary of some higher dimensional chain so that the vector space H_N has as its basis elements all of the "holes" present in the simplicial complex. Specifically we find that:

- zero dimensional holes correspond to connected components
- one dimensional holes correspond to loops
- two dimensional holes correspond to voids or "trapped volumes"

The dimension of the vector space - the number of basis elements - then counts the number of holes we also see that

$$\beta_N = \dim(H_N) = \dim(Z_N) - \dim(B_N) \tag{2.3}$$

So that the Betti number is equal to the number of N-cycles that can be constructed in our vector space minus the number of cycles that are already the boundary of a set

$N + 1$ simplices. For a more in depth explanation of the theory of homology and its place in algebraic topology see the book by Allen Hatcher Hatcher (2001).

Obviously in order to find the set of homology groups H_i we must first construct a simplicial complex that approximates the space - there are many examples of complexes that can do this so the complex we form typically depends on some kind of scale parameter, δ, so that we have some control on how many higher order simplices are included in the complex. This will become more clear with the following examples of simplicial complex. The idea of persistent homology is to use a whole range of values of scale parameter and analyse how the homology changes and as such we will typically be using a range of increasing scale parameters. From a theoretical point of view it can be shown Chung et al. (2021); Edelsbrunner and Harer (2010) that a construction called the Čech complex is a simplicial complex that ensures that the homology groups obtained are closest to the homology of the original space from which the data has been sampled. In this work the two simplicial complexes used are mainly the alpha complex and the Vietoris-Rips complex which can be shown to give a similar result to that which would be obtained using the Čech complex which as we will see is not a practical way of computing persistent homology. We will also discuss the similar concept of sublevel set persistent homology for which computations are carried out using the cubical complex.

2.1.1 Čech Complex

The Čech complex is defined in terms of a set of closed balls with radius ϵ (i.e. all points with a distance less than or equal to ϵ from some central point, for example a sphere in 3D) localised on each point in the dataset. A simplex is included if all the balls associated with each vertex of the simplex have a mutual intersection.

The Čech complex is impractical from a computational standpoint as the computation of the multiple intersections is difficult and often affords a very large number of simplices which can make the persistent homology calculation difficult in terms of both time and memory. To make matters worse this construction often results in simplices with a dimension higher than that of the underlying space which is not only impractical from a combinatorial point of view, but is also of limited geometric relevance.

2.1.2 Vietoris-Rips Complex

The Vietoris-Rips complex (VR complex) is the set of simplices for which a simplex is included if $d(v_i, v_j) \geq \epsilon$ for all i and j where ϵ is the scale parameter and $d(v_i, v_j)$ is the

Euclidean distance between vertices v_i and v_j. This is the same as the set of all simplices with diameter at least ϵ. We also note that if a simplex is in the VR complex all of its faces are also included. We typically specify both a maximum dimension of simplex and a maximum scale parameter to avoid an excessively large number of simplices which would make the computation of persistent homology intractable. It can be shown that the Čech complex is a subset VR-complex which in turn is a subset of the Čech complex with double the radius; we also know that the two complexes will have the same points and edges so as such we should expect the resulting homology to be quite similar.

Even when the maximum dimension of simplex is limited the number of simplices can grow very large so in practise we can employ a technique called sparsification to reduce the number of simplices while giving a similar result. The details of the technique are quite complex (see Sheehy (2013)) but the idea is that for large values of δ many points do not affect the homology of the complex so may be removed. The extent of sparsification is controlled by a parameter a: when a is close to zero the homology takes longer to compute but is a closer approximation to the that of the true complex, when a is increased the computation is much faster (and memory efficient) but is less likely to be accurate.

We can also consider the case for which the points are weighted, where some points are considered as ”larger” than others. This is useful for cases where atomic positions are used in the point cloud so that we can account for different atomic radii when constructing the simplicial complex. The idea here is that an edge will exist in the simplicial complex at a lower scale parameter when two points have low weight than if the two points had larger weights. This effect can be easily achieved in practice by dividing the distance between points by the sum of their respective weights - this has been shown the be a mathematically robust operation Bell et al. (2017).

2.1.3 Alpha Complex

The alpha complex is defined in the same way as the Čech complex above except instead of using balls centred on each point to find simplices we use the intersection of a ball centred on each point with its Voronoi cell.

We define the Voronoi cell associated with a particular element in a set of points, $u \in S$, as the region in space for which any point in this region is closer to point u than any other point in the set (S). More concretely if we have a set $S \subseteq \mathbb{R}^n$ then for a Voronoi cell centred at some $u \in S$ we have

$$V_u = \{x \in \mathbb{R}^n | \quad d(x,u) \leq d(x,v), \quad v \in S\} \tag{2.4}$$

given some distance function $d(x, y)$.

This gives the advantage of not generating any simplices of a dimension greater than the that of the space itself. Note that in the limit of infinite scale parameter we construct the complex using the Voronoi cells only and end up with a construction called the Delaunay complex which is the triangulation of the set of points for which no point in the set lies inside the circumcircle of any of the triangles Edelsbrunner and Harer (2010). This construction has many applications in computer science Dinas and Banon (2014); Liebeherr and Nahas (2001); Li et al. (2003); Weatherill (1992); Grise and Meyer-Hermann (2011). As such an alpha complex is always a subcomplex of the associated Delaunay complex so in practice the upper limit of the number of simplices that are included is not large (relatively speaking).

It is possible to consider the case of weighted points by using intersections of weighted Voronoi cells and weighted balls. The weighted Voronoi cell is defined as above expect we replace the metric in equation 2.1 with the power distance between x and weighted point u: $\pi_u(x) = d(u, x) - w_u$ where w_u is the weight of the point u. Similarly we give each weighted ball a radius $\sqrt{w + r^2}$. Note that for persistent homology calculations we increase the scale parameter such that the weighted ball with the smallest weight has zero radius at the start of the calculation. As such for weighted alpha filtration some of the complexes may then correspond to a negative scale parameter.

2.1.4 Cubical Complex

The application of this kind of simplicial complex will become more clear when we introduce sublevel set persistent homology but suffice it to say that when we compute the persistent homology of a function rather than of a set of points, it becomes more practical to work with cubes rather than triangles as the fundamental unit in our calculations. The resulting complex is not strictly a simplicial complex but has analogous properties Wagner et al. (2012).

We define an elementary cube with dimension N , Q_n as the product of N elementary intervals, $I_i \subset \mathbb{R}$, as follows

$$Q_N = I_1 \times I_2 \times ... \times I_{N-1} \times I_N \subset \mathbb{R}^N \tag{2.5}$$

Here any elementary interval is of the form of either $[n, n]$ or $[n, n+1]$ for integer n. This gives us constructions that correspond to edges, squares and cubes with increasing dimension as might be expected. We can then define the boundary of a cube in terms of the boundary of the constituent intervals as follows

$$\partial Q_N = \Big(\partial I_1 \times I_2 \times ... \times I_N\Big) + \Big(I_1 \times \partial I_2 \times ... \times I_N\Big) + ... + \Big(I_1 \times I_2 \times ... \times \partial I_N\Big) \tag{2.6}$$

Where we define the boundary of each interval I_i as

$$\partial I_i = \begin{cases} 0, & \text{if} \quad I_i = [n, n] \\ [n+1, n+1] - [n, n], & \text{if} \quad I_i = [n, n+1] \end{cases} \tag{2.7}$$

The cubical complex can then be defined as the set of elementary cubes such that the boundary of any cube already in the set is also in the set. Note for this particular case we must assign filtration values or "scale" to the individual cubes ourselves - in practice this is based on the value of the function for which we want to find the persistent homology at or around the points which define the elementary cubes.

2.2 Persistent Homology

The key construction for any persistent homology calculation is a sequence of nested simplicial complexes, K_i such that $K_1 \subset K_2 \subset K_3 \subset ... \subset K_i \subset ... \subset K_n = K$ called a filtration. A filtration can be constructed from any of the above examples of simplicial complex by increasing the scale parameter for each complex in sequence. Another option is to compute the persistent homology of a function $f : \mathbb{M} \to \mathbb{R}$ by using its sublevel sets $M_r = f^{-1}((-\infty, r])$ - this is the same as taking all values in the domain of a function with argument up to and including r; clearly this domain will get bigger as r is increased Chung et al. (2021); Mirth et al. (2020). The homology of the sets M_r can be computed either by constructing any of the simplicial complexes described above or, more often, by constructing the associated cubical complex which is more convenient to calculate on a function defined using a set of grid points.

We then find the homology of each complex in the filtration so that rather than obtaining a sense of the topological features of the space at a given scale parameter, we instead get a sense of how these features change as the scale is increased. Because each complex in the filtration is a subset of the next complex in the chain the number of simplices in the complex will increase as the scale is increased. While the rigorous definition Edelsbrunner et al. (2002); Zomorodian and Carlsson (2005) of persistent homology relies on the set of homology groups of dimension, i, at a given filtration step, j, $H_i(K_j)$ *and* the maps between these groups, $f_{kj} : H_i(K_k) \to H_i(K_j)$, in practice we interpret the output of a persistent homology calculation (and this can be shown to be a unique representation Zomorodian and Carlsson (2005)) by a set of intervals

(b_k, d_k) in different dimensions with associated multiplicities - these correspond to the birth and death of given homology features. An added simplex may give rise to new homology features (e.g. an added edge could form a ring) or could lead to the removal of a feature (e.g. an added 2-simplex could fill in an existing ring). We hence speak of a persistent homology feature, k, being born at scale b_k and dying at scale d_k. The lifetime of the feature is then $d_k - b_k$. If a feature is still alive at the maximum scale parameter then the feature "lives forever" and is considered to have $d_k = \infty$. For our work we carry out the calculations using the *gudhi* package Maria et al. (2014). A more in depth review of persistent homology can be found in references Edelsbrunner and Harer (2010) and Otter et al. (2017).

The two most common ways of visualising the output persistence intervals are persistence barcodes and persistence diagrams. A persistence barcode shows the intervals as a series of line segments with the beginning of line indicating the birth of a feature and the end of line indicating its death. Lines that reach the rightmost limit are understood to correspond to features that live forever. The intervals that correspond to features in different dimensions are often distinguished by different colours. Multiplicity is indicated by having multiple copies of the same line segment.

Another visualisation is the persistence diagram. In the persistence diagram the persistence intervals are plotted in the plane with birth (x) against death (y). Obviously as the lifetime of all persistent homology features must be greater than zero all points are above the line $y = x$. Again the maximum of the y axis is understood to denote $y = \infty$. The dimensions are likewise denoted with different colours. In this case the size of the data points is used to display the multiplicity of a given persistence interval. This construction is easier to visualise when we have very may persistent homology features.

The meaning of the two possible outputs is illustrated in figure 2.3 where the results of a persistent homology calculation on the vertices of a dodecahedron with edge length ϕ (where ϕ is the golden ratio) using the Vietoris-Rips filtration are shown in both formats (**e** : persistence diagram, **f** : persistence barcode). The 0D features are indicated by the red dots on the persistence diagram and the red bars on the barcode. We see that there are initially 20 connected components of which only one lives forever while the other 19 die at $\epsilon = \phi$ - this corresponds to when the scale parameter is equal to the edge length of the dodecahedron and the associated edges are included in the VR-complex. This transition is shown in **a** and **b**. The 1D features are represented by the blue dot on the persistence diagram and the blue bars on the barcode. We observe that there are eleven loops that are born at $\epsilon = \phi$ and die when $\epsilon = 2$. The loops are born once the edges of the dodecahedron are filled in (**b**) - they correspond to the twelve faces of the dodecahedron (we only count eleven loops as we can express one loop in terms of the others). The loops die at $\epsilon = 2$ (**c**) when they are

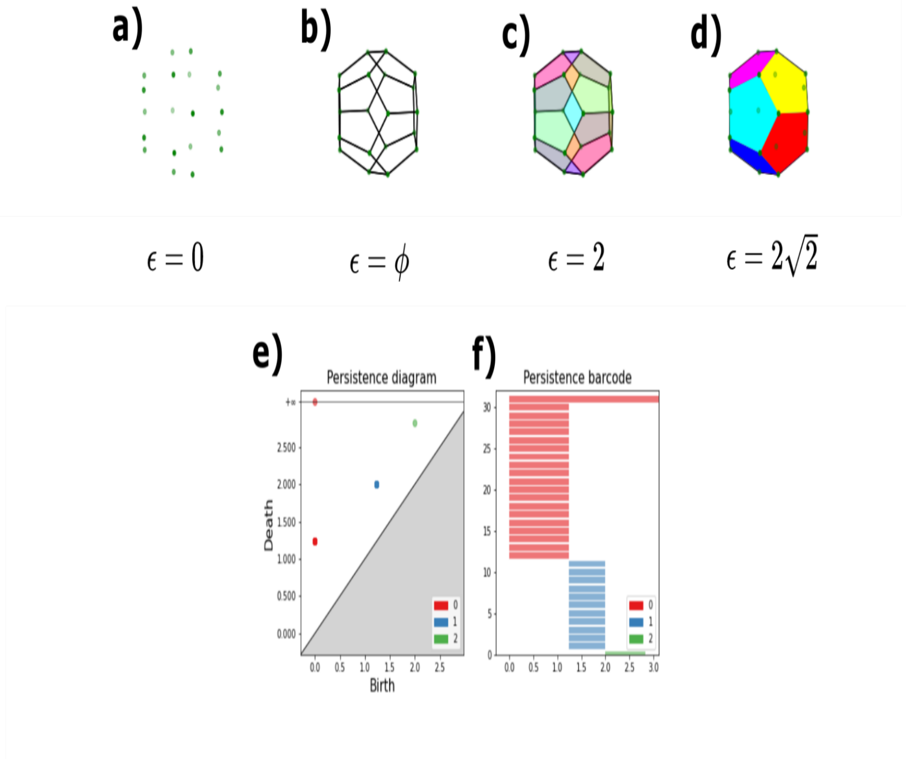


FIGURE 2.3: Example persistent homology calculation using the vertices of a dodecahedron as the input point cloud and using the Vietoris-Rips filtration.**(a-d)** show some examples simplicial complexes obtained during the course of the filtration: at $\epsilon = 0$ the simplicial complex only contains points; at $\epsilon = \phi$ (where ϕ is the golden ratio) the edges are joined; at $\epsilon = 2$ the distance between the faces is covered, as such the faces are "filled-in" with 1-simplices and finally at $\epsilon = s\sqrt{2}$ the trapped volumes within the dodecahedron are filled with 2-simplices. The associated persistence diagram is shown in **(e)** while the barcode is shown in **(f)**.

filled in with 1-simplices. There is only one 2D feature which is represented by the green dot on the persistence diagram and the corresponding bar on the barcode. The feature is born at $\epsilon = 2$ (**c**) and dies at $\epsilon = 2\sqrt{2}$ (**d**). At $\epsilon = 2$ the twenty faces already described act as the boundary of a trapped volume which is filled by tetrahedra once the internal edges are added at $\epsilon = 2\sqrt{2}$. We conclude by observing that while the persistence barcode represents the number of features more clearly it can quickly become cluttered; we see in figure 2.3 that the persistence diagram (**e**) is a much sparser representation of the associated persistence intervals.

2.3 The Representation of Persistence Diagrams

One of the central challenges of the application of persistent homology to practical problems in machine learning and data analysis is the difficulty in computing the statistical properties of persistence diagrams. The most rigorous way of comparing different persistence diagrams is by computing the p-Wasserstein distance between two diagrams D_1 and D_2 defined as follows:

$$W_p(D_1, D_2) = \inf_{\gamma : D_1 \to D_2} \left(\sum_{u \in D_1} \|u - \gamma(u)\|_\infty^p \right)^{\frac{1}{p}} \tag{2.8}$$

where inf denotes the *infimum* - the greatest lower bound - and γ denotes a bijection between the two diagrams D_1 and D_2. Since this map must, by definition be one-to-one we define persistence diagrams as also including every point along the diagonal ($y = x$) with infinite multiplicity - this does not affect the practical computation of persistent homology or the storage of the output persistence diagram - it is merely a technicality which ensures that the preceding equation makes sense mathematically. We also note that for the case in which $p = \infty$ we calculate the *supremum* (smallest upper bound) for each bijection - this construction is also called the bottleneck distance. Finally it is also worth noting that the stability results of persistent homology are proved with respect to the Wassestein metric: that is, two point clouds (or functions) which are very similar (for example, relative to the perturbations caused by the random vibrations of atoms and molecules) will give two persistence diagrams which are very close to each other with respect to the Wasserstein metric Cohen-Steiner et al. (2007).

While it is certainly possible to apply statistics and do machine learning on a general metric space (for example we can calculate the Fréchet mean of a set of persistence diagrams Turner et al. (2014) or convert the Wasserstein distance into a kernel for machine learning Carrière et al. (2017); Kusano et al. (2016)) this is seldom practical. The Wassersein distance is difficult to compute because it requires finding the perfect matching of a bipartite graph so finding the distance matrix for a large set of persistence diagrams each with very many points is impractical. Moreover many algorithms require the feature vectors to live in Banach space, that is, a space in which it is possible to compute both the length of a vector and the difference between two vectors. These are *not* defined for persistence diagrams in their current form. Thus in order to apply persistent homology to a wider variety of problems alternative methods for representing the output of a persistent homology calculation are highly desirable.

The most obvious way of converting the persistence diagram into a form more amenable to machine learning is to use histograms: in the case of 0D features (which all have the same birth ordinate) we need only find a 1D histogram, while for higher dimensional features we may use a 2D histogram.

The main problem with this approach is one of stability: a small change in the position of a birth-death pair may result in a large change in the underlying persistence representation if the point moves to a new bin. Another cause of instability can be when a new point emerges from the diagonal of the original persistence diagram Adams et al. (2017). The method of persistence images addresses this by first converting the diagram into a persistence surface by colvolving each point with a spherical Gaussian function to get a real valued function. To get a more efficient representation we first transform all pairs (b, d) in the diagram to $(b, d - b) = (b, l)$ (where l denotes the lifetime) which is equivalent to plotting all points by its distance from the diagonal. Then we define for each pair $u = (b, l)$ a function $g_u(x, y)$ as follows:

$$g_u(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{[(x-b)^2 + (y-l)^2]}{2\sigma^2}} \tag{2.9}$$

The persistence function for a diagram D is then

$$\rho_D(x,y) = \sum_{u \in D} f(u) g_u(x,y) \tag{2.10}$$

Here $f(u)$ is a weighting function which could be used in cases where we want to favour certain features when using statistical inference. In our work we set this to unity. Once we have defined the persistence diagram continuously we then construct a descriptor by discretising (up to a given resolution) the function into a set of pixels or boxes by integrating the persistence surface over the bounds of each pixel:

$$I_p(\rho_D) = \int \int_p \rho_D(x,y) \mathrm{d}x \mathrm{d}y \tag{2.11}$$

The grid is then then be flattened into a 1D representation. In order to account for the persistent homology features in different dimensions one need simply concatenate the persistence images calculated using the points from each dimension.

There are, of course, other methods of representing persistence diagrams such as persistence landscapes Bubenik (2015), Betti curves Umeda (2017) and complex polynomials Di Fabio and Ferri (2015) but these were found to be less effective at machine learning tasks on our data than the method of persistence images.

2.4 The Application of Persistent Homology on Crystal Structures

In order to apply persistent homology to crystal structures we must first deal with the periodicity of the crystal structure as persistent homology is typically applied to a point cloud with a finite number of points or over a function defined over given bounds. There are two possible ways to approach this. Firstly we could try to encode the periodicity into the simplicial complex - both periodic versions of the alpha complex (as an extension to the existence of the periodic Delaunay triangulation) Caroli and Teillaud (2009) and the cubical complex gudhi_periodic_cubical_complex_manual_page (accessed 2023) have been defined. Unfortunately the periodic Delaunay complex provably does not always exist for all point clouds Caroli and Teillaud (2009) although it *always* exists for a 27-sheeted covering of the space. Presumably a periodic alpha complex that uses a covering (and hence has duplicate points) is not amenable to persistent homology calculation. Similarly the periodic cubic complex, while it can always be defined, does not work when the periodic boundary conditions are not defined over a cube so will not work for the majority of crystal structures where the primitive cell is not cubic.

We hence adopt a second approach which is to use a finite cluster of points from the infinite crystal lattice. This approach may not be particularly limiting to the utility of the output persistence diagram: as a larger and larger fragment of the crystal structure is used the persistence diagram does not tend to change much. This is because as the periodic structure repeats, so too do the same homology features, so that if the point cloud is big enough virtually all of the important homology features will have been described albeit with different multiplicities. Another reason that the persistence diagram does not change much with increasing crystal fragment size is that cycles that occur over very large regions of a crystal fragment may be expressible in terms of those cycles that exist over smaller regions of the same fragment. These larger cycles would not be computed and so would not be included in the output persistence diagram. This is demonstrated in figure 2.4. A similar idea can be applied to the boundaries. That said we were not able to find any theoretical minimum size of crystal which captures all of the persistent homology features or indeed whether such a threshold even exists so being constrained to finite simplicial complexes is a definite but necessary limitation.

We demonstrate this size ”invariance” further in figure 2.5: here we use a 2d example as it is easier to visualise. The primitive cells whose centroid lies within the circle of increasing radius are shown on the left panels. The points included in these cells are used to find the persistent homology; the associated persistence diagrams are shown on the corresponding right panels. We see the diagrams are remarkably similar with

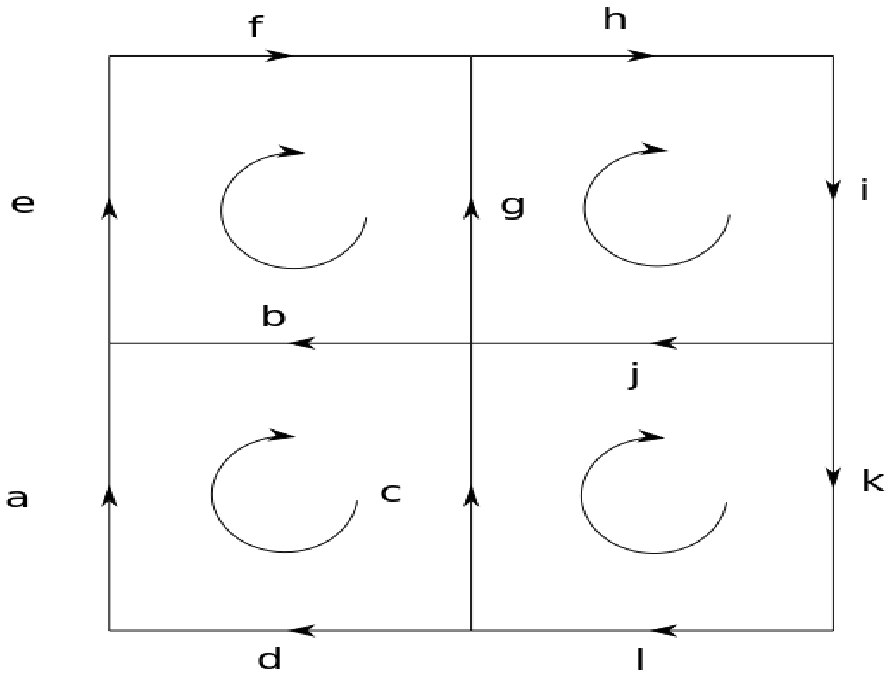


FIGURE 2.4: Example of decomposition of large cycle into smaller cycles. The large cycle $a + e + f + h + i + k + d + l$ can be written as the sum of the smaller cycles : $a - b - c + d, e + f - g + b, g + h + i + j$ and $c - j + k + l$.

increasing radius: the main difference being the commensurate increase in the multiplicity of each homology feature as the radius is increased.

In practice there are a number of ways to select a cluster of points for the persistent homology calculation: the most obvious is to construct a supercell - i.e. have some integer number of copies of each cell along each axis; another option is to include all molecules whose centroid lies within a specified sphere of given radius as in figure 2.5. We favour the latter approach as the primitive cell cannot be defined in a consistent manner.

We initially tried two different approaches - the first approach emphasised choosing a constant inclusion radius for the bounding sphere while the second approach ensured a constant number of molecules in the crystal fragment. Over time we found better results with the second approach. This corroborates earlier work Steinberg et al. (2019) that indicates that the number of points in each pointcloud has a significant effect on models which use persistent homology and hence that keeping the number of points constant in the input pointcloud across persistent homology calculations is highly desirable

The algorithm we use is shown in 2.6 and uses the Cambridge Structural Database's (CSD) API. The idea is that we make an arbitrarily large crystal fragment, define some centre of that fragment and then take exactly N molecules whose centroid lies closest to the centre of this arbitrarily large fragment. From there we can use every atom in

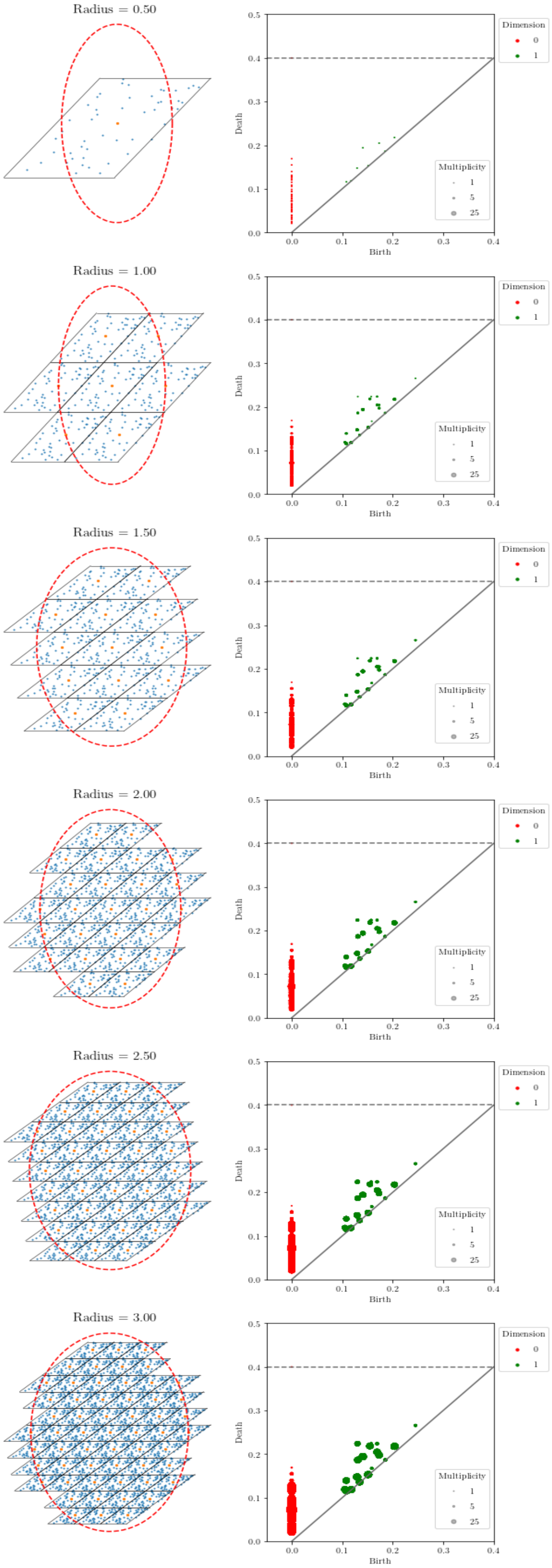
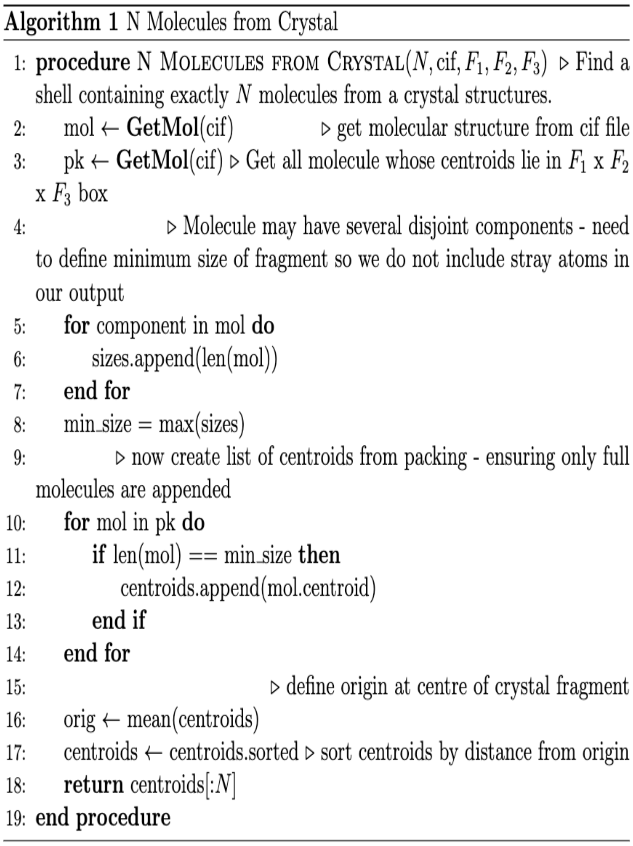


FIGURE 2.5: Persistence diagrams of periodic point sets using all primitive cells that lie the specified radius.



```
Algorithm 1 N Molecules from Crystal
procedure N MOLECULES FROM CRYSTAL($N$, cif, $F_1$, $F_2$, $F_3$)    ▷ Find a shell containing exactly $N$ molecules from a crystal structures.
    mol ← GetMol(cif)    ▷ get molecular structure from cif file
    pk ← GetMol(cif)    ▷ Get all molecule whose centroids lie in $F_1$ x $F_2$ x $F_3$ box
    ▷ Molecule may have several disjoint components - need to define minimum size of fragment so we do not include stray atoms in our output
    for component in mol do
        sizes.append(len(mol))
    end for
    min_size = max(sizes)
    ▷ now create list of centroids from packing - ensuring only full molecules are appended
    for mol in pk do
        if len(mol) == min_size then
            centroids.append(mol.centroid)
        end if
    end for
    ▷ define origin at centre of crystal fragment
    orig ← mean(centroids)
    centroids ← centroids.sorted    ▷ sort centroids by distance from origin
    return centroids[:$N$]
end procedure
```

FIGURE 2.6: Algorithm 1: For constructing a suitable crystal fragment with exactly N molecules from a cif file for the computation of persistent homology

the molecule list or the centroids only in order to compute the persistence diagram for each crystal structure. Sometimes the CSD python API incorrectly identifies some stray atoms as belonging to separate molecules, we avoid including these in our calculations as these adversely affect the results of the persistent homology calculations in the case in which we only use molecular centroids as our input pointcloud. We typically avoid this by ensuring that all molecules in the list have the same number of atoms as the molecule in the list with the largest number of atoms. Alternatively the size of the molecules can be specified beforehand by inspection of the molecular structure. This approach is clearly not applicable to crystal structures which correspond to co-crystals - in practise we found no problems with stray atoms in these examples.

Upon obtaining a set of persistence diagrams from crystal structures which warrant comparison, we typically find a distance matrix using the Wasserstein metric as described above or, more commonly, find a set of vector images associated with each crystal structure. Upon flattening the vector images we have a $N \times p$ matrix which describes the p dimensional space of crystal structures upon which further analysis

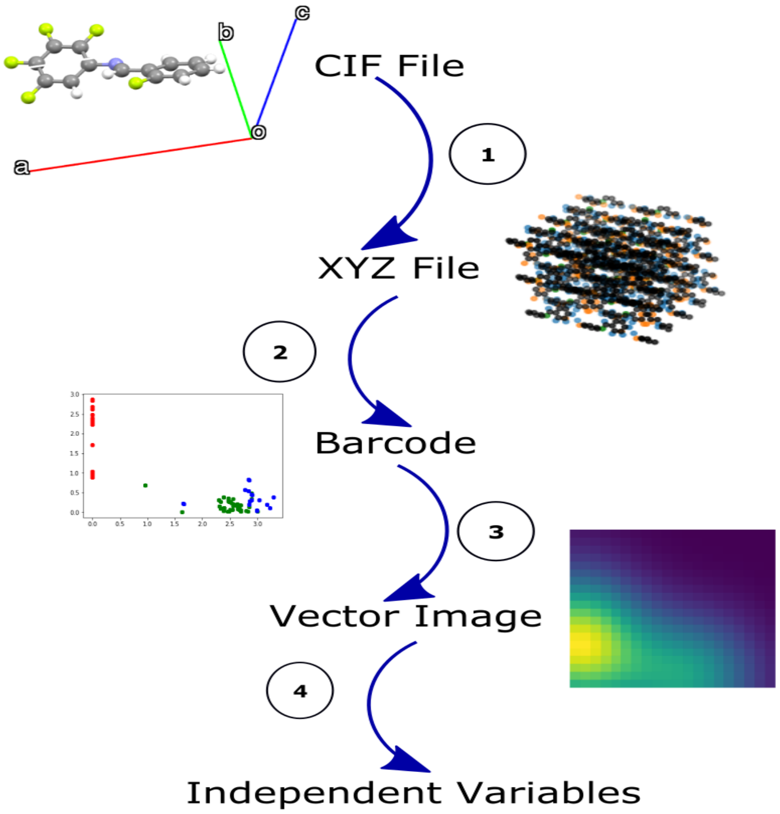


FIGURE 2.7: Flowchart for the processing of a cif file into a meaningful topological descriptor. Step 1: cif file is converted into a crystal fragment, which contains information about exactly N (typically around 50) molecules, typically this contains molecular centroids only but it can also contain atoms or other molecular invariants such as orientation vectors. This may be stored in an xyz file (if 3 co-ordinates are used). Step 2: the persistent homology is calculated using the coordinates provided in the crystal fragment by using one of persistent homology methods i.e. Vietoris-Rips or alpha filtration. We get a barcode or persistence diagram from this process. Step 4: Conversion of barcode into a vector image. As stated in the main body this is the technique most commonly used by us to convert the barcode to something useful. Other techniques are used by us as well such as the Wasserstein distance or the (1D) histogram of connected components. Step 4: the conversion of the vector image to a set of independent variables. This just involves flattening the $N \times N$ vector image into a set of N^2 values to be fed into our models.

may be carried out. In the case of the Wasserstein metric we are left with an $N \times N$ distance matrix for analysis.

The overall workflow used to generate a set of homology-based independent variables from a crystal structure (a cif file) is outlined in figure 2.7.

Chapter 3

Data Analysis Methods and Crystal Structure Prediction

The techniques we use to analyse the resulting data fall into three categories: dimensionality reduction, classification and regression.

Dimensionality reduction allows us to convert either the 400 dimensional space (or Wasserstein distance matrix) into a lower dimensional space - typically a two or three dimensional space- either to visualise the set of crystal structures to extenuate any trends in the data or to combine with classification or regression to improve the accuracy of these methods.

In classification we attempt to partition the set of data into discrete classes we can either do this in an unsupervised manner (this is called clustering) or in a supervised manner, that is, based on the fact that some or all of the data is already partitioned by other means. In this work most of the datasets which we work with have predefined labels predicted based on the intuitive packing scheme of the crystal structure and is largely based off analysing the crystal structures by eye.

Finally in regression we attempt to predict a continuous variable from the data. In our case this continuous variable is the calculated energy of each crystal structure. We will now describe each of the techniques which are used in this study according to the characteristics described above.

3.1 Dimensionality Reduction

3.1.1 Principal Component Analysis

The most common dimensionality reduction technique which we use is principal component analysis (PCA). The aim of PCA is to find a (smaller) set of new variables which are a linear projection of the original variables which capture most of the variance of the dataset and are also uncorrelated with each other.

More concretely let X be a $N \times p$ matrix, where N is the number of datapoints, describing the dataset and p is the number of variables that describe the data. Supposing that the mean of each set of datapoints in each dimension is zero (and if this is not the case we can accordingly recentre our data) then the sample covariance matrix is

$$S = \frac{1}{N-1} X X^T \tag{3.1}$$

which is a $p \times p$ symmetric matrix so may be written as

$$S = \frac{1}{N-1} U^T \Lambda U \tag{3.2}$$

by diagonalising the matrix. The matrix $\Lambda_{ij} = \begin{cases} 0 & \text{if} \quad i \neq j \\ \lambda_i & \text{if} \quad i = j \end{cases}$ is a (diagonal) matrix of eigenvalues, λ_i, which is basically the covariance matrix in the basis of (orthogonal) eigenvectors u_i which are the columns of U. These eigenvectors can be obtained from a linear combination of the vectors describing the data in our original basis as $u_i = \sum_{j=0}^{p} w_{ij} x_j$. Therefore we see that in this basis all of our vectors which describe the dataset are uncorrelated and hence we can choose the $q < p$ eigenvectors which correspond to the q largest eigenvalues as our principal components. For a number of different ways of reaching the same result and more information about PCA see Jolliffe and Cadima (2016); Greenacre et al. (2022); Bro and Smilde (2014). Note that for computational and stability reasons the matrix factorisation is carried out using singular value decomposition as opposed to standard diagonalisation Tipping and Bishop (1999).

3.1.2 Multidimensional Scaling

For the case in which the output of the persistent homology calculations is a distance matrix of Wasserstein distances as opposed to an p dimensional (Cartesian) space of vector images, PCA is not a valid method of dimensionality reduction. In these cases we use MultiDimensional Scaling (MDS) which is more suitable for data defined on distance matrices.

Given a (dis)similarity matrix between N data points the objective of MDS is to find a mapping (embedding) into a Cartesian space, $\mathbb{R}^m$, which preserves the distances between any two points as closely as possible. Here m can in principle be any positive nonzero integer although in practice values of 2 or 3 are chosen so that the resultant space can be visualised easily.

Specifically for an $N \times N$ similarity matrix $S_{ij} = s_{ij}$ whose elements denote the similarity between data points (in some fashion), an embedding into $\mathbb{R}^m$ is found with values $\{z_1, z_2, .., z_N\} \in \mathbb{R}^m$ which minimises the following stress function:

$$\text{stress}_S(z_1, z_2, .., z_N) = \sum_{i<j} \left[s_{ij} - \|z_i - z_j\| \right] \tag{3.3}$$

Where $\|...\|$ indicates the Euclidean distance between two embedded points in $\mathbb{R}^m$. In other cases this stress function may be normalised or modified to only preserve the ordering of the distances Hastie et al. (2009); Borg and Groenen (2005) but there are no applications of this in this work.

MDS may also be applied to a Cartesian space as in PCA by finding (as an example) the Euclidean distance matrix. This may yield more interesting results as the transformation of co-ordinates is inherently non-linear and such non-linear features may be revealed by an MDS-based approach (but at a higher computational cost).

3.1.3 Linear Discriminant Analysis

Sometimes it is desirable in our analysis to carry out *supervised* dimensionality reduction, that is, to find a projection of our high dimensional space which separates the set of points in the low dimensional space as much as possible according to some predefined partition or labelling. We mostly use such techniques to assess how very similar datasets behave under the same transformation. The most basic of supervised dimensionality reduction technique is perhaps Linear Discriminant Analysis (LDA).

There are a number of ways of formulating the ideas behind LDA (see Hastie et al. (2009); Sharma and Paliwal (2015) for example), but the method we use is that of the Fischer criterion Fischer (1936).

Suppose our dataset $X = (x_1, x_2, ..., x_N)^T$ (an $N \times p$ matrix) has assigned to it a set of class variables $y = (y_1, y_2, .., y_N)^T, y_i \in \{1, 2, .., c\}$ which can take values of 1 to c denoting c different classes each data point could belong to. The aim of LDA in this case is to find a projection of the dataset that *maximises* the variance *between* classes and *minimises* the variances *within* classes. The variance between classes is

$$S_b = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{3.4}$$

where there are N_i elements in class i and the mean of class i is μ_i while the mean of the the whole dataset is μ.

Meanwhile the variance within classes is

$$S_w = \sum_{i=1}^{c} (N_i - 1)\Sigma_i \tag{3.5}$$

that is, we simply take the average of the group variances $\{\Sigma_i\}$.

The problem of Linear Discriminant Analysis is then to find a projection U_{LDA} such that

$$U_{\text{LDA}} = \text{argmax}_U \left[\frac{\det(U^T S_b U)}{\det(U^T S_w U)}\right] \tag{3.6}$$

the determinant of the matrices here give a quantification of total variance arising from each projection of the variance matrix. It can be shown that the solution to this equation satisfies

$$S_w^{-1} S_b U = U\Lambda \tag{3.7}$$

that is U_{LDA} is composed of the eigenvectors of $S_w^{-1} S_b$ so note that we get at most $c - 1$ eigenvectors and therefore we can only project X to a dimension of at least $c - 1$ so in practice this technique is often combined with PCA to project to lower dimensions. Note also that S_w must be invertible in order for this to work. Since S_w has rank of at most $N - c$ so when N is less than $p + c$ there are serious stability issues with this algorithm. For this reason when dealing with datasets with a small number of datapoints relative to its dimension we either first reduce the dimension of the space with PCA or apply a technique called shrinkage where we approximate the covariance matrix as

$$S_w = (1 - \delta) S_w = \delta I \tag{3.8}$$

with shrinkage parameter δ. In this work the optimal shrinkage parameter is determined according to the method of Ledoit and Wolf Ledoit Wolf, Michael (2004).

3.1.4 Uniform Manifold Approximation

We also apply a more sophisticated non-linear dimensionality reduction technique to contrast with some of the linear methods described above. There are a number of non-linear options available to use such as MDS (which we have already discussed) and t-distributed Stochastic Neighbour Embedding (t-SNE) van der Maaten and Hinton (2008) but we found the best results both qualitatively and with respect to computation time with a new technique called Uniform Manifold Approximation (UMAP) McInnes et al. (2018) which has received a lot of attention in the literature and has found applications in a diverse set of scientific disciplines Diaz-Papkovich et al. (2021); Becht et al. (2018); Gensch et al. (2022); Rugard et al. (2021); Hozumi et al. (2021); Trozzi et al. (2021); Milošević et al. (2022); Lovrić et al. (2021); Vermeulen et al. (2021). This technique also has the advantage of being amenable to supervised dimensionality reduction in contrast to the more standard non-linear dimensionality reduction algorithms. The aim of UMAP is to provide a dimensionality reduction that respects the topology of local neighbourhoods as opposed to the set of absolute distances alone. UMAP has also been considered to provide a better description of the *global* topology of the dataset (than t-SNE, for example) Diaz-Papkovich et al. (2021); Becht et al. (2018) although this has been debated Wang et al. (2020); Kobak and Linderman (2019).

Like other non-dimensionality reduction techniques the UMAP algorithm boils down to finding an optimal matching between a weighted graph representing the points on the original manifold and a weighted graph in the lower dimension (embedded) Euclidean space. Unlike other non-linear dimensionality reduction algorithms the interpretation of the weighted graphs is heavily inspired by ideas from algebraic topology, fuzzy logic and category theory - the resulting structures are called fuzzy topological representations. These structures can be thought of as simplicial complexes built by the overlap of balls like the Čech complex described above - taking the 0 and 1 simplices only in this case - with a two crucial differences. Firstly each point has its local own metric so that all the balls have different sizes and, secondly, when building our complex the inclusion of an edge is not a binary outcome but a continuous variable such that the edge weight is related to the *probability* of inclusion. As the metrics are defined locally for each point, given two nodes labelled A and B, the edge weight from node A to node B may not necessarily be the same as the weight from node B to node A. As a result of this the bidirectional inclusion probabilities are combined in a mathematically robust way which we will define below. The fuzzy topological structure for the embedded space is somewhat simpler as the metric is the same for all balls (as we can be more confident in the uniform distribution of the data in the resulting Euclidean space) so as such we only need to compute the edge weight in one direction. The precise nature of these mathematical objects is beyond the scope of this work but for more information about these structures see McInnes et al. (2018); Allaoui et al. (2020); Ghojogh et al. (2021).

The mathematical objects that we will define are the edge weights (inclusion probabilities) that are actually computed for both the fuzzy representation of the high dimensional manifold and that of the embedded space.

For each point (in the high dimensional manifold) the probability weights are only found for the k nearest (Euclidean) neighbours of $x_i \in \mathbb{R}^p$ such that $p_{i|j} = 0$ if $x_j \notin \mathcal{N}_i = \{x_{i1}, x_{i2}, ..x_{ik}\}$ where x_{ik} is k^{th} nearest neighbour of x_i. In our applications k is set to 15 which is the default value. The probability of including the edge from node i (which corresponds to point $x_i \in \mathbb{R}^p$) to node j (which corresponds to point $x_j \in \mathbb{R}^p$) is then

$$p_{i|j} = \begin{cases} \exp\Big(-\frac{||x_i - x_j|| - \rho_i}{\sigma_i}\Big) & : \quad x_j \in \mathcal{N}_i \\ 0 \quad : \quad \text{else} \end{cases} \tag{3.9}$$

Where the $||...||$ indicates the Euclidean norm and ρ_i is the distance from x_i to its first nearest neighbour. The parameter σ_i is a scaling parameter which must be found to satisfy

$$\log_2 k = \sum_{j=1}^{k} \exp\Big(-\frac{||x_{ij} - x_j|| - \rho_i}{\sigma_i}\Big) \tag{3.10}$$

The weight of the edge from i to j , $p_{i|j}$, and the weight of edge from j to i, $p_{j|i}$ are combined to give an edge weight which is invariant to direction (this is just the probability that each one of the edges exist) as so

$$p_{ij} = p_{i|j} + p_{j|i} - p_{i|j}p_{j|i} \tag{3.11}$$

For the embedded space the edge weights are already invariant to direction as explained. For the edge connecting node i (which corresponds to point $y_i \in \mathbb{R}^q : q << p$) to node j (which corresponds to point $y_j \in \mathbb{R}^q : q << p$) we then have

$$q_{ij} = \left(1 + a\left\|y_i - y_j\right\|^{2b}\right)^{-1} \tag{3.12}$$

where a and b are hyperparameters. The optimal hyperparameters have been found to be $a \approx 1.929$ and $b \approx 0.7915$ McInnes et al. (2018) although there is some doubt about how much difference these make in practice Böhm et al. (2020).

To find the optimal graph matching a quantity called the cross entropy between distributions $P = \{p_{ij}\}$ and $Q = \{q_{ij}\}$ is minimised:

$$CE(P, Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1 - p_{ij}) \log \frac{1 - p_{ij}}{1 - q_{ij}} \tag{3.13}$$

In practice we start off with a best guess of the low dimensional representation of the dataset and then iteratively perturb the graph in the low dimensional space so that the cross entropy is minimised and the resultant graph is as similar as possible to the graph represented by the fuzzy topological representation in the high dimensional manifold. In the current algorithm this initial guess is provided with a technique called spectral embedding von Luxburg (2007) but this could also be accomplished with a similar technique such as isomap Tenenbaum et al. (2000) or MDS.

Lastly we note that in order to carry out the supervised version of this algorithm we need simply condition the probabilities in the fuzzy topological representation of the high dimensional manifold on the partition/labelling scheme UMAP_learn_documentation (acessed 2024). This will effectively force those nodes that correspond to the same label to be close together in our embedding.

3.2 Supervised Classification

3.2.1 Support Vector Classification

The only method for supervised classification that we describe here is Support Vector Classification (SVC) Cortes and Vapnik (1995) which, of the algorithms we attempted, provides the best combined classification for our data (and for all the problems to which we applied supervised classification).

Support Vector Classifiers are only strictly defined as a binary prediction algorithm, i.e. we can only predict whether a datapoint, $x_i \in X$ (X is a $n \times p$ matrix as before), belongs to one of two classes which we denote using a response variable $y_i \in \{-1, 1\}$ (i.e. the data point is in class A if $y_i = -1$ and in class B if $y_i = 1$). As such when we apply this technique to multi-class problems we apply several one-verses-one instances of SVC to the data, one for each pair of classes, and then classify the points according to a vote using the outcome of each and every binary classification Hsu and Lin (2002); Duan and Keerthi (2005).

The goal of SVC is to find a $p-1$ dimensional hyperplane that separates the p dimensional data into two classes denoted as above with the indicator variable $y_i \in \{-1, 1\}$ such that the the distance between the hyperplane and the points in each class is maximised. As the distance between a point $x_i \in \mathbb{R}^p$ and a $p-1$ dimensional hyperplane is $\frac{1}{\|\beta\|}(x_i^T\beta + \beta_0)$ (and we also note that $(x_i^T\beta + \beta_0)$ is negative when $y_i = -1$ and positive when $y_i = 1$) we can cast this as an optimisation problem

$$
\begin{aligned}
&\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 \\
\text{subject to} \quad & y_i(x_i^T\beta + \beta_0) \geq 1 \quad \forall i
\end{aligned}
$$

The solution can then be defined as $\hat{G}(X) = \text{sgn}(X^T\hat{\beta} + \hat{\beta}_0)$. This approach works for data that can be separated by a hyperplane however this is not always the case so in support vector classification we also allow a "slackness" by defining a set of slack variable $\{\xi_i\}$ which control how far each point may stray from the margin. The problem in this case is defined as

$$
\begin{aligned}
&\min_{\beta,\beta_0,\xi_i} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i \\
\text{subject to} \quad & \xi_i \geq 0, \quad y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \quad \forall i
\end{aligned}
$$

The parameter C acts as the reverse of regularisation, that is, the larger C is the less points are allowed to deviate from the boundary. Thus a smaller C can also prevent or alleviate overfitting problems at the expense of initial accuracy.

It can be shown using Lagrange multipliers $\{\alpha_i\}$ that the solution of the above is equivalent to solving the following *maximisation* problem (which is called the dual problem) Smola and Schölkopf (2004); Hastie et al. (2009)

$$\max_{\alpha_i} L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j$$
$$\text{subject to} \quad 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \quad \forall i$$

As the above expression is only in terms of the dot product between data points $x_i^T x_j = \langle x_i, x_j \rangle$ it is possible to solve this same optimisation problem using transformed variables $\phi(x_i)$ by replacing $\langle x_i, x_j \rangle$ (which we call the kernel) with $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ and noting that we express the solution

$$\hat{G}(X) = \operatorname{sgn}(\phi(X)^T \hat{\beta} + \hat{\beta}_0) \tag{3.14}$$
$$= \operatorname{sgn}\left(\sum_{i=1}^{N} (\alpha_i y_i \langle \phi(x_i), \phi(x_j) \rangle) + \beta_0 \right) \tag{3.15}$$

We call this technique the *kernel trick* Boser et al. (1992); Smola and Schölkopf (2004); Hastie et al. (2009). Transforming the variables in this way gives a non-linear boundary (by implicitly encoding a much higher dimensional space) that can vastly improve classification accuracy while the computational cost is low as the matrix of kernels is relatively easy to compute.

There are number of choices of kernel for this kind of problem Hastie et al. (2009); Smola and Schölkopf (2004) such as the polynomial kernel, the sigmoid kernel and even a kernel based off the Wasserstein distance between two persistence diagrams Carriere et al. (2017) (which sadly we found *not* to be very effective) but the kernel we found most effective (other than the linear kernel, $\langle x_i, x_j \rangle$, as above) was the radial kernel which is defined as

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{3.16}$$

As such for any classification problem that we carry out in this manner we must fit the hyperparameters C, *kernel* and γ (if we choose a radial kernel). As a result of this we need to carry out a computationally intensive hyperparameter grid search when optimal results for the classification are desired.

Note that as is standard best practise in data science and statistics if we wish to predict the set of labels for one of our datasets, we hold out around 10-25 % of the data as a test set. The model is then fitted on the remaining training set, typically with 5-10 folds of cross validation Ojala and Garriga (2010) which involves further splitting into test and train sets, before obtaining a final accuracy score by predicting labels on the test set with the best model thus obtained and comparing these to the true values. We can obtain either a combined accuracy score (i.e. the percentage correctly classified) or a confusion matrix which shows the number of data points in each class of the test set by its predicted category. Also note that any fitting of the hyperparameters is carried out on the the training set and should not involve the test set in order to prevent overfitting the hyperparameters.

3.3 Unsupervised Classification

3.3.1 K-means Clustering

Given N observations $(x_1, x_2, ..., x_N)$, the aim of K-means clustering is to partition this set into a K sets (clusters) $\mathcal{C} = (C1, C2, ..., C_K)$ where $K <= N$. For this algorithm the number of clusters, K, is specified beforehand.

The clusters are chosen such that the variance within each cluster is minimised so that we aim to solve

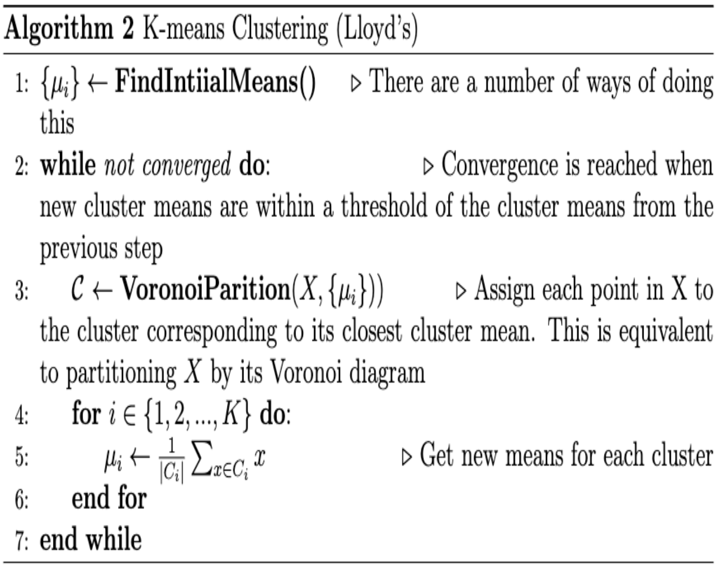
$$\text{argmin}_{\mathcal{C}} \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{3.17}$$

where μ_i is the centroid of the cluster C_i. This is the same as minimising the sum of all the pairwise distances within each cluster Hastie et al. (2009).

While it is theoretically possible to find the global optimum of such a calculation, this is normally intractable. The most common algorithm for finding a local minimum is Lloyd's algorithm Lloyd (1982) and this is the method used in this work.

The method is described in figure 3.1. There are number of ways to choose the initial set of cluster means which can have a drastic effect on the results. We use a method called k-means++ Arthur and Vassilvitskii (2007). This method finds a set of initial cluster centroids that are as far apart as possible by selecting points in a sequential manner such that the probability of picking a point is proportional to its distance from existing centroids.

It is often desirable to compare the output of such an unsupervised calculation with a predefined set of labels (or ground truth) that are known already to partition the dataset. The metric we use is called the adjusted mutual information (AMI) Vinh et al.



Algorithm 2 K-means Clustering (Lloyd's)

```
$\{\mu_i\} \leftarrow$ FindIntiialMeans()    ▷ There are a number of ways of doing this
while not converged do:    ▷ Convergence is reached when new cluster means are within a threshold of the cluster means from the previous step
    $\mathcal{C} \leftarrow$ VoronoiParition($X, \{\mu_i\}$))    ▷ Assign each point in X to the cluster corresponding to its closest cluster mean. This is equivalent to partitioning $X$ by its Voronoi diagram
    for $i \in \{1, 2, ..., K\}$ do:
        $\mu_i \leftarrow \frac{1}{|C_i|} \sum_{x \in C_i} x$    ▷ Get new means for each cluster
    end for
end while
```

FIGURE 3.1: Algorithm 2: Pseudo-code for k-means clustering by Lloyd's algorithm.

(2010). Given two partitions of our dataset $X = (x_1 x_2, ..., x_N), \mathcal{U} = \{U_1, U_2, ..., U_p\}$ and $\mathcal{V} = \{V_1, V_2, ..., V_q\}$ the mutual information is defined as follows

$$\mathrm{MI}(\mathcal{U}, \mathcal{V}) = \sum_{i=1}^{p} \sum_{j=1}^{q} P_{UV}(i,j) \log \frac{P_{UV}(i,j)}{P_U(i) P_V(j)} \tag{3.18}$$

This quantity has its roots in information theory Shannon (1948) and describes how much information about distribution $\mathcal{U}$ is obtained by observing $\mathcal{V}$. Here $P_U(i) = \frac{|U_i|}{N}$ is the probability of observing an object in class U_i (out of *all* objects) and similarly $P_V(j) = \frac{|V_j|}{N}$. $P_{UV}(i,j) = \frac{|U_i \cap V_j|}{N}$ is the probability of observing the intersection between two classes.

The *adjusted* mutual information is a modification of the above such that the expression yields 0 for completely unrelated partitions and 1 when the partitions are the same. Formally AMI is defined as

$$\mathrm{AMI}(\mathcal{U}, \mathcal{V}) = \frac{\mathrm{MI}(\mathcal{U}, \mathcal{V}) - E[\mathrm{MI}(\mathcal{U}, \mathcal{V})]}{\max(H(\mathcal{U}), H(\mathcal{V})) - E[\mathrm{MI}(\mathcal{U}, \mathcal{V})]} \tag{3.19}$$

where $E[...]$ denotes the expectation value and $H(\mathcal{U}) = -\sum_{i=1}^{p} P_U(i) \log P_u(i)$ is the entropy of partition $\mathcal{U}$.

Finally we note that while the choice of the number of clusters,K, has a large impact on the result of the clustering algorithm we typically set this to the number of labels in the ground truth - i.e. we want to see if we can replicate the same partition as expected

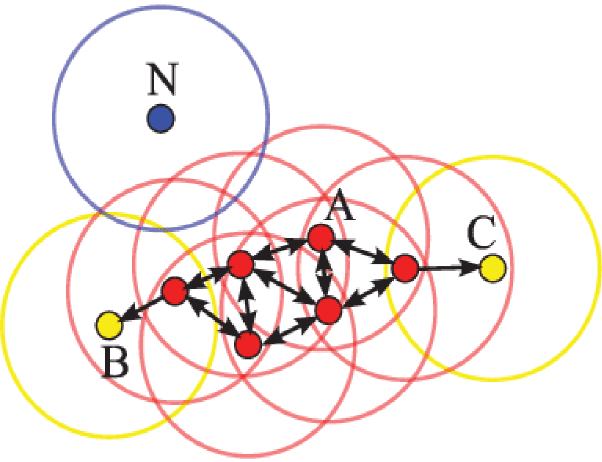


FIGURE 3.2: Example of core points, boarder points and noise Schubert et al. (2017). Here *MinPts* is 4. The red points are core points, the yellow points are boarder points and the blue point is noise.

by our domain knowledge. However as the goal here is compare the results of the same clustering algorithm on different descriptors the choice of K shouldn't matter too much as long as we are consistent between different datasets.

3.3.2 DBSCAN

In some cases it is desirable to cluster the data in such a way that not all data is assigned a label, that is, some of the data is labelled as either "undefined" or "outlier". We use a technique called density-based spatial clustering of applications with noise (DBSCAN) Ester et al. (1996) which aims to partition the data into regions of low and high density so that the high density regions correspond to clusters and points in the low density regions are more likely to be outliers. The algorithm also has the advantage that the number of clusters is not specified beforehand and also that the clusters need to not have a convex shape.

The two free parameters of most relevance to the result are the radius, ϵ, which determines the distance scale for cluster inclusion and the minimum number of samples,*MinPts*, which determines how many points need to be "close" to be constituted a cluster.

More formally we define any point with more than *MinPts* within a distance of ϵ as a *core point*. All points that are within ϵ of a core point are considered to belong to the same cluster as the core point. The neighbours of any core point of a cluster which is not itself a core point is designated a *boarder point*. These points intuitively correspond to the edges of the cluster. Any point that is not within ϵ of *any* core point is considered noise, that is, an outlier. The concept of core points, boarder points and noise is illustrated in figure 3.2.

The points are assigned into clusters by iterating through all the points in the dataset: if the point is not a core point it is labelled as noise; if the point *is* a core point we start building a cluster by iteratively finding its neighbours and the neighbours of any core points thus found and so forth until no more core points are found - any core points found in this manner are added to the cluster. Any object that has already been assigned a cluster is skipped as we iterate through the rest of the data. More details about the algorithm and pseudocode can be found in Ester et al. (1996) and Schubert et al. (2017).

As the parameters ϵ and *MinPts* have a substantial effect on the results of this algorithm, in this work we try a few sets of parameters and manually choose the ″best″ set of parameters. Since we only apply this technique on low dimensional data or data that is to be embedded in a low dimension checking this manually is quite easy as the clustering can be readily visualised.

3.4 Regression

3.4.1 Support Vector Regression

The first technique we use for regression is Support Vector Regression (SVR) Drucker et al. (1996) which is closely related to the technique of support vector classification described above.

By employing kernels as before the form of the regression model is $f(x) = \phi(x)^T\beta + \beta_0$ where $\phi(x)$ is a function that describes the kernel as in the SVC case. Note that when a linear kernel is used the solution is of the same form as the standard linear regression model (see Hastie et al. (2009),for example, for more information about standard linear regression).

By only allowing each prediction, $f(x)$, to lie within a region , ϵ, of the training value with the ″slackness″ allowed for each point defined by two variables ζ_i and ζ_i^* - these respectively define the slackness allowed below and above the target - we then define a very similar optimisation problem to SVC as

$$\begin{aligned} &\min_{\beta,\beta_0,\zeta_i,\zeta_i^*} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}(\zeta_i + \zeta_i^*) \\ \text{subject to} \quad &\zeta_i, \zeta_i^* \geq 0, \quad y_i - \phi(x)^T\beta + \beta_0 \leq \epsilon + \zeta_i \\ &\phi(x)^T\beta + \beta_0 - y_i \leq \epsilon + \zeta_i^* \quad \forall i \end{aligned}$$

As in the case of support vector classification we find the best kernels are the linear kernel and the radial kernel and, as such, for this case the hyperparameters which must be fit are γ, C, ϵ and *kernel* type (that is, radial or linear). Note that we carry out the same train test split procedures as mentioned above, although the fit of the model is typically assessed with the coefficient of determination, $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f(x))^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$, as befits a regression model.

3.4.2 Random Forest Regression

The technique of random forest regression works by combining the results of an *ensemble* of decision trees, that is, the results of several individual decision tree learning tasks are combined. As such we first describe how decision trees are used as regression models.

A decision tree creates a partition of the dataset into m regions $\{R_1, R_1, ..., R_m\}$ such that similar datapoints belong to the same region. This is done by recursively splitting the space according to decision rules such that we may think of the data being split according to a tree-like structure so that each condition on the data splits the tree into further branches. We only consider the case of *binary* trees such that each split results in *two* smaller regions of the dataset. The nodes at the bottom of the tree will then correspond to the regions $\{R_i\}$. The regression is then computed as

$$f(x) = \sum_{i=1}^{m} c_i I(x \in R_i) \tag{3.20}$$

where $I(condtion)$ is the indicator function and evaluates to one if the condition is false and evaluates to zero otherwise.

For each binary split the splits into regions $R_{\text{left}} = \{X : X_j \leq s\}$ and $R_{\text{right}} = \{X : X_j > s\}$ are chosen using the following minimisation

$$\text{min}_{j,s} \left[\text{min}_{c_1} \sum_{x_i \in R_{\text{right}}} (y_i - c_1)^2 + \text{min}_{c_2} \sum_{x_i \in R_{\text{left}}} (y_i - c_2)^2 \right] \tag{3.21}$$

so that we are minimising the sum of squares loss with the training data.

We find that the optimal c_i are the averages of the training variables over the corresponding region:

$$\hat{c}_i = \frac{1}{|X : x \in R_i|} \sum_{x_j \in R_i} y_j \tag{3.22}$$

while the optimal values of j and s must be found computationally.

These trees have a high propensity to overfit (i.e. the modal has low bias but high variance) to the training data, especially if a large number of splits (or equivalently a tree with a large *depth*) is used. There are a number of ways of mitigating this such as reducing the size of the tree after fitting (pruning) Hastie et al. (2009); Breslow and Aha (1997) or fitting successive trees with altered weights (boosting) Hastie et al. (2009); Freund and Schapire (1995) but the method of random forest reduces the variance of the model by combining the output of very many different trees (trained on different subsets of the dataset) in a manner which also minimises the correlation between trees.

The decision trees are combined using a technique called *bagging* (bootstrap aggregating) which involves fitting many instances of the same model on different sets of altered training data which is obtained by sampling with replacement. Suppose that we fit B models on different sets of training data $(X_1, y_1), (X_2, y_2), ..., (X_B, y_B)$ to obtain models $f_1(x), f_2(x), ..., f_B(x)$, the bagged predictor is the function

$$\hat{f}_{\text{bagged}}(x) = \frac{1}{B} \sum_{i=1}^{B} f_i(x) \tag{3.23}$$

It can be shown Hastie et al. (2009) that the variance of such a model is

$$\text{Var}(\hat{f}) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{3.24}$$

where σ is the variance of each tree and ρ is the correlation between trees. We thus see that even with a large number of bootstrap samples, the correlation between trees needs to be minimised.

This is achieved by modifying the splitting process of each tree in the bootstrap process: before every split is carried out randomly select $m < p$ out of the p variables - only these m variables may be used in the split condition. This prevents each tree from selecting similar variables in general and hence reduces the correlation between trees. Random forests were first introduced by Breiman Breiman (2001) while decision trees in general have been around much longer Belson (1959); Leo Breiman Jerome Friedman (1984).

This regression technique involves a large number of hyperparameters which we attempt to tune when using this model. We describe these below:

- **max_depth** - The maximum depth of any tree. The default setting is an unlimited depth for each tree and in this case the tree is split until each node has less than *min_samples_split* as described below.
- **max_features** - The numner of features, m, to use when reducing the paramter space before each splitting step (as discussed above), this often takes a function e.g. passing *sqrt* sets $m = \sqrt{p}$
- **max_leaf_nodes** - sets the maximum number of leaf nodes (at the base of the tree). The best nodes are selected as those that minimise square error as above. The default is an unlimited number of leaf nodes.
- **max_samples** - How many samples are used for each bootstrap dataset X_b
- **min_samples_split** - Minimum number of samples required to split an internal node
- **min_samples_leaf** - The minimum number of samples allowed in each leaf node, i.e. a split is *not* allowed (even if there are more than *min_samples_split* samples at this node) if the split results in one of the leaves (or children) from this split having less than *min_samples_leaf*.
- **n_estimators** - The number of trees in the random forest, i.e. the number of bootstrap samples, B

Note that due to the large number of hyperparameters in this case we sometimes opt for a random parameter search when fitting models for the larger datasets as checking every combination of hyperparameters is not feasible.

3.5 Crystal Structure Prediction

As well as dealing with real crystal structures, we also perform persistent homology calculations on predicted crystal structures. Crystal Structure Prediction (CSP) techniques typically generate a range of possible crystal structures for a given compound with calculated energies and densities. This is called a crystal structure landscape. Typically these structures are plotted as energy against density but different dimensionality reduction techniques may be used to plot the output in such a way that various properties of the crystal structure landscape are extenuated: for example one might want compounds with very similar crystal packing types to be close together in the reduced space. Using persistent homology to construct such plots that tease apart the regions of the crystal structure landscape that correspond to different crystal packing types is a key application of our work. As such we give a brief description of how these CSP techniques work.

Molecular organic crystal structure prediction algorithms typically consist of the following steps Bowskill et al. (2021); Day (2011):

- **Molecular Geometry Initialisation** The geometry of the starting structure must first be optimised such that the bond lengths and angles are representative of the real geometry of the molecule. Generally in CSP the molecular geometry is fixed (i.e. the molecules are assumed to be rigid) during the candidate generation and structure refinement stages to reduce computational load so it is important to have a sensible starting structure. Typically a simple calculation (for example B3LYP/6-311G** Density Functional Theory) in the gas phase is sufficient.
- **Candidate Generation** The optimised molecules are placed into the asymmetric cell with sets of differing positions and rotations either in a biased (such as a Monte Carlo search Kendrick et al. (2011)) or unbiased (such as a naive random search Karamertzanis and Pantelides (2005)) manner. The number of molecules per asymmetric cell, Z', is typically set to one in the first instance but this can be increased to explore more structures at the expense of a higher computational cost. Given a candidate asymmetric cell, the unit cell and hence an entire crystal structure can be found given its space group. In practice only the unit cell is needed for further calculations. While there are 230 space groups the vast majority of crystal structures belong to a very small set of space groups so typically only a small set of space groups (60 space groups account for 97.9% of all organic crystal structures Groom et al. (2016)) is used in the CSP algorithm which is specified beforehand. Sometimes computationally cheaper calculations are used to find and remove high energy structures. Also it is common to remove structures which have a high level of geometric similarity or in which molecules overlap: there are a variety of common approaches to this Bowskill et al. (2021).
- **Structure Refinement** The candidate structures are further optimised using varying levels of theory depending on the number of structures and the problem at hand.

In this work we use the techniques developed by Day *et al.* Case et al. (2016). After an appropriate geometry optimisation (the details vary but it typically involves a Density Functional Theory calculation using the Gaussian program Frisch et al. (2016)), the candidates are generated using the technique of quasirandom (*quasi*random because the random variables are not necessarily independent of each other) sampling Case et al. (2016). This technique is unbiased but ensures that random rotations of each molecule in the unit cell are appropriately sampled and that the unit cells are not unphysical (e.g. that the unit cell does not have a tiny cell angle and thus is long and thin). The positions of the molecule are determined using uniform random numbers while the random rotations of each molecule are determined using the method of

Shoemake Kirk (1992) which effectively gives uniform random rotations. The parameters for the unit cell (subject to the predefined spacegroup) are also generated at random. The angles (those not constrained by spacegroup symmetry) are sampled such that the cosine of the angle is evenly distributed and that the angles are not too acute or too obtuse (which gives rise to ”flat” unit cells which can present computational difficulties later on). The cell lengths are then sampled in such a way that the molecules which have already been defined always fit in the box provided. Structures for which molecules overlap are removed at this stage as well.

The lattice energy minimisation is carried out using the DMACRYS Price et al. (2010); Willock et al. (1995) program which applies the technique of distributed multipole analysis to crystal systems. Here the intermolecular interaction energy between two molecules, M and N, is modelled as

$$E_{MN}^{\text{intermolecular}} = \sum_{i,k} A^{pq} e^{-B^{pq} r_{ik}} - C^{pq} r_{ik}^{-6} + E_{ik}^{\text{elec}}(\text{DMA})$$

where i and k are respectively atoms of type p and q. The repulsive and attractive terms modelled by the first two terms are parameterised according to the revised version Pyzer-Knapp et al. (2016) of the Wiliams99 (W99) Williams (2001) force field. These correspond to the nonelectrostatic interaction (e.g. Van der Waals type interactions). The electrostatic term,$E_{ik}^{\text{elec}}(\text{DMA})$, is modelled using distributed multipole analysis Stone and Alderton (2002) of the B3LYP/6-311G** charge density. Typically (and in this work) multipoles of up to rank 4 (hexadecapoles) are used in the analysis.

Finally a clustering step is performed using a modified version of the COMPACK algorithm Chisholm and Motherwell (2005) in order to remove any duplicate crystal structures obtained at the end of the crystal structure prediction process.

Chapter 4

Experimental Data

We apply the topological data analysis methods described above to a range of datasets which shall be described in turn.

4.1 Fluorinated Benzylideneanilines

An example structure of a fluorinated benzylideneanine, or fluoroalanine for short, is shown in figure 4.1. The substitution pattern of fluorines varies across the dataset, if we assume the molecule is conjugated throughout then there are 1024 possible fluoroalanines else there are only 400 possible compounds. It is unclear whether the molecule is fully conjugated as in some crystal structures the rings are twisted while in others the molecule is planar. In either case we have crystal structures recorded for 118 fluoroalanine compounds with one set of polymorphs so we have 119 crystal structures in total.

In the work of Dodd *et al.* Dodd (2020) these structures were classified into two schemes of different packing motifs. The first of these was established by visual inspection while the second was based off analysis of the crystal symmetry,

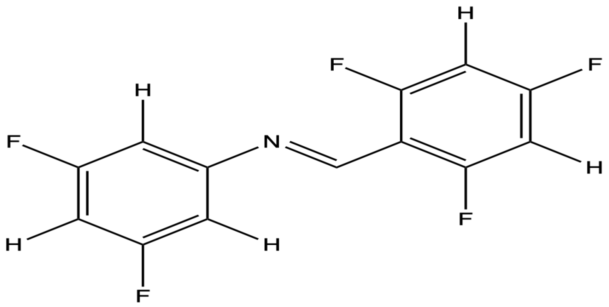


FIGURE 4.1: Representative fluoroalanine structure

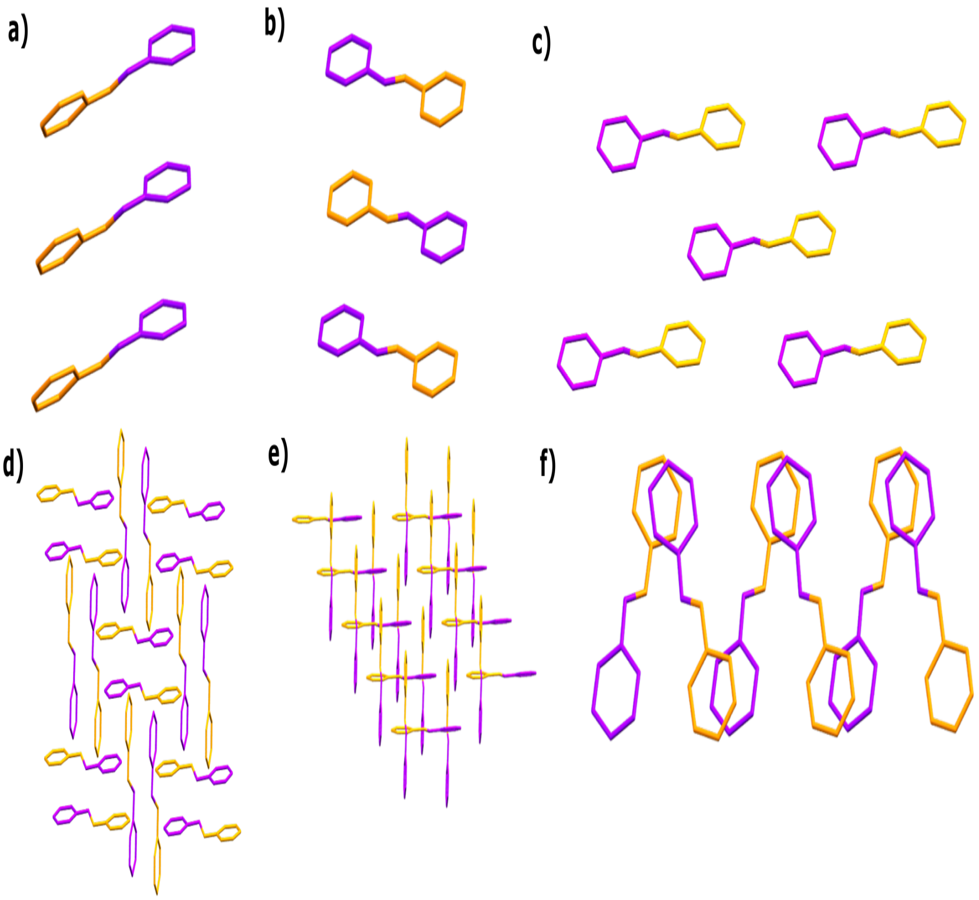


FIGURE 4.2: Example packing motifs of fluoroalanine crystal structures after Dodd *et al.*
Dodd (2020). The labels are as follow: a) Head-to-Head, b) Head-to-Tail, c) Staggered Overlap, d) Interwoven, e) Grid, f) Angled Overlap

intercentroid distances and dihedral angles. In the visual labelling scheme the structures are partitioned into the following motifs: head-to-head; head-to-tail; staggered overlap; angled overlap; interwoven; grid and ”other”. In the second packing scheme the head-to-head and head-to-tail classes are further subdivided into stacking groups 1,2,3 & 4. Some examples of these motifs are shown in Figure 4.2 .

We were not able to find any chemical applications for these sets of molecules in the literature; rather these crystal structures have been systematically analysed by Dodd *et al.* Dodd (2020) and others Kaur and Choudhury (2015); Kaur et al. (2012); Kaur and

Choudhury (2014) in order to understand how the C-F bond directs crystal packing - this is important as these kind of interactions play an important role in directing the crystal packing of many important molecules in medicinal chemistry, materials chemistry and elsewhere Berger et al. (2011).

Due to the quality of the labelling scheme this dataset is ideal for experimenting with different crystal structure descriptors - we have a firm concept of ”ground truth” for packing labels in this case. As such the bulk of our analysis is centred on this dataset.

4.2 Polyaromatic Hydrocarbons & Azapentacenes

The next class of compounds that are studied are polyaromatic hydrocarbons (PAHs) and the closely related azapentacenes. Like the fluoroalanines the packing structures of such compounds have been extensively studied in order to aid the field of crystal engineering Desiraju and Gavezzotti (1989b); Loots and Barbour (2012); Desiraju and Gavezzotti (1989a). Unlike fluoroalanines PAHs are well known to adopt one of four packing motifs known as beta (also known as sheet), gamma, herringbone and sandwich herringbone. Examples of these are shown in figure 4.3. The packing type depends on the interplay between edge-to-face interactions between the C-H bond and π systems and the face-to-face π-π stacking and holistic methods for identifying the motifs have existed for a long time Desiraju and Gavezzotti (1989b). The packing types adopted by different PAHs (or related compounds) influence the electronic conductivity properties (specifically the charge mobility as determined by intermolecular electronic coupling) Musil et al. (2018); Valeev et al. (2006) and these are in turn relevant to the possible applications of PAHs such as molecular organic semiconductors Sirringhaus et al. (1998).

There has been particular interest in the applications of pentacene Anthony (2008); Wang et al. (2012); Kitamura and Arakawa (2008) in the area of molecular semiconductors as well as its nitrogen substituted derivatives (azapentacenes) Winkler and Houk (2007). In work by Campbell *et al.* Campbell et al. (2017) and Musil *et al.* Musil et al. (2018) in which machine learning and dimensionality reduction techniques were used to investigate the properties of the crystal structure landscape of some of these compounds, a large set of crystal structures of these compounds were generated using the crystal structure prediction techniques described in the previous section. These calculations were performed using the 23 most commonly adopted space groups for organic molecules with $Z' = 1$ and the 12 most common space groups for molecules with $Z' = 2$. The crystal structure maps obtained were compared with some algorithmically calculated packing labels using a heuristic technique involving intermolecular angles. This algorithm was first designed by Campbell *et al.*

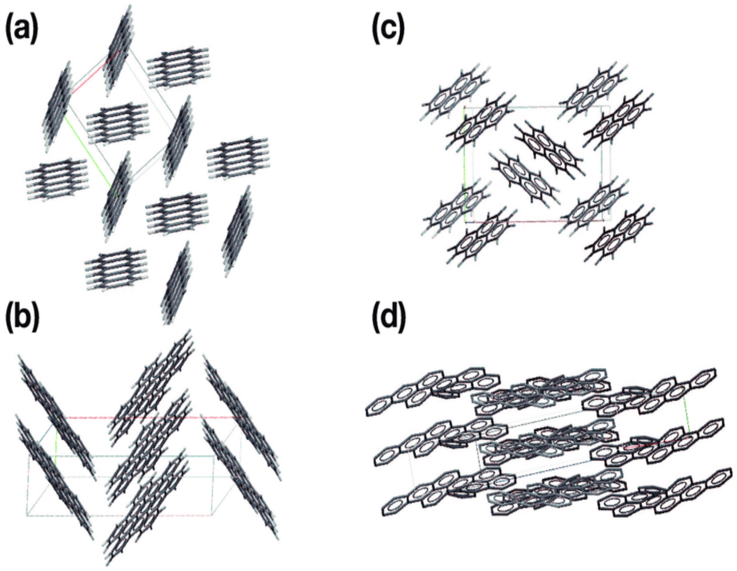


FIGURE 4.3: The four fundamental packing types for polyaromatic hydrocarbons Salzmann (2020). The structures are labelled as follows: a) Herringbone; b) Gamma ; c) Sandwich Herringbone and d) Beta

Campbell et al. (2017). This technique will be discussed in more detail later in the context of a more contemporary technique which we use later Loveland et al. (2020).

In this work we use the data from the paper of Musil *et al.* Musil et al. (2018) using the same labels as the authors. The compounds studied are shown in figure 4.4. Note that we have been kindly provided with additional structures not included in the supplementary information of the paper by the authors Musil et al. (2018). Note also that the labelling of the compounds we were given and those of the paper are not identical (5C is labelled as 5B in the paper).

We also study updated crystal structure landscapes (using more contemporary crystal structure prediction code provided by the same authors) for a subset of the compounds (pentacene, 5A, 5B and 5C) and also using updated labels which we found using the technique of Loveland *et al.* Loveland et al. (2020). Again the specifics and advantages of the new algorithm will be discussed in the results section.

Finally we also consider two small sets (with 28 and 172 compounds respectively) of experimental PAH structures,respectively described in Desiraju and Gavezzotti (1989b) and Loveland et al. (2020), the labels of which were found by inspection by a subject matter expert and the structures for which were found by querying the names or CSD codes provided in the papers on the Cambridge Structural Database (CSD) Groom et al. (2016). As such the labels in these datasets are of higher fidelity although the datasets are considerably smaller.

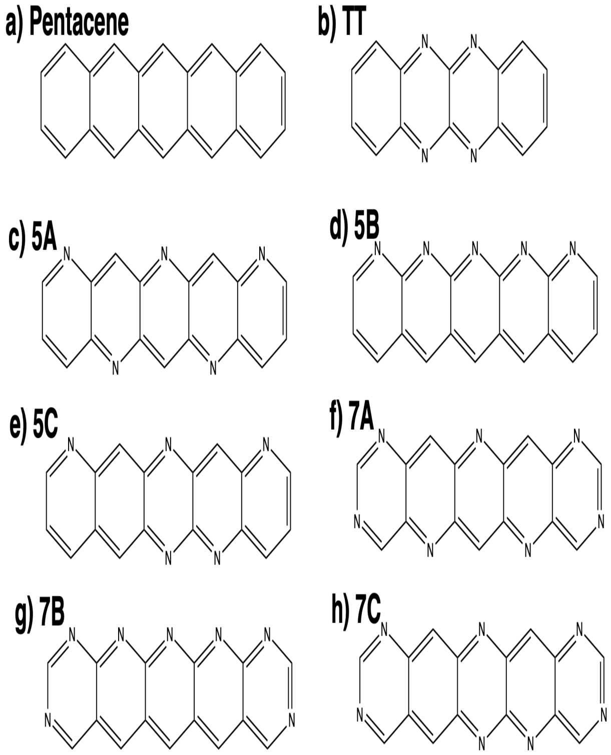


FIGURE 4.4: The azapentacene compounds used in this work.

4.3 Nicotinamide:Benzoic Acid (GAZCES) Co-Crystals

This dataset is after the paper by Yang *et al.* Yang and Day (2022). The structure of the molecules comprising the co-crystal labelled by the CSD as GAZCES (nictinomide and benzoic acid) are shown in figure 4.5. The GAZCES system was studied by Yang *et al.* due to its known polymorphism and rigid molecular structure. The aim of the study was to investigate the nature of the energy barrier between different polymorphs in the context of the global energy landscape, that is, the pathways from many different starting structures using threshold Monte Carlo methods J C Schön et al. (1996); Schön and Jansen (1996). The transition between two polymorphs of GAZCES was studied in this manner. [1]

In brief this process involves starting with a given structure (in the case of the work of Yang *et al.* on GAZCES, *both* polymorphs were chosen as starting points in different simulation tasks) and perturbing the molecules in the unit cell by a small amount to give a slightly different crystal structure with a different calculated energy. If this new energy is below a certain threshold, the lid energy, the move is allowed and this becomes the new structure for further Monte Carlo moves. Else the move is not allowed and we keep the existing structure. In this way only the local neighbourhood of the energy minimum in which the starting structure lies may be explored. The

[1] There are actually known to be *four* polymorphs of this co-crystal but only two have been resolved structurally Lukin et al. (2017).

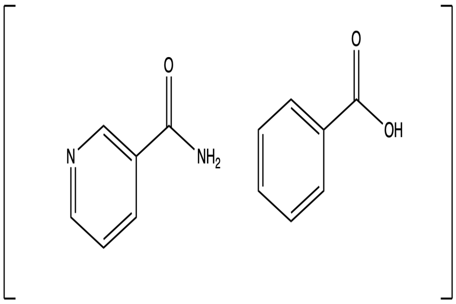


FIGURE 4.5: Structure of molecules in GAZCES co-crystal. Nicotinomide (left) and benzoic acid (right)

simulation proceeds in this manner for a certain number of steps corresponding to the resolution in which the landscape is to be explored, in the case of the work by Yang *et al.* on the GAZCES co-crystal, 3000 steps were used at each value of the lid energy Yang and Day (2022). The lid energy is then raised to allow further structures of the landscape to be explored which may lie on the other side of higher energy barriers. The process is repeated up to some maximum value of lid energy, in our case this is 150 kJmol^{-1} above the starting structures. In this way one can obtain a series of different energy basins which may be described using a disconnectivity graph Becker and Karplus (1997).

The upshot of this is that one obtains a potential energy landscape for GAZCES with two very deep basins. Each of the points on this disconnectivity graph has an associated crystal structure obtained at that Monte Carlo move. As such we have a large set of crystal structures with labels denoting two basins (which we simply call ″0″ and ″1″). In the work of Yang *et al.* conventional techniques (the SOAP REMatch kernel Bartók et al. (2013)) were used in an attempt to predict which basin a crystal structure might be associated with using only the crystal structure itself. This was unsuccessful. As such this is an interesting dataset to explore whether persistent homology can untangle these two datasets.

Chapter 5

Results

5.1 Fluorinated Benzylideneanilines

5.1.1 Pointclouds With Three Dimensions

We processed the 119 crystal structures according to the steps above in order to obtain, in the first instance, a corresponding set of vector images. In our initial approach we extracted all atoms in the crystal fragment (the fragment being obtained in a manner similar to Algorithm 1, 50 molecules were used in this case). As we shall show much better separation of crystal packing classes was achieved when only the molecular centroids were used. In figure 5.1 the Adjusted Mutual Information from a K-means clustering ($K = 8$, we are comparing the clusters obtained to the *visually* obtained packing scheme in this case and excluding the compounds labelled ”other”) on the space spanned by the set of vector images obtained using different crystal fragments: we compare crystal fragments using all atoms (using weighted persistent homology); carbon atoms only; ring centroids only and molecular centroids only. This was completed for different persistent homology filtrations and for different sets of feature dimensions for further comparisons such as the importance of crystal loops vs voids, for example. It is abundantly clear that persistent homology calculations involving sparser sampling from the underlying crystal structure yield a set of vector images that better describe the different packing types of these crystals. This is perhaps to be expected given that the persistence diagrams for the sparser crystal fragments are more likely to correspond to the useful intermolecular persistent homology features that describe the underlying packing type while the diagrams corresponding to denser sampling of the crystals contain very many intramolecular features (i.e. connected components, loops etc that live within the same molecule) which are irrelevant to our analysis: it is the positions in which the *molecules* are located that is most important. Furthermore these crucial intermolecular features may not occur at

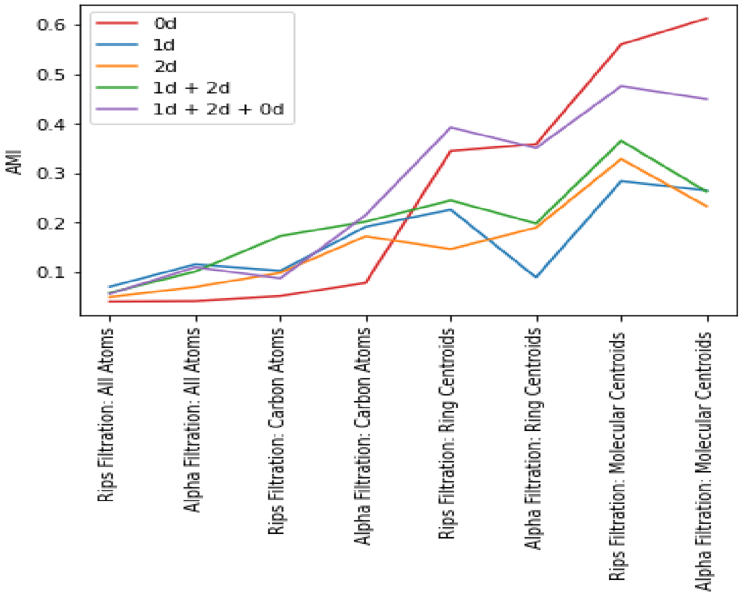
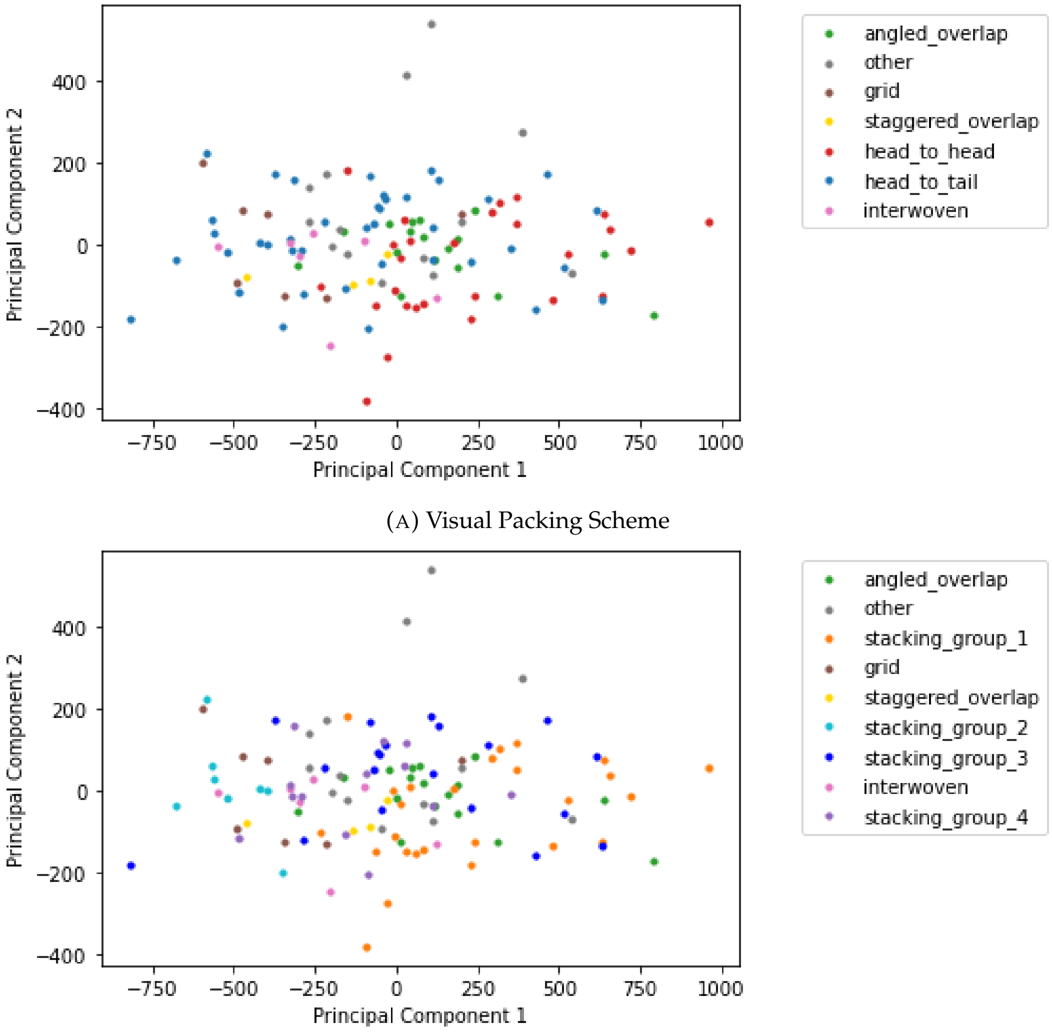


FIGURE 5.1: The comparison of different Adjusted Mutual Information scores for the clustering of a set of vector images corresponding to the persistent homology of a set of fluoroalanine crystals for different techniques for persistent homology calculation and crystal fragment generation. A high AMI indicates that the partition obtained by clustering the set of crystals is similar to that obtained by classifying the crystals according to packing types found by eye.

all on the persistence diagrams for these denser structures as these intermolecular cycles may be decomposed into smaller cycles involving nearby atoms in the manner described in figure 2.4. We also observe that the lower dimensional persistent homology features (i.e. connected components) are by far the most important in this instance. This is convenient for our purposes as these are much easier to calculate but we shall see later that this trend is not replicated in some of the other crystal systems that we investigate so it can certainly not be relied upon. It is unclear why connected components are much more important in this particular example although one reason may be that there are simply more connected components than the other higher dimensional features so there may simply be more data to work with. There does not appear to be any significant difference between the alpha and the Rips filtrations. We stick to the alpha filtration for these compounds as it is faster to compute.

To show the difference more concretely figures 5.2 and 5.3 show the set of vector images in a 2 dimensional subspace obtained by Principal Component Analysis for, respectively, the images obtained using crystal fragments using all atoms and those obtained using molecular centroids only. For each of these figures the points are labelled according to each of the two packing schemes found by Dodd *et al.* Dodd (2020) (i.e. that using visual analysis and that found using geometric arguments). It can be clearly seen here that the descriptor obtained using all atoms barely distinguishes the classes at all while the descriptor using molecular centroids separates the classes rather well. The only visually-defined classes which are not



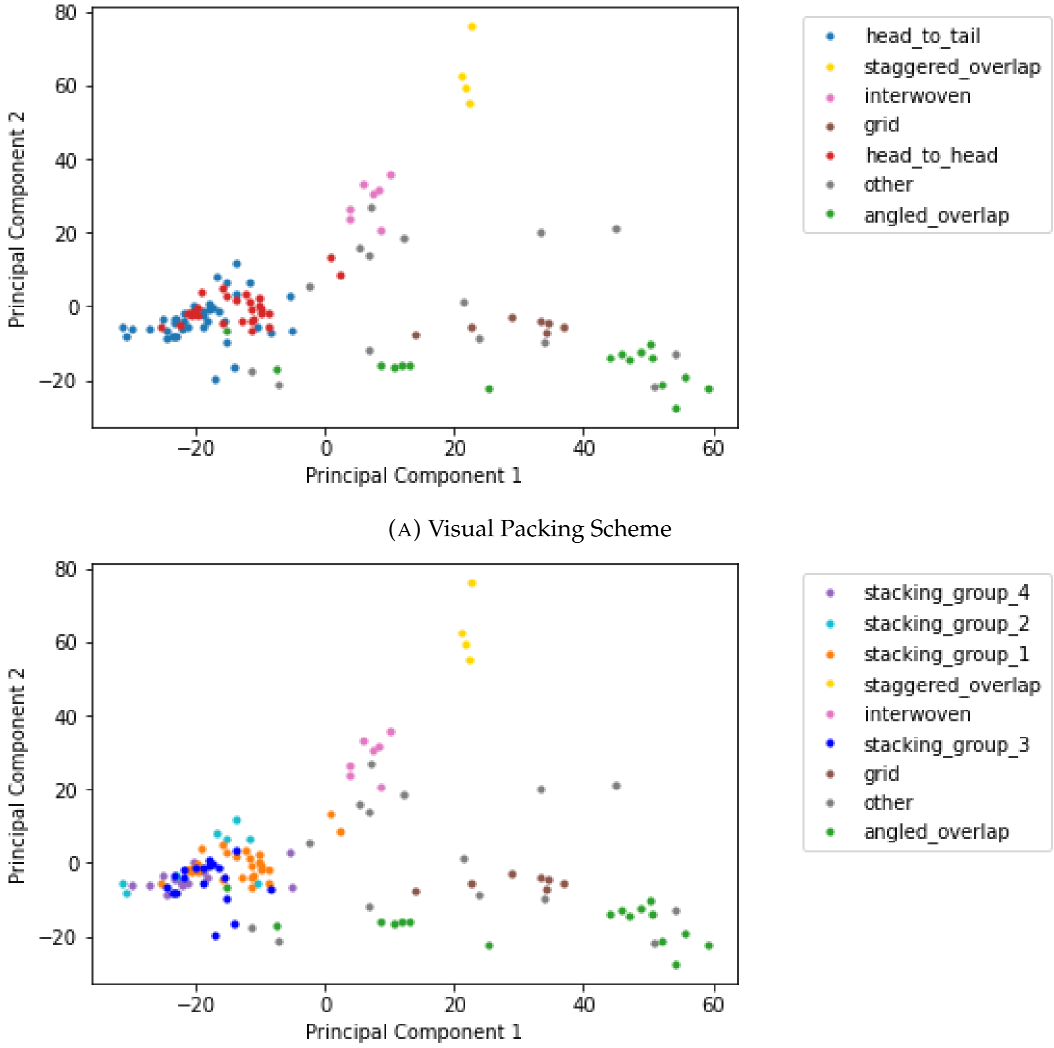
(A) Visual Packing Scheme

(B) Geometric Packing Scheme

FIGURE 5.2: Principal Components of the space of vector images (containing 0, 1 and 2 dimensional features) defined on the set of persistence diagrams of fluoroalanine compounds. Here the (weighted) persistent homology is calculated using all atoms from a crystal fragment containing 50 molecules. The structures are labelled according to the crystal packing observed either by eye or by geometric argument.

separated well are the head-to-head and head-to-tail classes. Similarly only stacking groups 1, 2, 3 & 4 are not distinguished in the case of the second packing scheme. Predictably the compounds labelled ”other” are at random positions on plots as these do not correspond to a single packing scheme; it is interesting to note that none of these compounds are significant outliers (that is, explain a significant amount of the variance of the data and cause the PCA plot to be heavily skewed) which could indicate that all of these compounds have quite similar crystal structures to the rest of the compounds insomuch that they do not ”discover” any new persistent homology features.

The vector image based approach is by no means the only way of distinguishing the

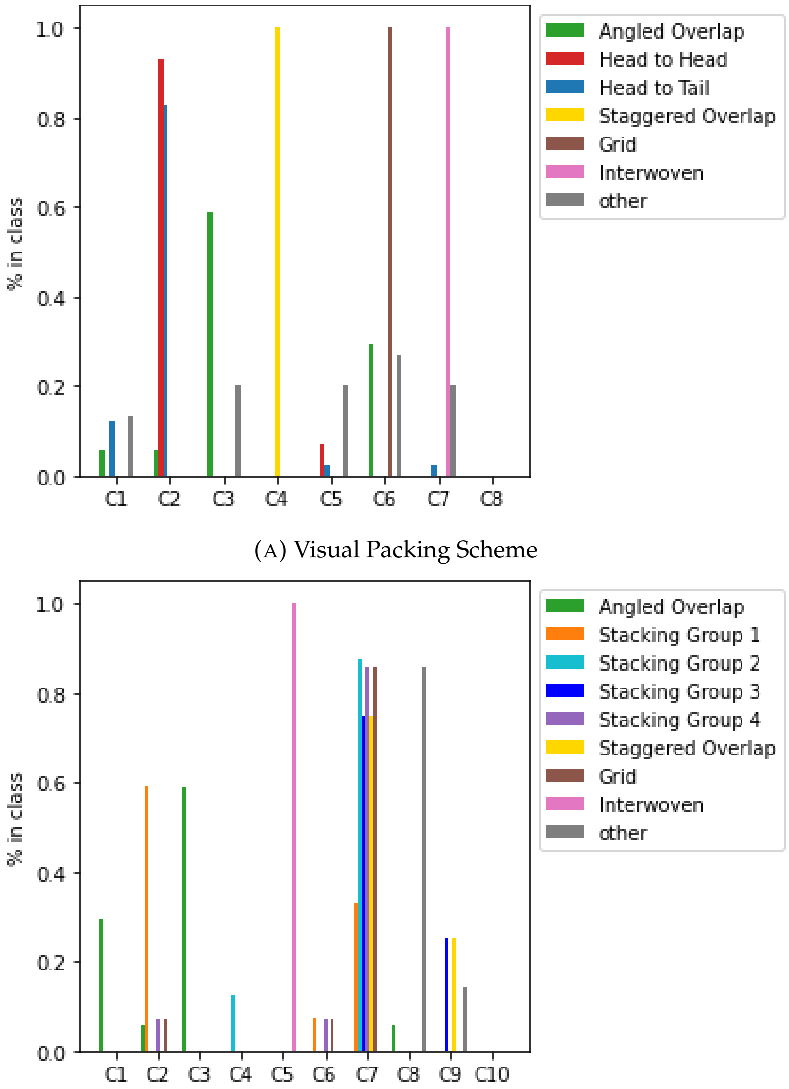


(A) Visual Packing Scheme

(B) Geometric Packing Scheme

FIGURE 5.3: Principal Components of the space of vector images (containing 0, 1 and 2 dimensional features) defined on the set of persistence diagrams of fluoroalanine compounds. Here the persistent homology is calculated using a set of molecular centroids from a crystal fragment containing 50 molecules. The structures are labelled according to the crystal packing observed either by eye or by geometric argument.

different packing types. The Wasserstein distances between persistence diagrams may be used to perform both clustering and dimensionality reduction (by Multidimensional Scaling) and in these cases we find respectively a high AMI and a good separation of the classes in the reduced space. This demonstrates that the persistence diagrams themselves serve as good representation of the packing classes - the vector images are simply a computationally convenient representation but are not necessary to the process *per se*. As discussed in the previous sections the Wasserstein distance is time consuming to compute so in practice it is not favoured by us. The MDS plots and the breakdown of the predicted clusters using K-means clustering (this time based on the Wasserstein metric) are shown in figures 5.5 and 5.4 respectively.



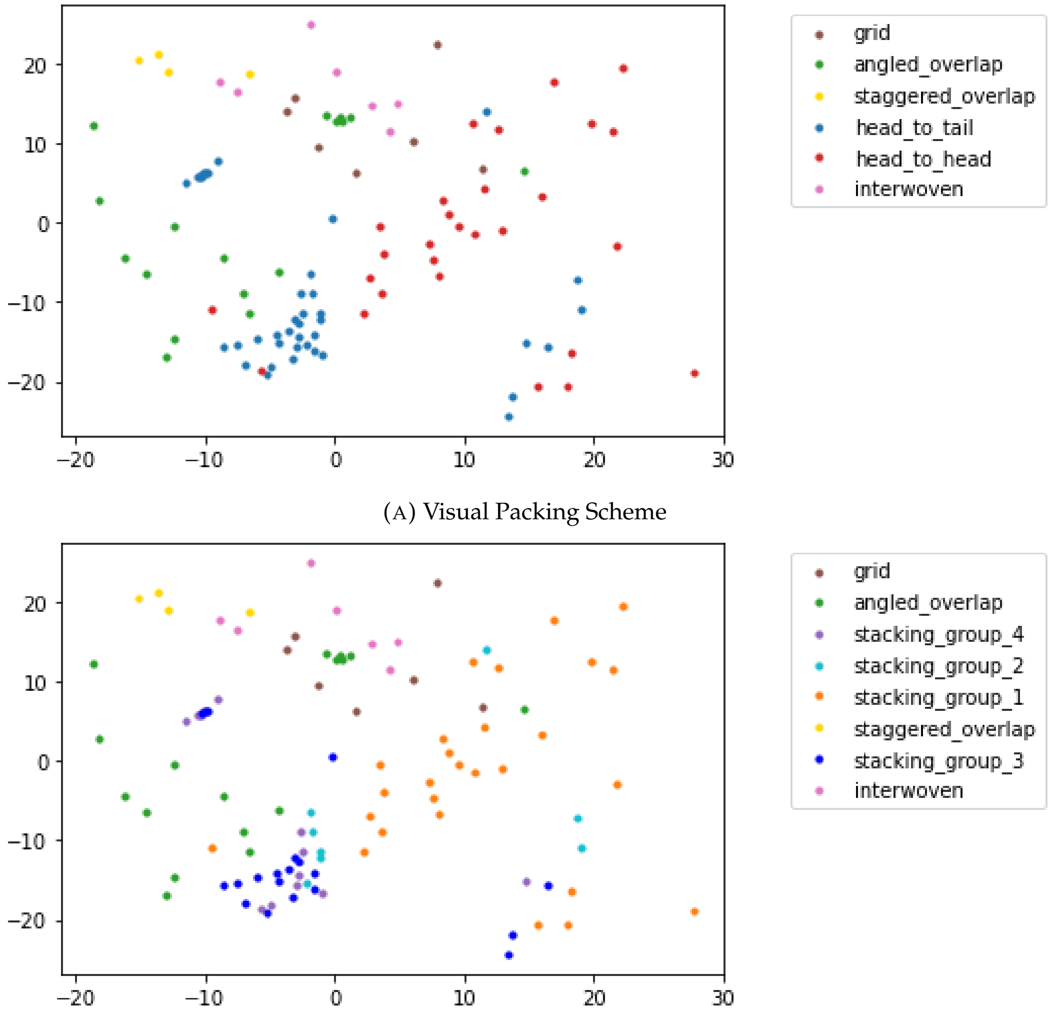
(A) Visual Packing Scheme

(B) Geometric Packing Scheme

FIGURE 5.4: Breakdown of the clusters obtained using K-means clustering using the Wasserstein distance between persistence diagrams of fluoroalanine crystal structures. The clustering is broken down by the packing types of crystal structures as obtained by both visual inspection and a geometric argument. The number of clusters to use in the K-means clustering algorithm, K, was set to match the predicted number of classes according to each packing scheme as such K is set to 8 for the visual packing scheme and to 10 for the geometric packing scheme. The clusters are labelled $C_1, C_2, ...$ and the ordering is irrelevant

Note that we do not have any meaningfully better results here although the process is technically more vigorous.

Lastly we briefly comment on the persistent homology ”invariance” phenomenon which we described in the previous sections. In figure 5.6 the AMI for the clustering based on vector images obtained using molecular centroids is plotted as function of the *number* of these centroids in the crystal fragment. We note that the AMI barely improves at all when more than around 50 molecules are used in the input crystal



(A) Visual Packing Scheme

(B) Geometric Packing Scheme

FIGURE 5.5: MDS embedding of the matrix of Wasserstein distances (containing 0, 1 and 2 dimensional features) between persistence diagrams of fluoroalanine compounds. Here the persistent homology is calculated using a set of molecular centroids from a crystal fragment containing 50 molecules. The structures are labelled according to the crystal packing observed either by eye or by geometric argument. The compounds labelled ″other″ have been removed for clarity as these are again randomly distributed across the reduced space. The embedding axes are unlabelled as they have no physical interpretation.

fragments. The presence of a cutoff past which the persistence diagrams do not change much (apart from multiplicity) and past which no new ″useful″ topological information can be obtained is consistent with the arguments and examples presented in the previous sections.

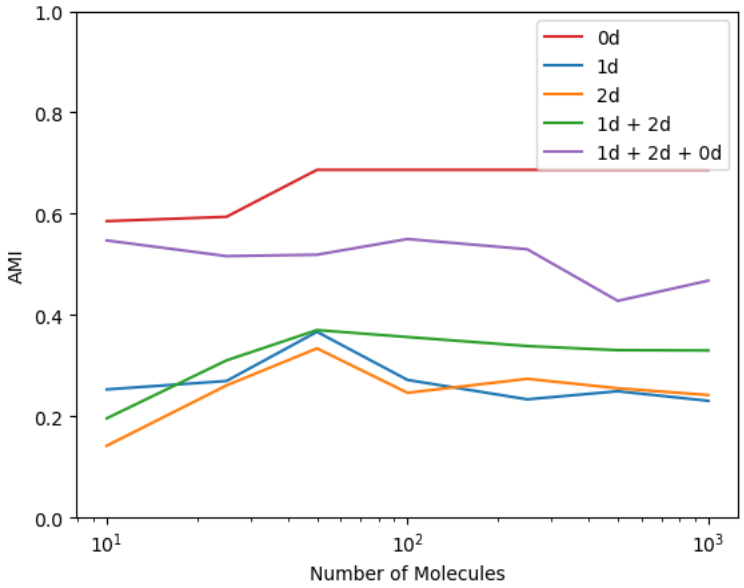


FIGURE 5.6: The variation of AMI for the K-means clustering of a set of vector images of fluoroalanine crystals with the number of molecular centroids in the input crystal fragment. A high AMI indicates that the partition obtained by clustering the set of crystals is similar to that obtained by classifying the crystals according to packing types found by eye.

5.1.2 Pointclouds With Six Dimensions

The results can be further improved by calculating the persistent homology using 6 variables rather than 3, the extra 3 variables indicating the orientation of a given molecule, that is, rather than computing the persistent homology on a set of $\{(x, y, z)\} \in \mathbb{R}^3$ where x, y and z are the coordinates of the molecular centroids, we instead compute on a set $\{(x, y, z, R_i, R_j, R_k)\} \in \mathbb{R}^6$ where R_i, R_j and R_k are respectively the x, y and z components of some vector $\vec{R}$ which describes the orientation of a given molecule. The crystal fragment hereby obtained will look similar to the vector field shown in figure 5.7, where we use the CN vector (see figure 5.8 below) on the fluoroalanine structure designated 0-0 (that is the crystal structure of the fluoroalanine with no fluorine substitutions). The idea here is that the topological spaces describing each crystal structure will also contain information about the ways in which each molecule is facing allowing us to more reliably distinguish the crystal structures involving head-to-head verses head-to-tail crystal packing. For the geometric packing scheme we are able to completely isolate the compounds corresponding to stacking group 1.

In the first instance we achieved this by setting this as the vector between the central carbon and nitrogen atoms of each fluoroalanine as shown in figure 5.8. Following a vector image approach analogous to that above we obtain the crystal structure landscape shown in figure 5.9 and get an AMI score of 0.836 (excluding the

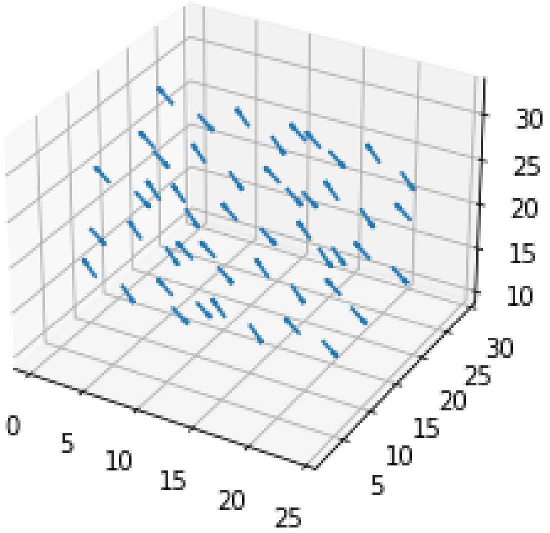


FIGURE 5.7: Example 6D crystal fragment obtained for the fluoroalanine molecule 0-0 encoding orientation using the central carbon-nitrogen vector. The x, y and z coordinates denote the positions of the centroids in 3D space while the vectors indicate the carbon-nitrogen vector (the length of which is just the length of the central carbon-nitrogen bond of the fluoroalaine). All distances should be understood to be in Ångstroms.

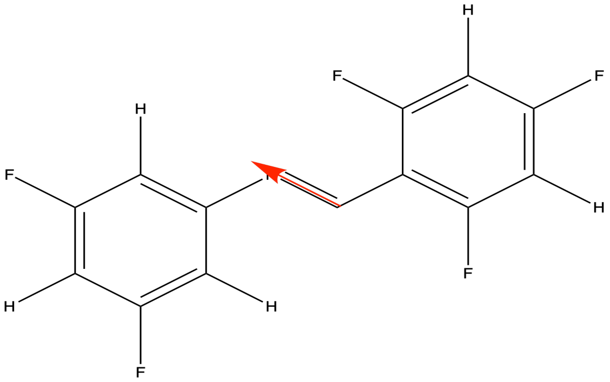
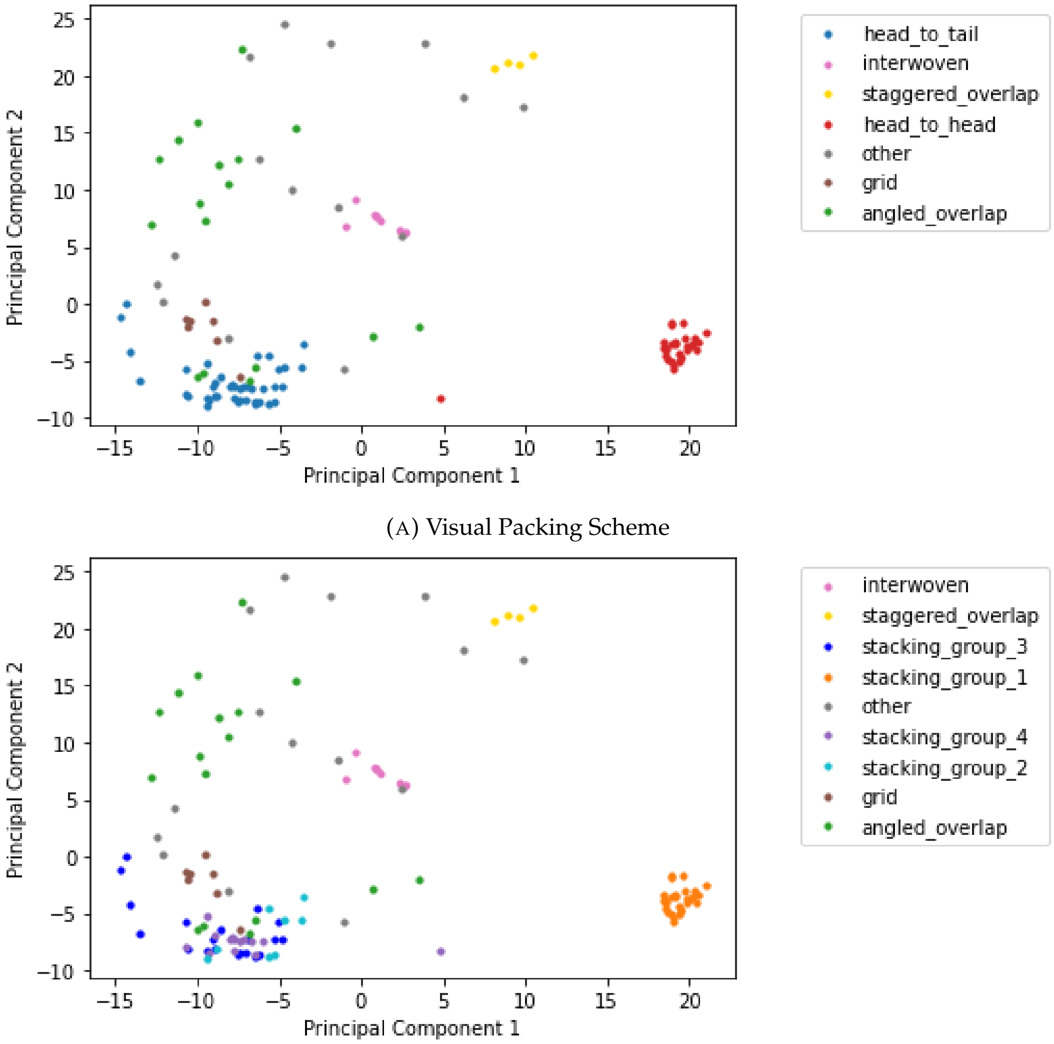


FIGURE 5.8: Example fluoroalanine molecule with CN vector illustrated

compounds labelled ″other″ for this calculation) when using the optimal set of vector images which is in this case the 0D, the 1D and the 5D set - as long as the 0D features are included the AMI is consistently higher than for the case when only the positions of the molecules are used, regardless of the combination of vector images used.

We see that the improvement is almost entirely due to the fact that the head-to-head and the head-to-tail classes are separated in our descriptor space. To illustrate why the descriptor based on the 6D pointcloud is able to separate these two packing classes while the descriptor based on the 3D pointcloud is not, consider figure 5.10. Here we only consider the 0D persistent homology features (which are all born at $\epsilon = 0$ so can be considered as a set of real numbers (deaths) rather than a set of birth-death pairs)



(A) Visual Packing Scheme

(B) Geometric Packing Scheme

FIGURE 5.9: Principal Components of the space of vector images (containing 0, 1 and 5 dimensional features) defined on the set of persistence diagrams of fluoroalanine compounds. Here the persistent homology is calculated using a set of molecular centroids *and* carbon-nitrogen vectors from a crystal fragment containing 50 molecules. The structures are labelled according to the crystal packing observed either by eye or by geometric argument.

and as such we may plot the number of homology features that occur at a given filtration value as a histogram to examine how the persistent homology changes across different packing classes more directly. In parts 5.10a and 5.10b the histograms of all persistent homology features (across all compounds with the packing type) for both head-to-head and head-to-tail packing classes are considered respectively for the case of the 3D and the 6D pointclouds. We observe that for the 3D case the prominent features for the head-to-head and head-to-tail packing classes occur at the same filtration value and hence it is difficult to readily distinguish the two classes. For the 6D case the two peaks of the head-to-head and head-to-tail histograms are well separated: this is owing to the fact that the distance (in 6D) between any two molecules in the crystal fragment is composed of the Euclidean distance between the two centroids *and* the difference between the components of the orientation vectors. More concretely the distance between two points p and q in 6D space with coordinates $(x, y, z, u, v, w) \in \mathbb{R}^6$ is simply

$$\begin{aligned} d_{\mathrm{6D}}(p,q)^2 =& (q_x - p_x)^2 + (q_y - p_y)^2 + (q_z - p_z)^2 + (q_u - p_u)^2 + (q_v - p_v)^2 + (q_w - p_w)^2 \\ =& d_{\mathrm{3D}}(p,q)^2 + d_{\mathrm{vector}}(p,q)^2 \end{aligned}$$

In the head-to-head case the vectors of nearby molecules are pointing in the same direction so this orientation contribution to the difference, $d_{\mathrm{vector}}(p,q)$, is zero so the distances between neighbouring molecules (and therefore the persistent homology features) are the same as in the 3D case. In the head-to-tail case this contribution is nonzero so that the distance between pairs of neighbouring molecules is increased so that, when we use a 6D pointcloud, we observe a clear difference between these two packing classes.

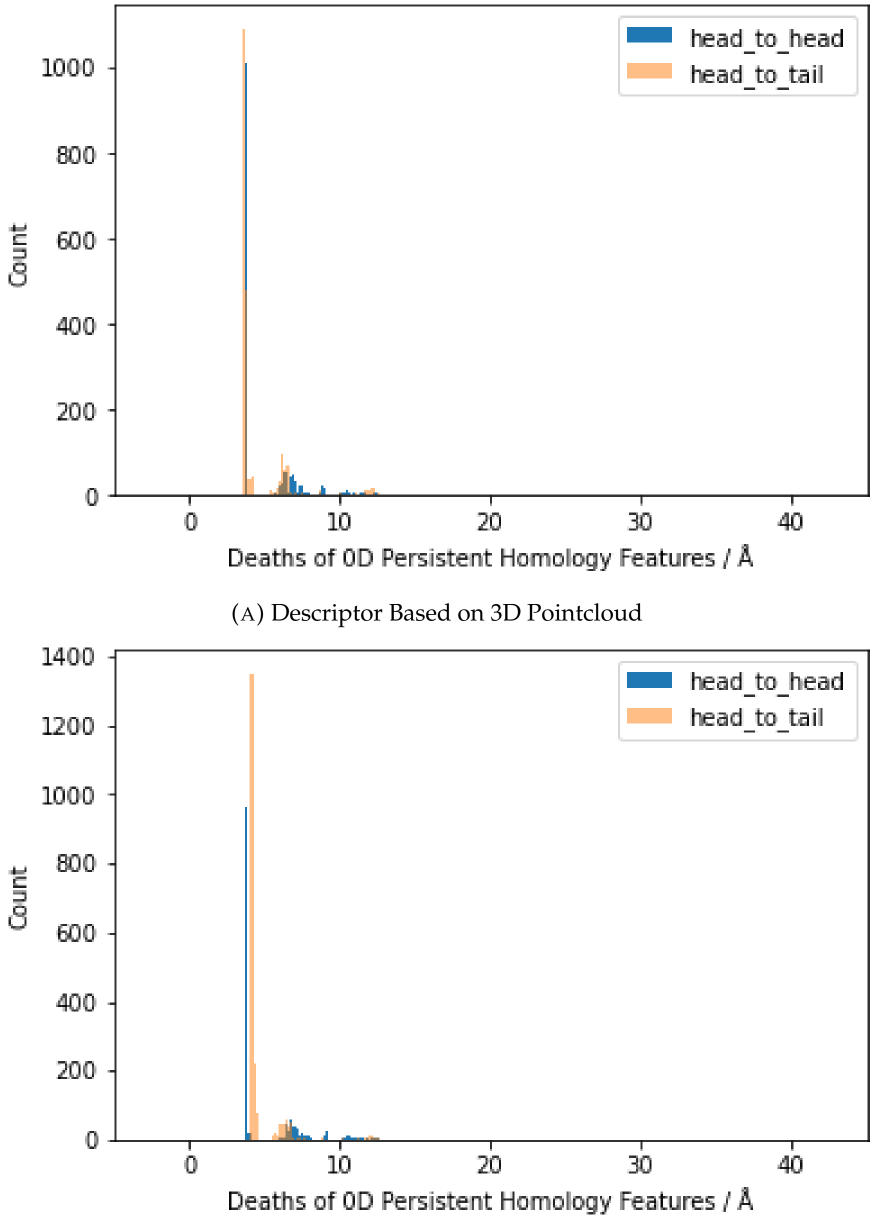
While encoding the orientation of each molecule using the carbon-nitrogen vector clearly works for the fluoroalanine compounds, this approach does not generalise to different crystal systems composed of different molecules for which there may not even be an easily defined vector between two sets of atoms which readily define the way in which the molecule is "facing". In order to address this we developed an alternative approach based on the eigenvectors of the inertia tensor. We bench-marked this technique against the fluoroalinine dataset (for which we know how a good 6D-based descriptor performs) in order to ensure that this methodology does indeed capture the desired set of molecular orientations and that we thus obtain the same separation of packing classes as above.

The inertia tensor is a 3x3 matrix which describes the moments of inertia about any axis of a rigid body composed of N point masses m_k. The elements of which are defined as follows:

$$I_{ij} = \sum_{k=1}^{N} m_k \left(\|\mathbf{r_k}\|^2 \delta_{ij} - a_i^{(k)} a_j^{(k)} \right) \tag{5.1}$$

where $\mathbf{r_k} = (a_1^{(k)}, a_2^{(k)}, a_3^{(k)})$ is the vector from the point about which the tensor is calculated to each point mass, m_k. In our case the tensor is always calculated about the molecular centre of mass so this vector is always the same as the position vector of each point mass.

By calculating the eigenvectors of this matrix we can find the principal axes of the



(A) Descriptor Based on 3D Pointcloud

(B) Descriptor Based on 6D Pointcloud

FIGURE 5.10: Histograms of 0D persistent homology features for fluoroalanine crystals with head-to-head and head-to-tail packing types. In a) the calculations were completed using 3D pointclouds using a set of molecular centroids from the crystal structures while in b) the calculations were completed with 6D pointclouds which additionally made use of the *orientation* of each molecule as encoded by the central carbon-nitrogen vector of each fluoroalanine molecule.

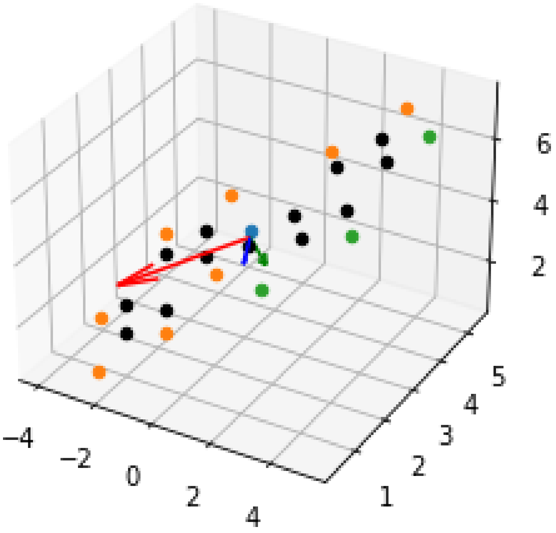


FIGURE 5.11: Example fluoroalanine molecule with inertia axes illustrated. The points denote the position of the constituent atoms in 3D space. All distances should be understood to be in Ångstroms. Note that the vectors have units kgm^2 and do not correspond to anything meaningful (cf. the carbon-nitrogen vector). In practise we often scale this vector to unity or some larger value as we shall see later.

body's rotation. Hence we can get vectors which describe the orientation for a molecule; one of these vectors may describe the long axis of a rod-shaped molecule for example. For the fluoroalanine molecule these orientation axes are illustrated in figure 5.11.

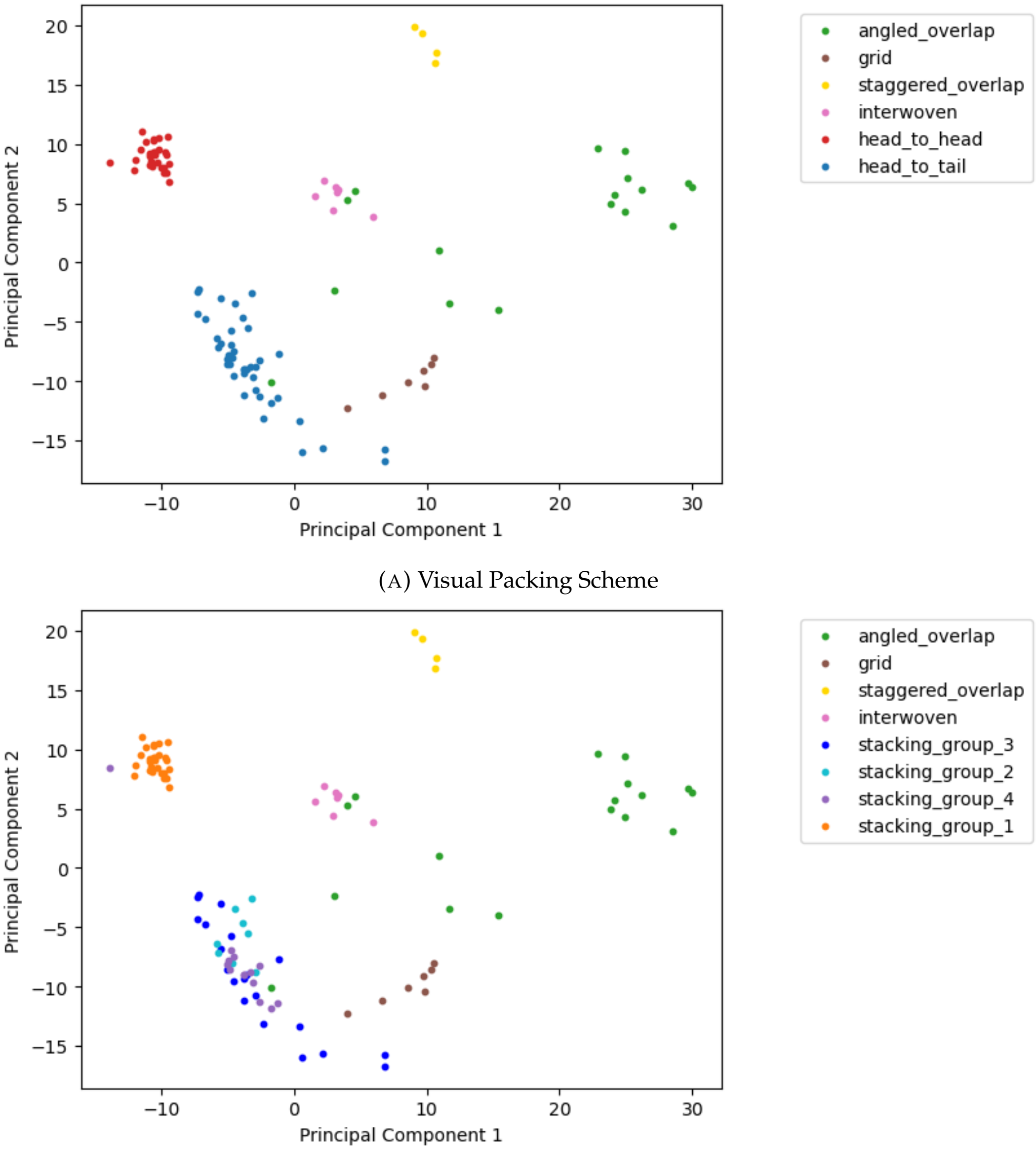
There are two issues with practically applying this method: the first is the fact that the inertia tensor is invariant to inversion (replace $\mathbf{r_k}$ with $-\mathbf{r_k}$ in equation 5.1) so we cannot define a consistent direction for each eigenvector; the second is choosing a vector from the set of three eigenvectors that best represents the molecular orientation.

For the first problem our strategy is to ensure that the inertia eigenvector points in the direction of the molecular mass gradient. That is, we can split the molecule by the plane perpendicular to the eigenvector in question at the molecule's centre of mass and count the fraction of the molecule's total mass on each side of the plane. If the mass fraction on the side in the direction of the unaltered vector is less than 0.5 then the vector is facing against the mass gradient. We then flip the vector, v, by setting $v := -v$. If the mass on each side of the boundary is equal we follow the same approach but take the sum of the mass moments on each side of the splitting plane, we seldom need to take this approach in practice however.

The second issue is due to the fact that not all molecules in the crystal packing have an identical geometry and hence that the eigenvectors and eigenvalues for the inertia tensor will all be different. This means that if we simply take the eigenvector with the largest eigenvalue in all cases we may not be selecting the same vector in all cases. In order to avoid this we take the convention in which we take the sum of the scalar products of each and every atomic position with a given eigenvector. We then choose the eigenvector with the smallest sum. The aim here is to select the vector that lies closest to the molecular plane. All the molecules considered in this work are planar, for nonplanar molecules this approach may not work. At any rate for datasets which contain only one kind of molecule (such as the large crystal structure landscapes we consider later) this problem is effectively eliminated as all molecules have the same (or very similar in the case of $Z' > 1$) geometry and we can choose any of the three eigenvectors without adversely affecting the results.

Using 6D pointclouds constructed in this manner gives very similar PCA plots to those obtained from the persistent homology calculations using the carbon-nitrogen vector. The results are shown in figure 5.12.

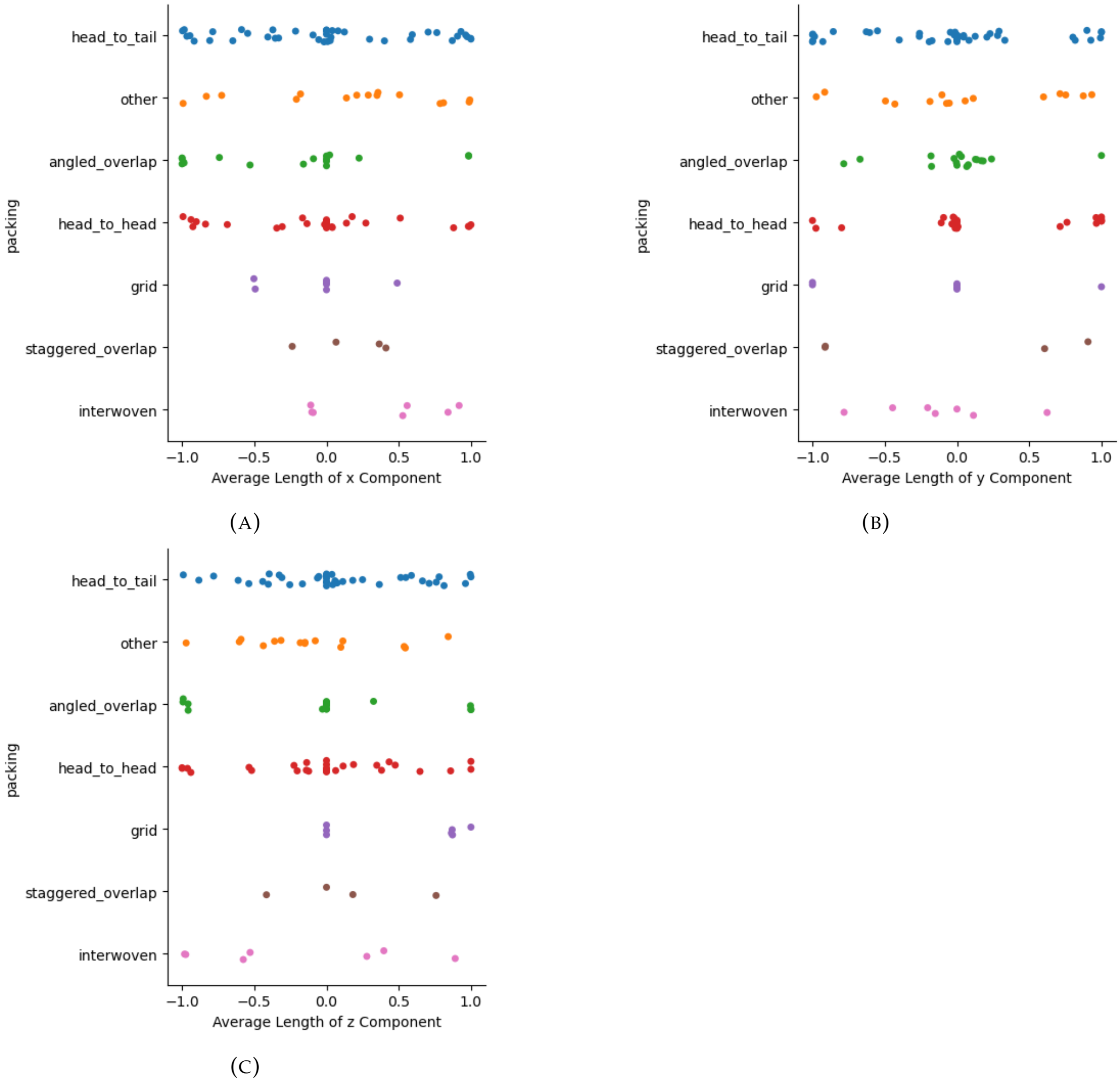
Lastly we note that the positions and the orientations are essential for successful separation of the packing classes: the set of molecular orientations on their own do not have much predictive power at all. In figure 5.13 we plot the average x, y and z components for the (normalised) CN vectors for each of the different (visually defined) packing classes. There is no clear trend in the average orientation of the fluoroalanine molecules between packing classes. Moreover figure 5.14 shows the PCA plots obtained when (Vietoris-Rips) persistent homology is applied to the matrix of scalar products between molecular orientation vectors: there is barely any separation of classes in this case.



(A) Visual Packing Scheme

(B) Geometric Packing Scheme

FIGURE 5.12: Principal Components of the space of vector images (containing 0, 1 and 5 dimensional features) defined on the set of persistence diagrams of fluoroalanine compounds. Here the persistent homology is calculated using a set of molecular centroids *and* a suitably chosen inertia eigenvector from a crystal fragment containing 50 molecules. The structures are labelled according to the crystal packing observed either by eye or by geometric argument.



(A)

(B)

(C)

FIGURE 5.13: The average x, y and z components of the central carbon-nitrogen vector of fluoroalanine molecules in different crystal packing motifs.

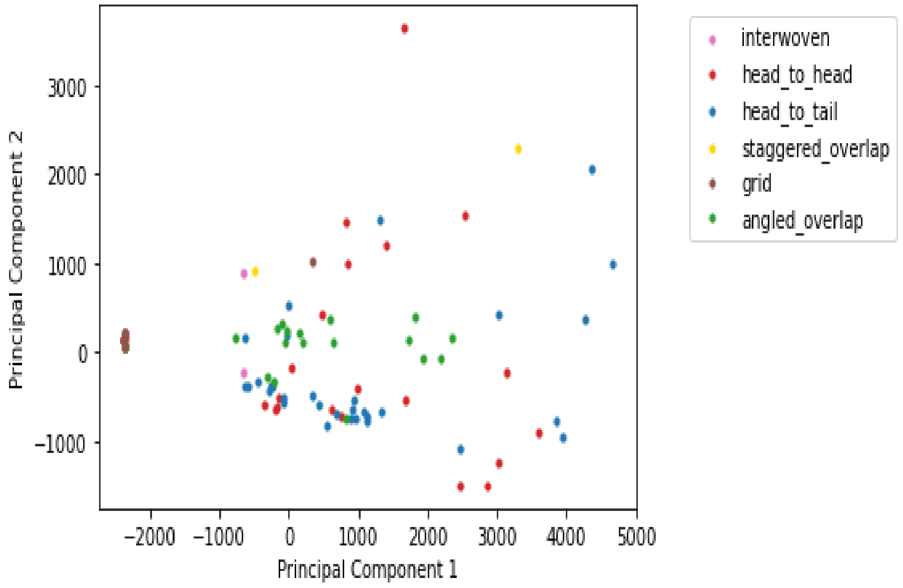


FIGURE 5.14: Principal Components of the space of vector images (containing 0, 1 and 2 dimensional features) defined on the set of persistence diagrams of fluoroalanine compounds. Here the persistent homology is calculated using the matrix of scalar products of the central carbon-nitrogen vectors between all pairs of molecules in a crystal fragment of 50 molecules via the Viertoris-Rips filtration . The structures are labelled according to the crystal packing observed by visual argument.

5.1.3 Density Based Approaches

In the previous examples the best descriptors of crystal packing class were obtained by minimising the contribution of individual atomic positions, the centroid being taken as a proxy of the position of the whole molecule in space and some molecular vector being used for a proxy of its orientation. Clearly the positions of the atoms *are* important - it is, after all, the intermolecular interactions centred at various atoms that determine the crystal packing. The atomic positions and bond lengths help constitute the shape of the molecules which is an important matter which we have not yet considered. It would hence be helpful to build a descriptor in which the atomic positions contribute to the persistent homology without breaking up important cycles or swamping out some of the more relevant features.

One possible way of achieving this is by using the technique of sublevel set persistent homology discussed in the previous sections. Rather than using the atomic positions directly we instead perform the calculation using information pertaining to the likelihood of atoms existing in a given cubic interval - the density. By tuning how small these intervals are, that is, how fine the grid that samples the density function, we may achieve a description of the crystal topology that uses all the atomic positions but does not necessarily contain too many persistent homology features and preserves the all-important intermolecular cycles.

This concept of ”density” also has clear parallels to electron density, taking the definition of the crystal structure to be one determined by the set of molecular orbitals rather than one determined by the ball and stick model, we can arrive at a picture of the crystal structure that, by definition, is not dependent on atomic positions but rather the probability of finding electrons in a given region of space - exactly the notion of ”density” we seek. Moreover the extraction of electron density surfaces is integral to the process of crystallography: the electron density being determined directly as the Fourier transform of the diffraction pattern itself Smyth and Martin (2000). The atomic positions are calculated later. Unfortunately for us this data is seldom provided in conventional crystal structure libraries and we struggled to obtain a large meaningful library of such data, particularly for the crystal structures in this work.

Instead, in order to get a feel for a how well a ”density” based approach for describing the topology of crystal structures performs, we relied on the technique of kernel density estimation which we shall describe below.

The aim of kernel density estimation (KDE) is to obtain a function which estimates the probability density function of a random variable Hastie et al. (2009); Rosenblatt (1956); Węglarczyk, Stanisław (2018); Chen (2017). In our case we are trying to find a probability density from a three dimensional random variable namely the Cartesian coordinates in our crystal structure. This is achieved by constructing a function from a set of *kernel* functions, $K(x)$, centred at each value of our known variables (i.e. the atomic positions). Given a set of N p dimensional (in our case $p = 3$) random variables $\{x_i \in \mathbb{R}^p\}$ we can write the density as

$$\hat{\rho}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K(\frac{x - x_i}{h})$$

where h is the bandwidth which controls the ”resolution” of the kernel: a smaller bandwidth will result in a ”smoother” function. [1] Note that the choice of kernel function and bandwidth in particular Hastie et al. (2009); Chen (2015) has a significant effect on the resulting density function as we shall see. There are a number of common choices for the kernel function: the functions we consider are the tophat kernel; the Epanechnikov kernel; the cosine kernel; the exponential kernel; the linear kernel and the Gaussian kernel. We define the form of these [2] in table 5.1.

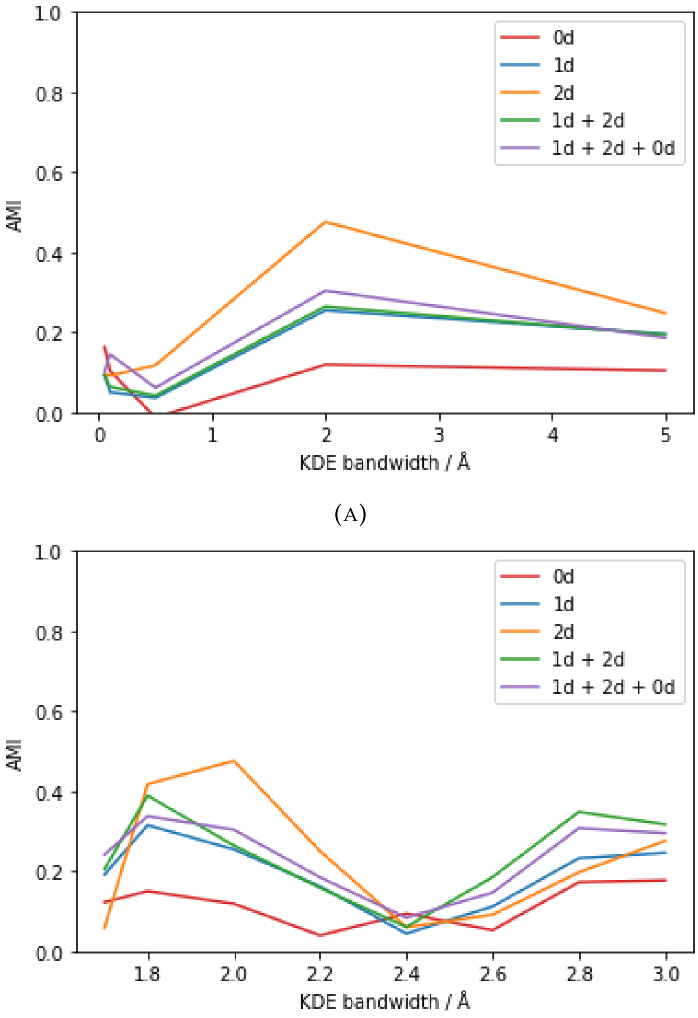
Kernel Name	Functional Form
tophat	$1 : x < h$
Epanechnikov	$1 - \frac{x^2}{h^2}$
cosine	$\cos\left(\frac{\pi x}{2h}\right)$
exponential	$e^{-\frac{x}{h}}$
linear	$1 - \frac{x}{h} : x < h$
Gaussian	$e^{-\frac{x^2}{2h^2}}$

TABLE 5.1: Mathematical Forms of Kernels used for Kernel Density Estimation

Using the technique of kernel density estimation and the sublevel set persistent homology in conjunction with the cubical complex as discussed in the methods section we obtained a set of vector images corresponding to the persistence diagrams of each the fluoroalanine crystal structures. The three main variables that need to be tuned in this case are: the kernel function; the kernel bandwidth and the number of gridpoints used to define the KDE function when constructing the cubical complex. We started by finding the optimal bandwidth for the Gaussian kernel, which is by far the most

[1] Strictly speaking in the case of multivariate kernel density estimation (that is for p > 1) the kernel need not be spherical and we can describe the bandwidth as a matrix, H, rather than a single variable Hwang et al. (1994). The kernel density is then given as $\hat{\rho}(x) = \frac{|H|^{-\frac{1}{2}}}{N} \sum_{i=1}^{N} K(H^{-\frac{1}{2}}(x - x_i))$. This reduces to the result above in the case that $H = hI$. We did not take this approach as this technique is not implemented in the software we use and using a symmetrical kernel makes more intuitive sense in our case, considering atoms to be approximately spherical.

[2] There is a proportionality constant also included in each of these but these are not provided in the scikit-learn software which we use for the calculation. The precise form of these kernels can vary between implementations.



(A)

(B)

FIGURE 5.15: Plot of the Adjusted Mutual Information of a K-means clustering on a set of vector images from persistence diagrams of fluoroalanine crystal structures as compared to packing classes defined by visual inspection against the KDE bandwidth used to construct the density function upon which the persistent homology is defined. We defined the density function using a Gaussian kernel on a grid of 100 points using the atoms of a crystal fragment of 50 molecules. Part a) shows a larger range of bandwidth while part b) shows the peak in more detail.

common choice of kernel and a nice approximation to an atomistic potential and we initially used 100 grid points. The AMI obtained for each KDE function with a given bandwidth is shown in figure 5.15. It is important to note that, in contrast to the point based homology, the 2 dimensional persistent homology features, the voids, are most important here. Also note how the optimal bandwidth is not far off the van der Waals radius of Carbon, 1.7 Å Pauling (1941), indicating that a density based approximation informed by the properties of the constituent atoms tends to have more useful topological features. This adds credence to the hypothesis that the electron density could carry useful topological properties. At this point the AMI is still lower than that obtained using the point-based approach (in both three and six dimensions).

While we did not optimise bandwidth with every kernel applying the optimal bandwidth of 2.0 to any other kernel yields poor results as shown in figure 5.16. The AMI for any non-Gaussian kernel is consistently lower than that obtained for a Gaussian kernel with a very sub-optimal bandwidth. We did not explore this issue further.

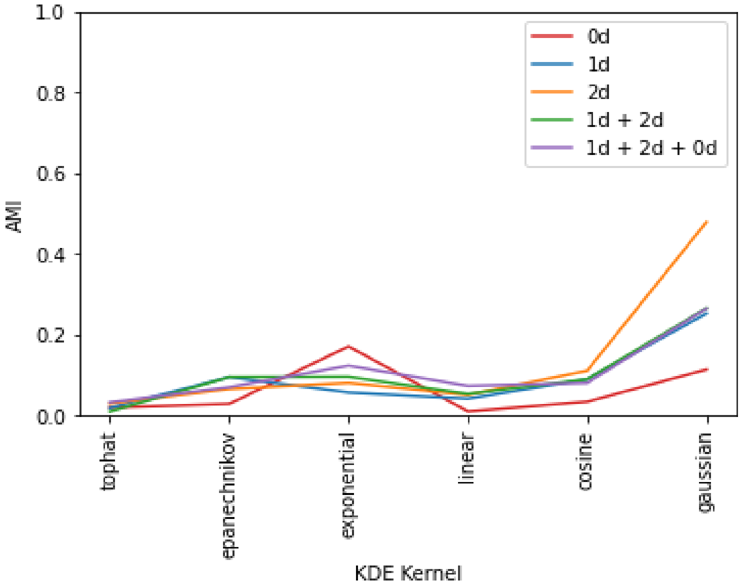


FIGURE 5.16: Plot of the Adjusted Mutural Information of a K-means clustering on a set of vector images from persistence diagrams of fluoroalanine crystal structures as compared to packing classes defined by visual inspection against the KDE kernel used to construct the density function upon which the persistent homology is defined. We defined the density function (bandwidth = 2.0) on a grid of 100 points using the atoms of a crystal fragment of 50 molecules.

When increasing the number of gridpoints we reach a similar plateau in the AMI as found when increasing the number of molecules in the point-based approach (figure 5.6) as shown in figure 5.17. Since a fairly coarse description (80 gridpoints) gives an optimal separation of packing classes it is questionable how much the finer atomic details really effect the underlying crystal topology. Moreover we note that even our best models using KDEs do not outperform the 3D centroid-based persistent homology model. The PCA representations of the two best sets of vector images (2D features only and 1+2D features only) for our best KDE model (100 gridpoints, Gaussian kernel, bandwidth = 2.0) are shown in figure 5.18. Note that while the packing classes are not as clearly well separated in space the head-to-head and head-to-tail classes are quite well separated in spite of not including any notion of molecular orientation in our model so the use of atomic positions is at least conferring *some* advantages. The other big disadvantage of our model is that the KDE functions themselves are quite laborious to compute (the KDE defined on 120 points took about 12 hours to compute, cf. a point-set persistent homology calculation that lasts the order of minutes). Therefore trying to expand this idea into a higher number of dimensions using atomic/molecular properties as we did above is almost certainly intractable. It would be desirable to compare these results with those obtained on real electron densities for the 119 crystal structures, either using experimental data or those calculated computationally, in further work.

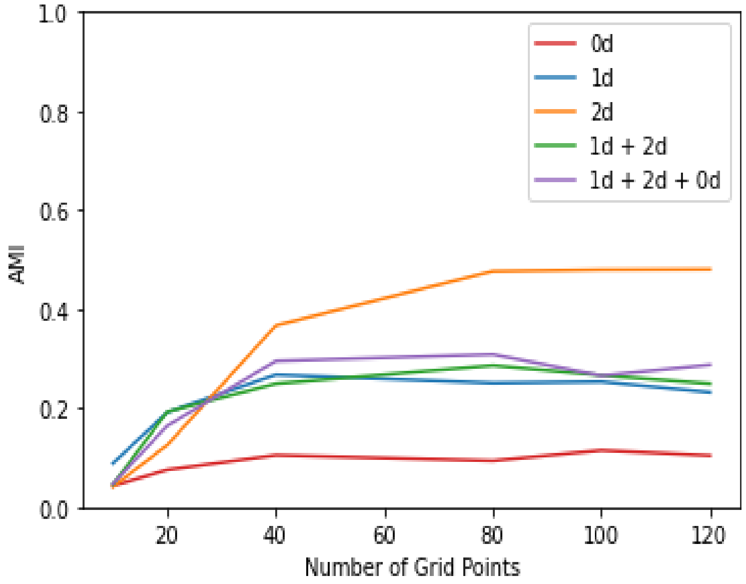
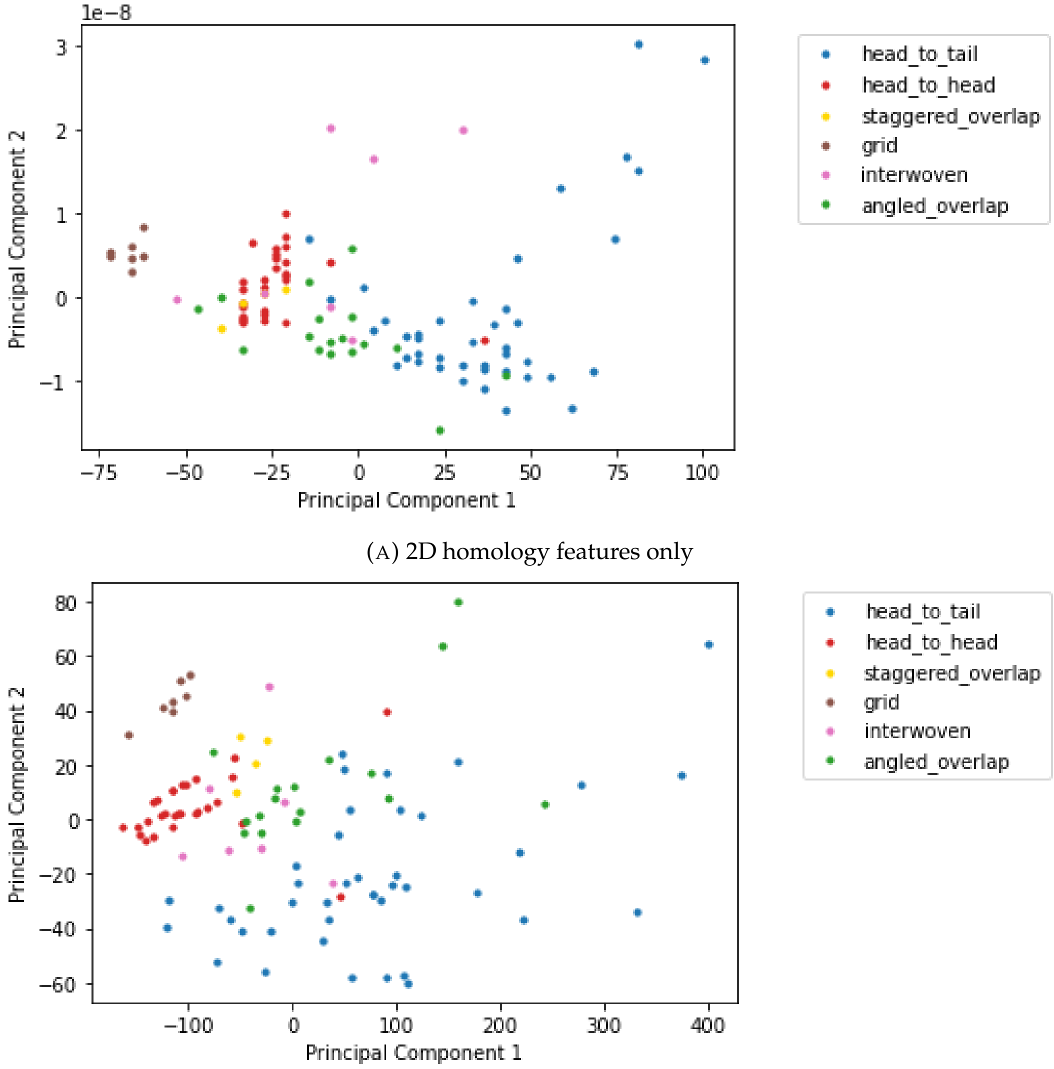


FIGURE 5.17: Plot of the Adjusted Mutural Information of a K-means clustering on a set of vector images from persistence diagrams of fluoroalanine crystal structures as compared to packing classes defined by visual inspection against the number of gridpoints used to apply persistent homology to a kernel density of a crystal fragment. We used a Gaussian kernel of bandwidth 2.0.



(A) 2D homology features only

(B) 1 and 2D homology features only

FIGURE 5.18: Plot of the principal components of a set of vector images constructed from the persistent homology of a set of fluoroalanine crystal structures using a KDE/-sublevel set based approach. The KDE estimator used a Gaussian kernel with a bandwidth of 2.0. The sublevel set persistent homology was calculated on a grid of 100 points extracted from the KDE function. The compounds are labelled according to the packing scheme assigned based on visual inspection of the crystal structures. In part a) only the vector images corresponding to to 2D persistent homology features are used. In part b) both the vector images from 1D features and 2D features are used.

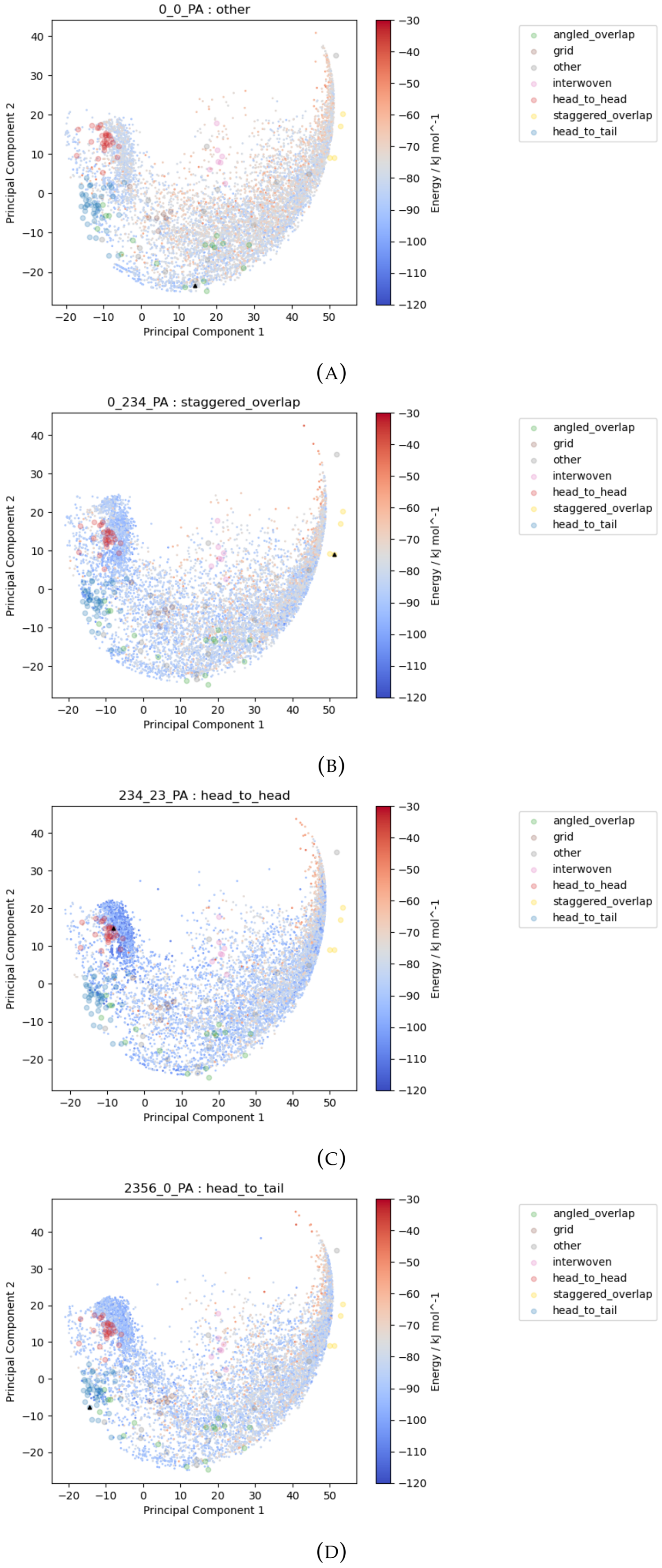
5.1.4 The Crystal Structure Landscapes of Fluoroalanines

We have already established that the set of fluoroalanine compounds is very amenable to the topological treatment showcased in the previous sections - mostly owing to the generic structure of the compounds and the well defined set of packing classes. We also augmented this dataset with some artificially generated crystal structures (see the section on crystal structure prediction). This was done for two reasons. Firstly it was interesting to view these labelled compounds within the context of a much larger set of compounds and establish whether the crystal structure predictions have similar packing classes (i.e. do not possess any sets of topological features identified as significantly different to those found in the existing structures). The second reason is to be able to use our existing knowledge of the topological features of this class of compounds (i.e. the set of vector images) to augment the crystal structure landscape of the generated structures to extenuate the features of this landscape which correspond to known packing classes. This can be done using both supervised and unsupervised methods as we shall see. In this manner a crystal structure landscape that corresponds to useful structural information may be obtained.

We used the crystal structure prediction software of Day *et al.* Case et al. (2016) using the methods outlined in the previous chapter. For each crystal structure prediction class we assumed one molecule per asymmetric unit and constrained cells to the 10 most common spacegroups Groom et al. (2016) - we were able to get around 1000 compounds in each instance. We were able to complete this calculation for 92 of the 118 experimental fluoroalanine compounds and as such 92 sets of crystal structure landscapes were generated. Owing to the large number of plots that were generated we will focus on four of the experimental compounds, the rest of the plots shall be provided in supplementary information in due course. The compounds are designated (according to the naming convention described earlier) ”0_0”, ”0_234”, ”234_23” and ”2356_0”. The compounds have respectively the packing types (visual scheme) ”other”, ”staggered overlap”, ”head-to-head” and ”head-to-tail”. In all cases the persistent homology was found using the 6D pointcloud approach with 50 molecules using a suitably chosen eigenvector of the inertia tensor. We use the Vietoris-Rips filtration so that we may only find the 0D features (the alpha filtration does not have this flexibility) in order to make this computationally tractable considering the large number of compounds that need to be processed.

In figure 5.19 we show the the first set of plots. Here we use PCA to find a subspace of the set of vector images as before. The PCA model was fit to the experimental data and the vector images obtained from the artificial structures were transformed accordingly. In this manner we can see all of the predicted structures for each compound in the context of figure 5.12a. The energies of the predicted structures are also shown to elucidate any relationship between the predicted energy of the crystal structures and its position on the crystal structure landscape in relation to those compounds with known packing types.

It is clear that both the predicted and the experimental compounds lie (almost) exactly on the same manifold in the reduced two dimensional space. This indicates that the features that explain most of the variance of the experimental structures are also present in the persistence diagrams of the computationally generated structures and that none of these structures are really different topologically when considering *these*



(A)

(B)

(C)

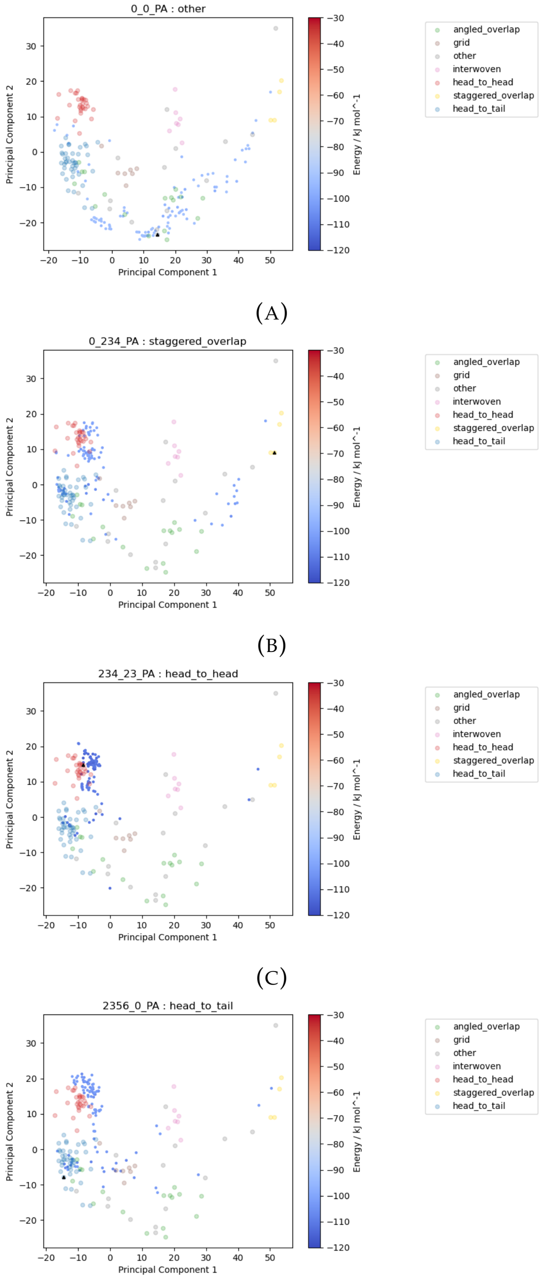
(D)

FIGURE 5.19: The crystal structure landscapes of four fluoroalanines obtained using the persistent homology of a set of experimental fluoroalanine structures with known packing class. In this case this was accomplished by transforming the vector images (obtained from the persistence diagrams) of the predicted structures according to the principal components of the vector images of the experimental structures. The energies of the predicted compounds are also labelled. The experimental compound which corresponds to the fluoroalanine which the landscape describes is indicated by the black triangle.

features at least. The plots seem to make sense intuitively as we might expect the CSP procedure to produce a large set of crystal structures which cover all the packing classes which we know about but also contains those structures that have sets of features that might constitute ”in between”, that is, the structures which could be created by a small perturbation of one of the existing packing classes. The plots have a relatively similar overall structure across the different crystal structure landscapes, the main difference (unsurprisingly) being that the ranges of energy are quite different. Examination of the four plots might lead one to consider the head-to-head and head-to-tail regions being the lowest in energy; this is in contrast to the position of the actual observed structure for each of these compounds (indicated by the black triangle). This is verified further in figure 5.20 where we show the same plots as figure 5.19 except with only the 100 lowest energy structures shown. It is possible that a better prediction of the energy of each structure (i.e. a higher level of theory) might yield better results. It is also possible that carrying out CSP calculations with $Z' > 1$ or with more space groups might yield more crystal structures which could have a lower energy and be closer to the observed experimental structure. Finally we have assumed that the fluoroalaine is completely rigid but there is actually considerable variation in the angle between the two rings of the fluoroalanine across the experimental data. This could indicate that the barrier to rotation across the central carbon nitrogen bond is small and can be overcome at room temperature in spite of the molecule being conjugated throughout. Allowing some flexibility within the constituent molecules within the CSP process might yield further structures which we have not yet explored albeit at a significant computational cost Day et al. (2007)

As suggested earlier crystal structure landscapes can be constructed that further extenuate the topological features that influence crystal packing type by using supervised dimensionality reduction techniques [3]. Our starting point was to use Linear Discriminant Analysis (LDA) as this is perhaps the most basic of these techniques and it is computationally cheap and relatively easy to interpret. As discussed previously this method performs poorly if the dimensionality of the data is much larger than the number of data (the number of data that the model if fitted on). This is the case for our data, there are 119 labelled fluoroalanine vector images each with a dimension of 400, so we need to adopt alternative approaches in order for this method to work. One easy option is to simply reduce the resolution of the vector images such that each vector image is a flattened 10x10 array rather than a 20x20 array. Using a coarser representation of the persistence diagram of course has its drawbacks with the potential loss of useful topological information when distinct features are unnecessarily combined. We see in figure 5.21 (we use compound ”0_0” as an example) that the LDA model provides meaningful results when this step is taken.

[3] When we carry out supervised dimensionality reduction techniques we take the data with the label ”other” to be unlabelled data and as such this data is not used to fit any of these models and is transformed in the same manner as those structures found using CSP.



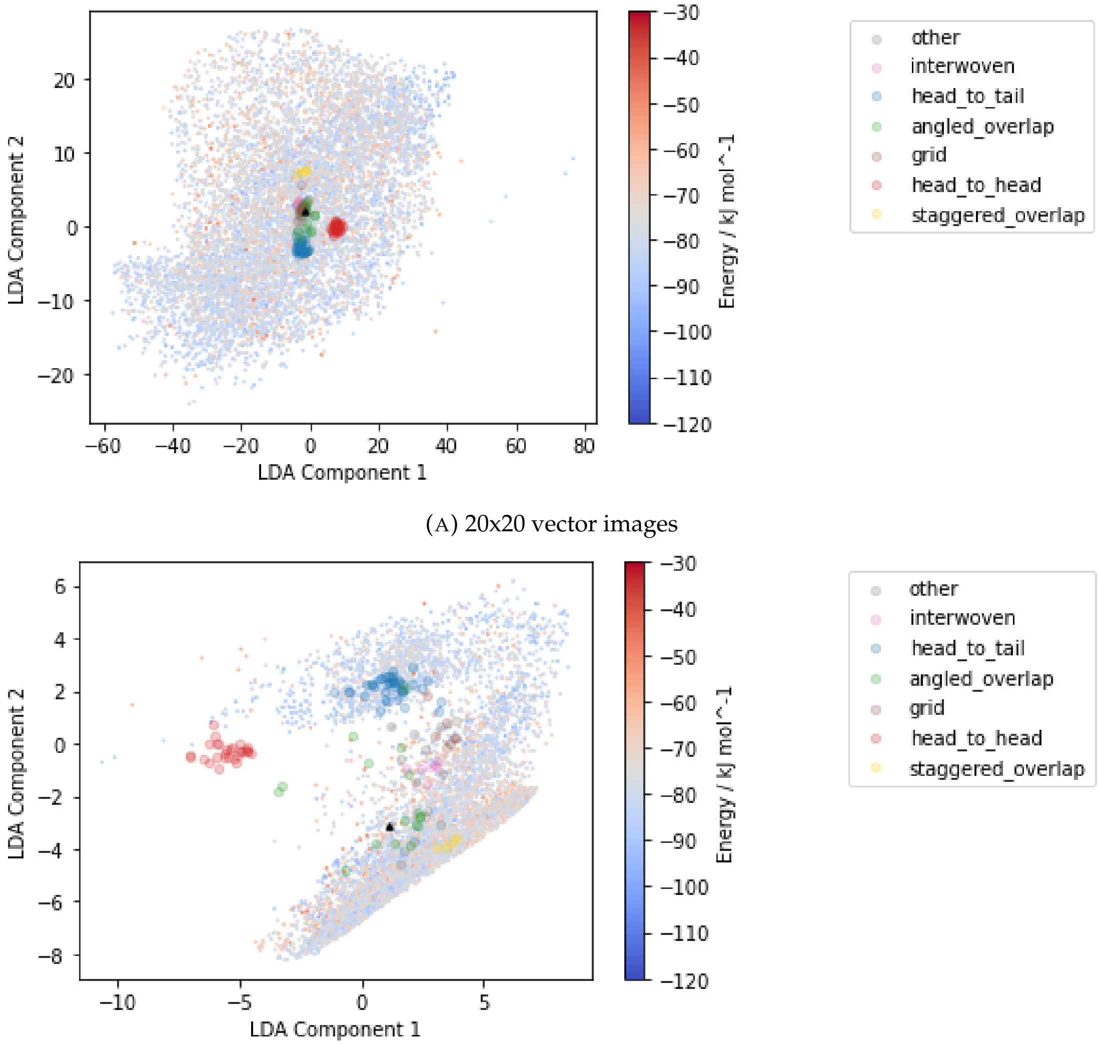
(A)

(B)

(C)

(D)

FIGURE 5.20: The crystal structure landscapes of four fluoroalanines obtained using the persistent homology of a set of experimental fluoroalanine structures with known packing class. The large points correspond to known experimental fluoroalanine structures while the small points indicate those crystal structure generated using crystal structure prediction algorithms. In this case this was accomplished by transforming the vector images (obtained from the persistence diagrams) of the predicted structures according to the principal components of the vector images of the experimental structures. The energies of the predicted compounds are also labelled. The experimental compound which corresponds to the fluoroalanine which the landscape describes is indicated by the black triangle. In these plots we only display the 100 predicted crystal structures with the lowest energy of the overall set.



(A) 20x20 vector images

(B) 10x10 vector images

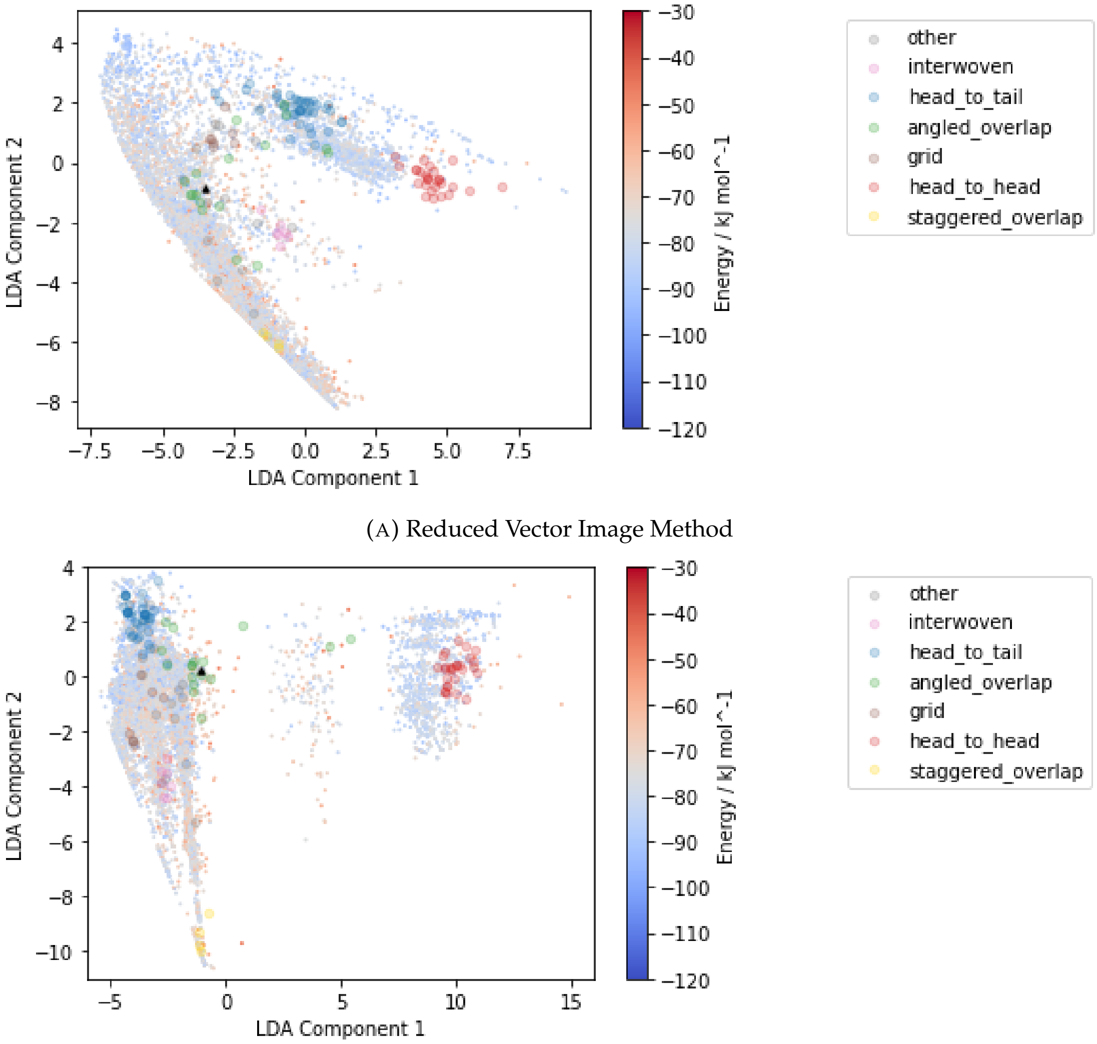
FIGURE 5.21: The crystal structure landscapes for the fluoroalanine designated "0_0" obtained by transforming the vector images using a Linear Discriminant Model fit on a set of experimental fluoroalanine structures. The large points correspond to known experimental fluoroalanine structures while the small points indicate those crystal structure generated using crystal structure prediction algorithms. In part a) we use our usual vector image resolution of 20x20 while in part b) we reduce the resolution to 10x10 such that the LDA model is fit on a dataset for which the dimensionality is less than the number of datapoints.

Figure 5.21b shows meaningful information but the experimental fluoroalanines are not much better separated than in the PCA case. Moreover the manifold containing the generated structures does not match up with the experimental structures as well in this case although this could simply be due to the fact that the differences between the topology of the experimental and computational data is clearer here and thus we can establish which packing classes are *not* being adopted.

In order to establish that the resolution loss of the vector images is not too adversely affecting the results for the LDA model, we also build LDA models fit on the experimental data using two further methods. In the first of these we simply reduce the dimension (to five dimensions) of the 20x20 vector images using PCA before the LDA model is fit. The idea here is to preserve any finer topological features which explain the variance in the set of experimental vector images but which might have been missed with the vector images with reduced resolution. For further comparison in the second method we do not use vector images at all and instead fit the models directly with the histograms of the 0D homology features with 30 bins. While we have generally found these models inferior to those based on vector images (and the 1D histograms cannot be applied to homology features with dimension greater than zero which limits the range of applicability to other systems where higher order features are more important), this makes an interesting point of contrast as the dimensionality of this descriptor is inherently lower than that of the vector image.

The two plots are shown respectively in figures 5.22a and 5.22b. In both cases the separation of the experimental fluoroalnines is not much different from the unsupervised case with the head-to-head class being further distinguished, particularly in the plot using the histogram descriptor. How the predicted crystal structures behave according to the respective transformations is the biggest difference however. Note that the manifold on which the generated structures lie matches the experimental structures more closely for the high resolution vector images which were reduced with PCA before fitting than for the lower resolution vector images. This could be due to the loss of important homology features when the resolution is lowered or it could be due to the fact that the 100 dimensional space used to fit the LDA model for the low resolution vector images is still quite large relative to the number of available data points so the model may simply be more successful for the case when the dimensionality of the dataset is *much* lower. The crystal structure landscape obtained using histograms seems to perform best especially when comparing the crystal structure landscapes using supervised LDA model (figure 5.22b) vs unsupervised PCA model (figure 5.23). It appears that applying supervised dimensionality reduction is providing a tangible advantage when histogram based descriptors are used, while the advantage of LDA in the case of vector images is less clear. The inconsistency of the results that are obtained using LDA, with the possible need to radically change our descriptor in order to get a useful crystal structure landscape means that this approach to supervised dimensionality reduction is perhaps not appropriate for our use case - even in figure 5.22b the crystal packing classes are not that well separated compared to the unsupervised approach and it is difficult to assign any of the unknown crystal structures into one category or another based on these results. We hence turn to nonlinear dimensionality reduction techniques in order to tease out further structure in our data and establish a supervised dimensionality reduction methodology that can be more easily applied to our data.

The nonlinear dimensionality reduction technique used by us is Uniform Manifold



(A) Reduced Vector Image Method

(B) Histogram Method

FIGURE 5.22: The crystal structure landscapes for the fluoroalanine designated "0_0" obtained by transforming the persistent homology descriptor using a Linear Discriminant Model fit on a set of experimental fluoroalanine structures. The large points correspond to known experimental fluoroalanine structures while the small points indicate those crystal structure generated using crystal structure prediction algorithms. In part a) The homology descriptor is a 20x20 vector image reduced in dimension with PCA to get a 5D descriptor. In part b) the descriptor is a histogram of 0D peristent homology features with 30 bins.

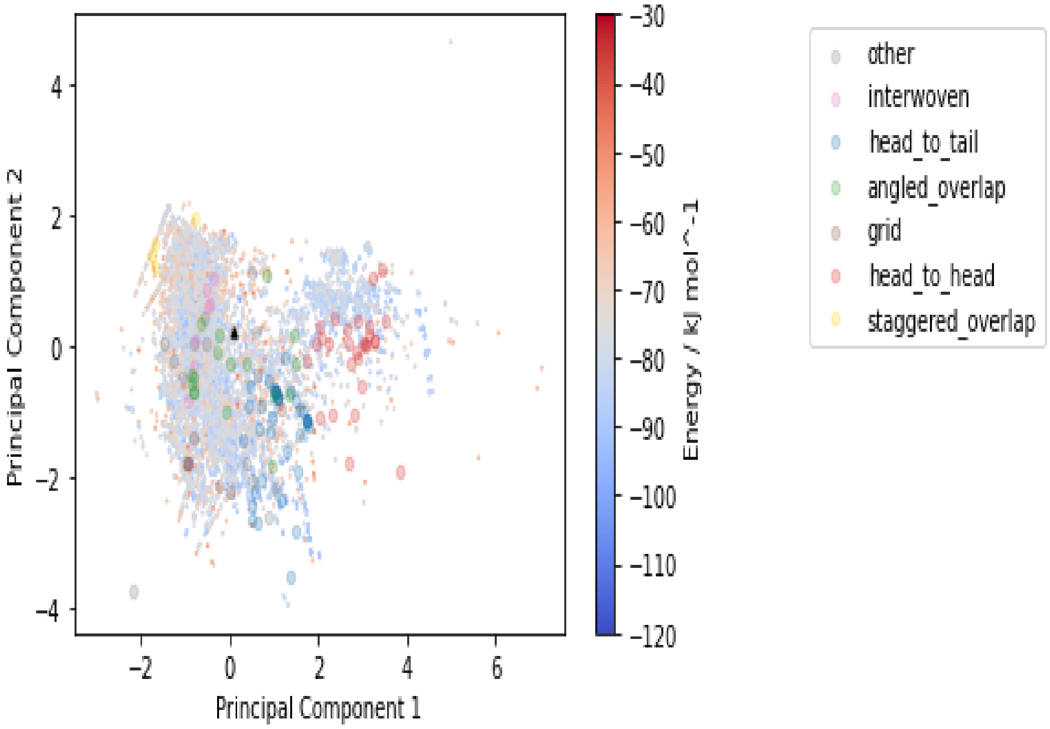
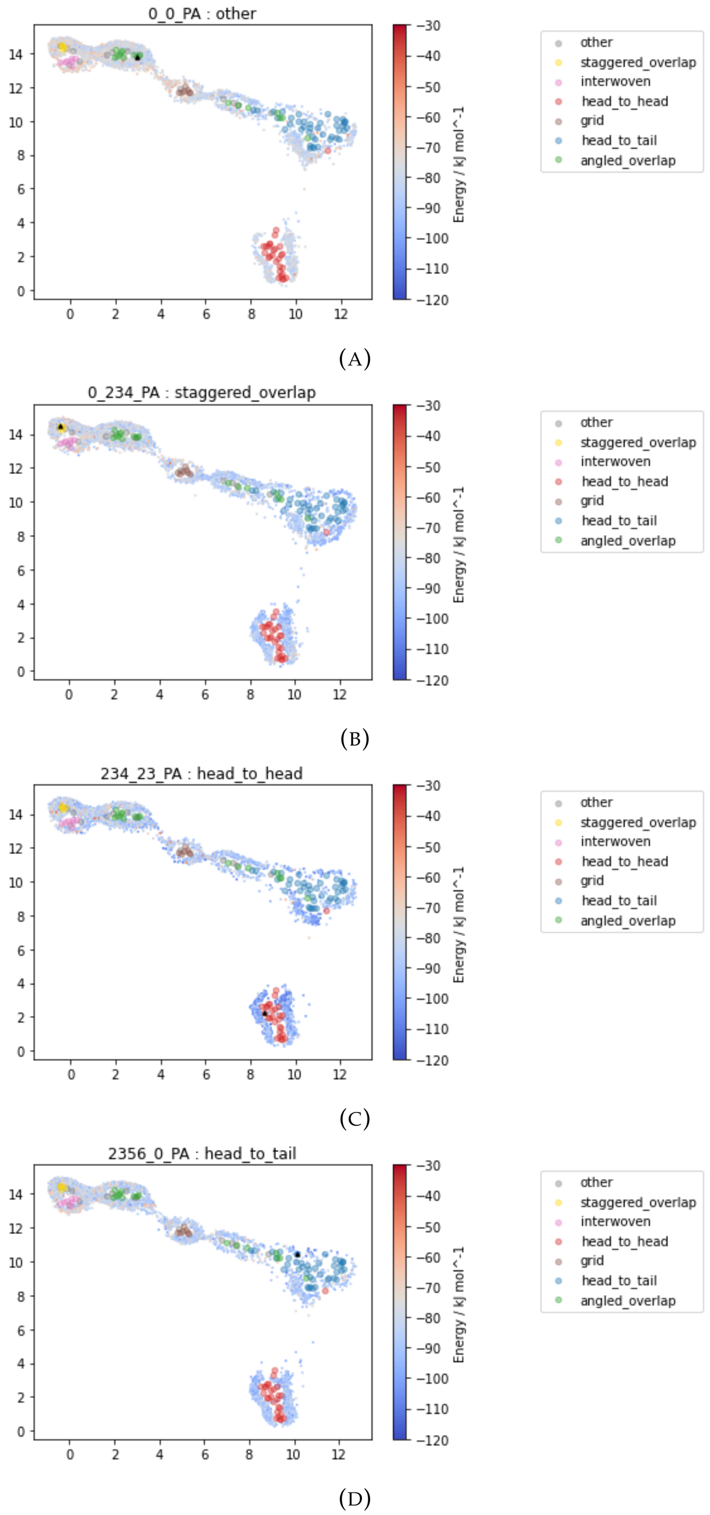


FIGURE 5.23: The crystal structure landscape for the fluoroalanine designated "0_0" obtained by transforming the histogram (30 bins) of connected components using a Principal Component Analysis fit on a set of experimental fluoroalanine structures. The large points correspond to known experimental fluoroalanine structures while the small points indicate those crystal structure generated using crystal structure prediction algorithms.

Approximation (UMAP). This algorithm has both supervised and unsupervised implementations. We first examine the crystal structure landscapes obtained when the unsupervised version of the algorithm is used. These are shown in figure 5.24 for the same exemplar structures that were used in our previous discussion. In this case the different class are more obviously separated in reduced space with the separation of the head-to-head class being the most stark. The presence of the islands of generated structures around each of the packing classes gives us a much clearer interpretation of which of the generated compounds are considered to be topologically similar to the fluorolanine compounds with a known packing type.

The partition of the classes are stronger still if the supervised dimensionality reduction algorithm is used. As explained in the methods section this just amounts to conditioning the probability weighted graph describing the data based on the known packing classes. These plots are shown in figure 5.25. What is particularly satisfying here is the fact that the compounds that lie between the two clusters occupy a single straight line implying a continuum where compounds belonging to one packing class may be slowly "deformed" into those belonging to a second packing class. We also note that the stark separation of the crystal packing classes is maintained even if only



(A)

(B)

(C)

(D)

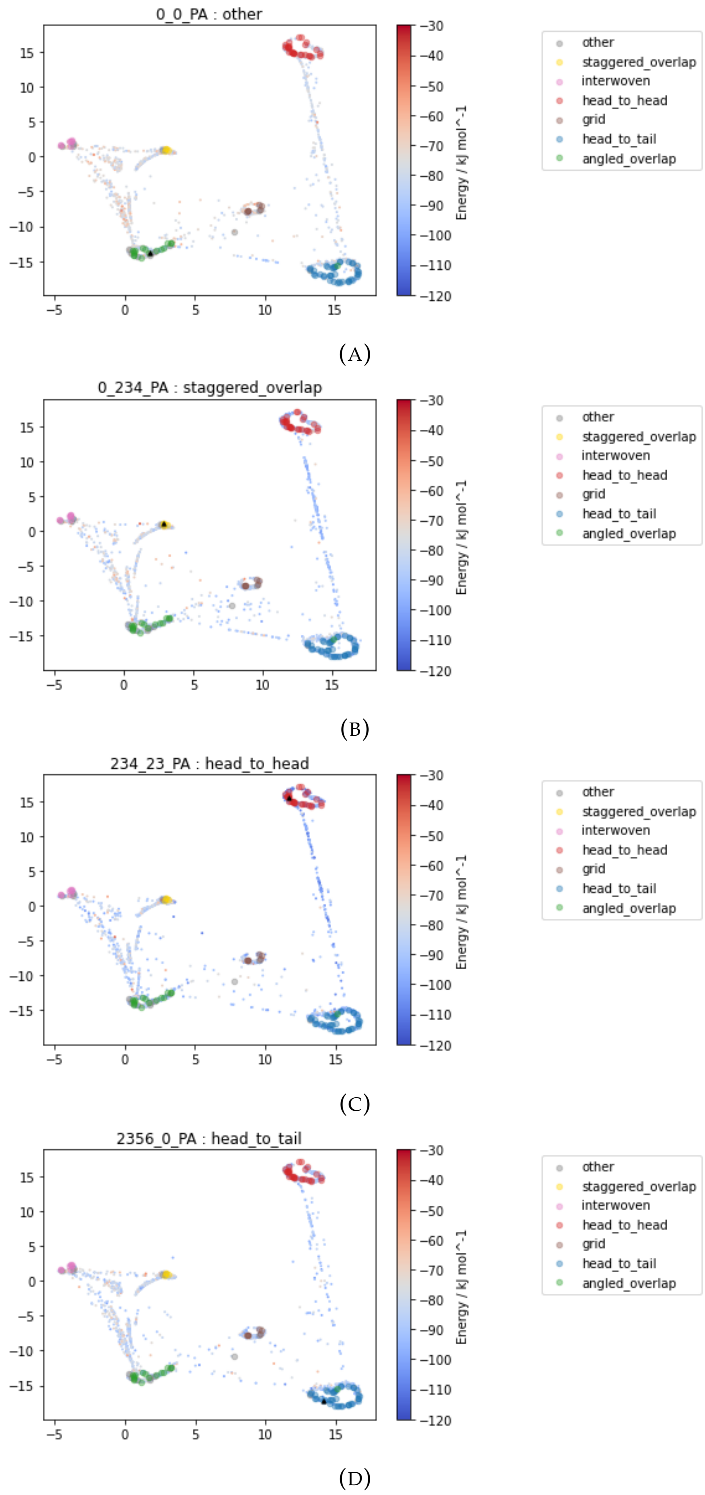
FIGURE 5.24: The crystal structure landscapes of four fluoroalanines obtained using the persistent homology of a set of experimental fluoroalanine structures with known packing class. In this case this was accomplished by transforming the vector images (obtained from the persistence diagrams) of the predicted structures according to the unsupervised UMAP algorithm. The energies of the predicted compounds are also labelled. The experimental compound which corresponds to the fluoroalanine which the landscape describes is indicated by the black triangle. The axes of the plots have no physical interpretation.

a relatively small portion of the labelled fluoroalanine structures are used to fit the UMAP model as shown in figure 5.26. Note that even in cases where 40% of the labelled data is hidden from the model, most of the known crystal structures *not* used to fit the model are assigned to the sensible regions of the crystal structure landscape which implies that there is at least some physicality in the landscapes produced with the supervised UMAP algorithm.

Owing to the relative density of the clusters it can be quite hard to visualise the relationship between the predicted energy of the crystal structures and the expected packing class. As the lowest energy structures are of the most interest we display the same plots as figure 5.25 with only the 100 compounds with the lowest energy indicated (we removed the energy scale to make the CSP data contrast with the experimental data more). This is shown in figure 5.27.

Inspection of these plots appears to suggest a tendency for the crystal structures of lowest energy to be located at any one of the clusters - i.e. have a similar topology to a packing type identified experimentally. There seem to be comparatively fewer structures of low energy that exist in the space between the clusters which do not have any apparent topological similarity with an existing packing class. Can we verify this visual trend quantitatively? For each crystal structure landscape the CSP data is transformed in exactly the same way, that is, according to the model trained on the experimental fluoroanaline crystal structures. This means that the clusters in figure 5.25 and 5.27 will be located in the same position in the reduced space for each and every crystal landscape. We can use the DBSCAN clustering algorithm [4] to partition the data into either data belonging to a particular cluster or no cluster at all as shown in figure 5.28. The centroids of each cluster and a characteristic radius were found for each of these clusters such that we could classify the predicted structures without having to fit a suitably parameterised clustering model in each case: if a crystal structure lay within some characteristic radius (defined empirically for each cluster) of that cluster′s centroid the structure would be assigned the label of that cluster. In this way we can directly compare the energy of a predicted crystal structure and its predicted packing class (or lack thereof) based on the UMAP landscape. In figure 5.29 the energies of the predicted crystal structures for each of our four examples are plotted as a histogram partitioned by the predicted packing class. Unfortunately no clear trend in the unclassified compounds towards higher energy is observed with most of these compounds being of middling energy rather than high energy. This trend is repeated across the histograms not shown here. For the compounds ″234_23″ and ″2356_0″ it appears as though the lowest energy regions of the histogram are dominated by compounds with the same packing class as the experimental compound. To investigate this further we also split the data into ten tranches of

[4] This was parameterised by inspecting the plot and ensuring that the clusters the algorithm came up with matched those expected. The parameters were eps = 0.4 and min_samples = 7. We only labelled the clusters which correspond to a packing class and manually labelled the other clusters as unclassified.



(A)

(B)

(C)

(D)

FIGURE 5.25: The crystal structure landscapes of four fluoroalanines obtained using the persistent homology of a set of experimental fluoroalanine structures with known packing class. In this case this was accomplished by transforming the vector images (obtained from the persistence diagrams) of the predicted structures according to the supervised UMAP algorithm. The energies of the predicted compounds are also labelled. The experimental compound which corresponds to the fluoroalanine which the landscape describes is indicated by the black triangle. The axes of the plots have no physical interpretation.

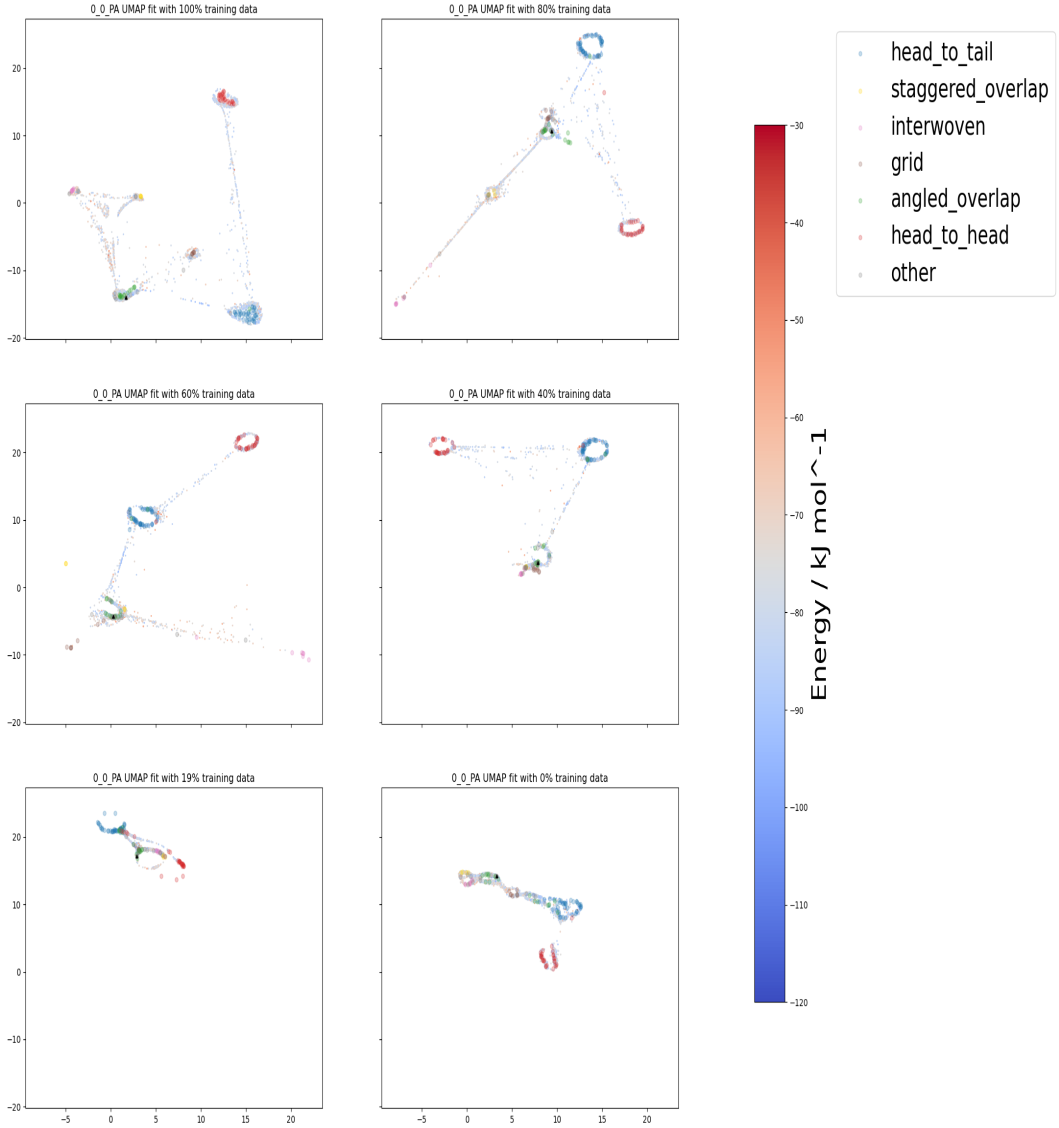
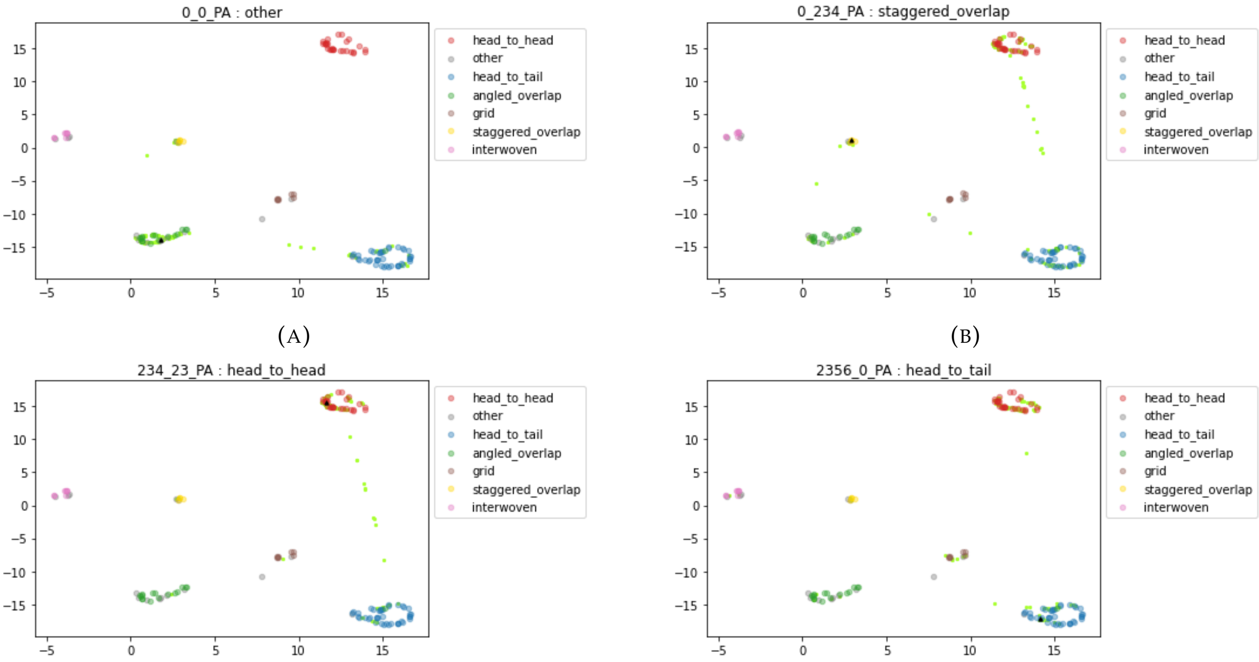


FIGURE 5.26: The crystal structure landscapes for the fluoroalanine designated ”0_0” obtained by transforming a set of vector images using the supervised UMAP algorithm on a set of experimental fluoroalanine structures with a varying portion of the known experimental data being used to fit the model (the rest being transformed in the manner of the unknown data). The large points correspond to known experimental fluoroalanine structures while the small points indicate those crystal structure generated using crystal structure prediction algorithms. The axes of the plots have no physical interpretation.



(A)

(B)

(C)

(D)

FIGURE 5.27: The crystal structure landscapes of four fluoroalanines obtained using the persistent homology of a set of experimental fluoroalanine structures with known packing class. In this case this was accomplished by transforming the vector images (obtained from the persistence diagrams) of the predicted structures according to the supervised UMAP algorithm. The energies of the predicted compounds are also labelled. The experimental compound which corresponds to the fluoroalanine which the landscape describes is indicated by the black triangle. Of the generated crystal structures only the structures with the 100 lowest values of predicted energy are shown. The axes of the plots have no physical interpretation.

increasing energy by calculating percentiles. In figure 5.30 we plot the ratio of compounds of a given packing class in each tranche and compare this to the ratio of packing types for all the data. In this way we can analyse whether structures of a given packing type are over- or underrepresented at a given energy range. We now see that only the compound ”234_23” over-represents the experimentally observed packing class at lower energies. We also don’t tend to see any over-representation of the ”unclassified” crystal types at higher energies as might have been expected in figure 5.27. We do not, in general, observe any of these trends in the other 88 crystal structure landscapes either.

It is possible that if these calculations were repeated with a more accurate CSP calculation (that explored more space groups and asymmetric cell types) and the final crystal energy were found at an higher level of theory the trends in packing class might be more reflective of experimental data. It could also be that the crystal structure landscapes generated with the UMAP algorithm do not appropriately represent the topology of the packing types and that some of the clusters are artefacts or at the very least contain more data points than correspond to the desired packing

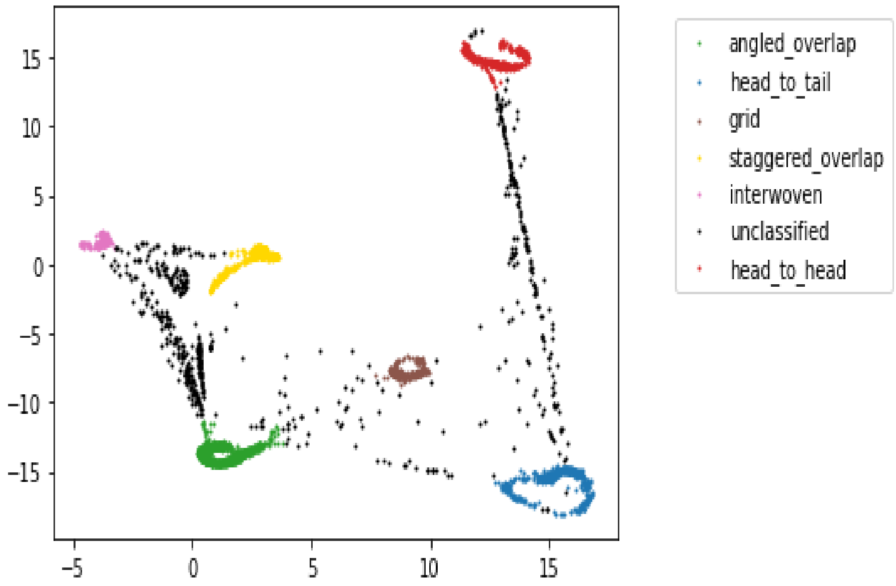
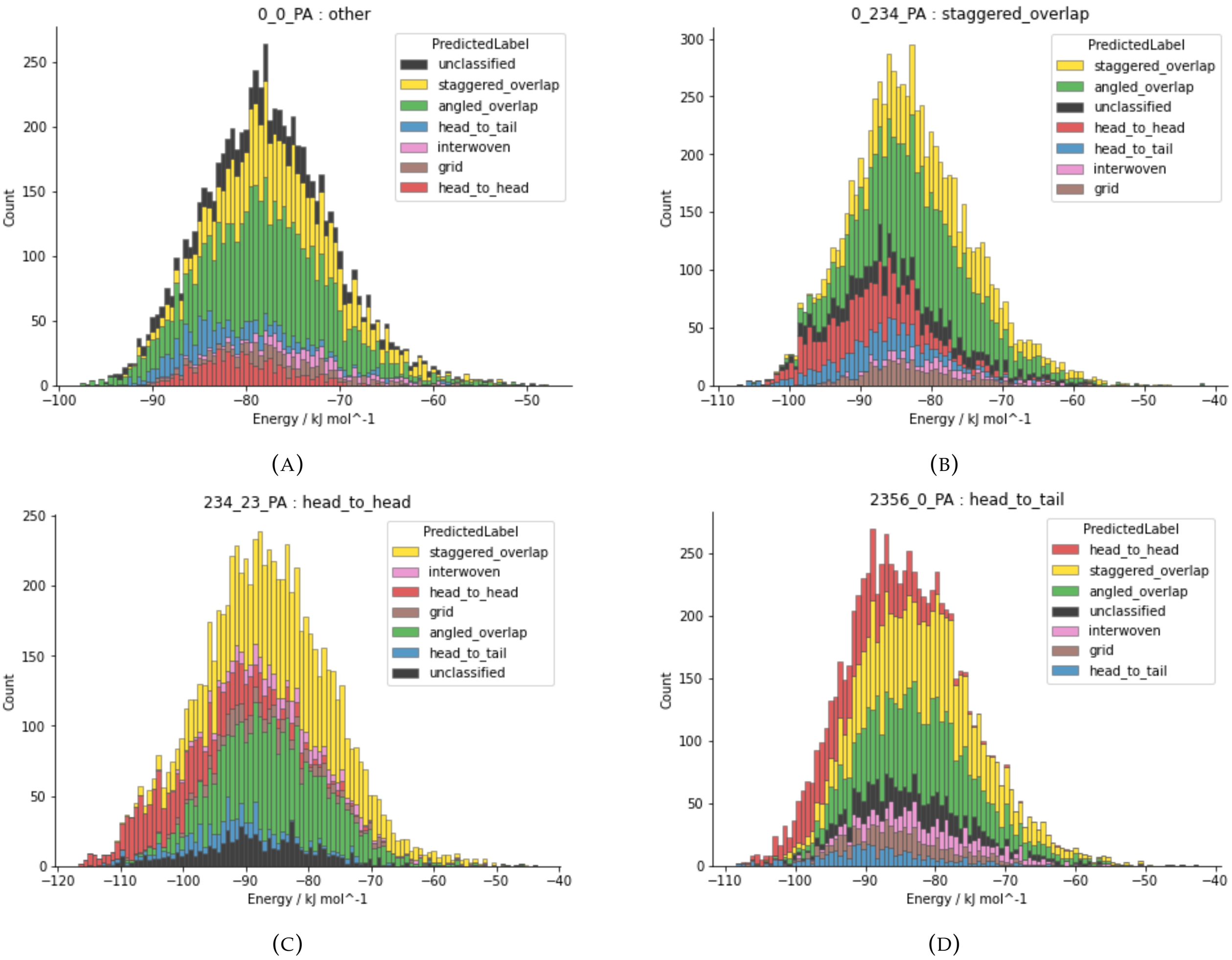


FIGURE 5.28: Classification of the crystal structure landscape (supervised UMAP model) of the fluoroalanine labelled ”0_0” by DBSCAN (eps = 0.4 and min_samples = 7). The axes of the UMAP landscape do not have a physical interpretation.

type. While the semi-supervised learning example of figure 5.26 certainly suggests that packing classes are being assigned to the right places it is important to emphasise the non-linearity of the UMAP plot, that is, the sense of ”closeness” varies significantly across the landscape, the implication being that structures which are quite different could be very close to each other in some parts of the landscape while some structures which are quite similar could be relatively far apart. If some features which do not correspond to a given packing class are ”pulled” into one of the clusters or, indeed, if some features which should be unclassified are not assigned as such *because of* this phenomenon, this could radically affect the conclusions one makes about the relationship between energy and packing class which we described above. We demonstrate this non-linearity in figure 5.31 where we plot the clusters obtained using DBSCAN/UMAP for compound ”0_0” on the PCA plot. We contrast this with the classification obtained from attempting to predict the packing class from the vector images using a support vector classifier [5]. The labels obtained using the SVC are plotted on both the crystal landscape obtained using PCA and the crystal landscape found using supervised UMAP. We see that when the UMAP crystal structure landscape is calculated many very disparate sections of the manifold are in effect cleaved apart. There is scope for error here because compounds could be pulled from the wrong regions and clusters that look very far apart in the UMAP plot could in fact be very similar and have strongly overlapping topological features. This also makes the inherent non-linearity of the UMAP method abundantly clear which can make the

[5] Vector images were converted to standard scalars before classification. We used a linear kernel and C was set to 100.



(A)

(B)

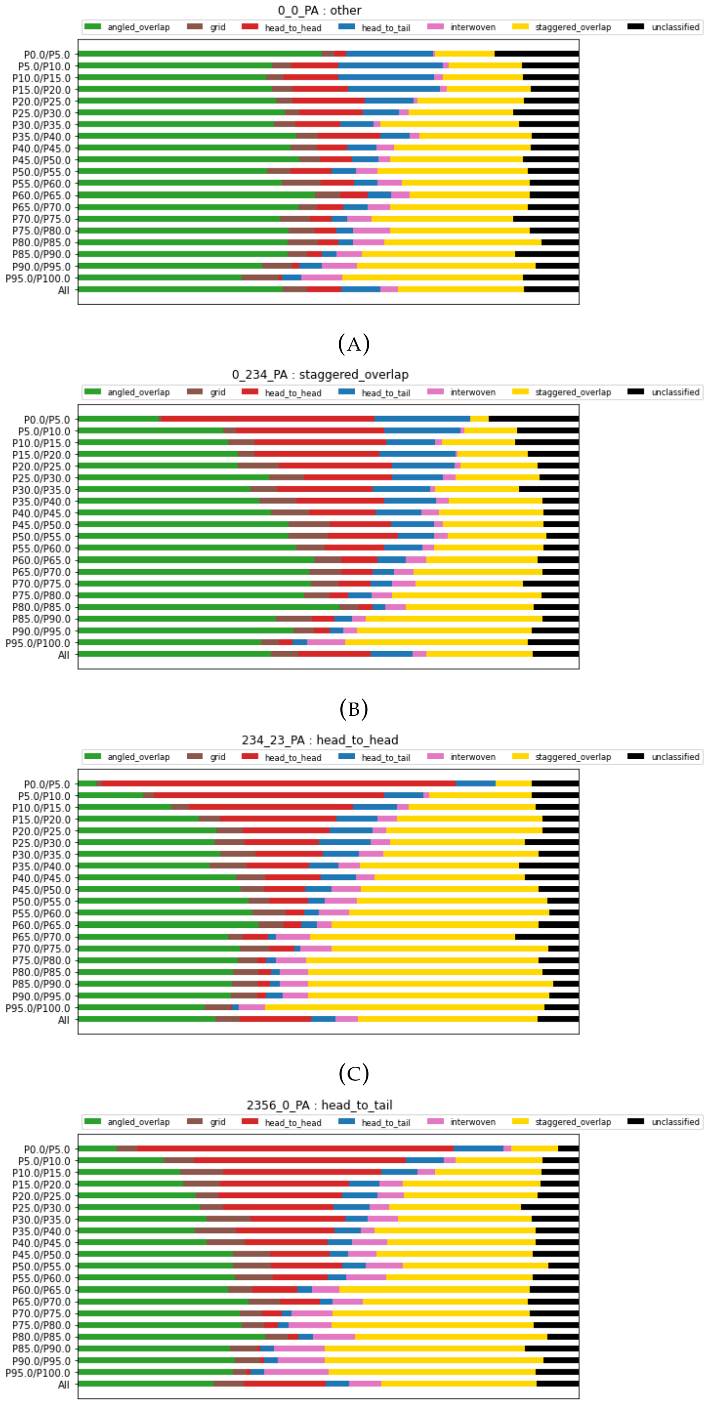
(C)

(D)

FIGURE 5.29: Histogram of the predicted energies of a set of crystal structure predictions for fluoroalanine crystals stratified by predicted packing type according to a supervised UMAP/DBSCAN method using a small set of classified experimental fluoroalanine crystals to fit the model.

figure challenging to interpret. To conclude all of the different sets of crystal structure landscapes (that is, those based on PCA, LDA and UMAP) have potential advantages and disadvantages when attempting to understand the data in the context of a smaller high fidelity labelled dataset. It might be most wise to use these techniques in conjunction with one another rather than relying on a single one of these pictures and being subject to the inherent pitfalls of that viewpoint.

We finish this section by considering the problem of property prediction for these compounds. Topological data analysis itself only provides information about the connectivity of the data points, so at the very most we have encoded information pertaining to molecular orientations and various factors relating to how the molecules may pack together. There is nothing here related to the internal structure of these compounds, indeed in none of these models do we even encode the *type* of atoms involved let alone consider ways of thinking about the interaction of different atomic or molecular orbitals or multipoles or intermolecular bonding patterns. Therefore an



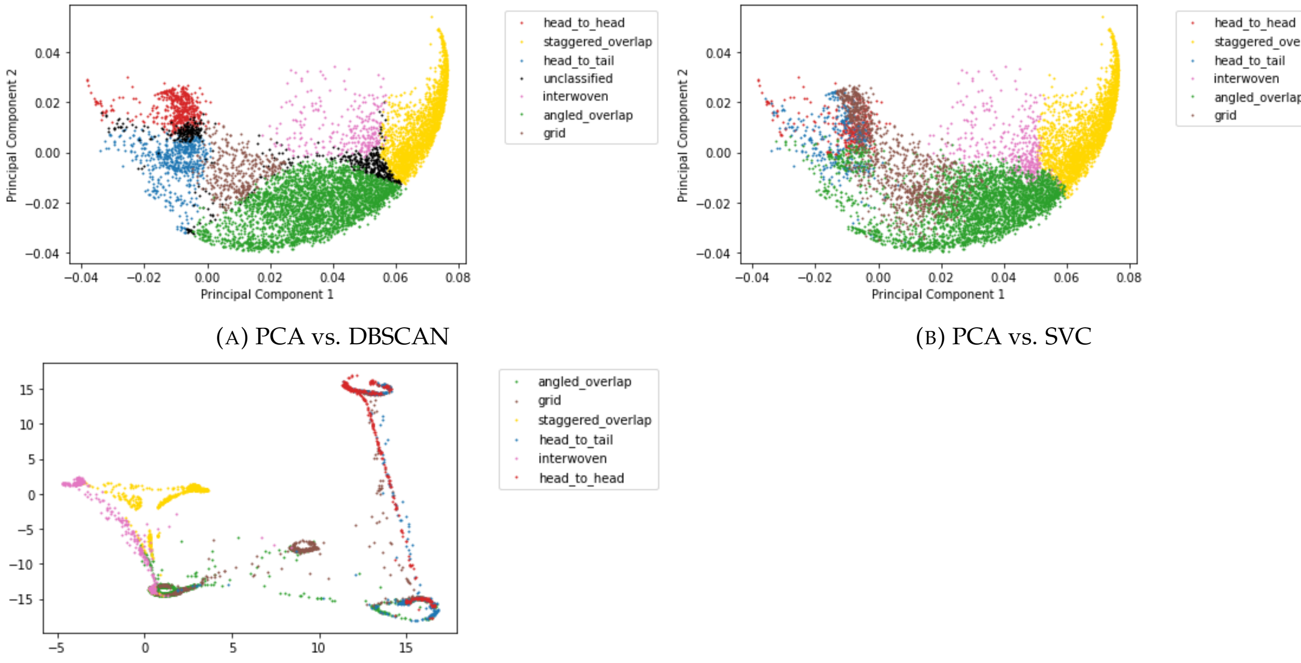
(A)

(B)

(C)

(D)

FIGURE 5.30: The distribution of predicted packing type according to a supervised UMAP/DBSCAN method using a small set of classified experimental fluoroalanine crystals to fit the model for a set of crystal structure predictions for fluoroalanine crystals stratified by predicted energy. The structures are split into tranches based on percentiles. Here Px/Py refers to the set of compounds whose energies are less than the bottom y percentile and greater than the bottom x percentile so that for example $P0/P10$ contains the bottom 10% of energies and $P10/P20$ contains the next 10% etc.



(A) PCA vs. DBSCAN

(B) PCA vs. SVC

(C) UMAP vs SVC

FIGURE 5.31: A set of crystal structure landscapes for the predicted structure of the fluoroalanine ”0_0”. In the parts a) and b) the crystal structure landscape was found by transforming the vector images according to the principal components of the vector images of a set of experimental fluoroalanine compounds with known packing type. In part c) the crystal structure landscape was obtained using the supervised UMAP model, the model being fit on the same set of known fluoroalanines as above. There are two labelling schemes the first was obtained by applying DBSCAN (eps $= 4$, min_samples $= 7$) on the UMAP based landscape. This labelling scheme is used in part a). In parts b) and c) the labelling scheme comes from fitting a support vector classifier (linear kernel, $C = 100$) to the (scaled) vector images of the known fluoroalanines and predicting the labels of the unknown fluoroalanines accordingly. Again the axes of the UMAP plots have no physical interpretation.

accurate prediction of any properties of the crystal structure without further information seems a little far fetched. There is however always some link between the geometry of the sets of atoms and molecules and the chemical properties of the aggregate so it is worth briefly investigating how well the persistent homology can act as proxy for these features. It might be possible to combine this with other descriptors to obtain more accurate results, for example.

Since we already have a large set of predicted crystal structures with a predicted energy and a persistent homology descriptor which we know has strong predictive power when it comes to crystal packing type, carrying out energy prediction on these data using our persistent homology based descriptor seems an interesting problem.

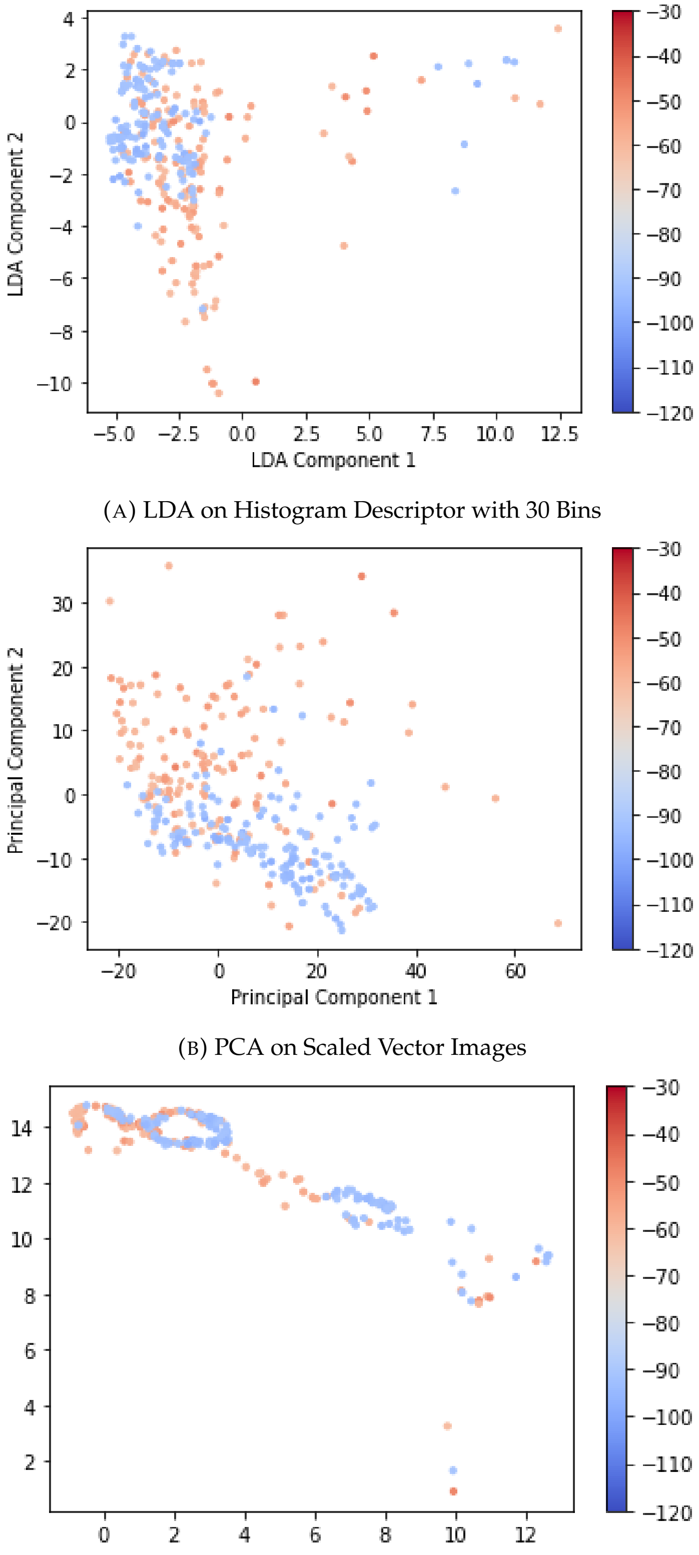
We start by noting that there is at least *some* evidence of trends in the energy across the crystal structure landscapes we have already considered although this might not be clear from the previous figures. Figure 5.32 shows a set of crystal structure landscapes for which only the 100 lowest energy and 100 highest energy structures are plotted. These three landscapes all show regions of the landscape for which there are high energy structures but no low energy structures or *vice versa*. There is not a clear *overall* trend insomuch that all regions of crystal landscapes have crystal structures with a range of energies and that low and high energy structures may be found in any part of the crystal structure landscape.

With the above in mind a good starting point might be to consider whether we can predict which structures have either high or low energy, that is, to frame the problem as a classification problem. The technique we chose to use for this is support vector classification and we applied this to both vector images and histograms of connected components. These were either scaled (converted to a standard scalar before fitting) or unscaled and either in their full (i.e. high dimensional) form or with the dimension reduced with PCA, LDA and supervised and unsupervised UMAP. In all cases the models were fit using 75% of the vector images (or histograms) as training data - the rest being withheld for validation. The optimal hyperparameters for each model were found using 5-fold cross validation. The energy was split into either 2 classes or 5 classes using the energy percentile function (i.e. in the 2 class case, the classes were the top 50% and bottom 50% of energies).

The best (highest accuracy) results were found using vector images and the only dimensionality reduction technique that yielded decent results was PCA. In figure 5.33 we show the confusion matrices obtained using 2 energy classes and 3 different methodologies. In the first example the unaltered vector image was used with a radial kernel with $C = 10000$. In the second example this vector image was converted to a standard scalar before the model was fit with a radial kernel with $C = 100$. In the third case the model was reduced to 50 dimensions [6] with PCA before the model was fit. In this case a radial kernel was also used and C was set to 1000. The three confusion matrices are very similar so it is difficult to say definitively which approach to choosing appropriate independent variables for the classification model is best. The models also perform very poorly with a result only slightly better than a random guess with the accuracy of these models being, respectively, 0.666, 0.660 and 0.668.

Unsurprisingly the models are not much better when 5 classes are used. We fit with the same input variables as before. The confusion matrices are shown in figure 5.34. A radial kernel was found to be the best in all cases. The optimal values of C also turn out to be the same, that is respectively $C = 10000$, $C = 100$ and $C = 1000$. The accuracies of the models are (respectively) 0.34, 0.34 and 0.35. Once again the model is only sightly better than a random guess.

[6] It turns out the accuracy of the final model is pretty much invariant the actual number of dimensions chosen provided that it is greater than 5.

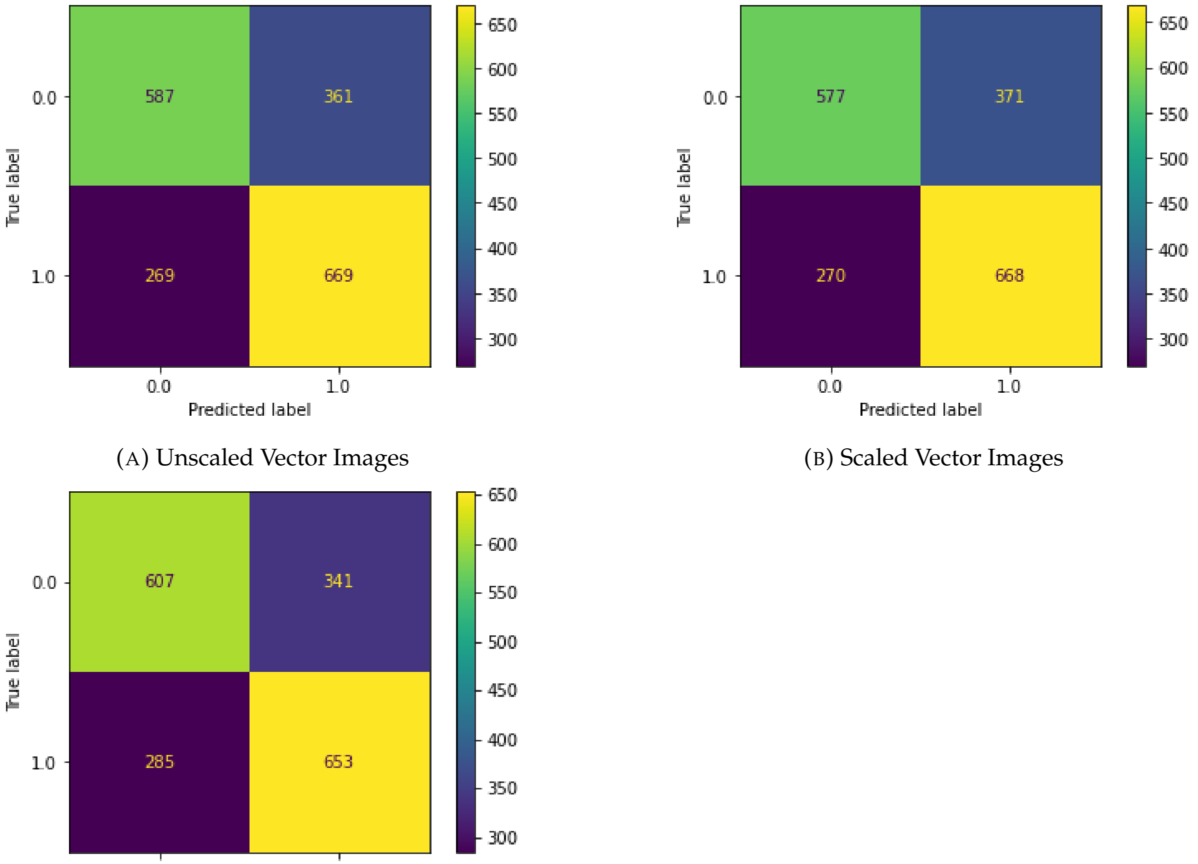


(A) LDA on Histogram Descriptor with 30 Bins

(B) PCA on Scaled Vector Images

(C) Unsupervised UMAP on Unscaled Vector Images

FIGURE 5.32: A set of crystal structure landscapes of the fluoroalanine labelled ”0_0” with only the 100 lowest and 100 highest energy structures shown. Again the axes of the UMAP plots have no physical interpretation.



(A) Unscaled Vector Images

(B) Scaled Vector Images

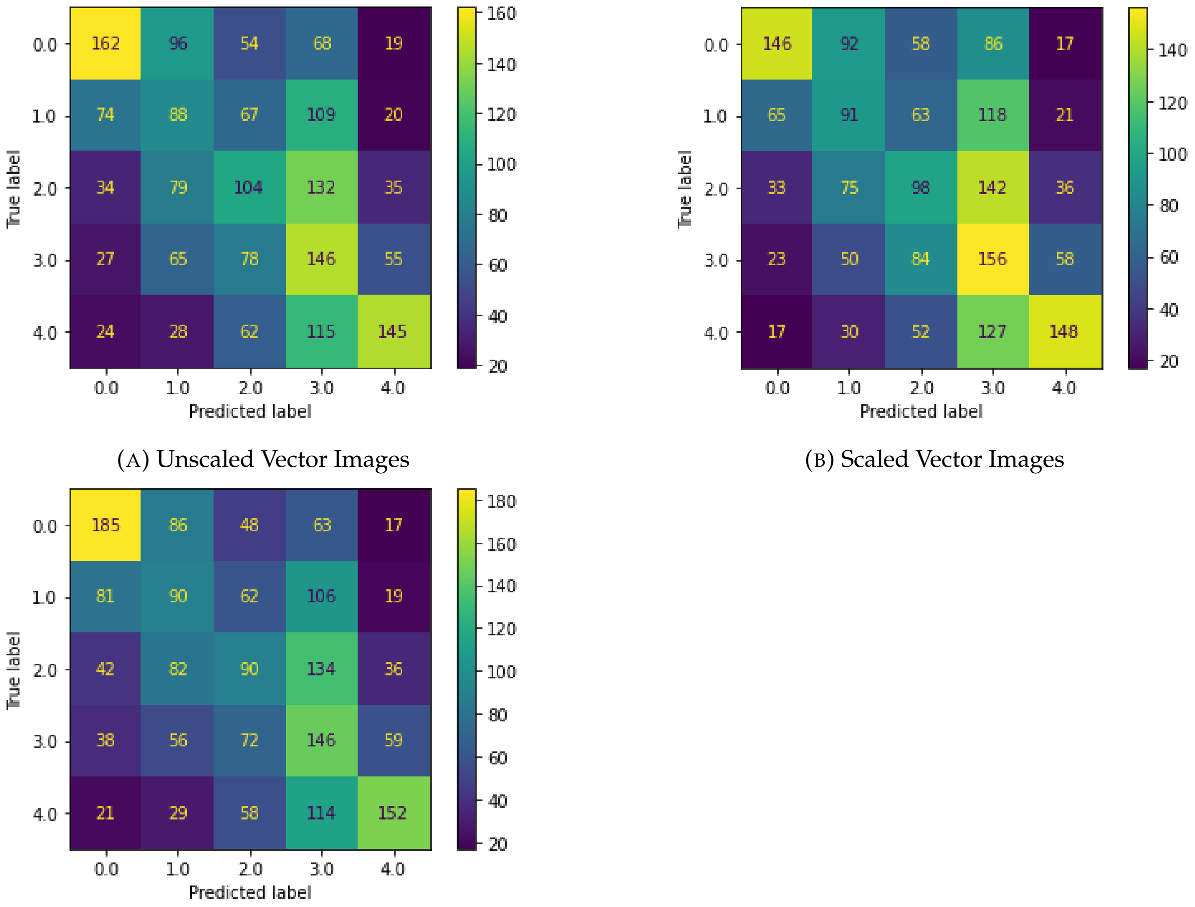
(C) Reduced Dimension Vector Images

FIGURE 5.33: Confusion matrices for a set support vector classifiers for the prediction of the energy of predicted structures for the fluoroalanine ”0_0” into two classes (split by 50th percentile). The confusion matrices pertain to the predictions on the 25% holdout set. Part a) corresponds to a model fit on the unaltered vector image. Part b) corresponds to a model fit on the vector images converted to standard scalars. Part c) corresponds to the model fit on the first 50 principal components of the vector images. In all cases we found that the radial kernel was the best choice of kernel. The regularisation parameter C was set respectively to 10000, 100 and 1000

We also attempted to fit an SVC model on a 3 class case with the top and bottom 20 % of energies in two classes and the middle 40 % in the remaining class [7] with the view of establishing if we can detect low and high energy examples. This model performed badly in all cases. The confusion matrix of the model (radial kernel, $C = 1000$) fit on the unaltered vector image in shown in figure 5.35.

For completeness we also attempted to fit a regression model for the energy. For each of these models we used the unscaled vector images reduced in dimension with PCA to 50 dimensions. The data was split into training and test sets with a 75:25 ratio as before. Any hyperparamters were found using successive 5-fold cross validations

[7] We used the technique of stratified sampling to ensure that the training set (and the sets in cross validation) had sufficient examples of each class during training.



(A) Unscaled Vector Images

(B) Scaled Vector Images

(C) Reduced Dimension Vector Images

FIGURE 5.34: Confusion matrices for a set support vector classifiers for the prediction of the energy of predicted structures for the fluoroalanine ”0_0” into five classes (split by 20th percentile). The confusion matrices pertain to the predictions on the 25% holdout set. Part a) corresponds to a model fit on the unaltered vector image. Part b) corresponds to a model fit on the vector images converted to standard scalars. Part c) corresponds to the model fit on the first 50 principal components of the vector images. In all cases we found that the radial kernel was the best choice of kernel. The regularisation parameter C was set respectively to 10000, 100 and 1000

with different parameters sets. In the case of the random forest regressor we sampled a random subset of these parameters as trying every combination was not computationally feasible. Figure 5.36 shows plots of predicted vs actual energy for compounds in the test set for three models: a linear model, a support vector regression model and a random forest model. The optimal parameters for the SVR model were $\epsilon = 0.1, C = 10000,$ kernel $=$ radial. The parameters chosen for the random forest model were max_depth’= 20, max_features = ”auto”, max_leaf_nodes = ”None”, max_samples = None, min_samples_leaf = 2, min_samples_split = 6 and n_estimators = 200. The R^2 values found for the linear model, support vector regression and random forest regression were respectively 0.16, 0.31 and 0.29. It can seen that while the fits are very poor in all cases there is at least some correlation between the predicted and true energies suggesting that these models have at least some predictive power and hence

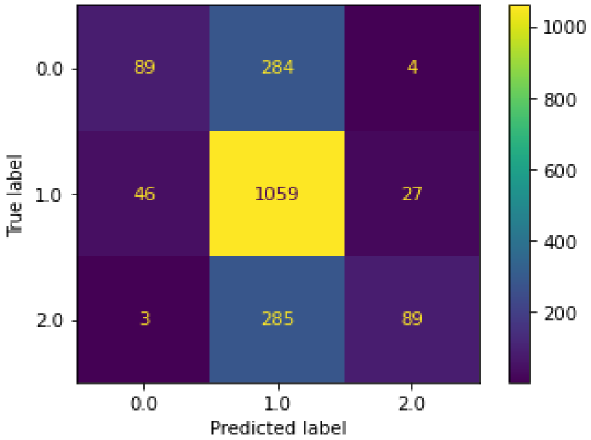
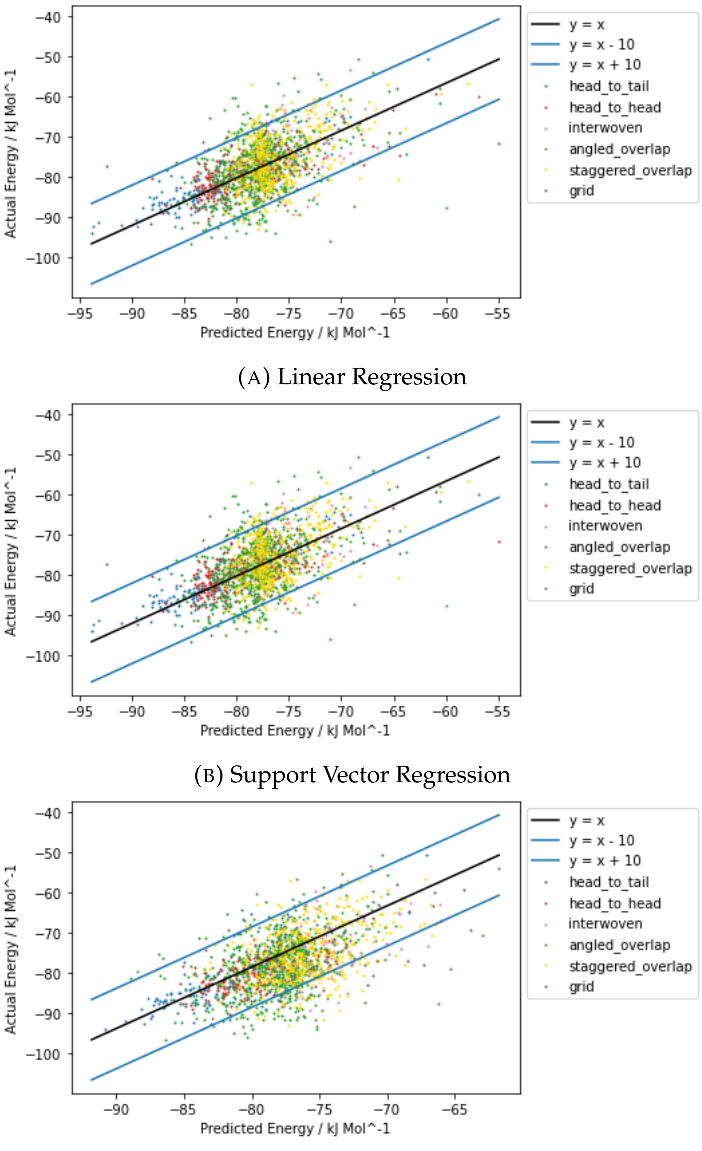


FIGURE 5.35: Confusion matrix of a support classifier model fit on the vector images corresponding to the persistent homology of some predicted crystal structures of the fluoroalanine ″0_0″. We attempted to predict if the energy of the crystal was in the top or bottom twentieth percentile or otherwise. A radial kernel was used and C was set to 1000.

that there is some connection between the persistent homology of the underlying crystal structures and their energies. The topological information we extracted could perhaps be useful when refining existing models which already take in to account the underlying chemistry (by consideration of potential intermolecular interactions, for example) by providing extra information pertaining to the topological connectivity of the molecules in 3D space. At any rate the discussion above highlights the utility of this descriptor in creating and interpreting useful crystal structure landscapes for sets of related molecules but perhaps its lack thereof when directly predicting any physical or chemical properties.



(A) Linear Regression

(B) Support Vector Regression

(C) Random Forest Regression

FIGURE 5.36: The prediction of the energy of some predicted structures of the fluoroalanine labelled ”0_0” using a vector image descriptor derived from the persistent homology of the crystal structures. The vector image dimension was reduced to 50 using PCA before each model was fit. The figures show the plot of predicted vs actual energy for the test set (25% of the data). The first model is a linear model. The second is a support vector machine ($\epsilon = 0.1$, $C = 10000$, kernel $=$ radial). The third model is a random forest model. The parameters chosen for the random forest model were max_depth’= 20, max_features = ”auto”, max_leaf_nodes = ”None”, max_samples = None, min_samples_leaf = 2, min_samples_split = 6 and n_estimators = 200.

5.2 Nicotinamide:Benzioic Acid (GAZCES) Co-Crystals

We start by considering the persistent homology of the GAZCES co-crystal using molecular centroids only. We found a similar size invariance as in the fluoroalanine case: 50 centroids were sufficient for our calculations. We used the Vietrois-Rips filtration for the calculation of persistent homology and as independent variables used only the 0D persistent homology features which like in the fluoroalanine case turn out to be by far the most important for understanding crystal packing. The reduced space obtained using PCA is shown in figure 5.37. Unlike the fluoroalanines a descriptor based on the homology of the set of molecular centroids does not separate the classes definitively at all (at least not in the space defined by the principal components) with the possible exception of the dense ″cluster″ of compounds on the right of the plot of figure 5.37, which outside of any additional context would simply be considered an artefact of the data.

Owing to the work completed earlier on the generalised method for finding molecular orientations it is trivial to augment the set of molecular centroids with a set of suitably chosen inertia eigenvectors to obtain a 6D space for which to find the persistent homology. The PCA plot obtained in this manner is found in figure 5.38. This is almost exactly the same as the plot obtained from 3D persistent homology in figure 5.37 apart from the notable difference that the ″cluster″ of points (belonging to basin

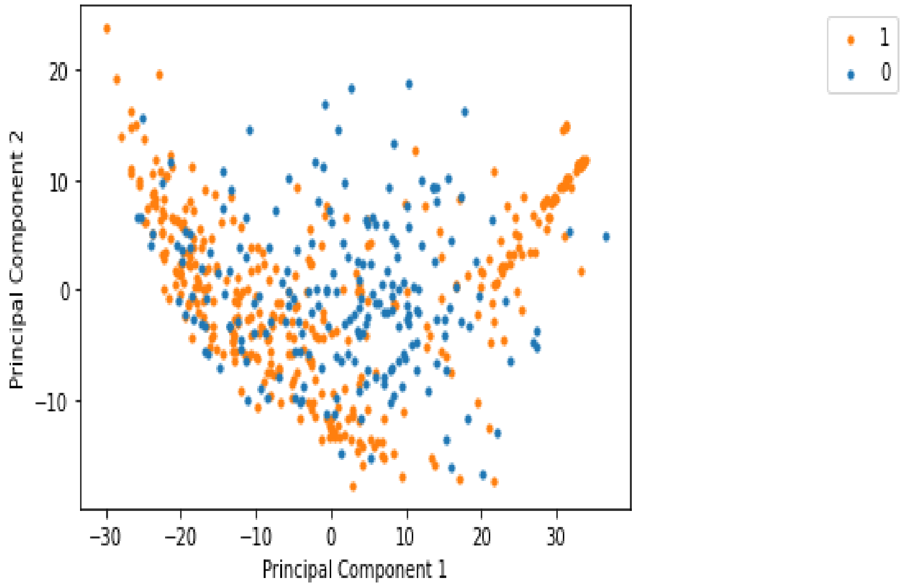


FIGURE 5.37: Plot of the first two principal components of the space spanned by a vector image descriptor based on the 0D persistent homology of the (3D) set of molecular centroids belonging to GAZCES co-crystal structures. The data is partitioned according to whether the crystal structures were found to lie in two deep basins (corresponding to experimental polymorphs) of the crystal structure landscape. The basins are simply labelled ″0″ and ″1″.

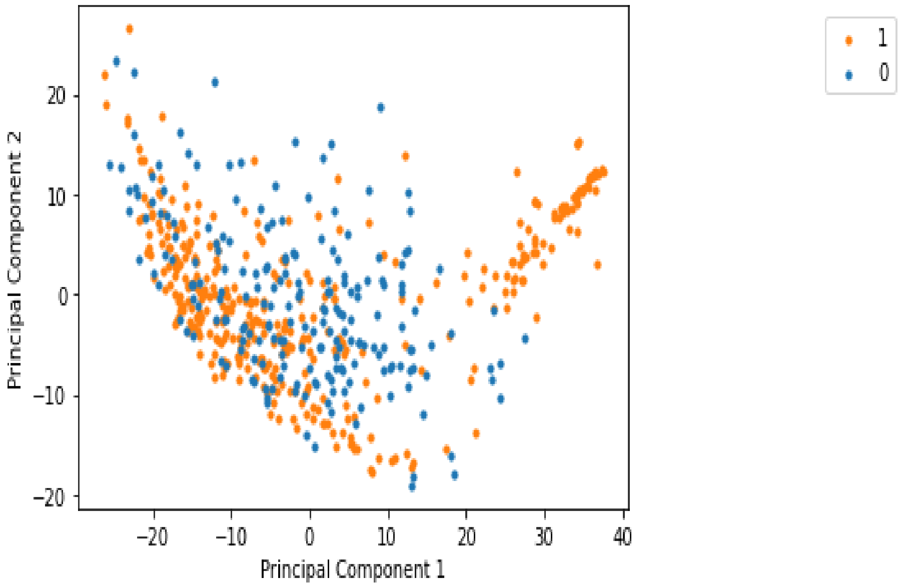
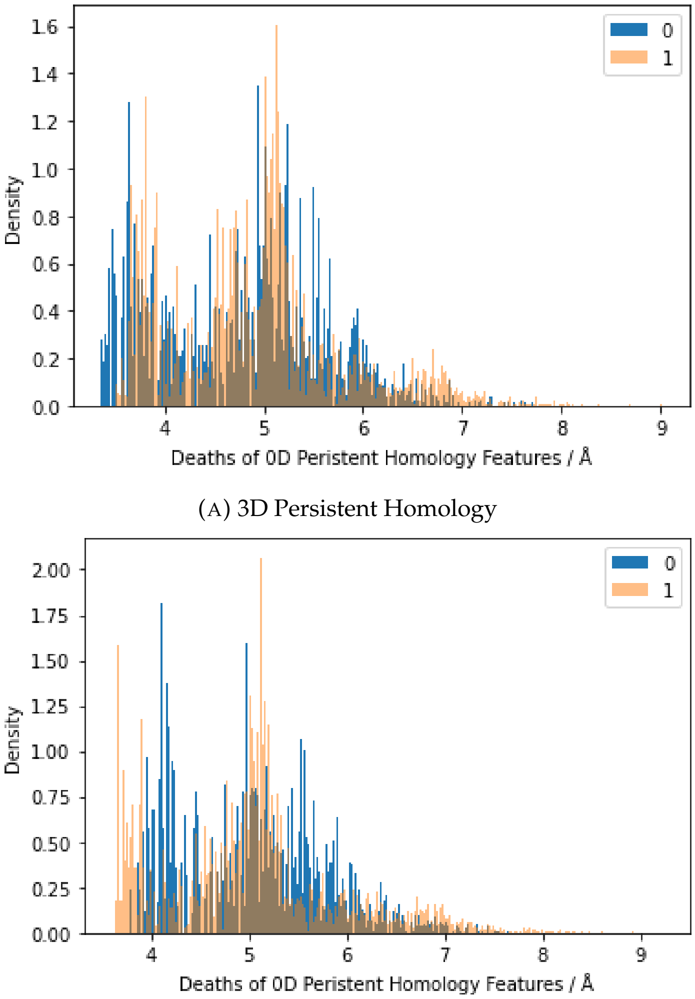


FIGURE 5.38: Plot of the first two principal components of the space spanned by a vector image descriptor based on the 0D persistent homology of the (6D) set of molecular centroids and suitably chosen inertia vectors belonging to GAZCES co-crystal structures. The data is partitioned according to whether the crystal structures were found to lie in two deep basins (corresponding to experimental polymorphs) of the crystal structure landscape. The basins are simply labelled ″0″ and ″1″.

1) is now much more prominent. This may give us tenuous insight that the connectivity information encoded in the orientation vectors are important for describing the packing of these crystal systems. The effect is even more stark when we examine the histograms of the set of connected components (across all persistence diagrams belonging to given basin) as shown in figure 5.39. We see that for the 3D case the histograms can almost be superimposed while in the 6D case the histogram corresponding to the homology features of the crystals in potential energy basin ″0″ have shifted to the right.

Can we augment this effect with modification of the orientation vector? It turns out that all we need to do to achieve this is to scale the orientation vector so that it is large relative to the intermolecular distances. Figure 5.40 shows the PCA plots that are obtained when the 6D persistence homology is calculated on the space of centroids and vectors for which the length of each vector is increased by a factor of 20. The effect on the crystal structure plot is incredibly striking. Not only is the ″artefact″ in figure 5.37 now a clearly defined cluster but almost all of the compounds that belong in funnel ″1″ are separated from those that belong in funnel ″0″.

This striking effect is also evident in the histogram of connected components as shown in figure 5.41. We see that the histogram corresponding to homology feautrs of crystal structures which belong to funnel ″0″ has shifted to the right substantially so as to be almost disjoint to the histogram corresponding to funnel ″1″.



(A) 3D Persistent Homology

(B) 6D Persistent Homology

FIGURE 5.39: Histogram of the deaths of 0D persistent homology features for a set of GAZCES co-crystals belonging to a given potential energy minimum of the wider crystal structure landscape. In a) the persistent homology was found using molecular centroids only. In b) the persistent homology was found using both molecular centroids and a suitably chosen inertia eigenvector which describes the molecular orientation.

The implication of the rescaled orientation vector is interesting: the distances between molecules in the new 6D space are dominated by the differences in their orientations and *not* by the differences in their xyz coordinates. The implication of this is that we can imagine our space as having all molecules with the same orientation grouped together so that when the resulting 6D points are connected together in the homology filtration, the points with the same orientation are always connected first while the simplices composed of sets of molecules with different orientations are joined at the end. The structure of the 6D space of one of the GAZCES structures with unit orientation vectors verses augmented orientation vectors is demonstrated in 2 dimensions with Multidimensional Scaling (MDS) in figure 5.42.

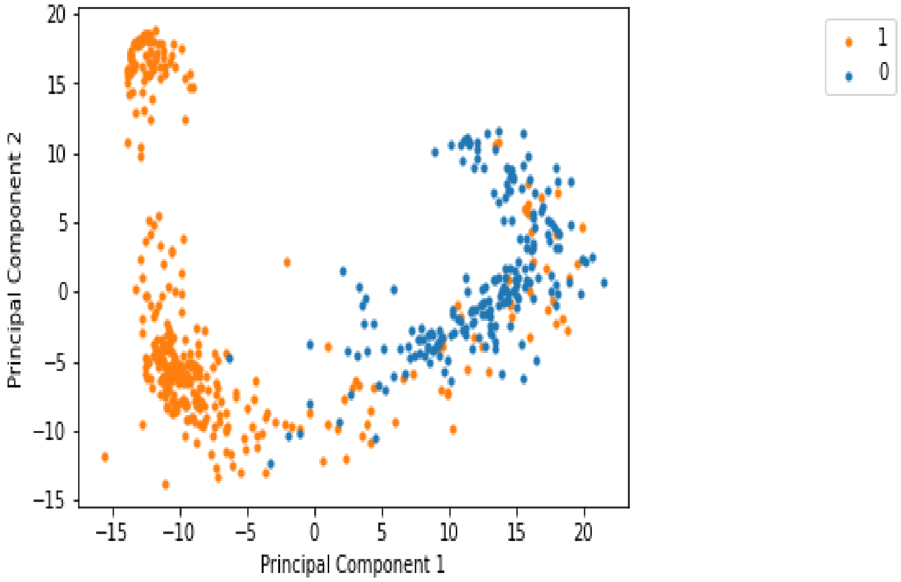


FIGURE 5.40: Plot of the first two principal components of the space spanned by a vector image descriptor based on the 0D persistent homology of the (6D) set of molecular centroids and modified inertia vectors belonging to GAZCES co-crystal structures. The data is partitioned according to whether the crystal structures were found to lie in two deep basins (corresponding to experimental polymorphs) of the crystal structure landscape. The basins are simply labelled ”0” and ”1”. The inertia vectors have been modified by increasing their length by a factor of 20.

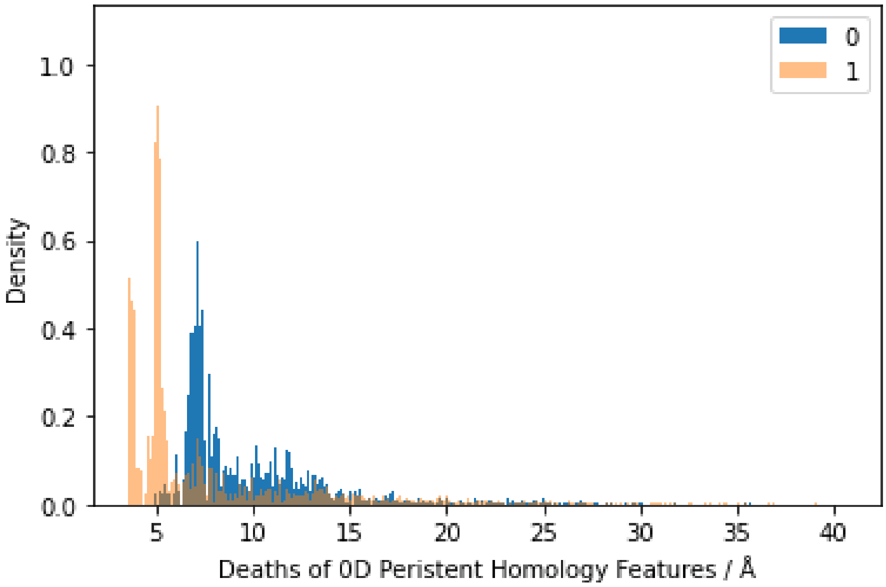


FIGURE 5.41: Histogram of the deaths of 0D persistent homology features for a set of GAZCES co-crystals belonging to a given potential energy minimum of the wider crystal structure landscape. The persistent homology was found using both molecular centroids and a rescaled inertia eigenvector (increased by a factor of 20).

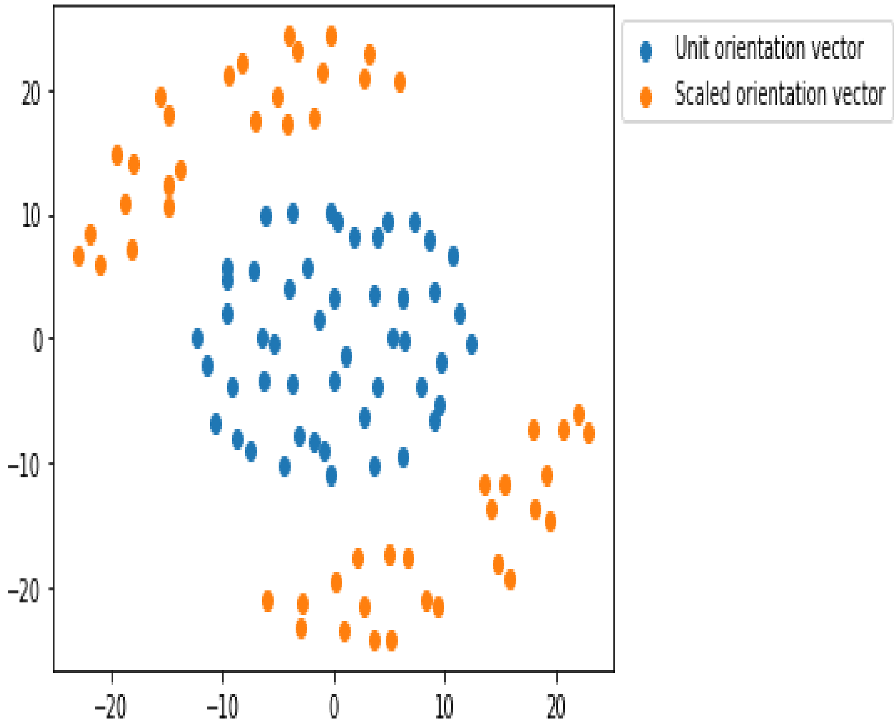
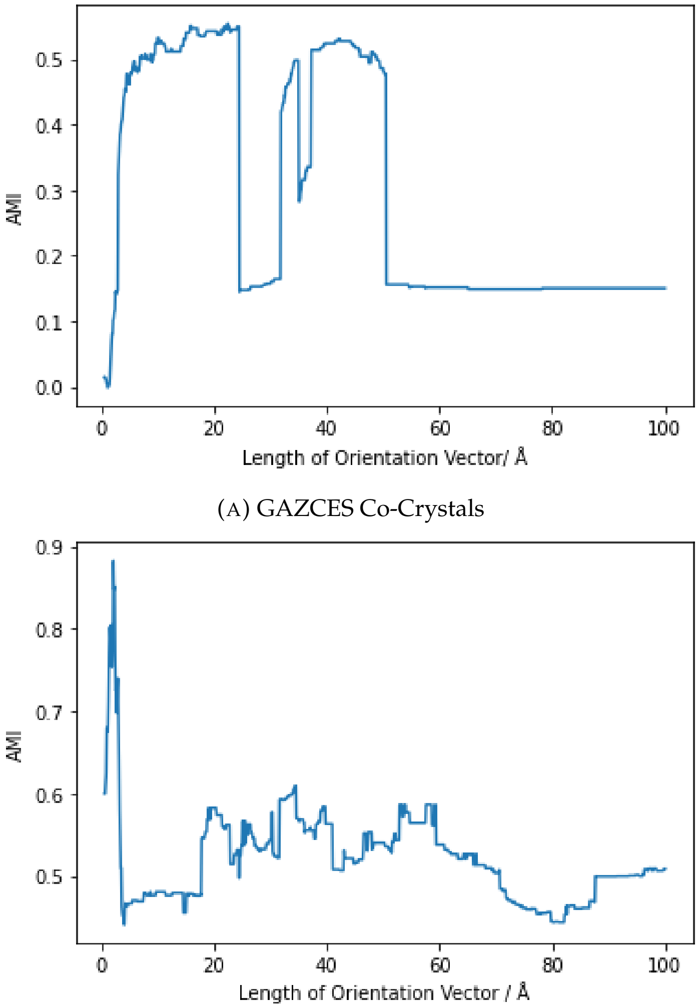


FIGURE 5.42: 2D MDS embedding of an example 6D GAZCES input structure using both unit orientation vectors and augmented orientation vectors. We see the space is partitioned into clusters according to vector orientation in the augmented case. The axes of the MDS embedding have no physical interpretation.

The length of the inertia vector has a pronounced effect on the AMI of the clustering of the high dimensional space obtained from the underlying vector images. A plot of the AMI against the orientation vector length is shown in figure 5.43a. A clear feature here is the sharp increase in AMI when the length reaches 5 Å steadily increasing until 20 Å where we reach the maximum. We then have a very sharp drop in the fit. This is followed by another period of increased AMI from 30-50 Å before another drop. It is not yet clear what causes this behaviour. Contrast this with the equivalent plot for the fluorolanines in figure 5.43b. Increasing the length of the orientation vector actively makes the fit worse with the optimal length of orientation vector being close to unity. This is the only dataset we have encountered thus far for which the length of the orientation vector has such a stark effect. One hypothesis is that because the GAZCES structures are composed of two different kinds of molecule (nicotinamide and benzoic acid) the augmented orientation vector 6D space effectively separates these molecules in the resulting 6D space and then the relative configuration of nicotinamide and benzoic acid molecules is what determines the polymorph type. However examination of figure 4.5 suggests that the inertia vectors of the two components of GAZCES should be rather similar which undermines this hypothesis. It remains to be seen if there are more examples of sets of crystal structures (which are not entirely composed of co-crystals) that exhibit this behaviour surrounding the magnitude of the



(A) GAZCES Co-Crystals

(B) Fluoroalanines

FIGURE 5.43: Variation of clustering AMI when predicting packing types of crystal structures from persistent homology calculations with the length of the vector used to describe molecular orientation for the 6D space for which the persistent homology is found. Part a) corresponds to the set of fluoroalanines labelled by a packing scheme found by eye while part b) corresponds to a set of GAZCES co-crystals labelled by which (of two) potential energy basins on the crystal structure landscape they belong to.

orientation vector.

5.3 Polyaromatic Hydrocarbons & Azapentacenes

5.3.1 Experimental Polyaromatic Hydrocarbons

Firstly we examine the set of PolyAromatic Hydrocarbons (PAHs) obtained from the CSD and the paper by Loveland *et al.* Loveland et al. (2020). In the first instance we found the persistent homology using both the 3D and the 6D pointcloud using the

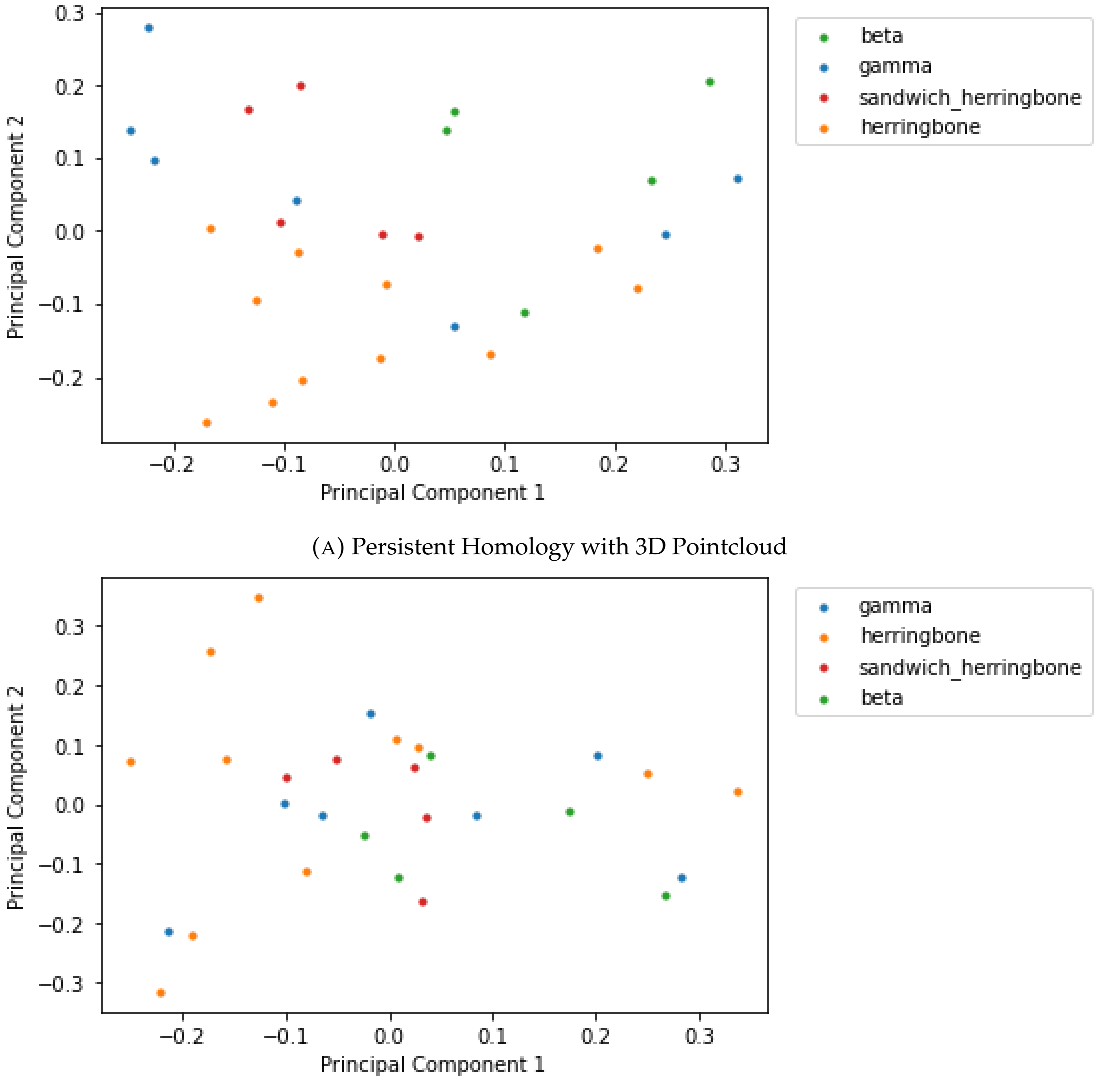
Rips filtration and using the vector images of the 0D features only (once again these turn out to be the most important). The PCA plots thus obtained are shown in figure 5.44 for the first set of PAHs (28 structures from the CSD) and in figure 5.45 for the second set (172 structures from Loveland *et al.*).

For the first set of PAHs the classes do not seem to be well separated in the reduced space although it could be that with more data a trend would emerge. Note also that incorporating the molecular orientations does not seem to make things better, and possibly makes things slightly worse.

For the larger set of PAHs we also do not see any clear separation of classes. There are a few loose trends: for example there is a small zone mostly occupied by structures with the gamma packing type and most of the herringbone structures occur at the edge of the set of points but these are the only trends observed. Also note that in this case there is literally no improvement in the class separation with the incorporation of orientation vectors (cf. the GAZCES co-crystal) into the homology calculation.

It is also worth remarking that when these datasets are combined the data points occur at sensible locations relative to one another which highlights that these two datasets do not have radically different crystal structures and thus persistent homology as expected. This is shown in figure 5.46 (we only show the 6D case).

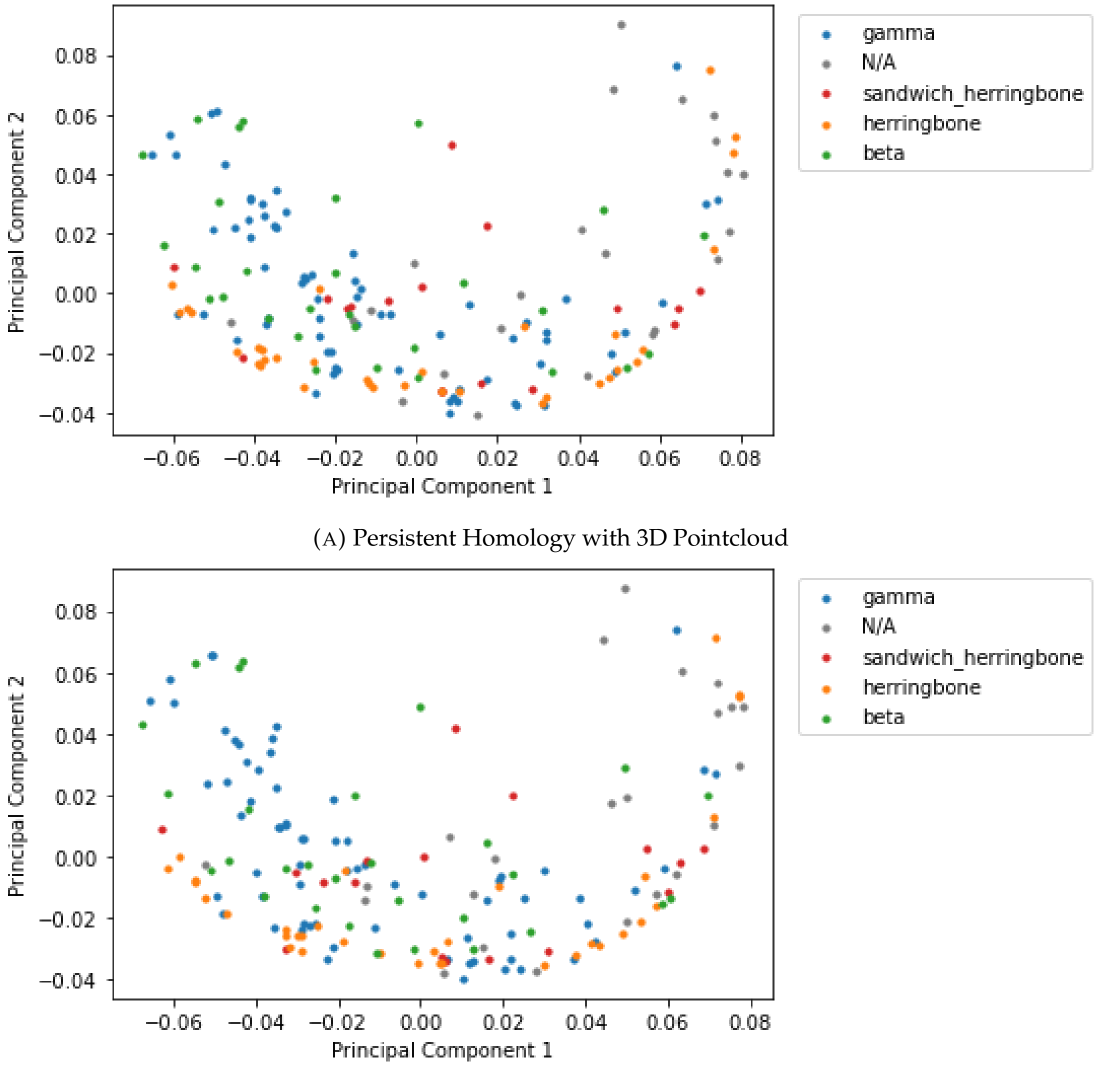
Are there any features of this dataset that might be complicating our analysis? One immediate difference between this set and the proceeding two datasets is that the constituent molecules have radically different shapes and sizes. The GAZCES dataset is composed of different packing types of the same compound while in the case of the fluoroalanines all the compounds are very closely related the only difference being the fluorine substitution pattern. The reason this is very important in our context is that we are using molecular centroids to record the position of the molecules in each crystal structure and if the consistent molecules in a crystal structure happen to be larger or have a very different shape (like a longer long axis for example) then the intercentroid distances will be increased and the persistent homology features which we observe will not all occur on the same scale. Observe figure 5.47. Here the distances between molecules and their first nearest neighbour are plotted in a histogram across all molecules in all crystals in a given dataset. One can see that in the case of the fluoroalanines (labelled emd), while there is a large range of first nearest neighbour distances, there is a very large peak indicating that the majority of distances are occurring at the same scale. Contrast this with the PAHs. The distances occupy a larger range and moreover the distances are scattered across this range suggesting that we are seeing a lot of different geometric features occurring at different scales. We also show the histogram for the distances to the second nearest neighbour which shows a similar trend.



(A) Persistent Homology with 3D Pointcloud

(B) Persistent Homology with 6D Pointcloud

FIGURE 5.44: Reduced Vector Images from the persistent homology of a set of classified polyaromatic hydrocarbons (into four packing types) from the CSD. The persistent homology was found with the Vietoris-Rips filtration and only the 0D features were used. The vector images were reduced in dimension with PCA. In part a) the persistent homology was calculated on a set of 50 molecular centroids extracted from the crystal structure and in part b) both the molecular centroids and suitably chosen inertia eigenvectors were used to find the persistent homology of the resultant 6D space.



(A) Persistent Homology with 3D Pointcloud

(B) Persistent Homology with 6D Pointcloud

FIGURE 5.45: Reduced Vector Images from the persistent homology of a set of classified polyaromatic hydrocarbons (into four packing types and N/A) from Loveland *et al.* Loveland et al. (2020). The persistent homology was found with the Vietoris-Rips filtration and only the 0D features were used. The vector images were reduced in dimension with PCA. In part a) the persistent homology was calculated on a set of 50 molecular centroids extracted from the crystal structure and in part b) both the molecular centroids and suitably chosen inertia eigenvectors were used to find the persistent homology of the resultant 6D space.

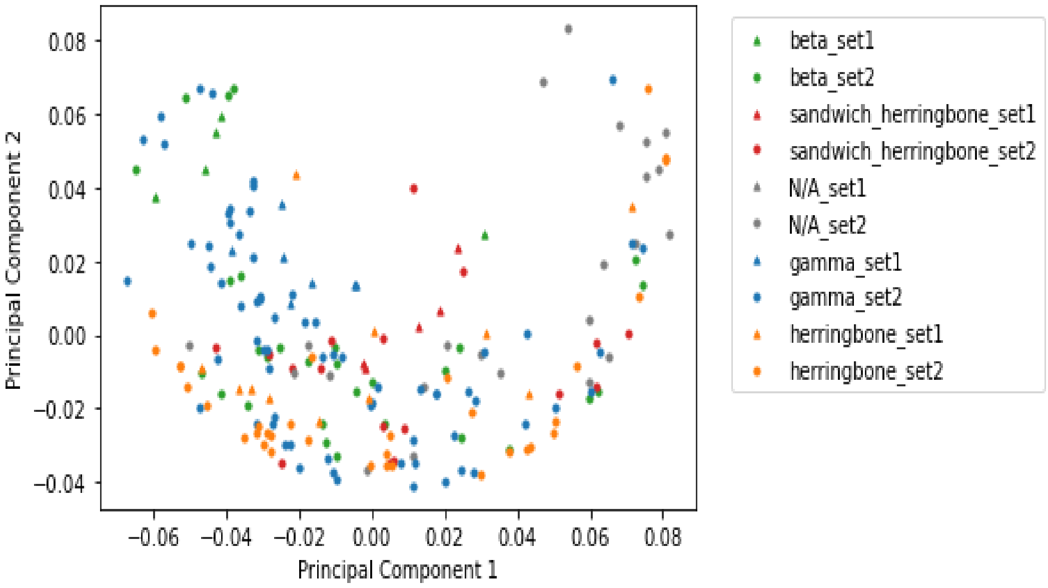
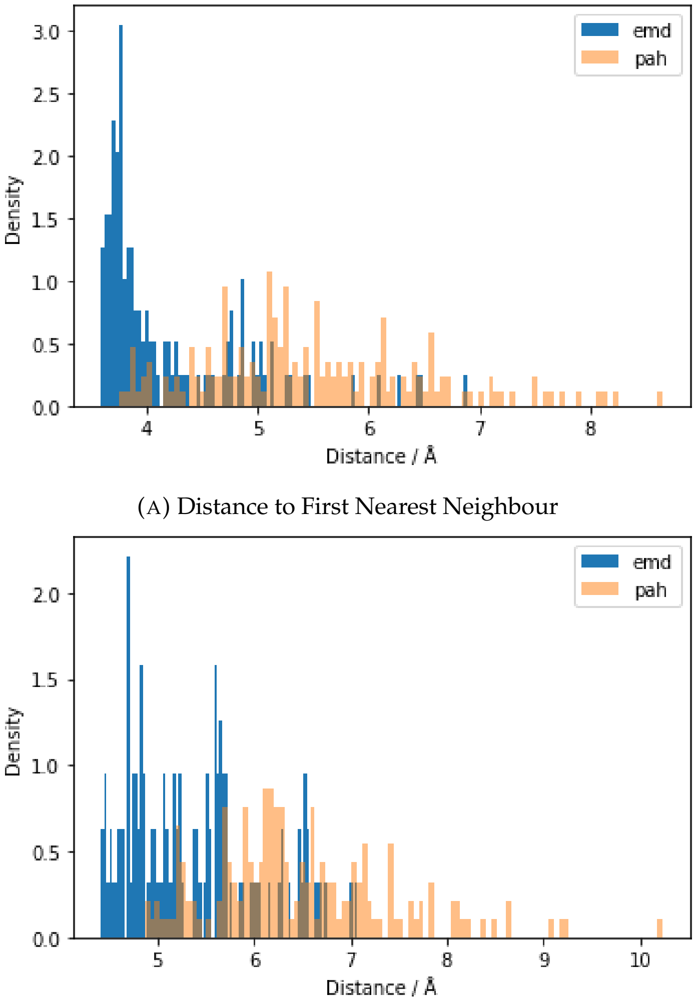


FIGURE 5.46: Reduced Vector Images from the persistent homology of *both* sets of classified polyaromatic hydrocarbons (into four packing types and N/A) from (respectively) the CSD (set 1) and from Loveland *et al.* Loveland et al. (2020) (set 2). The persistent homology was found with the Vietoris-Rips filtration and only the 0D features were used. The vector images were reduced in dimension with PCA. Both the molecular centroids and suitably chosen inertia eigenvectors were used to find the persistent homology of the resultant 6D space.

One possible way of accounting for the different molecular sizes may be to normalise the distances such that for each and every crystal structure the average distance to its first nearest neighbour is one. The histograms which we get for the two sets of crystal structures after applying this transformation are shown in figure 5.48. We observe that the set of intercentroid distances now occupy a similar range and are both distributed around a peak suggesting that features are now occurring at a similar length scale across different compounds in the dataset.

Because we have elected to use the Vietoris-Rips filtration for computing the persistent homology, the homology can be calculated using the distance matrix alone as opposed to the set of 3D or 6D co-ordinates. Thus we can apply this normalisation and calculate persistent homology such that all the distances between molecules occur at the same scale.

The results we obtain are shown in figure 5.49. Here we only show those plots obtained using 3D pointclouds (once again the figures look almost identical when using 3D vs. 6D pointclouds). We observe that while the PCA plot for the first set of PAHs now has the polyaromatic hydrocarbons quite well separated, the plot for the second set of PAHs is only moderately improved (some of the herringbone compounds are closer together, as are some of the sandwich herringbone structures and some of the gamma structures appear to occupy their own region of the plot).



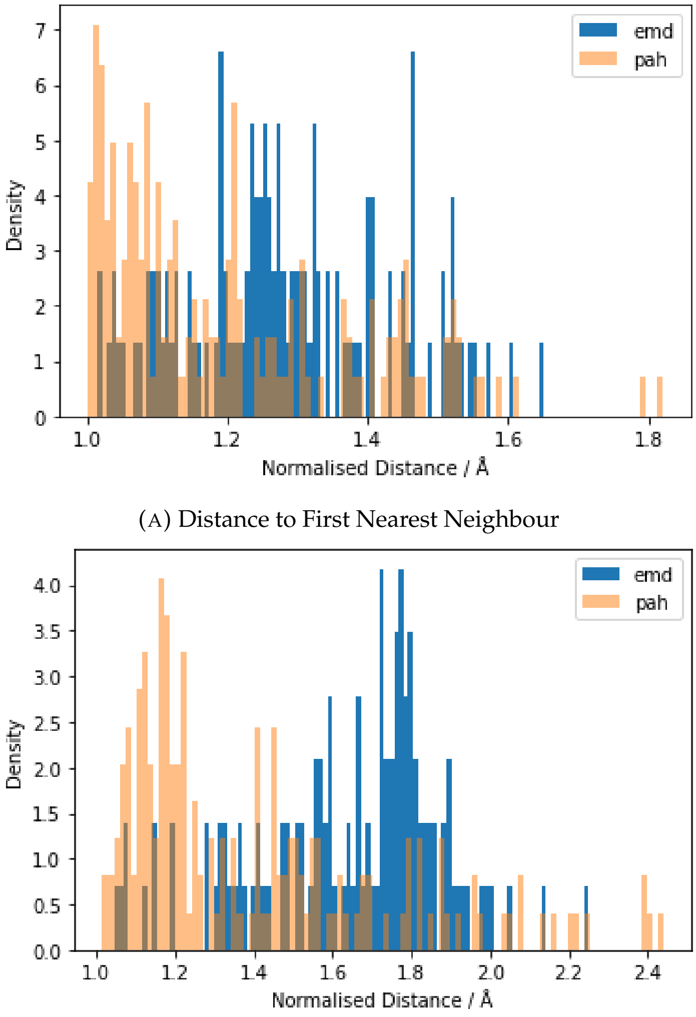
(A) Distance to First Nearest Neighbour

(B) Distance to Second Nearest Neighbour

FIGURE 5.47: Histogram of intercentroid distances to first (part a) and second (part b) nearest neighbours for all molecules of all crystal structures in the set of fluoroalanines (labelled emd) and polyaromatic hydrocarbons (labelled pah).

Any improvement here is probably offset by the fact that it appears now that some crystal structures have homology features that explain a large amount of variance of the data and thus warp the PCA plot and make it more difficult to interpret. The two datasets when plotted together seem once again to coincide in sensible places which may indicate that the ”improved” class separation of the first set of polyaromatic hydrocarbons is merely a result of the small sample size and the inclusion of a more diverse set of polyaromatic hydrocarbons undermines any trend in this first dataset.

We attempt one further method to use persistent homology to describe the packing type of these data (which we shall use again later on). In the work of Loveland *et al.* Loveland et al. (2020) the importance of defining the correct plane upon which to describe the neighbours of a given polyaromatic hydrocarbon (i.e. by viewing the crystal structure from different angles one can ”see” different motifs in the crystal

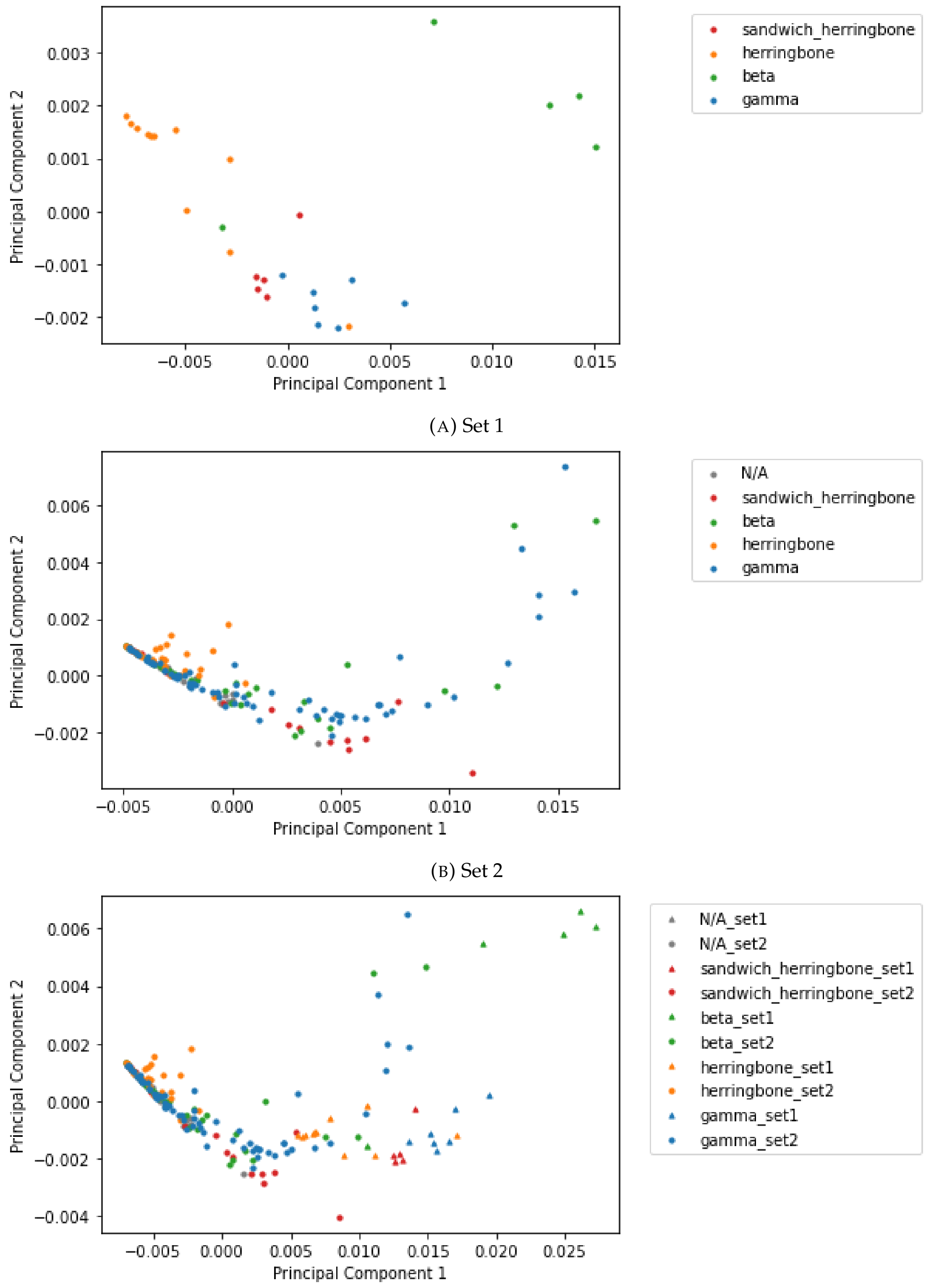


(A) Distance to First Nearest Neighbour

(B) Distance to Second Nearest Neighbour

FIGURE 5.48: Histogram of the normalised intercentroid distances to first (part a) and second (part b) nearest neighbours for all molecules of all crystal structures in the set of fluoroalanines (labelled emd) and polyaromatic hydrocarbons (labelled pah). Here the normalisation is undertaken by dividing all intercentroid distances in each crystal structure by the average distance to its *first* nearest neighbour *before* any of the *k*-nearest neighbour distances for the histograms are calculated.

structure) is highlighted. Finding the correct such plane to describe the beta, gamma, herringbone and sandwich herringbone motifs is the key idea behind the Autopack algorithm which we shall describe later. One of the things that characterises polyaromatic hydrocarbons is their flat often long and thin structure, clearly we cannot capture this shape by using the molecular centroid alone (although the orientation vector may help with this). The implication of this is when the filtration which describes the persistent homology is constructed, the molecules which are closer in terms of their intercentroid distance will be connected first, not those molecules whose interactions best describe the packing motif: in effect some of the persistent homology features most crucial for the description of the four polyaromatic hydrocarbon packing classes will be ”lost” as they are filled in by some of the less



(A) Set 1

(B) Set 2

(C) Sets 1 & 2

FIGURE 5.49: Reduced Vector Images from the persistent homology of *both* sets of classified polyaromatic hydrocarbons (into four packing types and N/A) from (respectively) the CSD (set 1) and from Loveland *et al.* Loveland et al. (2020) (set 2). The persistent homology was found with the Vietoris-Rips filtration using the normalised distance matrix between molecules and only the 0D features were used. All distances between molecules in each crystal structure were normalised by dividing through by the average distance of each molecule to its *first* nearest neighbour. The vector images were reduced in dimension with PCA.

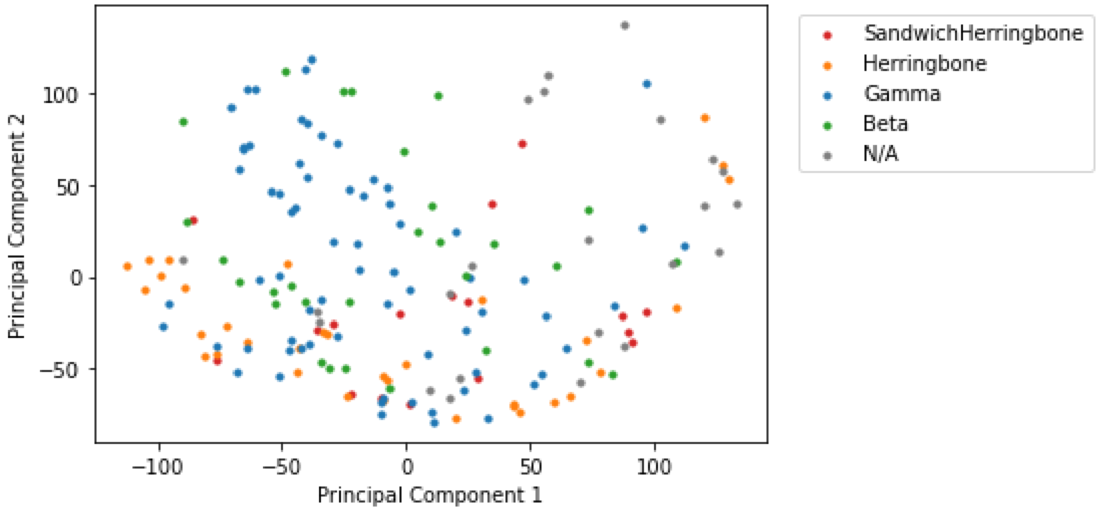
important interactions which take precedence by virtue of their smaller intercentroid distance.

In order to get a feel for the importance of this effect rather than defining the optimal plane using the technique of Loveland *et al.*, we take a much simpler brute force approach and find persistent homology on all three planes defined by the cell axes. The hope here is that by defining point clouds where the crystal unit cells are expanded in one plane only, those intermolecular distances which are shorter and disrupt other important homology cycles are *not* included as all interactions between molecules across neighbouring unit cells above or below this plane are suppressed. Of course this might not suppress problematic interactions within unit cells if $Z' > 1$ but this is not the case for the majority of these compounds.

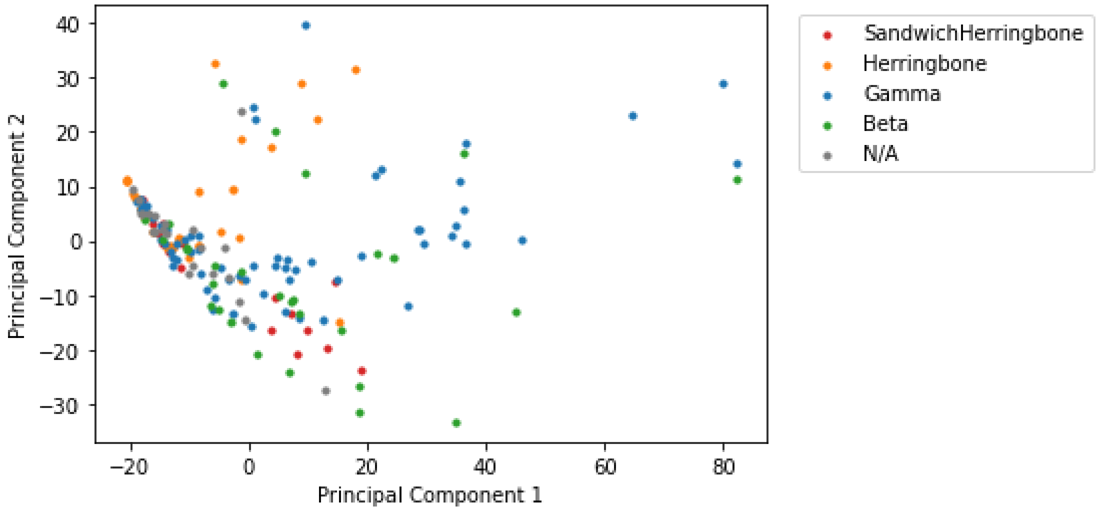
In practice we achieve this by altering the algorithm which we use to generate the crystal fragment (figure 2.6): rather than initially forming a large crystal fragment (from which to select the N closes molecules) we generate a large $R \times R \times 1$ (or $R \times 1 \times R$ or $1 \times R \times R$) supercell and select the molecules from this structure. Here R is chosen to be sufficiently large such that we can always choose exactly N molecules from the desired crystal fragment. This gives us a "planar" crystal fragment so that when we compute the persistent homology some features will be suppressed but other new homology features will appear. We obtain our altered descriptor by simply concatenating the four vector images that we get from the persistent homology of the fragment defined from each supercell in addition to our usual descriptor (which is effectively constructed using a $R \times R \times R$ supercell). Even if we do not know the plane that best describes the crystal motif we can guarantee that it appears in at least one of our crystal fragments.

When we use this expanded descriptor we get the PCA plots described in figure 5.50. Here we focus on the second set of polyaromatic hydrocarbons by Loveland *et al.* as these are the structures which seem difficult to describe by persistent homology. In figure 5.50a we use the unnormed distances in the Vietrois-Rips filtration while in figure 5.50b we use the normalised distances. The figures do not seem to show an obvious separation of the different crystal packing types, we shall see that we have more success with this technique in a later dataset.

While we were not able to find any meaningful trends in the persistent homology of a large part of these data, we have outlined some of the pitfalls that may be involved when applying distance-based molecular descriptors to molecules of different shapes and sizes and possible ways to address this. The next two datasets are also PAHs, but these particular structures were generated from crystal structure prediction calculations so each dataset will contain crystal structures of the *same* molecule and this issue will be eliminated. We will hence be able to establish whether the issues involved in this dataset are specific to the nature of the dataset (all the molecules are



(A) Persistent Homology with Unaltered Distances



(B) Persistent Homology with Normalised Distances

FIGURE 5.50: Reduced Vector Images from the persistent homology of a set of classified polyaromatic hydrocarbons (into four packing types and N/A) from Loveland *et al.* Loveland et al. (2020). The persistent homology was found with the Vietoris-Rips filtration and only the 0D features were used. The vector images were reduced in dimension with PCA. Here the vector images of four different persistent homology calculations are concatenated, each vector image corresponding to a persistence diagram calculated using a crystal fragment after a given supercell. In part a) the distances between molecules is computed using the 6D Euclidean distance defined by the centroids and orientation vectors while in part b) these distances are normalised such that for each crystal structure in our dataset the average distance of a molecule to its *first* nearest neighbour is exactly 1

different) or specific to the nature of PAH crystal structures themselves which could be to do with the difficulties in capturing the ″correct″ intermolecular interactions when computing the persistent homology according to the discussion above.

5.3.2 Azapentacene CSP Calculations: Part 1

The following discussion is centred around data pertaining to the work of Musil *et al.* Musil et al. (2018) on the crystal structure landscapes of azapentacenes. The compounds in question have already been described in the experimental section.

We now describe the technique used by Musil *et al.* Musil et al. (2018) to algorithmically label the azapentacene compounds; this technique was developed by Campbell *et al.* Campbell et al. (2017). The technique is described in figure 5.51 and is loosely based off the seminal work of Gavezzotti *et al.* Desiraju and Gavezzotti (1989b) who identified the importance of inter-planar angles between the neighbouring rings of PAH molecules in determining their experimentally derived packing type (herringbone, gamma, beta and sandwich herringbone). It can be seen that the algorithm in figure 5.51 broadly relies on the principal of finding intermolecular angles between ″nearest neighbours″. In the results of Musil *et al.* there are also compounds labelled ″other″ - it is unclear from either the paper by Musil *et al.* or the one from Campbell *et al.* what this means so they shall be ignored in our analysis.

The 8 crystal structure landscapes for azapentacenes studied (5A. 5B, 5C, 7A, 7B, 7C, TT and pentacene - see figure 4.4) have radically different ratios of assigned crystal packing type. This is demonstrated in figure 5.52. It is interesting to note that only pentacene seems to possess compounds with the herringbone, sandwich herringbone and ″other″ categories. Also note that the CSP landscapes 7A, 7B and 7C are completely dominated by crystal structures with the beta (sheet) label. We do not choose to discuss these compounds further as the structures in these landscapes with the gamma packing class are so sparse that it is difficult to ascertain any trends in their positions on any crystal structure landscape which we create.

Thus we shall discuss the properties of five sets of crystal structure landscapes which we can interpret by means of the persistent homology descriptors which we have developed.

The first structure we consider is the compound labelled 5A. At this point we are only interested in the persistent homology that arises from the 3D pointcloud, that is, the set of molecular centroids (once again we use 50). As usual we find the principal components of the vector images obtained from the associated persistence diagram (alpha filtration). For reasons that shall become clear we remove all diagrams that contain any features with a maximum death past a certain cutoff value (i.e. involving

```
Algorithm 3 Polyaromatic Hydrocarbon Packing Determination (Campbell)
procedure CLASSIFY CRYSTAL($\{\mathrm{mol}_i\} \in$ crystal)    ▷ Label a PAH as herringbone, gamma, beta or sandwich herringbone
    Choose a reference molecule, $\mathrm{mol}_r$, from the crystal
    Define $\{N_i\} = \{\mathrm{mol} \in \mathrm{crystal} : d(\mathrm{mol}, \mathrm{mol}_r) < 20\text{Å}\}$ ▷ Select all mols in crystal within 20 Å radius from reference
    if $\angle(\mathrm{mol}_r, N_i) < 25 \quad \forall \quad i$ then:    ▷ If interplanar angle is small for all neighbours, classify as beta (or sheet)
        return beta
    end if
    Consider four nearest neighbours of $\mathrm{mol}_r$ $n_0, n_1, n_2, n_3$
    if $\angle(\mathrm{mol}_r, N_i) > 25^\circ$ for $n_0, n_1, n_2, n_3$ then :    ▷ If none of four nearest neighbours are approx. parallel, structure is herringbone
        return herringbone
    end if
    if Only one of $n_0, n_1, n_2, n_3$ has $\angle$ with $\mathrm{mol}_r$ $< 25^\circ$ then:
        return sandwich herringbone
    end if
    if Two or three of $n_0, n_1, n_2, n_3$ have $\angle$ with $\mathrm{mol}_r$ $< 25^\circ$ then:
        return gamma
    end if
end procedure
```

FIGURE 5.51: The algorithm used by Campbell *et al.* Campbell et al. (2017) to classify PAH crystals into one of four packing classes using interplanar angles. The Azapentacene dataset which we are using has been classified in the same manner.

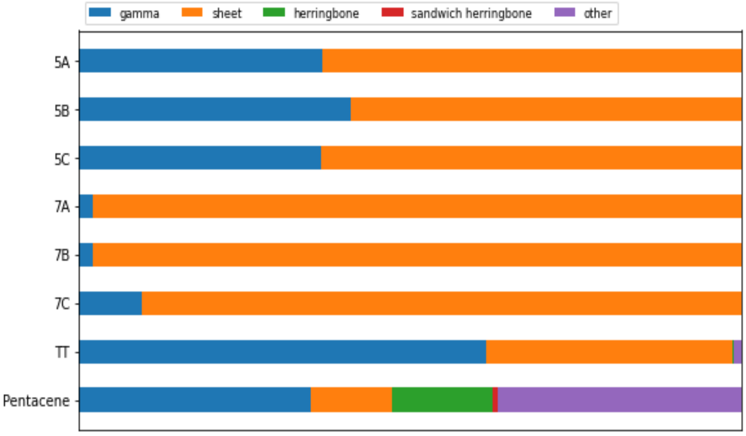


FIGURE 5.52: The ratio of compounds in the crystal structure landscapes of eight azapentacenes assigned different packing labels by a heuristic algorithm which uses the interplanar angles within the crystal structure to assign the crystal structure to one of five packing classes - beta (sheet), gamma, herringbone,other, and sandwich herringbone.

any simplex with edge length greater than this value). We consider the PCA plots obtained when 0D, 1D, 2D, 1+2D and 0+1+2D features are used, each with a set of 5 different maximum deaths (or cutoffs): 30 Å,40 Å, 50 Å, 60 Å and an infinite maximum death (this corresponds to just keeping all the crystal structures as normal). As such we end up with the matrix of 25 PCA plots shown in figure 5.53.

Note how the PCA plots are very uninformative when the infinite maximum death is used due to the presence of compounds whose persistent homology features dominate the variance of the total set of vector images. Further analysis of these outlier structures reveals that these correspond to the unphysical crystal structure predictions, which have long and thin unit cell, alluded to in our previous discussion. This is due to the fact that these data were generated using an older version of the CSP code for which the pseudo-random sampling technique was not yet fully optimised such that sensible random cell parameters were chosen. The upshot of this is that these compounds end up having persistent homology features that occur over much larger length scales than any other features in the set of crystal structures and which thus end up dominating the variance on the set of vector images. An easy way of dealing with this without having to resort to outlier detection methods (with which we had more limited success) is to set the maximum value for the death of any given persistent homology feature. If the persistence diagram contains a feature with a maximum death above this threshold, the diagram is deemed unphysical and is removed [8] It can be seen in figure 5.53 that even at a cutoff of 60 Å, most of the effects of these unphysical data are removed. It is difficult to ascertain, however, how much of the persistent homology information should be jettisoned in order to be sure of the removal of all unphysical cycles. For example while it is clear in figure 5.53 that while setting the cutoff to 60 Å clearly removes any significant ”outlier” points on our crystal structure landscapes, perhaps the ”best” landscape - in terms of class separation and lack of the clumping of points in a particular region (which may make the landscape harder to interpret) - is obtained when the cutoff is much lower at 40 Å. There seems to be a trade-off between losing topological information by excluding useful homology features and the skewing the crystal structure landscape unnecessarily by the inclusion of unphysical homology features into our landscapes.

This is examined in more detail in figure 5.54 . Here the AMI of the (K-means) clustering of the persistence diagrams against the four packing classes is plotted against the maximum allowable death value for each set of homology features. Also plotted is the percentage of crystal structures that are included in our model at each value of maximum death (so that when the maximum death is infinite 100 % of

[8] One might ask why the death values of the persistent homology features are used for the filtering out of unphysical diagrams and not the birth values, all features clearly being described by both of these numbers. We found empirically that when we analysed the homology features that skew the crystal structure landscape, these were invariably those with a high death value and not those features born at a high filtration value

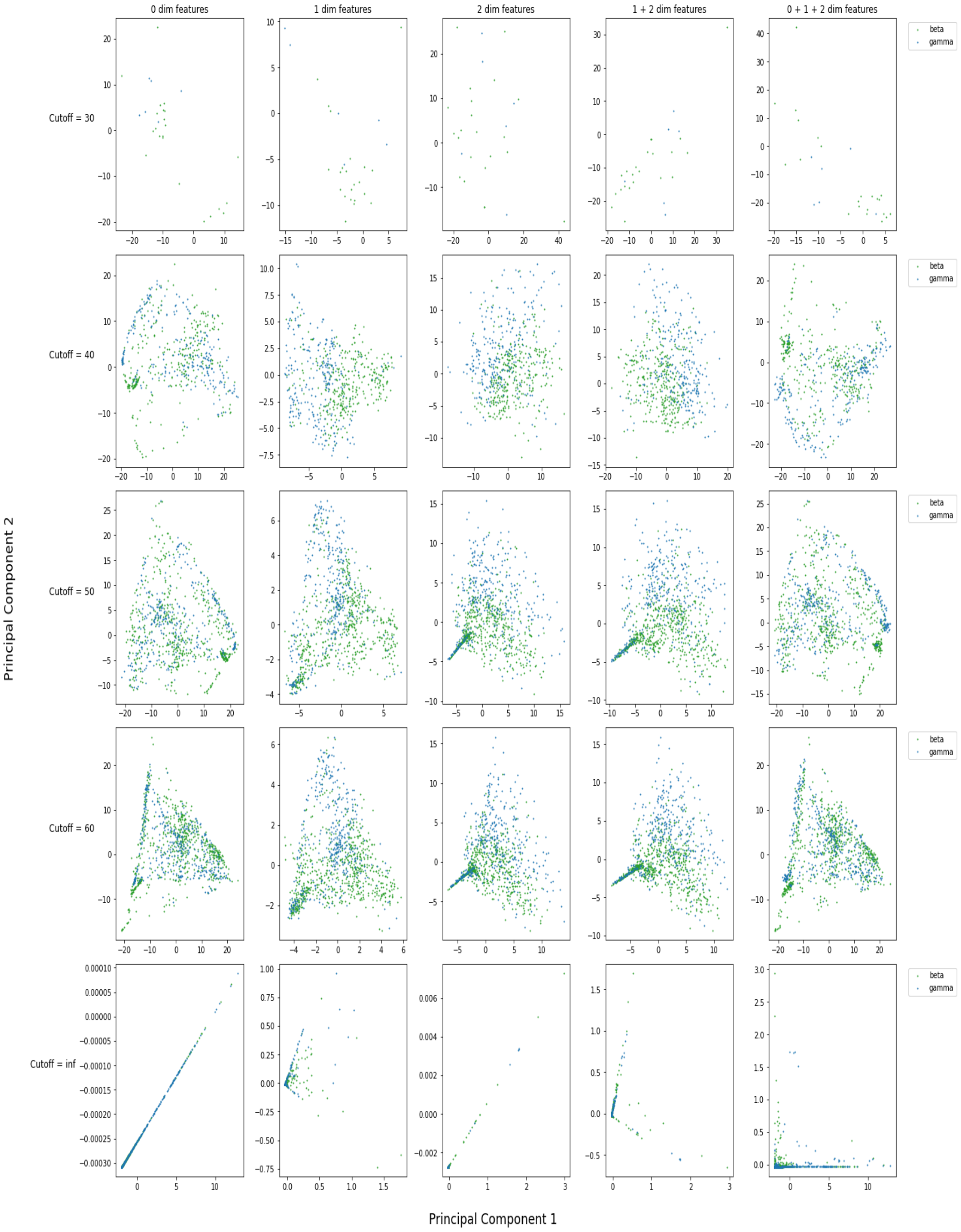


FIGURE 5.53: The set of plots obtained by dimensionality reduction of the vector images corresponding to the persistence diagrams of a large set of predicted crystal structures of the azapentacene named here as 5A. These structures are labelled according to a heuristic algorithm involving the interplanar angles of sets of neighbouring molecules. There are 25 plots corresponding to both the dimension of persistent homology features to be considered and the maximum death value allowed on any given diagram (those diagrams which have a feature that dies past this value are removed).

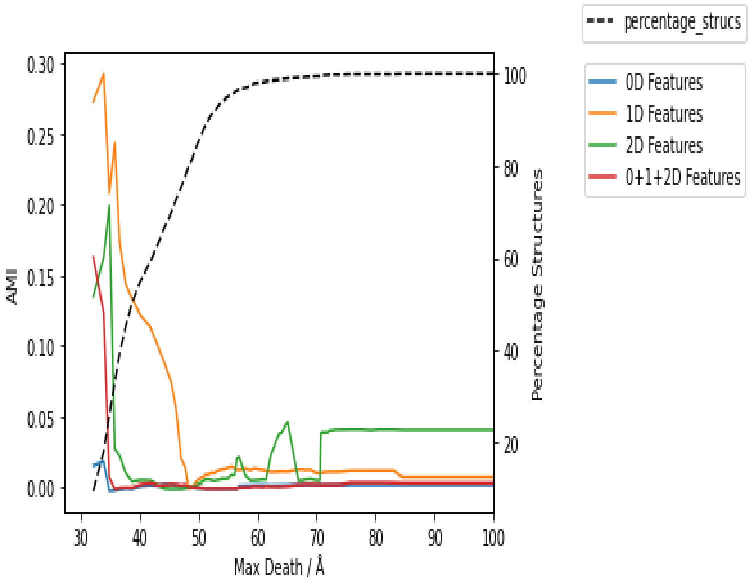


FIGURE 5.54: The Adjusted Mutual Information of a clustering of vector images (obtained from persistence diagrams of azapentacene 5A) against four packing classes (described by a heuristic geometric argument) plotted against the maximum death value (any persistence diagram which contains a homology feature with a death value past this maximum are *not* considered in our analysis) for each dimension of homology feature (or combinations thereof). Also plotted is the percentage of crystal structures that are included in our model at each value of maximum death

structures are used). First note how much lower the AMI is in general for this dataset compared with the fluoroanalines - persistent homology seems to have a lot less predictive power over packing type in this case (in spite of the fact that crystal structures in this dataset correspond to *the same compound* in contrast with the other datasets we have considered thus far). Also note that we get the best fit when the cutoff is low and there are comparatively fewer crystal structures in the dataset. While the removal of unphysical structures is clearly important note that the higher AMI found when the maximum death is low may simply be due to the fact that when there are numerically fewer structures in the dataset, any arbitrary labelling is likely to match whatever ground truth you can define by pure chance - there are simply fewer combinations for the packing labels at this point. It is also interesting to note that the 0D persistent homology features do not seem to carry the most predictive power in this case.

Lastly we can also circumvent the problem of the unphysical structures by constructing the crystal structure landscape by embedding the matrix of Wasserstein distances [9] between persistence diagrams into 2D space using MultiDimensional

[9] The Wasserstein distance is only strictly defined for homology features of a given dimension, that is one can only find the Wasserstein distance for 0D features only, or 1D features only etc. In practise we just define a structure with the birth-death pairs for all dimensions and feed this into the algorithm. Another, perhaps a more robust approach is to define the distance between persistence diagrams as the sum of the Wasserstein distances in each dimension. In practise this gives us pretty much the same results.

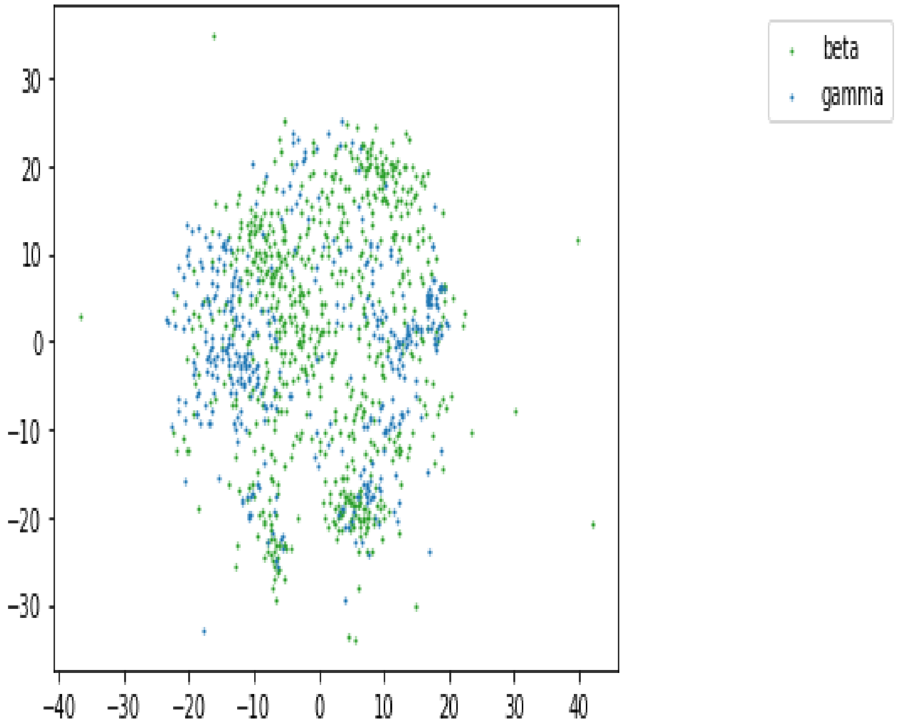


FIGURE 5.55: MDS embedding of the matrix of Wasserstein distances between sets of persistence diagrams which correspond to predicted crystal structures for the azapentacene labelled 5A. The compounds are labelled according to their packing type predicted by a heuristic algorithm using interplanar angles of neighbouring molecules. The axes of the space into which the data are embedded have no physical meaning.

Scaling (MDS). Any of the homology features which arise in the unphysical structures has a very limited impact on the Wasserstein metric (of order 2) presumably due to the fact that the single homology features have a very small contribution to the sum of mappings between points while one can imagine a single homology feature at an extreme value having a large effect on the vector image. While this takes a very long time to compute compared to the vector images the results are much more promising as shown in figure 5.55. The packing classes are not separated *per se* as in the case of the fluorolanines but there are certain regions of the embedding that have a clear preference for structures of a given packing type which indicates that the persistent homology is telling us something about the packing type even if it cannot necessarily predict it.

We now repeat these plots for the remaining azapentacenes in the dataset.

The set of PCA plots obtained with vector images, the max death/number of structures trade-off and the Wasserstein embedding for azapentacene 5B are shown respectively in figures 5.56, 5.57 and 5.58. The set of plots are relatively similar but note that first of all the PCA plots with a large cutoff show less concentration of points in a given region than those of 5A making these plots easier to interpret. Figure 5.57

reveals that the clustering on this data is more faithful to the heuristic packing labels than in the case of 5A. Lastly note that the Wasserstein embedding shows a similar structure to that of 5A, that is various ″lobes″ which mostly correspond to one packing class. It must again be noted that this was very time consuming to plot (around 1 day on a MacBook Pro) so these embeddings may not be a practical method for exploring the crystal structure landscapes of these compounds.

We repeat this process for azapentacene 5C.

The set of PCA plots obtained with vector images, the max death/number of structures trade-off and the Wasserstein embedding for azapentacene 5C are shown respectively in figures 5.59, 5.60 and 5.61. These data are generally harder to interpret due to the smaller size of this dataset - we nevertheless see similar trends to 5B with the distribution of packing classes across the PCA and MDS plots. We observe that the AMI is general quite poor and note that the spike in AMI at low cutoff is almost certainly due to the increased probability of a random labelling being correct when the sample size is small.

We repeat this process for azapentacene TT.

The set of PCA plots obtained with vector images, the max death/number of structures trade-off and the Wasserstein embedding for azapentacene TT are shown respectively in figures 5.62, 5.63 and 5.64. Both the AMI and the PCA plots reveal that this compound does not seem particularly amenable to a description of packing by persistent homology - it is unclear why the results for this compound are poor with respect to the others. Note especially that the Wasserstein embedding in figure 5.64 no longer has the nice ″lobe″ structures observed in the above examples.

Finally we examine the results we obtain for the predicted structures of the pentacene.

The set of PCA plots obtained with vector images, the max death/number of structures trade-off and the Wasserstein embedding for pentacene are shown respectively in figures 5.65, 5.66 and 5.67. Perhaps mostly due to the more diverse set of predicted packing types for this compound, the results are also poor and it is hard to ascertain any underlying trends in the data.

While these results show some interesting trends and there is some evidence that persistent homology can indeed at least partially describe the (predicted) packing types of some of these crystal structures (particularly compounds 5A and 5B), the presence of the unphysical crystal structures is concerning and having to artificially prune these out of the dataset is far from ideal. Are there other unphyscial structures adversely effecting the results in other ways? It is hence desirable to consider the persistent homology of the crystal structure predictions for the same set of compounds but with more up to date code where some of these problems are avoided.

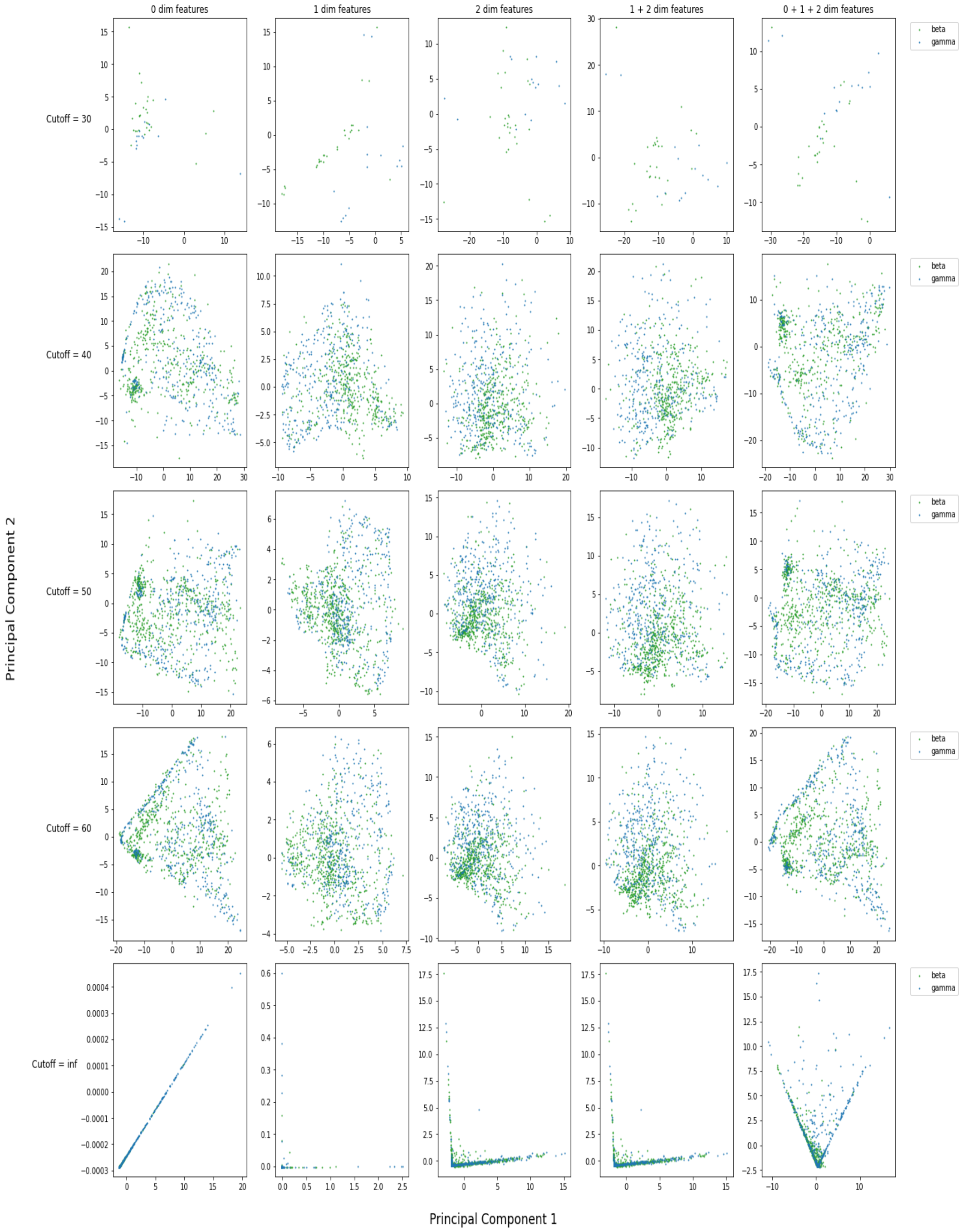


FIGURE 5.56: The set of plots obtained by dimensionality reduction of the vector images corresponding to the persistence diagrams of a large set of predicted crystal structures of the azapentacene named here as 5B. These structures are labelled according to a heuristic algorithm involving the interplanar angles of sets of neighbouring molecules. There are 25 plots corresponding to both the dimension of persistent homology features to be considered and the maximum death value allowed on any given diagram (those diagrams which have a feature that dies past this value are removed).

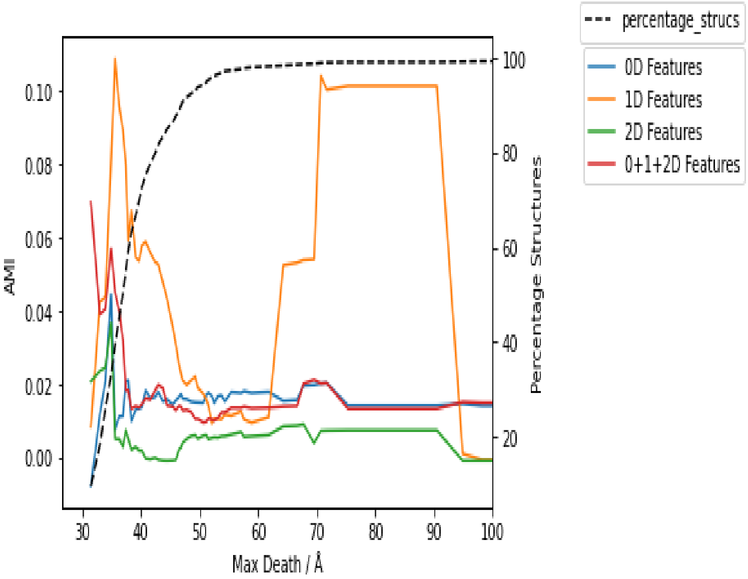


FIGURE 5.57: The Adjusted Mutual Information of a clustering of vector images (obtained from persistence diagrams of azapentacene 5B) against four packing classes (described by a heuristic geometric argument) plotted against the maximum death value (any persistence diagram which contains a homology feature with a death value past this maximum are *not* considered in our analysis) for each dimension of homology feature (or combinations thereof). Also plotted is the percentage of crystal structures that are included in our model at each value of maximum death

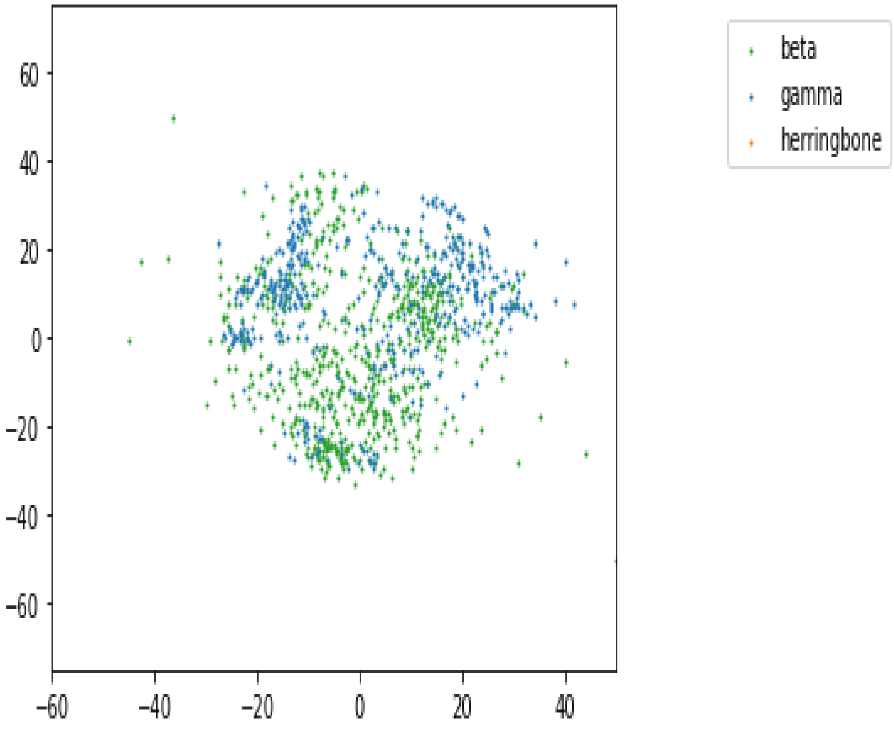


FIGURE 5.58: MDS embedding of the matrix of Wasserstein distances between sets of persistence diagrams which correspond to predicted crystal structures for the azapentacene labelled 5B. The compounds are labelled according to their packing type predicted by a heuristic algorithm using interplanar angles of neighbouring molecules. The axes of the space into which the data are embedded have no physical meaning.

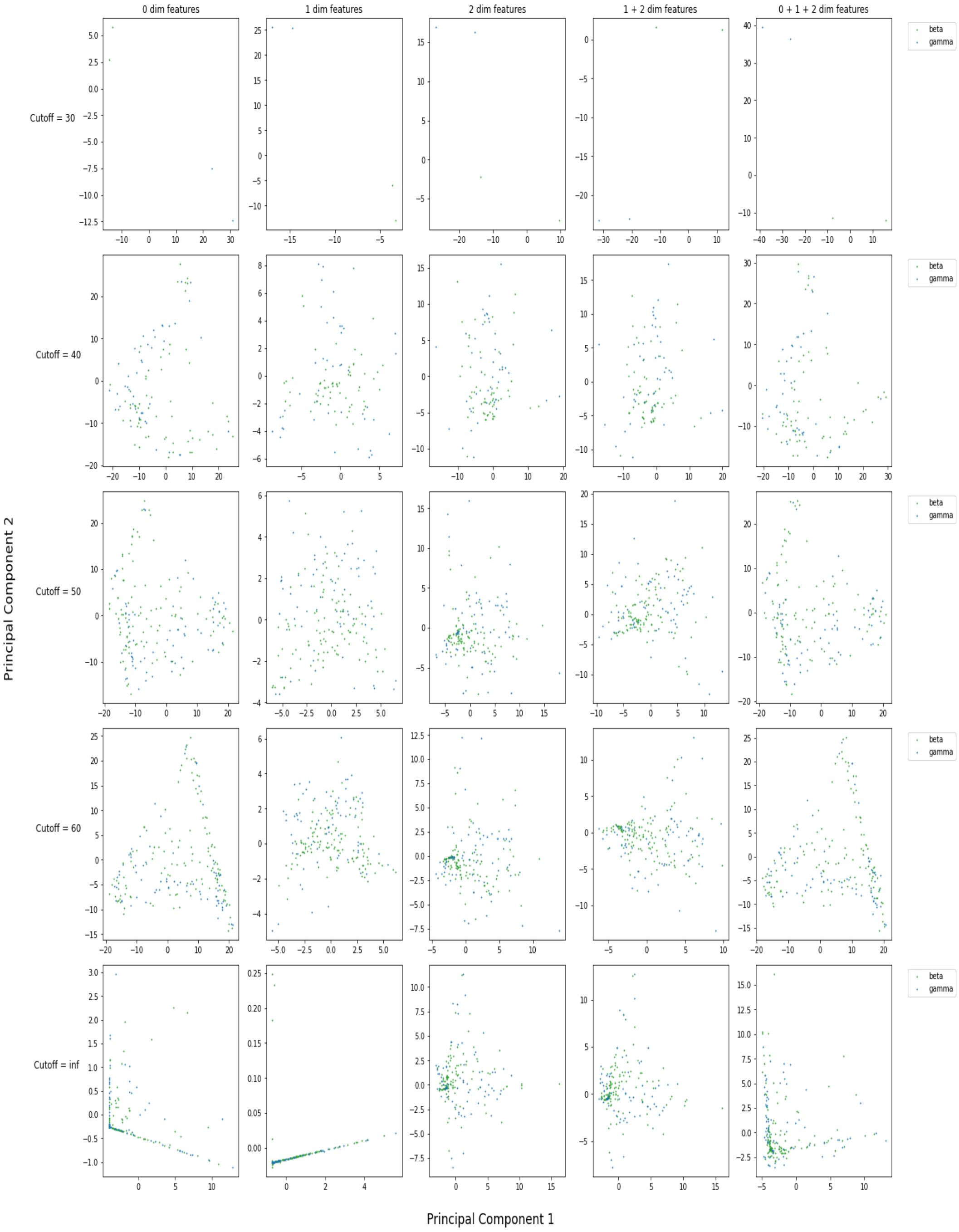


FIGURE 5.59: The set of plots obtained by dimensionality reduction of the vector images corresponding to the persistence diagrams of a large set of predicted crystal structures of the azapentacene named here as 5C. These structures are labelled according to a heuristic algorithm involving the interplanar angles of sets of neighbouring molecules. There are 25 plots corresponding to both the dimension of persistent homology features to be considered and the maximum death value allowed on any given diagram (those diagrams which have a feature that dies past this value are removed).

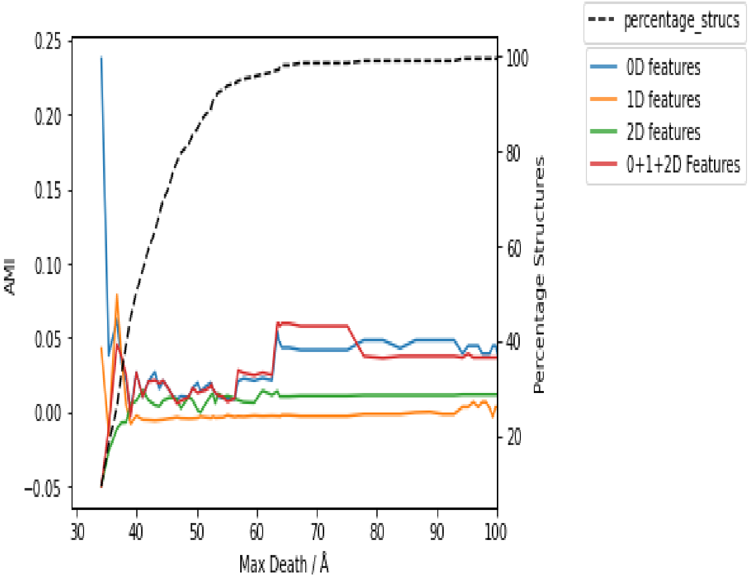


FIGURE 5.60: The Adjusted Mutual Information of a clustering of vector images (obtained from persistence diagrams of azapentacene 5C) against four packing classes (described by a heuristic geometric argument) plotted against the maximum death value (any persistence diagram which contains a homology feature with a death value past this maximum are *not* considered in our analysis) for each dimension of homology feature (or combinations thereof). Also plotted is the percentage of crystal structures that are included in our model at each value of maximum death

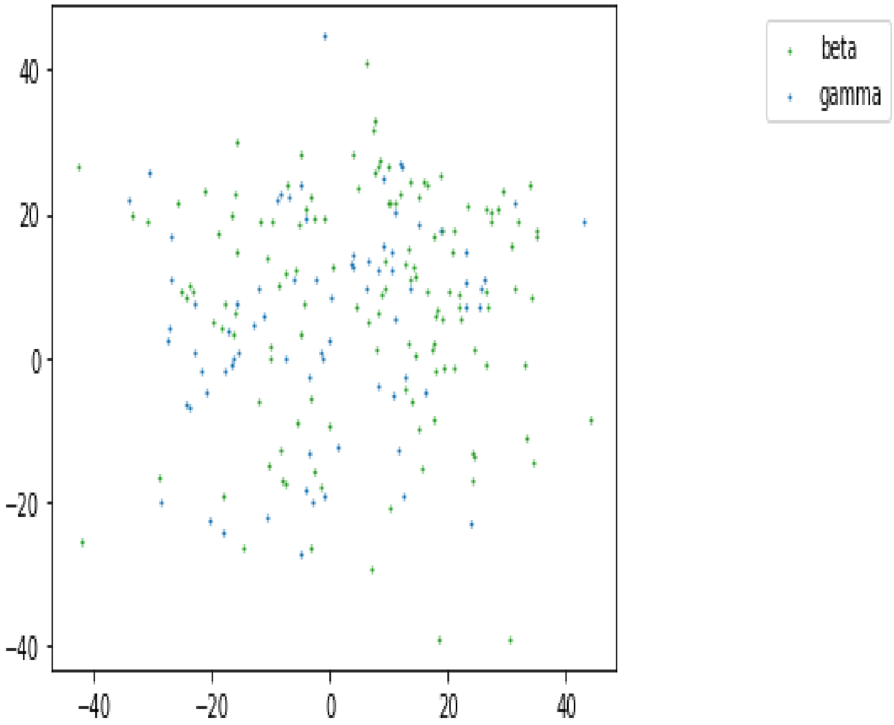


FIGURE 5.61: MDS embedding of the matrix of Wasserstein distances between sets of persistence diagrams which correspond to predicted crystal structures for the azapentacene labelled 5C. The compounds are labelled according to their packing type predicted by a heuristic algorithm using interplanar angles of neighbouring molecules. The axes of the space into which the data are embedded have no physical meaning.

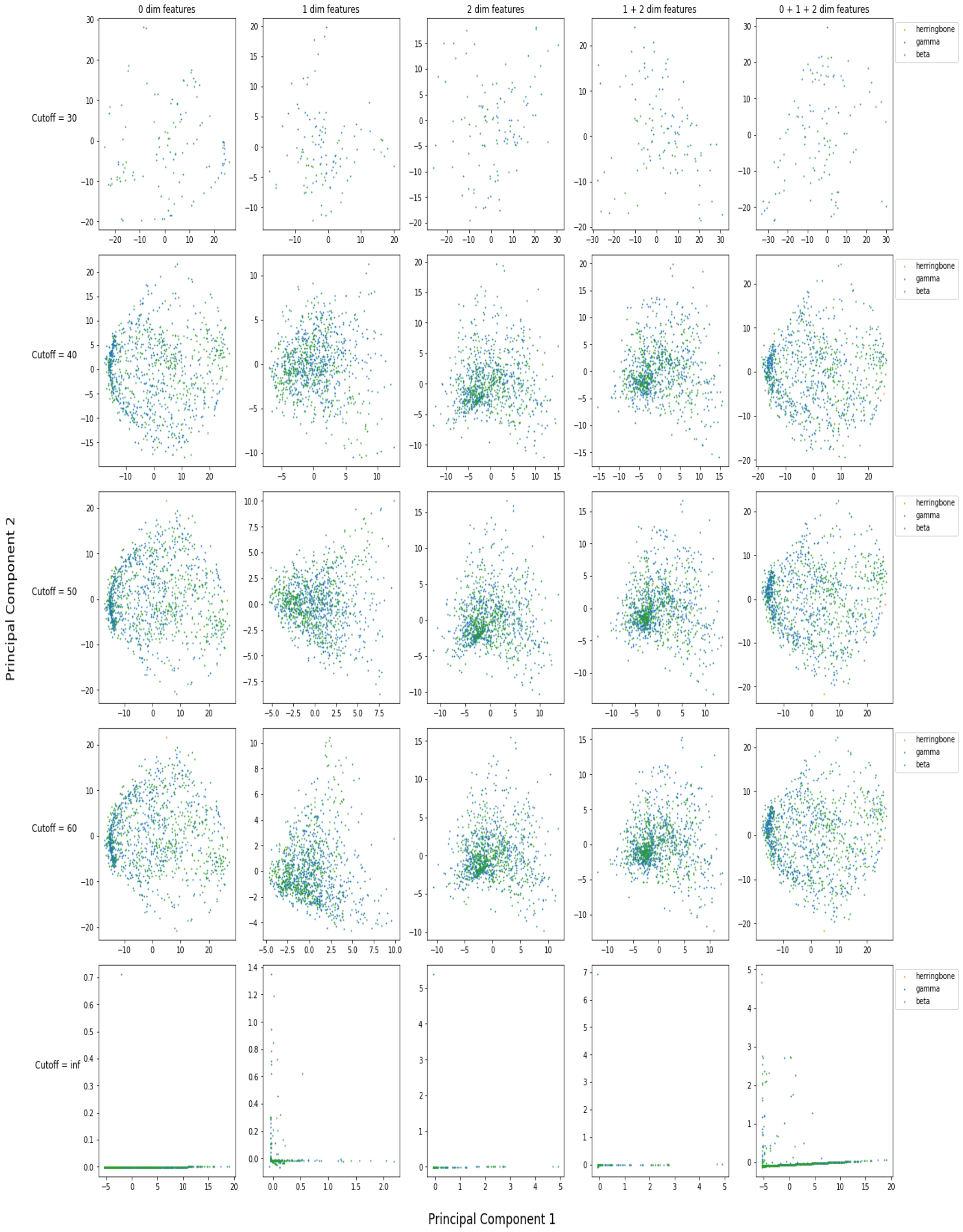


FIGURE 5.62: The set of plots obtained by dimensionality reduction of the vector images corresponding to the persistence diagrams of a large set of predicted crystal structures of the azapentacene named here as TT. These structures are labelled according to a heuristic algorithm involving the interplanar angles of sets of neighbouring molecules. There are 25 plots corresponding to both the dimension of persistent homology features to be considered and the maximum death value allowed on any given diagram (those diagrams which have a feature that dies past this value are removed).

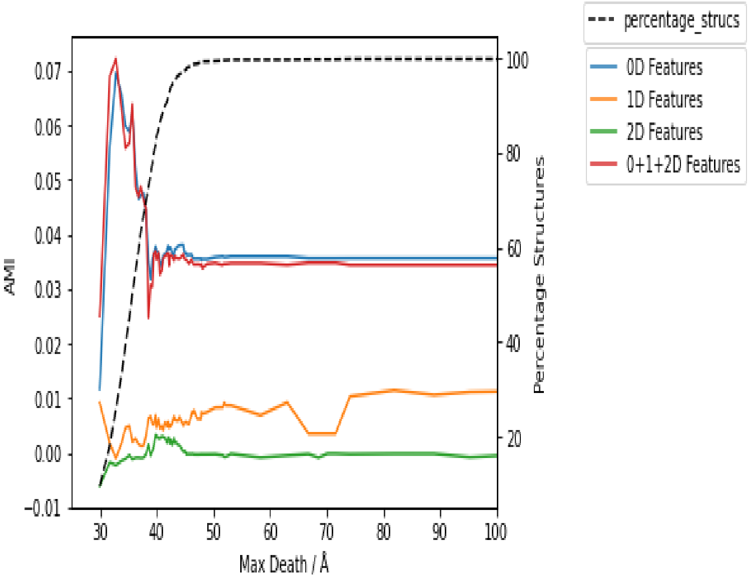


FIGURE 5.63: The Adjusted Mutual Information of a clustering of vector images (obtained from persistence diagrams of azapentacene TT) against four packing classes (described by a heuristic geometric argument) plotted against the maximum death value (any persistence diagram which contains a homology feature with a death value past this maximum are *not* considered in our analysis) for each dimension of homology feature (or combinations thereof). Also plotted is the percentage of crystal structures that are included in our model at each value of maximum death

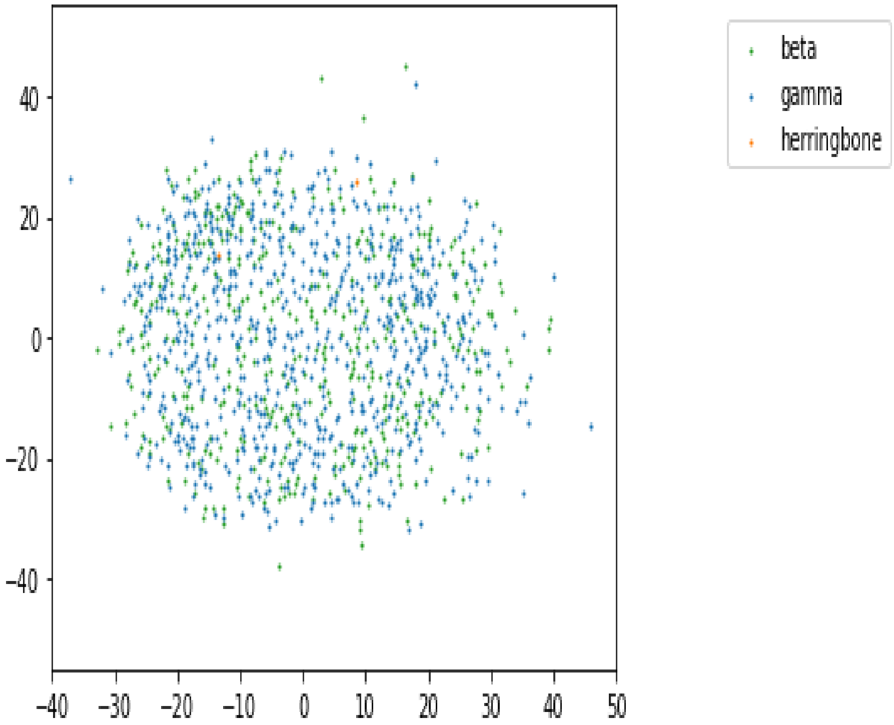


FIGURE 5.64: MDS embedding of the matrix of Wasserstein distances between sets of persistence diagrams which correspond to predicted crystal structures for the azapentacene labelled TT. The compounds are labelled according to their packing type predicted by a heuristic algorithm using interplanar angles of neighbouring molecules. The axes of the space into which the data are embedded have no physical meaning.

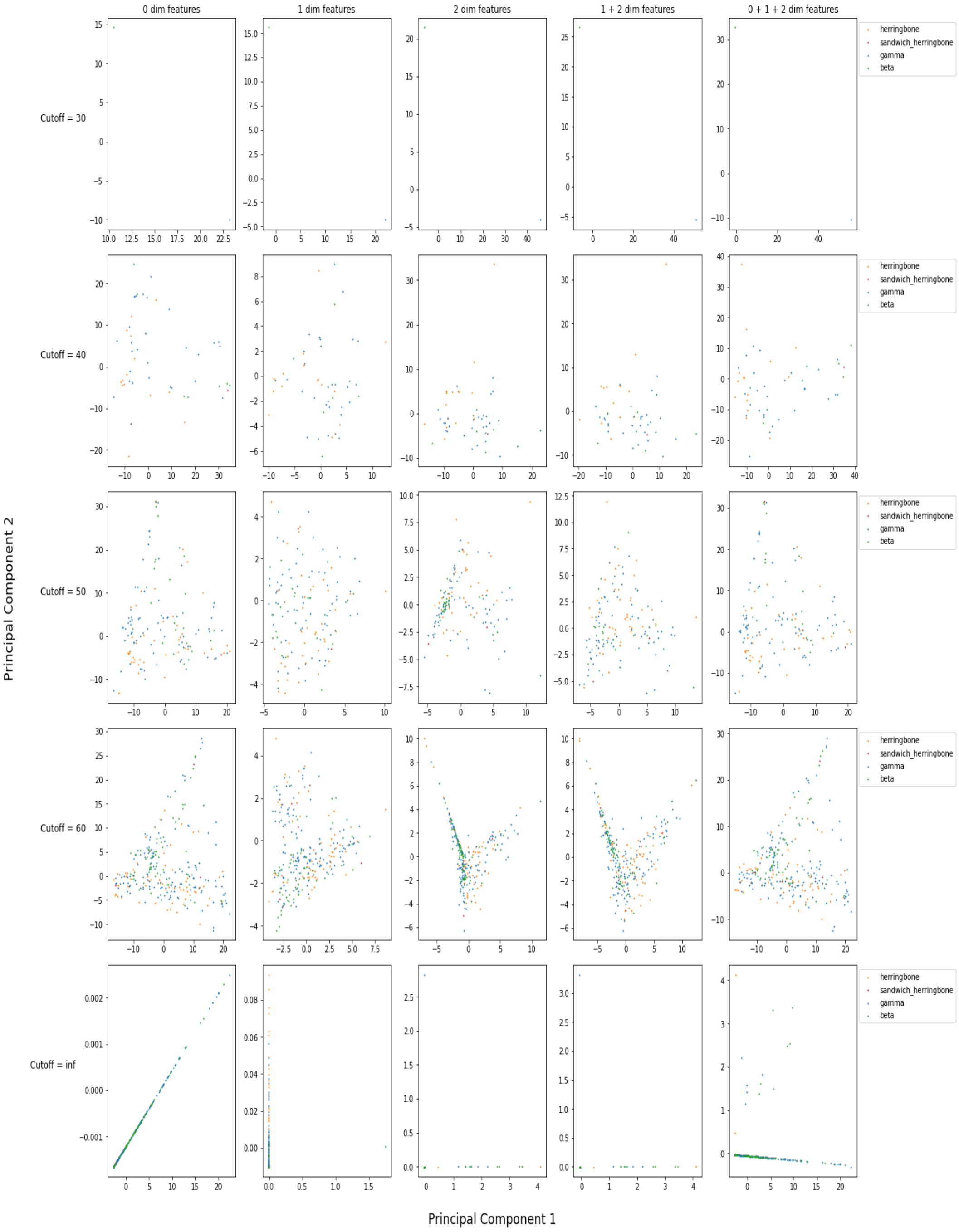


FIGURE 5.65: The set of plots obtained by dimensionality reduction of the vector images corresponding to the persistence diagrams of a large set of predicted crystal structures of pentacene. These structures are labelled according to a heuristic algorithm involving the interplanar angles of sets of neighbouring molecules. There are 25 plots corresponding to both the dimension of persistent homology features to be considered and the maximum death value allowed on any given diagram (those diagrams which have a feature that dies past this value are removed).

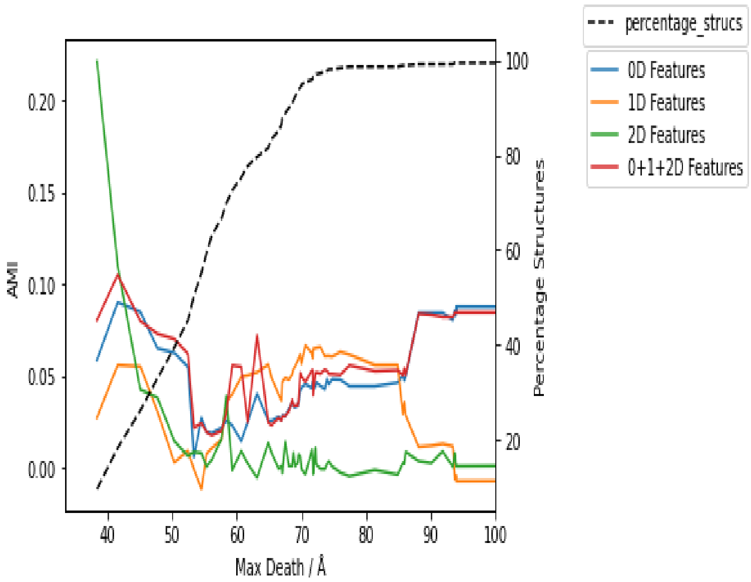


FIGURE 5.66: The Adjusted Mutual Information of a clustering of vector images (obtained from persistence diagrams of pentacene) against four packing classes (described by a heuristic geometric argument) plotted against the maximum death value (any persistence diagram which contains a homology feature with a death value past this maximum are *not* considered in our analysis) for each dimension of homology feature (or combinations thereof). Also plotted is the percentage of crystal structures that are included in our model at each value of maximum death

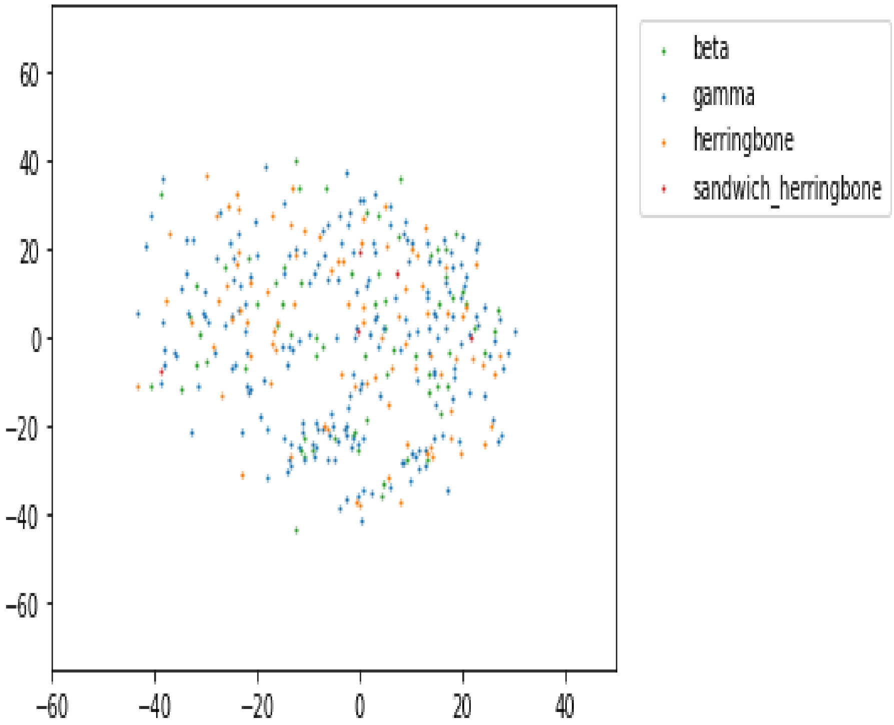


FIGURE 5.67: MDS embedding of the matrix of Wasserstein distances between sets of persistence diagrams which correspond to predicted crystal structures for pentacene. The compounds are labelled according to their packing type predicted by a heuristic algorithm using interplanar angles of neighbouring molecules. The axes of the space into which the data are embedded have no physical meaning.

Moreover there are reasons to doubt the fidelity of the labelling algorithm of Campbell *et al.* Campbell et al. (2017). In the first instance it seems rather odd that there are almost no examples in most of the data of compounds adopting the herringbone and sandwich herringbone structures: since structures of this packing type are very common it seems surprising that there were *no* higher energy structures on the crystal structure landscape of 5A and 5B that adopt this packing type. At any rate more data encompassing all four packing types would be desirable for us as we need to establish the interplay of all four of the common packing classes of these compounds and their persistent homology. It is hence very worthwhile to consider alternative (and more contemporary) algorithms which can solve the same problem. A very good candidate for this is the *Autopack* algorithm of Loveland *et al.* Loveland et al. (2020) which is built on the same principles of the algorithm of Campbell *et al.* Campbell et al. (2017) but which attempts to resolve some inherent limitations of this algorithm. We shall discuss this method in detail in the next section.

5.3.3 Azapentacene CSP Calculations: Part 2

As mentioned earlier we only carried out further CSP calculations for four compounds: 5A, 5B, 5C and pentacene. For compounds 5A, 5B and 5C we only computed the crystal structres with $Z' = 1$ while for pentacene we also computed structures where $Z' = 2$. In contrast to the Musil *et al.* Musil et al. (2018) we only computed the CSP landscape for the 10 most common spacegroups - the default option - as a starting point - it is possible that we missed out on some of the richness of the resulting crystal landscapes by making this choice. Nevertheless this resulted in 9384 structures for 5A; 3971 structures for 5B; 14918 structures for 5C and 8289 & 5271 structures for pentacene with $Z' = 1$ and $Z' = 2$ respectively so hopefully we have enough data to showcase a large variety of different packing types.

These structures were classified into the four canonical packing types using the *Autopack* algorithm Loveland et al. (2020) which we describe below.

The central issue with the algorithm of Campbell *et al.* Campbell et al. (2017) identified by the authors of *Autopack* Loveland et al. (2020) is that the nearest neighbours of the reference molecule (see figure 5.51) are chosen on the basis of intercentroid distance alone. This can lead to an erroneous assignment as these nearest neighbours may not live in the plane that identifies the packing motif. We exemplify this in figure 5.68 (reproduced from Loveland *et al.* Loveland et al. (2020)). Observe that the compound shown (4,5-diphenylbenzo[e]pyrene) would be incorrectly assigned the label gamma by our current algorithm when this structure clearly takes the sandwich herringbone motif. This is because the wrong neighbours are chosen which lie in a different plane to that describing the crystal motif. Thus the wrong sets of angles would be used to

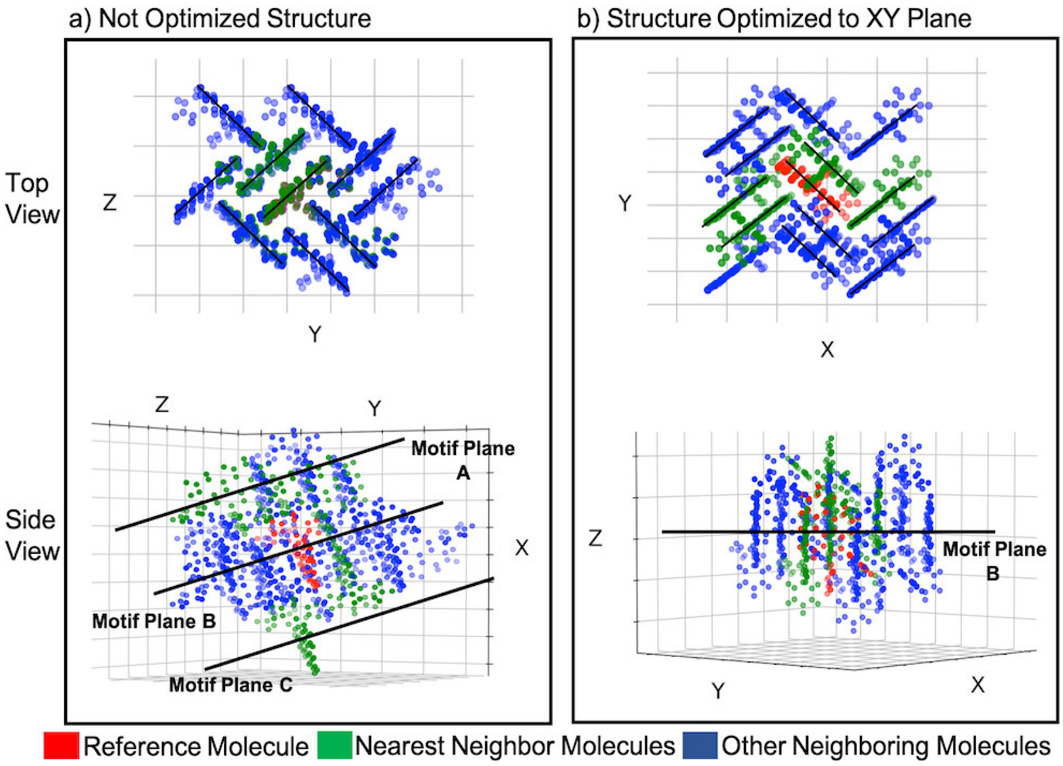
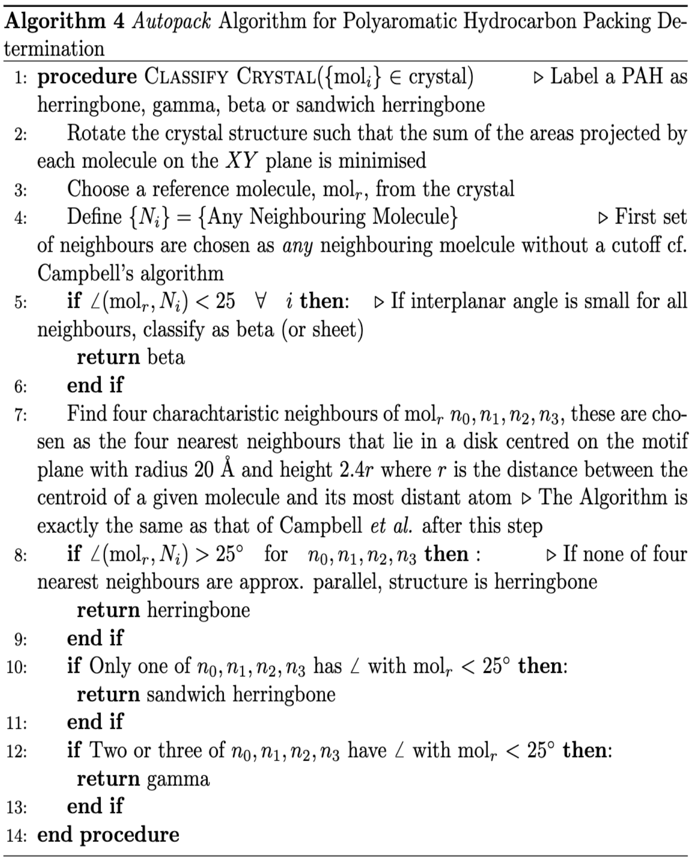


FIGURE 5.68: Figure reproduced from Loveland *et al.* Loveland et al. (2020). This figure shows the crystal structure of 4,5-diphenylbenzo[e]pyrene which takes the sandwich herringbone packing motif. If the crystal is not properly optimised as in part a) the nearest neighbours chosen relative to the reference molecule which are used for packing assignment may not lie in the plane which describes the packing motif and as such will result in an incorrect assignment of this structure to the gamma packing class. In part b) this crystal is rotated such that the plane that describes the crystal motif is parallel to the XY plane. If nearest neighbours are chosen relative to this plane then the correct set of nearest neighbours are chosen and this structure is correctly labelled as sandwich herringbone.

assign the structure. This problem is avoided if the crystal structure is rotated so that the motif plane is parallel to the XY plane (which the authors take to be the viewing plane) and then only molecules in the same plane as the reference molecule are eligible to be classed as neighbours. This is the method used by the *Autopack* algorithm to resolve these potential mischaracterisation issues.

In practice this optimal rotation is found by choosing the rotation that minimises the sum of the projected areas of each the molecules (these being modelled themselves as planes) onto the viewing plane (the XY plane). A fragment based approach is used here - typically about 50 molecules are used in this process Loveland et al. (2020). Once the crystal structure is rotated all neighbours are chosen according to this motif plane. These so-called characteristic neighbours do not actually have to lie on the motif plane itself as for some crystal structures the molecules in a stack can be tilted so that they deviate from the plane significantly. The upshot is that, in practice, the characteristic neighbours are restricted to a 3D disk around the reference molecule of



Algorithm 4 *Autopack* Algorithm for Polyaromatic Hydrocarbon Packing Determination

```
procedure CLASSIFY CRYSTAL($\{\mathrm{mol}_i\} \in$ crystal)    ▷ Label a PAH as herringbone, gamma, beta or sandwich herringbone
    Rotate the crystal structure such that the sum of the areas projected by each molecule on the $XY$ plane is minimised
    Choose a reference molecule, $\mathrm{mol}_r$, from the crystal
    Define $\{N_i\} = \{$Any Neighbouring Molecule$\}$    ▷ First set of neighbours are chosen as *any* neighbouring moelcule without a cutoff cf. Campbell's algorithm
    if $\angle(\mathrm{mol}_r, N_i) < 25 \quad \forall \quad i$ then:    ▷ If interplanar angle is small for all neighbours, classify as beta (or sheet)
        return beta
    end if
    Find four charachtaristic neighbours of $\mathrm{mol}_r$ $n_0, n_1, n_2, n_3$, these are chosen as the four nearest neighbours that lie in a disk centred on the motif plane with radius 20 Å and height $2.4r$ where $r$ is the distance between the centroid of a given molecule and its most distant atom    ▷ The Algorithm is exactly the same as that of Campbell *et al.* after this step
    if $\angle(\mathrm{mol}_r, N_i) > 25°$ for $n_0, n_1, n_2, n_3$ then :    ▷ If none of four nearest neighbours are approx. parallel, structure is herringbone
        return herringbone
    end if
    if Only one of $n_0, n_1, n_2, n_3$ has $\angle$ with $\mathrm{mol}_r < 25°$ then:
        return sandwich herringbone
    end if
    if Two or three of $n_0, n_1, n_2, n_3$ have $\angle$ with $\mathrm{mol}_r < 25°$ then:
        return gamma
    end if
end procedure
```

FIGURE 5.69: The *Autopack* Loveland et al. (2020) algorithm for the classification of the crystal structures of polyaromatic hydrocarbons into one of four canonical packing classes (beta, gamma, herringbone and sandwich herringbone). This algorithm mostly differs from that of Campbell *et al.* in that the crystal is rotated to find an optimal view of the crystal *before* any neighbours are computed and in that the four canonical neighbours for the classification into the classes of gamma, herringbone and sandwich herringbone are fixed into a disk centred on the motif plane.

radius 20 Å and with height $2ar$ where r is taken as the distance between the centroid of a given molecule and its most distant atom, and a is a constant multiplier - with optimal value empirically found to be 1.2. This information can be used to refine the algorithm in figure 5.51 to give the *Autopack* algorithm shown in figure 5.69.

When the new azapentacene crystal structure landscapes were classified into packing types, the different azapentacene landscapes have a very different ratio of packing classes as compared to the previous dataset (figure 5.52). This is shown in figure 5.70. Note the increased presence of the herringbone and sandwich herringbone classes for the azapentacenes with respect to the previous distribution. Also note that the pentacene crystal structure landscape appears to posses very different packing labels depending on the Z' used in the CSP algorithm. Assessing the interplay of Z' and crystal packing classes should be a topic for further study.

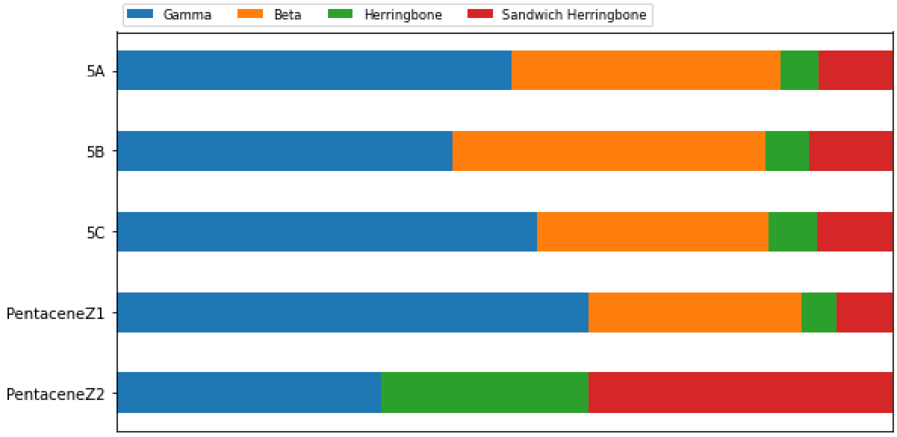


FIGURE 5.70: The ratio of compounds in the crystal structure landscapes of three azapentacenes and two sets of pentacene compounds (generated with different Z') assigned different packing labels by the *Autopack* algorithm Loveland et al. (2020) which uses the interplanar angles between neighbouring molecules on (or near) a specially selected motif plane to assign the crystal structure to one of four packing classes - beta (sheet), gamma, herringbone, and sandwich herringbone.

For each of the four compounds we use persistent homology to generate the following descriptors:

- A set of vector images from the persistent homology that is found from the set of 3D molecular centroids in a crystal fragment
- A set of vector images from the persistent homology that is found from the set of 3D molecular centroids *and* a suitably chosen inertia vector - so is computed in 6D
- A 6D persistent homology descriptor as above except the inertia vector is scaled by a factor of 25
- The persistence diagrams themselves (for the 6D unscaled case), that is we use the the Wasserstein distance (of order 2) to make an embedding of the distance matrix of persistence diagrams in a low dimensional space

For the compound 5A we also find the following two descriptors:

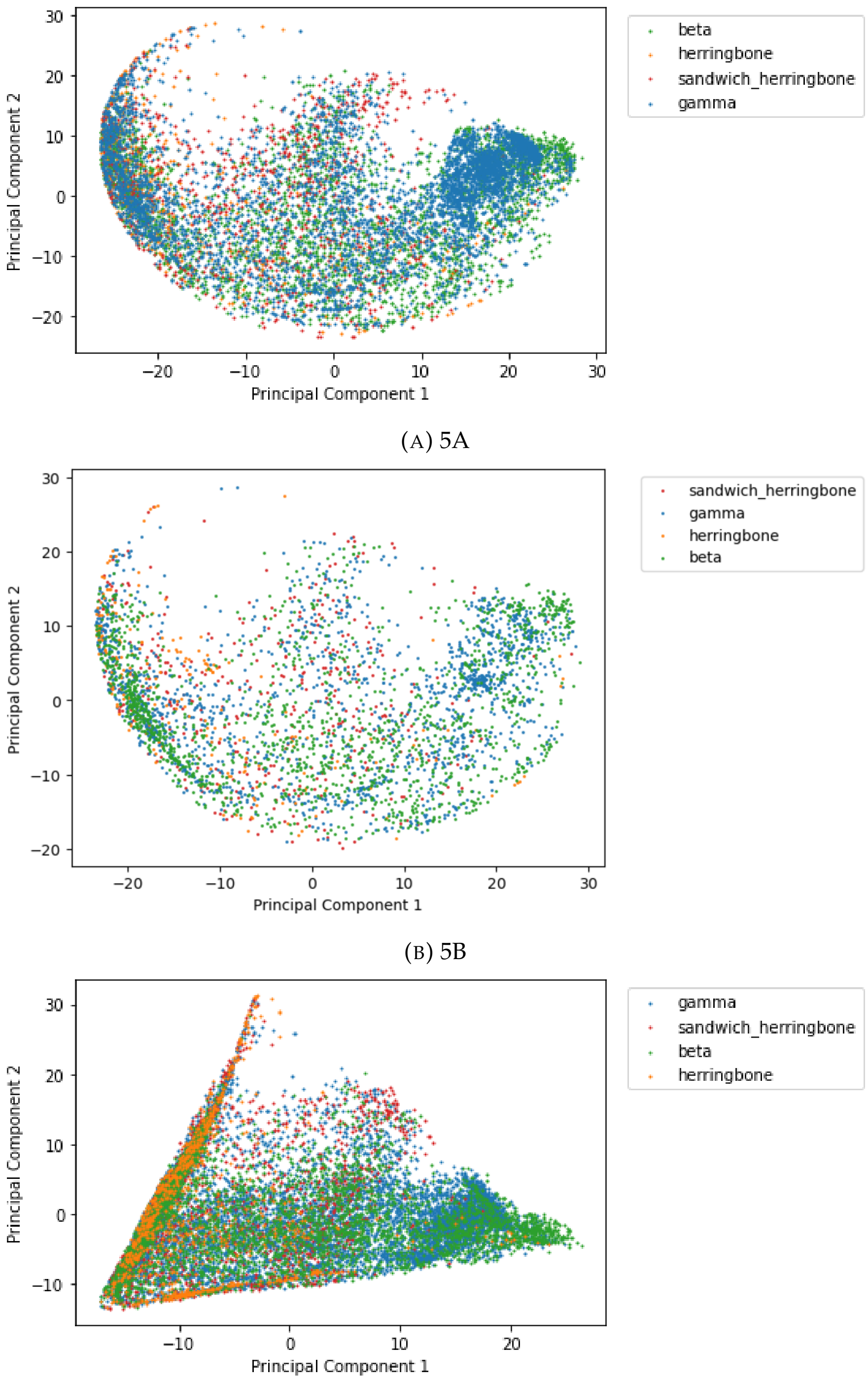
- The concatenation of four sets of vector images obtained from the persistence diagrams (from 6D) using crystal fragments built (respectively) around a $(R \times R \times 1)$, $(R \times 1 \times R)$, $(1 \times R \times R)$ and $(R \times R \times R)$ supercell (with arbitrarily large R).
- A descriptor that uses both the scaled inertia vector *and* the restricted supercells

In all cases we use only the 0,1 and 2 dimensional homology features. We use the Vietoris-Rips filtration so that we only calculate features up to 2D to save computation time.

We plot the two principal components of the first of these descriptors - the persistent homology from a 3D pointcloud - for 5A, 5B and 5C in figure 5.71. Clearly there is absolutely no discernible trend in the locations of the different packing types on each crystal structure landscape. There do appear to be certain regions which have a very large number of compounds with the same packing type but there are so many data points in this region it is difficult to establish if these are the *only* packing types in this region - some points are plotted very close together or on top of each other in the dense regions.

We repeat this process for the pentacene crystal structures: we plot the crystal structure landscape for $Z' = 1$, the crystal structure landscape for for $Z' = 2$ and the crystal structure landscape for all the pentacene structures (that is both for $Z' = 1$ *and* for $Z' = 2$). The results are shown in figure 5.72. The plots once again show pretty much no trends in the packing label but note that for the $Z' = 1$ case there are at least some regions that appear to have a high concentration of structures of a given packing type, while this is not the case at all for $Z' = 2$. Furthermore any structure that exists in the landscape for $Z' = 1$ is lost when the additional structures for $Z' = 2$ are added. We do not know what causes the structures for $Z' = 1$ to have a packing structure that is more readily described with persistent homology.

Adding in the inertia vectors into our model does not necessarily improve things much - in fact the plots look almost identical as those generated from 3D persistent homology. Figure 5.73 shows the PCA plots obtained for each of the azapentacens (5A, 5B and 5C) using this descriptor. We see a similar effect for the pentacene data - figure 5.74. It could be that molecular orientations are simply unimportant for the description of the packing type but this seems very unlikely - it is, after all, the interplanar angles between sets of neighbouring PAH molecules that determines the packing type. Hence our next step is to scale the vectors by a factor of 25 to increase the importance of the set of molecular orientations vs. their positions in Cartesian space. The crystal structure landscapes thus obtained are shown in figure 5.75 for the azapentacens and figure 5.76 for the pentacene data. These results are more promising. For each of the azapentacens there appear to be regions that favour certain packing types. This trend is by no means universal, however. Also note that, as before, in some of these regions there is a very large number of structures and hence data that do not belong to a given packing class are masked by those that do giving us the illusion that these regions of the crystal landscape favour one packing variety more strongly than is the case in reality. Examination of the plots one packing class at a time revealed that this masking behaviour is prominent, that is, most of the dense regions of these plots

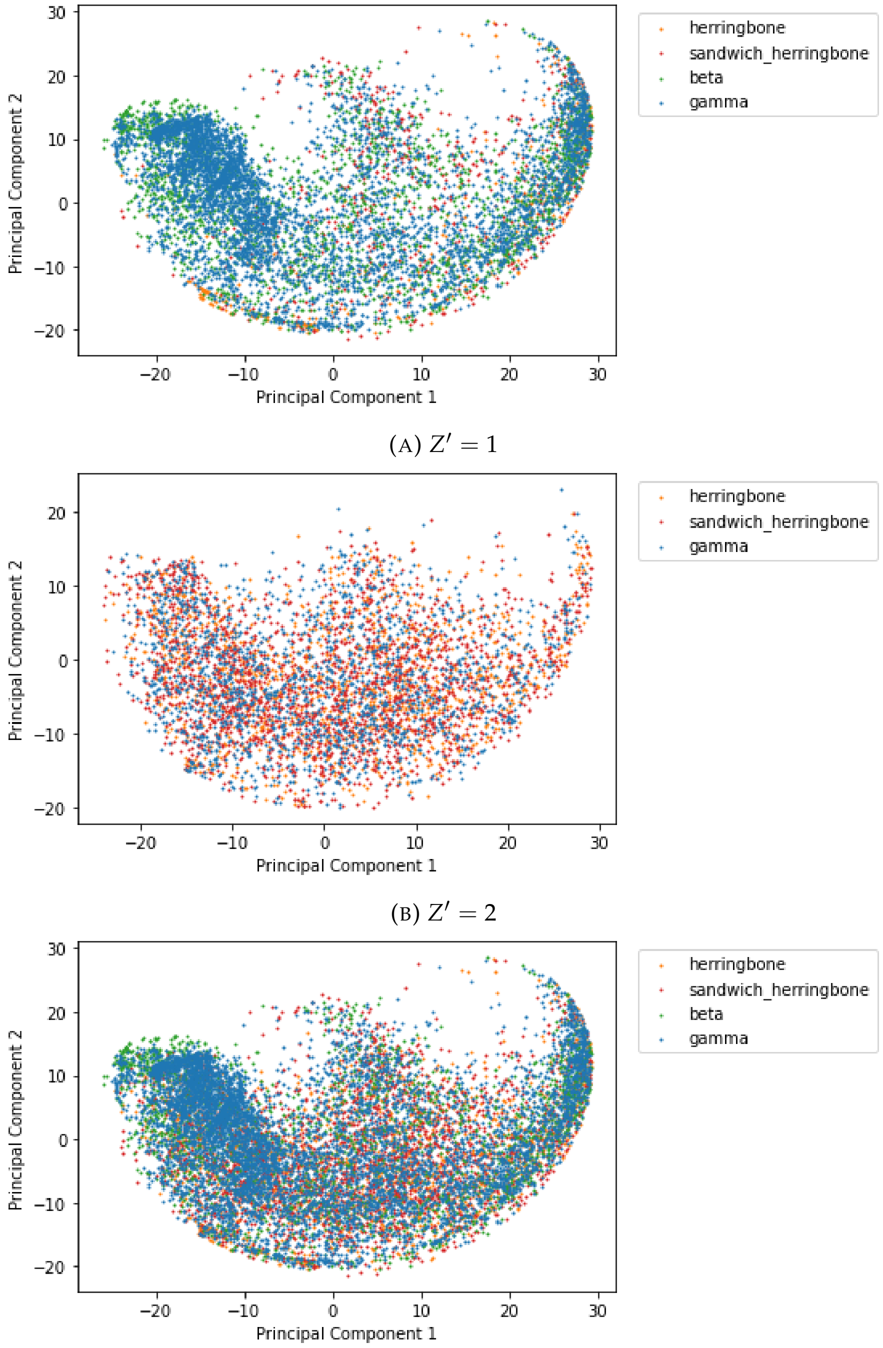


(A) 5A

(B) 5B

(C) 5C

FIGURE 5.71: Plots of the first two principal components for the vector images from the persistent homology of the predicted crystal structures for the azapentacens (a) 5A, b) 5B, c) 5C) using the molecular centroids only to compute the homology. The structures are labelled according to the predicted canonical packing labels (beta, gamma, herringbone and sandwich herringbone) using the *Autopack* algorithm Loveland et al. (2020).



(A) $Z' = 1$

(B) $Z' = 2$

(C) All Structures

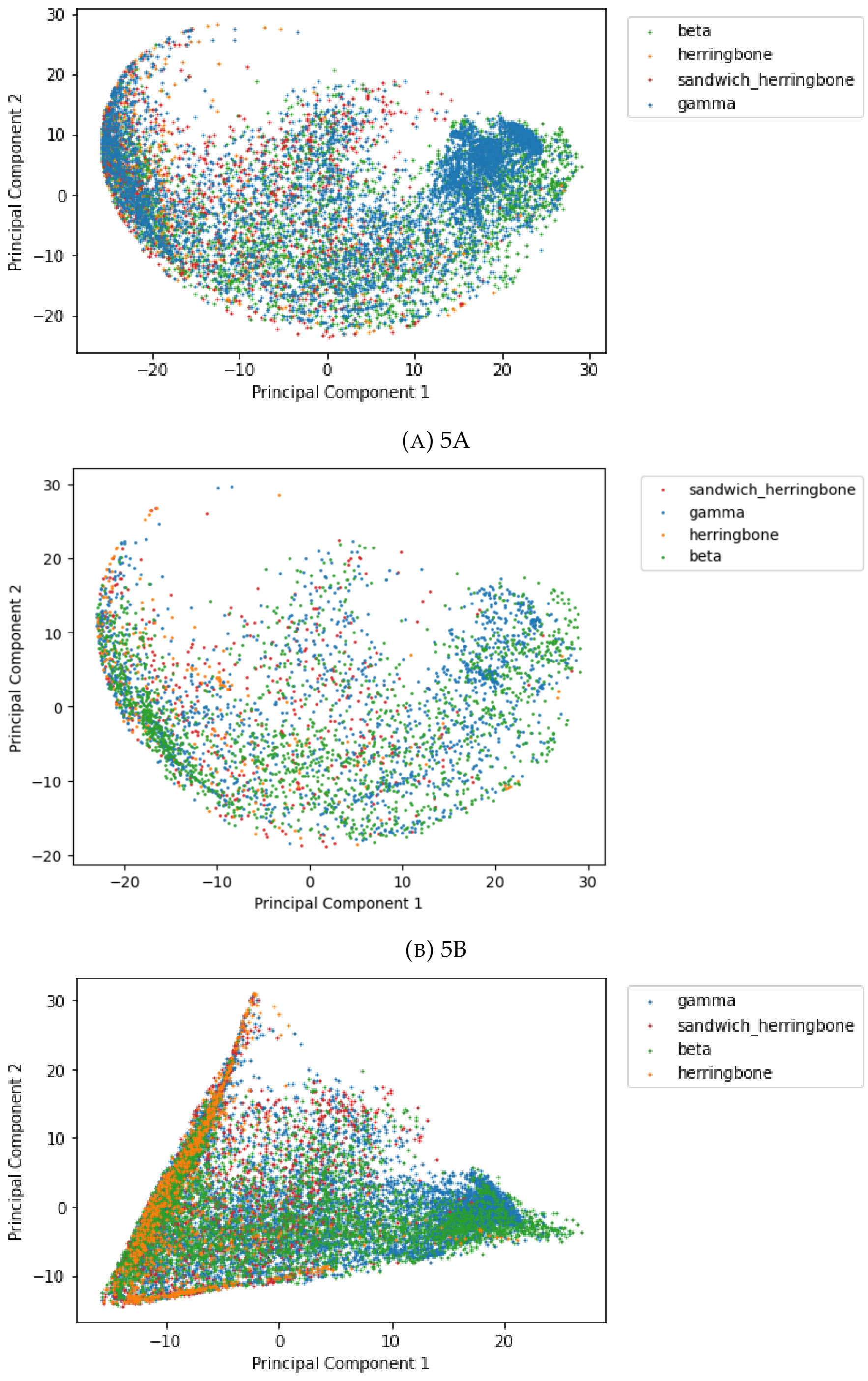
FIGURE 5.72: Plots of the first two principal components for the vector images from the persistent homology of the predicted crystal structures for pentacene (a) structures with $Z' = 1$, b) structures with $Z' = 2$, c) all structures) using the molecular centroids only to compute the homology. The structures are labelled according to the predicted canonical packing labels (beta, gamma, herringbone and sandwich herringbone) using the *Autopack* algorithm Loveland et al. (2020).

contain compounds of each and every packing class, but also showed that there were still trends in the data and that regions in figure 5.76 do indeed favour the packing types as indicated albeit less than suggested in the figure. This behaviour is replicated for pentacene but once again the compounds from crystal structure predictions with $Z' = 2$ do not follow the trends as starkly and undermine the data for $Z' = 1$.

Finally we also consider the Wasserstein embedding that we get for each of these compounds. For the azapentacenes we get the plots in figure 5.77 while the embeddings for pentacene are shown in figure 5.78. The embeddings clearly do not separate the packing classes anywhere near as well as in the previous dataset. There are a set of structures in the beta class that seem to be localised at a particular region of the embedding - by no means all compounds in this class. It could be that the embeddings in the previous section do have a more clear structure as to what regions of the landscape correspond to which packing class by virtue of there simply being less packing classes so it is easier to infer structure into these embeddings even when there is none. Or it could be the case that persistent homology has some predictive power in understanding the difference between some of the packing classes but not all of them at once - it seems that some of the beta and gamma structures are positioned in the embedding in a manner reminiscent of the previous section but the herringbone and the sandwich herringbone structures are much less predictable. In further work it would be wise to consider how these landscapes change when the parameters of the CSP calculation are altered. Note how the embedding for pentacene for $Z' = 2$ has by far the least structure across the packing classes. The influence of altering the CSP methodology between the datasets should not be understated.

There are two further factors we have yet to consider: the effects of the persistent homology features that might occur when the crystal fragment is only defined on a restricted supercell and the interplay between the crystal structure landscape and the energy. For the azapentacene, 5A, we construct three further plots. The PCA plot of the descriptor obtained when the restricted supercell approach is used *and* the vector on the orientation vector is *not* scaled and the plot obtained with the same calculation when the orientation vector *is* scaled. These two plots are shown in figure 5.79. Both of these figures show some structure in the relative positions of the packing classes but neither of these plots necessarily provide any more insight than the equivalent plots where the restricted supercells are not used (figures 5.73a and 5.75a respectively).

For compound 5A we also replot the reduced vector images for the 6D case with both the scaled and unscaled orientation vector but display only those structures with the 1000 lowest energy values. This is simply to ascertain whether the persistent homology is more closely related to the packing type for the more stable crystal structures. It can be seen from figure 5.80 that this is not really completely the case as the structures for the beta and gamma classes (by far the most common at this energy

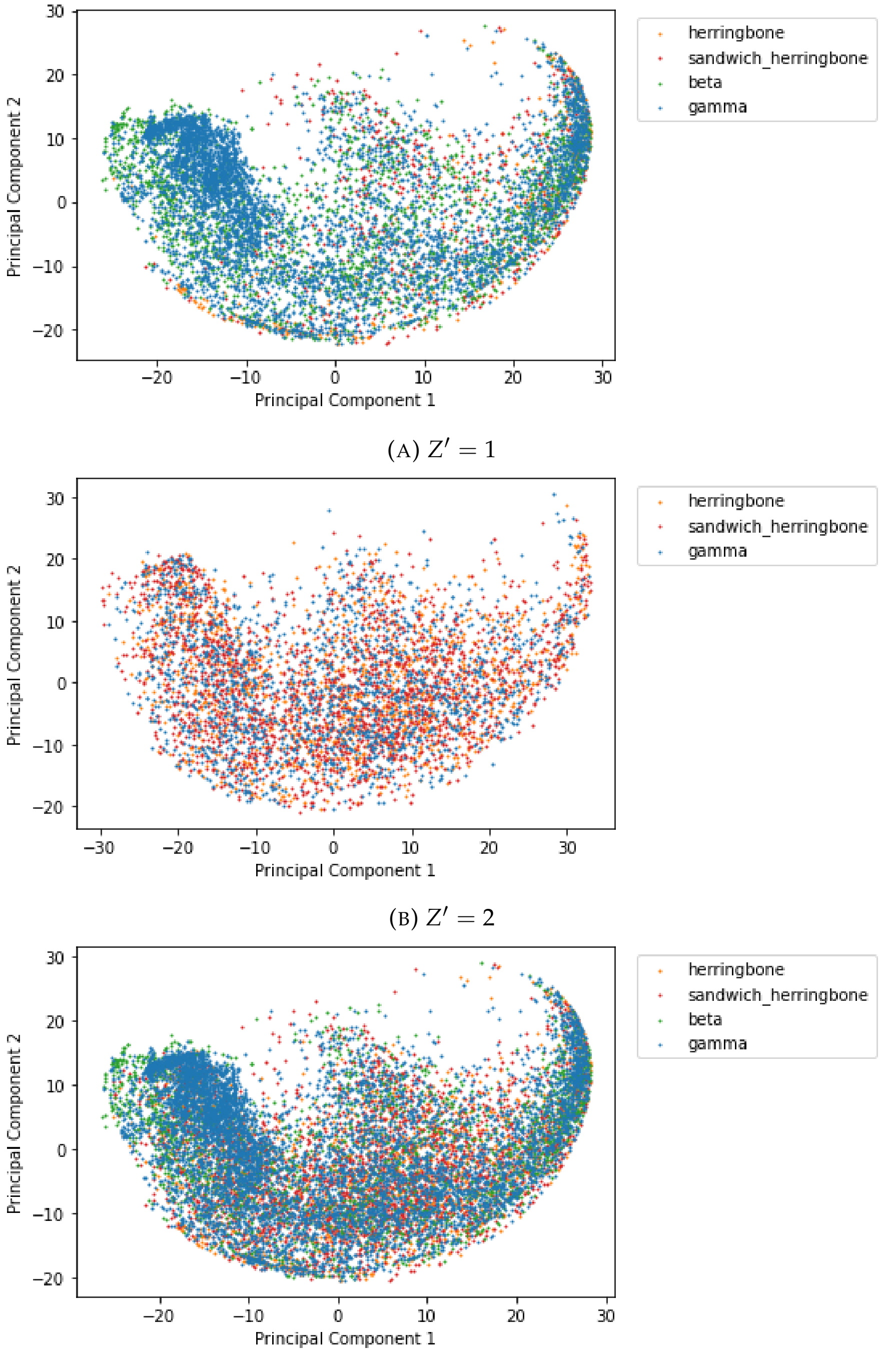


(A) 5A

(B) 5B

(C) 5C

FIGURE 5.73: Plots of the first two principal components for the vector images from the persistent homology of the predicted crystal structures for the azapentacens (a) 5A, b) 5B, c) 5C) using the molecular centroids and suitably chosen inertia vectors to compute the homology (in 6D). The structures are labelled according to the predicted canonical packing labels (beta, gamma, herringbone and sandwich herringbone) using the *Autopack* algorithm Loveland et al. (2020).

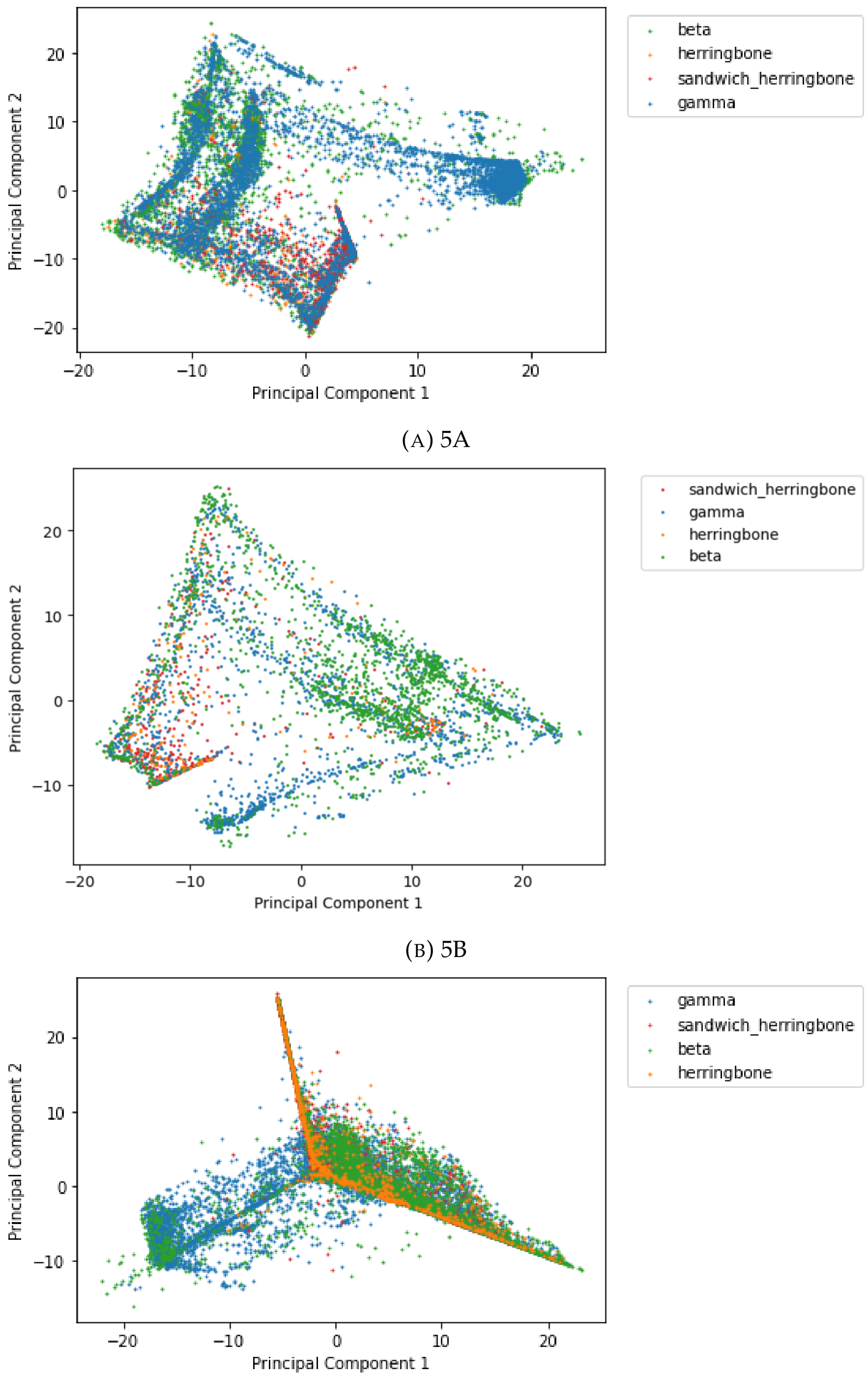


(A) $Z' = 1$

(B) $Z' = 2$

(C) All Structures

FIGURE 5.74: Plots of the first two principal components for the vector images from the persistent homology of the predicted crystal structures for pentacene (a) structures with $Z' = 1$, b) structures with $Z' = 2$, c) all structures) using the molecular centroids and a suitably chosen inertia vector to compute the homology (in 6D). The structures are labelled according to the predicted canonical packing labels (beta, gamma, herringbone and sandwich herringbone) using the *Autopack* algorithm Loveland et al. (2020).

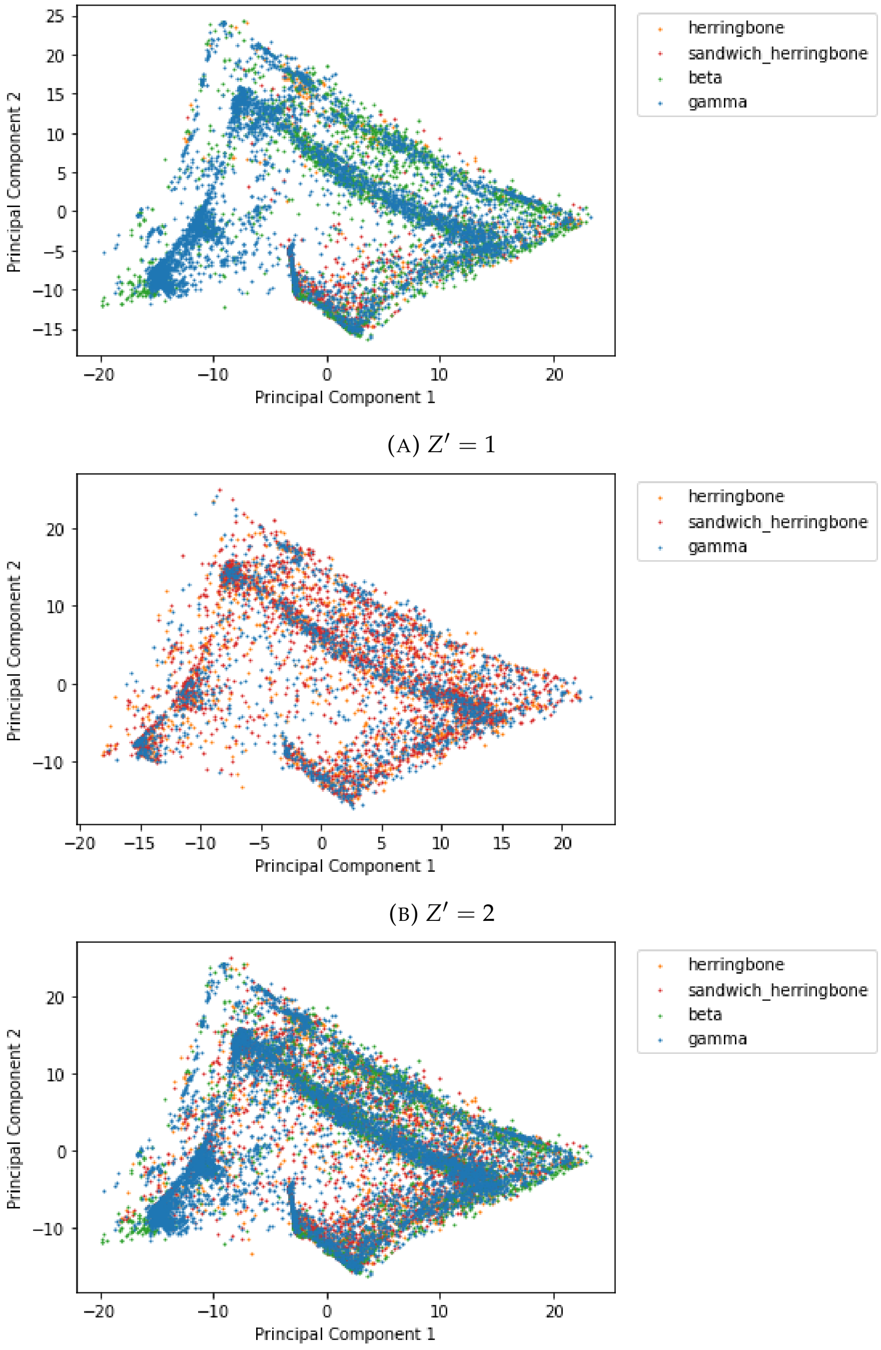


(A) 5A

(B) 5B

(C) 5C

FIGURE 5.75: Plots of the first two principal components for the vector images from the persistent homology of the predicted crystal structures for the azapentacens (a) 5A, b) 5B, c) 5C) using the molecular centroids and a suitably chosen inertia vector (which is scaled in by a factor of 25) to compute the homology (in 6D). The structures are labelled according to the predicted canonical packing labels (beta, gamma, herringbone and sandwich herringbone) using the *Autopack* algorithm Loveland et al. (2020).

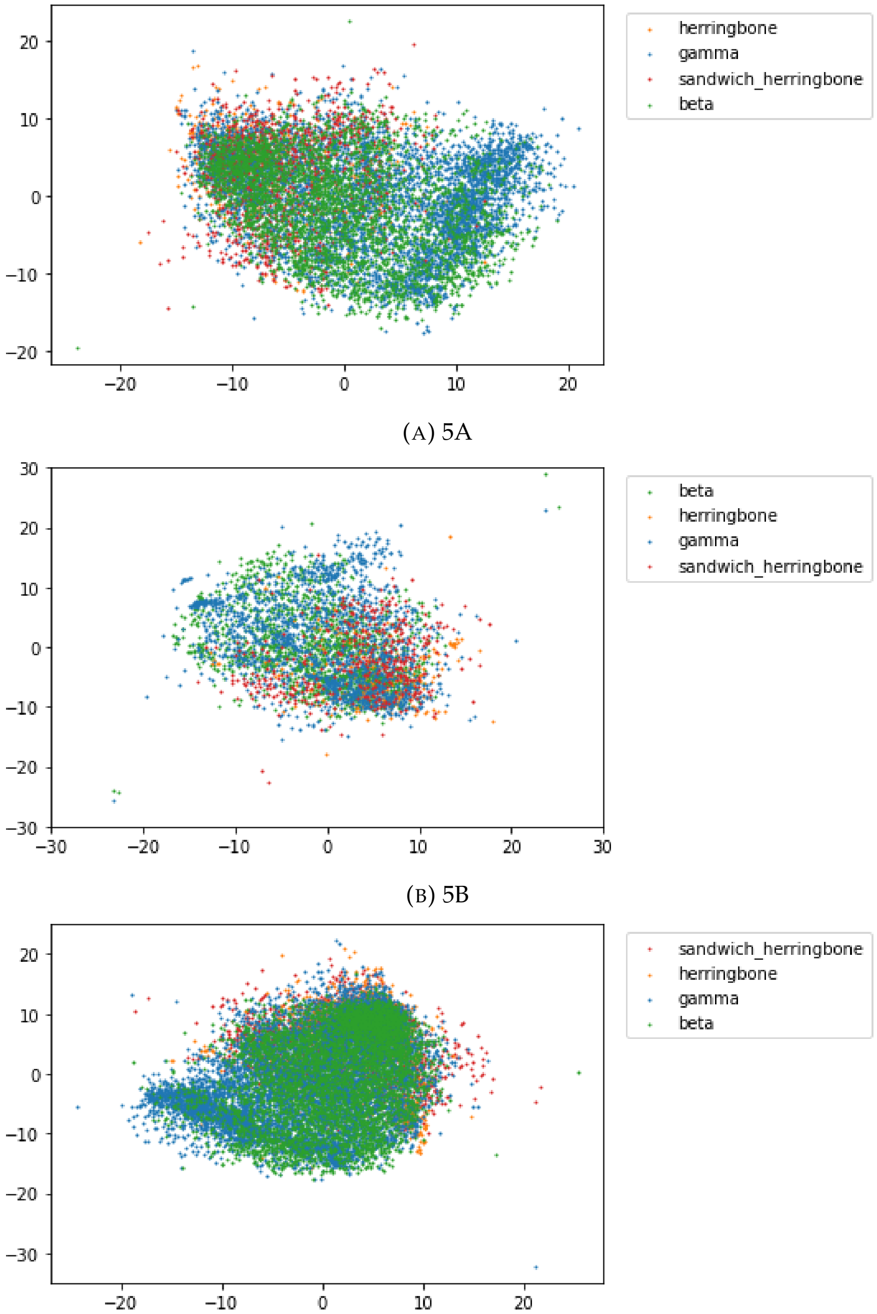


(A) $Z' = 1$

(B) $Z' = 2$

(C) All Structures

FIGURE 5.76: Plots of the first two principal components for the vector images from the persistent homology of the predicted crystal structures for pentacene (a) structures with $Z' = 1$, b) structures with $Z' = 2$, c) all structures) using the molecular centroids and a suitably chosen inertia vector (which has been scaled by a factor of 25) to compute the homology (in 6D). The structures are labelled according to the predicted canonical packing labels (beta, gamma, herringbone and sandwich herringbone) using the *Autopack* algorithm Loveland et al. (2020).

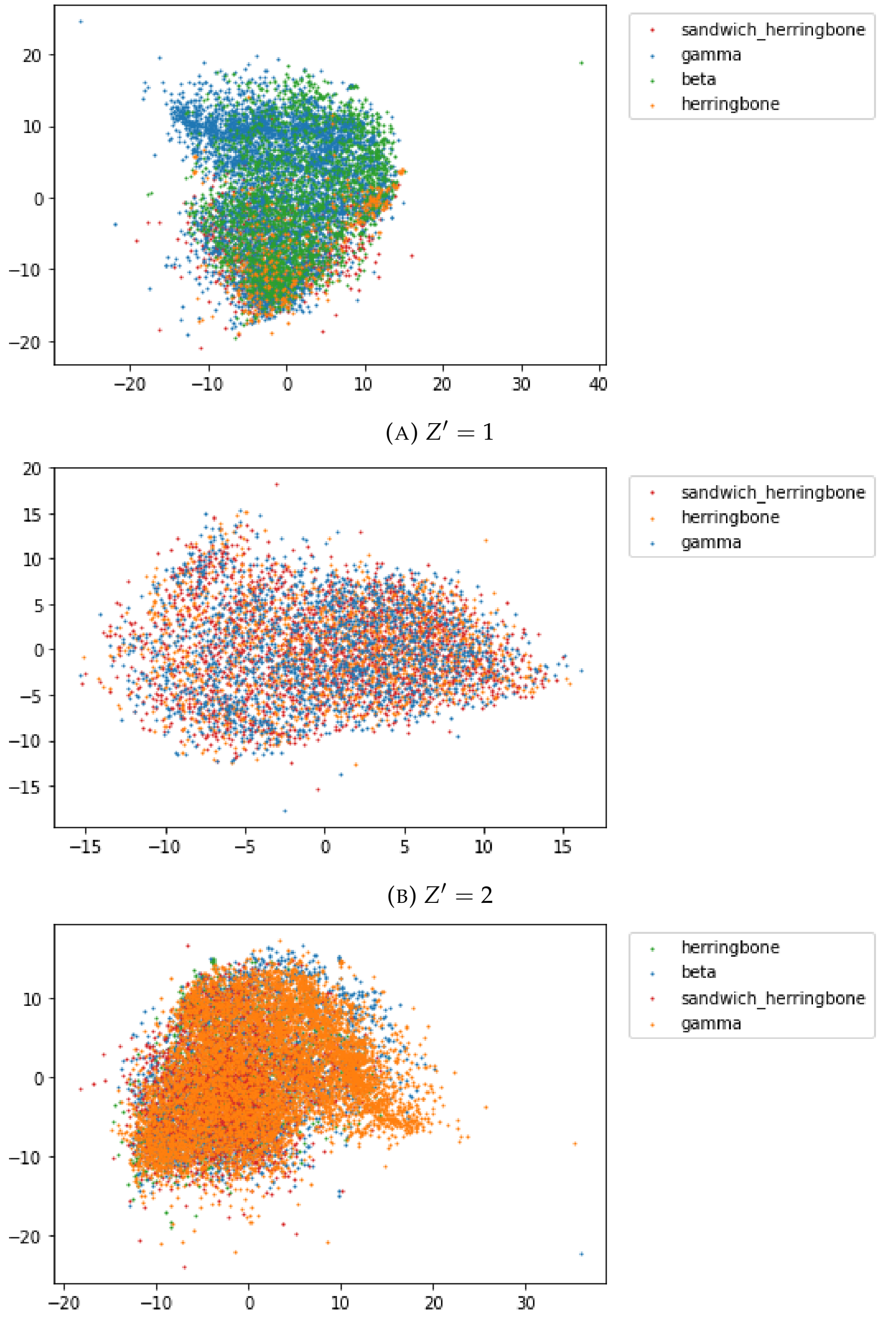


(A) 5A

(B) 5B

(C) 5C

FIGURE 5.77: MDS embedding of the set of persistence diagrams of the predicted crystal structures for the azapentacens (a) 5A, b) 5B, c) 5C). Th persistence diagrams were found using the molecular centroids and suitably chosen inertia vectors. The structures are labelled according to the predicted canonical packing labels (beta, gamma, herringbone and sandwich herringbone) using the *Autopack* algorithm Loveland et al. (2020). As before the axes of the space into which the data are embedded have no physical interpretation and are hence unlabelled.



(A) $Z' = 1$

(B) $Z' = 2$

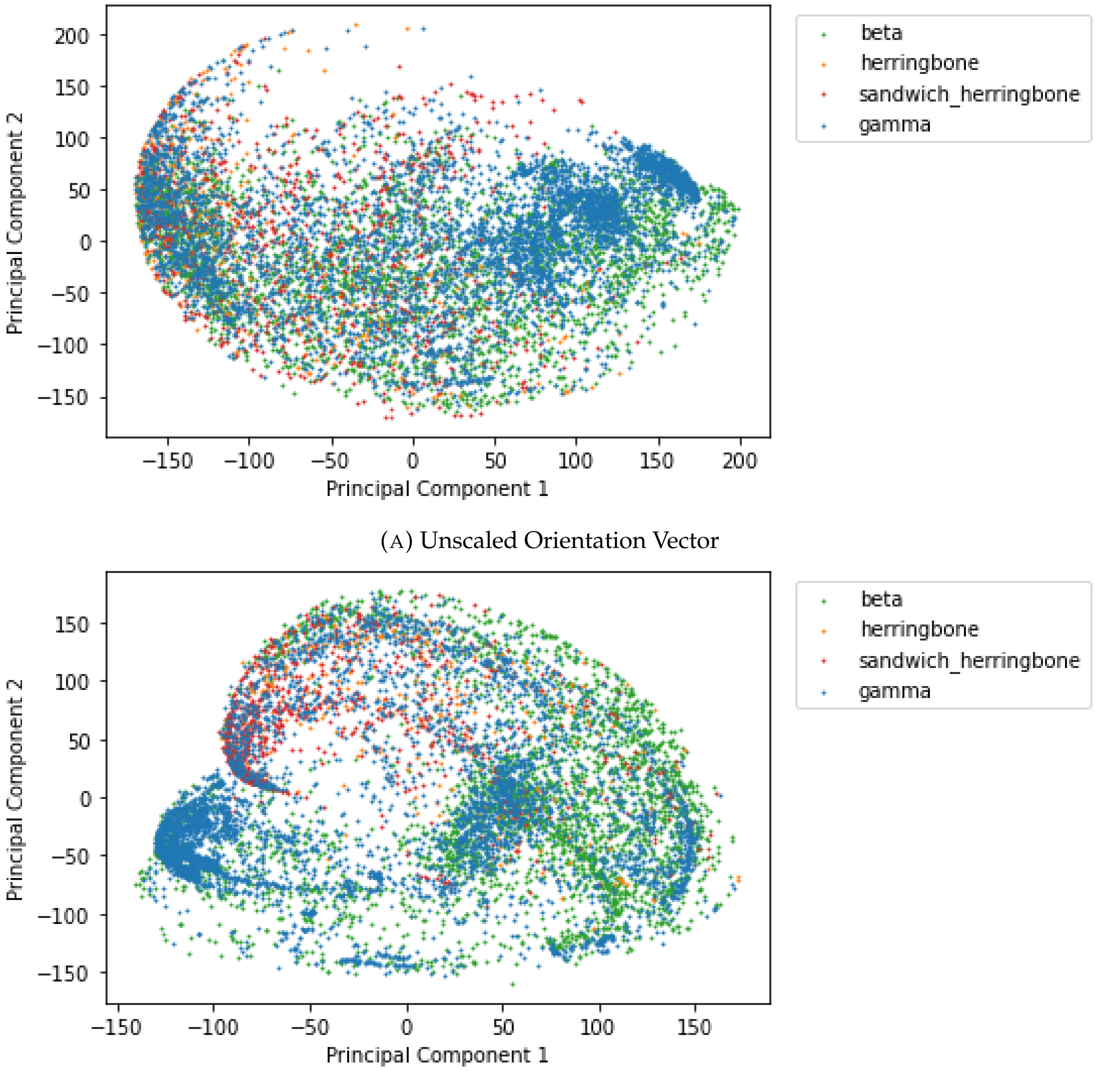
(C) All Structures

FIGURE 5.78: MDS embedding of the set of persistence diagrams of the predicted crystal structures for pentacene (a) structures with $Z' = 1$, b) structures with $Z' = 2$, c) all structures).The persistent homology was computed using the molecular centroids and a suitably chosen inertia vector. The structures are labelled according to the predicted canonical packing labels (beta, gamma, herringbone and sandwich herringbone) using the *Autopack* algorithm Loveland et al. (2020). As before the axes of the space into which the data are embedded have no physical interpretation and are hence unlabelled.

threshold) do overlap significantly, although this overlap is remarkably reduced for the case in which the orientation vector is scaled. It may not be the case that the lower energy structures can be more accurately described by persistent homology *per se*, it is equally likely that the beta packing class is more likely to possess different persistent homology than the other packing classes, particularly for the case when the orientation vector is scaled. One can imagine that the fact that pretty much all of the molecules are parallel in these structures is something that is encoded into the (exaggerated) vector differences, while for the other classes exactly *which* molecules have very different orientations is important, which is harder to encode in the 6D persistent homology.

In order to establish if one can separate all four packing classes *at all*, that is whether there is *any* difference in the persistent homology between these structures, we also carry out *supervised* dimensionality reduction, specifically LDA, to see if we can tease these packing classes apart with a more direct approach. Note that due to the large number of data to which the LDA models are being fit we do not need to worry about our model being too high dimensional for LDA to be effective. For each compound we produced an LDA based crystal structure landscape both for the 6D persistent homology descriptor and for the 6D persistent homology descriptor with the rescaled orientation vector. The resulting plots for each compound (5A, 5B, 5C and pentacene) with the unscaled 6D persistent homology descriptor are shown in figure 5.81. Note that for pentacene we combine the data for $Z' = 1$ and $Z' = 2$. The same plots obtained using persistent homology with the *scaled* orientation vector are shown in figure 5.82. As implied in the unsupervised case, we see that using the scaled orientation vector when calculating the persistent homology significantly improves results. Also consistent with our earlier discussion is the fact that - particularly when the rescaled vector is used - the beta class is by far the class that is most distinguishable using this method. This is consistent with the fact that this class may be easier to clearly distinguish with the properties of the orientation vectors alone. Even in the best cases we cannot completely separate the classes - the overlap of the different packing classes is significant and there are many regions of the landscapes where lots of structures are on top of each other. This may imply that some of the structures have the same packing type and very different persistent homology and that equally some structures have very different persistent homology but the same predicted packing type. There are a few things that could cause this.

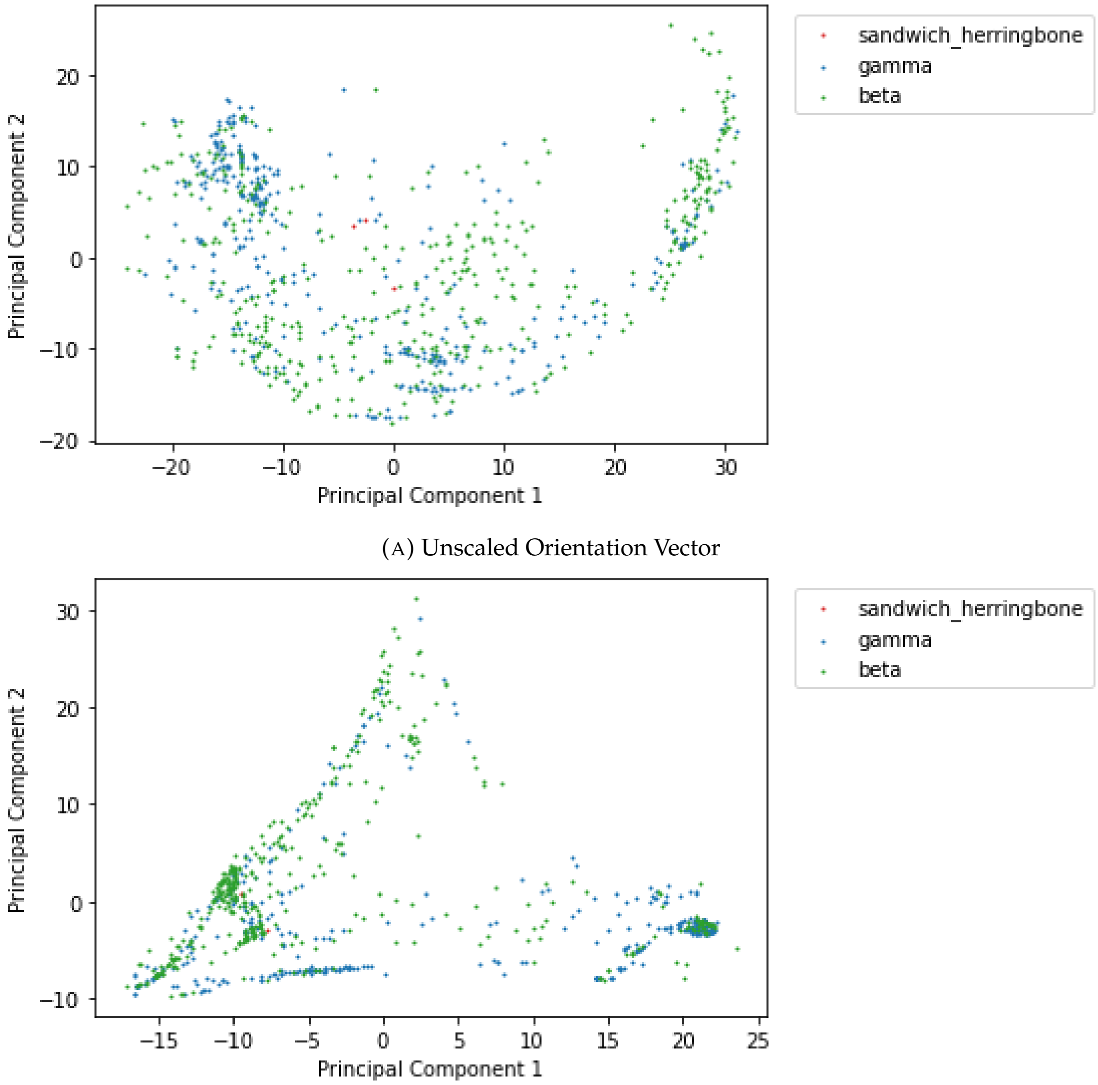
Firstly its possible that the labels that can be predicted by the *Autopack* algorithm do not fully reflect the full diversity of structures these compounds can take and/or it is possible that there are subclasses within the packing motifs not taken account of, for example Campbell *et al.* Campbell et al. (2017) discuss a sub-variety of the gamma packing class described there as ”slipped gamma”. Notice that in the LDA plots many of the packing class occur in at least two different regions which might correspond to



(A) Unscaled Orientation Vector

(B) Rescaled Orientation Vector

FIGURE 5.79: Plots of the first two principal components for the descriptor from the persistent homology of the predicted crystal structures for the azapentacene labelled 5A, using the molecular centroids and a suitably chosen inertia vector (which is either unscaled and has a length of 1 (part a) or is rescaled and has a length of 25 (part b)) to compute the homology (in 6D). The descriptor is composed of four vector images constructed from four persistence diagrams which are respectively calculated using crystal fragments after the $(1 \times R \times R)$, $(R \times 1 \times R)$, $(R \times R \times 1)$ and the $(R \times R \times R)$ supercells (with arbitrarily large R) . The structures are labelled according to the predicted canonical packing labels (beta, gamma, herringbone and sandwich herringbone) using the *Autopack* algorithm Loveland et al. (2020)



(A) Unscaled Orientation Vector

(B) Rescaled Orientation Vector

FIGURE 5.80: Plots of the first two principal components for the descriptor from the persistent homology of the predicted crystal structures for the azapentacene labelled 5A, using the molecular centroids and a suitably chosen inertia vector (which is either unscaled and has a length of 1 (part a) or is rescaled and has a length of 25 (part b)) to compute the homology (in 6D). In this case we only display the 1000 structures of lowest energy. The structures are labelled according to the predicted canonical packing labels (beta, gamma, herringbone and sandwich herringbone) using the *Autopack* algorithm Loveland et al. (2020)

further subclasses of the same packing type. Perhaps many of the compounds located in the very dense central regions of these plots which contain every packing class might contain examples of crystal structures which have been assigned as beta, gamma, herringbone and sandwich herringbone when these structures might best be labelled ”other” or ”unclassified”. All descriptions of packing classes are somewhat arbitrary even if they are the most physically and chemically relevant - it might be quite presumptuous to assume that all structures can be assigned into these classes and that any crystal structure landscape could be transformed to respect these arbitrary definitions.

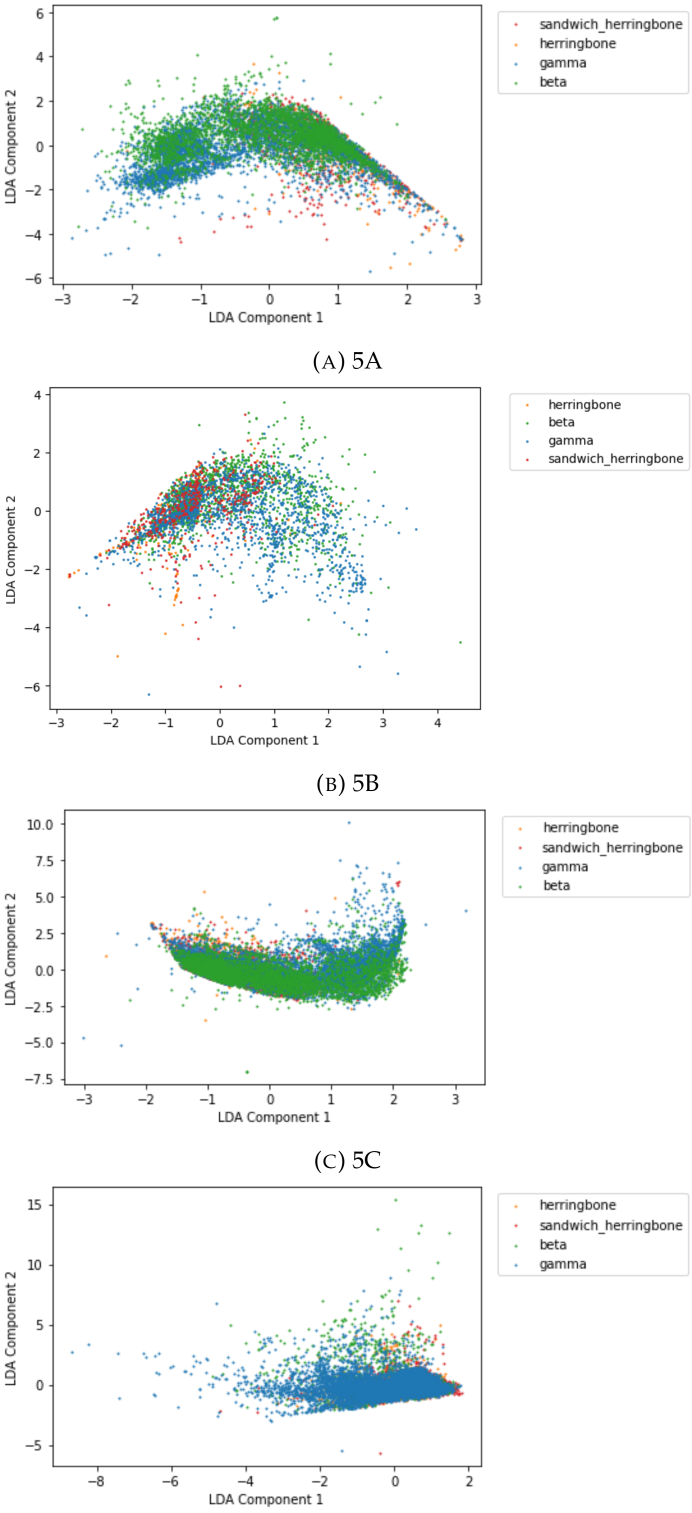
The second reason for the substantial class overlap could simply be deficiencies in our model. One significant issue is that by taking the molecular centroid and a single vector to represent a molecule, one loses a lot of geometric richness and might fail to capture the rod-like nature of some of these molecules or the all important face-to-face π-stacking interaction. Further work should investigate whether it is possible to extract topological invariants (using persistent homology or otherwise) from the graph that can be built up between the network of short contacts between the organic molecules in the crystal (which can be found quite easily with software provided by the CSD Groom et al. (2016)). Better still perhaps one might even construct a graph that specifically focuses on π-stacking or the edge-to-face interactions between the C-H bond and π systems (which are most relevant to the packing types of polyaromatic hydrocarbons) and compute the resulting invariants.

We conclude this section by briefly visiting the subject of energy prediction. As we have a large set of compounds and a descriptor that describes the topology of the packing and a set of energies found during the CSP algorithm, it is interesting to build models to assess the fidelity of our descriptor through the lens of property prediction. Recall that we have already tried this with the set of fluorolanines to pretty abysmal results, this might not necessarily mean that the descriptor will not work in this case. Indeed the connection between the molecular packing geometry and the crystal energy could be stronger for these compounds.

We adopted the same methodology as in the fluoroalanine energy prediction, that is we used a 25% : 75% test:train split and we used support vector regression and random forest regression.

The fits we obtained for each of the compounds with the support vector model are shown in figure 5.83. In all cases a radial kernel was used with $C = 1 \quad \text{and} \quad \epsilon = 0.1$. The R^2 values for the model fit on 5A, 5B, 5C and pentacene are respectively $0.394, -0.0496, 0.309 \quad \text{and} \quad 0.323$.

For the random forest model we used the following sets of parameters in all cases: ”n_estimators”=100, ”max_depth”=None, min_samples_split=2, min_samples_leaf=1,



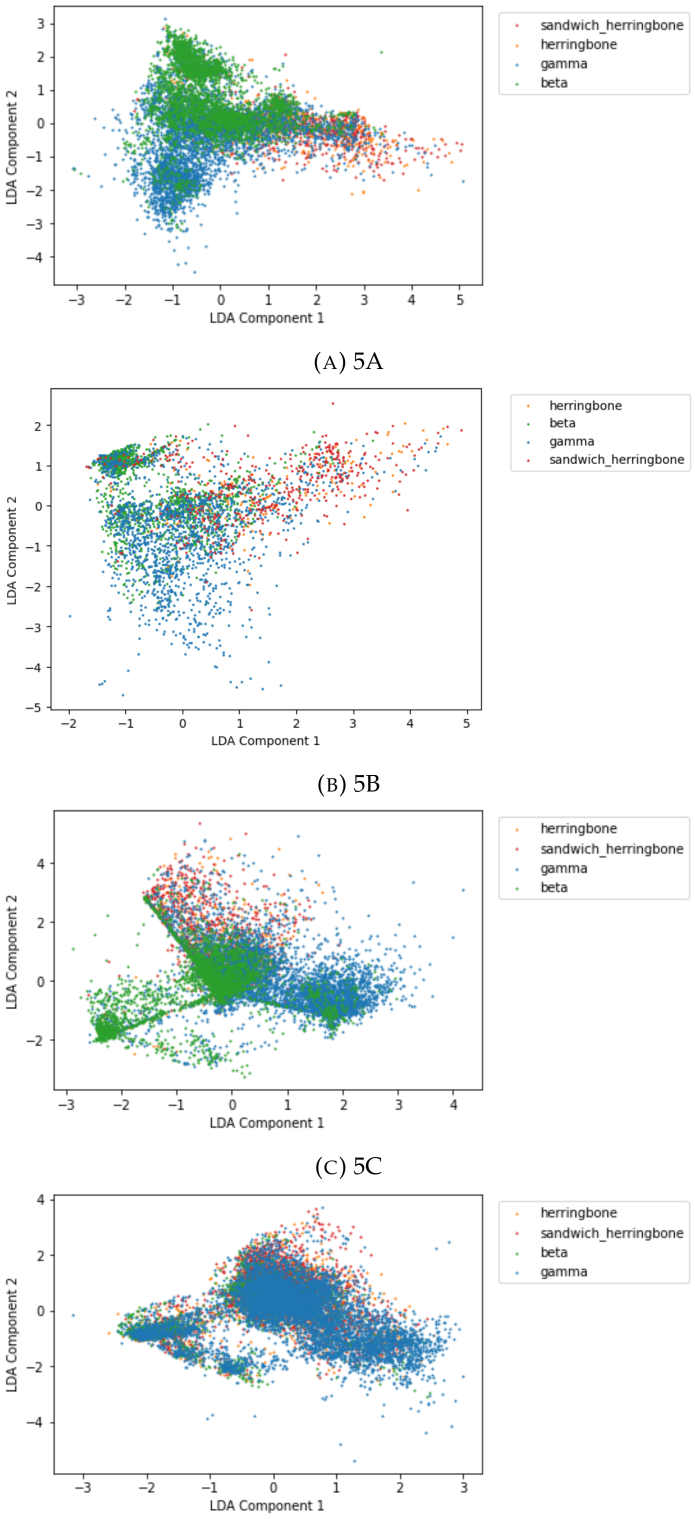
(A) 5A

(B) 5B

(C) 5C

(D) Pentacene

FIGURE 5.81: Supervised dimensionality reduction (LDA) of the set of vector images corresponding to the persistent homology (with 6D pointcloud and *unscaled* orientation vector) of the predicted structures of three azapentacenes (5A, 5B and 5C) and pentacene. The labels used for the LDA model are the predicted canonical packing labels (beta, gamma, herringbone and sandwich herringbone) using the *Autopack* algorithm Loveland et al. (2020).



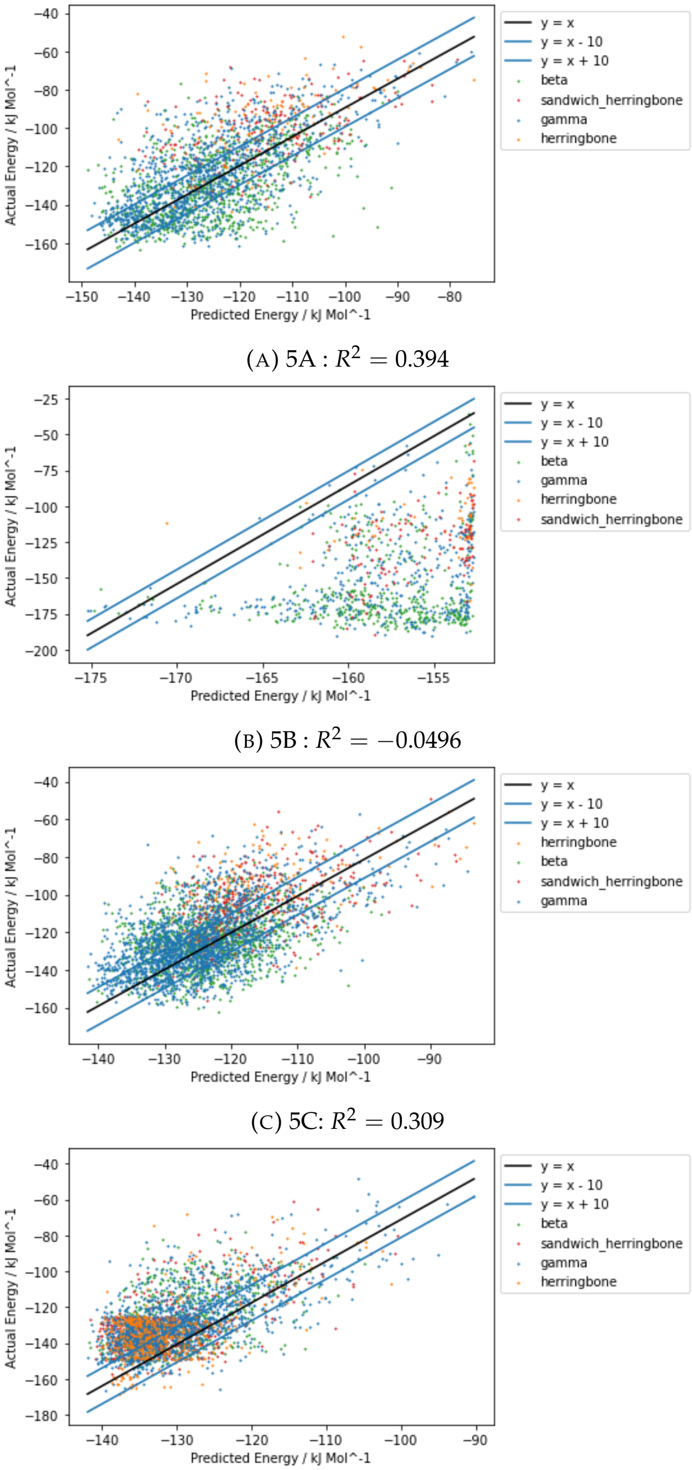
(A) 5A

(B) 5B

(C) 5C

(D) Pentacene

FIGURE 5.82: Supervised dimensionality reduction (LDA) of the set of vector images corresponding to the persistent homology (with 6D pointcloud and *scaled* (to length 25) orientation vector) of the predicted structures of three azapentacenes (5A, 5B and 5C) and pentacene. The labels used for the LDA model are the predicted canonical packing labels (beta, gamma, herringbone and sandwich herringbone) using the *Autopack* algorithm Loveland et al. (2020).



(A) 5A : $R^2 = 0.394$

(B) 5B : $R^2 = -0.0496$

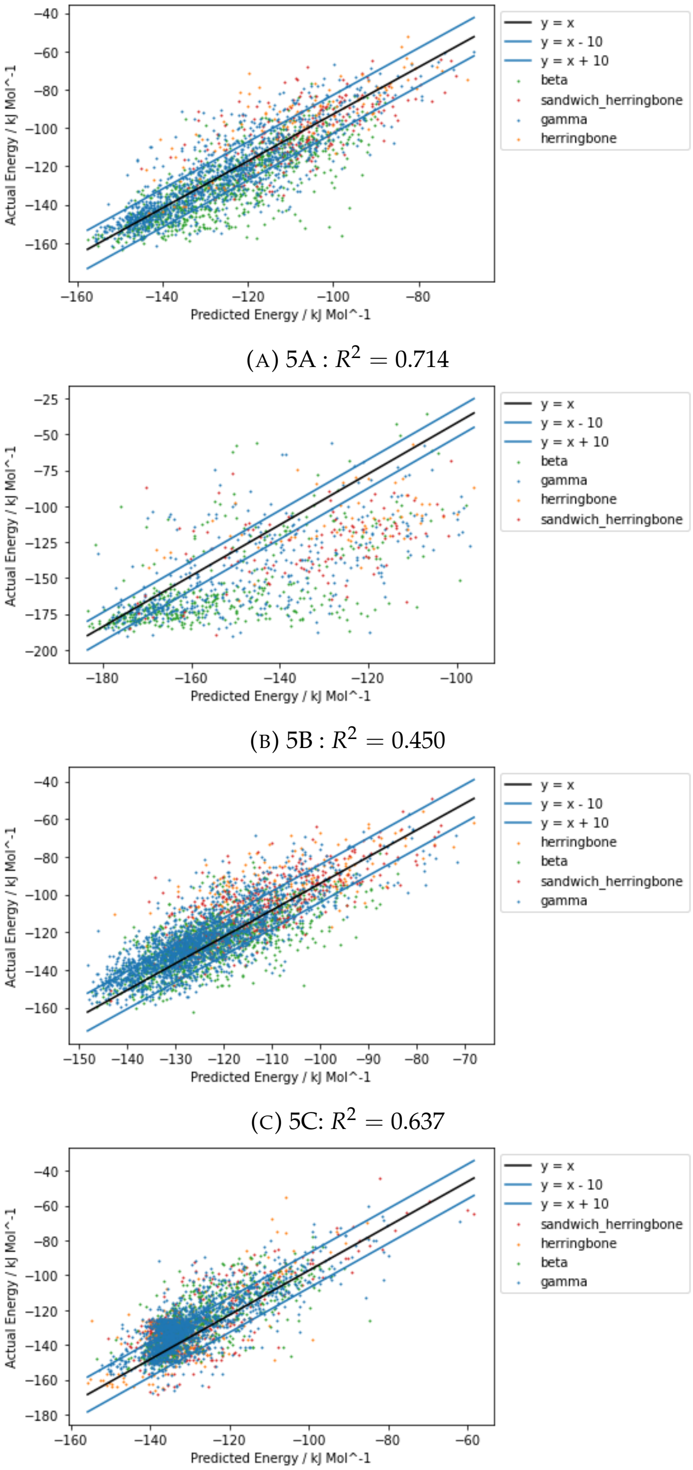
(C) 5C: $R^2 = 0.309$

(D) Pentacene : $R^2 = 0.323$

FIGURE 5.83: Predicted vs. true energy for the (25 %) test set for a support vector model trained on a descriptor based on the persistent homology of a set of predicted azapentacene (and related) structures using the 6D pointcloud method (vector unscaled). The compounds studied are 5A (a), 5B (b), 5C (c) and pentacene (d). In all cases a radial kernel was used with $C = 1, \epsilon = 0.1$.

”max_features”=1.0, ”max_leaf_nodes”=None . The R^2 values for the model fit on 5A, 5B, 5C and pentacene are respectively $0.714, 0.450, 0.637 \quad \text{and} \quad 0.592$. The fits are shown in figure 5.84.

The support vector models provide pretty poor results which are not much better than the results of the support vector machine on the fluoroalanine crystal structure landscape ($R^2 = 0.314$). The results for the random forest model are in fact quite good although the fit is clearly better for some azapentacene compounds than others (5A vs. 5B, for example) - these models vastly outperform the equivalent models trained on the fluorolalanine crystal structure landscape ($R^2 = 0.298$). It is interesting that while in this case our descriptor does a worse job at describing the packing classes, it does a better job at predicting the energy. It may be that for this class of compounds the correspondence between packing type and energy is stronger (the energy contributions being dominated by the face-to-face and edge-to-face interactions of the aromatic rings) than for the fluoroalanine compounds (the energy being influenced by a much more complex set of intermolecular interactions, see the work of Dodd *et al.* for example Dodd (2020)). The fact that we can use the persistent homology to predict the energy of these compounds on some level implies that the persistent homology *is* capturing some important geometric features - perhaps not those that correspond to the canonical packing classes. By refining the energy predictions of each of these compounds with more sophisticated quantum mechanical calculations and by better optimising the set of hyperparamters for each of our models, it will be possible to get a much more accurate prediction of the energy of these compounds. This should be a topic for further work.



(A) 5A : $R^2 = 0.714$

(B) 5B : $R^2 = 0.450$

(C) 5C: $R^2 = 0.637$

(D) Pentacene : $R^2 = 0.592$

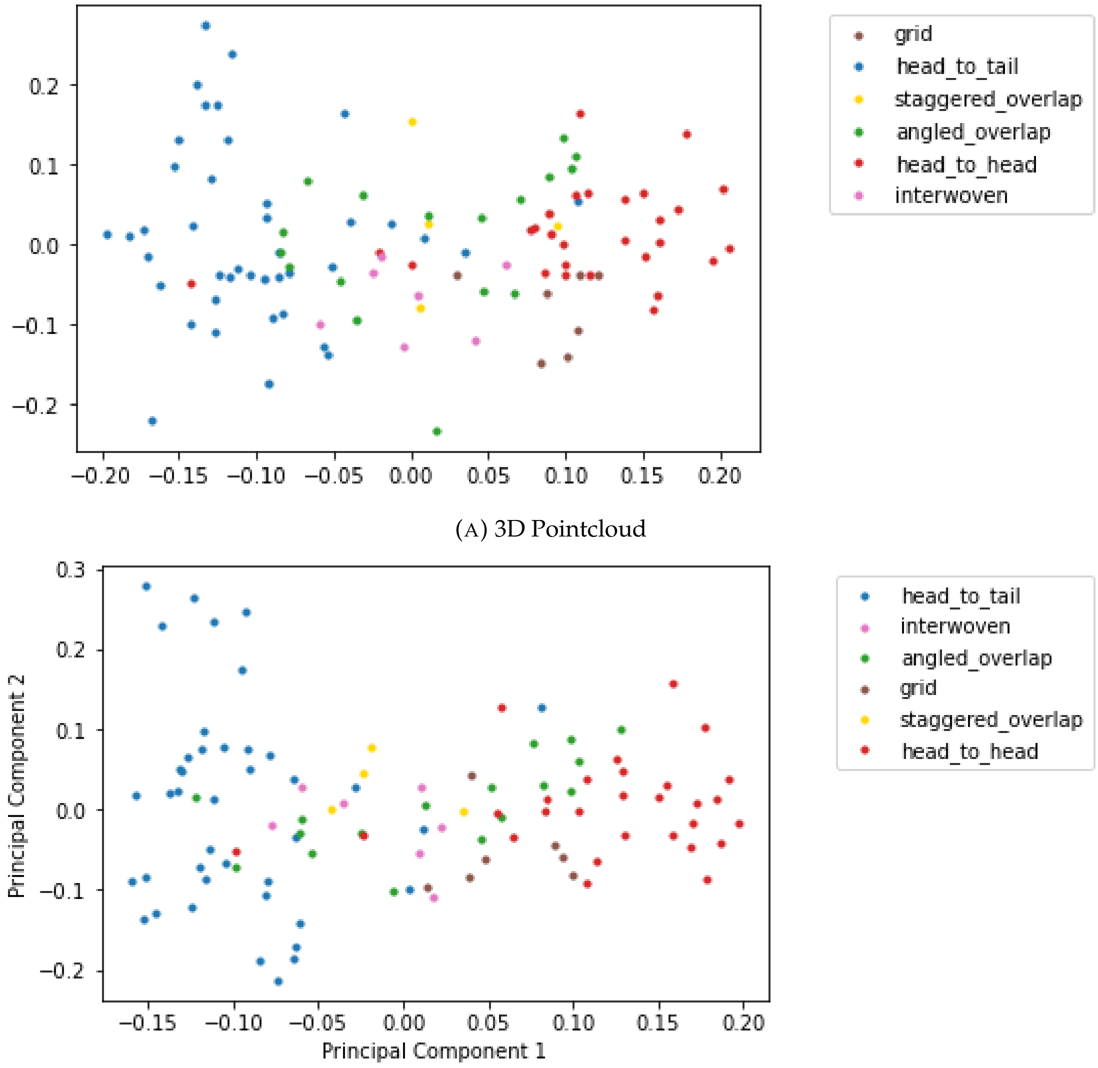
FIGURE 5.84: Predicted vs. true energy for the (25 %) test set for a random forest model trained on a descriptor based on the persistent homology of a set of predicted azapentacene (and related) structures using the 6D pointcloud method (vector unscaled). The compounds studied are 5A (a), 5B (b), 5C (c) and pentacene (d). We used the following sets of parameters in all cases: ”n_estimators”=100, ”max_depth”=None, min_samples_split=2, min_samples_leaf=1, , ”max_features”=1.0, ”max_leaf_nodes”=None

5.4 Do We Really Need Persistent Homology?

Over the course of this work we have established geometric structures which can be extracted from the periodic crystal structure and to which persistent homology may be applied in order to obtain a descriptor that in turn conveys topological information that relates to the packing structure of the crystal. It might be reasonable to ask whether the relationship to the crystal packing stems from the topological information extracted by the persistent homology or simply from the geometric structure which was constructed in the first instance, that is, the set of molecular positions and orientations. In this final section we explore whether more basic, conventional or intuitive descriptors applied to the molecular positions and/or orientations can generate similar results to persistent homology. In so doing we can also learn about what features of the persistent homology makes it a powerful descriptor for these crystal systems. We centre our discussion around the set of fluoroalanine structures for which we know persistent homology provided a good description of the crystal packing. In all the examples below we work with the set of 50 molecular centroids and suitably chosen inertia vectors (if used) that are the starting point for our persistent homology calculations.

Perhaps the simplest way of constructing a descriptor from the geometric entity described above (that satisfies the invariance constraints for descriptors) is the distance matrix which is actually a rather widely used descriptor in chemical and materials informatics Li et al. (2023b); Randić and Pompe (2001); Takemura et al. (2021); Musil et al. (2021) . An easy way to give this descriptor a reasonable dimension (it is not normally advisable to have a larger dimension than the number of data points if this can be avoided Hastie et al. (2009)) is to summarise the matrix as a histogram. In figures 5.85a and 5.85b we condense the distance matrix of the fluoroalanine crystal fragments used to compute persistent homology into a histogram with 100 bins for both the 3D and the 6D fragment. The separation of the packing classes is actually pretty good considering the simplicity of the descriptor, although there are not clear clusters that correspond to packing classes as when we apply persistent homology. Note that the inclusion of the vectors into the pointcloud does not improve the model much.

The molecular orientations can be more usefully encoded in the descriptor by altering our notion of ”distance” in the constructed distance matrix. That is, rather than finding the distance between each six dimensional point, one can instead find a suitable interaction term inspired by something physical. For example the interaction potential between two dipoles $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ separated by vector $\mathbf{R}$ can be modelled as Stone (2013):



(A) 3D Pointcloud

(B) 6D Pointcloud

FIGURE 5.85: The principal components of a histogram of the distance matrix with 100 bins from pointclouds defined from a set fluoroalanine crystals with visually defined packing schemes. The pointclouds in question contain either a) 50 molecular centroids (3D case) or b) 50 molecular centroids with a suitably chosen inertia vector which describes the molecular orientation (6D case). Although the distances themselves have units - the histogram of these distances is a set of dimensionless quantities (counts) so that the axes of this plot are also dimensionless.

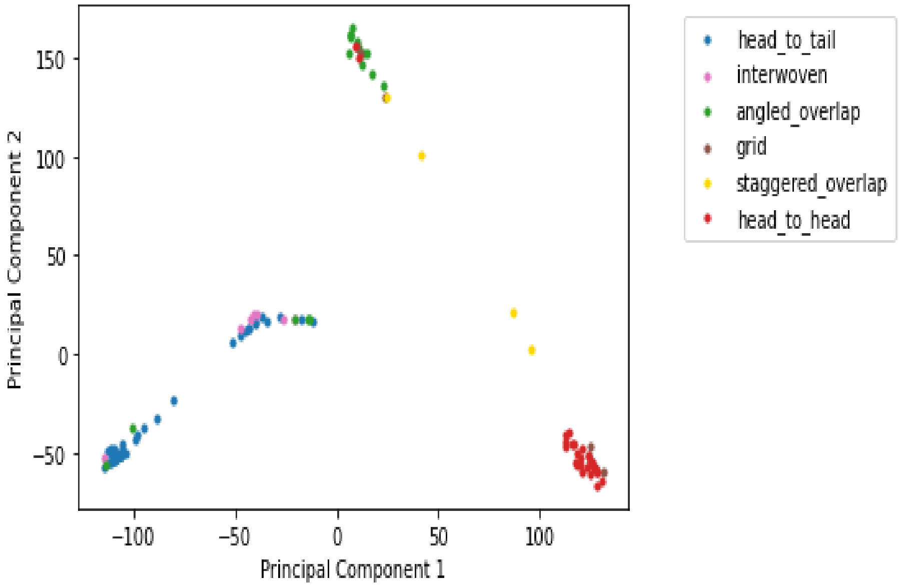


FIGURE 5.86: The principal components of a histogram of the dipole-dipole interaction matrix with 100 bins from pointclouds defined from a set fluoroalanine crystals with visually defined packing schemes. We model the orientation vectors in the 6D pointcloud as dipoles and use the Cartesian positions of the molecular centroids to work out the separation of the dipoles. Although the interaction terms themselves have units - the histogram of these interactions is a set of dimensionless quantities (counts) so that the axes of this plot are also dimensionless.

$$V = -\frac{\boldsymbol{\mu_1} \cdot \boldsymbol{\mu_2} - 3(\boldsymbol{\mu_1} \cdot \mathbf{R})(\mathbf{R} \cdot \boldsymbol{\mu_2})}{4\pi\epsilon_0 \|\mathbf{R}\|^3} \tag{5.2}$$

if we populate the interaction matrix with this term instead of the Euclidean distance, we have a set of interactions which combine the differences in position and the differences in orientation in a physically sensible way. Once again we convert the interaction matrix into a histogram with 100 bins. The first two principal components of this descriptor for the fluoroalanine dataset are shown in figure 5.86. This descriptor does a great job at separating out the head-to-head and head-to-tail classes but mixes up some of the other classes such as the grid class and the interwoven class.

Another method for extracting invariants from the crystal fragments is to generate histograms of *three* body interactions rather than two body interactions as above. One physical model for such interactions is the Axilrod-Teller potential which models the van der Waals interaction between three atoms Axilrod and Teller (1943). The form of the interaction is

$$V_{ijk} \sim \frac{1 + 3\cos\gamma_i \cos\gamma_j \cos\gamma_k}{(r_{ij} r_{jk} r_{ik})^3} \tag{5.3}$$

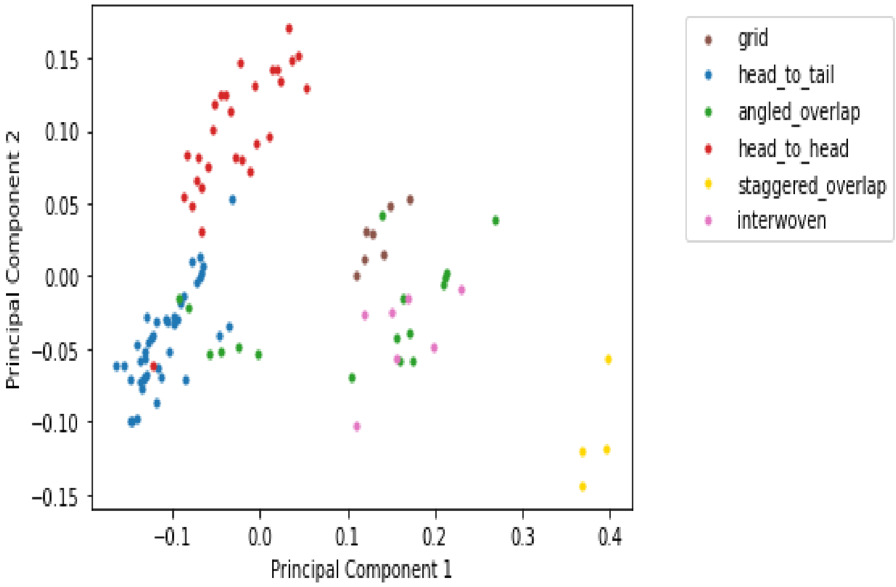


FIGURE 5.87: The principal components of a histogram of the log of 3-body Axilrod-Teller interaction tensor with 100 bins, calculated from pointclouds defined from a set fluoroalanine crystals with visually defined packing schemes. We use the molecular centroids from the crystal fragment only. Although the interaction terms themselves have units - the histogram of these interactions is a set of dimensionless quantities (counts) so that the axes of this plot are also dimensionless.

where r_{ij} is the distance between atoms i and j and γ_i is the angle between vectors $\mathbf{r_{ij}}$ and $\mathbf{r_{ik}}$. The first two principal components of a descriptor based on this interaction are shown in figure 5.87. Note that we condense the tensor of three body interactions into a histogram with 100 bins. Owing to the wide range of values of the interactions across the dataset, we take the log of the interaction before converting it into a histogram. We see that we get a separation of classes almost comparable to the 3D persistent homology descriptor. It is telling that the descriptor improves significantly going from two-body interactions to three body interactions. Perhaps the strength of the persistent homology descriptor vs. more conventional descriptors is that by extracting cycles that can be obtained from different points in the dataset one is encoding many body interactions which are geometrically richer than simple pairwise terms. If the effectiveness of persistent homology really comes from the implicit description of many body terms, we could cut out the middle man and build descriptors that focus on these terms directly.

In order to get a more thorough understanding of the advantages of including higher order interactions into materials descriptors we consider the Atom Centred Symmetry Functions (ACSF) class of descriptors. These descriptors, introduced by Behler *et al.* Behler (2011), are *local* descriptors, i.e. they describe the local environment around a given atom in the system, as such these are mostly used in the creation of neural networks for potential energy surfaces Behler (2011); Anstine and Isayev (2023); Brezina et al. (2023); Wang et al. (2024) however these descriptors have numerous applications across chemistry from spectroscopy de Armas-Morejón et al. (2023) to molecular dynamics Glielmo et al. (2021) to materials science Gou et al. (2024) and catalysis Chowdhury et al. (2024). There is a lot of flexibility in how these functions are used in practice. For our purposes the set of functions which are fit to each atom can be summarised into a histogram (i.e. a histogram of coefficients) which, we shall see can also predict the packing type of crystal systems. Note that while the points in our crystal fragment correspond to molecular not atomic positions, we can still extract geometric information in this manner, especially if we choose function hyperparameters which cause the functions to decay to zero at a slower rate. What is most attractive however is that these descriptors are composed of the coefficients of very many symmetry functions, some of which focus on two body interactions and some of which focus on three body interactions. By comparing the performance of these descriptors with and without the three body terms we can get further insight into how important these many body terms really are when it comes to understanding the underlying crystal packing.

As previously indicated there are several forms of ACSFs which incorporate different kinds of atomic interaction, in total there are five kinds of ACSF, $G_i^j : j \in \{1, 2, 3, 4, 5\}$. We shall describe each in turn.

All of the functions have in common a cutoff function, $f_c(R_{ij})$ so that atoms which are a past a given radius from the given centre are ignored and therefore that the local interactions are considered only. It has the form

$$f_c(R_{ij}) = \begin{cases} \frac{1}{2}\left[\cos \pi \frac{R_{ij}}{R_c} + 1\right] & : \text{ if } \quad R_{ij} \leq R_c \\ 0 & : \text{ if } \quad R_{ij} > R_c \end{cases} \tag{5.4}$$

We then define

$$G_i^1 = \sum_j f_c(R_{ij}) \tag{5.5}$$

and

$$G_i^2 = \sum_j f_c(R_{ij}) e^{-\eta(R_{ij}-R_s)^2} \tag{5.6}$$

and

$$G_i^3 = \sum_j f_c(R_{ij}) \cos(\kappa R_{ij}) \tag{5.7}$$

These functions model the two body interactions. The first function simply counts how many atoms live within the cutoff radius, the second and third multiply this by, respectively, a Gaussian function and a cosine function in order to model the overlap of radially symmetric orbitals. The parameters R_c, η, R_s and κ are the hyperparamters of the model. We typically fit very many of these function to each atom so as such we feed the model a large number of combinations of the hyperparameters.

There are yet more hyperparamters for the two three body terms.

$$G_i^4 = 2^{1-\zeta} \sum_{j,k \neq i} (1 + \lambda \cos\theta_{ijk})^{\zeta} e^{-(R_{ij}^2 + R_{ik}^2 + R_{jk}^2)} f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}) \tag{5.8}$$

and

$$G_i^5 = 2^{1-\zeta} \sum_{j,k \neq i} (1 + \lambda \cos\theta_{ijk})^{\zeta} e^{-(R_{ij}^2 + R_{ik}^2)} f_c(R_{ij}) f_c(R_{ik}) \tag{5.9}$$

where ζ and λ are hyperparameters and θ_{ijk} is the angle between $\mathbf{R_{ij}}$ and $\mathbf{R_{ik}}$.

In the first instance we generated descriptors using the functions $\{G_i^1\}$ and $\{G_i^2\}$ only. Owing to the very large number of possible hyperparameter combinations we fit models over a large range of these values and recorded the AMI of the clustering of the resultant descriptor with the visual packing scheme of the fluoroalanines. Only molecular centroids were used in this calculation. The cutoff values $5, 20, 50, 75$ and 100 were used in combination with 1000 random sets of the pair η, R_s in which η could take the value 1, 2, 3 & 4 and R_s could take the values $-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5$ and 4. Each set contained a random combinations of G^2 functions which one can build from the above values. In all cases 50 bins were used in the histogram. As such we produced a set of 5000 AMI values for 5000 different clusterings of the ACSF descriptor using functions $\{G_i^1\}$ and $\{G_i^2\}$ only. The results are plotted in the histogram in figure 5.88. The best model gives an AMI of 0.685 which is actually better than that of our best 3D persistent homology model (0.612) although note that a slightly higher AMI might not necessarily mean a better separation of classes and hence a better descriptor. The optimal parameters were $R_c = 20$ and the set of pairs η, R_s : (1, -1), (1, 0), (1, 0.5), (1, 2), (1, 3), (1, 3.5), (1, 4), (2, -2), (2, -1), (2, -0.5), (2, 1), (2, 3), (2, 3.5), (2, 4), (3, -2), (3, -1.5), (3, 0), (3, 0.5), (3, 1.5), (3, 3.5), (3, 4), (4, -0.5), (4, 0.5), (4, 1.5), (4, 2), (4, 2.5). The PCA plot we get with the optimal $\{G_i^1\}$ and $\{G_i^2\}$ functions is shown in figure 5.89.

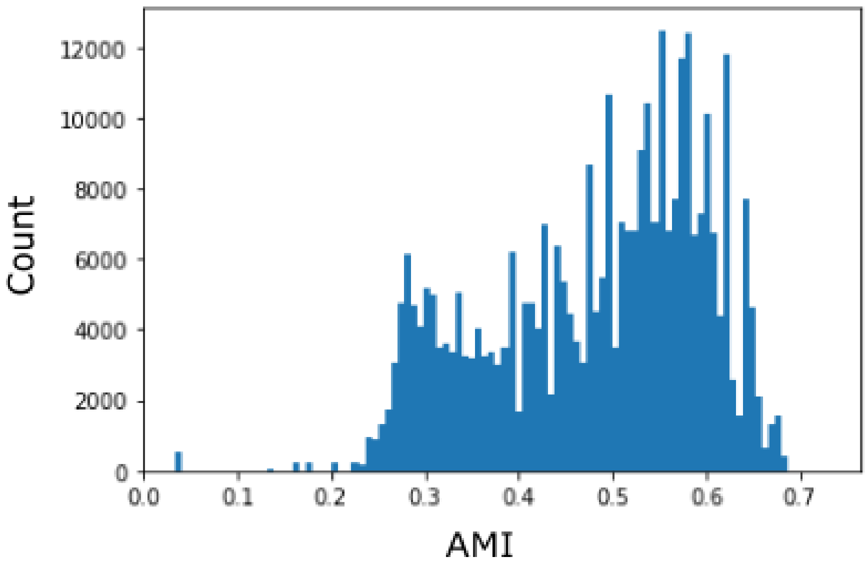


FIGURE 5.88: Variation of Adjusted Mutual Information with respect to the packing scheme of fluoroalanine compounds for an ACSF descriptor applied to a set of 50 molecular centroids extracted from the crystal structure using G_i^1 and G_i^2 functions only.

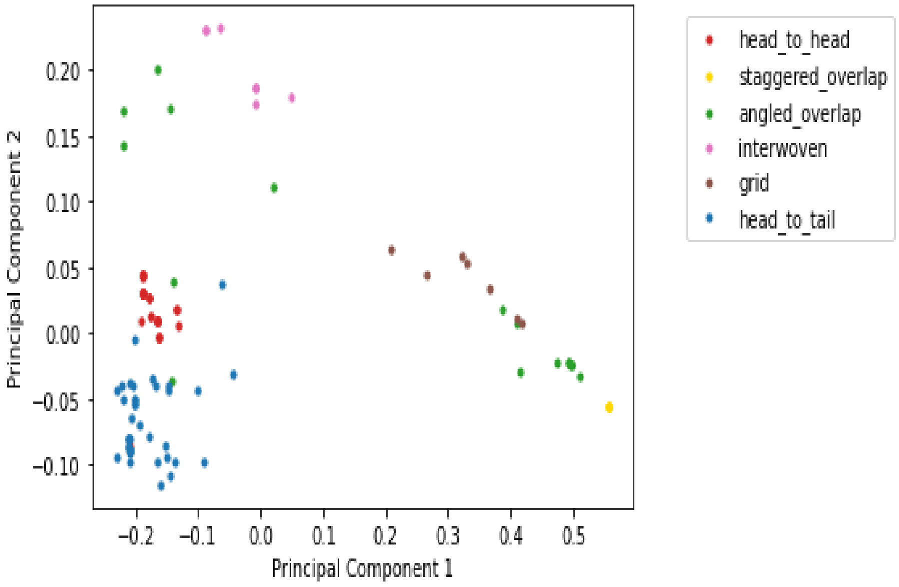


FIGURE 5.89: Principal Components for the best ACSF descriptor using $\{G_i^1\}$ and $\{G_i^2\}$ functions only for a set of fluoroalanine crystals. The compounds are labelled by the packing scheme of these crystals identified by visual argument. The optimal parameters were $R_c = 20$ and the set of pairs η, R_s : (1, -1), (1, 0), (1, 0.5), (1, 2), (1, 3), (1, 3.5), (1, 4), (2, -2), (2, -1), (2, -0.5), (2, 1), (2, 3), (2, 3.5), (2, 4), (3, -2), (3, -1.5), (3, 0), (3, 0.5), (3, 1.5), (3, 3.5), (3, 4), (4, -0.5), (4, 0.5), (4, 1.5), (4, 2), (4, 2.5).

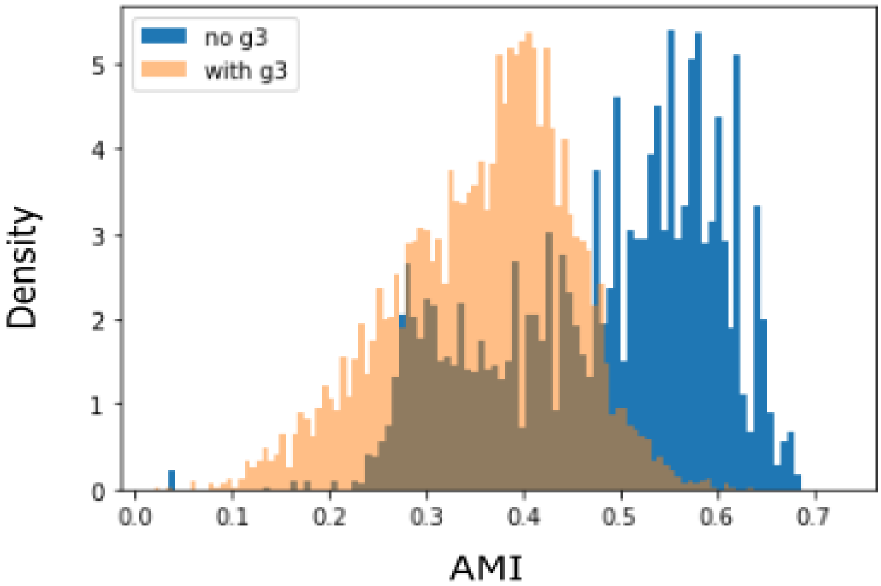


FIGURE 5.90: Variation of Adjusted Mutual Information with respect to the packing scheme of fluoroalanine compounds for an ACSF descriptor applied to a set of 50 molecular centroids extracted from the crystal structure using G_i^1, G_i^2 and G_i^3 functions only. This is superimposed with the histogram obtained using G_i^1 and G_i^2 only.

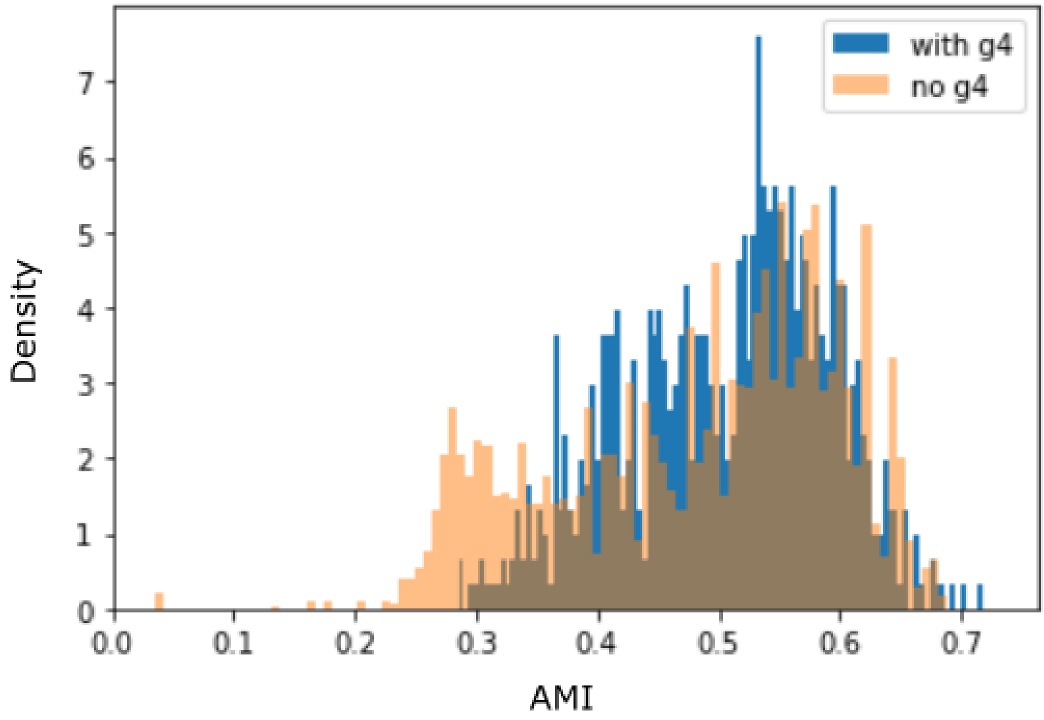
We produced a further set of models using G_i^3 functions as well. Here we used 1000 random sets of functions with random cutoff and G_i^2 parameters chosen as before and G_i^3 parameters which could take 50 values of κ from 1 to 10. The histogram of AMIs is plotted in figure 5.90 superimposed with that from figure 5.88 - we see that adding G_i^3 functions does not tend to improve the model. We do not consider these functions further.

We performed similar comparisons with ACSF functions using G_i^4 and G_i^5 functions. The values ζ, λ and η were allowed respectively to take the values $1, 2, 3, 4$ & 5; the values 0.1 to 10^6 (sampled logarithmically) and all values between -0.99 and 0.99 with increment 0.1. Note because these calculations are much more computationally expensive we were only able to sample 700 and 768 combinations for G_i^4 and G_i^5 respectively. The histograms of AMI thus obtained are shown in figure 5.91. We see a very moderate improvement in AMI when the three body terms are included indicating that it is, perhaps, in general better to include three body terms than not although the difference is hardly night and day. The best descriptors using G_i^4 and G_i^5 functions yield an AMI of (respectively) 0.717 and 0.706. Not a massive difference. The PCA plots for the best models with G_i^4 and G_i^5 functions are shown in figure 5.92. Considering the increased computation time it is debatable whether this slightly improved descriptor is of practical value. Clearly we have only sampled a very small portion of the space of possible descriptors with different ACSF functions so a larger set of calculations with further combinations of hyperparameters might yield better results. The salient point here, however, is that a descriptor that conveys many body interactions will not always significantly outperform a well optimised descriptor that only considers pairwise interactions. We also note that one of the clear limitations of the ACSF descriptor is the very large number of hyperparameters. For this reason a more rigorous descriptor was developed by Bartók *et al.* Bartók et al. (2013) called the Smooth Overlap of Atomic Potentials (SOAP). This is also a local descriptor which, in summary, for each atom in a structure computes a density function $\rho(\mathbf{r})$ (which is itself a sum of Gaussian functions), so that a rotationally averaged set of overlaps can be used to compare local atomic environments - this only computes pairwise overlaps. It turns out that there is a rigorous method (using optimal transport theory see De et al. (2017)) for combining these local atomic similarities into a single similarity kernel which can compare two sets of atomic environments, that is, one can compare two atomic or molecular structures. This is called the SOAP-reMATCH kernel. We can compute this kernel for our set of crystal fragments (centroids only) and use MDS embedding by invoking the kernel induced distance [10]. We get the reduced space shown in figure 5.93. Note that the plot bears striking resemblance to that obtained using the Axilrod-Teller interaction in spite of this method involving no three body interactions.

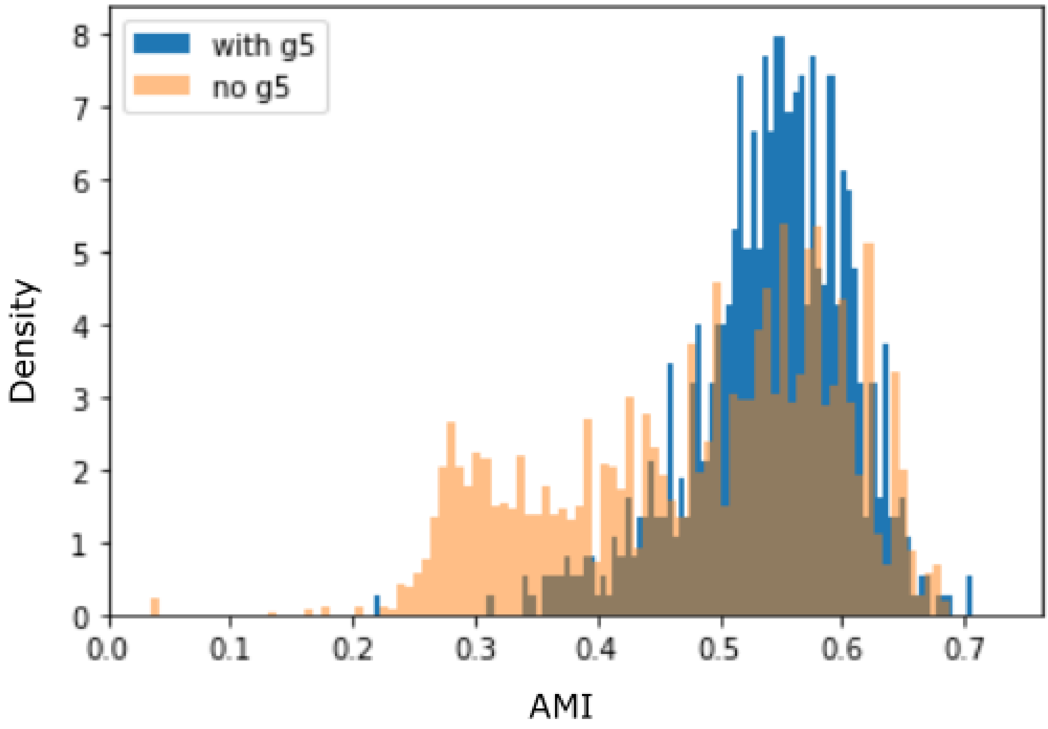
Lastly we consider an alternative formulation of the use of pairwise Euclidean distances we discussed at the start of this section. Rather than considering *all* pairwise distances, we can instead focus on the distances of any given point to its k^{th} nearest neighbour. We have already made use of such distances when we were attempting to rescale the intercentroid distances during our discussion of the set of polyaromatic hydrocarbons from the CSD - they are very easy to compute - the kD tree is one very efficient method Bentley (1975). Widdowson *et al.* Widdowson et al. (2022) have introduced two simple descriptors that focus on these distances. For the case of an infinite lattice and using atomic positions they show that these descriptor are *isometry invariants* which means that if two structures are the same the descriptors are the same (and *vice versa*) and that the differences between the values of these descriptors are such that they are continuous with respect to perturbation of the crystal structure. They are also independent of choice of unit cell. We are less concerned with the precise mathematical properties of these descriptors as the application of this descriptor on to *molecular* positions is not the original use case. We nevertheless find that one of these descriptors performs very well on our sets of molecular centroids.

The definition of these descriptors are actually quite simple. The first descriptor is called PDD_k or the k^{th} pointwise distance distribution. This is a local descriptor and is just the list of the k-nearest neighbours of a given atom. Like we did with the ACSF descriptor earlier we can just convert this into a histogram for our use case. In figure

[10] Given a kernel $K(A,B)$ between environments A and B. An expression that satisfies the mathematical conditions for a distance metric is $D(A,B) = \sqrt{\left[K(A,A) + K(B,B) - 2K(A,B)\right]}$. This is the kernel induced distance.

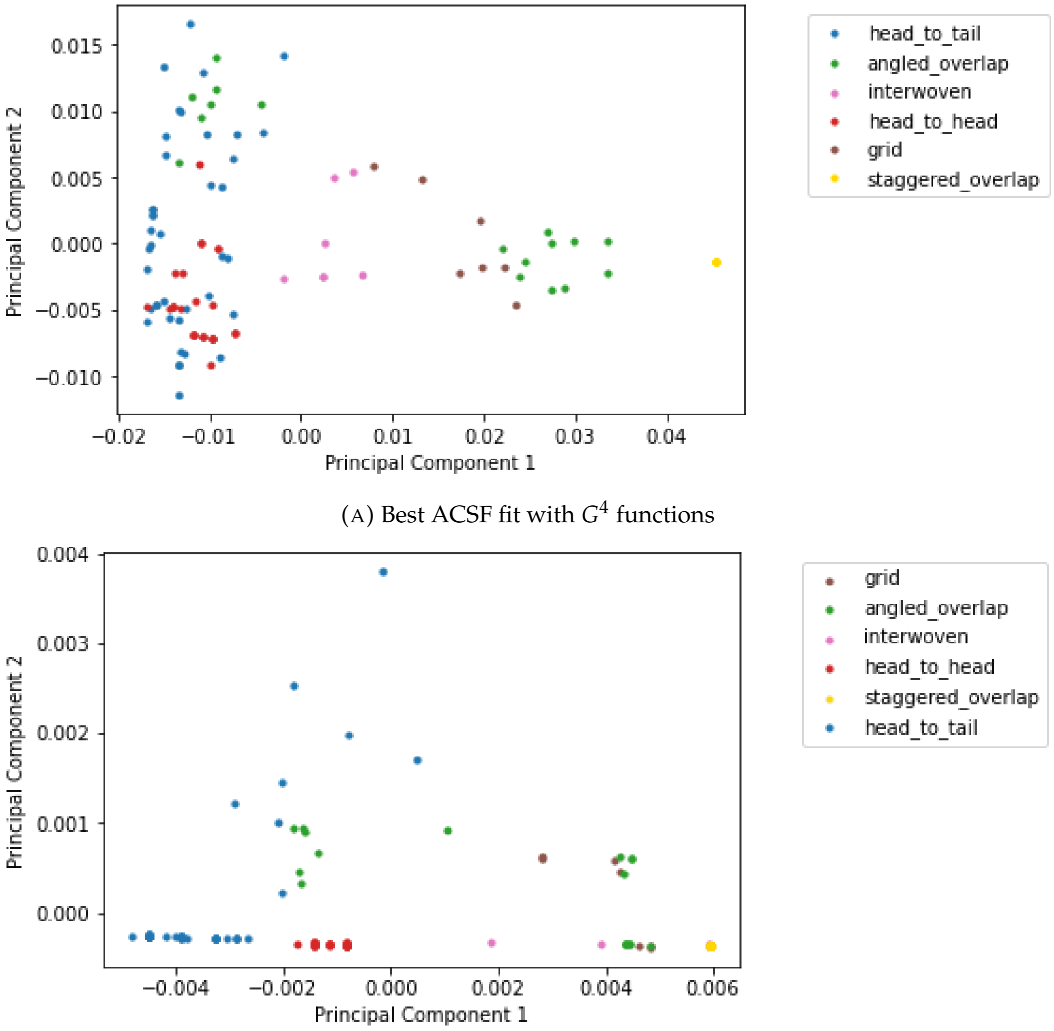


(A) ACSF fits with G^4 functions



(B) ACSF fits with G^5 functions

FIGURE 5.91: Variation of Adjusted Mutual Information with respect to the packing scheme of fluoroalanine compounds for an ACSF descriptor applied to a set of 50 molecular centroids extracted from the crystal structure using G_i^1, G_i^2 and G_i^4 functions only for part a) and using G_i^1, G_i^2 and G_i^5 functions only for part b) . These are superimposed with the histograms obtained using G_i^1 and G_i^2 only.



(A) Best ACSF fit with G^4 functions

(B) Best ACSF fits with G^5 functions

FIGURE 5.92: Principal Components for the best ACSF descriptor using G_i^1, G_i^2 and G_i^4 (part a) or G_i^5 (part b) functions only for a set of fluoroalanine crystals. The compounds are labelled by the packing scheme of these crystals identified by visual argument. The list of parameters for these models is too long to list here.

5.94 we show the first two principal components of the PDD_k descriptor with increasing value of k (we use 10 bins in the histograms) for our set of fluoroalanine pointclouds.

The separation of classes is pretty good but we find that the second descriptor proposed by Widdowson *et al.* is better. The descriptor AMD_k or the k^{th} average minimum distance is related to the previous descriptor: it is simply the average of the PDD for each atom. Put differently AMD_k is a list of k numbers and the i^{ith} $(i \leq k)$ element of which is the average distance of any given atom to its i^{th} nearest neighbour. This is a global descriptor so we do not need to mess around with histograms. In figure 5.95 we show the first two principal components of the AMD_k descriptor with increasing value of k for our set of fluoroalanine pointclouds. We have seen once again

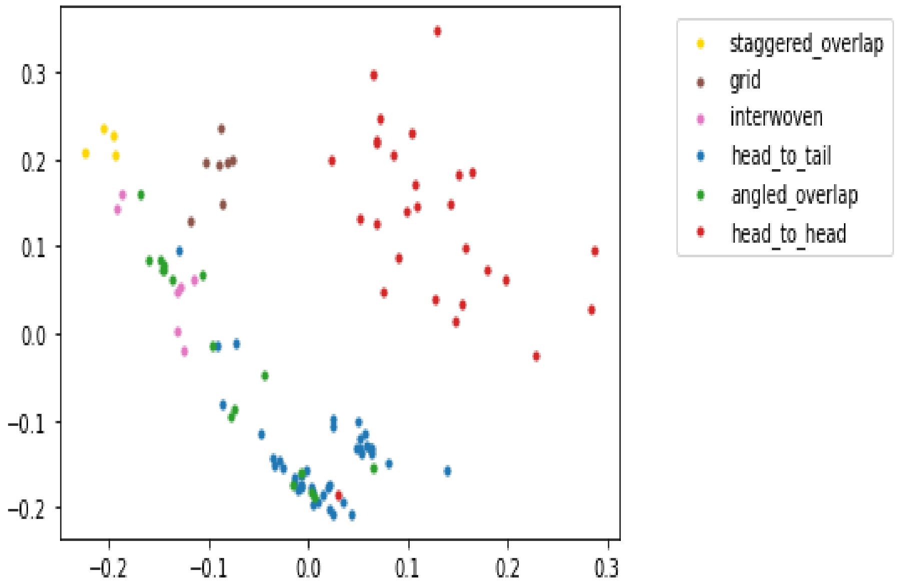
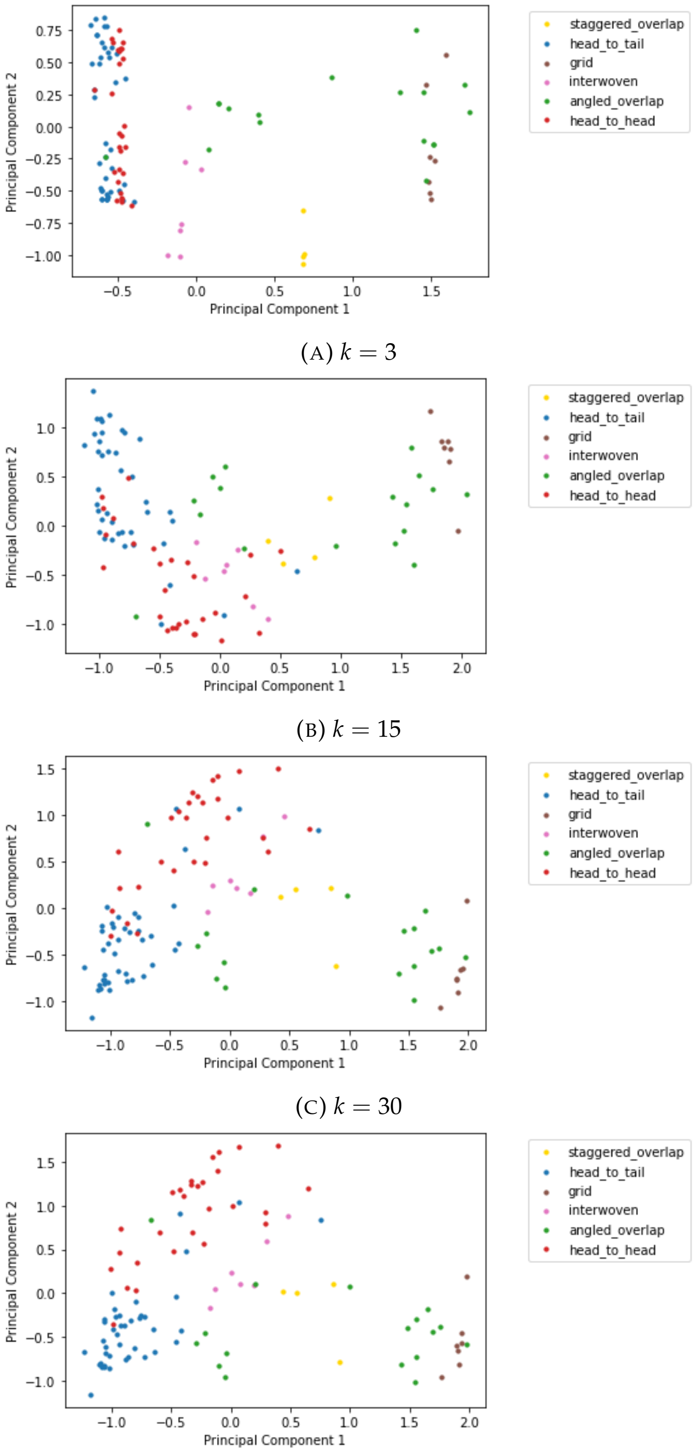


FIGURE 5.93: MDS embedding of the set of kernel induced distances obtained from the SOAP-reMATCH kernel between sets of crystal fragments containing 50 molecular centroids from fluoroalanine crystal structures. The structures are labelled by the packing scheme of these crystals identified by visual argument. Once again the axes of the space into which we are embedding have no physical meaning so remain unlabelled.

that we can separate the fluoroalaine packing classes at least as well as the Axilrod-Teller descriptor which gives us further evidence that many body terms, while useful, are not necessarily key to a good descriptor for organic crystal structures. Also note the similarity between the reduced spaces obtained for AMD_{50}, the SOAP-reMATCH kernel and the Axilrod-Teller tensor. This could be a coincidence but it is also possible that certain features of this particular dataset are quite prominent so we see certain features in the reduced descriptor space regardless of the descriptor.

We started this section by posing a question as to whether our successful description of the packing classes of organic crystals was truly a result of our persistent homology methods or simply just a result of a judicious choices of information to be extracted from the crystal structures in the first instance. We have seen that we can describe the packing classes rather well using the point cloud of molecular centroids with many different descriptors. This does *not* mean that the answer to the question ”Do we really need persistent homology?” is ”no”. For one thing the 3D persistent homology descriptors we generated earlier are still among the best of the descriptors we explored and further none of the models proposed here were able to get a better result than the persistent homology using a 6D pointcloud. It might be better to pose the question ”Should we *only* use persistent homology?”, the answer to which is *definitely* ”no”. We have seen that lots of different equivalent geometric features can help us understand the properties of crystal packing. Descriptors which are simpler and that can be readily interpreted in terms of various kinds of many-body terms or k^{th} nearest



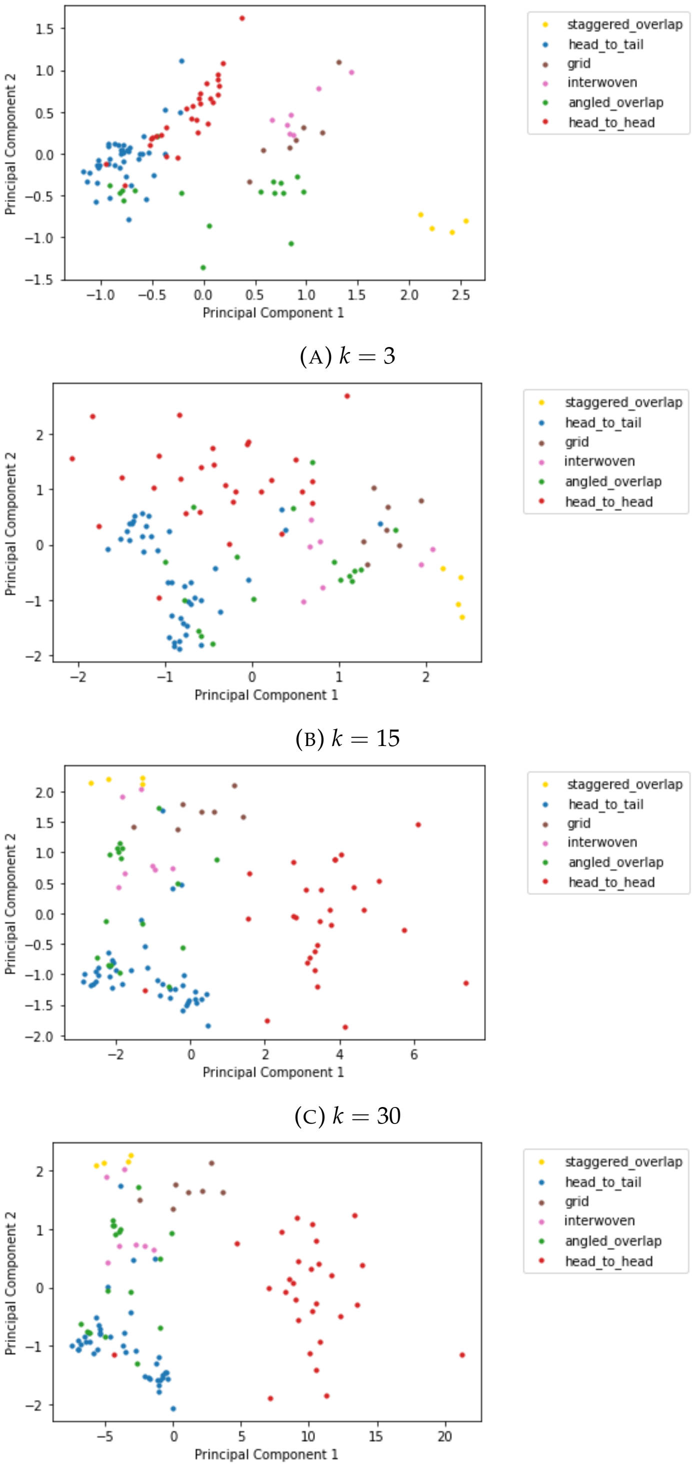
(A) $k = 3$

(B) $k = 15$

(C) $k = 30$

(D) $k = 50$

FIGURE 5.94: The first two principal components of the histogram PDD_k descriptor with 10 bins and for increasing k for a set of 50 molecular centroids extracted from fluorolanine crystals. The structures are labelled by the packing scheme of these crystals identified by visual argument.



(A) $k = 3$

(B) $k = 15$

(C) $k = 30$

(D) $k = 50$

FIGURE 5.95: The first two principal components of the AMD$_k$ descriptor for increasing k for a set of 50 molecular centroids extracted from fluorolanine crystals. The structures are labelled by the packing scheme of these crystals identified by visual argument.

distances are invaluable while one of the clear disadvantages of persistent homology is the fact that it can come across as a black-box. It is very difficult indeed, for example, to find an important feature from a persistence diagram and reverse engineer the geometric structure to which it corresponds (see Obayashi (2018) for example). We also note that we have not in this section found many convincing methods for encoding the molecular orientation vectors in non-topological descriptors. This would be a very desirable outcome of further work. Finding a mathematically sensible way to model the interaction between *three* vector like objects would be very desirable, for example. One attractive avenue might be using sets of spherical harmonics, the multipoles which describe the charge distribution of a molecule are modelled by these functions and the interactions thereof modelled by their overlap. By representation the points in our pointcloud as a series of multipoles (whether or not these multipoles correspond to the actual charge distribution of the molecule which could be time consuming to calculate). This way we could not only model the interaction between three point-wise terms and three vector terms, we could also model the interaction between two or more mathematical objects with a more complex directionality - perhaps we could model the molecular shape. This kind of descriptor even if not better than persistent homology might be more interpretable as one could look at the coefficients of the terms pertaining to given kinds of interactions when evaluating a model. This is not unprecedented: Zhu *et al.* Zhu et al. (2022) have recently published work on using spherical harmonic expansions within crystal structures to understand the packing classes of hydrocarbons.

Chapter 6

Conclusion and Further Work

We have used persistent homology to understand the packing structure of four distinct crystal systems: fluoroalanines, nicotinamide:benzoic acid co-crystals, azapentacenes and general polyaromatic hydrocarbons. In all of these systems we were able to gain at least some insight into the packing system through the suitable conversion of the persistent homology into a descriptor - insight which generally matches the description of crystal packing which was already well established.

The fluoroalanine dataset proved to be an ideal sandbox to test new ideas for the application of persistent homology to crystal structures owing to its small size and high fidelity packing labels. The first key insight was the importance of using a coarser description of the crystal structure as the input for the persistence calculation - calculations involving molecular centroids vastly outperformed those using atomic positions even when the differing atomic radii were taken into account. A further advantage to this approach is that it means that an accurate description of the persistent homology can be found from a pointcloud involving few points. This makes the computation more tractable from both a time and memory perspective and allowed us to calculate the persistent homology of progressively larger crystal fragments so that we could verify that the persistent homology does not change much past a certain threshold. This helped us allay concerns about the minimal structure that should be used for persistent homology as larger structures are always computationally tractable and are likely to contain *exactly the same* topological features just with different multiplicity.

Another key insight that was gained during this investigation was the possibility of encoding information about orientations into the crystal descriptor by the means of, in the first instance, a suitably chosen interatomic vector in the molecule and then later a suitably chosen inertia eigenvector. This approach proved to be very powerful and we were able to generate a low dimensional representation of the set of fluorolainines that has all of the main packing classes (as identified by a subject matter expert) almost entirely separated in this space. We established later that a similar level of class separation was not achievable with a wider set of conventional and unconventional descriptors. For the GAZCES co-crystal (after augmenting the relative contribution of the orientational information vs. the positional) we were able to almost completely separate the structures of co-crystals that were identifiable with one of two potential energy wells in the wider energy landscape of this crystal structure. The original researchers Yang and Day (2022) were not able to achieve this with conventional descriptors. Finally although the azapentacene crystal structure landscapes were generally less amenable to full description by persistent homology we were able to use an augmented 6D persisting homology descriptor to generate crystal structure landscapes that begin to isolate certain subsets of the canonical packing classes and begin to show concrete trends in the data. An important topic of further work would be to establish if the idea of encoding molecular orientation or other relevant chemical properties into the geometric structure of a crystal system could be used to improve existing descriptors or indeed design new ones.

During the course of this project we also demonstrated the value of supervised dimensionality reduction techniques in gaining insight into larger crystal structure landscapes in the context of smaller datasets of the same or similar compounds. The supervised UMAP algorithm has not yet found much use in the literature but has allowed us to generate crystal structure landscapes that, by construction, heavily emphasise chemically relevant information. These techniques should always be used in the context of other supervised dimensionality reduction techniques such as LDA and in the context of predictions from conventional classification and clustering algorithms as, due to the highly non-linear nature of the algorithm, the plots can be a little misleading. Within a sensible context this technique could certainly have more applications in this field. The power of these techniques illustrates the importance and value of small highly curated datasets with high fidelity annotation by subject matter experts. These data can be used to contextualise and understand much larger volumes of data which lack such curation.

Another, slightly more unexpected result, was that the persistent homology descriptors do have some predictive power for the energy of the azapentacene molecules. This has implications on the further use cases of our methodology so this should be studied further by retraining our models with more accurate energy calculations and with a more extensive choice of models and hyperparameters.

A key area of further work should be the acquisition of a dataset of the electron density surfaces of a curated library of crystal structures, either by experimental means (hopefully incidentally while recording the structure) or computationally. The predictive power of the topological characteristics of this function, extracted by sublevel set persistent homology, would be fascinating. As the quantum mechanical structure of the molecules would be encoded in the electron density data, the topological invariants might be more likely to have predictive power for chemical properties than our current pointset approach. Moreover establishing a surface that captures the shape of each molecule without needing to include every atom in the persistent homology filtration (which tends to generate redundant intramolecular homology features while suppressing useful intermolecular features) would be very desirable and perhaps help us address the fact that the correct intermolecular interactions are not always encoded in the homology filtration - this could be the reason that our current persistent homology descriptors do not completely describe the packing classes of polyaromatic hydrocarbons.

The most overwhelming need however in order to be able to find further applications of persistent homology in organic crystal systems and to be able to generate new descriptors (using homology or otherwise) is to have more access to higher quality data. In particular, even small amounts of data which have been labelled by subject matter experts are invaluable and can allow us to refine our models further, establish different use cases and explore the limitations of our methods. In particular it would be interesting to explore the application of persistent homology to more co-crystal structures in order to establish if the near complete separation of the two classes of GAZCES crystals by our augmented homology descriptor was simply a fluke or a natural consequence of the co-crystal structure, the orientation vector acting as proxy for the two different sub-units and thus allowing us to build a descriptor that compares the positions of these sub-units relative to one another. If this is true then our descriptors might have particular predictive power for these systems.

Finally we note that while we have shown in multiple cases that persistent homology can be a very useful descriptor for organic crystal systems it is by no means a panacea for describing them and we should be clear eyed about its limitations. There are three key limitations that should be borne in mind during further study.

Firstly while ostensibly topological data analysis is something that extracts inherent properties of the data and does not involve hyperparameters that need tuning like other methods, the techniques we have developed have *a lot* flexibility and there is a lot of nuance in *how* topological data analysis is applied in the first instance and then how the resulting persistence diagram is processed. During our particular workflow we must decide:

- What information should be extracted from the crystal structure, atoms, centroids, vectors etc.
- How many of these features to extract

- Do we want to also include crystal fragments with different shapes as in the restricted supercell approach we employed for the polyaromatic hydrocarbons
- What persistent homology filtration will we use
- Will we fix the maximum dimension or edge length during the homology filtration
- What technique will we use to convert the persistence diagram to a descriptor - using the vector images was an arbitrary choice in some ways
- *How* to carry out the conversion of the persistence diagram into a descriptor with the given method, for example with the vector images we need to decide on a resolution, we also need to decide on any weighting function (we do not do this in our work but we certainly did not experiment with this for every dataset).
- Do we want to normalise/rescale any of the features or prune unwanted/unphysical features
- What homology features do we actually want to use

and all of this takes place *before* we even decide on which data analysis techniques should be applied. The point here is simple. Persistent homology is *not* easy to use for a non subject matter expert and there are very many things that can and should be tweaked before, during and after the workflow.

Secondly persistent homology is *not* necessarily easily interpretable: even if one can establish exactly what geometric structure a homology feature corresponds to, this could be difficult to visualise and/or explain. While it is eminently true that often these descriptors are not directly interpreted, the direct interpretablity of persistent homology is often sold as one of its strengths Musil et al. (2021) which does not necessarily reflect the reality of doing persistent homology in practice.

Finally oftentimes the application of persistent homology seems to involve the scientific process in retrograde - one starts with a model and decides what problem to apply it to! There is an adage that topological data analysis is a solution looking for a problem. This need not be the case, but it is very important that the system to be studied is carefully considered first. After all most of the successes, but by no means all, of our technique came from identifying which key features of the structural chemistry to *encode* into our topological data analysis approach. We saw in the last section that other descriptors that are fed the same data do pretty well, even if persistent homology was the best model. Therefore for any further work and for all further datasets the most pertinent task is to identify which features of the crystal geometry are most important to the chemistry and figure out how to encode that with topological data analysis or otherwise. For example, further work on the azapentacenes and polyaromatic hydrocarbons should focus on methods for encoding the rod-like shape of these molecules as well as the precise face-to-face and edge-to-face interactions which are most important of determining the packing class. This may or may not involve persistent homology. The methods that will arise out of such analysis will not only have the advantage of being more likely to be correct but also have the advantage of being more interpretable and useful from the perspective of a chemist.

Bibliography

Ch B Aakeröy. Crystal Engineering: Strategies and Architectures. *Acta Crystallographica Section B: Structural Science*, 53(4):569–586, 1997.

Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence Images: A Stable Vector Representation of Persistent Homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017. URL `http://jmlr.org/papers/v18/16-337.html`.

Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study BT - Image and Signal Processing. In Abderrahim El Moataz, Driss Mammass, Alamin Mansouri, and Fathallah Nouboud, editors, *Image and Signal Processing*, pages 317–325. Springer International Publishing, 2020. ISBN 978-3-030-51935-3.

Dylan M Anstine and Olexandr Isayev. Machine Learning Interatomic Potentials and Long-Range Physics. *The Journal of Physical Chemistry A*, 127(11):2417–2431, March 2023. ISSN 1520-5215 (Electronic).

John E. Anthony. The Larger Acenes: Versatile Organic Semiconductors. *Angewandte Chemie International Edition*, 47(3):452–483, January 2008. ISSN 1433-7851. URL `https://doi.org/10.1002/anie.200604045`.

David Arthur and Sergei Vassilvitskii. k-means++: the Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 9780898716245.

B M Axilrod and E Teller. Interaction of the van der Waals Type Between Three Atoms. *The Journal of Chemical Physics*, 11(6):299–300, June 1943. ISSN 0021-9606. URL `https://doi.org/10.1063/1.1723844`.

Paolo Bajardi, Matteo Delfino, André Panisson, Giovanni Petri, and Michele Tizzoni. Unveiling Patterns of International Communities in a Global City Using Mobile

Phone Data. *EPJ Data Science*, 4(1):3, 2015. ISSN 2193-1127. URL https://doi.org/10.1140/epjds/s13688-015-0041-5.

Albert P. Bartók, Risi Kondor, and Gábor Csányi. On Representing Chemical Environments. *Physics Review B*, 87:184115, May 2013. URL https://link.aps.org/doi/10.1103/PhysRevB.87.184115.

Albert P Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R Kermode, Gábor Csányi, and Michele Ceriotti. Machine Learning Unifies the Modeling of Materials and Molecules. *Science Advances*, 3(12), April 2024. URL https://doi.org/10.1126/sciadv.1701816.

Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality Reduction for Bisualizing Single-Cell Data Using UMAP. *Nature Biotechnology*, December 2018. ISSN 1087-0156. URL https://www.nature.com/articles/nbt.4314.

Oren M Becker and Martin Karplus. The Topology of Multidimensional Potential Energy Surfaces: Theory and Application to Peptide Structure and Kinetics. *The Journal of Chemical Physics*, 106(4):1495–1517, January 1997. ISSN 0021-9606. URL https://doi.org/10.1063/1.473299.

Jörg Behler. Atom-centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials. *The Journal of Chemical Physics*, 134(7):74106, Febuary 2011. ISSN 0021-9606. URL https://doi.org/10.1063/1.3553717.

Greg Bell, Austin Lawson, Joshua Martin, James Rudzinski, and Clifford Smyth. Weighted Persistent Homology. *Involve, a Journal of Mathematics*, 12, 2017.

William A Belson. Matching and Prediction on the Principle of Biological Classification. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 8(2): 65–75, January 1959. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2985543.

Jon Louis Bentley. Multidimensional Binary Search Trees Used for Associative Searching. *Communications of the Association of Computing Machinery*, 18(9):509–517, September 1975. ISSN 0001-0782. URL https://doi.org/10.1145/361002.361007.

Ricarda Berger, Giuseppe Resnati, Pierangelo Metrangolo, Edwin Weber, and Jürg Hulliger. Organic Fluorine Compounds: a Great Opportunity for Enhanced Materials Properties. *Chemical Society Reviews*, 40(7):3496–3508, 2011. ISSN 0306-0012. URL http://dx.doi.org/10.1039/C0CS00221F.

Omer Bobrowski and Primoz Skraba. Homological Percolation and the Euler Characteristic. *Physical Review. E*, 101, March 2020.

Jan Böhm, Philipp Berens, and Dmitry Kobak. A Unifying Perspective on Neighbor Embeddings along the Attraction-Repulsion Spectrum. *ArXiv*, abs/2007.08902, 2020. URL https://api.semanticscholar.org/CorpusID:220633055.

Ingwer Borg and Patrick J F Groenen. *Modern Multidimensional Scaling: Theory and Applications, 2nd ed.* Springer Series in Statistics. Springer Science + Business Media, New York, NY, US, 2005. ISBN 0-387-25150-2 (Hardcover); 978-0387-25150-9 (Hardcover).

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 144–152, New York, NY, USA, 1992. Association for Computing Machinery. ISBN 089791497X. URL https://doi.org/10.1145/130385.130401.

David H Bowskill, Isaac J Sugden, Stefanos Konstantinopoulos, Claire S Adjiman, and Constantinos C Pantelides. Crystal Structure Prediction Methods for Organic Molecules: State of the Art. *Annual Review of Chemical and Biomolecular Engineering*, 12:593–623, June 2021. ISSN 1947-5446 (Electronic).

Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565. URL https://doi.org/10.1023/A:1010933404324.

Leonard A. Breslow and David W. Aha. Simplifying Decision Trees: A Survey. *The Knowledge Engineering Review*, 12(1):1–40, 1997.

Krystof Brezina, Hubert Beck, and Ondrej Marsalek. Reducing the Cost of Neural Network Potential Generation for Reactive Molecular Systems. *Journal of Chemical Theory and Computation*, 19(19):6589–6604, October 2023. ISSN 1549-9626 (Electronic).

Rasmus Bro and Age K Smilde. Principal Component Analysis. *Analytical Methods*, 6 (9):2812–2831, 2014. URL http://dx.doi.org/10.1039/C3AY41907J.

Peter Bubenik. Statistical Topological Data Analysis using Persistence Landscapes. *Journal of Machine Learning Research*, 16(3):77–102, 2015. URL http://jmlr.org/papers/v16/bubenik15a.html.

Josh E Campbell, Jack Yang, and Graeme M Day. Predicted Energy–Structure–Function Maps for the Evaluation of Small Molecule Organic Semiconductors. *Journal of Materials Chemistry C*, 5(30):7574–7584, 2017. URL http://dx.doi.org/10.1039/C7TC02553J.

Zhuo Cao, Yabo Dan, Zheng Xiong, Chengcheng Niu, Xiang Li, Songrong Qian, and Jianjun Hu. Convolutional Neural Networks for Crystal Material Property Prediction Using Hybrid Orbital-Field Matrix and Magpie Descriptors. *Crystals*, 9 (4):191, 2019.

Manuel Caroli and Monique Teillaud. Computing 3D Periodic Triangulations. In *ESA Symposia- 17th European Symposium on Algorithms*, pages 59–70, 2009.

Mathieu Carriere, Marco Cuturi, and Steve Oudot. Sliced Wasserstein kernel for persistence diagrams. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 664–673. JMLR.org, 2017.

Mathieu Carrière, Marco Cuturi, and Steve Oudot. Sliced Wasserstein Kernel for Persistence Diagrams. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 664–673. PMLR, August 2017. URL https://proceedings.mlr.press/v70/carriere17a.html.

David H Case, Josh E Campbell, Peter J Bygrave, and Graeme M Day. Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling. *Journal of Chemical Theory and Computation*, 12(2):910–924, Febuary 2016. ISSN 1549-9626 (Electronic).

Michele Ceriotti, Gareth A Tribello, and Michele Parrinello. Simplifying the Representation of Complex Free-Energy Landscapes Using Sketch-Map. *Proceedings of the National Academy of Sciences*, 108(32):13023–13028, August 2011. URL https://doi.org/10.1073/pnas.1108486108.

Su Chen. Optimal Bandwidth Selection for Kernel Density Functionals Estimation. *Journal of Probability and Statistics*, 2015, 2015. ISSN 1687-952X. URL https://doi.org/10.1155/2015/242683.

Yen-Chi Chen. A Tutorial on Kernel Density Estimation and Recent Advances. *Biostatistics Epidemiology*, 1(1):161–187, January 2017. ISSN 2470-9360. URL https://doi.org/10.1080/24709360.2017.1396742.

James A Chisholm and Sam Motherwell. COMPACK: a Program for Identifying Crystal Structure Similarity Using Distances. *Journal of Applied Crystallography*, 38: 228–231, 2005. URL https://api.semanticscholar.org/CorpusID:94406296.

Jawad Chowdhury, Charles Fricke, Olajide Bamidele, Mubarak Bello, Wenqiang Yang, Andreas Heyden, and Gabriel Terejanu. Invariant Molecular Representations for Heterogeneous Catalysis. *Journal of Chemical Information and Modeling*, 64(2):327–339, January 2024. ISSN 1549-960X (Electronic).

Pierre Christian, Chi-kwan Chan, Anthony Hsu, Feryal Özel, Dimitrios Psaltis, and Iniyan Natarajan. Topological Data Analysis of Black Hole Images. *Physical Review D*, 106(2):23017, July 2022. URL https://link.aps.org/doi/10.1103/PhysRevD.106.023017.

Yu-Min Chung, Chuan-Shen Hu, Yu-Lun Lo, and Hau-Tieng Wu. A Persistent Homology Approach to Heart Rate Variability Analysis With an Application to Sleep-Wake Classification. *Frontiers in Physiology*, 12, 2021. ISSN 1664-042X. URL https://www.frontiersin.org/articles/10.3389/fphys.2021.637684.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of Persistence Diagrams. *Discrete and Computational Geometry*, 37(1):103–120, 2007. ISSN 1432-0444. URL https://doi.org/10.1007/s00454-006-1276-5.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20 (3):273–297, 1995. ISSN 1573-0565. URL https://doi.org/10.1007/BF00994018.

James Damewood, Jessica Karaguesian, Jaclyn R. Lunger, Aik Rui Tan, Mingrou Xie, Jiayu Peng, and Rafael Gómez-Bombarelli. Representations of Materials for Machine Learning. *Annual Review of Materials Research*, 53:399–426, 2023. ISSN 1545-4118. URL https://www.annualreviews.org/content/journals/10.1146/annurev-matsci-080921-085947.

Ian Dance. Distance Criteria for Crystal Packing Analysis of Supramolecular Motifs. *New Journal of Chemistry*, 27(1):22–27, 2003.

G M Day, W D S. Motherwell, and W Jones. A Strategy for Predicting the Crystal Structures of Flexible Molecules: the Polymorphism of Phenobarbital. *Physical Chemistry Chemical Physics*, 9(14):1693–1704, 2007. URL http://dx.doi.org/10.1039/B612190J.

Graeme M Day. Current Approaches to Predicting Molecular Organic Crystal Structures. *Crystallography Reviews*, 17(1):3–52, 2011. URL https://doi.org/10.1080/0889311X.2010.517526.

Sandip De, Felix Musil, Teresa Ingram, Carsten Baldauf, and Michele Ceriotti. Mapping and Classifying Molecules from a High-Throughput Structural Database. *Journal of Cheminformatics*, 9(1):6, 2017. ISSN 1758-2946. URL https://doi.org/10.1186/s13321-017-0192-4.

Carlos Manuel de Armas-Morejón, Luis A Montero-Cabrera, Angel Rubio, and Joaquim Jornet-Somoza. Electronic Descriptors for Supervised Spectroscopic Predictions. *Journal of Chemical Theory and Computation*, 19(6):1818–1826, March 2023. ISSN 1549-9626 (Electronic).

G R Desiraju and A Gavezzotti. Crystal Structures of Polynuclear Aromatic Hydrocarbons. Classification, Rationalization and Prediction from Molecular Structure. *Acta Crystallographica Section B*, 45(5):473–482, October 1989a. ISSN 0108-7681. URL https://doi.org/10.1107/S0108768189003794.

Gautam R Desiraju. Approaches to Crystal Structure Landscape Exploration. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 73(5): 775–778, 2017.

Gautam R Desiraju and A Gavezzotti. From Molecular to Crystal Structure; Polynuclear Aromatic Hydrocarbons. *Journal of the Chemical Society, Chemical Communications*, (10):621–623, 1989b. ISSN 0022-4936. URL `http://dx.doi.org/10.1039/C39890000621`.

Barbara Di Fabio and Massimo Ferri. Comparing Persistence Diagrams Through Complex Vectors. In Vittorio Murino and Enrico Puppo, editors, *Image Analysis and Processing - ICIAP 2015*, pages 294–305, Genova, Italy, 2015. Springer International Publishing. ISBN 978-3-319-23231-7.

Alex Diaz-Papkovich, Luke Anderson-Trocmé, and Simon Gravel. A Review of UMAP in Population Genetics. *Journal of Human Genetics*, 66(1):85–91, 2021. ISSN 1435-232X. URL `https://doi.org/10.1038/s10038-020-00851-4`.

Simena Dinas and José Mar´Banon. A review on Delaunay Triangulation with Application on Computer Vision. *International Journal of Computer Science Engineering*, 3:9–18, 2014.

Pawel Dlotko, Lucy Minford, Simon Rudkin, and Wanling Qiu. An Economic Topology of the Brexit vote, 2019. URL `https://arxiv.org/abs/1909.03490`.

Eleanor Marie Dodd. *A Systematic Study into the Influence of Aromatic Stacking Interactions and Fluorine Substituent Effects on Molecular Organic Crystal Assembly*. PhD thesis, University of Southampton, Chemistry Department, 2020.

Harris Drucker, Christopher J C Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support Vector Regression Machines. In M C Mozer, M Jordan, and T Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. URL `https://proceedings.neurips.cc/paper_files/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf`.

Kai-Bo Duan and S Sathiya Keerthi. Which Is the Best Multiclass SVM Method? An Empirical Study BT - Multiple Classifier Systems. In Nikunj C Oza, Robi Polikar, Josef Kittler, and Fabio Roli, editors, *MCS: International Workshop on Multiple Classifier Systems*, pages 278–285, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31578-0.

Edelsbrunner, Letscher, and Zomorodian. Topological Persistence and Simplification. *Discrete and Computational Geometry*, 28(4):511–533, 2002. ISSN 1432-0444. URL `https://doi.org/10.1007/s00454-002-2885-2`.

H Edelsbrunner and J L Harer. *Computational Topology An Introduction*. American Mathematical Society, Providence, RI, USA, 2010.

Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

Michelle Feng and Mason A Porter. Persistent Homology of Geospatial Data: A Case Study with Voting. *SIAM Review*, 63(1):67–99, 2021. URL `https://doi.org/10.1137/19M1241519`.

R A Fischer. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188, 1936. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x`.

Alastair J Florence, Andrea Johnston, Sarah L Price, Harriott Nowell, Alan R Kennedy, and Norman Shankland. An Automated Parallel Crystallisation Search for Predicted Crystal Structures and Packing Motifs of Carbamazepine. *Journal of Pharmaceutical Sciences*, 95(9):1918–1930, 2006.

Yoav Freund and Robert E Schapire. A Desicion-Theoretic Generalization of On-Line Learning and an Application to Boosting BT - Computational Learning Theory. In Paul Vitányi, editor, *Second European Conference on Computational Learning Theory*, pages 23–37, Berlin, Heidelberg, 1995. Springer Berlin Heidelberg. ISBN 978-3-540-49195-8.

M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian˜16 Revision C.01, 2016. Gaussian Inc. Wallingford CT.

Tobias Gensch, Gabriel dos Passos Gomes, Pascal Friederich, Ellyn Peters, Théophile Gaudin, Robert Pollice, Kjell Jorner, AkshatKumar Nigam, Michael Lindner-D'Addario, Matthew S Sigman, and Alán Aspuru-Guzik. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *Journal of the American Chemical Society*, 144(3):1205–1217, January 2022. ISSN 0002-7863. URL `https://doi.org/10.1021/jacs.1c09718`.

Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Uniform Manifold Approximation and Projection (UMAP) and its Variants: Tutorial and Survey. *ArXiv*, abs/2109.02508, 2021. URL https://api.semanticscholar.org/CorpusID:237420947.

Marian Gidea. Topological Data Analysis of Critical Transitions in Financial Networks. In Erez Shmueli, Baruch Barzel, and Rami Puzis, editors, *3rd International Winter School and Conference on Network Science*, pages 47–59. Springer International Publishing, 2017. ISBN 978-3-319-55471-6.

Aldo Glielmo, Brooke E Husic, Alex Rodriguez, Cecilia Clementi, Frank Noé, and Alessandro Laio. Unsupervised Learning Methods for Molecular Simulation Data. *Chemical Reviews*, 121(16):9722–9758, August 2021. ISSN 1520-6890 (Electronic).

Qiaolin Gou, Jing Liu, Haoming Su, Yanzhi Guo, Jiayi Chen, Xueyan Zhao, and Xuemei Pu. Exploring an Accurate Machine Learning Model to Quickly Estimate Stability of Diverse Energetic Materials. *iScience*, 27(4):109452, April 2024. ISSN 2589-0042 (Electronic).

Michael Greenacre, Patrick J F Groenen, Trevor Hastie, Alfonso Iodice D'Enza, Angelos Markos, and Elena Tuzhilina. Principal Component Analysis. *Nature Reviews Methods Primers*, 2(1):100, 2022. ISSN 2662-8449. URL https://doi.org/10.1038/s43586-022-00184-w.

Graziela Grise and Michael Meyer-Hermann. Surface Reconstruction Using Delaunay Triangulation for Applications in Life Sciences. *Computer Physics Communications*, 182(4):967–977, 2011.

Colin R Groom, Ian J Bruno, Matthew P Lightfoot, and Suzanna C Ward. The Cambridge Structural Database. *Acta Crystallographica Section B*, 72(2):171–179, April 2016. URL https://doi.org/10.1107/S2052520616003954.

Martin C Grossel, Michael B Hursthouse, and James B Orton. Structural Investigation of x,y-bis-(chlorocarbonyl) Pyridines Derivatives: "Strength in Diversity"—a Disparity of Supramolecular Packing Motifs. *CrystEngComm*, 7(45):279–283, 2005. URL http://dx.doi.org/10.1039/B501034A.

gudhi_periodic_cubical_complex_manual_page. Gudhi periodic cubical complex manual page, accessed 2023. URL https://gudhi.inria.fr/python/latest/periodic_cubical_complex_ref.html. [Online; accessed 21-Mar-2023 - no citation given but code works].

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2 edition, 2009. URL http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

A Hatcher. *Algebraic Toplogy*. Cambridge University Press, 2001.

Tony Hey, Stewart Tansley, Kristin Tolle, and Jim Gray. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, October 2009. ISBN 978-0-9825442-0-4. URL https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/.

Sven Heydenreich, Benjamin Brück, and Joachim Harnois-Déraps. Persistent Homology in Cosmic Shear: Constraining Parameters with Topological Data Analysis. *Astronomy and Astrophysics*, 648, 2021. URL https://doi.org/10.1051/0004-6361/202039048.

Yasuaki Hiraoka, Takenobu Nakamura, Akihiko Hirata, Emerson G Escolar, Kaname Matsue, and Yasumasa Nishiura. Hierarchical Structures of Amorphous Solids Characterized by Persistent Homology. *Proceedings of the National Academy of Sciences*, 113(26), June 2016. URL http://www.pnas.org/content/113/26/7035.abstract.

Yuta Hozumi, Rui Wang, Changchuan Yin, and Guo-Wei Wei. UMAP-Assisted K-means Clustering of Large-Scale SARS-CoV-2 Mutation Datasets. *Computers in Biology and Medicine*, 131:104264, 2021.

Chih-Wei Hsu and Chih-Jen Lin. A Comparison of Methods for Multiclass Support Vector Machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

Jenq-Neng Hwang, Shyh-Rong Lay, and A Lippman. Nonparametric Multivariate Density Estimation: a Comparative Study. *IEEE Transactions on Signal Processing*, 42 (10):2795–2810, 1994.

Olexandr Isayev, Denis Fourches, Eugene N Muratov, Corey Oses, Kevin Rasch, Alexander Tropsha, and Stefano Curtarolo. Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints. *Chemistry of Materials*, 27(3):735–743, Febuary 2015. ISSN 0897-4756. URL https://doi.org/10.1021/cm503507h.

Salima Z Ismail, Clare L Anderton, Royston C B Copley, Louise S Price, and Sarah L Price. Evaluating a Crystal Energy Landscape in the Context of Industrial Polymorph Screening. *Crystal Growth & Design*, 13(6):2396–2406, 2013.

J C Schön, H Putz, and M Jansen. Studying the Energy Hypersurface of Continuous Systems - the Threshold Algorithm. *Journal of Physics: Condensed Matter*, 8(2):143, 1996. ISSN 0953-8984. URL https://dx.doi.org/10.1088/0953-8984/8/2/004.

Kevin Maik Jablonka, Daniele Ongari, Seyed Mohamad Moosavi, and Berend Smit. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chemical Reviews*, 120(16):8066–8129, August 2020. ISSN 0009-2665. URL https://doi.org/10.1021/acs.chemrev.0c00004.

Yi Jiang, Dong Chen, Xin Chen, Tangyi Li, Guo-Wei Wei, and Feng Pan. Topological Representations of Crystalline Compounds for the Machine-Learning Prediction of Materials Properties. *npj Computational Materials*, 7(1):28, 2021.

Ian T Jolliffe and Jorge Cadima. Principal Component Analysis: a Review and Recent Developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 2016. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2015.0202`.

Hyunwook Jung, Sina Stocker, Christian Kunkel, Harald Oberhofer, Byungchan Han, Karsten Reuter, and Johannes T Margraf. Size-Extensive Molecular Machine Learning with Global Representations. *ChemSystemsChem*, 2(4), July 2020. URL `https://doi.org/10.1002/syst.201900052`.

Panagiotis G Karamertzanis and Constantinos C Pantelides. Ab Initio Crystal Structure Prediction—I. Rigid molecules. *Journal of Computational Chemistry*, 26(3): 304–324, Febuary 2005. ISSN 0192-8651. URL `https://doi.org/10.1002/jcc.20165`.

Gurpreet Kaur and Angshuman Roy Choudhury. Understanding of the Weak Intermolecular Interactions Involving Halogens in Substituted N-Benzylideneanilines: Insights from Structural and Computational Perspectives. *Crystal Growth Design*, 14(4):1600–1616, April 2014. ISSN 1528-7483. URL `https://doi.org/10.1021/cg401573d`.

Gurpreet Kaur and Angshuman Roy Choudhury. A Comprehensive Understanding of the Synthons Involving C–HF–C Hydrogen Bond(s) from Structural and Computational Analyses. *CrystEngComm*, 17(15):2949–2963, 2015. URL `http://dx.doi.org/10.1039/C5CE00215J`.

Gurpreet Kaur, Piyush Panini, Deepak Chopra, and Angshuman Roy Choudhury. Structural Investigation of Weak Intermolecular Interactions in Fluorine Substituted Isomeric N-Benzylideneanilines. *Crystal Growth Design*, 12(10):5096–5110, October 2012. ISSN 1528-7483. URL `https://doi.org/10.1021/cg3010294`.

John Kendrick, Frank J J Leusen, Marcus A Neumann, and Jacco van de Streek. Progress in Crystal Structure Prediction. *Chemistry (Weinheim an der Bergstrasse, Germany)*, 17(38):10736–10744, September 2011. ISSN 1521-3765 (Electronic).

David Kirk, editor. *Graphics Gems III*. Academic Press Professional, Inc., USA, 1992. ISBN 0124096719.

Masatoshi Kitamura and Yasuhiko Arakawa. Pentacene-Based Organic Field-Effect Transistors. *Journal of Physics: Condensed Matter*, 20(18):184011, 2008. ISSN 0953-8984. URL `https://dx.doi.org/10.1088/0953-8984/20/18/184011`.

Dmitry Kobak and George C Linderman. UMAP Does not Preserve Global Structure Any Better Than t-SNE When Using the Same Initialization. *bioRxiv*, 2019. URL https://www.biorxiv.org/content/early/2019/12/19/2019.12.19.877522.

Genki Kusano, Yasuaki Hiraoka, and Kenji Fukumizu. Persistence Weighted Gaussian kernel for Topological Data Analysis. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2004–2013, New York, New York, USA, June 2016. PMLR. URL https://proceedings.mlr.press/v48/kusano16.html.

Peter Lawson, Andrew B Sholl, J Quincy Brown, Brittany Terese Fasy, and Carola Wenk. Persistent Homology for the Quantitative Evaluation of Architectural Features in Prostate Cancer Histology. *Scientific Reports*, 9(1):1139, 2019. ISSN 2045-2322. URL https://doi.org/10.1038/s41598-018-36798-y.

Olivier Ledoit Wolf, Michael. Honey, I Shrunk the Sample Covariance Matrix, 2004. URL https://www.pm-research.com/content/iijpormgmt/30/4/110.

Jonggul Lee, Jungho Shin, Tae-Wook Ko, Seunghee Lee, Hyunju Chang, and YunKyong Hyon. Descriptors of Atoms and Structure Information for Predicting Properties of Crystalline Materials. *Materials Research Express*, 8(2), 2021.

Yongjin Lee, Senja D Barthel, Paweł Dłotko, Seyed Mohamad Moosavi, Kathryn Hess, and Berend Smit. High-Throughput Screening Approach for Nanoporous Materials Genome Using Topological Data Analysis: Application to Zeolites. *Journal of Chemical Theory and Computation*, 14(8):4427–4437, August 2018. ISSN 1549-9618. URL https://doi.org/10.1021/acs.jctc.8b00253.

Charles J Stone R A Olshen Leo Breiman Jerome Friedman. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

Qinyang Li, Rongzhi Dong, Nihang Fu, Sadman Sadeed Omee, Lai Wei, and Jianjun Hu. Global Mapping of Structures and Properties of Crystal Materials. *Journal of Chemical Information and Modeling*, 63(12):3814–3826, June 2023a. ISSN 1549-9596. URL https://doi.org/10.1021/acs.jcim.3c00224.

Shunning Li, Yuanji Liu, Dong Chen, Yi Jiang, Zhiwei Nie, and Feng Pan. Encoding the Atomic Structure for Machine Learning in Materials Science. *WIREs Computational Molecular Science*, 12(1), January 2022. ISSN 1759-0876. URL https://doi.org/10.1002/wcms.1558.

Xiang-Yang Li, Gruia Calinescu, Peng-Jun Wan, and Yu Wang. Localized Delaunay Triangulation with Application in Ad Hoc Wireless Networks. *IEEE Transactions on Parallel and Distributed Systems*, 14(10):1035–1047, 2003.

Zian Li, Xiyuan Wang, Yinan Huang, and Muhan Zhang. Is Distance Matrix Enough for Geometric Deep Learning?, 2023b.

Jörg Liebeherr and Michael Nahas. Application-layer Multicast with Delaunay Triangulations. In *GLOBECOM'01. IEEE Global Telecommunications Conference (Cat. No. 01CH37270)*, volume 3, pages 1651–1655. IEEE, 2001.

S. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

Leigh Loots and Leonard J Barbour. A Simple and Robust Method for the Identification of π–π Packing Motifs of Aromatic Compounds. *CrystEngComm*, 14 (1):300–304, 2012. URL http://dx.doi.org/10.1039/C1CE05763D.

Donald Loveland, Bhavya Kailkhura, Piyush Karande, Anna M Hiszpanski, and T Yong-Jin Han. Automated Identification of Molecular Crystals' Packing Motifs. *Journal of Chemical Information and Modeling*, 60(12):6147–6154, December 2020. ISSN 1549-9596. URL https://doi.org/10.1021/acs.jcim.0c01134.

Mario Lovrić, Tomislav čić, Han T N Tran, Hussain Hussain, Emanuel Lacić, Morten A Rasmussen, and Roman Kern. Should we Embed in Chemistry? A Comparison of Unsupervised Transfer Learning with PCA, UMAP, and VAE on Molecular Fingerprints. *Pharmaceuticals*, 14(8):758, 2021.

Stipe Lukin, Tomislav Stolar, Martina Tireli, Maria Valeria Blanco, Darko Babić, Tomislav Friščić, Krunoslav Užarević, and Ivan Halasz. Tandem In Situ Monitoring for Quantitative Assessment of Mechanochemical Reactions Involving Structurally Unknown Phases. *Chemistry – A European Journal*, 23(56):13941–13949, October 2017. ISSN 0947-6539. URL https://doi.org/10.1002/chem.201702489.

Clément Maria, Jean-Daniel Boissonnat, Marc Glisse, and Mariette Yvinec. The Gudhi Library: Simplicial Complexes and Persistent Homology. In Hoon Hong and Chee Yap, editors, *Mathematical Software – ICMS 2014*, pages 167–174, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg. ISBN 978-3-662-44199-2.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861, 2018. URL https://doi.org/10.21105/joss.00861.

Djuradj Milošević, Andrew S Medeiros, Milica Stojković Piperac, Dušanka Cvijanović, Janne Soininen, Aleksandar Milosavljević, and Bratislav Predić. The Application of Uniform Manifold Approximation and Projection (UMAP) for Unconstrained Ordination and Classification of Biological Indicators in Aquatic Ecology. *Science of The Total Environment*, 815, 2022.

Emi Minamitani, Ippei Obayashi, Koji Shimizu, and Satoshi Watanabe. Persistent Homology-Based Descriptor for Machine-Learning Potential of Amorphous Structures. *The Journal of Chemical Physics*, 159(8), 2023.

Joshua Mirth, Yanqin Zhai, Johnathan Bush, Enrique G Alvarado, Howie Jordan, Mark Heim, Bala Krishnamoorthy, Markus Pflaum, Aurora Clark, Y Z, and Henry Adams. Representations of Energy Landscapes by Sublevelset Persistent Homology: An Example With n-Alkanes, 2020.

Félix Musil, Sandip De, Jack Yang, Joshua E Campbell, Graeme M Day, and Michele Ceriotti. Machine Learning for the Structure–Energy–Property Landscapes of Molecular Crystals. *Chemical Science*, 9(5):1289–1300, 2018. ISSN 2041-6520. URL `http://dx.doi.org/10.1039/C7SC04665K`.

Felix Musil, Andrea Grisafi, Albert P Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti. Physics-Inspired Structural Representations for Molecules and Materials. *Chemical Reviews*, 121(16):9759–9815, August 2021. ISSN 0009-2665. URL `https://doi.org/10.1021/acs.chemrev.1c00021`.

Takenobu Nakamura, Yasuaki Hiraoka, Akihiko Hirata, Emerson G Escolar, and Yasumasa Nishiura. Persistent Homology and Many-Body Atomic Structure for Medium-Range Order in the Glass. *Nanotechnology*, 26(30), 2015. ISSN 0957-4484. URL `http://dx.doi.org/10.1088/0957-4484/26/30/304001`.

Ippei Obayashi. Volume-Optimal Cycle: Tightest Representative Cycle of a Generator in Persistent Homology. *SIAM Journal on Applied Algebra and Geometry*, 2(4):508–534, January 2018. URL `https://doi.org/10.1137/17M1159439`.

Markus Ojala and Gemma C. Garriga. Permutation Tests for Studying Classifier Performance. *Journal of Machine Learning Research*, 11:1833–1863, August 2010. ISSN 1532-4435.

Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A Roadmap for the Computation of Persistent Homology. *EPJ Data Science*, 6(1):17, 2017. ISSN 2193-1127. URL `https://doi.org/10.1140/epjds/s13688-017-0109-5`.

Linus Pauling. The Nature of the Chemical Bond. *Philosophy of Science*, 8(1):133–133, 1941.

Mariam Pirashvili, Lee Steinberg, Francisco Belchi Guillamon, Mahesan Niranjan, Jeremy G Frey, and Jacek Brodzki. Improved Understanding of Aqueous Solubility Modeling Through Topological Data Analysis. *Journal of Cheminformatics*, 10(1):54, 2018. ISSN 1758-2946. URL `https://doi.org/10.1186/s13321-018-0308-5`.

Sarah L Price. From Crystal Structure Prediction to Polymorph Prediction: Interpreting the Crystal Energy Landscape. *Physical Chemistry Chemical Physics*, 10 (15):1996–2009, 2008.

Sarah L Price, Maurice Leslie, Gareth W A Welch, Matthew Habgood, Louise S Price, Panagiotis G Karamertzanis, and Graeme M Day. Modelling Organic Crystal Structures Using Distributed Multipole and Polarizability-Based Model Intermolecular Potentials. *Physical Chemistry Chemical Physics*, 12(30):8478–8490, 2010. URL http://dx.doi.org/10.1039/C004164E.

Sarah (Sally) L Price. Computed Crystal Energy Landscapes for Understanding and Predicting Organic Crystal Structures and Polymorphism. *Accounts of Chemical Research*, 42(1):117–126, 2009.

Edward O Pyzer-Knapp, Hugh P G Thompson, and Graeme M Day. An Optimized Intermolecular Force Field for Hydrogen-Bonded Organic Molecular Crystals Using Atomic Multipole Electrostatics. *Acta Crystallographica Section B*, 72(4):477–487, August 2016. URL https://doi.org/10.1107/S2052520616007708.

Milan Randić and Matevž Pompe. The Variable Molecular Descriptors Based on Distance Related Matrices. *Journal of Chemical Information and Computer Sciences*, 41 (3):575–581, May 2001. ISSN 0095-2338. URL https://doi.org/10.1021/ci0001029.

Murray Rosenblatt. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832–837, September 1956. URL https://doi.org/10.1214/aoms/1177728190.

Marylène Rugard, Thomas Jaylet, Olivier Taboureau, Anne Tromelin, and Karine Audouze. Smell Compounds Classification Using UMAP to Increase Knowledge of Odors and Molecular Structures Linkages. *PloS one*, 16(5), 2021.

Ingo Salzmann. *Structural and Energetic Properties of Pentacene Derivatives and Heterostructures 2004-07-01 - 2008-12-21*. PhD thesis, Humboldt University, Faculty of Mathematical Sciences I, September 2020.

Johann Christian Schön and Martin Jansen. First Step Towards Planning of Syntheses in Solid-State Chemistry: Determination of Promising Structure Candidates by Global Optimization. *Angewandte Chemie*, 35:1286–1304, 1996. URL https://api.semanticscholar.org/CorpusID:93750381.

Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Trans. Database Systems*, 42(3), July 2017. ISSN 0362-5915. URL https://doi.org/10.1145/3068335.

C E Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, July 1948. ISSN 0005-8580. URL https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Alok Sharma and Kuldip K Paliwal. Linear Discriminant Analysis for the Small Sample Size Problem: an Overview. *International Journal of Machine Learning and Cybernetics*, 6:443–454, 2015. URL https://api.semanticscholar.org/CorpusID:13858274.

Donald R Sheehy. Linear-Size Approximations to the Vietoris-Rips Filtration. *Discrete Computational Geometry*, 49:778–796, 2013. URL https://doi.org/10.1007/s00454-013-9513-1.

Gurjeet Kaur Chatar Singh, Facundo Mémoli, and Gunnar E. Carlsson. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In *PBG@Eurographics*, 2007.

Henning Sirringhaus, Nir Tessler, and Richard H Friend. Integrated Optoelectronic Devices Based on Conjugated Polymers. *Science*, 280(5370):1741–1744, June 1998. URL https://doi.org/10.1126/science.280.5370.1741.

Alex J Smola and Bernhard Schölkopf. A Tutorial on Support Vector Regression. *Statistics and Computing*, 14(3):199–222, 2004. ISSN 1573-1375. URL https://doi.org/10.1023/B:STCO.0000035301.49549.88.

M S Smyth and J H Martin. X-Ray Crystallography. *Molecular Pathology : MP*, 53(1): 8–14, Febuary 2000. ISSN 1366-8714 (Print).

Lee Steinberg. *Topological Data Analysis and its Application to Chemical Systems*. PhD thesis, University of Southampton, 2019.

Lee Steinberg, John Russo, and Jeremy Frey. A New Topological Descriptor for Water Network Structure. *Journal of Cheminformatics*, 11(1):48, 2019. ISSN 1758-2946. URL https://doi.org/10.1186/s13321-019-0369-0.

A J Stone. *The Theory of Intermolecular Forces*. The Theory of Intermolecular Forces. OUP Oxford, 2013. ISBN 9780199672394. URL https://books.google.co.uk/books?id=dfMpYkacvy8C.

A J Stone and M Alderton. Distributed Multipole Analysis Methods and Applications. *Molecular Physics*, 100(1):221–233, January 2002. ISSN 0026-8976. URL https://doi.org/10.1080/00268970110089432.

Shota Takemura, Takashi Takeda, Takayuki Nakanishi, Yukinori Koyama, Hidekazu Ikeno, and Naoto Hirosaki. Dissimilarity Measure of Local Structure in Inorganic Crystals Using Wasserstein Distance to Search for Novel Phosphors. *Science and Technology of Advanced Materials*, 22(1):185–193, April 2021. ISSN 1468-6996 (Print).

Robin Taylor and Clare F Macrae. Rules Governing the Crystal Packing of Mono-and Dialcohols. *Acta Crystallographica Section B: Structural Science*, 57(6):815–827, 2001.

Robin Taylor and Peter A Wood. A Million Crystal Structures: The Whole Is Greater than the Sum of Its Parts. *Chemical Reviews*, 119(16):9427–9477, August 2019. ISSN 0009-2665. URL https://doi.org/10.1021/acs.chemrev.9b00155.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000. URL https://www.science.org/doi/abs/10.1126/science.290.5500.2319.

Igor V Tetko and Ola Engkvist. From Big Data to Artificial Intelligence: Chemoinformatics Meets New Challenges. *Journal of Cheminformatics*, 12(1):74, 2020. ISSN 1758-2946. URL https://doi.org/10.1186/s13321-020-00475-y.

Michael E. Tipping and Christopher M. Bishop. Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation*, 11(2):443–482, Febuary 1999. ISSN 0899-7667. URL https://doi.org/10.1162/089976699300016728.

Chad Topaz, Lori Ziegelmeier, and Tom Halverson. Topological Data Analysis of Biological Aggregation Models. *PloS one*, 10, December 2014.

Francesco Trozzi, Xinlei Wang, and Peng Tao. UMAP as a Dimensionality Reduction Tool for Molecular Dynamics Simulations of Biomacromolecules: a Comparison Study. *The Journal of Physical Chemistry B*, 125(19):5022–5034, 2021.

Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet Means for Distributions of Persistence Diagrams. *Discrete and Computational Geometry*, 52 (1):44–70, 2014. ISSN 1432-0444. URL https://doi.org/10.1007/s00454-014-9604-7.

UMAP_learn_documentation. UMAP for Supervised Dimension Reduction and Metric Learning, acessed 2024. URL https://umap-learn.readthedocs.io/en/latest/supervised.html.

Yuhei Umeda. Time Series Classification via Topological Data Analysis. *Transactions of the Japanese Society for Artificial Intelligence*, 32(3), 2017.

Edward F Valeev, Veaceslav Coropceanu, Demetrio A da Silva Filho, Seyhan Salman, and Jean-Luc Brédas. Effect of Electronic Polarization on Charge-Transport Parameters in Molecular Organic Semiconductors. *Journal of the American Chemical Society*, 128(30):9882–9886, August 2006. ISSN 0002-7863. URL https://doi.org/10.1021/ja061827h.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Marc Vermeulen, Kate Smith, Katherine Eremin, Georgina Rayner, and Marc Walton. Application of Uniform Manifold Approximation and Projection (UMAP) in Spectral Imaging of Artworks. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 252:119547, 2021.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 11(95):2837–2854, 2010. URL `http://jmlr.org/papers/v11/vinh10a.html`.

Oliver Vipond, Joshua A Bull, Philip S Macklin, Ulrike Tillmann, Christopher W Pugh, Helen M Byrne, and Heather A Harrington. Multiparameter Persistent Homology Landscapes Identify Immune Cell Spatial Patterns in Tumors. *Proceedings of the National Academy of Sciences*, 118(41), 2021.

Ulrike von Luxburg. A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4): 395–416, 2007. ISSN 1573-1375. URL `https://doi.org/10.1007/s11222-007-9033-z`.

Hubert Wagner, Chao Chen, and Erald Vuçini. Efficient Computation of Persistent Homology for Cubical Data. In *Topological Methods in Data Analysis and Visualization II*, pages 91–106, 2012.

Chengliang Wang, Huanli Dong, Wenping Hu, Yunqi Liu, and Daoben Zhu. Semiconducting π-Conjugated Systems in Field-Effect Transistors: A Material Odyssey of Organic Electronics. *Chemical Reviews*, 112(4):2208–2267, April 2012. ISSN 0009-2665. URL `https://doi.org/10.1021/cr100380z`.

Yanxing Wang, Théo Jaffrelot Inizan, Chengwen Liu, Jean-Philip Piquemal, and Pengyu Ren. Incorporating Neural Networks into the AMOEBA Polarizable Force Field. *The Journal of Physical Chemistry B*, 128(10):2381–2388, March 2024. ISSN 1520-5207 (Electronic).

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data Visualization. *Journal of Machine Learning. Research*, 22, 2020. URL `https://api.semanticscholar.org/CorpusID:227745109`.

Węglarczyk, Stanisław. Kernel Density Estimation and its Application. *ITM Web Conference*, 23:37, 2018. URL `https://doi.org/10.1051/itmconf/20182300037`.

Nigel P Weatherill. Delaunay Triangulation in Computational Fluid Dynamics. *Computers & Mathematics with Applications*, 24(5-6):129–150, 1992.

Daniel Widdowson, Marco M Mosca, Angeles Pulido, Vitaliy Kurlin, and Andrew I Cooper. Average Minimum Distances of Periodic Point Sets - Foundational Invariants for Mapping Periodic Crystals. *MATCH Communications in Mathematical and in Computer Chemistry*, 87(3):529–559, 2022.

Donald E. Williams. Improved Intermolecular Force Field for Crystalline Oxohydrocarbons Including O - H. . .O Hydrogen Bonding. *Journal of Computational Chemistry*, 22(1):1–20, January 2001. ISSN 01928651. URL https://doi.org/10.1002/1096-987X(20010115)22:1%3C1::AID-JCC2%3E3.0.COhttp://2-6.

D J Willock, S L Price, M Leslie, and C R A Catlow. The Relaxation of Molecular Crystal Structures Using a Distributed Multipole Electrostatic Model. *Journal of Computational Chemistry*, 16(5):628–647, May 1995. ISSN 0192-8651. URL https://doi.org/10.1002/jcc.540160511.

Michael Winkler and K N Houk. Nitrogen-Rich Oligoacenes: Candidates for n-Channel Organic Semiconductors. *Journal of the American Chemical Society*, 129(6): 1805–1815, Febuary 2007. ISSN 0002-7863. URL https://doi.org/10.1021/ja067087u.

Kelin Xia, Zhixiong Zhao, and Guo-Wei Wei. Multiresolution Persistent Homology for Excessively Large Biomolecular Datasets. *The Journal of Chemical Physics*, 143(13), October 2015. ISSN 0021-9606. URL https://aip.scitation.org/doi/abs/10.1063/1.4931733.

Shiyue Yang and Graeme M Day. Global Analysis of the Energy Landscapes of Molecular Crystal Structures by Applying the Threshold Algorithm. *Communications Chemistry*, 5(1):86, 2022. ISSN 2399-3669. URL https://doi.org/10.1038/s42004-022-00705-4.

Qiang Zhu, Weilun Tang, and Shinnosuke Hattori. Quantification of Crystal Packing Similarity from Spherical Harmonic Transform. *Crystal Growth Design*, 22(12): 7308–7316, December 2022. ISSN 1528-7483. URL https://doi.org/10.1021/acs.cgd.2c00933.

Afra Zomorodian and Gunnar Carlsson. Computing Persistent Homology. *Discrete and Computational Geometry*, 33(2):249–274, 2005. ISSN 1432-0444. URL https://doi.org/10.1007/s00454-004-1146-y.