UNIVERSITY OF SOUTHAMPTON
Faculty of Engineering and Physical Sciences
SCHOOL OF CHEMISTRY

Topological Data Analysis and its Application to Chemical Systems

by

Lee Steinberg

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

October 2019

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

Faculty of Engineering and Physical Sciences
SCHOOL OF CHEMISTRY

Doctor of Philosophy

by Lee Steinberg

Topological data analysis techniques are applied to distinct problems in chemistry, to determine their efficacy and gain new understanding of chemical systems. The mapper algorithm is utilised to understand the underlying descriptor space of a solubility prediction data set. Insight from the resulting topological summaries was able to create more consistent solubility models. Persistent homology is then used to create a series of metric spaces for molecular shape. It is shown that these metric spaces correlate with other molecular descriptors, and also allow for the accounting of molecular flexibility.

This molecular flexibility is further explored with persistent homology. By constructing a point cloud of individual conformers, a technique to characterise the conformational spaces of various molecules is developed. Alanine dipeptide is shown to have a toroidal conformational space, and persistence is then used to locate extrema on its torsional free energy surface. Pentane is then studied, and shown to also have a toroidal conformational space, or a Möbius band when symmetry is taken into account. The conformational space of cyclooctane is shown to be non-manifold, and the separate manifold components separated. It is found that there are separate spherical and Klein bottle components, before the single point energy landscape of the sphere is also analysed and extrema located.

Finally, simulated water networks are analysed through persistent homology. The general use of persistence to analyse simulations is studied, and persistence is shown to be a well-behaved descriptor. A size-agnostic persistence descriptor is generated, and used with a support vector machine to understand the differences in simulated water networks. Atomistic and coarse-grained water potentials are compared, and similarities between potentials are related to topological features.

Contents

List of Figures

List of Tables

Acknowledgements

Although it only has one name on the title page, a PhD is by no means the work of an individual. Throughout my PhD, and for many years before, I have been helped by countless people. This work is theirs as much as mine.

Firstly, I thank my supervisor, Professor Jeremy Frey. Since we met in Easter 2016, and I turned up saying I wanted to do a different project to the one which we had originally agreed, he has supported me, and done whatever he could to facilitate my work. His ability to walk into a room filled with people studying totally different topics, and give each one the insight they needed, is unmatched.

I also thank my collaborators. Professor Jacek Brodzki, for letting a chemist enter the world of mathematicians. Ingrid and Mariam, I hope we have done some high-quality work together. I also have to thank Kiko, for answering all of my mathematical questions, but mostly for being a friend to me these past few years. Thanks to everyone I worked with at GlaxoSmithKline in Stevenage, and in particular Dr David Marcus for making my time so enjoyable - even if I spent 5 hours each day on trains during the hottest summer in memory.

Thanks go to the rest of the Frey group office over the years. A special mention goes to Frank and Sam - I couldn't have asked for better colleagues. Also, the various project students we have had, for all of their questions and support. I want to name Rahma and Dani specifically - they will know why.

I thank the Theory and Modelling in Chemical Sciences Centre for Doctoral Training. Thanks to the rest of my cohort specifically and to St. Aldate's crew, and those I shared a kitchen with that year in Oxford.

Thanks to all of the friends I have made in Southampton, both inside and outside the department. Thank you to the Berkeley Boys, to Jon, David, and the Toms (both Classic and Crystal), for movie nights, and for evenings spent at the Brewhouse and the Bookshop. They will not be forgotten.

Thank you to my family, to my parents and grandparents, for supporting me and pretending they have vague interest in chemistry and data science. Thank you to both my brothers. To Elliot, for reminding me that their are other things to life than science, and to Adam, for reminding me of the exact opposite.

I thank the EPSRC for funding, and the University of Southampton for resources. I also want to thank Khaled Abdel Maksoud and Shawn Martin for supplying data that appears in this thesis. Lastly I thank all of those who I forgot to directly mention here. I am sure there are a lot of you, but I could not have done this work without the help you gave me along the way.

For Mum and Dad

Nomenclature

General Spaces

$\mathbb{F}$	An arbitrary field
I	The unit interval $[0, 1]$
R	An arbitrary ring
$\mathbb{R}^n$	An n-tuple of real numbers
$\mathbb{R}^+$	The non-negative real numbers
$\mathbb{R}P^2$	The real projective plane
S^n	The topological n-sphere.
	The set of points $x \in \mathbb{R}^{n+1}$ where $\|\|x\|\| = 1$
T^n	The topological n-torus, equivalent to be the set of n-tuples
	where each component is a point on the circle S^1.
	T^2 is topologically equivalent to the surface of a donut
$\mathcal{V}$	An arbitrary vector space
$\mathbb{Z}$	The integers: $\{0, 1, 2, \ldots\}$
$\mathbb{Z}_p$	The integers modulo p: $\{0, 1, 2, \ldots, p\}$

Algebraic Topology

β_n	The n^{th} Betti number
C_p	The vector space generated by p-simplices
d	A metric function
∂_p	The boundary map from C_p to C_{p-1}
δ	The persistence parameter
H_p	The p^{th} homology group
im	The image of a map. The set of all output values of the map
K	A simplicial complex
ker	The kernel of a map.
	The elements of the domain which map to zero
Λ	A persistence landscape
λ	A single persistence landscape function
PD	A persistence diagram
σ	An arbitrary k-simplex, made of $k + 1$ vertices
v	A vertex

Conformational Spaces

$\mathcal{C}$	A conformer
$\mathcal{G}$	A molecule (as defined by a molecular graph)
$\mathcal{M}$	A conformational space
$SO(3)$	The group of all rotations about the origin in 3-dimensions

Water Networks

$C_{PD}(t)$	The pseudo-autocorrelation of a persistence diagram through a simulation
N_O	The number of oxygen atoms (water molecules) in a simulation
L1NPI	An l_1-normalised persistence image
λ	The parameter defining the relative strength of the three body term of the Stillinger-Weber potential

Chapter 1

Introduction

1.1 Chemistry as a Data Driven Science

Chemistry has *always* been a data driven science. The year of writing of this thesis (2019) is the 150^{th} anniversary of the publication of Mendeleev's periodic table (Figure 1.1), and his work was a triumph in data science. Mendeleev was not the first to notice that the elements showed a periodicity when arranged in mass order, but he also speculated that this property was fundamental in nature. He argued that discrepancies between predicted and actual behaviour were due to experimental measurement errors, or the fact that certain elements were undiscovered [1]. These predictions were later proven to be correct, with the discovery of the elements scandium, gallium, and germanium.

Reihen	Gruppo I. — R^1O	Gruppo II. — RO	Gruppo III. — R^2O^3	Gruppo IV. RH^4 RO^2	Gruppo V. RH^3 R^2O^5	Gruppo VI. RH^2 RO^3	Gruppo VII. RH R^2O^7	Gruppo VIII. — RO^4
1	H=1							
2	Li=7	Be=9,4	B=11	C=12	N=14	O=16	F=19	
3	Na=23	Mg=24	Al=27,3	Si=28	P=31	S=32	Cl=35,5	
4	K=39	Ca=40	—=44	Ti=48	V=51	Cr=52	Mn=55	Fe=56, Co=59, Ni=59, Cu=63.
5	(Cu=63)	Zn=65	—=68	—=72	As=75	Se=78	Br=80	
6	Rb=86	Sr=87	?Yt=88	Zr=90	Nb=94	Mo=96	—=100	Ru=104, Rh=104, Pd=106, Ag=108.
7	(Ag=108)	Cd=112	In=113	Sn=118	Sb=122	Te=125	J=127	
8	Cs=133	Ba=137	?Di=138	?Ce=140	—	—	—	— — — —
9	(—)	—	—	—	—	—	—	
10	—	—	?Er=178	?La=180	Ta=182	W=184	—	Os=195, Ir=197, Pt=198, Au=199.
11	(Au=199)	Hg=200	Tl=204	Pb=207	Bi=208	—	—	
12	—	—	—	Th=231	—	U=240	—	— — — —

FIGURE 1.1: Mendeleev's 1871 published periodic table. His first table was published in 1869, but the 1871 version predicted the existence of scandium, gallium and germanium, as evidenced by the gaps present.

Obviously, the field has greatly changed since 1871 - although as of 2019 there is still no officially recommended IUPAC periodic table [2]. However, chemistry, like all sciences, has at its heart the notion of gaining insight from data. Furthermore, recent years have seen an explosion within the field of 'big data', and chemistry has certainly benefited

from this. This thesis seeks to make a dent within the field, in particular exploring the use of relatively new mathematical techniques, known collectively as topological data analysis, and their application to a series of chemical problems.

1.2 Big Data, Big Chemistry

There has been a global explosion of data, and as mentioned chemistry has not been immune to this. Chemical databases have become a useful tool for chemists of all types, allowing them to easily access information of all kinds [3]. Such databases are naturally different chemical spaces, which can then be endowed with mathematical properties, and studies can be performed. For example, when designing a new drug to target a particular protein it is reasonable to begin the search with the molecules that are known to act against this target, and use a notion of 'nearness', or similarity, to determine which molecules should be tested. Thankfully, the advent of open chemical databases has enabled researchers to more easily access this information through the internet. Rather than provide an exhaustive list of these databases (which will no doubt be out of date by the time this is being read), instead a brief introduction to the following list of databases will be discussed:

- The Protein Data Bank
- The Cambridge Structural Database
- PubChem
- ChEMBL
- ZINC
- GDB

The Protein Data Bank (PDB) [4] is a database for large biomolecules, containing their three-dimensional structures. Established in 1971, the PDB now contains approximately 150,000 depositions, often found using X-ray crystallography or NMR spectroscopy. The PDB is often used to find starting structures for molecular dynamics simulations of complex protein systems.

The Cambridge Structural Database (CSD) [5] is a database for small-molecule crystal structures. In mid 2019, the CSD had its millionth deposition, since its establishment in 1965. The database has various uses in computational chemistry, for example screening favourable interactions to determine if a potential new molecule has a similar crystal structure.

PubChem [6] is a large multi-purpose chemical database. It can be considered as a series of smaller databases, with three of these considered primary: Compounds, Substances and BioAssay. Compounds contains approximately 100,000 pure chemical compounds and their physical properties. Substances is a database of approximately 200,000 mixtures, complexes, and uncharacterised substances. BioAssay data is the output of high-throughput screening programs, containing over 250,000 bioactivity endpoints. The sheer variety and volume of PubChem data leads to its wide use in computational chemistry.

ChEMBL is a database of assay data [7]. ChEMBL is manually curated, containing approximately two million compounds in 1,100,000 assays. ChEMBL is often used in the development of screening libraries in lead identification, or as an input for molecule generation algorithms.

The ZINC database [8] is a database of commercially available compounds. This is useful for virtual screening, as it can be guaranteed that the desired compound can be purchased from standard vendors. However, as all of ZINC is commercially available, it can lead to issues with novelty when designing new therapeutics or agricultural compounds.

The GDB is a series of databases of small molecules. The GDB is a database created by the group of Reymond at the University of Bern, Switzerland [9], and in contrast to the other databases discussed, can be considered as a graph enumeration database. For example, the GDB-17 contains all organic molecules up to 17 atoms of C, N, O and F. This contains approximately 50,000,000 molecules. Such data sets are useful when designing algorithms to efficiently search 'chemical space'. Furthermore, data sets such as the GDB are useful to be sampled. For example, the QM9 data set [10] is a subset of the GDB-17 data set. This data set contains geometric, electronic and thermodynamic properties of 134,000 molecules, which could be used for benchmarking methods of predicting properties of small organic molecules.

1.3 Traditional Statistical Techniques and their Breakdown

The sheer size of chemical databases, coupled with the exponential increase in computing power, has led to an explosion in the use of data science and statistical techniques within chemistry. Broadly, a data science project can be split into three main themes:

1. Data Collection & Preparation
2. Model Building
3. Prediction

In all three of these themes, statistical techniques are important. When preparing data, statistical methods are used to impute missing values. Machine learning models are all inherently statistical in nature, and therefore the construction of useful models requires an understanding of their underlying statistical assumptions. Lastly, assessing the performance of model prediction is often performed using statistical metrics and heuristics.

However, focusing on the model building aspect in more detail, a problem becomes apparent. In particular, construction of a useful and efficient model requires an inherent understanding of the underlying structure of the data set - and a misunderstanding of this can make constructed models worthless.

This problem is perhaps most famously illustrated with Anscombe's quartet [11] or the more recent 'Datasaurus Dozen' [12] These data sets are pathological, designed to illustrate the flaws in simply calculating summary statistics without first visualising the data. The Datasaurus Dozen can be seen in Figure 1.2. Various statistics are identical with all of the data sets, even though it is clear that the data sets themselves vastly differ. For example, the Pearson coefficient r is identical, and small (0.07). However, it would be foolish to state there are no relationships between the dimensions of the data sets.

For each data set within the dozen, different models would find use for a hypothetical prediction. For some data sets, it would be possible to transform the data onto a straight line. For others, a clustering algorithm may be more useful. Furthermore, some data sets could be analysed using a combination of clustering and regression. However, this insight would not be possible without first visualising the data.

However, visualising the data set is only half of the story. Most data is inherently high dimensional, and the manifold hypothesis states that a high-dimensional data set tends to lie in the vicinity of a low-dimensional manifold [13]. Understanding the data therefore requires some sort of dimensionality reduction, or projection. This projection will in general lead to a loss of information from the original data (although this is not guaranteed). An example of this can be seen in Figure 1.3, where a torus is projected onto two dimensions such that it looks like an annulus.

The problems of projection and visualisation directly oppose each other. On the one hand, to create useful data models, the underlying structure of the data set should be understood, and visualisation aids this. On the other, visualisation of high dimensional data can lead to incorrect conclusions being made regarding the data's structures. Clearly, there is a need to develop methods that work in high dimension, that assist in the elucidation of data structure information. Recently, the mathematical field of topology has found use in this problem.

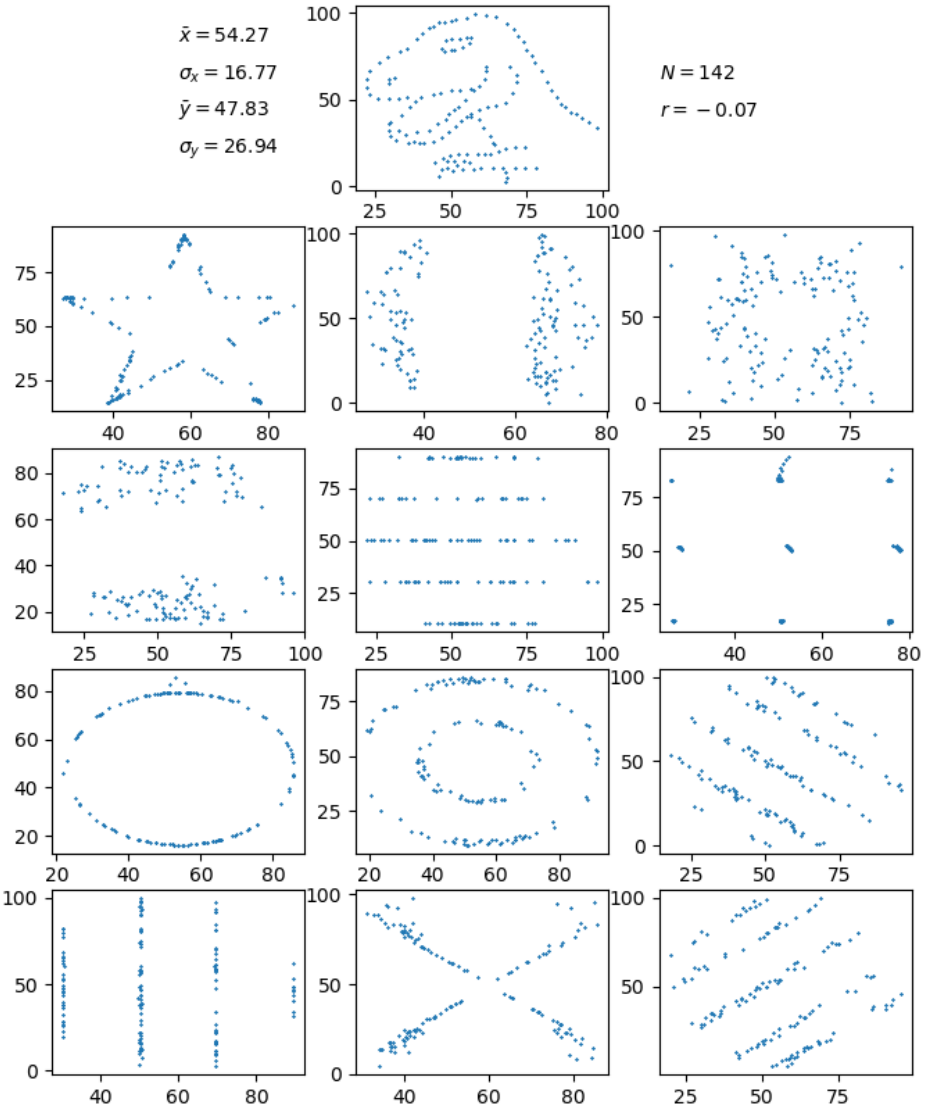


FIGURE 1.2: The 'Datasaurus Dozen' series of data sets. σ_x, σ_y, $\bar{x}$, $\bar{y}$, and the Pearson coefficient r are identical (to two decimal places) between all of the data sets, although they clearly have markedly different relationships.

1.4 Topology: A Brief Introduction

At its heart, topology is about what it means to be 'connected' - in a mathematical sense. Historically, topology has always been viewed as a particularly pure subject, with very few direct applications. Although one of the first published papers in the field was

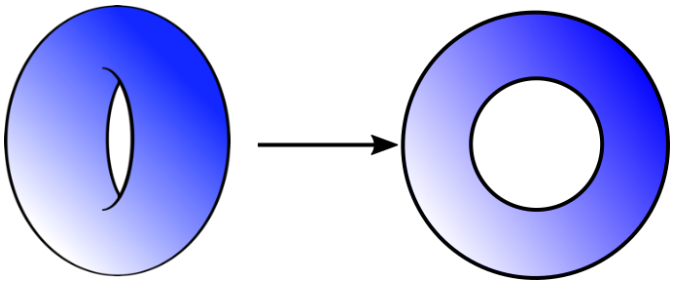


FIGURE 1.3: Illustration of the projection process applied to a torus. The torus (left) is projected to look like an annulus (right). This leads to a loss of information regarding the topology of the original object.

Euler's 1736 work on the Königsberg seven bridges problem, toppology has since evolved to be less of an applied mathematical field.

Topology studies the properties of (often geometric) objects that do not change through continuous deformations. Depending on which classes of deformations are allowed, different topological equivalences are defined. The most famous of topological equivalences is the 'homeomorphism':

> A homeomorphism between two topological spaces X and Y is a continuous function that maps each point in X to a distinct point in Y (it is injective), the entirety of Y is mapped onto (it is surjective), and whose inverse is also continuous. If such a function exists, the two spaces X and Y are said to be homeomorphic.

The homeomorphism is a flexible equivalence relation. Within a homeomorphism, all objects are constructed of the world's most malleable dough - and the only things that are disallowed are gluing and tearing. It should be obvious that all convex polyhedra (such as the Platonic solids) are homeomorphic to the sphere S^2. Another famous pair of homeomorphic spaces are the torus and coffee cup - the handle of the cup is the ring of the torus. However, some spaces have less intuitive homeomorphisms, for example it can be shown that a homeomorphism exists between S^n with a single point removed and $\mathbb{R}^n$. For example, $S^1/\{*\}$ is homeomorphic to $\mathbb{R}$ through the trigonometric tangent function. Homeomorphic spaces have identical topological invariants. There are innumerate topological invariants, but a few important ones are listed below:

- Cardinality: The number of elements of a space
- Betti numbers β_n: The number of n-dimensional holes in a space
  - Related to β_n is the Euler number χ, the alternating sum of Betti numbers. For convex polyhedra, this is often written as $\chi = n_{vertices} - n_{edges} + n_{faces} = 2$

- Path-connected: A space is path connected if any two points can be connected by a path

Discussing objects through their topology rather than geometric properties is actually an everyday occurence. A common set of examples are maps of rail networks, such as the London Underground or Paris Metro. Here, the relative distance between stations is not relevant - if you want to use the rail network you simply want to know how to get to your destination. This is a topological property - it is how the network is connected. This work uses a modern branch of topology known as topological data analysis to understand how various chemical objects are connected: chemical shape space, conformational spaces, and water networks.

1.4.1 Topological Data Analysis

Chapter 2 contains the theory and mathematics behind topological data analysis. Here a general introduction regarding motivation is presented.

Before the topology of data can be discussed, it is important to first comment on the data collection process. Figure 1.4 shows an idealised cartoon of data collection. In this example, the data being sampled exists on a two-dimensional manifold, with a single hole. A series of data points is collected, the lower half of the image. From the sampled

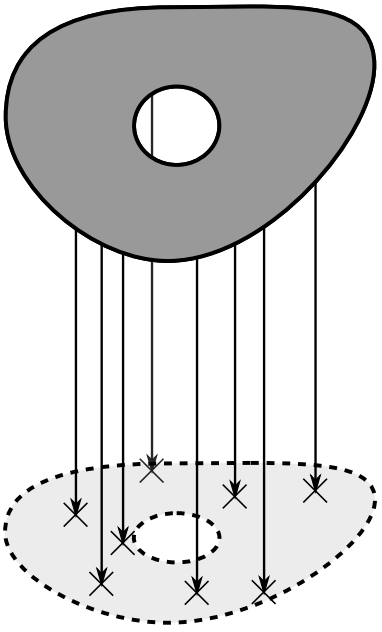


FIGURE 1.4: A cartoon illustrating the data collection process. Individual points are sampled from some high dimensional topological structure.

points, is it possible to understand the structure of the original data set? For example, could one conclude that there is a hole in the data? If it were possible to sample an

infinite number of points, the answer is yes - an infinite sample of points would recreate the topology of the original manifold exactly, provided the dimensionality was high enough. The sampling of an infinite number of points is clearly impossible. However, without an infinite number of points, there is no way to say for certain what the original data structure was.

In fact, the situation is even more bleak. Without the limit of infinite sampling, there is a potentially infinite number of points between any two sampled points, and each sampled point is isolated. To be able to discuss the presence of a hole, it is necessary to create some higher structure on the data set, to join it up in some way. In particular, the joining up method shouldn't fill in the hole, but give it the correct boundary.

Topological data analysis deals with the question: how can the data be joined up, and what can be learned from it? The two main threads of topological data analysis differ in their methods of joining the data. Persistent homology [14] constructs a series of topological structures on the data set, and seeks to understand which topological features are common. Mapper [15] creates a single topological space, by clustering various local sets of data points.

1.5 Topology and Chemistry

Ideas in topology have found use in chemistry through history. For example, Hückel theory, used to calculate the molecular orbitals of π systems, reduces to calculating adjacency matrices and other graph theory properties - and graph theory is certainly a cousin of topology, even if it is its own rich area of study.

Using this definition, any chemical descriptor derived from a skeletal formula, or molecular graph, can be described as a topological descriptor of the molecule. However, when referring to a topological descriptor, chemical informaticians often mean properties such as the Wiener index. Given a molecule, and the graph distance metric (i.e. the distance d_{ij} between two atoms is the number of bonds between them), the Wiener index is defined as [16]:

$$W = \sum_{i<j} d_{ij} \tag{1.1}$$

The Wiener index, although simple, correlates well with various physical properties such as alkane boiling points [17], and their van der Waals surface area [18]. However, the Weiner index is fairly degenerate, with different molecules having identical values. Furthermore, it does not take into account properties such as functional groups.

There are many other uses of topology in chemistry, outside of informatics. For example, Flapan uses three tiers of chirality [19]

1. Geometric chirality: No rigid motion takes a molecule to its own mirror image

2. Chemical chirality: A molecule cannot transform itself (through conformational changes) to its own mirror image at room temperature

3. Topological chirality: A molecule cannot be transformed into its mirror image assuming complete flexibility

A molecule which is chemically chiral may not be topologically chiral. However, topologically chiral molecules are chemically chiral. This hierarchy is particularly useful when studying molecules such as Möbius ladders [20], or those containing non-planar molecular graphs (where planar is used in the graph-theoretical sense) [21].

The language of topology is also useful when studying proteins and other macromolecules. For example, some proteins contain knots of various complexity. These knots have been used to explain structural stability, in that active sites become slightly more favourable [22]. Within polymers in general, circuit topology is used to explain its intra-molecular contacts, which has implications for folding kinetics [23].

The topology of potential energy surfaces has also been widely studied. Peterson *et al* relate the number of critical points of a potential energy surface to its global topology through the relationship between Betti numbers and Euler characteristic [24]. Furthermore, work such as those by Karplus and Becker [25], and Doye [26], attempting to transform potential energy surfaces into network representations of transitions between basin states can be considered topological - they reduce the potential energy surface into a discussion about how it is connected.

However, the regularity with which topology is mentioned in chemistry has caused concern, as written in a *Nature* editorial by Michelle Francl [27]. For example, some of the instances within which topology is mentioned are actually geometrical properties. These properties include the coiling of DNA, which does not change the topology of DNA in the mathematical sense [28]. Although sometimes a little too pessimistic - Francl states that potential energy surfaces all have the same topology as a piece of paper, which can be seen to be untrue from a hypothetical molecule with a single free torsion angle - it is certainly important to ensure that chemists believe they are in the correct branch of mathematics. In particular, one does not want to be searching topological literature when actually the work of interest is instead one of differential geometry. This work endeavours to ensure that all of its uses of topology would be recognised as such by a mathematician.

1.6 Structure of Thesis

The general aim of this work is to detail some of the uses of topological data analysis to chemical problems.

Chapter 2 details the main theory of topological data analysis. Beginning with persistent homology, the chapter introduces the notion of simplicial complexes, which can be used to generate topological structures on sampled data points. The methods of linear algebra are then used to define homology groups, which allow the calculation of holes in these topological structures. Then, the 'persistent' is put into persistent homology, by showing that a nested sequence of topological structures enable the elucidation of holes in a data set, as well as showing their significance. To allow the calculation of statistics on persistence, various representations of persistent homology are introduced, as well as metrics defined, allowing the persistent homology of different data sets to be calculated. Leaving persistence behind, the mapper algorithm is then presented as an alternative topological data analysis technique. The mapper algorithm is shown to be able to generate a simplicial complex representation of a data set, but requires a wide range of parameters.

The main body of this work is split into three general sections: using topological data analysis to generate a space comparing molecular shapes, the application of topological techniques to understand the underlying conformational space (and energy landscape) of molecules, and finally the use of persistent homology to understand the structure of simulated water networks.

Chapter 3 introduces the notion of chemical space in general, as well as describing how this space changes depending on the description of the molecules used. The mapper algorithm is used to understand a general descriptor space. Insights from mapper are then used to improve solubility modelling. Persistent homology is then used as a descriptor for chemical shape, and then used to create a series of metric spaces for molecular shape. These metric spaces are then analysed through dimensionality reduction, and combined using data science techniques. The effect of different variants of persistent homology on the resulting metric space is analysed, as well as the use of persistent homology to create metric spaces dependent on an ensemble of molecular conformation, rather than a single low energy conformer.

Moving onto a problem which is more fundamental, Chapter 4 utilises persistent homology in the characterisation of molecular conformational spaces. Firstly, the notion of a conformer and conformational space is formalised, before two different representations of a conformational space, for the same conformer set, are detailed. Persistent homology is shown to be able to study both the conformational space and energy landscape itself, by the use of different functions on simplicial complexes. Through the use of three examples, alanine dipeptide, pentane, and cyclooctane, persistent homology is shown to be able to compare the two different representations, and verify that the expected conformational space can be found. The rigid geometry hypothesis is tested, showing that the conformational space is indeed independent of bond stretching and bending. With a combination of data science techniques, extrema on energy landscapes are located, and shown to correspond to those that are expected.

The use of persistent homology to analyse water networks is the focus of Chapter 5. Firstly, persistence is shown to be a well-behaved descriptor, that varies smoothly over time. A new representation of persistent homology is created, which is described as 'size-agnostic'. This descriptor is used to compare simulated water networks of different molecular mechanics models, sizes, and temperatures, using machine learning techniques. Furthermore, the descriptor is analysed through dimensionality reduction, allowing the differences of various simulations to be analysed in terms of their persistent homology.

Some of the work discussed in this thesis has been published in peer-reviewed articles or on preprint servers. Chapter 3 discusses some work previously published in [29], although it extends this work in places. Chapter 4 has been partially published in [30], and the published article describes the mathematics in more detail. Finally, results from Chapter 5 can also be found in [31].

Chapter 2

Theory of Topological Data Analysis

This chapter is designed to present the mathematical preliminaries of topological data analysis to a chemist. Therefore, it does not claim to be an exhaustive description, and in particular formal proofs are eschewed. The chapter follows a series of arguments, aiming to convince the reader that topological data analysis is on sound logical footing.

The chapter first introduces the idea of a simplicial complexes, combinatorial building blocks of topological spaces. Algebra on these complexes is then defined, leading to the notion of homology groups, a mathematical definition of holes. This is then extended to data, through persistent homology. The discussion of persistent homology is then continued to different representations of persistence, so all of the mathematical topics of persistence used in this work are found in one location.

The chapter then moves onto the mapper algorithm. Explained via examples, the mapper algorithm is shown to create single network-like summaries of data sets. The different parameters for the mapper algorithm are discussed, demonstrating the difficulty in creating useful mapper networks. Finally, the chapter closes with discussions regarding previous uses of topological data analysis in chemistry.

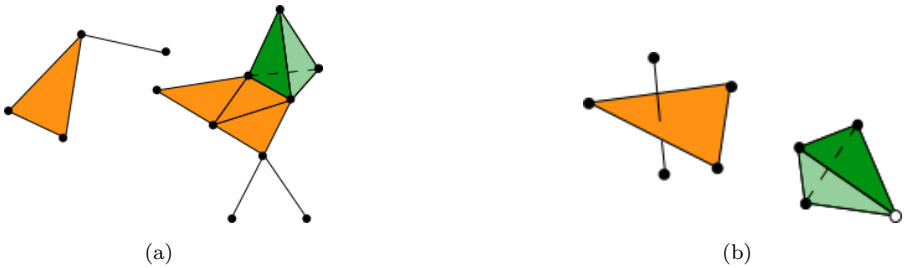
There are a range of texts that can be used for a more complete treatment of the concepts within this chapter. With regard to homology theory, the reader is directed to Hatcher [32], a standard reference book. A treatment of persistence can be found in [33], and [34] contains a series of algorithms for its computation. A perspective of algebraic topology, with reference to its use in biomolecular systems, can be found in [35]. A more recent introduction to persistent homology, designed for neuroscientists, can be found in [36].

2.1 Persistent Homology

2.1.1 Simplices and Simplicial Complexes

Within algebraic topology, the notion of a k-simplex is fundamental, as they can be considered as combinatorial 'building blocks' of topological spaces. A k-simplex, denoted σ, is the smallest convex set in a given Euclidean space $\mathbb{R}^d$ that contains $k+1$ vertices $\{v_i\}$, $i \in \mathbb{Z}$, $0 \leq i \leq k$, where each pair of vertices is linearly independent. These have familiar geometrical interpretations - a 0-simplex is a point, a 1-simplex a line, a 2-simplex a triangle, and a 3-simplex a tetrahedron. When it is necessary to keep track of individual vertices, a k-simplex will be represented as $[v_0, v_1, \ldots, v_{k-1}, v_k]$, and the removal of the j^{th} vertex is denoted as $[v_0, v_1, \ldots, \hat{v}_j, \ldots, v_{k-1}, v_k]$. The $k-1$-simplex created by removal of a vertex from a k-simplex will be referred to as a *face* of the k-simplex.

An *abstract simplicial complex* K can be considered a set of simplices, where it is required that any face of σ in K is also in K. In other words, there are no missing 'building blocks' in K. The *geometric realisation* of K is the embedding of K in some $\mathbb{R}^n$, where it is also required that the intersection between any two simplices $\{\sigma, \sigma'\} \in K$ is either empty or a shared face of both σ and σ'. The geometric realisation therefore differs from the abstract simplicial complex by restricting the 'embedding' of simplex building blocks, where the intersection restriction is trivial for the abstract simplicial complex. In general, the geometric realisation of an abstract simplicial complex will be used, and this will be referred to as a simplicial complex. Figure 2.1 shows two sets of simplices. Figure 2.1(a) is a valid simplicial complex. On the other hand, Figure 2.1(b) is not a valid geometrical representation of a simplicial complex - the left structure has an intersection which is not in a shared simplex face. Furthermore, the right structure is missing a 0-simplex, therefore it is not even a valid abstract simplicial complex.



(a) (b)

FIGURE 2.1: A simplicial complex (a) and a set of simplices that do not form a [geometric realisation of a] simplicial complex (b)

Simplicial complexes are the general combinatorial objects that will be discussed in this chapter. However, before the study of holes in simplicial complexes can be formalised, it is first necessary to introduce methods of linear algebra. For the reader less equipped

with such methods, various textbooks are available, such as [37]. Furthermore, obscure terms are formally defined in the text.

2.1.2 The Boundary Map

To perform linear algebra on the simplices of K, the notion of a general vector space is required.

> A vector space $\mathcal{V}$ over a field $\mathbb{F}$ is a set of elements $\{\alpha, \beta, \gamma, \ldots\} \in \mathcal{V}$ along with the following operations:
>
> 1. A vector addition, satisfying:
>   (a) $\alpha + \beta = \beta + \alpha$ (Associativity)
>   (b) $\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma$ (Commutativity)
>   (c) There exists an element $0 \in \mathcal{V}$ such that $\alpha + 0 = \alpha$ (Existence of an identity)
>   (d) For every $\alpha \in \mathcal{V}$ there exists an element $\alpha' \in \mathcal{V}$ such that $\alpha + \alpha' = 0$ (Existence of an inverse)
> 2. For $c \in \mathbb{F}$ and $\alpha \in \mathcal{V}$, the scalar product $c\alpha \in \mathcal{V}$ is defined and satisfies:
>   (a) $c(\alpha + \beta) = c\alpha + c\beta$ (Distributivity with respect to vector addition)
>   (b) $(a + b)\alpha = a\alpha + b\alpha$ (Distributivity with respect to field addition)
>   (c) $(ab)\alpha = a(b\alpha)$ (Compatibility with field multiplication)
>   (d) $1\alpha = \alpha$ (The multiplicative identity in $\mathbb{F}$ is the identity of scalar multiplication)

In textbooks, rather than the creation of vector spaces over fields, the objects used for homological algebra are modules over a ring R, often the ring of integers $\mathbb{Z}$. However, this ring is not suitable for computation (it is infinite), and the use of a ring leads to other complications such as torsion within algebraic topology. Further details regarding the differences between fields and rings can be seen in Appendix C.

Computation therefore uses the notion of a vector space over a finite field $\mathbb{F}$. This is most commonly the field $\mathbb{Z}_p \equiv \mathbb{Z}/p\mathbb{Z}$ where p is prime. In this work, the field chosen is $\mathbb{Z}_2 = \{0, 1\}$, unless otherwise stated. This choice allows computation to become more efficient, and ensures that simplicial orientation does not need to be defined (within $\mathbb{Z}_2$, -1 and 1 are equivalent). Furthermore, all illustrations and discussions in this chapter will use $\mathbb{Z}_2$ as the field of coefficients. Physical scientists are often familiar with the notion of vector spaces over the field $\mathbb{R}$.

Alongside the field of coefficients, the elements $\{\alpha, \beta, \ldots\}$ must be determined. The elements are chosen to be the set of p-simplices, and the resulting vector space will be denoted $C_p(K)$. Elements of $C_p(K)$ are referred to as p-chains.

Consider the simplicial complex in Figure 2.2. Formally, this is the complex $K = \{[v_0], [v_1], [v_2], [v_0, v_1], [v_1, v_2]\}$. $C_0(K)$ is the vector space generated by the vertices of K: $\{0, [v_0], [v_1], [v_2], [v_0] + [v_1], [v_0] + [v_2], [v_1] + [v_2], [v_0] + [v_1] + [v_2]\}$. Similarly $C_1(K) = \{0, [v_0, v_1], [v_1, v_2], [v_0, v_1] + [v_1, v_2]\}$. As there are no higher order simplices in K, $C_p(K) = \{0\}$ for $p > 1$. From this example, it should be clear that elements of $C_p(K)$ can be identified as taking a (potentially zero) selection of p-simplices from K in the case of coefficients in $\mathbb{Z}_2$.

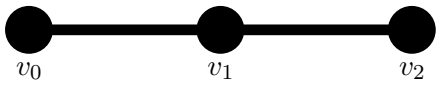


FIGURE 2.2: An example simplicial complex with labelled vertices

Next, a linear map between vector spaces of consecutive dimension $C_p(K) \to C_{p-1}(K)$ is defined. This map is the algebraic equivalent of the geometric notion of boundary. The boundary map is defined on individual simplices:

$$\partial_p \sigma = \sum_{i=0}^{p} [v_0, \ldots, \hat{v}_i, \ldots, v_p] \tag{2.1}$$

with an example seen in Figure 2.3. Here $\partial_1([v_0, v_1]) = [v_0] + [v_1]$. Through linearity, it

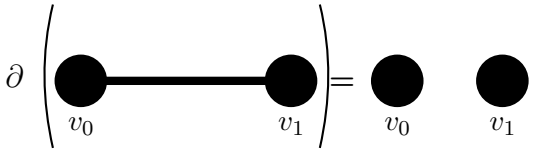


FIGURE 2.3: A cartoon illustrating the operation of the boundary map on a 1-simplex. A 1-simplex is mapped to its two endpoint vertices.

is possible to define the effect of the boundary map on p-chains, and this is illustrated in Figure 2.4. In this case $\partial_1([v_0, v_1] + [v_0, v_2]) = [v_0] + [v_1] + [v_0] + [v_2]$ However, in $\mathbb{Z}_2$ $1 + 1 = 0$, therefore $\partial_1([v_0, v_1] + [v_0, v_2]) = [v_1] + [v_2]$. In a similar way, for any simplex

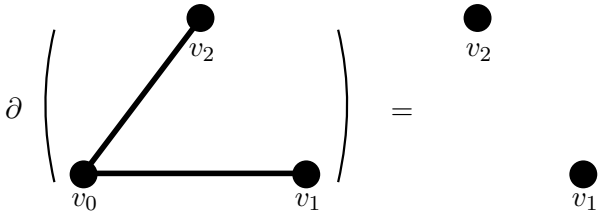


FIGURE 2.4: A cartoon illustrating the operation of the boundary map on a 1-chain. The boundary coincides with geometric intuition.

K, the boundary map ∂_p can be calculated for any element of $C_p(K)$.

The boundary operator takes elements of $C_p(K)$ to $C_{p-1}(K)$. It might be natural to ask: what does successive operation of the boundary operator achieve? It is a standard

exercise to show that $\partial_{p-1} \circ \partial_p = 0$, illustrated for the field $\mathbb{Z}_2$ in Appendix E, with the more general case of the ring $\mathbb{Z}$ in Hatcher [32].

2.1.3 Homology Groups

Now the boundary map has been introduced, it is possible to discuss how it can be used to detect holes in simplicial complexes. In particular, subclasses of p-chains based on the result of the application of the boundary map are defined. The first subclass, p-cycles, are defined as:

$$Z_p = \ker \partial_p = \{\sigma \in C_p \,|\, \partial(\sigma) = 0\}$$

i.e. the set of p-chains which are sent to 0 by the boundary map. The second class, p-boundaries, are defined as:

$$B_p = \operatorname{im} \partial_{\mathrm{p}+1} = \{\partial(\tau) \,|\, \tau \in C_{p+1}(K)\}$$

the set of p-chains which are the result of applying the boundary map to $p+1$-chains. From the relation $\partial_{p-1} \circ \partial_p = 0$, it is clear that every element of B_p is an element of

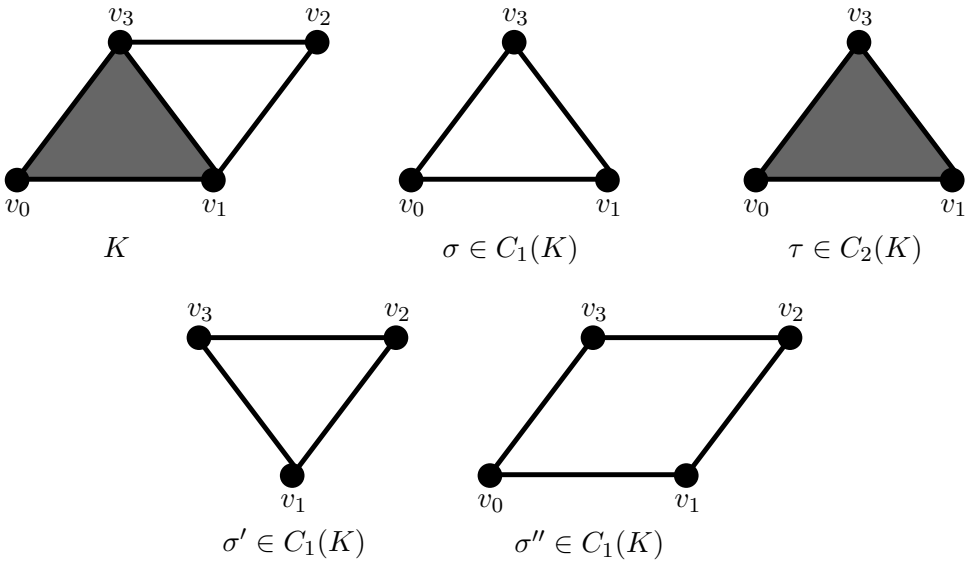


FIGURE 2.5: A simplicial complex K, and a series of subcomplexes $\{\sigma, \tau, \sigma', \sigma''\} \subseteq K$ Shaded areas represent filled simplices, whereas a white background signifies that it is unfilled - K contains only one 2-simplex. The operation of the boundary map on the p-chains present in K allows homology groups to be defined.

Z_p. The vector spaces B_p and Z_p contain all of the necessary information to compute the holes in K. This is perhaps best illustrated by example (Figure 2.5). Begin with the simplicial complex K. This is clearly a simplicial complex with a hole. Consider the chain σ. This is sent to zero by the boundary map. However, σ does not enclose a hole - it is actually the result of the operation of ∂_2 on τ, a boundary of a 2-chain.

In contrast, the chain σ' is also in the kernel of ∂_1, and it is not an element of the image of ∂_2. This is because it encloses a hole, and is also true of σ'', which encloses the same hole. Algebraically, this is dealt with by noting that $\sigma'' - \sigma' = \sigma = \partial_2(\tau)$. This implies the quotient vector space $\frac{\ker \partial_1}{\text{im } \partial_2}$ contains the same class representing σ'' and σ'.

In general, this quotient is the vector space generated by the p-cycles of K that *are not* boundaries. The following quantity:

$$H_p(K) = \frac{Z_p(K)}{B_p(K)} \tag{2.2}$$

is defined, the p^{th} homology group of K.

2.1.4 Betti Numbers

The homology groups themselves are fairly unwieldy objects, but can be easily related to geometric quantities. In particular, the p^{th} Betti number β_p:

$$\beta_p(K) = \dim_{\mathbb{F}} H_p(K) \tag{2.3}$$

carry topological information, counting p-dimensional holes. The notation can be described as the amoount of copies of $\mathbb{F}$ needed to describe $H_p(K)$. Betti numbers of small degree have obvious geometrical interpretations:

1. $\beta_0(K)$ is the number of connected components of K
2. $\beta_1(K)$ is the number of holes bounded by a loop in K, such as those found in a ring
3. $\beta_2(K)$ is the number of holes bounded by a surface in K, such as those found in a sphere

With some simple spaces and their associated Betti numbers (up to second degree) described in Table 2.1. Other spaces and their Betti numbers are introduced as needed.

Space	β_0	β_1	β_2
$\mathbb{R}^n$	1	0	0
S^1	1	1	0
S^2	1	0	1
T^2	1	2	1
Borromean rings	3	3	0

TABLE 2.1: Some topological spaces and their associated Betti numbers

An example calculation of the Betti numbers of the Klein Bottle and real projective plane can be found in Appendix D. These examples begin with a complex homeomorphic to the space of interest, and work through the process of calculating the homology groups themselves. Furthermore, they also illustrate the importance of the choice of $\mathbb{F}$, with the resulting Betti numbers shown to be dependent on this choice. This result is later shown to be useful when classifying spaces through homology.

2.1.5 Applying Homology to Data

Having discussed the calculation of homology groups and Betti numbers, it is of interest as to how we can apply these methods to data. A data set of m measurements of $n_{features}$ variables can be considered to be m points in some subset of $\mathbb{R}^{n_{features}}$. Of interest here is therefore: what structure can be added to these isolated points such that a useful topology is found? Clearly the answer here depends on the exact meaning of 'useful' - but a possible construction has already been discussed: the simplicial complex. In general, the sampled data points can be considered as vertices, and relationships between $k+1$-tuples of points determine the presence of k-simplices. This is all perhaps best understood via example.

2.1.5.1 The Vietoris-Rips Complex

The Vietoris-Rips (Rips) complex $VR_r(S)$ was originally developed by Leopold Vietoris as a means of calculating the homology of metric spaces [38]. To construct the Rips complex on a finite subset of points S, the following procedure is used:

1. Define a parameter r
2. For all subsets $s \subseteq S$
3. If diam $s \leq 2r$, include the simplex with vertices in s

Geometrically, this is equivalent to creating balls of radius r around the points in s, and including the simplex if there is a non-zero intersection between all pairs of balls. The Rips complex is completely defined by its 1-skeleton, as it only depends on pairwise relationships. It is clear why the Rips complex is commonly used, as it can be calculated directly from a distance matrix, and there are various algorithms for its efficient computation [39, 40]. This simplicity has topological consequences. Most importantly, the Rips complex does not necessarily have the same topology as the union of balls used in its creation (the nerve theorem [32]). An example Rips complex can be seen in Figure 2.6, for a 2D point cloud. The Betti numbers for this complex are $\beta_0 = 2$, $\beta_1 = 1$, and $\beta_p = 0$ for $p \geq 2$.

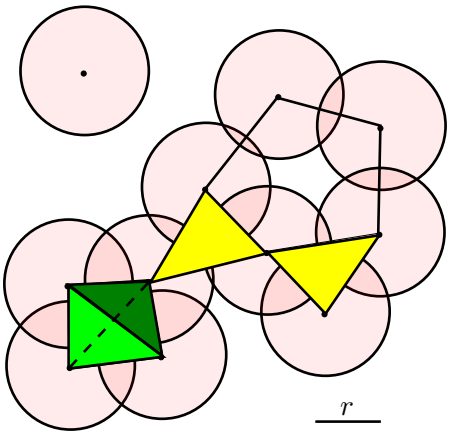


FIGURE 2.6: An example Vietoris-Rips complex constructed on a 2D point cloud. Yellow and green are used to denote 2- and 3-simplices respectively. Circles are used to illustrate the r-balls from which the Vietoris-Rips complex is constructed, and do not actually appear in the complex.

2.1.6 The Choice of r

The Vietoris Rips complex is highly dependent on the value of r. At $r = 0$, every point is isolated, and β_0 is the number of points in the set. In contrast, $\lim_{r\to\infty} \beta_0 = 1$, and $\lim_{r\to\infty} \beta_n = 0$ for $n \geq 1$. Therefore, a natural question would be: what value of r best recreates the topology of the point cloud?

Again, discussion is motivated by example. In this case, the point cloud of interest is the set of vertices of a regular hexagon, of nearest neighbour distance R. A series of Rips complexes for this data set created for $r \in \{0, r, \sqrt{3}R\}$ can be seen in Figure 2.7. The value of r that leads to the Rips complex recreating the topology of the hexagon is $r = R$. However, if there was a small amount of noise in the sampling procedure, the optimal value of r would likely change. Furthermore, in the general case, the homology groups of the underlying space is unknown - we are trying to estimate them through the simplicial complex. However, this notion of altering the parameter r can be used to define persistent homology.

2.1.7 Persistent Homology

For persistent homology, a nested sequence of simplicial complexes is required:

$$K_0 \hookrightarrow K_1 \hookrightarrow \ldots \hookrightarrow K_N \tag{2.4}$$

where $K_0 \subseteq K_1 \subseteq K_2$ etc., and $\hookrightarrow$ denotes an inclusion map. It is clear that this property is satisfied by a sequence of Vietoris-Rips complexes with increasing r.

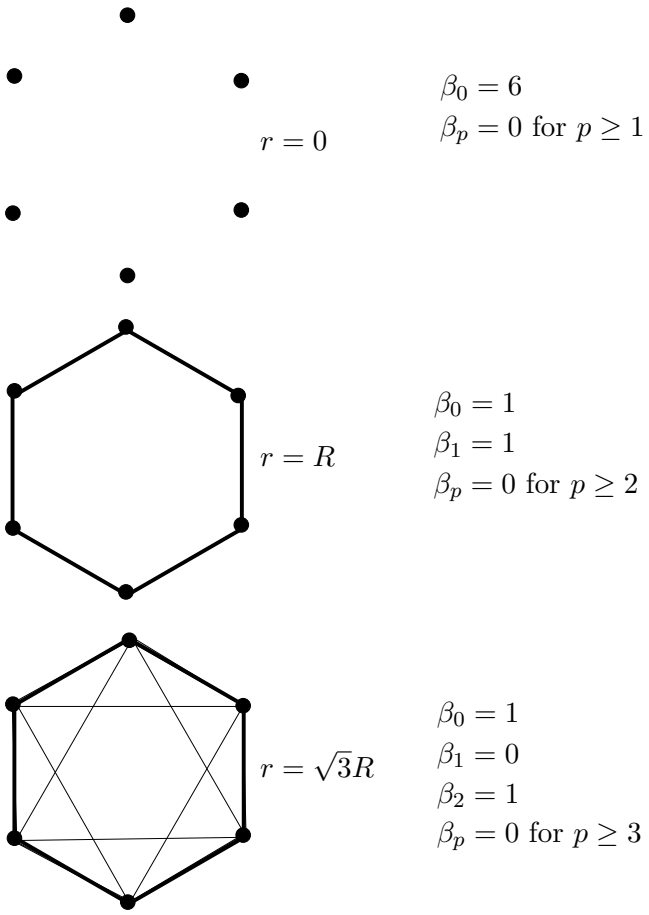


FIGURE 2.7: A series of Rips complexes constructed upon the point cloud of the vertices of a regular hexagon, with a nearest neighbour distance of R. Betti numbers are also seen. The second degree feature seen at $r = \sqrt{3}R$ is a consequence of the Rips complex not satisfying the nerve theorem.

Let $0 \leq i \leq j \leq N$. The inclusion maps of Equation 2.4 lead to induced maps in homology:

$$f_p^{i,j} : H_p(K_i) \to H_p(K_j) \tag{2.5}$$

The p of $f_p^{i,j}$ is now made implicit for clarity. Each degree of homology is studied independently. Three classes within homology groups are defined based on these $f^{i,j}$:

- Classes α that are *born* at i. These are classes where $\alpha \neq 0$ and $\alpha \notin \operatorname{im} f^{i-1,i}$
- Classes β that *persist* from $i \to j$. These are classes where $f^{i,j}(\beta) \notin \operatorname{im} f^{i-1,j}$
  - This implies that β also persists from $i \to i + \epsilon$ if $i + \epsilon < j$
- Classes γ that *die* at j. These are classes where $\gamma \in \ker f^{j-1,j}$ or $f^{j-1,j}(\gamma) = f^{j-1,j}(\gamma')$ and γ' was born before γ
  - The first requirement defines classes that are 'filled in' at j, whereas the second defines classes that merge with older classes. The older class is given

preference to ensure that a suitable basis can always be found for the merged class.

- There is no certainty that all classes will die. For example, the limiting behaviour of the Rips complex as $r \to \infty$ is to have one connected component, therefore a zeroth degree homology class does not die. Such a class may be said to live *to infinity*.
- The number of points that live to infinity of degree p is referred to as the p^{th} persistent Betti number. This work will also use the term 'persistent' to refer to features that live far longer than others in persistent homology.

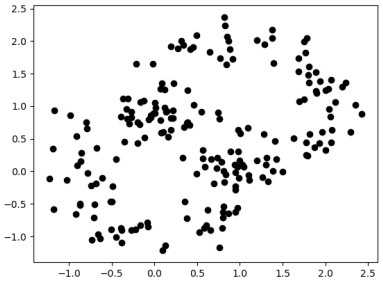
Given a sequence of nested simplicial complexes parameterised by some δ (often referred to as *time*), the persistent homology procedure returns a set of homological classes. Each homology class can be identified with a birth and death time (δ_b and δ_d respectively). Features that are born and die soon after are often considered to be topological noise, whereas classes that persist for an extended period are considered to be true features of the underlying structure. However such definitions should be used as guidance only. A video showcasing the calculation of the persistent homology of the Rips complex on the vertices of a regular hexagon can be found at [41], although it is important that the persistence diagram representation is first understood to make best use of the video.

2.1.8 Persistence Diagrams and Barcodes

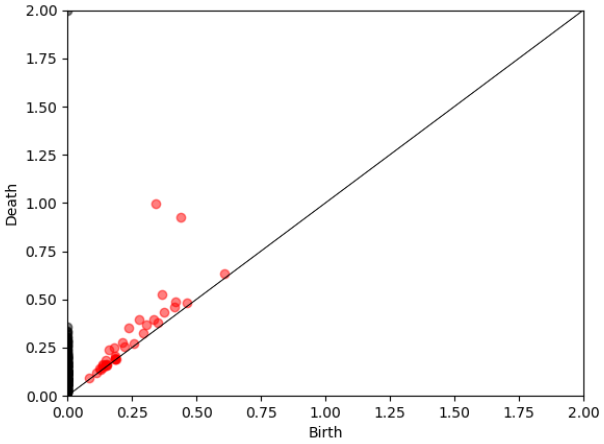
The most common methods for representing the information given through the persistent homology procedure are known as [persistence] diagrams and barcodes. The persistence diagram is a plot of δ_b vs δ_d, whereas the barcode is a series of lines, one for each feature, stretching from δ_b to δ_d.

Figure 2.8(a) contains a data set sampled from two circles, centred at $(0, 0)$ and $(1, 1)$ with radius 1, with normally distributed random noise. Persistent homology was calculated for the sequence of Rips complexes with the standard Euclidean (l_2) metric, and the corresponding persistence diagram and barcode can be seen in Figures 2.8(b) and 2.8(c) respectively. Although the persistence diagram and barcode contain the same information, related persistent homology concepts are easier to understand depending on representation. For example, it is easier to extract the number of features that persist between δ and δ' from the barcode, as this is simply the number of lines that pass through both ends of the interval. In contrast, relationships between δ_b and δ_d are more obvious in the diagram representation, as well as persistent homology metrics being easier to define.

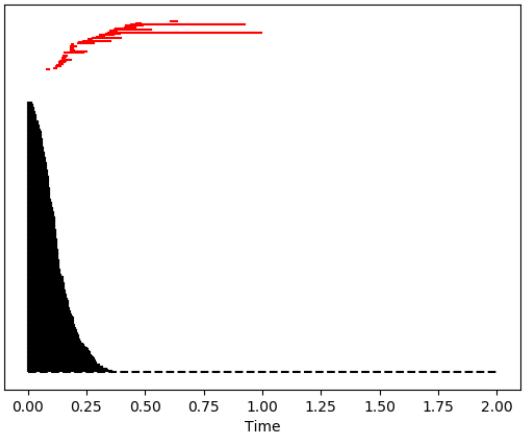
In this work, the persistence diagram representation is preferred to the barcode representation. Zeroth, first, and second degree homological features are coloured black, red, and blue respectively.



(a) Intersecting circle data



(b) Persistence Diagram



(c) Persistence Barcode

FIGURE 2.8: A (synthetic) data set of two intersecting circles with noise, and the data set's corresponding Rips persistence diagram and barcode. Black and red features correspond to zeroth and first degree homology respectively.

2.1.9 Metrics on Persistence Diagrams

A persistence diagram can be considered to be a multiset of points $\in \mathbb{R}^2$, where each point is also given a multiplicity. By giving each point a multiplicity, the persistence diagram is able to handle two persistent features with identical birth and death times This enables the definition of metrics on the space of persistence diagrams, in particular the p-Wasserstein metrics:

$$d_{Wp}(PD_1, PD_2) = \inf_{\phi: PD_1 \to PD_2} \left[\sum_{x \in PD_1} ||x - \phi(x)||_p \right]^{\frac{1}{p}} \tag{2.6}$$

Where ϕ is a bijection (i.e. a matching between every unique point in PD_1 to a unique point in PD_2). The p-Wasserstein metric can therefore be considered as finding the optimal matching between persistence diagrams. However, this matching contains an important caveat. In particular, as ϕ is a bijection, the two persistence diagrams must contain the same number of points (including multiplicities) - which is not the case in general.

This issue is resolved by the presence of the diagonal, which can be considered to be the multiset of points that are born and die at the same time. In principle, there can be an infinite number of points with this property. An infinite degeneracy is therefore associated to each point on the diagonal (δ, δ). This therefore allows bijections to be defined between persistence diagrams with different numbers of features. Furthermore, this ensures that points in a persistence diagram do not appear and disappear out of nowhere. Instead, they are considered to come from the diagonal.

In practice, this increases the number of points that need to be computed, and the number of bijections that need to be considered. Furthermore, this definition can handle points that live to infinity, and lead to an infinite distance if the two persistence diagrams have different persistent Betti numbers.

The bottleneck metric can be considered as the limit of p-Wasserstein distances as $p \to \infty$, and can be written as:

$$d_B(PD_1, PD_2) = \inf_{\phi: PD_1 \to PD_2} \left[\sup_{x \in PD_1} ||x - \phi(x)||_\infty \right] \tag{2.7}$$

The bottleneck metric is computationally cheaper than the p-Wasserstein metrics, and in particular the stability theorem for persistence diagrams utilises the bottleneck metric. The stability theorem ensures that slight changes for persistent homology input only causes slight changes in the resulting output [42].

2.1.10 Average Persistence

In principle, the process of data collection is inherently noisy. Therefore, it is desirable to be able to perform statistics on results, persistent homology included. For example, if persistent homology is to be used to study equilibrium properties of chemical simulation, it is important to be able to in some sense 'average' the persistent homology, and make conclusions from the average behaviour. This will ensure small deviations are removed from the persistence. Given a set of persistence diagrams PD_i, a hypothetical average persistence diagram $\overline{PD}$ could minimise the following:

$$\sum_i d_B(\overline{PD}, PD_i) \tag{2.8}$$

i.e. the average persistence diagram is the one closest (in the bottleneck metric sense) to the individual persistence diagrams. However, such a definition is not unique. This effect is illustrated in Figure 2.9, where there are two potential choices for the average persistence diagram that minimise Equation 2.8.

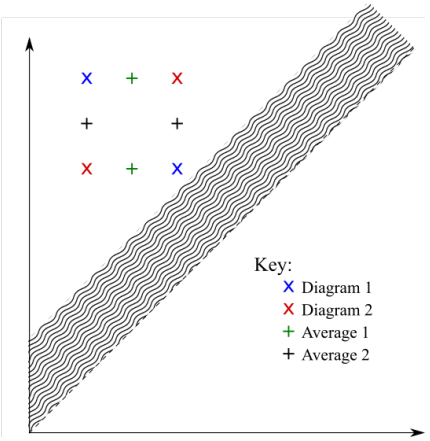


FIGURE 2.9: An example where the mean defined in 2.8 is not unique. For two persistence diagrams (blue and red) the green and black sets are both equidistant. Wavy lines are used to indicate that the points are not close to the diagonal, and therefore bottleneck distance matching will not send any points to the diagonal.

As diagrams and barcodes both lead to non-unique means, other persistence representations, more amenable to statistics, have been developed. These are designed to have one or more of the following properties:

- Averaging is possible and unique
- Representation is a metric space, allowing computation of distances
- Representation is an inner-product space, allowing computation of scalar products
- Space allows vectorisation, for use in machine learning techniques

- Space allows computation of real-valued useful characteristics of initial persistent homology

In this work, persistence landscapes [43] and persistence images [44] are discussed. These representations fulfill all of the above criteria, and are used to accomplish different tasks. Here, a brief introduction to their computation is presented. Other representations include persistence entropy [45, 46, 47] and persistence vineyards [48].

2.1.11 Persistence Landscapes

The persistence landscape was first defined by Bubenik in [43]. The persistence diagram is transformed into a sequence of nested functions, and inherited properties are used to define means and distances.

For a given degree homology, each point in the persistence diagram (δ_b, δ_d) is transformed to a piecewise linear function:

$$f_{(\delta_b,\delta_d)}(x) = \begin{cases} 0 & x \notin (\delta_b, \delta_d) \\ x - \delta_b & x \in \left(\delta_b, \frac{\delta_b+\delta_d}{2}\right] \\ -x + \delta_d & x \in \left(\frac{\delta_b+\delta_d}{2}, \delta_d\right) \end{cases} \tag{2.9}$$

This can be considered to be a rotation such that the diagonal becomes the x-axis, before lines of gradient 1 are drawn to and from each persistent point. The persistence landscape, denoted Λ is then the sequence of functions $\lambda_k(x)$ where $\lambda_k(x)$ is the k^{th} largest value of $f_{(\delta_b,\delta_d)}(x)$ for all persistent points. To ensure that algebra can be defined between two different persistence landscapes, $\lambda_k = 0$ if the k^{th} largest value does not exist. The persistence landscape for the first degree homology of the data set found in Figure 2.8(a) can be seen in Figure 2.10. As the persistence landscape is a set of functions, several useful properties are inherited. The mean persistence diagram $\overline{\Lambda}$ can be defined on a set of landscapes Λ^i by finding the mean of the individual landscape functions λ_k^i.

$$\overline{\lambda_k} = \frac{1}{N} \sum_{i=1}^{N} \lambda_k^i(x) \tag{2.10}$$

with $\overline{\Lambda}$ being the set of $\overline{\lambda_k}$. Furthermore, distances between landscapes can be defined using standard l_p distances:

$$d_p(\Lambda_1, \Lambda_2) = \left[\sum_{k=1}^{\infty} \int |\lambda_k^1(x) - \lambda_k^2(x)|^p \, \mathrm{d}x \right]^{\frac{1}{p}} \tag{2.11}$$

A useful property of persistence landscapes, proven in [49], is that *any* real, linear functional, when applied to a persistence landscape results in a real-valued random

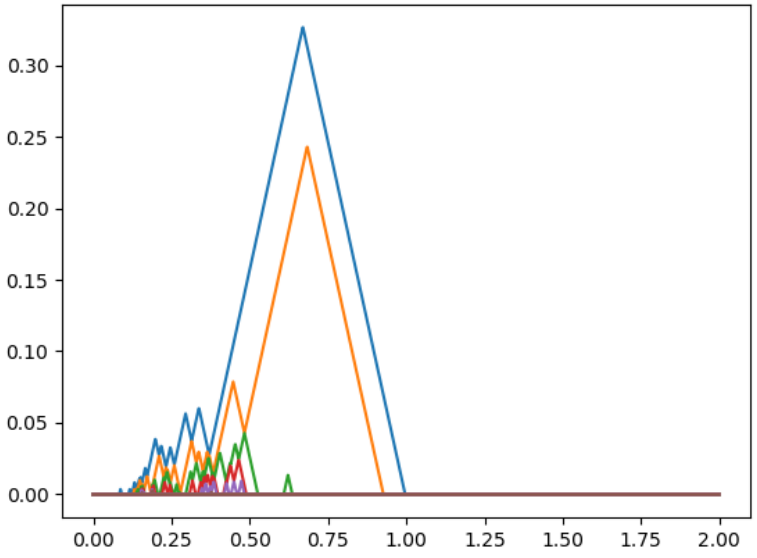


FIGURE 2.10: The first degree persistence landscape for the data found in Figure 2.8(a).

variable that is approximately normally distributed.

$$\begin{aligned} Y &= \int f\Lambda \, \mathrm{d}x \\ \sqrt{n}(\bar{Y}_n - E[Y]) &\xrightarrow{d} N(0, \mathrm{Var}(Y)) \end{aligned} \tag{2.12}$$

Where $\xrightarrow{d}$ is used to denote a limit in distribution. Often, this functional is set to 1:

$$Y = \sum_{k=1}^{\infty} \int \lambda_k(x) \, \mathrm{d}x \equiv ||\Lambda||_1 \tag{2.13}$$

Persistence landscape computation can be performed using two methods, both detailed in [49]. The first is a rigorous calculation of persistence landscapes using its critical points. The second calculates persistence landscapes over a grid. Although non-exact, the second method is considerably faster, and allows for easier computation of landscape statistics. However, care must be taken when comparing landscapes, as operations need to be performed to ensure the grids match.

2.1.12 Persistence Images

In contrast to persistence landscapes, which transforms a diagram into a set of functions, the persistence image transforms a diagram into a matrix in $\mathbb{R}^{n \times n}$. This representation can be represented as a single-channel image, as an $n \times n$ grid of pixels with intensity in $\mathbb{R}$. Such objects are well-suited for machine learning techniques. For example, the

original implementation of persistence images [44] constructed a classifier on the well-known MNIST handwritten letter/number data set [50].

Similarly to the persistence landscape, a persistence image is calculated for each degree of homology. Firstly, points are transformed from a birth-death representation (δ_b, δ_d) to birth-persistence coordinates $(\delta_b, \delta_p \equiv \delta_d - \delta_b)$. Each point is then represented by a Gaussian function, centred at (δ_b, δ_p):

$$g_i(x, y) = \frac{1}{2\pi\sigma^2} e^{-[(x-\delta_b)^2+(y-\delta_p)^2]/2\sigma^2} \tag{2.14}$$

Where i has been introduced as an index to iterate over features in the persistence diagram. The various functions g_i are transformed into a single surface:

$$\rho(x, y) = \sum_i f(x, y) g_i(x, y) \tag{2.15}$$

Where the weight function $f(x, y)$ is necessary to account for the diagonal, with properties discussed later. $\rho(x, y)$ is transformed into the persistence image by discretising $\mathbb{R}^2$ into pixels, and integrating $\rho(x, y)$:

$$I_\rho(p) = \iint_p \rho(x, y) \, \mathrm{d}y \, \mathrm{d}x \tag{2.16}$$

The presence of the weight function may at first seem unnecessary, but becomes clear when the presence of the diagonal is considered. As previously discussed, each point of the diagonal (δ, δ) has infinite multiplicity. If this was not appropriately dealt with, the pixels of the persistence image containing the x-axis would have a value of infinity. The weight function is therefore defined to remove this by being equal to 0 at the x-axis.

The choice of weight function is a parameter for the user. In this work, the weight function is defined as:

$$f(x, y) = \frac{y}{y_{max}} \tag{2.17}$$

Where y_{max} is the maximum value of the filtration parameter used in the original persistence calculation. This definition fulfils the general intuition of 'persistent features that live for a while are topologically relevant, whereas features which are born and die quickly are topological noise'. However, when using this definition, care must be taken when comparing persistence images that y_{max} is the same, or the weight function would differ. The first degree persistence image for the data of Figure 2.8(a) can be found in Figure 2.11.

2.1.12.1 Comparison of Landscapes and Images

Both landscapes and images satisfy the 'useful properties' of a persistence representation as defined in Chapter 2.1.10. However, they both have different strengths and

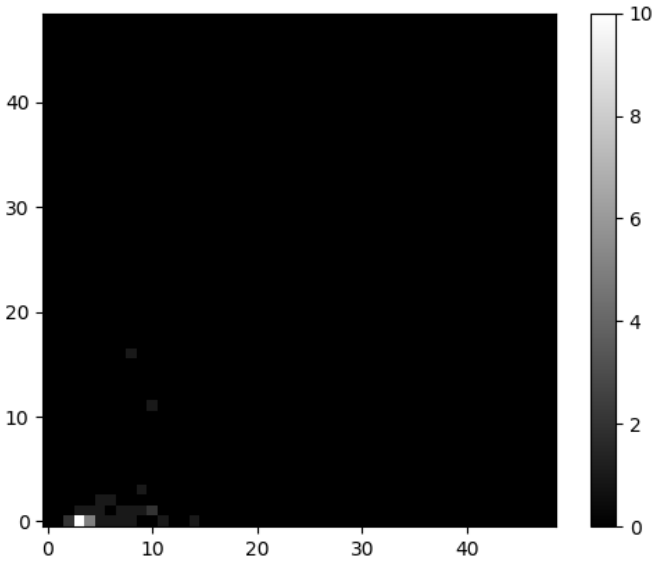


FIGURE 2.11: The first degree persistence image for the data found in Figure 2.8(a), coloured by image intensity. The vast majority of the image contains no features.

weaknesses.

Firstly, the transformation from a persistence diagram to its persistence landscapes is entirely reversible. The same is not true for the persistence image. Furthermore, the central limit theorem defined in Equation 2.12 provides a wide variety of quantities that could be used in the comparison of landscapes, alongside estimates of statistical significance.

On the other hand, the weighting function $f(x, y)$ of an image allows control of which points are seen as significant. For example a sigmoidal function would treat all points on its plateau equally. However, this flexibility does lead to the added complication that comparison of images is only reasonable if they were constructed in precisely the same manner.

2.1.13 Persistent Homology Software

Several software suites have been developed for the computation of persistent homology, and this work does not endeavour to be an exhaustive list. However, it would be remiss to not include a brief comparison of the software used in throughout this work.

The R package TDA [51] provides some tools for persistent homology computation in the R statistical software package. This includes Rips persistent homology on arbitrary distance matrices, as well as the more computationally expensive alpha shape complex [52]. Furthermore, the TDA package contains routines that can calculate approximate

confidence intervals for persistence diagrams [53], allowing estimates as to which features are topologically relevant. The package is also able to calculate Wasserstein and bottleneck metrics between persistence diagrams.

The C++ program Ripser is designed for the efficient computation of Rips complexes of arbitrary distance matrices [54]. It is time and memory efficient, but severely limited by its lack of other complexes. Unlike the R package TDA, Ripser is able to compute persistent homology over any $\mathbb{Z}_p$ basis.

The python library GUDHI is designed to be the most flexible of the persistent homology softwares discussed [55]. As well as Rips and alpha complex persistence, GUDHI is able to calculate persistent homology of any user defined sequence of simplicial complexes, through the underlying simplex tree object [56]. GUDHI can also perform persistent homology calculations over any $\mathbb{Z}_p$ basis. Lastly, GUDHI can be used via the underlying C++ library, which contains more features such as the Cech complex [57], as well as software to produce persistence landscapes through the persistence landscape toolbox [49].

2.2 The Mapper Algorithm

Mapper is another technique of topological data analysis, distinct yet related to persistent homology. Rather than the creation of a series of nested subcomplexes, and analysing how the topology changes, the mapper algorithm [58] is a method designed to produce a single low-dimensional simplicial complex from which information about the underlying data may be extracted. Firstly, the 1-dimensional mapper algorithm will be discussed, with reference to an example. From this, the general mapper algorithm can be introduced.

The 1-dimensional mapper algorithm requires the following as input:

- A data set X
- A metric d on X
- A filter function $f : X \to \mathbb{R}$
  - This is often referred to as the *lens*
- A covering of $\mathbb{R}$ by overlapping intervals (a_i, b_i)

Figure 2.12 contains a cartoon illustrating the entire algorithm. X is represented by the 2D point cloud in the centre of the image. d is the Euclidean distance in $\mathbb{R}^2$. f is the height function, taking every point in X to its y-coordinate. The covering is seen

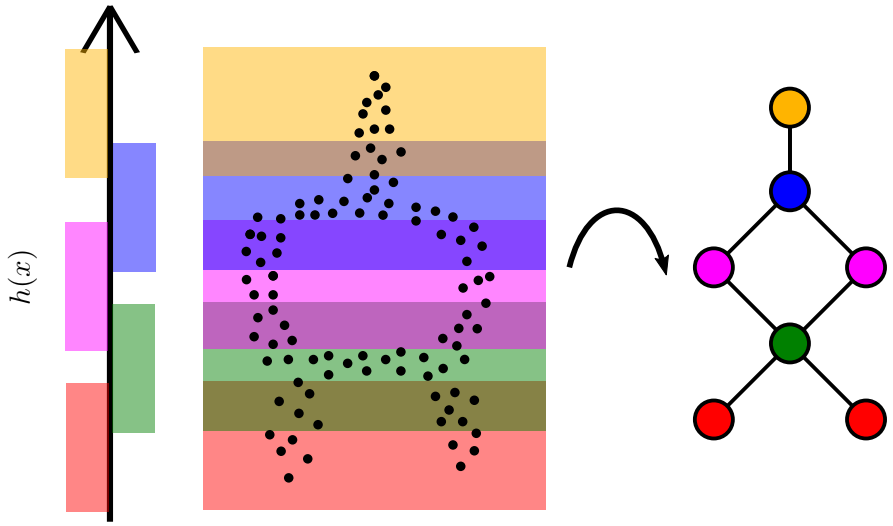


FIGURE 2.12: A cartoon illustrating the 1-dimensional mapper algorithm. The point cloud is seen in the centre, and the filter function (height) alongside its covering found on the left. The pre-images of the covering patches are found, and clustering performed on these pre-images. If two clusters from distinct patches share the same point, they are then joined by a connection. This results in the network seen on the right.

by the overlapping rectangles. Given a 'patch' in the covering, the mapper algorithm firstly determines the points in the pre-image of the patch. For example, the points in X inside the red interval. A clustering algorithm is performed on this pre-image, using the metric d as the clustering parameter. For the red pre-image, this would result in 2 clusters. When this is performed for all of the patches, the result would be the nodes found on the right of Figure 2.12. Two nodes n_1 and n_2 are connected if $n_1 \cap n_2 \neq \varnothing$. By definition, this cannot occur between nodes from the same patch, leading to linkages between patches being found. The resulting network of Figure 2.12 clearly captures information about the relationships in X. In particular, the network contains the hole present in X, and shows how different regions of X merge and split as a function of height.

The general mapper algorithm is defined for any filter function $f : X \to \mathcal{P}$, rather than the 1-dimensional $\mathbb{R}$. The only requirement of $\mathcal{P}$ is that it is able to admit a similar covering of overlapping intervals. The most intuitive way of creating this space is to simply use a sequence of 1-dimensional filter functions f such that $\mathcal{P}$ is some $\mathbb{R}^n$, but this is not by any means the only method - for example, the filter function could be defined as $f : X \to S^1$, with a cover of angular patches.

The remaining step to define the general mapper algorithm is to note that $\mathcal{P}$ could have a topological dimension different to 1. In turn, this leads to each patch of the cover being parameterised by two variables, with more than two patches able to overlap. Rather than only connections possible between any nodes, higher order simplices can occur. In general, the p-simplex between nodes n_i where $1 \leq i \leq p$ is present if $\bigcap_i n_i \neq \varnothing$.

2.2.1 Mapper Parameters

Mapper is a highly parametrised algorithm, and therefore care must be taken to ensure that the resulting networks will be useful. The most important parameters are the choice of metric and lens. Changing these could feasibly lead to entirely different networks, as it is akin to fundamentally changing how the data is understood. The covering of $\mathcal{P}$ can also be changed, with the number of patches and percentage overlap parameters often referred to as the resolution and gain of the cover respectively. In contrast to changing the metric or lens, work has been carried out to understand how the choice of resolution and gain impacts the network [59, 60], often referred to as investigating the stability of mapper. The last parameter for mapper is the choice of clustering algorithm (and its associated hyperparameters). It can be shown that the choice of clustering algorithm can lead to different partitions, and therefore a different resulting network [61].

The main method for determining if observations gained from mapper networks are real is to create networks with different parameters, and to see if the conclusions still hold. Conclusions that can be made from a wide variety of network parameters are considered to be real relationships within the data.

2.2.2 Mapper Software

There are a few software implementations of the Mapper algorithm. The most well known is commercial software produced by Ayasdi [62, 63]. However, this software is proprietary, with the lenses in particular sometimes behaving as a black box. There are also issues with regards to data ownership when using Ayasdi's implementation of mapper, and for this reason large companies with sensitive data may choose to use a different implementation if terms with Ayasdi are unable to be agreed.

The open-source implementations of mapper tend to be written in python. Mullner and Babu's version, known as 'Python Mapper' was written in 2013 [64]. However, this has been superseded by KeplerMapper [65], which is more fully featured and part of the growing set of python libraries for topological data analysis, Scikit-TDA [66].

2.3 Previous Applications of Topological Data Analysis in Chemistry

Within materials science, persistent homology has already found a wide range of applications. Recent work includes pore-geometry recognition [67, 68], where persistent homology was shown to be a powerful descriptor for both pore shape and size in crystalline materials. This success is perhaps unsurprising - homology is the study of holes,

after all. However, the persistence descriptor described in this work is dependent on the size of a hypothetical probe, and it is this feature that makes the work so powerful. Probes of different size can be used, enabling materials to be screened for their efficacy at storing different molecules. Persistence has also been used to study covalent networks supported on crystalline materials [69], and used to create a descriptor quantifying the homogeneity of the network. Furthermore, persistent homology has been used to create topological descriptors for fullerene stability [70]. Persistence landscapes have also been used to study models for phase transitions [71], and were able to study topological properties of the configurational space. There is a wide body of work studying the persistent homology of time-series [72, 73], but there does not yet appear to be any application of this to chemical systems, particularly lose with interesting underlying dynamics.

A large body of work applying persistent homology to materials has been carried out by Hiraoka *et al*. This work includes the detection of different phases of materials [74, 75, 76], and order-disorder transitions [77]. A recent review of this work has been written [78]. In contrast to the work discussed previously on crystals, the methods developed by Hiraoka have found use studying amorphous and glassy materials. Although the studies performed in this work are similar to those discussed, it is noted that Hiraoka *et al* study single frames of simulation, and draw conclusions from lone persistence diagrams - such as relating curves of persistent points to geometric constraints of the system. In this work, efforts are made to use average properties of persistence diagrams.

There have also been many successes in the application of persistent homology based techniques within the study of biomolecules - a field suggested as an area of study with topological data analysis early in its lifetime [79]. Guo-Wei Wei and colleagues have developed topological descriptors for protein rigidity, based on barcode lengths [80, 70, 81]. They have also developed various filtration parameters (such as normalisation of distances by charge) which have proven useful in deep learning frameworks to predict biomolecular properties [82, 83, 84], or a classifier of proteins [85]. A review of these methods can be found in [86]. Furthermore, their element-specific persistent homology has been shown to be suitable for protein-ligand interactions [87]. Outside of this group, Emmett *et al* used a more classical persistence approach to characterise the scale and conformation of loops in protein folding [88, 89], and derived persistent homology variables have been related to quantities such as protein compressibility [90]. Folding pathways have been analysed using mapper, elucidating transitions between low-density states [91]. Also, two-dimensional persistence has been successful in virtual screening, by studying filtrations of both Rips complexes and atomic charges [92, 93].

Persistence landscapes have also had a series of applications to biomolecules, in particular to protein binding [94, 95]. Using the well-defined statistics on persistence landscapes, a simple classifier was able to detect conformational changes of a 370 amino acid maltose-binding protein, from its open to closed form. Furthermore, using short-loop [96] software designed to recover the shortest basis for first degree homology classes, the

authors were able to identify the active site residues themselves. Similar work has also been performed by Haspel *et al* [97]

Persistent homology has been previously used to study water networks. Xia *et al* used persistent homology to analyse the difference of water networks of simulated NaCl and KSCN solutions [98], and were able to relate features of persistent barcodes to two morphological types of aggregation. Recently, this work has been extended to osmolyte solutions [99]. Weighted persistent homology has also been developed in a similar manner to the element specific persistent homology mentioned earlier, and has been used to study hydrogen bonding networks [100, 101]. For this work, the average persistence entropy was used to take into account statistical fluctuations of the simulated systems. However, such properties are difficult to interpret and relate back to network properties of the water.

Interestingly, chemistry has also inspired work in TDA itself - with the nudged elastic band method being used to create filtration functions by altering the elastic band hyperparameters [102].

Chapter 3

Topological Data Analysis of Chemical Space

3.1 Introduction

3.1.1 What is Chemical Space?

Chemical space is a particularly broad concept. A general definition, found in [103], is 'the ensemble of all possible molecules'. Even this broad definition can still lead to useful notions. For example, the process of lead optimisation within drug discovery can be considered to be the creation of a path in chemical space. An understanding of chemical space could therefore prove a powerful tool for chemists.

To make progress with understanding chemical space, two further problems can be defined:

1. What molecules are possible?
2. How are these molecules described?

Estimates for the number of drug like (i.e. obeying Lipinski's rule of five [104]) have been as high as 10^{60} [105]. This number is clearly far too high to be of practical use with modern computing technology. In practice, chemical space is restricted, depending on the nature of the problem that is being studied - chemical space is rarely studied in its entirety. This is something that was earlier explored in Chapter 1, when discussing databases. Furthermore, the same set of molecules can lead to an entirely different space depending on how they are described. For example, a chemical space of molecules described by their element counts will contain different information to a space of the same molecules defined by their molecular shape. Different descriptions may also yield

spaces with different mathematical properties - it could be a vector space, a metric space, or even just a topological space. Again, the description of the space will depend on the exact nature of the problem. This work will use the phrase 'chemical space', regardless of these factors.

This chapter will explore two fundamentally different problems with chemical space. The first problem is the use of pictures of chemical space to improve solubility prediction. This task will utilise a description of chemical space constructed from molecular graph descriptors. Topological data analysis will be used to create pictures of the chemical space, and insight from these pictures will be used to create models for solubility. The second problem is that of using topological data analysis to create a metric space of molecules. In particular, topological data analysis is utilised to create a 'chemical shape space', where proximity between molecules implies similar shapes. The shape space will then be related to other chemical descriptors.

3.2 Description of Data Sets

There are two data sets of small organic molecules that are used in this chapter. The first data set originates from a study in predicting water solubility, by Wang *et al* [106]. This data set consists of various subsets of solubility data, that were collated for Wang's work.

1. The low molecular weight subset of Delaney's ESOL data set [107] ($n = 1312$)
2. The Huuskonen data set [108] ($n = 354$)
3. Hou data set [109] ($n = 25$)
4. Personal Correspondence from Wang ($n = 9$)
5. Jain and Yalkowsky's data set [110] ($n = 545$)
6. A subset of the Beilstein data set [111] ($n = 1210$)
7. The solubility challenge data set [112] ($n = 90$)
8. A set of molecules with experimentally determined melting point, from literature (references 20-29 in [106]) ($n = 119$)

A full description of this data set can be found in [106] and the supporting info of that publication. Originally, Wang provided a set of molecules labelled by source, with the Sybyl Line Notation (SLN) [113] used as a chemical identifier. Each molecule also had a measured solubility value ($\log S$), and the predictions from the various models presented

in [106]. This data set is used as an example of data that may be used in a project predicting a molecular property, and will be referred to as the Wang data set.

The second data set is used as an example of a data set that may be used in a molecule generation task. In 2014, Ruddigkeit *et al* performed a study on the structure of what the authors termed 'fragrance like' chemical space [114]. Using the SuperScent data base [115], the authors determined the following properties as classifying a molecule as 'fragrance like':

- Fewer than 21 heavy atoms
- Only containing carbon, hydrogen, oxygen and sulphur atoms
- The total number of oxygen and sulphur atoms combined is less than 3
- At least one hydrogen bond donor

the authors then generated a series of subsets of large chemical databases, containing only the fragrance like molecules. In this work, the subset used is that of the ChEMBL data base [7] and will be referred to as the ChEMBL-FL data set.

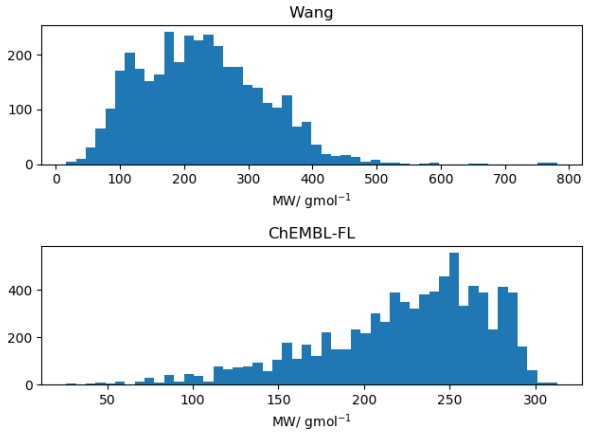
3.2.1 Comparison of Data Sets

A series of summary statistics for each data set can be seen in Table 3.1. The Wang data set contains fewer molecules, but a wider range of atoms. This discrepancy is due to their original purposes. The Wang data set was constructed as a data base for solubility prediction. This would need a diverse data set, to lead to the model with the widest domain of applicability (ignoring issues regarding the performance of the models themselves). In contrast, the ChEMBL-FL data set was designed to create a set of molecules which could be used in a molecule generation task. The main drawback of this data set for such a task would be that it is unlikely that any molecule generator trained on this data set would be able to generate a molecule that contains a phosphorus atom, as there are no phosphorus atoms in the rest of the data set.

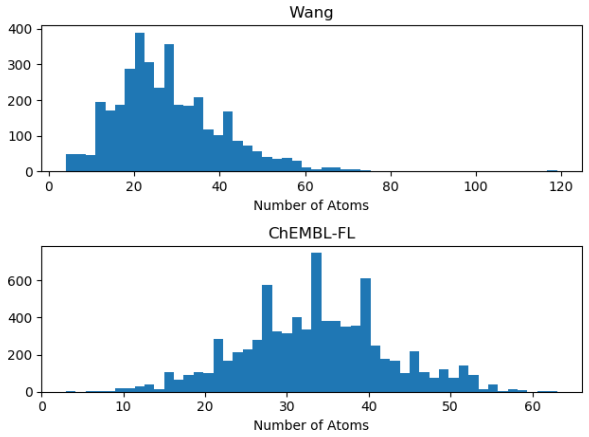
	Wang Data Set	ChEMBL-FL Data Set
Number of Molecules	3664	8144
Atoms in Data Set	H, C, N, O, F, P, S, Cl, Br, I	H, C, O, S
Mean Molecular Weight /gmol^{-1}	223.94	227.17
Average Number of Atoms	33.43	28.09

TABLE 3.1: Comparison of Wang and ChEMBL-FL data sets.

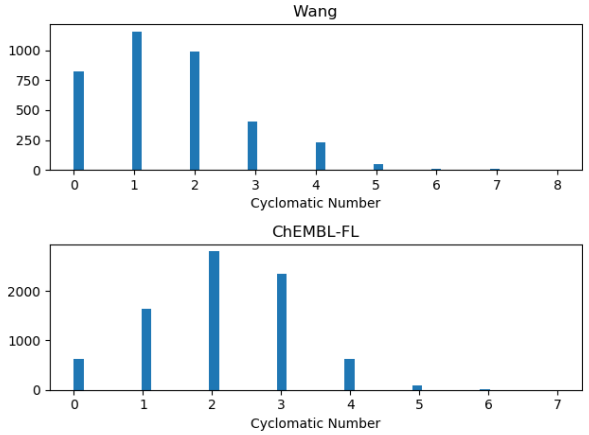
The distributions of molecular weight, number of atoms, and number of cycles can be seen in Figure 3.1. It is clear that the Wang data set is the more diverse, with longer tailed distributions. In contrast, the ChEMBL-FL data set is less diverse.



(a) Molecular Weight



(b) Number of Atoms

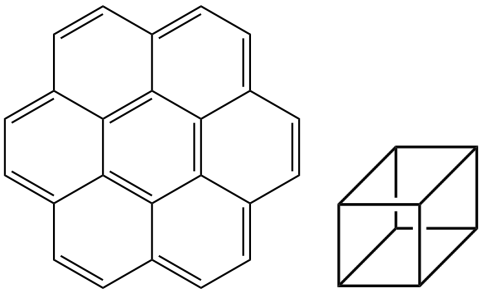


(c) Number of Cycles

FIGURE 3.1: Distributions of commonly used descriptors for both the Wang and ChEMBL-FL data sets. The differences in distributions are largely down to the original purposes of the data sets.

3.2.1.1 How is the number of cycles in a molecule calculated?

For most molecules, the question of how many cycles are contained is 'obvious' from the molecular graph. For example, phenol contains one cycle. However, this is not always the case, particularly with highly symmetric molecules, or molecules with fused rings. Take the examples of coronene and cubane, seen in Figure 3.2. There are several cycles in coronene - depending on how a cycle is defined. For example, if a cycle is defined as a closed path on the molecular graph, there are several cycles with six nodes. There are also cycles that could be made by walking around larger closed paths, such as those created by fused rings. Cubane can be seen to have a similar problem, in that the number of cycles should be unambiguously defined. Historically, algorithms such as the Smallest Set of Smallest Rings (SSSR) algorithm have been used to find the number (and location) of rings in a molecule [116]. However, it has been noted that this set of rings is not unique, and often does not match 'intuition'. For example, the SSSR of cubane contains five elements, where it may be expected that there are six. Furthermore, there are clearly several options for the elements of the SSSR (although this is not a problem within this work, where only the number of cycles is important as opposed to whether a given atom is an element of a cycle). It is therefore important to discuss what is meant



(a) Coronene (b) Cubane

FIGURE 3.2: Chemical graphs of coronene and cubane molecules. Due to differences in the definition of 'the number of cycles', a user may get an unexpected result.

in this work by 'the number of cycles' in a molecule, when using it as a descriptor. The following properties are defined:

- SSSR - The number of elements of the smallest set of smallest rings
- SSSRsym - The number of elements of the symmetric smallest set of smallest rings (from the RDKit 2019.a implementation [117])
- nCIC - The cyclomatic number of the molecular graph. This is equal to its first Betti number
- nCIR - The number of circuits of the molecular graph. This is equal to the number of all closed, self avoiding walks on the graph

For the ChEMBL-FL data set, these descriptors were calculated (SSSR and SSSRsym in RDKit, nCIC and nCIR from DRAGON [118]). Pairwise comparisons between these descriptors can be seen in Figure 3.3.

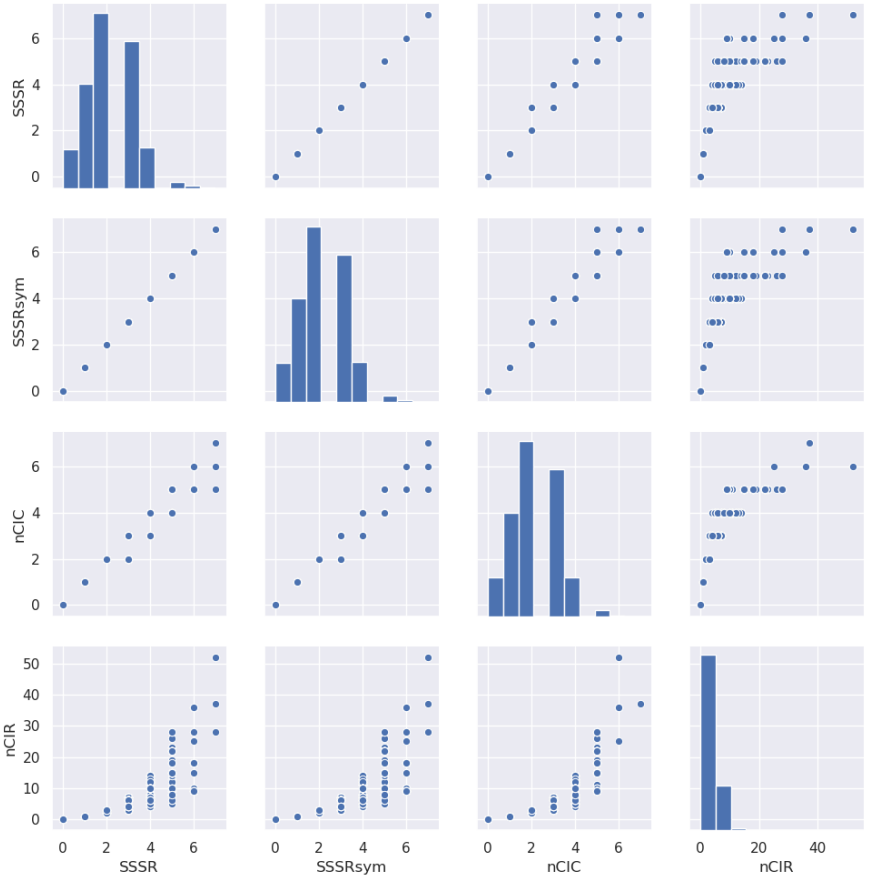


FIGURE 3.3: Pairwise comparison of descriptors used to calculate the number of cycles in a molecule. The diagonal elements show a rough histogram, with off-diagonal elements showing scatter plots of pairs of descriptors. Although correlated, it is clear that different definitions for the number of cycles lead to different results.

The nCIR descriptor shows the most deviation from the other descriptors. This is the expected behaviour, as it has been seen that fused rings lead to a larger number of circuits. For the ChEMBL-FL data set, SSSR and SSSRsym are identical - this suggests that there are no cubane-type molecules in the data set. nCIC is always the lowest descriptor, with the other descriptors always being equal or higher than nCIC. This is actually unexpected, as one would expect SSSR to contain the smallest set of rings. It is unclear if this is an implementation issue.

As nCIC is a topological descriptor of molecular graphs (i.e. it is the first Betti number, which for connected graphs is equal to $n_{edges} - n_{nodes} + 1$), it is the descriptor that is used for this work when referring to 'the number of cycles'.

3.3 Mapper Algorithm on Descriptor Space

3.3.1 Solubility Modelling

As a large part of this work was designed to improve the accuracy of solubility modelling, it is important to discuss the potential trials and pitfalls of the field. Given a molecular structure, is it possible to accurately predict its solubility in water? This is an important test in the drug discovery process, as it has previously been estimated that up to 40% of drug discovery programs fail due to an issue with bioactivity, including solubility [119].

To this end, many different methods have been calculated to predict solubility. Hussko-nen used linear regression, and early neural networks, to predict solubility [108, 120]. The earlier work used a set of neural networks on a data set of 211 compounds. Descriptors were a combination of electronic and topological, and included simple hydrogen bond donor/acceptor counts. This method was able to achieve an R^2 of 0.86, with a standard deviation of 0.5 log units on the test set. Later, this method was extended to a larger data set of 1297 compounds, and although the standard deviation increased to 0.6, the predictive R^2 improved to 0.92. A standard multilinear regression model using the same descriptors performed comparably ($R^2 = 0.88$, $\sigma = 0.71$). From this, it can be concluded that more sophisticated solubility models were unable to outperform more simple ones - a theme that has occured many times within this field.

The general solubility equation (GSE) was developed in 1980 to be a widely applicable, minimal parameter model [121]. These descriptors were melting point T_m, $\log P_{octanol}$ ($\log P$), and $\Delta_{fusion}S$. The original model was moderately successful, however required the calculation of $\Delta_{fusion}S$, which to this day is difficult to do with high accuracy. Therefore, an alternative form of the GSE was proposed by Jain and Yalkowsky in 2001 [110]. This was able to obtain an R^2 of 0.97, and an absolute average error of 0.45 on their own set of 580 compounds.

Alternatively to the pure informatics-based approach previously mentioned, Jorgensen and Duffy attempted to build a solubility model using parameters calculated by simulation [122]. These were derived from Monte Carlo simulations of single solvent molecules in a cube of 500 TIP4P molecules. This allowed the calculation of parameters such as interaction energies and solvent accessible areas. Further, hydrogen bonds were able to be directly measured, through geometric constraints. Their model was able to achieve an R^2 of 0.82 on their data set of 150 compounds. There have been attempts to compare both the informatics-based GSE and the MC based method from Jorgensen and Duffy. In 2001, Ran and Yalkowsky used the revised GSE to estimate the solubility of the same molecules as Jorgensen and Duffy [123]. This led to a smaller average absolute error (0.43 compared to 0.56).

In 2004, Delaney published ESOL [107]. This was a 4 parameter model ($\log P$, molecular weight MW, percentage of aromatic atoms AP and rotatable bond numbers RBN), resulting in the following functional form:

$$\log S = 0.16 - 0.63 \log P - 0.0062MW + 0.66RBN - 0.74AP \tag{3.1}$$

On a set of 3402 molecules, an R^2 0f 0.69 was achieved. Although perhaps a less accurate model than those previously discussed, ESOL was designed to be accurate over a wider range of molecules, and therefore be more applicable to new molecules. When the revised GSE was applied to this data set, it was also found to have an R^2 of 0.67. Further analysis showed that the GSE outperformed ESOL for smaller molecules, however for molecules with molecular weights in excess of 300gmol^{-1}, ESOL was the better model. This emphasises the difficulty in creating a single model that performs well on a wide range of chemical space.

In their review of the field, Jorgensen and Duffy stated that the average uncertainty in experimental $\log S$ measurements was no better than 0.6 log units. For example, rotenone has been measured as having $\log S$ values of -4.42 [108], and -6.29 [124]. Additionally, when guanine had its solubility measured as -3.58, it had been found difficult to accurately predict [25], but when this experiment was repeated and $\log S$ measured to be -1.86, this value was found easier to model [125]. This reflects the inherent difficulty in solubility prediction, as well as the difficulty in obtaining the full provenance of solubility data. Here, it is simply noted that all of the previously noted models were within or close to this 0.6 log unit measurement of accuracy, and that work needs to be carried out to determine whether 'better quality' predictions are simply fitting to noise in measurements.

To determine the state of the field, the 'Solubility Challenge' was created [112]. A data set containing 100 measured solubilities was released, alongside 32 structures with unknown solubilities. Researchers were invited to construct a model for solubility based on the 100 known compounds, and to submit predictions for the solubility of the oither 32. The findings were announced the following year [126], with eight major conclusions:

1. Defining a prediction as 'correct' if the calculated value of S is within 10% of the measured value, the entrants were correct between $0 - 33\%$ of the time.
2. The R^2 value on S for the test set ranged from 0 to 0.642.
3. Defining a prediction as 'correct' if the calculated value of $\log S$ is within $0.5 \log S$ of the measured value, the entrants were correct between $15.6 - 62.5\%$ of the time.
4. The R^2 value on $\log S$ ranged between 0.018 and 0.650

5. No entrant discussed the possibility of polymorphism, or predicted solubility for more than one polymorphs of a compound. This has been shown to be important, such as in [127].

6. Entrants tended to more accurately predict the solubility for an intermediate range of compounds ($\log S$ between 0.5-3), than they did for compounds of high or low solubility.

7. The accuracy of prediction was inconsistent between molecules. The easiest molecule to predict (imiprimane) was predicted correctly by 81% of entrants. In contrast, napthoic acid had a correct prediction from only 37% of entrants.

8. Within the intermediate range, some molecules were still difficult to predict. Both probenecid and indomethacin were predicted correctly by $< 5\%$ of entrants.

The original creators of the solubility challenge did not release any specifics of individual entrants. However, entrants were free to submit their models now the results were known. Hewitt *et al* submitted 4 entrants to the challenge [128]. These approaches included linear regression, neural networks, and category formation (alongside some commercially available techniques). It was found that the linear regression was the best performing model. The neural network was found to have overfitted to the training data - which is unsurprising due to the low number of data points in the set. The category approach also performed poorly, again due to the lack of data, Three categories contained fewer than ten compounds, and Hewitt *et al* concede their categories may have been too broad. The failure of the commercial models was attributed to their training sets - which were thought to be more likely to contain a wider variety of molecules than the druglike compounds of the solubility challenge set. However, as these training sets were not publicly available, it is hard to agree with these conclusions.

Hewitt *et al* conclude that current methods likely do not fail due to inadequate methodology, but instead an insufficient appreciation for the complexibility of the solubility process. They make a set of recommendations:

1. Simple modelling approaches should be preferred to more complex ones - particularly with small data sets

2. Knowledge of data quality should be more prominent in model design and publication

3. The applicability domain of a model is information that should be included in all published models, including those commercial packages

4. Available solubility models, although imperfect, are useful for initial screening.

- If more accurate results are desired, an ensemble of models may allow the user to estimate reliability with more confidence

5. Despite current models being of use in screening, high-quality predictions are as yet unavailable. Alternative non-statistical models could be explored more, such as those including a mechanistic reasoning

The poor results of the solubility challenge did not deter the creation of new models. Lusci *et al* designed a molecular graph-based neural network [129]. It was suggested that this approach could lead to interpretation of functional groups that improve or hinder solubility. Although the average absolute error was 0.43, it has to be concluded that the model was fitting to noise, due to the aforementioned high errors in experimental measurement. However, this approach could be useful for other prediction tasks.

In 2014, Palmer and Mitchell investigated whether it was indeed this experimental noise that made solubility challenging to predict, or problems in the approach itself [130]. The solubility challenge data set was used due to its high accuracy (widely accepted to be 0.05 log units), and compared to noisy data (up to 0.6 log units error), taken from various literature sources. The authors argued that it is not the data itself that was flawed, but instead the descriptor sets themselves, and concluded that this was because the high-accuracy data was still difficult to consistently predict. However, this data set is considerably smaller, and this effect was not properly considered.

A recent study performed by Boobier *et al* studied the proficiency of consensus-type prediction for these problems [131]. It was found that a consensus prediction outperformed all individual predictors on a relatively small (100 molecule) set. However, it was found that a consensus of human predictors, found by a survey, was able to outperform all of the machine learning models. This suggests that a consensus type prediction is potentially a useful tool in solubility prediction. Furthermore, there is perhaps a problem with the descriptors currently being used, as they are not able to describe information that is known to a consensus of human predictors.

One of the main criticisms of solubility prediction is that the biomolecular environment is not pure water [132]. Despite attempts to construct modified media to better reflect the cellular environment (such as in [133]) *in silico* approaches are almost entirely based on the solubility of neutral compounds in pure water. This is likely due to the difficulty of modelling the influence of additives, such as phospholipids, on solubility. This work does not consider these more complex properties.

To commemorate ten years since the original solubility challenge [112], a new solubility challenge has been developed [134]. Participants in the new challenge were invited to use their own training sets to predict $\log S$ on two new test sets, one with solubility values measured with high accuracy, and the second data set with an average standard

deviation of 0.6 log units. At the time of writing of this work, this second challenge is still underway.

3.3.1.1 Solubility as a Thermodynamic Cycle

The solvation of a molecule can be considered as a thermodynamic cycle, such as in Figure 3.4.

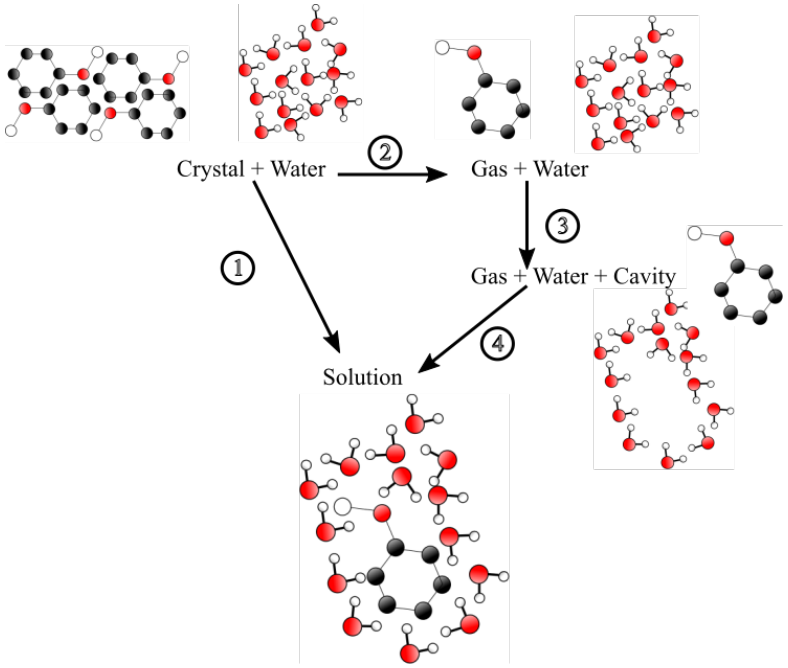


FIGURE 3.4: The thermodynamic cycle of the solvation of a molecule in water. The various steps are illustrated by number: total solvation, molecular dissociation, cavity formation and cavity hydration.

1. Total solvation
2. Dissociation of the molecule from the crystal
3. Formation of a cavity in the solvent
4. Hydration of the molecule in the cavity

Steps (3) and (4) are often considered to occur together, and are collectively known as 'hydration'. Methods have been designed to calculate each step in this cycle individually, such as the work performed by Palmer et al in 2007 [135]. The free energy was related to the solubility by the following:

$$\Delta_{sol}G^{\ominus} = \Delta_{sub}G^{\ominus} + \Delta_{hyd}G^{\ominus} = -RT\ln(S_0V_m) \tag{3.2}$$

Where V_m is the molar volume of the crystal, and S_0 is the intrinsic solubility, in molL^1. By selecting single polymorphs from the Cambridge Structural Database, the various $\Delta G^{\ominus}$ were estimated, and solubility calculated for a set of 60 molecules (34 training). An R^2 of 0.83 was achieved (RMSE = 0.63), reflecting the potential for such methods in the field. In particular, if better estimates for the various ΔG were obtained, such a model would have a strong predictive and explanatory power. Luder *et al* released a series of papers in 2007, aiming to better calculate these free energies [136, 137, 138], in particular aiming to quantify the difference between solvation of crystals and amorphous structures. It was concluded that the electrostatic interactions are much larger in the crystalline materials, whereas it was the Lennard-Jones type interactions that were important for the amorphous structures.

3.3.2 Networks

For all of the presented networks in this section, the colour gradient is such that: *blue* $\rightarrow$ *red* $\implies$ *low* $\rightarrow$ *high*. All networks were made with the Ayasdi implementation of mapper. It is important to emphasise that mapper networks are topological in nature, conclusions should be made only about properties of connectivity, rather than distance. In essence, networks should be considered to be infinitely flexible.

A series of mapper networks, created with various combinations of filter functions and metrics, can be seen in Figure 3.5. The choices of lens and metric are as follows:

(a) 2D MDS lens, correlation metric

(b) 2D Neighbourhood lens, Euclidean metric

(c) 2D PCA lens, variance normalised Euclidean metric

Each of the networks looks markedly different - as expected with mapper. As mentioned in the Theory of Mapper section (Chapter 2), it is important to focus on the analysis of the observations which are consistent between networks. Henceforth, all conclusions in this work can be assumed to be independent of choice of input parameters, but the network studied will be the MDS lens, Figure 3.5(a).

The important observations can be summarised as being links between network location and chemical descriptors, and what can then be inferred about solubility. The two most obvious descriptors, are molecular weight and number of cycles. The MDS lens, coloured by these features, can be seen in Figure 3.6.

There is a clear gradient for these colourings. The molecular weight varies smoothly across the network, whereas the number of cycles creates clear groupings of nodes. In particular, nodes can be classified as follows:

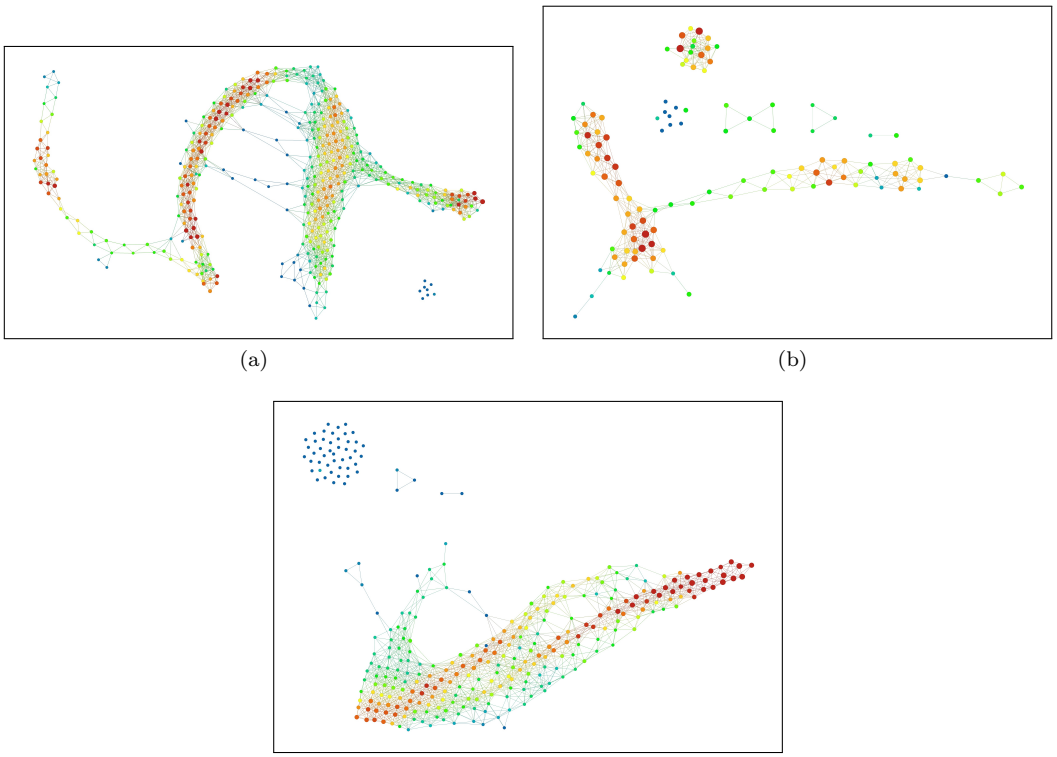
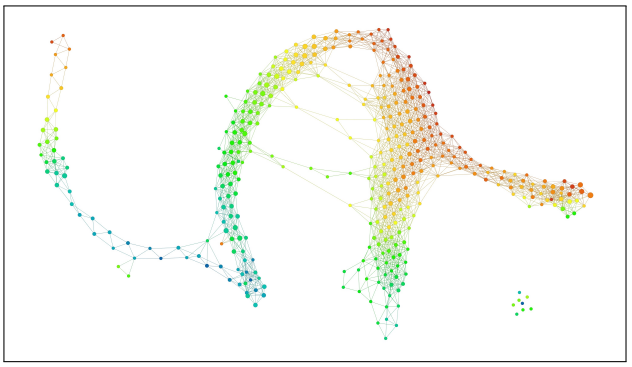


FIGURE 3.5: A series of mapper output networks. Full details as to the parameters used for each network can be found in the main text. Networks are coloured by the number of observations (molecules) in each node. Boxes are used to make clear which outliers belong to each network. The hyperparameters for the mapper algorithm lead to markedly different networks. The gradient is such that blue to red is equivalent to low to high.

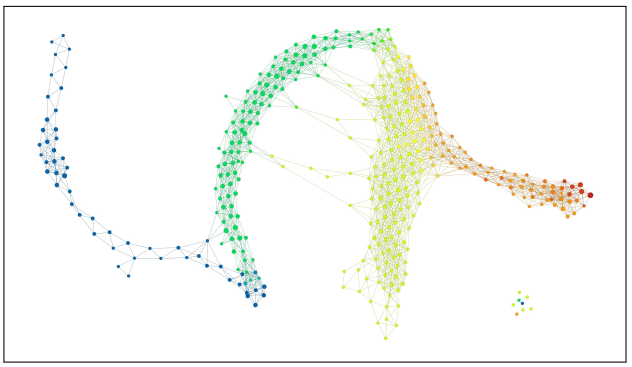
- 0 cycles (blue)
- 1 cycle (green)
- 2 cycles (yellow)
- 3 or more cycles (red)

It is perhaps expected that the mapper algorithm would create a network with these colour gradients, as these two variables are expected to be well correlated with a wide range of other descriptors. Interestingly, the outliers in the mapper network all seem to have intermediate molecular weights, but vary in the number of cycles. In general, statistical tests such as the Kolmogorov-Smirnov (KS) test can be used to determine what features separate different subgroups of nodes. This has been done with other studies of the mapper algorithm [62, 139, 140] but this is not the focus of this work.

Instead, the goal of this project is to improve solubility prediction, using insights gained from mapper networks. In this way, topological data analysis can be considered to be describing how to *look* at the data, rather than directly answering questions. The MDS



(a) Coloured by molecular weight



(b) Coloured by number of cycles

FIGURE 3.6: The MDS mapper network, coloured by molecular weight and number of cycles. There are clear trends with respect to these variables, illustrating that the mapper networks contain chemically useful information. The gradient is such that blue to red is equivalent to low to high.

network, coloured by $\log S$, can be seen in Figure 3.7. Although the $\log S$ variable was not used in the network's construction, there is a clear trend, with similar solubilities tending to be grouped together. This is again expected, as there is a reasonable correlation between molecular weight and solubility. It is also noted that there appears to be no relationship between the outlier nodes and their solubilities, suggesting that the properties that separate them from the main network are independent of this.

The trend in solubility is not consistent throughout the mapper network, it does not have an unbroken colour gradient. Perhaps the most prominent breaking of this trend is in the group of molecules with two cycles, which have a much lower solubility than their neighbours. Analysis of this subgroup via KS testing did not yield descriptors that were thought to be chemically relevant (in particular, they tended to be the more esoteric descriptors calculated by DRAGON). However, the breaking of the gradient in solubility is not too dissimilar to one found in the network coloured by molecular weight. Therefore, it was decided to study the effect of halogen substitution, as this leads to larger molecular weights for molecules with similar graphs.

The MDS network, now coloured by the number of chlorine atoms, can be seen in Figure 3.8. The anomalous trend in $\log S$ previously mentioned is almost entirely matched by

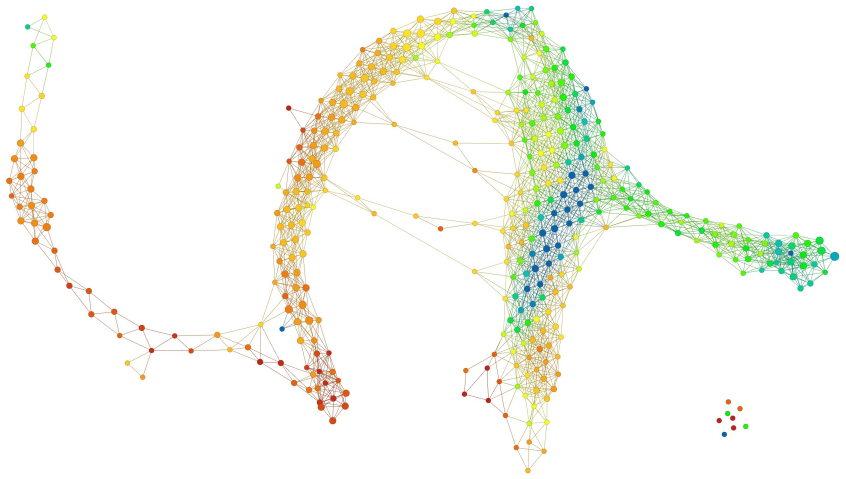


FIGURE 3.7: The MDS mapper network, coloured by $\log S$. The gradient is such that blue to red is equivalent to low to high.

nodes with high chlorine numbers. This effect is only seen for systems with two rings - for systems with fewer than two cycles the presence of chlorine atoms does not particularly impact solubility. Each outlier node also contains no chlorine atoms - but as mentioned it is difficult to say if this is what truly separates the outliers, as the same property is seen for the number of bromine descriptor, amongst others.

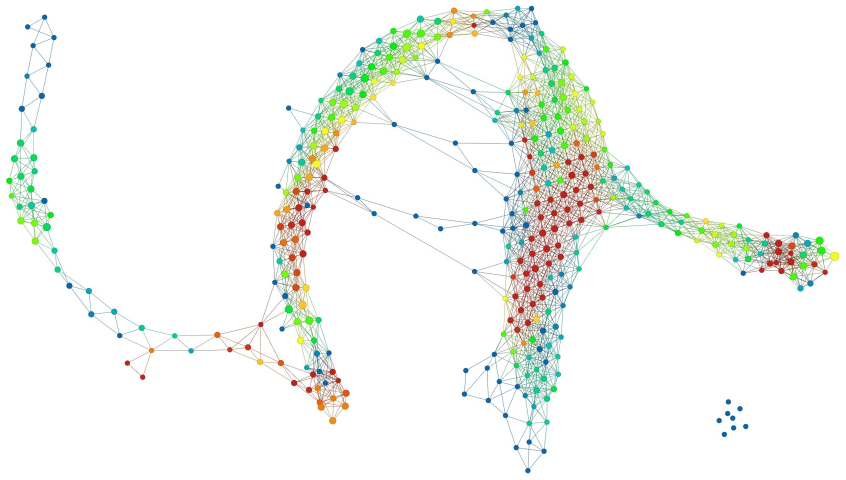


FIGURE 3.8: The MDS mapper network, coloured by the number of chlorine atoms. The anomolous region of low solubility for cycles with two molecules clearly correlates with a region of several chlorine atoms. The gradient is such that blue to red is equivalent to low to high.

The network allows inference of information regarding solubility. For small to medium sized molecules, the gradient in solubility is largely matched by that of molecular weight. This implies that solubility and molecular weight are highly correlated for these molecules. However, as molecular size increases, other features become important. For systems with two cycles, the presence of chlorine appears to impact solubility more than

with any other number of cycles. The effect of chlorine substitution on solubility has been well studied (see, for example, [141]), and it is surprising that the effect is only seen here for systems with two cycles. For example, an analysis of the solubility data sets of benzene and chlorobenzene found in [142] shows that no data source predicts chlorobenzene to be more soluble, with chlorobenzene being at least 0.67 log units less soluble. It is therefore unclear why the effect is more prominent here for those molecules with two cycles. To see if this is due to a sampling effect, it is important to understand the distribution of chlorine numbers, as a function of cycles. This can be seen in Figure 3.9.

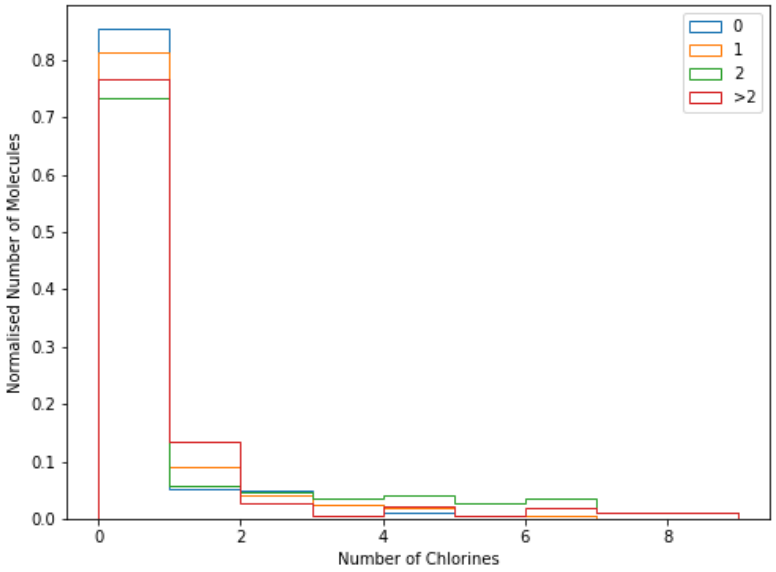


FIGURE 3.9: The distributions of chlorine numbers as a function of the number of cycles. As this distribution is consistent between the number of cycles, it can be said that the effect seen on solubility is not simply due to differences in chlorine distributions.

The vast majority of molecules do not contain any chlorine atoms. Furthermore, the distributions are relatively consistent, independent of the number of cycles. This implies that the different behaviour observed from the mapper network regarding the relationship between chlorine and solubility is likely a real effect.

It is worth noting that although the mapper network was constructed independent of solubility, conclusions can be made about the behaviour of this property. In principle, this could therefore be achieved for any useful chemical property.

3.3.3 Creation of Models

A standard model for solubility prediction is Delaney's ESOL model [107]. ESOL predicts solubility to within the limits of the previously discussed experimental measurements - see Chapter 3.3.1 for an introduction. As a reminder, for the molecule with index i, ESOL calculates $\log S$ through a standard linear model:

$$\log S_i = \beta_0 + \sum_j \beta_j X_{ij} \tag{3.3}$$

where the set of descriptors is CLogP, Molecular Weight, Rotatable Bond Number, Aromatic Ratio. The original β parameters for ESOL can be found in Equation 3.1

The mapper network can provide useful insight when constructing machine learning models for chemical properties, in this case solubility. Firstly, the strongest gradients in the network are that of molecular weight, and the number of cycles. This matches 'chemical intuition', and there is no model that would not take these two descriptors into account in some fashion. Within ESOL, molecular weight is directly taken into account, whereas information about cycles is held within the rotatable bond number and aromatic ratio.

The mapper network suggests that there is a relationship between the number of chlorines and solubility, which varies as a function of the number of cycles. Therefore, it was chosen to create models which were able to take this relationship into account. Using ESOL as a base model, a linear model was developed of the form found in Equation 3.3, where $X \in \{$MLogP, Molecular Weight, Rotatable Bond Number, Aromatic Ratio, Number of Chlorines$\}$. MLogP was used rather than CLogP due to its availability in DRAGON.

This linear model yielded an RMSE of 0.56 on the Wang data set. This appears to be reasonable, and considering it is around the experimental uncertainty of solubility measurement, it might be unclear as to why this model may require improvement. However, a problem with this model becomes clear when looking at the residuals of this model:

$$\text{residual} = \log S_{experiment} - \log S_{model} \tag{3.4}$$

As a function of $\log S$, the residuals for this model can be seen in Figure 3.10. For larger values of $\log S$, the model seems to over- and under-estimate the true value with equal probability. However, as $\log S$ is reduced, the distribution of residuals becomes skewed, with over-estimates for $\log S$ becoming more likely. Mapper networks imply that it may be sensible to group the molecules by the number of cycles. Here, the classes used are the ones defined in the previous section:

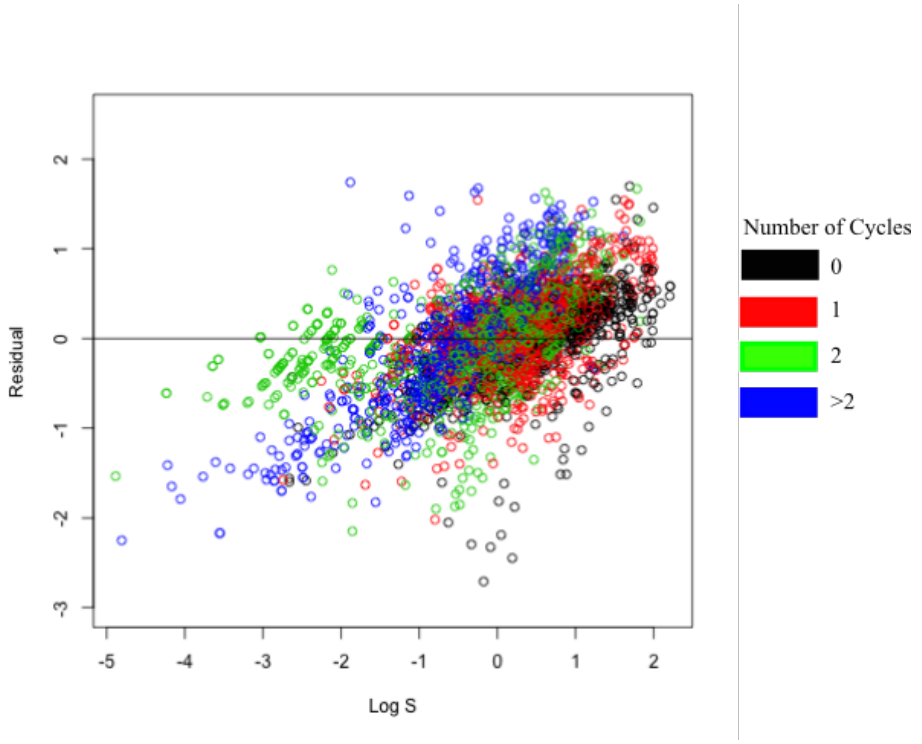


FIGURE 3.10: Residuals vs $\log S$ for the linear model. Points are coloured by the number of cycles in the molecule. Using the results of the previous mapper analyses, it is possible to target specific regions of the residuals to improve - such as those molecules with two cycles.

Number of Cycles	Number of Molecules
0	820
1	1157
2	988
> 2	698

TABLE 3.2: The classes used regarding numbers of cycles, as well as the number of molecules in each class.

Firstly, studies were performed understanding the behaviour of $\log S$ for the different classes. Figure 3.11 contains a series of box plots, used to study the distribution of the $\log S$ variable.

The distributions of $\log S$ are different, for different classes. This suggests that, rather than a linear model, a linear *mixed* model should be used. Whereas a linear model optimises the set of parameters β_i over all observations, a linear mixed model allows for a subset of β to be optimised for a set of classes individually. Such a model could be useful for hierarchical data, for example when assessing the performance of students over time. In this case, a linear mixed model would allow for each student to have a separate baseline performance β_0.

Firstly, as the mean values of $\log S$ vary for each class, a linear mixed model was created where the intercept was allowed to vary. This led to a slightly improved RMSE (0.550),

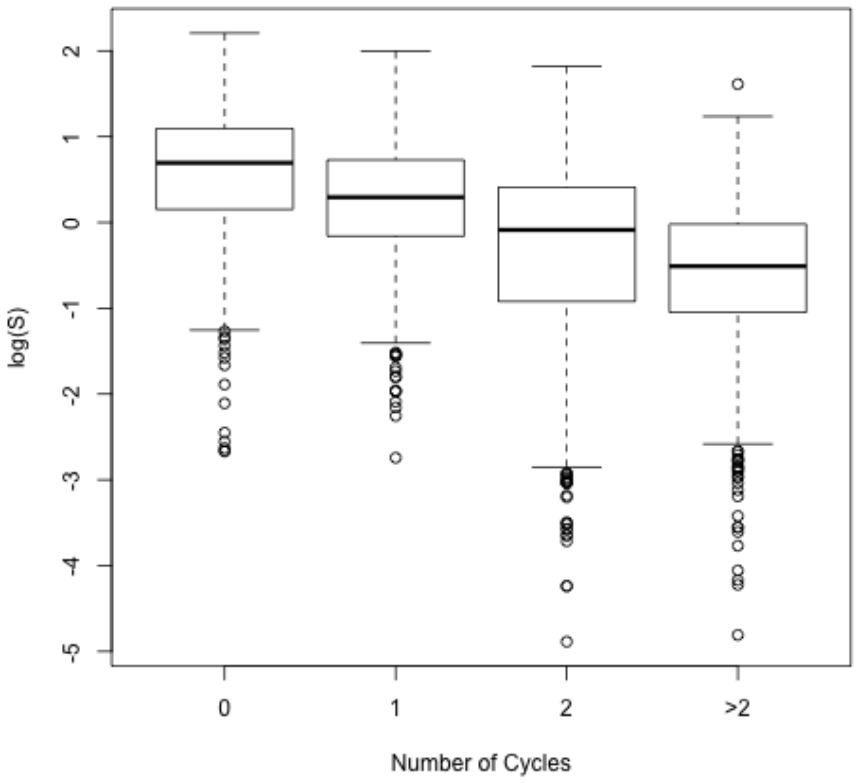


FIGURE 3.11: Box plots showing the distribution of $\log S$ when classified by the number of cycles in a molecule. As expected, the mean $\log S$ decreases as a function of the number of cycles. This suggests using a linear mixed model with varying intercepts.

where the intercept reduced as the number of cycles increased - matching the intuition that larger molecules tend to be less soluble. However, the residuals were not significantly improved (Figure 3.12).

However, the mapper network did imply that the dependence of solubility on the number of chlorine atoms varied with the number of cycles. Therefore, a mixed model was created where both the intercept β_0 and the coefficient on the number of chlorines β_{Cl} were allowed to vary between classes. This led to an RMSE of 0.544, and the residual behaviour of Figure 3.13. The residuals of the two cycle class in particular are improved, with the distribution being less skewed

β_{Cl} as a function of the number of cycles can be seen in Table 3.3.3. Regarding an error for β_{Cl}, within the linear mixed model framework errors on random effects such as this are ill defined, and instead it is advised to use the overall error as an estimate, which in this case is approximately equal to 0.05

In some cases, the value of β_{Cl} reflects what is seen in the mapper network. The presence of chlorine seems to have little or no effect on the solubility of molecules with one or

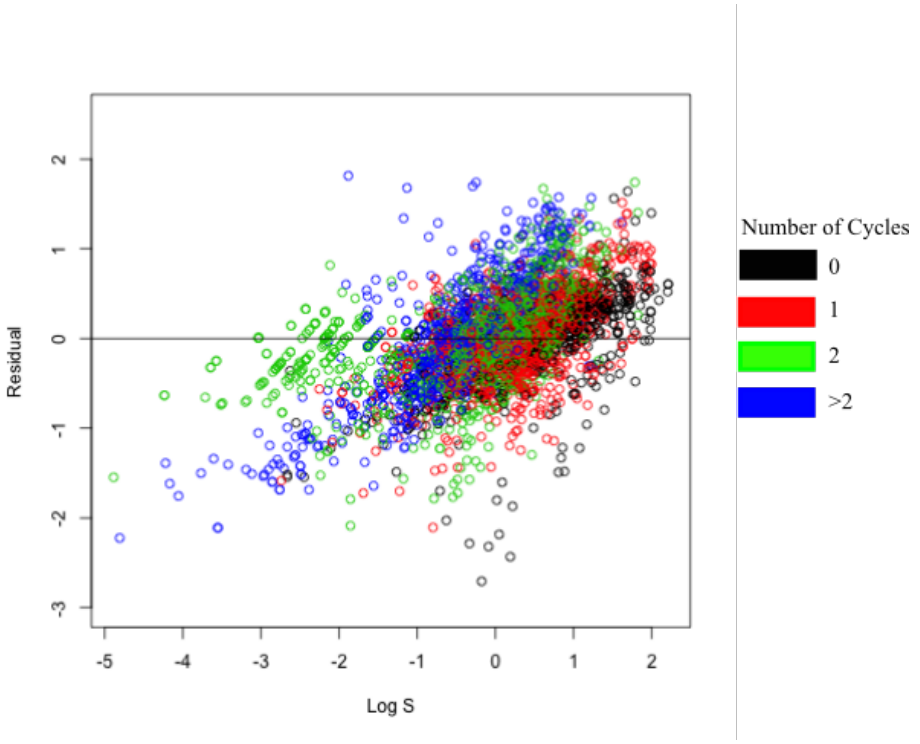


FIGURE 3.12: Residuals vs $\log S$ for the linear mixed model with random intercepts. Points are coloured by the number of cycles in the molecule. The residual distribution is largely unchanged when compared to the standard linear model.

Number of Cycles	β_{Cl}
0	0.00
1	-0.07
2	-0.22
> 2	-0.21

TABLE 3.3: The values of the coefficient for the number of chlorine descriptor β_{Cl} for the random intercept/chlorine coefficient model, as a function of the number of cycles.

fewer cycles. However, there is a negative correlation between the number of chlorines and solubility for larger molecules, implying the presence of chlorine hinders solubility, as expected for these molecules. This effect does appear to be equal in magnitude for systems with two rings as it is for systems with three or more rings. It is not thought that this effect would continue if the data set contained a larger quantity of systems with three or more rings. In particular, if it were possible to use a set with an appreciable number of molecules with this size, the correlation could disappear, as the size of these molecules becomes the dominant factor in their solubility.

Model comparison can be performed between the two linear mixed models created for this study. Using ANOVA it is possible to understand if differences in the residual sum of squares between these models are significant. This was performed using a χ^2 test, and was found to be significant ($p < 0.001$).

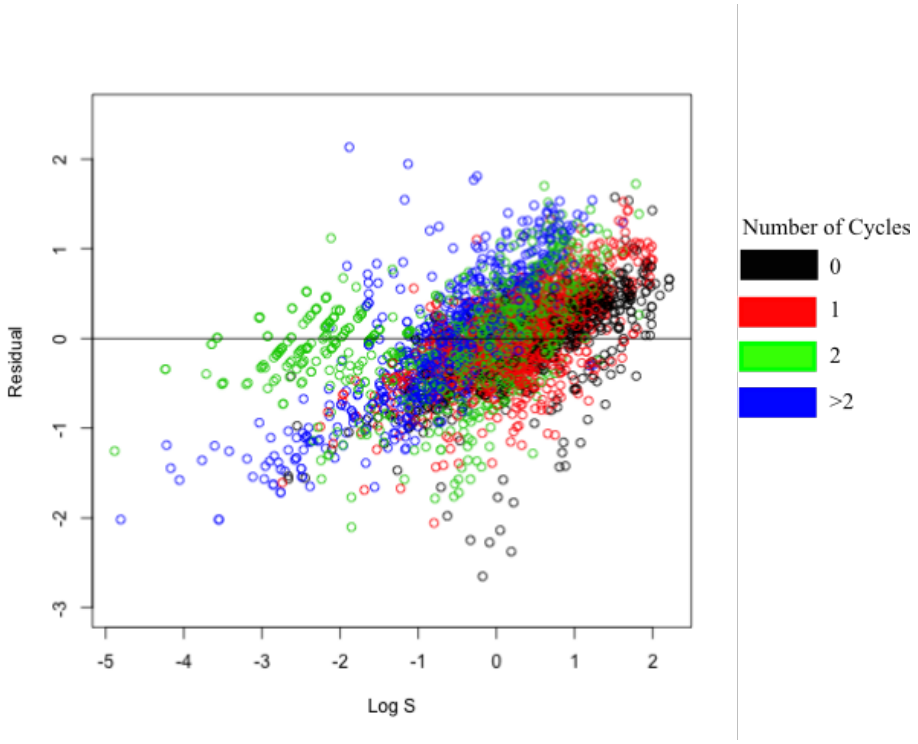


FIGURE 3.13: Residuals vs $\log S$ for the linear mixed model with random intercepts and chlorine coefficients. Points are coloured by the number of cycles in the molecule. The distribution of residuals for molecules with 2 cycles has now improved, and is less skewed to be negative when compared to the previous models.

As the final point of discussion regarding mapper and its use in creating more efficient models, it is noted that there are various descriptors that correlate well with the number of chlorines, and would therefore have a similar effect on these systems. An example of this is the mean atomic volume, which leads to a similar mapper output to Figure 3.8, and a similar behaviour of the correlation coefficient with linear mixed models. However, the purpose of this work was to study the information that could be gathered from a mapper analysis, and investigate its potential application to the design of QSPR-type models. It was found that there was an anomalous region of solubility in systems with two cycles, and that there was a strong correlation with that region with the number of chlorine atoms. When taken into account, this correlation was able to improve the residuals of these molecules specifically. Mapper can therefore be used as a reasonable feature selection tool, enabling the creation of better models.

3.4 An Atlas of Chemical Shape Space

3.4.1 Molecule Shape Similarity

The notion of molecular shape similarity is particularly important in drug design and discovery [143]. For example, the relative orientation between a ligand and target protein

greatly influences the binding affinity [144, 145]. Large databases can be screened to find molecules with similar shapes to an active compound, yielding potential candidate molecules [146]. Alternatively, scaffold hopping can use a shape similarity as a metric for similar purposes [147].

However, there are many potential candidates for 'shape similarity'. In general, these can be separated into those that require molecular alignment, and those that do not. Alignment based methods calculate the optimal superposition between two molecules, and use a metric based on this, such as the RMSD (defined in Chapter 4). These methods include ROCS, based on finding the maximum volume overlap between two molecules [148].

Non-alignment based methods do not need to calculate this superposition, instead using differences between intra-molecular descriptors (such as atomic distance distributions). Non-alignment methods include Ultrafast Shape Recognition [149], which utilises atom distances from four reference positions. The persistent homology methods that will be used in this work fall into the non-alignment based category.

There are various potential applications of persistent homology to the study of chemical shape spaces. This work explores the use of persistent homology as a descriptor for the shape of molecules, to define a metric space of chemical shape. In particular, it studies what chemical properties are apparent from a persistent homology space. Alternatively, although not explored in this work, is the use of persistent homology to find holes in chemical space. It is reasonable to think that a hole in chemical space describes a molecule with a set of associated parameters that is missing from the data (such as an unobtainable combination of $\log P$ and molecular weight). These studies would certainly be performed in future.

3.4.2 Methodology

The procedure for the analysis performed in this section can be found in Figure 3.14. Given a data set containing a set of molecules (for example represented as a SMILES string), a 'chemical shape space' can be defined. Firstly, the single conformer case is discussed. For each molecule, an ensemble of conformations is created, before they undergo independent minimisation of their (MMFF94 [150]) energy. The lowest energy conformation is then chosen as the single conformer. Persistent homology is calculated, leading to a single persistence diagram for each molecule. A metric space is then created by calculating the bottleneck metric between all pairs of persistence diagrams, for each degree of homology. These metric spaces are inherently high-dimensional and non-Euclidean, so optimal low-dimensional Euclidean representations are found via multidimensional scaling (MDS). Alternatively, the metric spaces for each degree of homology

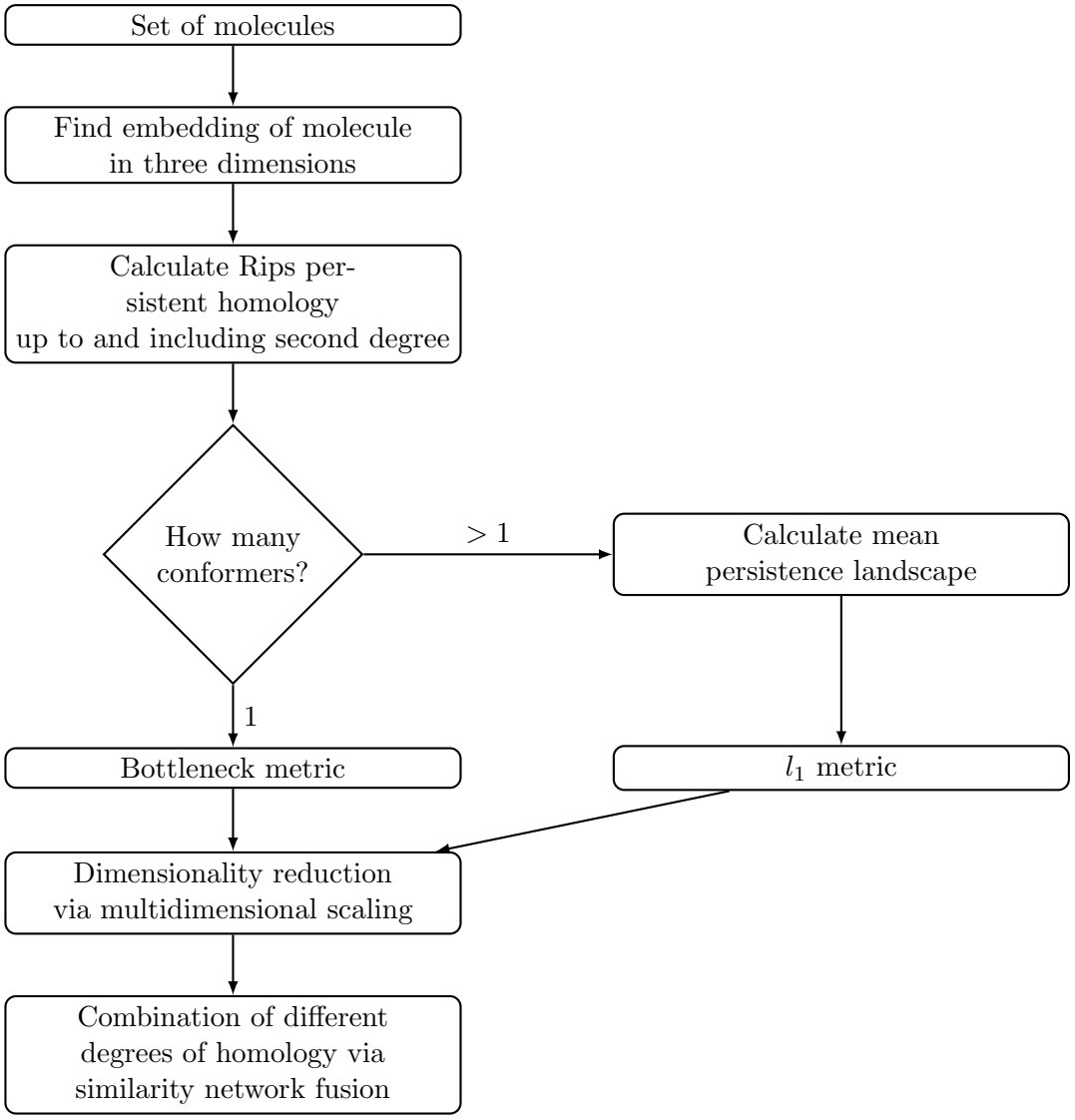


FIGURE 3.14: The general procedure used for the analysis of chemical shape space via persistent homology.

can be combined via similarity network fusion (SNF), leading to a space which ought to contain information about all degrees of homology.

If a single conformer is used, there is a potential issue when defining a chemical shape space. In particular, there could be several energy minima, with wildly differing shapes. Therefore, an alternative methodology is defined. Rather than finding a single minimum energy conformer, an ensemble of conformers are minimised, before they are subject to an RMSD pruning. This pruning seeks to ensure that only a single conformer is found for each minimum, which would potentially bias the 'average shape' of the molecule. Persistence is calculated for each minimum, leading to a series of persistence diagrams for each molecule (note that there is no reason that any two molecules have the same number of resulting persistence diagrams). Instead of defining a metric based on persistence diagrams themselves, the persistence diagrams are converted into landscapes, and a

mean landscape for each molecule is found. A metric space is then defined using the l_1 metric on persistence landscapes, before MDS and SNF can be performed in a similar manner to the single conformer case.

3.4.3 Persistence through Kernels

Thus far, the discussion for this section has avoided the topic of the point cloud on which the persistent homology is calculated. In the water networks chapter (Chapter 5), persistent homology is only calculated on the coordinates of the oxygen atoms in Euclidean space, and it would not be unreasonable to do similar here. However, work carried out by Guo-Wei Wei, Kelin Xia *et al* has suggested the use of kernel-based persistent homology. A simple example of this can be found in their 2014 work on predicting fullerene stability [70]. It is well-known that various physical properties of fullerenes can be linked to their shape [151], and the authors calculated Rips (and other geometrically focused) simplicial complex persistence on kernel-transformed distances:

$$C_{ij} = w_j \Phi(r_{ij}, \eta_{ij}) \tag{3.5}$$

The terms w and η allow the definition of different length scales for different interactions, and the filtration is performed over the output C. The two kernels used are the exponential and Lorentz kernels:

$$\Phi(r, \eta) = e^{-(r/\eta)^\kappa} \tag{3.6}$$

$$\Phi(r, \eta) = \frac{1}{1 + (r/\eta)^\nu} \tag{3.7}$$

Where the κ and ν parameters allow the definition of a whole family of kernels. Persistence was then performed on a 'correlation matrix':

$$M_{ij} = 1 - C_{ij} \tag{3.8}$$

From the persistent homology, the authors were able to create a model for predicting heats of formation and curvature energies. However, it is unsurprising that persistence is useful in this application. The authors themselves acknowledge that the descriptor is essentially capturing the number of hexagons per carbon atom - and even decide to remove features corresponding to the central fullerene hole.

For the data sets used here, it is worth investigating whether the use of kernel-based persistence alters the shape space. For simplicity, in this work the following forms of M_{ij} were used:

$$M_{exp}(r_{ij}) = 1 - e^{-r_{ij}^2} \tag{3.9}$$

$$M_{lor}(r_{ij}) = 1 - \frac{1}{1 + r_{ij}^2} \tag{3.10}$$

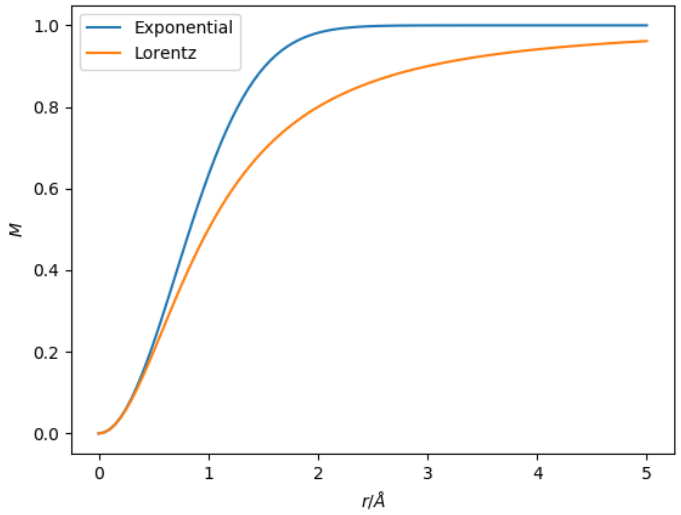


FIGURE 3.15: The exponential and Lorentz kernels used for kernel-based persistent homology in the creation of a chemical shape space. Although they have the same limiting behaviour, the kernels clearly have different shapes.

i.e. η is considered to be equal to 1 distance unit (Å). The correlation function can also be seen in Figure 3.15. From the concavity and monotonic increasing nature of the correlation function, it can be shown that $M(r_{ij})$ is a valid metric and can therefore be used for Rips complex persistence on point clouds. The use of the kernels as defined in this work will alter the persistence diagrams in a fairly predictable way. In particular, as the kernels are concave and monotonic increasing, as well as being defined using all points in the point cloud (unlike the element specific persistent homology found in [152]), the total number of features, and the order where they appear will remain unchanged. Instead, features that are born later are pushed closer to the diagonal, and are born closer towards the end of the persistence process. The kernels therefore give a stronger weighting to features that are born early in the persistence process.

3.4.4 Chemical Shape Space: Single Conformer Studies

3.4.4.1 Coordinate-Based Shape Space

The first chemical shape space explored in this work will be the bottleneck metric space calculated on the Wang data set. The simplest example of the shape spaces calculated were the single conformer spaces, which can use either the bottleneck or landscape metrics for each kernel. Of these parameters, the most obvious place to begin would be the bottleneck metric on the persistence diagrams for the distance functions themselves.

The zeroth degree shape space, projected onto two dimensions by MDS, can be seen in Figure 3.16. The first thing to note is that the space appears to be disconnected. This

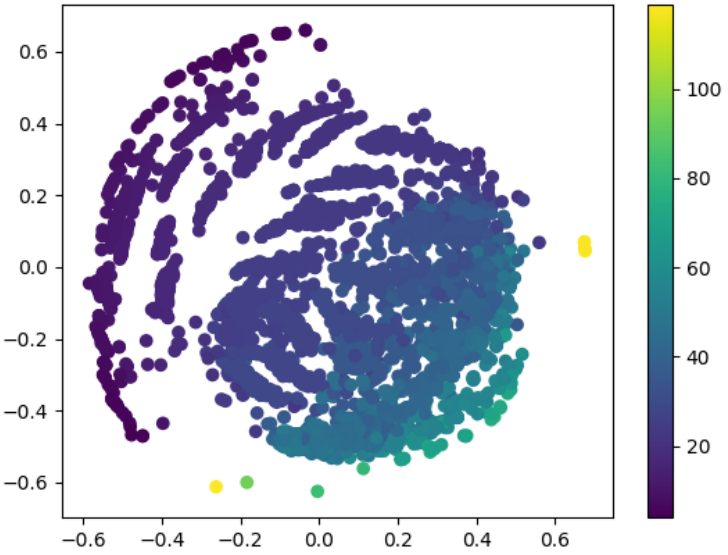


FIGURE 3.16: The two dimensional projection of the zeroth degree chemical shape space for the Wang data set. Coloured by the number of atoms in the molecule. There is a clear correlation between location and the number of atoms.

can be explained by considering the behaviour of zeroth degree Rips complex persistence diagrams. There are n_{atom} points in the diagram, all born at $\delta = 0$. They merge at various times, approximately corresponding to the nearest neighbour distance. Lastly, one point will persist to infinity. The bottleneck metric depends on the optimal matching between pairs of persistence diagrams, and therefore it is not largely surprising that there are gaps in this chemical shape space. Also, it is clear that this leads to location being highly correlated with the number of atoms in the molecule, which is also seen in the projection.

The first degree shape space, again projected onto two dimensions, can be seen in Figure 3.17. Again, the behaviour of the first degree shape space can be explained by considering the persistence diagrams themselves. The first degree persistent homology counts the number of loops in a space. The number of first degree features should therefore match the number of cycles, and therefore a strong correlation between location and the number of cycles is seen in this space. It might be expected that all molecules with 0 cycles should therefore coincide in this space. However, this is not what is observed in the shape space. This is likely a result of the persistent homology being calculated on all atoms, including hydrogen. This could lead to artefacts in the persistence diagram, cycles in the filtration of the simplicial complex that are not 'true' cycles in the molecule (such as a benzene ring). Furthermore, cycles of three atoms (such as cyclopropane or epoxide rings) would not be seen, as three point cycles do not appear in the Rips filtration by definition. These two effects lead to a correlation between position and the number of cycles, that is strong yet imperfect.

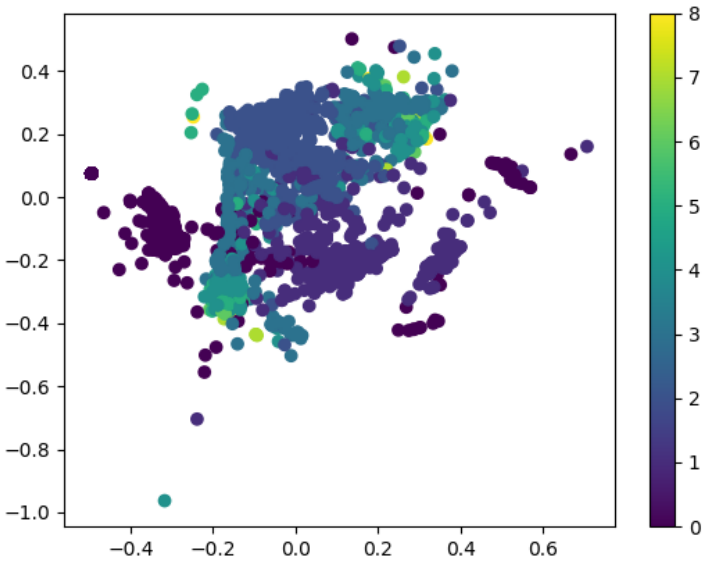


FIGURE 3.17: The two dimensional projection of the first degree chemical shape space for the Wang data set. Coloured by the number of cycles in the molecule. In first degree homology, there is a clear correlation between location and number of cycles

The second degree shape space projected onto two dimensions is found in Figure 3.18. This space contains what appears to be three clusters. These clusters can be seen to clearly correlate with the number of second degree features in the persistence diagram for each molecule, in Figure 3.19. Cluster is therefore determined by the number of second degree features, with location within the cluster corresponding to the differences in birth and death values. Furthermore, the clusters do not distinguish between molecules with ≥ 2 second degree features. This suggests that a large number of these features are close to the diagonal, leading to small bottleneck distances.

It has been seen that the clusters themselves depend on the number of second degree features. Therefore, it is important to determine what molecular property the features correspond to. Within the persistence diagram for a molecule, there are three types of second degree features:

1. Second degree features caused by an enclosed space, such as those found in a fullerene. These are 'real' holes within the molecule.
   (a) As a result of the persistence construction, this type of feature could also be found in a molecule which is not entirely closed, such as a hemispherical molecule.
2. Second degree features caused by conformation. This could be a result of two distinct regions of the molecule being close together, where the Rips construction used in this work does not distinguish between regions that are not directly bonded.

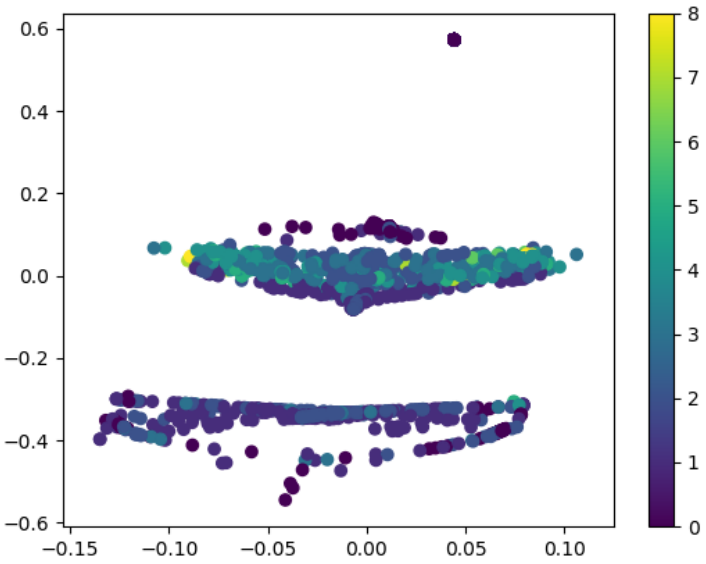
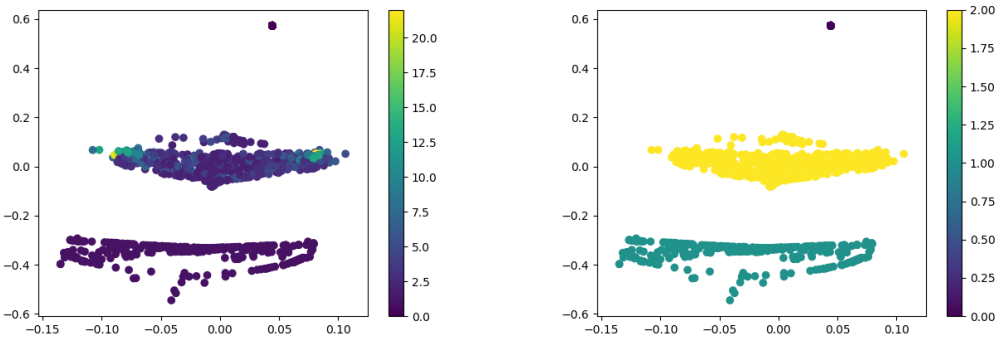


FIGURE 3.18: The two dimensional projection of the second degree chemical shape space for the Wang data set. Coloured by the number of cycles in the molecule. The relationship between location and the number of cycles has now disappeared.



(a) Coloured by Actual Value

(b) Coloured by class {0,1,2+}

FIGURE 3.19: Two dimensional projection of second degree shape space, coloured by the number of second degree features in a molecule's persistence diagram. Clearly, the space is separated by the number of second degree features, which is harder to relate to chemical properties.

Simplices could appear between these distinct regions, which then could lead to second degree persistent features.

3. Second degree features caused by the death of first degree features, such as the artefact found in the hexagon Rips complex filtration (see Figure 2.7). These features are 'artefact' features, that only exist due to the Rips construction themselves - however this does not mean that they do not contain useful information.

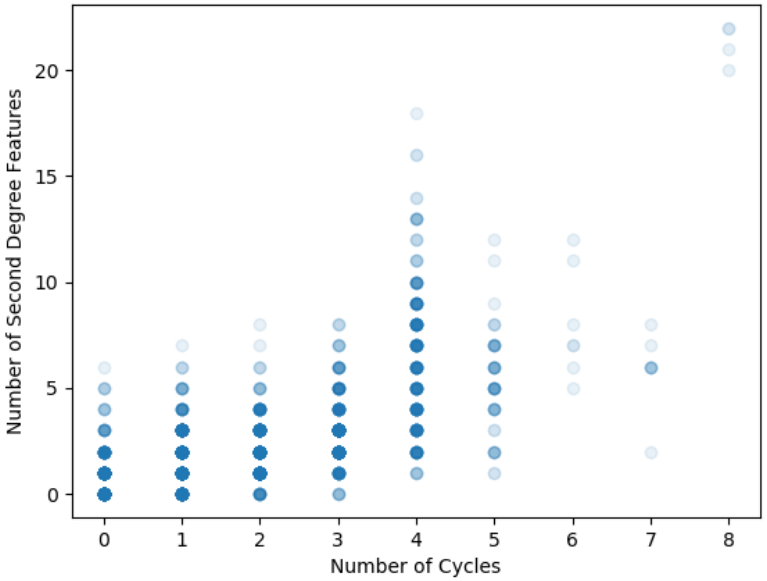


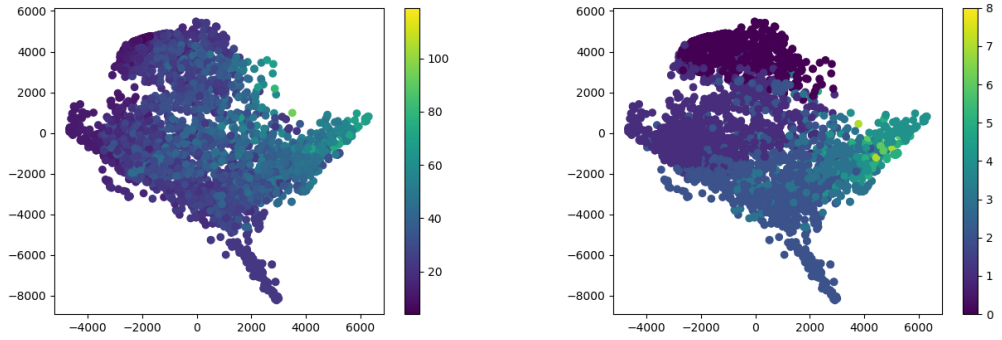
FIGURE 3.20: The number of second degree features as a function of the number of cycles in a molecule. There is a wide scatter in this relationship, illustrating that cycles and second degree features are difficult to relate.

Features of all kinds may appear in the persistence diagram. This can be further understood through Figure 3.20, showing the number of second degree features in a persistence diagram as a function of the number of cycles in a molecule. Points below the diagonal imply the number of cycles is more than the number of features. This could correspond to cycles of length three or four, which do not lead to second degree (artefact) features in the Rips construction. In contrast, points above the diagonal are molecules where the number of second degree features is greater than the number of cycles. Some of these features *must* be the result of 'real' holes, or conformation, as there are not enough cycles to account for the 'artefact' features.

The bottleneck metrics for various degrees of homology can be combined via SNF. This results in a distance matrix, which can then also be analysed by MDS. This should give the most complete description of the chemical shape space defined by persistent homology. The two dimensional projection of the SNF space can be seen in Figure 3.21. The space has correlations with both the number of cycles and the number of atoms, as expected.

3.4.4.2 Effect of Kernel

The two dimensional projection of the zeroth degree persistent homology shape space created with the exponential and Lorentz kernels can be seen in Figure 3.22. They are both qualitatively similar in shape, as well as to the original coordinate-based shape



(a) Coloured by Number of Atoms

(b) Coloured by Number of Cycles

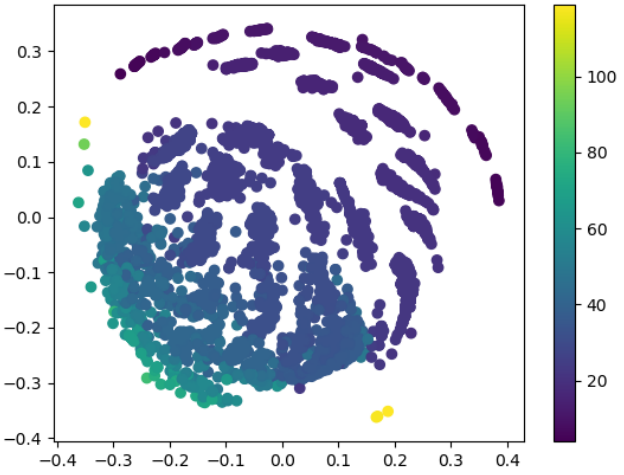
FIGURE 3.21: Two dimensional projection of SNF of the bottleneck shape space. Both the number of cycles and number of atoms are relevant to location in this combined SNF plot.

space Figure 3.16. This is expected, considering that the effect of the kernels is to impact features more heavily as their birth time increases. For zeroth degree homology, all features are born at the same time ($\delta = 0$). The small differences between kernels are as a result of the variation in death times. However, the space still clearly correlates with the number of atoms.

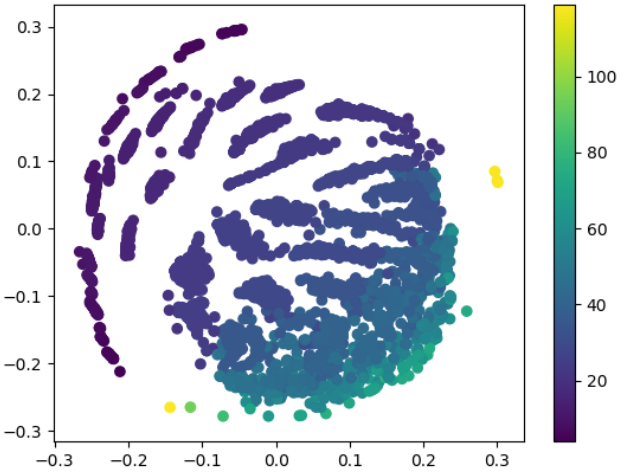
The same projection of the first degree bottleneck shape space is found in Figure 3.23. Again, the kernel-based shape spaces are qualitatively similar, as well as similar to the coordinate-based shape space. However, the distributions of points has narrowed, with the exponential kernel being the most narrow. This can be explained when considering the functional form of the kernels, as seen in Figure 3.15. The exponential kernel decays quickly, leading to persistence diagrams that are more similar, and therefore closer in the resulting shape spaces. Again, there is a clear correlation with the number of cycles, as previously seen.

Similar conclusions can be made regarding the second degree kernel-based bottleneck shape space 3.24. The second degree spaces again form three clusters (which correspond to the number of second degree features). However, the clusters are tighter than previously, with the exponential kernel leading to clusters which appear to collapse to almost a single point. This is still a result of the functional form of the kernels themselves.

The effect of the two kernels chosen in this work appears to be to tighten the resulting shape spaces. This is a result of the kernels treating all interactions with the same weighting. In particular, if different values for w and η were chosen for different interactions (for example, treating C-C distances differently to C=C, or non-bonded interactions differently altogether), the kernels would have an impact reflecting this. The resulting shape space would then look fundamentally different to those studied in this work.



(a) Exponential Kernel

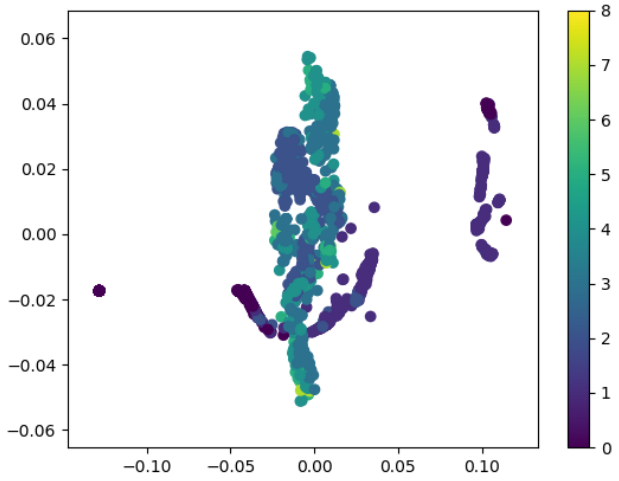


(b) Lorentz Kernel

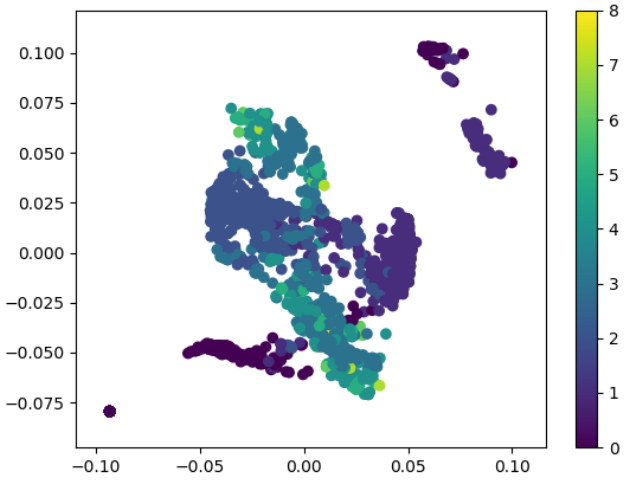
FIGURE 3.22: Two dimensional projection of zeroth degree shape space. Coloured by the number of atoms. The use of kernel does not largely affect the resulting chemical shape space.

3.4.4.3 Comparison to ChEMBL-FL Data Set

The ChEMBL-FL data set contains a different distribution of molecules to those found in the Wang data set, in particular containing molecules that are more similar, as discussed previously in this chapter. It is therefore interesting to study the differences in the resulting shape spaces between these data sets. The zeroth degree bottleneck space can be seen in Figure 3.25. There are similarities between this space and the shape space of the Wang data set, in Figure 3.16. Firstly, the correlation between the number of atoms and location is present, in this case arguably stronger. It is thought that



(a) Exponential Kernel

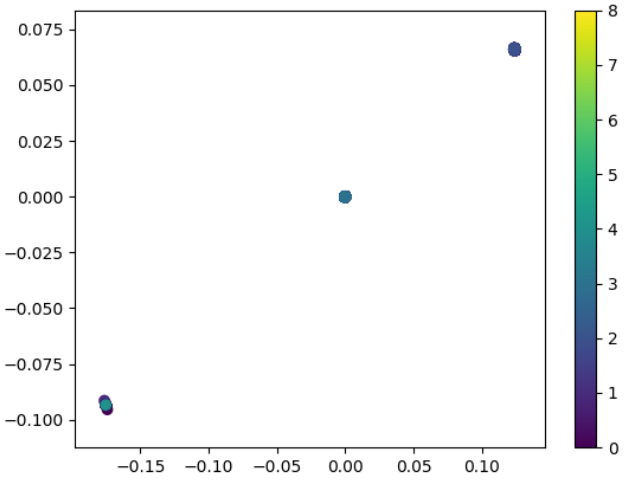


(b) Lorentz Kernel

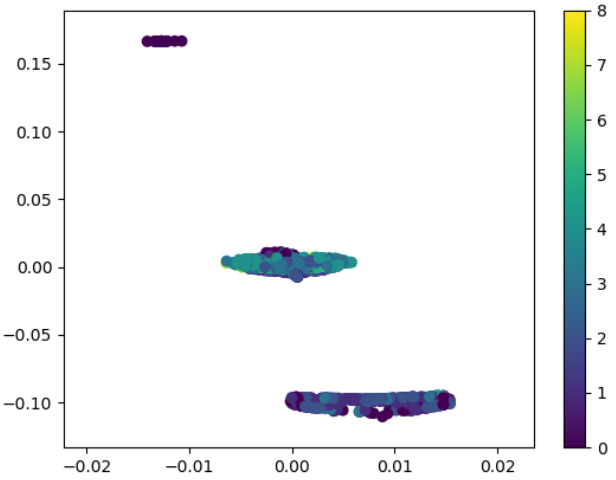
FIGURE 3.23: Two dimensional projection of first degree shape space. Coloured by the number of cycles. Again, the choice of kernel does not largely affect the qualitative features of the chemical shape space.

this is due to the differences in distributions of the number of atoms between data sets (Figure 3.1(b)). The Wang data set appears to have an outlier molecule, and a much more skewed distribution, whereas the ChEMBL-FL data set has a far less skewed distribution. This lack of outliers leads to a shape space that is much more regular, with a far more pronounced correlation between the number of atoms and location.

With first degree homology (seen in Figure 3.26), similar conclusions can be made. Although the shape space looks markedly different to that found for the Wang data set, similar correlations with the number of cycles can be seen. However, it is worth seeing



(a) Exponential Kernel



(b) Lorentz Kernel

FIGURE 3.24: Two dimensional projection of second degree shape space. Coloured by the number of cycles. The short decay-scale of the exponential kernel leads to a point-like distribution of the chemical shape space.

if this number of cycles correlation is actually a correlation with the number of first degree features. This can be seen in Figure 3.27. Molecules tend to group depending on the number of first degree features, with resolution being lost when $n_{features} \geq 2$. However, for this low resolution subset, molecules appear to cluster based on the actual number of cycles in the molecule. This behaviour can be explained by considering the underlying persistence diagrams. The true cycles in the molecule would be expected to cause features that are far from the diagonal, whereas other first degree features would likely be noise. This would lead to bottleneck distances correlating with the number of cycles in a molecule, for these larger molecular systems.

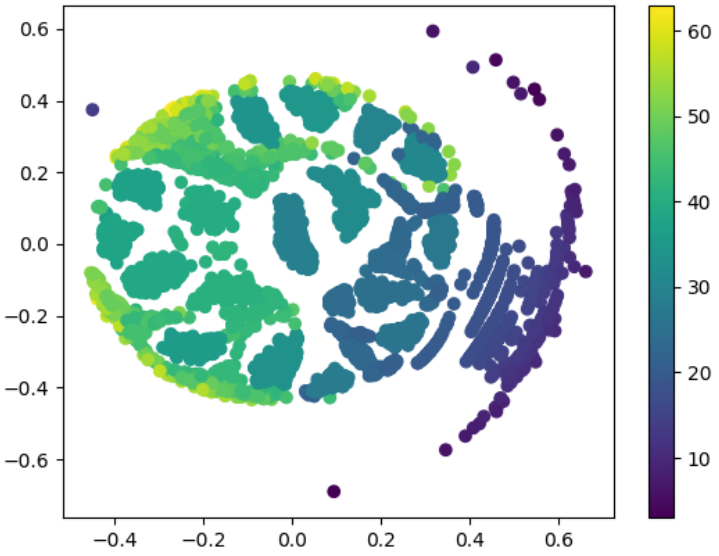


FIGURE 3.25: The two dimensional projection of the zeroth degree chemical shape space for the ChEMBL data set. Coloured by the number of atoms in the molecule. Although the shape space looks different for the new data set, the relationship between the number of atoms and location in zeroth degree homology shape space is retained.

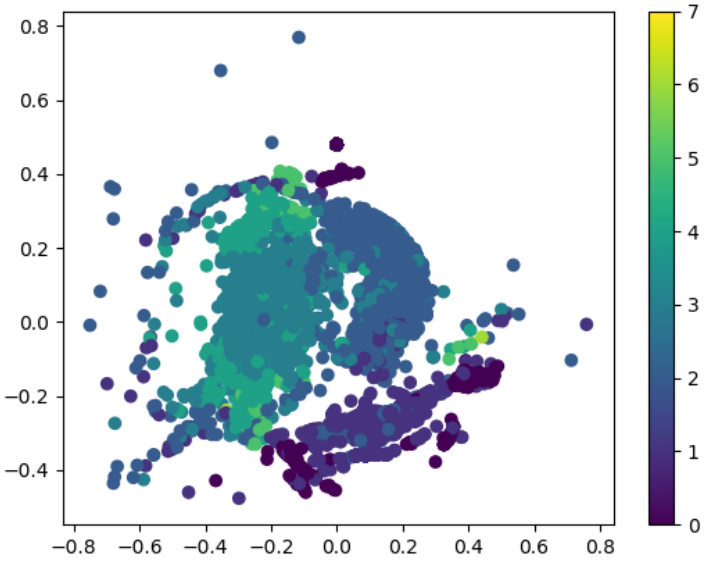
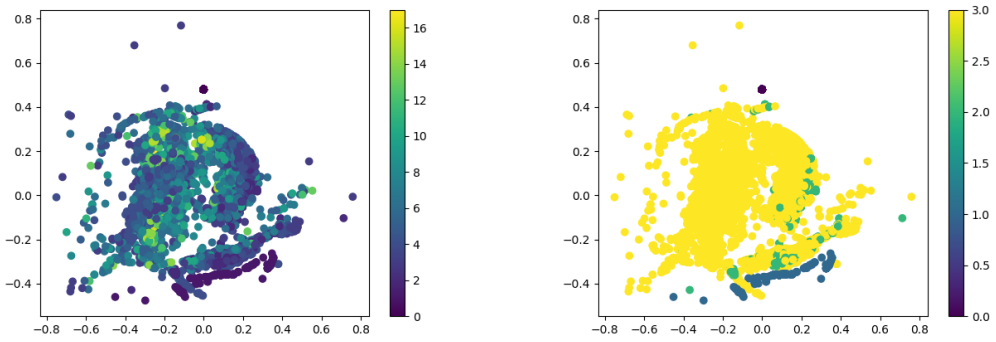


FIGURE 3.26: The two dimensional projection of the first degree chemical shape space for the ChEMBL data set. Coloured by the number of atoms in the molecule. Although the shape space looks different for the new data set, the relationship between the number of cycles and location in first degree homology shape space is retained.



(a) Coloured by Actual Value

(b) Coloured by class {0,1,2,3+}

FIGURE 3.27: Two dimensional projection of first degree shape space, coloured by the number of first degree features in a molecule's persistence diagram. A relationship between number of first degree features and location is now observed.

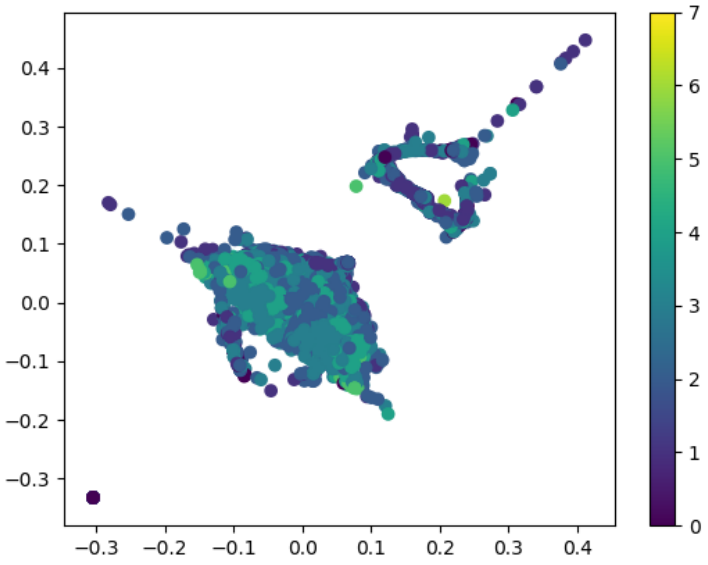


FIGURE 3.28: The two dimensional projection of the second degree chemical shape space for the ChEMBL data set. Coloured by the number of cycles in the molecule. As before, it is difficult to determine the relationship between location and number of cycles in second degree homology.

Finally, the second degree shape spaces can be compared (Figure 3.28). The familiar three clusters are seen, corresponding to the number of second degree features. Unlike the first degree space, the correlation with the number of cycles is no longer seen. This is a result of the second degree features tending to be closer to the diagonal, with the cycle 'artefact' features mentioned previously in particular having similar birth and death values.

Although all of the ChEMBL-FL spaces *look* different to those found for the Wang data

set, the broad conclusions regarding location and molecular descriptors are the same. In particular, the zeroth degree space is intimately connected to the number of atoms, and the first degree space is connected to the number of cycles. The second degree space relates strongly to the number of second degree features, which in turn is again linked to the number of cycles. However, as most second degree features are actually artefacts, rather than true voids found within molecular conformations, it is difficult to determine which 'real' chemical property determines the second degree space.

3.4.5 Chemical Shape Space: Effect of Multiple Conformations

As an example for how the persistence landscape is able to capture conformational flexibility, some conformers and their individual first degree landscapes for 11-aminoundecanoic acid can be found in Figure 3.29. The mean landscape for the 31 low energy conformations found in the conformer generation procedure is seen in Figure 3.30. Although it is difficult to determine what features of the molecule lead to features of the persistence landscape, it is clear that the landscapes from different conformations can markedly differ, and the mean landscape contains information about all of the conformations. Although there are no loops within the molecule, the first degree landscapes contain features caused by non-bonded groups of atoms.

The shape space induced by the l_1 metric on persistence landscapes (Equation 2.11), and in particular how this can be used to develop shape spaces for multiple conformers is investigated. For the Wang data set, a set of low-energy conformations (with RMSD pruning) were created for each molecule. Persistence landscapes were calculated, and the l_1 metric between mean persistence landscapes were found for each degree of homology separately. MDS can then be calculated on the distance matrices, and projected into two dimensions. For zeroth degree homology, this can be seen in Figure 3.31.

The space again correlates with the number of atoms. The space is now connected - with disconnections caused by gaps in the distribution of the number of atoms. To understand if this is a feature of the multiple conformations, or instead a property of the landscape metric, the same procedure was performed on the minimum energy conformer landscape, as opposed to the mean landscape. This can be seen in Figure 3.32.

The two projections are similar, and they are indeed both connected. The connectedness is therefore not a result of the multiple conformations and mean landscape in some sense 'smoothing' the distance metric. In fact, this similarity is not unsurprising. The zeroth degree homology is essentially measuring nearest neighbour information. Bond lengths would not particularly be expected to vary between minimum energy conformations. This would therefore lead to the different conformations having similar zeroth degree homology and persistence landscapes, and the mean landscape to reflect this similarity.

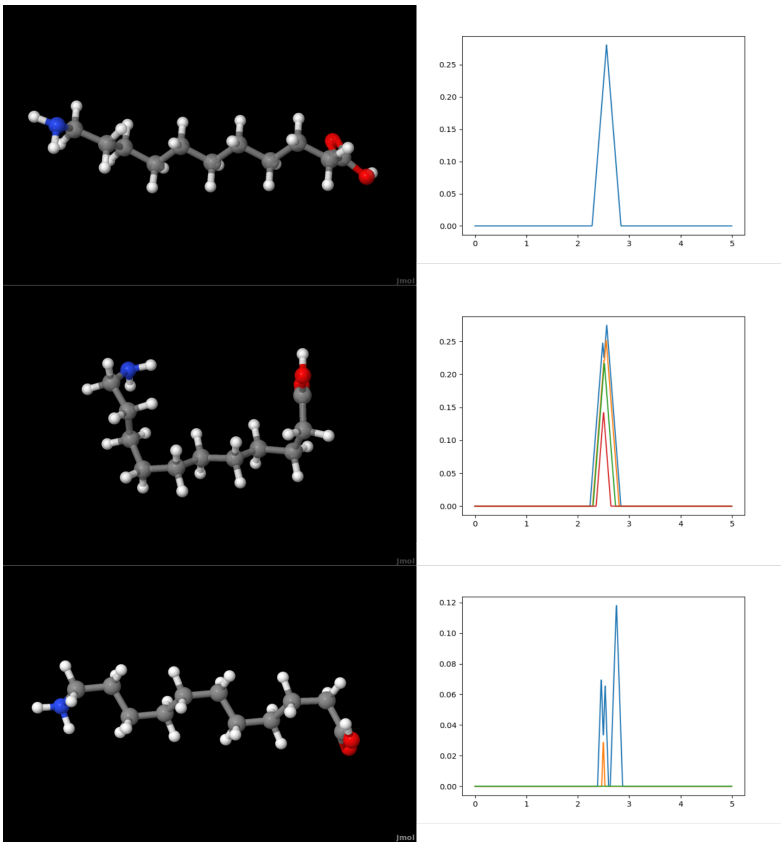


FIGURE 3.29: Some low energy conformations of 11-aminoundecanoic acid and their first degree persistence landscapes. The different conformations can lead to different persistence landscapes, which can be combined to create a single persistence landscape reflecting molecular flexibility.

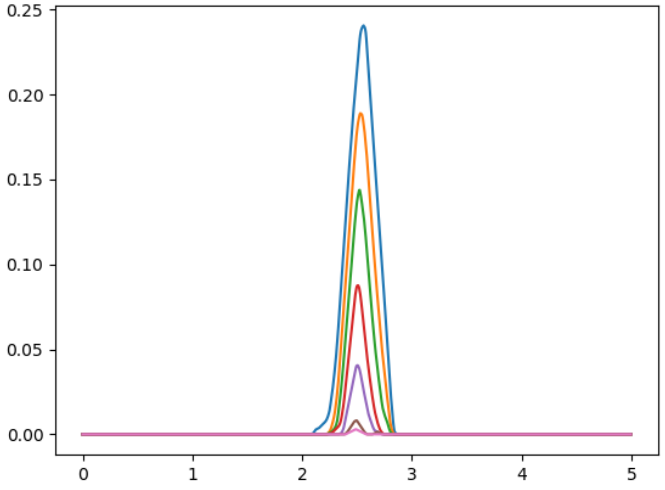


FIGURE 3.30: The mean first degree persistence landscape for the 31 low energy conformations of 11-aminoundecanoic acid. This landscape can be used to create a shape space reflecting the inherent flexibility of molecules.

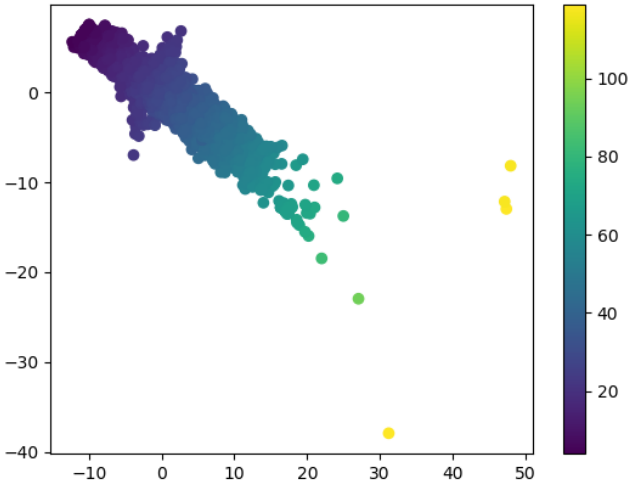


FIGURE 3.31: The two dimensional projection of the zeroth degree chemical shape space for the Wang data set, multiple conformations. Coloured by the number of atoms in the molecule. There is a relationship between number of atoms and location, as before.

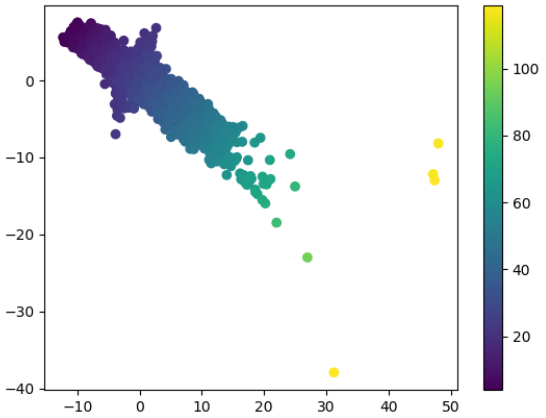
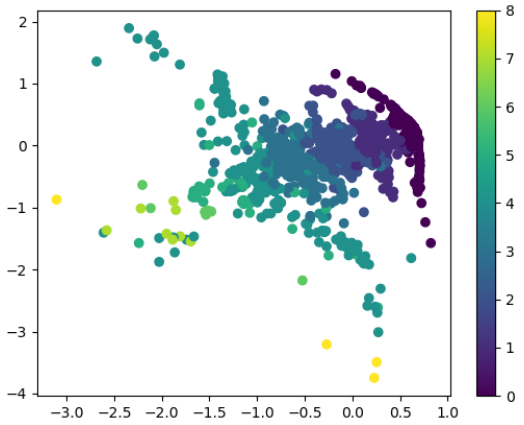
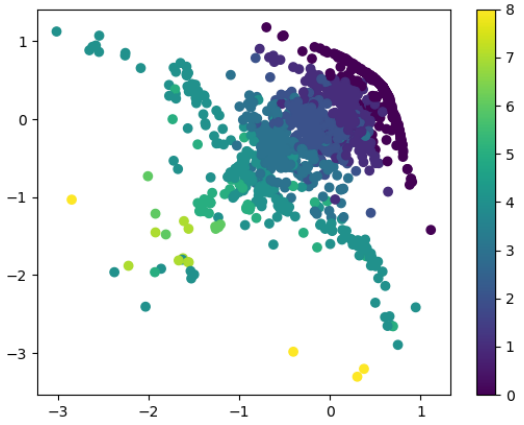


FIGURE 3.32: The two dimensional projection of the zeroth degree chemical shape space for the Wang data set, minimum energy conformation. Coloured by the number of atoms in the molecule. The space is unchanged when compared to the landscape space utilising multiple conformations. Reasons for this are discussed in the text.

The first degree homology shape space projections can be seen in Figure 3.33, for both the mean persistence landscape and the minimum energy conformation landscape. The two spaces are again similar, but not as similar as the two zeroth degree projections. This is due to two effects:



(a) Multiple conformations, mean landscape



(b) Lowest energy conformation, single landscape

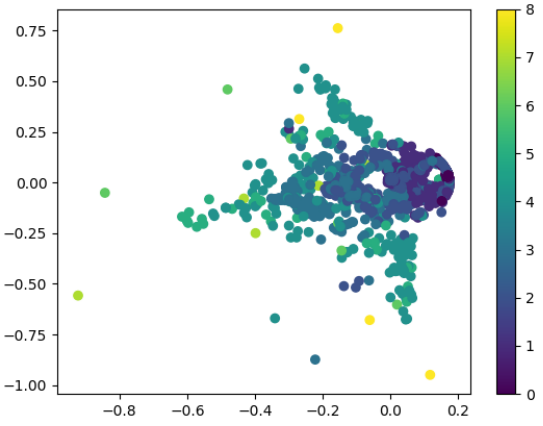
FIGURE 3.33: Two dimensional projection of first degree shape space with the landscape metric, for both mean landscapes and minimum energy landscapes. Coloured by the number of cycles in a molecule. The relationship between number of cycles and location is again observed, and there is a difference in the shape spaces when multiple conformations are considered.

- Changes in the ring conformation themselves, leading to changes in position of pre-existing first degree features
- Changes in overall conformation that lead to new (short lived) first degree features

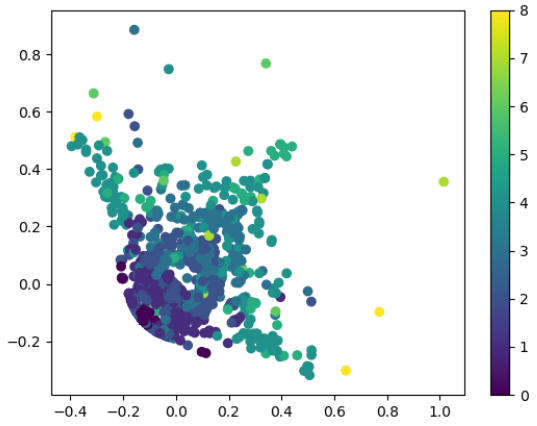
The large-scale differences are likely due to the former. This is because these features are likely to be more long lived, and therefore lead to larger differences in the persistence

landscapes themselves. In contrast, the new features are topological noise, caused by simplices between distinct regions of the molecule. These would be short lived, and therefore lead to small changes in the landscape and the overall shape space.

The same projections, for the second degree landscape space are found in Figure 3.34. The conclusions to be made regarding second degree homology are largely similar to



(a) Multiple conformations, mean landscape



(b) Lowest energy conformation, single landscape

FIGURE 3.34: Two dimensional projection of second degree shape space with the landscape metric, for both mean landscapes and minimum energy landscapes. Coloured by the number of cycles in a molecule. The relationship between location and number of cycles is again harder to find in second degree homology.

those made previously, including the similarity to second degree homology. Again, this is due to second degree features caused by the death of first degree features. Furthermore, the space does not seem dependant on whether a single conformer, or multiple conformers are used when defining the shape space.

3.5 Conclusions and Future Directions

Topological data analysis has been applied to two tasks in the analysis of chemical space. The mapper algorithm has been used to understand a descriptor space of a set of molecules, and the conclusions from the resulting networks used to better understand solubility prediction. It was found that there is a correlation between the solubility and the number of chlorines in a molecule, but only for molecules with two rings. This correlation was then taken advantage of with the creation of linear mixed models, that were able to improve the consistency of solubility prediction.

Topological data analysis was then used to create a molecular shape space. Persistent homology was used to understand the topology of molecules from their atomic distances, before the bottleneck metric was used to create a notion of similarity. Different degrees of homology then created different molecular shape spaces. The zeroth degree homology space was found to be closely related to the number of atoms in a molecule, and the first degree space was found to correlate with the number of rings. Through similarity network fusion, these spaces were combined to create a single molecular shape space for a data set. As well as atomic distances, persistent homology was then calculated on functions of the distances themselves through kernels. However, as the kernel did not take into account atom or bond types, it was found that there was little effect of the kernel. This persistence methodology was then extended to persistence landscapes, to account for the effect of multiple low energy conformations, which may alter the shape space. The same correlations were found between different degrees of homology and molecular properties, and it was found that the effect of conformation was negligible.

In future, the mapper algorithm could be applied to other, difficult to predict molecular properties, such as drug activity. This enables the creation of simpler models, which are more interpretable than deep learning methodologies. With regard to solubility prediction, the Box Cox transformation on the Wang data set has recently suggested that the log transformation may not lead to normally distributed data (although there are physical reasons why the log transformation is used). The impact of this effect on the residuals studied should be investigated. Regarding chemical shape space, in future the effect of kernels could be investigated further, through the use of different characteristic length scales for different interactions. Also, a comparison of this description of chemical space, and others such as the 'Molecular Quantum Number' description would be interesting [153]. In particular, the use of two dimensional persistence, such as that suggested by Keller [92], would enable charge information to be included in a persistent homology description of chemical space.

Chapter 4

Topological Data Analysis of Conformational Space

4.1 Introduction

4.1.1 Configurational Spaces

The notion of a conformational space is related to that of a configuration space in physics. The configuration space of a system can be thought of as the space defined by all of its possible positions, subject to its constraints. This is distinct from the phase space of the same system, which also describes momenta. The phase space therefore also describes the dynamics of the system, as opposed to the statics described by the conformational space.

Consider a single particle, moving with the influence of no external forces in Euclidean 3-space. Its position at any moment t can be written as:

$$r(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}$$

with the associated configuration space of the system being $\mathbb{R}^3$. In principle, this description can actually be reduced further. The system is under no external forces, and therefore every location $r \in \mathbb{R}^3$ is identical. This is an underlying symmetry of the system, and if this symmetry could somehow be taken into account, the configuration space could be in some sense 'simplified'. For this example this is trivial - all points in $\mathbb{R}^3$ are identical, leading to a 'simplified' conformational space of the singleton $\{*\}$.

Now consider a system of two identical particles joined by a rigid rod. Clearly, the positions of the particles can be written as:

$$q(t) = \begin{pmatrix} x_1(t) \\ y_1(t) \\ z_1(t) \\ x_2(t) \\ y_2(t) \\ z_2(t) \end{pmatrix}$$

therefore the configurational space is *at most* $\mathbb{R}^6$. However, in the same sense as previously, symmetries can be removed from the system. Firstly, the position of one particle can be fixed as the origin. This leads to a configurational space of $\mathbb{R}^3$. Further, the two particles are a fixed distance apart, due to the rigid rod. This implies that the only coordinate needed to describe the location of the second particle (relative to the first) is some angle θ - the configurational space of this system can therefore be reduced to the circle S^1.

Turning to an example which is perhaps more familiar to chemists, consider the system above with a flexible rod. It is clear that such a system is analogous to a classical approximation to a diatomic molecule. Now, the system can be described by the length of the rod, and the angle. This leads to a configurational space of $\mathbb{R}^+ \times S^1$, a cylinder of infinite extent. Provided the bond has a maximum length, the configurational space can be considered to have the topology of $I \times S^1$

This chapter focuses on the classification and analysis of configurational spaces of molecules, via topological data analysis. However, it is already apparent the need to define the notion of a molecule exactly, as well as which symmetries may be considered. A full, mathematically rigorous treatment can be found in [30]. This work presents a definition lacking in the complete rigour, but designed to match notions familiar to chemists.

4.2 Mathematical Definitions

4.2.1 Molecules and Conformers

Molecules are inherently quantum objects. However, the use of classical approximations to molecules is ubiquitous in chemistry, for example in classical molecular dynamics simulations. This classical treatment of molecules is utilised in this work, although it would be an interesting region of future study to extend the methods of analysis described here to a quantum treatment.

Firstly, a classical description of a molecule must be defined. This definition should be as general as possible, in particular the definition should not make reference to a set of coordinates describing the molecule in $\mathbb{R}^3$ - as this is now moving into defining a conformer. The notion of a molecular graph is useful here, and familiar to chemists. A molecular graph is a tuple $\mathcal{G} = (V, E, c_v, L, \Theta)$ with the following data:

1. $\Gamma = (V, E)$ is a finite, undirected graph. V is a finite set of vertices, and E is a set of unordered pairs $(v, w) \in V$ detailing edges between vertices. Chemically, these correspond to atoms and bonds respectively.
2. $c_v : V \to \mathbb{N}$ is a vertex colouring, where for every vertex in V, c_V describes the element of any given atom
3. $L : E \to (0, \infty)$ is a set of length constraints, describing bond lengths
4. $\Theta : E_2 \to (0, \pi]$ is a set of angle constraints, where:

$$E_2 = \{(v, w_1, w_2) \in V \times V \times V | (v, w_1), (v, w_2) \in E, w_1 \neq w_2\}$$

   is the set of adjacent bonds, i.e. $\Theta(v, w_1, w_2)$ is the angle between bonds (v, w_1) and $(v, w_2) \in E$.

A single conformer $\mathcal{C}$ can be considered to be a geometric realisation of $\mathcal{G}$ into $\mathbb{R}^3$, such that for two bonded atoms (v, w) their Euclidean distance is equal to $L(v, w)$, and for each pair of adjacent bonds (v, w_1) and (v, w_2) their angle (as defined by the dot product) is equal to $\Theta(v, w_1, w_2)$. The *classical* energy of a conformer can be calculated via the use of molecular mechanics forcefields, of the general form:

$$\begin{aligned} E(\mathcal{C}) = &\sum_{(v,w)\in E} k_{v,w} \left(d(v, w) - \bar{d}_{v,w}\right)^2 \\ &+ \sum_{(v,w_1,w_2)\in E_2} K_{v,w_1,w_2} \left(\Theta(v, w_1, w_2) - \bar{\Theta}_{v,w_1,w_2}\right)^2 \\ &+ \sum_{t\in\text{torsions}} E_t(\tau) \end{aligned} \tag{4.1}$$

Where $\tilde{d}$ and $\tilde{\Theta}$ can be considered to be equilibrium values of a given bond length and angle respectively. The first term treats bond stretching in a harmonic manner, the second treats angle stretching similarly. The third term describes how the energy of a given torsion angle (i.e. between 4 adjacent atoms) varies:

$$E_t(\tau) = \sum_n \frac{1}{2} V_n (1 + \cos(n\tau + \delta_n)) \tag{4.2}$$

Neither the definition of the molecule, or definition of the conformer, make reference to torsion angles τ. This torsional flexibility leads to the definition of a conformational space.

4.2.2 Conformational Spaces

The most general definition of a conformational space $\mathcal{M}$ is as the set of *all* permissible embeddings of $\mathcal{G}$. However, there are various symmetries that could now be taken into account. The first symmetry is translational - there is nothing in the previous definition prohibiting the same conformer embedding in different regions of $\mathbb{R}^3$. This symmetry can be taken into account via a centre of mass alignment.

The second symmetry that could be taken into account is rotational. After a centre of mass alignment, two conformers may be identical except for a transformation in $SO(3)$. This can be dealt with by aligning principal axes of inertia for the conformer set - or in this work via an RMSD alignment (see Chapter 4.4.1 for more details).

The final symmetry that can be taken into account is the symmetry inherent to the molecule itself. Specifically, a subset of vertices in $\mathcal{G}$ may be equivalent. This can be analysed by studying the permutation group of the molecular graph - which vertices can be swapped without changing the graph. This symmetry is more general than the point group symmetry often studied in chemistry, which details the symmetry of an embedding of the molecular graph (i.e. a conformer). Furthermore, this symmetry is less general than the complete permutation inversion group studied by Longuet-Higgins [154], which details the symmetry of a quantum molecule (i.e. all nuclei of the same element are equivalent, all electrons are equivalent etc.). Symmetries of a molecule, and how the RMSD can take them into account, are covered in detail in [155].

Regardless of the symmetries taken into account, this work will use the term 'conformational space'. It is clear from context which symmetries in particular have been used for each example.

4.2.2.1 Properties of Conformational Spaces

There are some properties that a conformational space must require to ensure they match 'chemical intutition'. They are detailed below:

- Path-connected: A conformational space must be path connected. This ensures that any conformer in the space can be transformed into any other. This allows physical changes (bond stretching etc.) but not chemical changes to the molecule (i.e. stereoisomerism)
- Metric space: There is some notion of similarity d between conformers. This allows conformers to be compared - in particular to ensure that symmetries have been removed
- Bounded: The limits of the conformational space exist

These properties are not automatically fulfilled by the definition of a molecule above. The definition of a molecular graph does not describe chirality, for example, and it will be discussed later how this can be a potential issue. On the other hand, the bounded property is satisfied by ensuring that there is a maximum/minimum bond length or angle. This is not implied by the definition of $\mathcal{C}$, but is necessary to ensure that bonds cannot become infinitely long (dissociate), which is unphysical within this classical approximation.

4.3 Characterisation of Conformational Spaces

Conformational spaces of molecules have been studied previously, often in the context of the creation of efficient methods for their enumeration. For example, RDKit uses a distance geometry approach to conformer generation [117, 156]. Such a method is particularly useful when generating small sets of low-energy conformers, such as those described in Chapter 3. However, enumeration methods are often necessary, such as for ring molecules. Work has been carried out by Porta *et al*, developing methods for enumeration of molecular loop conformational spaces [157], which again can be reduced to a distance geometry approach. Some comparisons of conformational space generation methods can be found in [158, 159].

The body of work on the characterisation of conformational spaces is often focused on a select few molecules, rather than general methods such as those outlined in this work. Often this is because characterisation is difficult, whereas the calculation of low-energy conformers is much simpler, and often all that is needed. However, the goal of characterisation is still important. For example, features of the energy landscape can be immediately learned, without calculation, from the conformational space itself. An elementary use of this is through the Borsuk-Ulam theorem, which states that if the map $f : S^n \to \mathbb{R}^n$ is continuous, there are two antipodal points on S^n which map to the same point in $\mathbb{R}^n$. If the energy map is assumed to be continuous (which is not an extreme assumption), this implies that there should be two antipodal points on the conformational space of butane that have the same energy (given caveats on the assumption of butane's conformational space). The notion of treating the potential energy surface as a map from the conformational space itself is not new, and has been seen previously for alanine dipeptide [160]. Furthermore, the notion of studying trajectories through dimensionality reduction [161, 162], has an intrinsic dependence on the properties of the underlying conformational space.

As mentioned, characterisation methods have previously been restricted to the analysis of the conformational space of single molecules, or molecules of a particular class. For example, Crippen studied the conformational space of cyclo-alkanes through distance geometry [163]. However, this method is heavily restricted, as the conformers within

the set are strongly restricted by the bounds on their constraint matrix. This leads to cyclohexane having a disconnected conformational space in his analysis, as only torsion pseudorotations were allowed. As explained later in the text, an n-site loop has $n-6$ torsion degrees of freedom - leaving cyclohexane with none. Transitions between chair and boat conformations require bonds to bend and stretch - as can be seen with a basic model kit.

For the conformational spaces of small flexible molecular loops, Porta *et al* used ideas from robotics known as higher-dimensional continuation [164]. This method creates local charts to efficiently sample the conformational space. Their work could certainly be used in conjunction with the topological methods of this thesis to characterise the conformational spaces of complicated molecules. This work is strongly related to the work carried on by Martin *et al* [165, 166], which was able to characterise the conformational space of cyclooctane - this work is discussed in more detail in this thesis.

4.4 Analysis Methodology

The general procedure for the generation and analysis of conformational spaces can be seen in Figure 4.1. The initial set of conformers is generated via RDKit's distance geometry methods. This leads to a set of randomly generated conformations, which it is hoped covers all of the degrees of freedom of the space (including bond length and angle variation). Furthermore, by allowing all degrees of freedom to vary, it is possible to study the effect of this on the resulting conformational space.

However, there is a chance that the degrees of freedom are not entirely covered, or covered in an unphysical way. For example, amide bonds may become too flexible, or torsional degrees of freedom become too narrowly distributed. Therefore these degrees of freedom are always checked, and if necessary manually altered in a stochastic manner, to ensure that the conformer sets match chemical intuition. As a further check, conformers are filtered by their energies. This is designed to remove conformers with unphysical atom overlaps.

Once the conformer sets are generated, they are transformed into two conformational space representations. These are described in more detail in Chapter 4.4.2, but they lead to two different metrics between conformations. One of the aims of this work is to determine if these two metrics are equivalent, and in general to contrast between these representations. For each representation, different analyses can be performed, using persistent homology. The persistent homology analyses are described in more detail later in the chapter.

The specifics of the procedures for each molecule studied in this work are presented in more detail later, but all follow the general methodology defined above.

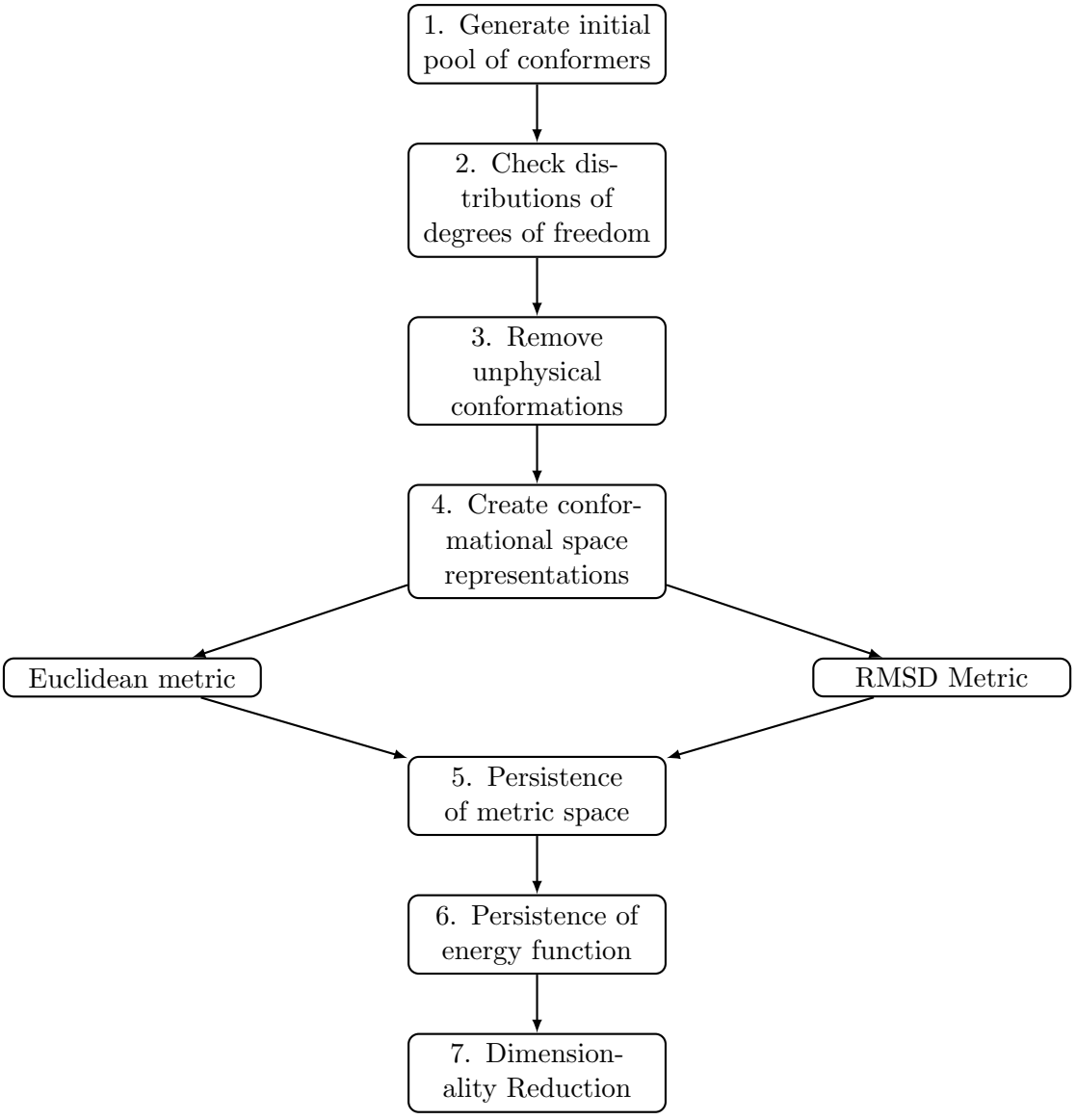


FIGURE 4.1: The general procedure used for the analysis of conformational spaces via persistent homology.

4.4.1 Conformational Space Representations

Given a set of conformers, various representations can be defined on their conformational spaces. These representations could in principle lead to different topologies of the conformational space. This work investigates two representations, the first using the Euclidean metric and the second the RMSD metric. This section details how these two metrics are generated, how they influence their corresponding representations, and how persistence can be done on them in principle.

A single conformer $\mathcal{C}$, consisting of n atoms, can be written as the following matrix or equivalent vector in $\mathbb{R}^{3n}$:

$$\mathcal{C} = \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ & \vdots & \\ x_n & y_n & z_n \end{pmatrix} \simeq \begin{pmatrix} x_1 & y_1 & z_1 & x_2 & y_2 & z_2 & \dots & x_n & y_n & z_n \end{pmatrix}$$

A conformational space $\mathcal{M}$ set of m conformers can therefore be written as a matrix in $\mathbb{R}^{3n\times m}$:

$$\mathcal{M} = \begin{pmatrix} \mathcal{C}_1 \\ \mathcal{C}_2 \\ \vdots \\ \mathcal{C}_m \end{pmatrix} = \begin{pmatrix} x_{1,1} & y_{1,1} & z_{1,1} & x_{2,1} & y_{2,1} & z_{2,1} & \dots & x_{n,1} & y_{n,1} & z_{n,1} \\ x_{1,2} & y_{1,2} & z_{1,2} & x_{2,2} & y_{2,2} & z_{2,2} & \dots & x_{n,2} & y_{n,2} & z_{n,2} \\ & \vdots & & & \vdots & & \ddots & & \vdots & \\ x_{1,m} & y_{1,m} & z_{1,m} & x_{2,m} & y_{2,m} & z_{2,m} & \dots & x_{n,m} & y_{n,m} & z_{n,m} \end{pmatrix}$$

The Euclidean metric is then defined as:

$$d_E(\mathcal{C}_1, \mathcal{C}_2) = \sqrt{\sum_{i=1}^{n} \left(\sum_{\alpha\in\{x,y,z\}} |\alpha_{i,1} - \alpha_{i,2}|^2 \right)} \tag{4.3}$$

The RMSD metric can be defined similarly:

$$RMSD(\mathcal{C}_1, \mathcal{C}_2) = \sqrt{\frac{1}{N}\sum_{i=1}^{n} \left(\sum_{\alpha\in\{x,y,z\}} |\alpha_{i,1} - \alpha_{i,2}|^2 \right)} \tag{4.4}$$

The metrics are clearly similar, and if both representations used the same alignment of conformers, they would lead to equivalent conformational spaces. However, the representations differ in how the conformers are aligned. For the RMSD representation, each pair of conformers are pairwise aligned:

$$\bar{d}_R(\mathcal{C}_1, \mathcal{C}_2) = \min_{g\in SO(3)} d_R(\mathcal{C}_1, g(\mathcal{C}_2)) \tag{4.5}$$

In contrast, the Euclidean representation is defined as follows:

$$\bar{d}_E(\mathcal{C}_1, \mathcal{C}_2) = d_E(g_1(\mathcal{C}_1), g_2(\mathcal{C}_2)) \tag{4.6}$$

Where:

$$g_n = \arg\min_{g\in SO(3)} d_R(\mathcal{C}_{ref}, g(\mathcal{C}_n)) \tag{4.7}$$

i.e. the alignment takes place before the metric is calculated, and is to some reference conformer. In this work, the reference conformer is found via a quantum mechanical optimisation using Gaussian 09 [167], with the B3LYP functional [168] using a 6-311g(d,p) basis set and dispersion correction, unless otherwise stated.

The question now becomes: *why would two representations be necessary?* The Euclidean metric requires a total of m alignments, and only requires storage of a $3n \times m$ matrix. This is computationally efficient. In contrast, the RMSD metric would require $m(m-1)/2$ alignments, and this many values must be stored. In particular, there is no equivalent coordinate matrix for the RMSD representation. To ensure good coverage of the conformational space $n \ll m$ and therefore the RMSD representation requires the storage of large amounts of data in memory.

In essence, the Euclidean representation is desired for ease of computation and use, whereas the RMSD representation is required to ensure the most accurate description of the conformational space. One of the goals of this work is to assess the situations where each representation is required in practice.

4.4.2 Persistent Homology Details

In contrast to the rest of this work, this chapter explores more than just Rips persistent homology. This section details the specifics regarding the various flavours of persistence in this chapter. Please refer to Chapter 2 for more details regarding the mathematics behind persistent homology.

4.4.2.1 Persistence of Conformational Spaces

This persistent homology method is the most similar to those studied in the rest of the work. The Euclidean and RMSD representations define a conformational metric space. Rips persistent homology approaches enable understanding of holes in the conformational space. These holes take two forms:

1. Holes due to a lack of sampled conformers
2. Holes due to the topology of the conformational space

Provided that the conformer generation procedure defined above is correct, the holes of the first type should be small, and therefore close to the persistence diagram. In contrast, the second type of holes should all be long lived. This work investigates the use of persistent homology in determining the conformational space of a molecule.

4.4.2.2 Persistence of Energy Functions

Previously, the Rips procedure has been used to define a sequence of filtered simplicial complexes from a set of vertices (in this case, conformers). However, given a well-behaved function defined on the vertices v, there exists a natural extension to any p-simplex σ:

$$f(\sigma) = \max_{v \in \sigma} f(v) \tag{4.8}$$

The sublevel sets of f are defined as:

$$L_c(f) = \{\sigma | f(\sigma) \leq c\} \tag{4.9}$$

From the definition of f it is clear that, for $a < b$, $L_a \subseteq L_b$. This implies that the inclusion maps of Equation 2.4 are well-defined and lead to induced maps in homology. Persistent homology can therefore be calculated of the function $f(\sigma)$ defined on the simplicial complex.

In general, persistent homology of functions defined on simplicial complexes contains information regarding critical values of the function itself. It is strongly related to the mathematical field of Morse theory [91]. Here, persistence is explained via a series of examples.

To understand what the persistence of f implies, consider the example one-dimensional simplicial complex found in Figure 4.2, with its associated height function. It is clear that there would be no interesting first degree homology or higher, therefore only zeroth degree homology (i.e. connected components) is discussed for this complex.

In this example, the notion of height level persistence can be thought of as sweeping up a line parallel to the x-axis, and seeing what is underneath. As the complex is discrete, the only values of y that need to be discussed are the height values of the simplices, denoted y_α, y_β, y_γ, y_δ, y_ϵ, and y_ζ in ascending order.

At $y = y_\alpha$, there is one connected component, so a zeroth degree feature is born. At $y = y_\beta$, another zeroth degree feature is born. $y = y_\gamma$ does not change the number of connected components, and therefore the persistent homology does not contain information regarding v_2. At $y = y_\delta$, the connected components born at y_α and y_β merge. Using the elder rule, the one that persists is the component at y_α. Also at y_δ, a new component is formed by v_4. y_ϵ does not change the number of connected components, and finally y_ζ sees the merging of the two alive components by v_6, which then lives to infinity. The zeroth degree persistent homology can therefore be summarised as the following three features:

1. A feature born at y_α, that lives to infinity
2. A feature born at y_β, that dies at y_δ

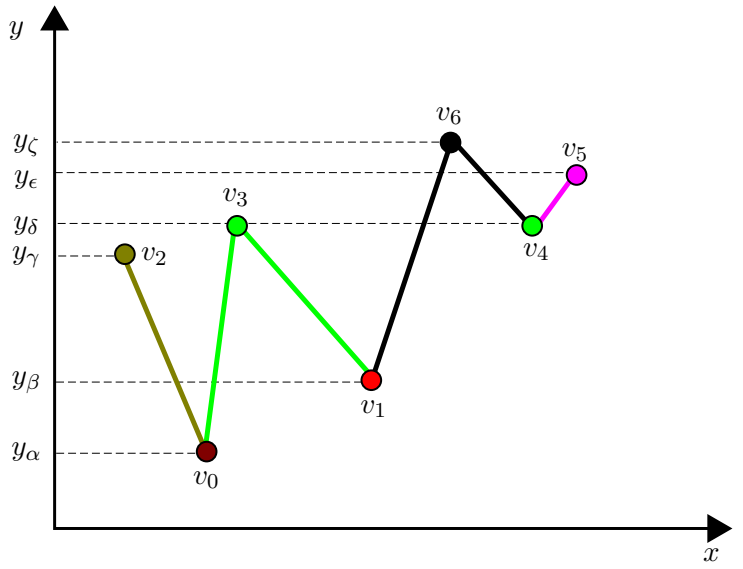


FIGURE 4.2: A one-dimensional simplicial complex, with geometric realisation matching its associated height function. The colour of the simplex is determined by the value of its height function. By observing how the homology of sub- and super-level sets of the simplicial complex change as a function of height, critical points of the height function can be found.

3. A feature born at y_δ, that dies at y_ζ

Upon further inspection, it is apparent that this filtration contains information regarding the maxima of the height function. In particular, the non-trivial features detail the value of the maximum, as well as its closest minimum. Information about minima can also be gained by inverting the height function.

$$\tilde{f}(v) = \max_{\nu} f(\nu) - f(v) \tag{4.10}$$

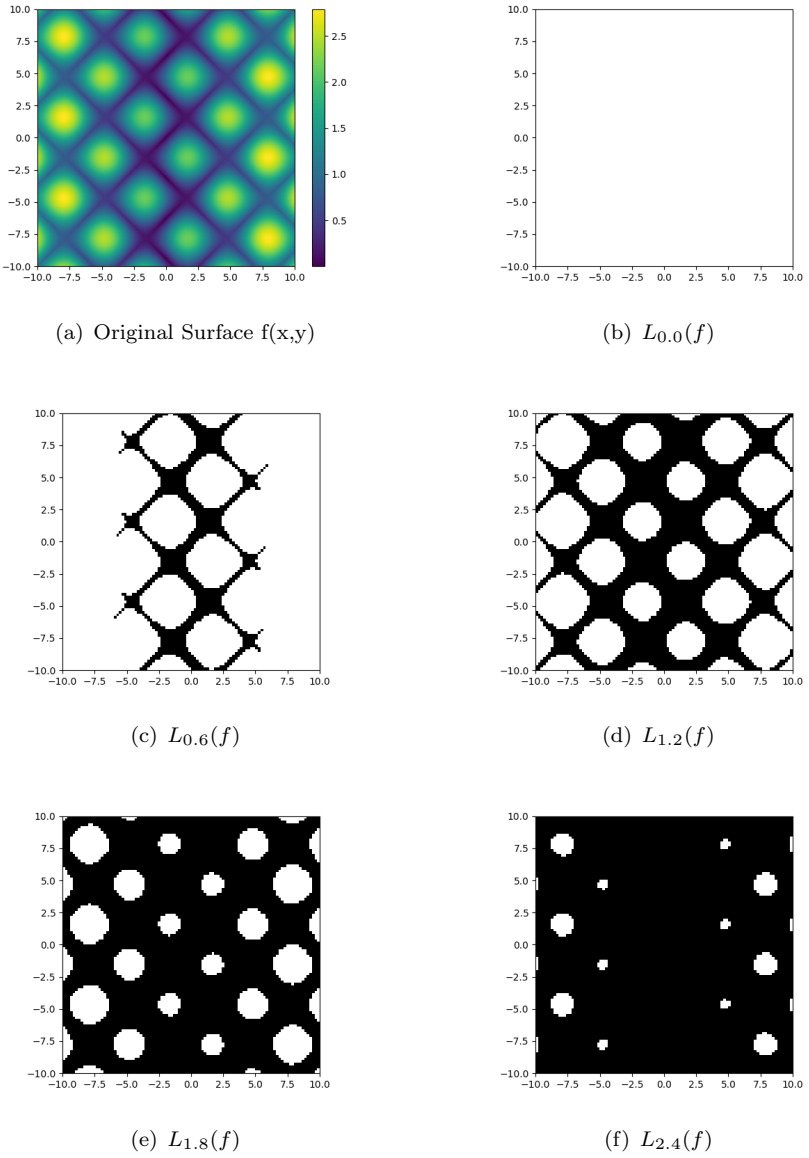
where to avoid confusion, ν is used to represent the vertices when finding the maximum. $\tilde{f}$ can be extended to higher order simplices in a similar manner to previously, and persistence can therefore be taken of $\tilde{f}(\sigma)$.

This discussion can be extended to higher order simplicial complexes, such as those equivalent to surfaces. An example can be seen in Figure 4.3, which uses the two dimensional function:

$$f(x, y) = |\sin(x) + \sin(y)| + \frac{|x|}{10}$$

chosen due to its different height minima and maxima.

By $L_{0.6}(f)$, there is a single zeroth degree component, and several first degree components. Each first degree component encloses a maximum of f. These components have disappeared by $L_{2.4}(f)$, implying that the enclosed maximum is within the sublevel set. Notice that this is different to the one-dimensional case, where criticality was found using



(a) Original Surface f(x,y)

(b) $L_{0.0}(f)$

(c) $L_{0.6}(f)$

(d) $L_{1.2}(f)$

(e) $L_{1.8}(f)$

(f) $L_{2.4}(f)$

FIGURE 4.3: A function defined over a surface, and a series of its sublevel sets. Within the sublevel sets, black regions should be considered as 'real', whereas the white regions are those not within the set. Again, differences in homology allow critical points to be determined.

zeroth degree homology. A reasonable rule of thumb is that for an m-dimensional region of a simplicial complex, critical values are contained in $(m-1)$-dimensional information.

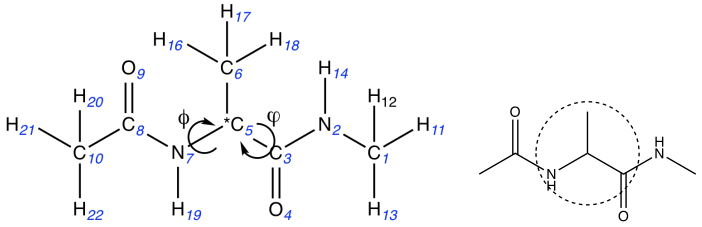
Returning to conformational spaces, the general function of importance is the single point energy of a conformer. The underlying simplicial complex, and the energy function chosen, are discussed in more detail for each analysis individually.

4.5 Molecule Sets

This section will detail the molecules studies in this chapter. In particular, it will explain the rationale of choosing each molecule, as well as detail which features of conformational spaces will be studied with each molecule. This section will also explain the nuances in generating each conformer set.

4.5.1 Alanine Dipeptide

Alanine dipeptide (Figure 4.4) is a commonly studied molecule, in particular within enhanced sampling [169, 170, 171, 172, 173, 174, 175]. This is due to its 'well understood' conformational space, and asymmetric free energy landscape. The conformational space



(a) With atoms, torsion angles and chiral centre labelled

(b) With alignment core highlighted

FIGURE 4.4: The alanine dipeptide molecule, with chiral centre and alignment core highlighted.

of alanine dipeptide is widely stated to be that of a torus [176, 162]. This can be understood from Figure 4.4(a). The bond between atoms 5 and 7 is a free torsion, likewise with the bond between atoms 5 and 3. These bonds can rotate around a full circle. In contrast, the C-N bonds (between atoms {7,8} and atoms {2,3}) have a restricted rotation, due to the amide bond resonance. These bonds are fixed to a planar geometry. As there are two free torsions, the conformational space is considered to be their product: $S^1 \times S^1 = T^2$, the torus. However, this assumes that other degrees of freedom, such as bond bending and stretching, do not contribute to the conformational space - the rigid geometry hypothesis [157].

When generating the conformational space for alanine dipeptide, care has to be taken to ensure the conformational space is both physically correct and matches chemical intuition. Firstly, the chiral centre on atom 5 has to be dealt with. Proposition 3.6 of [30], found by Ingrid Membrillo-Solis, is that different each chiral centre leads to a pair of path connected components of the space. This corresponds to the chemical notion of a lack of interchange between enantiomers. Within this work, a conformational space has to be path connected, and therefore the chirality of the molecule must be fixed. This work fixes the chiral centre to have the S chirality.

Another issue with generating the conformational space is due to the presence of the amide bond. As mentioned, the amide bond is found in a planar geometry. The *trans*-isomer is more favoured, however both isomers are found [177]. If a hypothetical conformer set contained both isomers, there would not automatically be a path between them. This could be rectified if a path was manually added in some manner, with the conformational space changing accordingly. For example, a path could be added by manually adding conformers with the amide torsion angles being distributed on $[0, 2\pi)$, and this would add another pair of circles to the conformational space product. Rather than deal with this added complication, this work simply fixes the amide bonds to the *trans*-isomer.

The conformer set for alanine dipeptide is generated using the procedure described in Figure 4.1. Specifically:

- **Step 1**: Limit initial pool to include only the S-enantiomer.
- **Step 2**: Ensure free torsions are distributed on $[0, \pi)$. Also ensure that amide bond is fixed to trans isomer
- **Step 4**: For the Euclidean representation, align only core of the conformers (as shown in Figure 4.4(b)). This *is not* the case for the RMSD representation.

The resulting conformer set has 9112 conformations of alanine dipeptide. The alanine dipeptide molecule is used to test the following:

- Do the presence of hydrogens significantly alter the topology of the conformational space?
- Does the difference in alignment procedure between the Euclidean and RMSD representation lead to differences in the conformational space topology?

4.5.2 Pentane

The pentane molecule (Figure 4.5) is used in this work due to its structural simplicity. Under the rigid geometry hypothesis, there are again two free torsions. This leads to the same toroidal topology as previously. Unlike alanine dipeptide, there is a plane of symmetry within pentane, leading to the two free torsions to be identical. The effect of

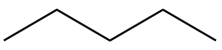
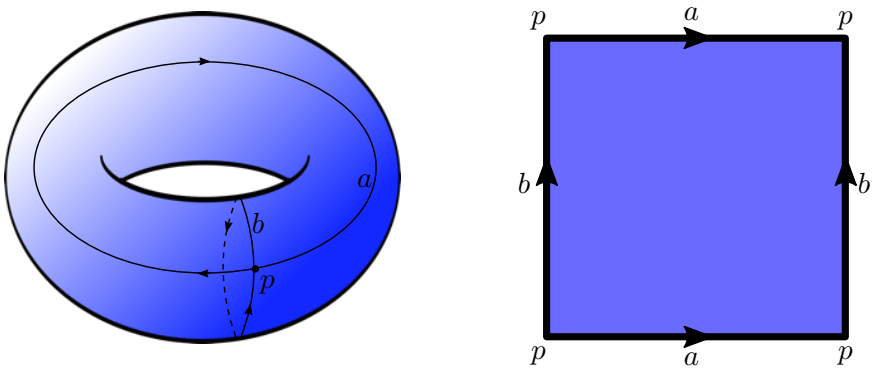


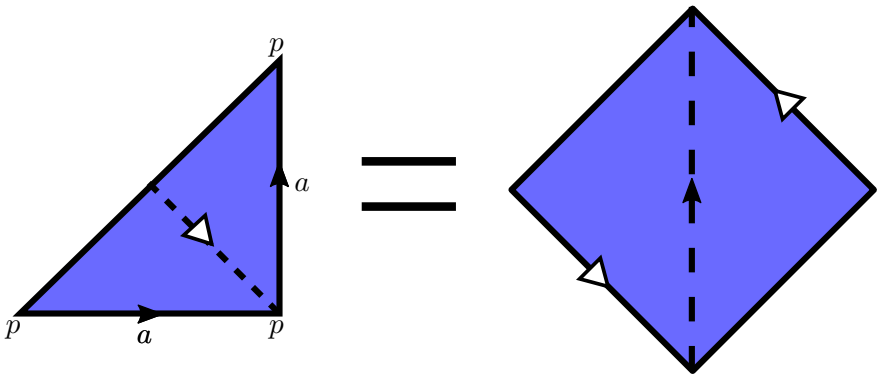
FIGURE 4.5: Skeletal formula of pentane

this symmetry can be understood pictorially. The image below shows a torus. Labelled on the torus are a single point p, and a basis for the two first degree homological features

that intersect p, labelled a and b. The torus can be 'flattened' by cutting along a and b. This is the process of creating a CW-complex for the torus, and leads to the square on the right.



The next step is to recognise that the symmetry being added identifies the two loops a and b as identical. This is analogous to folding the square along the diagonal, leading to the triangle. Finally, the shape is stretched, resulting in the final space. By gluing the edges back together, such that they match orientations, the resulting space is topologically the Möbius band, seen in Figure 4.6.



The conformer generation procedure for pentane is again as described in Figure 4.1. However, the specifics vary when compared to the alanine dipeptide procedure.

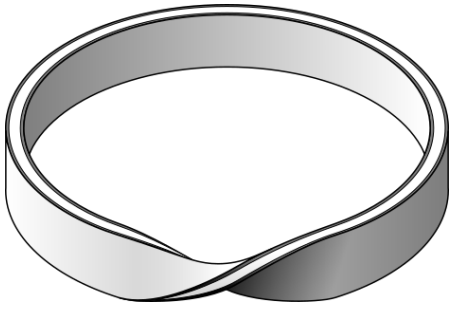


FIGURE 4.6: The Möbius band

- **Step 2**: Ensure free torsions are distributed on $[0, \pi)$.

- **Step 4**: Perform two alignments, to the entire carbon chain:
  - Align conformers such that their indices match
  - Align conformers such that the reverse of the indices match
- **Step 4** cont: Create two of each representation:
  - *Index Align*: The representation with the matching index alignment
  - *Min Align*: The representation where the minimum of the two alignments is used

The above procedure led to a set of 9108 pentane conformations. The pentane molecule is used to test the following:

- Is persistent homology able to detect the presence of symmetry?
- How does this vary between the two representations?

4.5.3 Cyclooctane

The skeletal formula for cyclooctane can be seen in Figure 4.7, and it may seem a strange molecule of which to study the conformational space when compared to the other two described in this work. However, the conformational space of cyclooctane has previously been studied by Shawn Martin *et al* [166, 165], and therefore provides a useful molecule with which to test persistent homology techniques.

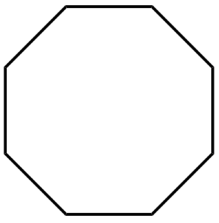


FIGURE 4.7: Skeletal formula of cyclooctane

Under the rigid geometry hypothesis, it is well established that a loop of n nodes has $(n-6)$ degrees of freedom [178]. This can be explained when it is considered that the first 3 nodes define a plane, resulting in $(n-3)$ free torsion angles. However, the process of joining the ends of the loop restricts 3 degrees of freedom (2 angles and the length), resulting in $(n-6)$ total degrees of freedom. For cyclooctane, it is therefore expected that there are two degrees of freedom. The most common set of topological objects with this property are closed surfaces, and the classification of such objects is well-established. Although a proof is not discussed here, the classification theorem of closed surfaces states that any closed surface is homeomorphic to one of the following:

1. The sphere S^2
2. The connected sum of g tori T^2
3. The connected sum of k real projective planes $\mathbb{R}P^2$

These surfaces have well-known Betti numbers, found in Table 4.1. Thus it would be

Surface	β_0	β_1	β_2
S^2	1	0	1
Connected sum of g tori	1	$2g$	1
Connected sum of k real projective planes*	1	$k-1$	0

TABLE 4.1: The Betti numbers β_n for the classification of closed surfaces. *Technically this depends on the field, see Appendix D.

hoped that the conformational space of cyclooctane could be classified with persistent homology.

Before the discussion on how this work aims to study cyclooctane, it is worth describing the two articles [166, 165] in more detail. For these works, the authors use other techniques to determine the conformational space of cyclooctane (persistent homology was a fairly recent development at the time of writing of these articles, and it is unclear if the authors were aware of its existence). The authors firstly generate a conformer set using a loop closure algorithm [179], before using the Euclidean representation (as named in this thesis) to create a point cloud in $\mathbb{R}^{72}$. The authors then analyse this representation using a combination of Isomap [180], and the generation of a single triangulation of the conformational space - this is a particular area where persistence would have been of assistance. The authors argue that the resulting low dimensional representation is actually *not* one of the surfaces defined above. In particular, they find that the conformational space of cyclooctane is in fact the union of S^2 and the connected sum of two copies $\mathbb{R}P^2$, topologically the Klein bottle KB. The authors found that the two different closed surface intersect in two copies of S^1, the circle. The Isomap embedding found in [165] can be seen in Figure 4.8. It was Hendrickson who found in his 1967 work

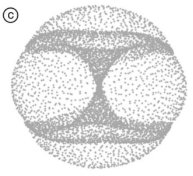


FIGURE 4.8: The Isomap embedding of the conformational space of cyclooctane. Reproduced from Figure 1 of [165]. The hypothesised spherical component can be seen, with the Klein bottle component twisted in such a way as to make it look like an hourglass.

on molecular conformation the existence of ten particular cyclooctane conformations [181]. Martin *et al* were able to use the geometry of the cyclooctane to understand the

rates of relevant transitions between these states, and in particular why the boat-chair conformation is significantly more common than the crown conformation, even though there is a difference of approximately 1kcal/mol.

However, this work, and in particular the conclusions regarding the nature of the conformational space, require further study. Firstly, it is worth checking that the results obtained are not due to the representation chosen. Secondly, the authors fixed a triangulation of the space, which may impact their results. Previously, it had been thought that the conformational space of cyclooctane could be respresented as the union of a sphere and a torus [182] - Martin *et al* argued this was due to a limited sample of conformations (supplementary information of [165]), and also the the choice of Cremer-Pople ring puckering coordinates for the representation [183].

The cyclooctane set used in this work is a subset of the original data set used in [165]. In particular, 6040 conformations, where no two conformations have an RMSD < 0.05 were obtained from Shawn Martin directly. For cyclooctane, the following is tested:

- Is the conformational space that of a Klein bottle intersecting a sphere?
- Is this representation dependent?

The study of this cyclooctane data set has been of recent interest to other researchers in the field of topological data analysis. For example, Stolz *et al* have developed topological methods for geometric anomalies (such as those supposedly found in the circular intersections of the conformational space), and tested it on this data set [184]. However, the authors do not draw any chemical conclusions from their studies, instead just describing their methodology.

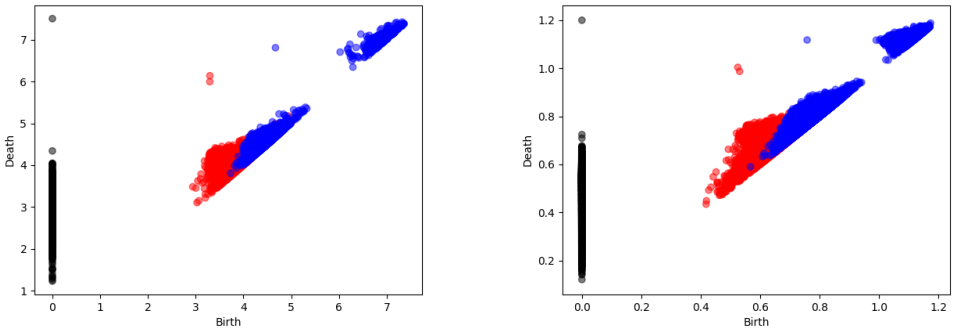
4.6 Results

4.6.1 Alanine Dipeptide

The persistence diagrams for the different representations of conformers can be seen in Figure 4.9. For this molecule, both representations are similar. This is due to the alignment procedure in the Euclidean representation - by aligning to the defined core, as opposed to the entire molecule, the motion of the free torsions are emphasised. Within both representations, there are multiple discussion points from the persistence diagram.

The first set of features to note are the longer lived features, which have multiplicity $(1, 2, 1)$. These are the Betti numbers of a torus - as suggested by the rigid geometry hypothesis. This is perhaps unsurprising, as by focusing on the heavy atoms only, a large number of the degrees of freedom are missing, which may have contributed a large

amount of noise. The two first degree long-lived features do not have matching birth and death values. This reflects the asymmetry of the two flexible torsions in the alanine dipeptide molecule - which leads to circles of two different radii. This can lead to issues in molecular simulations, such as in creating data-driven collective variables [175].

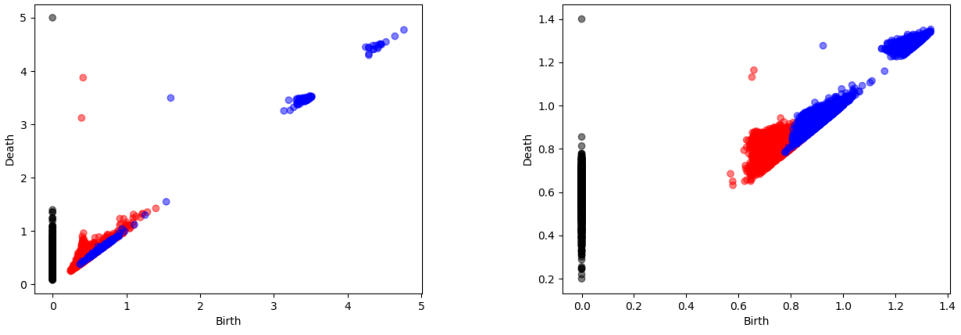


(a) Euclidean Representation (b) RMSD Representation

FIGURE 4.9: Persistence diagrams for the two different representations of the all-atom conformational space of alanine dipeptide. Black, red and blue correspond to zeroth, first and second degree homology respectively.

There are two large clusters of second degree features. The cluster born earlier corresponds to simplices between adjacent conformations. However, the second cluster is due to the 'filling in' of the basis circles of the torus. This can be seen from the fact that the birth values of this second cluster match the death values of the long-lived first degree features.

It is worth seeing if these results are reflected with the heavy-atom conformational space. The corresponding persistence diagrams can be found in Figure 4.10. The RMSD rep-



(a) Euclidean Representation (b) RMSD Representation

FIGURE 4.10: Persistence diagrams for the two different representations of the heavy-atom conformational space of alanine dipeptide. Black, red and blue correspond to zeroth, first and second degree homology respectively.

resentation is largely unchanged, as the relative contribution to the RMSD of hydrogen

atoms is small. However, this is not the case for the Euclidean representation the asymmetry of the two free torsions is much more pronounced. This is due to two factors, firstly, hydrogen number 12 of Figure 4.4(a) sweeps a far larger circle than the corresponding hydrogen number 19. The second difference is due to computational difficulties. In particular, the heavy-atom Euclidean representation had to be subsampled, as the persistence procedure was found to be too memory intensive to be performed on the supercomputers available. These two factors lead to the pronounced difference in the all-atom and heavy-atom persistence diagrams for the Euclidean representation.

However, it is also noted that the persistent Betti numbers appear to be unchanged for the all-atom system. This suggests that the presence of hydrogen atoms do not seem to add too much noise to the metric spaces of the RMSD and Euclidean representations - i.e. the conformational space of molecules is not influenced by the extra degrees of freedom caused by including hydrogen atoms.

4.6.1.1 Persistence of the Free Energy Surface

Using molecular simulation, it is possible to create the free energy surface for the two torsional degrees of freedom of alanine dipeptide. In particular, metadynamics [185] over the two torsion degrees of freedom allows recovery of the free energy. Using the GROMACS [186] and PLUMED [187] software packages, the free energy surface was calculated with assistance from Khaled Abdel-Maksoud (Figure 4.11), using the methodology defined in [188]. This matches the free energy surface found in sources such as [172, 189, 162]. Such a surface can have its critical points analysed via persistent homology.

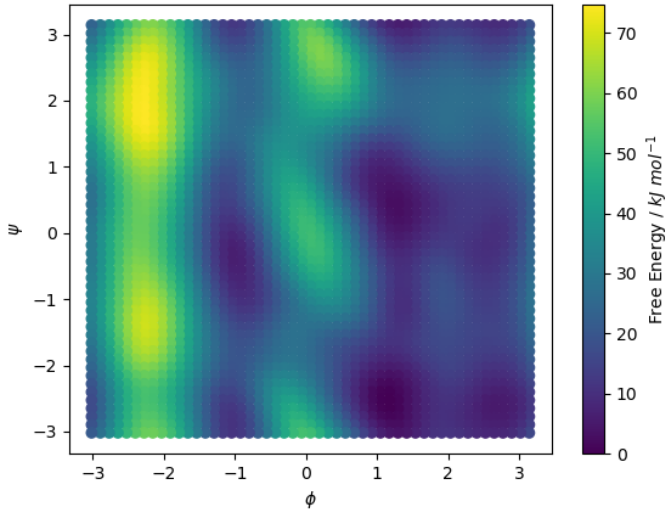


FIGURE 4.11: The free energy surface of the two free torsions in alanine dipeptide, as calculated with metadynamics.

Firstly, the simplicial complex must be defined. Beginning with the vertices, a (50×50) grid of (ϕ, ψ) coordinates was generated, before the free energy calculated on these vertices. Next, 1- and 2-simplices must be determined in a way that recreates the toroidal topology of the (ϕ, ψ) space. The method for 1-simplices can be seen in Figure 4.12. Each vertex is connected to its four neighbours, as well as to its north-eastern neighbour. The boundary of the grid is also connected such that rather than a plane, the topology is that of the torus. 2-simplices are then defined by filling in all of the smallest triangles, and the energy of higher order simplices is found using the relationship in Equation 4.8.

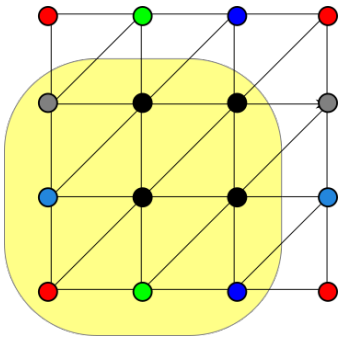
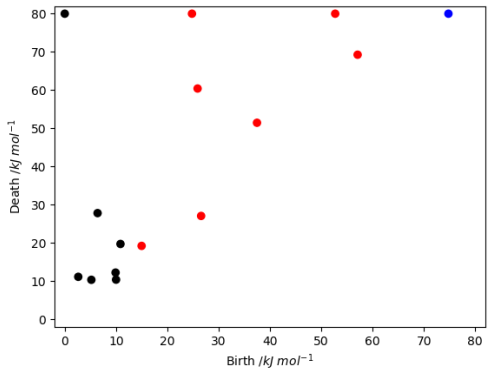


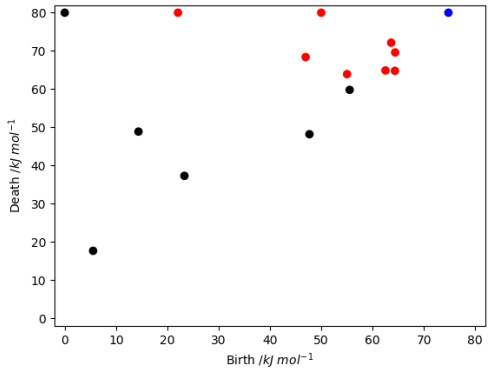
FIGURE 4.12: The basic method for creating a simplicial complex with toroidal topology. Nodes of the same colour (excluding black) are identified. The basic unit is within the shaded area. Provided the basic unit has at least 3 vertices connected in this way, this is a valid simplicial complex.

Persistence diagrams of the free energy surface and its related inverted form (Equation 4.10) can be seen in Figure 4.13. Both persistence diagrams have the persistent Betti numbers (1,2,1). Here, this is a result of the construction of the simplicial complex itself, rather than an emergent property of the data set as previously, such as in Figure 4.9(b). The first degree persistent features do not have matching birth times, which is due to two effects. The first is the use of the maximum in Equation 4.10, which essentially sets the 'zero' of energy. This choice is essentially arbitrary, but is used to ensure $\tilde{f}$ is strictly positive. Altering this would shift all values birth and death times.

The second effect can be understood by considering the nature of these first degree features. In particular, these features correspond to the basis for the cycles on the torus, and are born when a complete path can be made around its periodic boundary. Specifically, with the persistence of the free energy surface itself, the birth value of the first of these features that are born describes the minimum energy path that crosses a periodic boundary of the torus. In contrast, the same feature of the inverted surface describes the maximum energy path - however care must be taken when extracting this information that $\tilde{f}$ is transformed back into the free energy surface itself. In summary, whereas the short-lived first degree features contain information about extrema, the persistent feature describes a complete loop on the free energy surface, in particular over the periodic boundary.



(a) Free Energy Surface



(b) Inverted Free Energy Surface

FIGURE 4.13: Persistence diagrams of the free energy surface and its inverted form. The persistence of the free energy surface contains information about maxima, whereas the inverted form describes minima. Black, red and blue correspond to zeroth, first and second degree homology respectively.

Looking at the other first degree features in more detail, a few things can be noted. Firstly, rather than use the distance from the diagonal as a metric for significance, instead the persistence $p = \delta_d - \delta_b$ is used. Basic geometry shows that these quantities are proportional, however the persistence directly measures a useful quantity regarding the extrema. For example, in the case of the surface itself, the persistence describes how much higher a maximum is from its lowest energy minimum. Also, notice that the matching between maxima and minima is not trivial. This is much simpler in the one dimensional complex case, the matching can often be done by eye.

However, it is sometimes possible to match topological features to features of the energy landscape, by seeing which vertices (and higher order simplices) are inserted into the complex at any given time. Take, for example, the feature in Figure 4.13(a), at coordinate $(26, 60)$. From the free energy surface in Figure 4.11, it can be seen that this is likely to be the feature at $\phi \approx 0$, $\psi \approx 2.7$. The two maxima at $\phi \approx -2.3$, although they are the highest free energy, are not particularly persistent, with the persistence

diagram feature at $(57, 70)$ corresponding to the maximum at $(-2.3, -1.4)$. This is due to the presence of the saddle point at $(-2.3, 0)$ - which leads to complications for morse persistence. Furthermore, the global extrema are often not found in this procedure, as rather than closing an $(n-1)^{th}$ degree hole, they close the surface itself, corresponding to a n^{th} degree feature.

Similarly, the feature in Figure 4.13(b) at $(47, 68)$ can be seen to correspond to the minimum at $(-1, -0.4)$ through the same procedure. The maxima and minima found are highlighted in Figure 4.14, where it can be seen that the persistence of these features is not simply dependent on the extremum value, but the relative difference between extrema values.

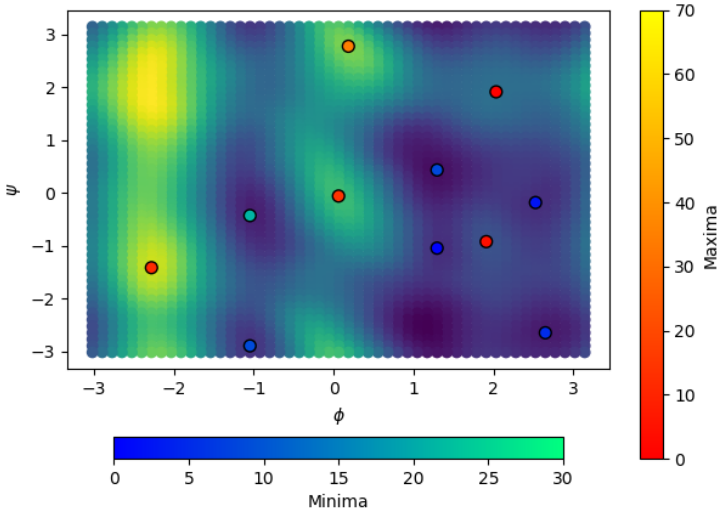
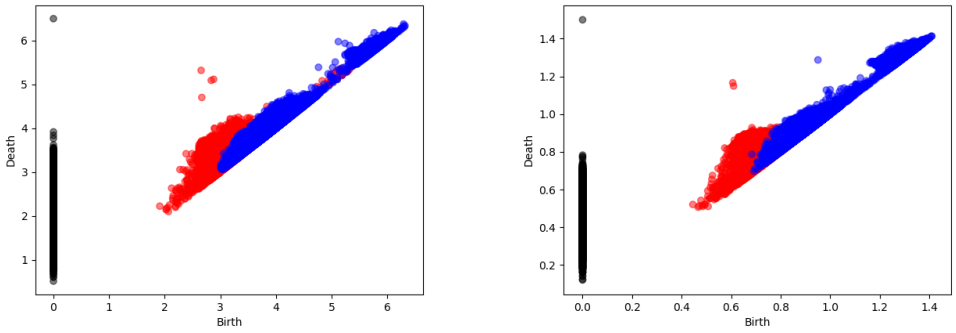


FIGURE 4.14: The free energy surface of the two free torsions in alanine dipeptide, with extrema found by persistence highlighted. Points are coloured by their persistence values, with larger values corresponding to minima which are much deeper than their respective maxima and vice-versa.

4.6.2 Pentane

The persistence diagrams for the different representations, *index align* can be seen in Figure 4.15. The RMSD representation has qualitatively the correct persistent Betti numbers, of (1,2,1). However, the persistent Betti numbers are not the same for the Euclidean representation. β_2 is harder to discern, with β_0 and β_1 appearing to be 1 and 4 respectively. The conformational space of pentane, without symmetry taken into account, should be toroidal. This is what is seen in the RMSD representation, which was previously stated to be the most accurate. Clearly, the methodology of the Euclidean representation has led to an incorrect conformational space. One potential cause is that the Euclidean representation has led to each rotatable bond contributing a circle to

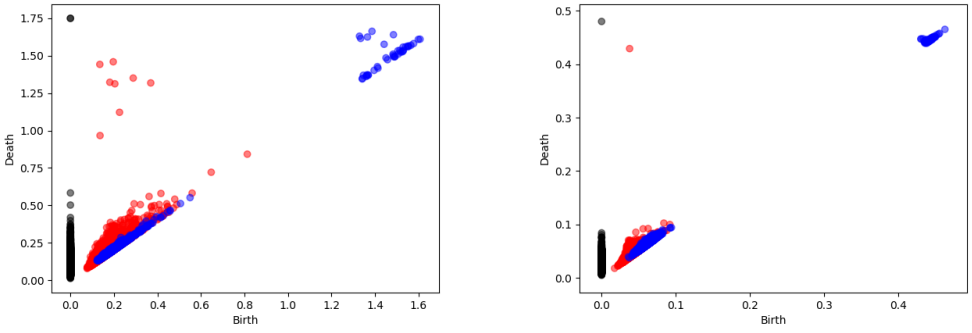


(a) Euclidean Representation (b) RMSD Representation

FIGURE 4.15: Persistence diagrams for the two different representations of the heavy-atom conformational space of pentane, without molecular symmetry taken into account. Black, red and blue correspond to zeroth, first and second degree homology respectively. The Euclidean representation does not correctly identify the expected toroidal conformational space.

the conformational space, as opposed to the two truly free torsions. This could lead to a space with $\beta_1 = 4$. However, it is emphasised that this conformational space is fundamentally incorrect, and therefore any interpretation of it should be viewed with scepticism.

The persistence diagrams for the different representations, *min align*, can be seen in Figure 4.16. Here, the Euclidean representation has broken down entirely. The persistent β_0 is now equal to two, i.e. there are two connected components in the space. These components are clearly showing which alignment to the reference has been found - with matching or opposite indices. If the opposite index matching is used, it is not surprising



(a) Euclidean Representation (b) RMSD Representation

FIGURE 4.16: Persistence diagrams for the two different representations of the heavy-atom conformational space of pentane, with molecular symmetry taken into account. Black, red and blue correspond to zeroth, first and second degree homology respectively. The Euclidean representation again fails to capture the correct topology of the space. The RMSD representation now has different persistent Betti numbers, illustrating that the presence of symmetry changes the underlying topology.

that this leads to a much greater Euclidean distance in this representation, as can be seen from the previous discussion on conformational space representations. This is an artefact resulting from the fact that the Euclidean representation fundamentally requires the indices of atoms to match in its coordinate definition.

In contrast, the RMSD representation has persistent Betti numbers that appear to be $(1, 1, 0)$. Although this matches the Möbius band, as was the expected topology, the Betti numbers also match a circle S^1. However, the conformational space can be verified using MDS, in three dimensions (as both of these manifolds can be embedded in this dimension). This can be seen in Figure 4.17, where there is a clear twist in the embedding. This suggests the RMSD representation can correctly identify the conformational space, even when taking molecular symmetry into account.

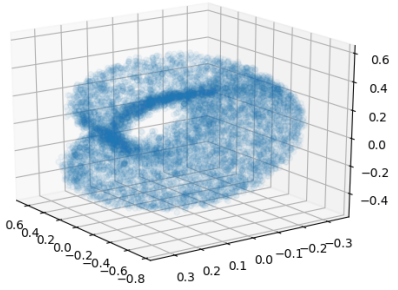


FIGURE 4.17: MDS projection of pentane's RMSD metric with symmetry taken into account. A twist is visible, suggesting a Möbius band topology.

In summary, from the analysis of the conformational space of pentane, it can be seen that when possible, the RMSD representation should be used to ensure the correct topology of the conformational space. The studies on alanine dipeptide did however show that the Euclidean representation can capture the same topology, but it should never be assumed. Therefore, considering that the previous work on cyclooctane [165] utilised the Euclidean representation, it is important to validate their work, and see if their conclusions on the non-manifold nature of the conformational space is correct. Furthermore, it is also now possible to investigate the effect of symmetry on the conformational space of cyclooctane.

4.6.3 Cyclooctane

The persistence diagram for the RMSD representation of the conformational space of cyclooctane can be seen in Figure 4.18. The persistent Betti numbers are seen to be $(1, 1, 2)$. Immediately, this shows that the conformational space for cyclooctane is non-manifold. This is a consequence of Poincaré duality, which states that for an n-dimensional manifold M, the homology groups $H_p(M)$ and $H_{n-p}(M)$ are isomorphic, and therefore have

matching Betti numbers. In this case, one would expect β_0 to be equal to β_2 (as has been found for all of the previously studied conformational spaces). As this is not the case, the conformational space of cyclooctane does not appear to be a manifold.

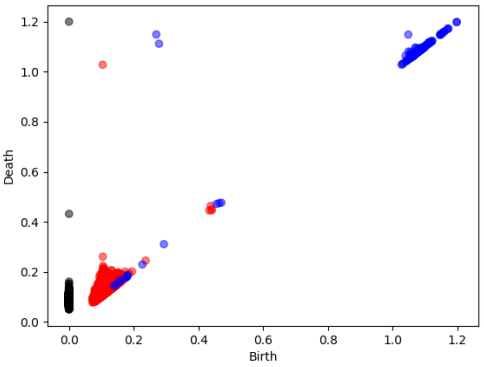


FIGURE 4.18: Persistence diagram for the RMSD representation of the conformational space of cyclooctane. Black, red and blue correspond to zeroth, first and second degree homology respectively. The persistent Betti numbers of (1,1,2) suggest a non-manifold topological structure.

4.6.3.1 Separation of Manifold Components

Given that the conformational space is non-manifold (as expected from previous work), the next logical step is to separate it into manifold components. Local PCA is a potential route to achieving this, with an example seen in Figure 4.19. For each point, define a neighbourhood according to some metric. Then, PCA is calculated for this subset, and in particular information can be extracted from the number of non-zero singular values of the correlation matrix to understand the local dimension around that point. If the local dimension is different from what is expected, the point is removed from the original set. This may separate the space, but each component can then be analysed.

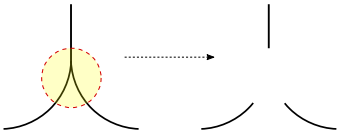


FIGURE 4.19: Cartoon illustrating how non-manifold points can be removed to leave a (disconnected) set of manifold structures. The yellow circle is used to demonstrate a neighbourhood which is non-manifold, which when removed from the left image leads to three manifold components on the right.

In the case of cyclooctane, the local dimension is expected to be two, as previously discussed. However, the local dimension analysis found points with local dimensions of two and three. These *singular* points were extracted from the set, and PCA was performed on this set. Clustering was also performed on the high dimensional representation, with the projection coloured by cluster seen in Figure 4.20. It is clear that there are two

circular paths. This compares favourably to the conclusions of [165], that the sphere and Klein bottle components (which are yet to be demonstrated in this work) intersect in two circles.

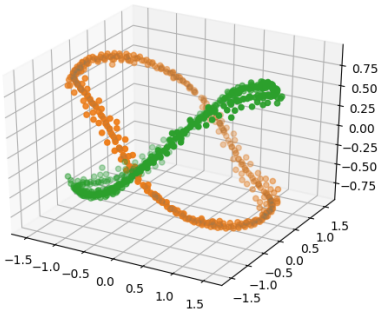


FIGURE 4.20: Three-dimensional PCA of the Euclidean representation of cyclooctane conformers with local dimension three. Points are coloured based on the result of a clustering analysis, performed in the high-dimensional space. These match the hypothesised intersection circles from Martin *et al*'s original cyclooctane work.

Having removed the singular points from the set, the next step is to separate the remainder of points into their individual components. First, it is hypothesised that the removal of the singular points leads to regions of low point density. This suggests that a density-based clustering method (such as DBSCAN [190]) could be used to separate the space. However, DBSCAN requires an estimate for the density *a priori*, which is a major drawback. In particular, it assumes that all clusters are of the same density, and that all regions of the cyclooctane conformational space are sampled with the same efficiency. Instead a hierarchical DBSCAN (HDBSCAN [191]) was used. In essence, HDBSCAN allows clusters of different density to be found, by finding regions of relatively high density, and allowing the definition of 'relative' to change locally. HDBSCAN was performed on the set of non-singular points, before PCA was used for visualisation. This can be seen in Figure 4.21, where the hypothesis of using relative drops in density to find clusters can be seen to be validated.

Perhaps remarkably, the system seems to separate into four neat clusters (clusters 1-4), as well as a group of points that is less organised (cluster 5). The number of conformers for each cluster can be found in Table 4.2. Having found these clusters, the next step

Cluster Number	n_{confs}
1	426
2	3381
3	432
4	810
5	176

TABLE 4.2: The number of points found for each cyclooctane conformational space cluster. The remainder of points are found in the singular clusters.

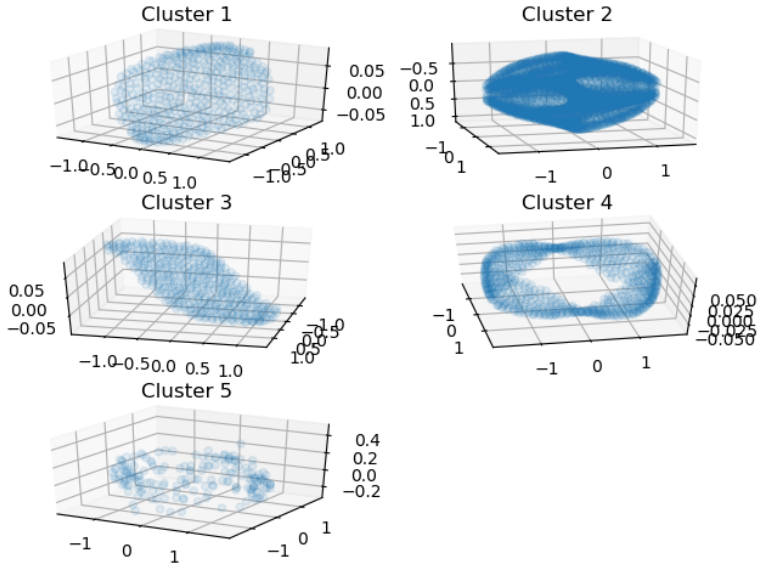


FIGURE 4.21: Clusters of the non-singular points of the cyclooctane conformational space, as found by HDBSCAN. Visualised by performing a 3-dimensional PCA on the set of non-singular points, before viewing each cluster separately. These clusters can then be matched to separate the original space into manifold components.

would be to 'glue them back together'. The process of removing the singular points would have not only separated the underlying manifold components, but it may have split them further than necessary. In this case, rather than through a process of trial and error, it is fairly straightforward to understand by eye which clusters slot into each other. Clusters 1, 3 and 4 appear to fit together, and cluster 2 appears to stand alone. Therefore, these were chosen to be glued back together. Further, the singular points discovered previously were also included in each group.

To understand what structures are expected after the gluing procedure, rather than perform PCA on the entire set, PCA was performed on each cluster separately. This is found in Figure 4.22. Firstly, it is apparent that cluster 5 matches the pattern found for the singular points in Figure 4.20. Given that the paths have already been established and clustered, cluster 5 is removed from the set and is not further analysed. On the other hand, clusters 1 and 3 seem to be spherical caps and cluster 4 could be the band between the singular points. It is harder to determine the topology of cluster 2. Fortunately, persistent homology provides a route for analysis.

4.6.3.2 Analysis of Clusters

The persistence diagram for the RMSD representation on the group of molecules formed by clusters 1, 3 and 4 can be found in Figure 4.23(a). The persistent Betti numbers are (1,0,1), suggesting that the sphere suggested in [165] has been found. Similarly cluster 2 (persistence in Figure 4.23(b)) was speculated to be a Klein bottle. The persistent

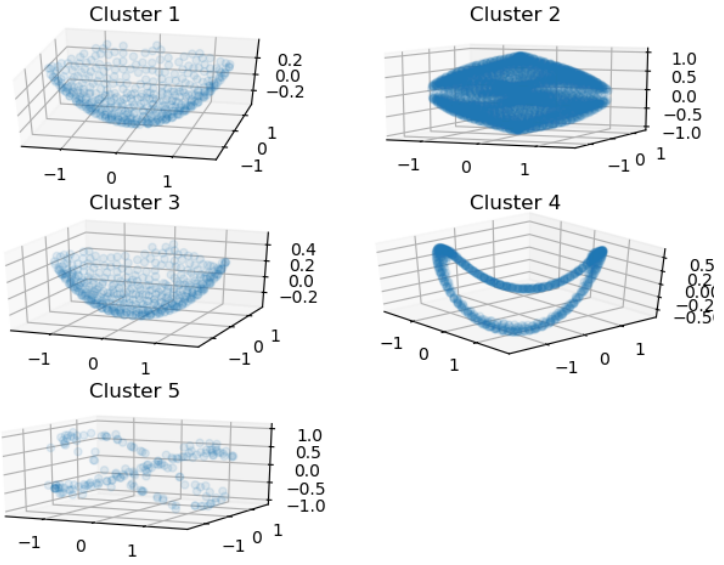
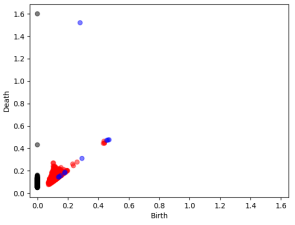
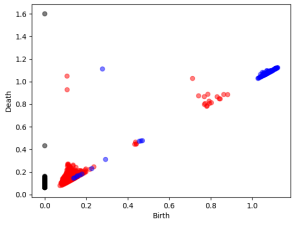


FIGURE 4.22: Clusters of the non-singular points of the cyclooctane conformational space, as found by DBSCAN. Visualised by performing a 3-dimensional PCA on each cluster separately.

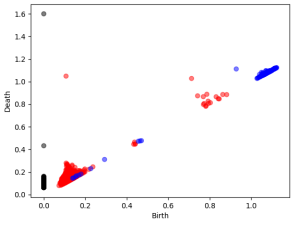
Betti numbers, (1,2,1) suggest either a torus or Klein bottle. By using a different field of coefficients, in this case $\mathbb{Z}_3$, persistent homology can now distinguish between these two manifolds (see Appendix D for details). The persistent Betti numbers, now (1,1,0), support the Klein bottle hypothesis. Therefore, the original results found in [165] have



(a) Clusters 1, 3 and 4



(b) Cluster 2, coefficients in $\mathbb{Z}_2$



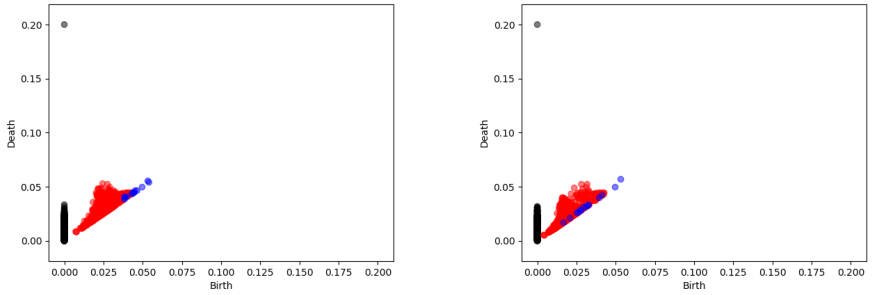
(c) Cluster 2, coefficients in $\mathbb{Z}_3$

FIGURE 4.23: Persistence diagrams of the RMSD representations of different groups of clusters in the cyclooctane conformational space. The results suggest the hypothesised sphere and Klein bottle components of the cyclooctane conformational space.

been validated using the stronger toolbox of persistent homology, allowing for fewer

assumptions. Furthermore, there is no reason why the methods outlined in this chapter could not be used for other molecules with more complicated conformational spaces.

The effect of symmetry on the different regions of conformational space can now be studied. The symmetries allowed in cyclooctane can be deduced from the molecular graph. Numbering the carbons around the chain $\{1, 2, 3, 4, 5, 6, 7, 8\}$ is identical to the numbering $\{2, 3, 4, 5, 6, 7, 8, 1\}$ and so on. Furthermore, the order can also be entirely reversed, equivalent to 'flipping' the molecule. This group can be identified as the dihedral group D_8, which has 16 elements. There are therefore 16 alignments for each pair of conformers, with the minimum pairwise RMSD metric used as the representation.



(a) Clusters 1, 3 and 4 (The Sphere) (b) Cluster 2 (The Klein Bottle)

FIGURE 4.24: Persistence diagrams of the RMSD representations of different groups of clusters in the cyclooctane conformational space, with the molecular symmetry taken into account. For completeness, the coefficients are taken in $\mathbb{Z}_2$, but the overall features are unchanged in $\mathbb{Z}_3$. The inclusion of symmetry effects leads to conformational spaces with trivial homology groups.

Both the Klein bottle and spherical components no longer appear to have any interesting topological features. The 'flip' symmetry of the molecule is identifying antipodal points on the sphere as identical. The resulting space is homeomorphic to the real projective plane $\mathbb{R}P^2$, which has Betti numbers $(1, 1, 1)$ with coefficients in $\mathbb{Z}_2$. Therefore, the extra symmetries of the cyclooctane molecule have changed the topology of the conformational space even further than the simple antipodal map.

4.6.3.3 Persistence of the Energy Landscape

Rather than using a new conformer set, such as the sampled points of a free energy surface seen previously, this section details a general method that could work on any system. Here, the conformational space is restricted to purely the spherical component, however there is no reason why this logic could not be extended to other components, or even the conformational space as a whole.

Firstly, a method of defining a simplicial complex with the correct topology, using the conformers of the spherical component as vertices, must be created. This can be accomplished using the persistence diagrams previously seen. Figure 4.25 demonstrates that at $\delta = 0.6$, the Rips complex has the topology of the Klein bottle. Therefore, if all the simplices with birth time less than 0.6 are used in its construction, the simplicial complex has the correct topology.

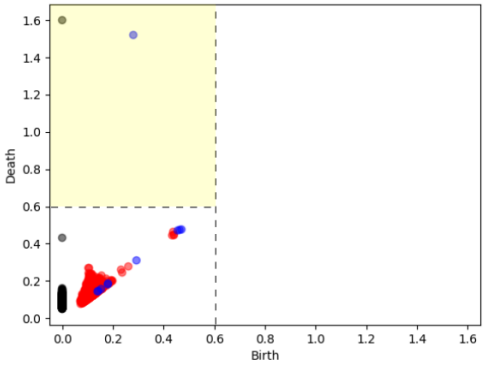
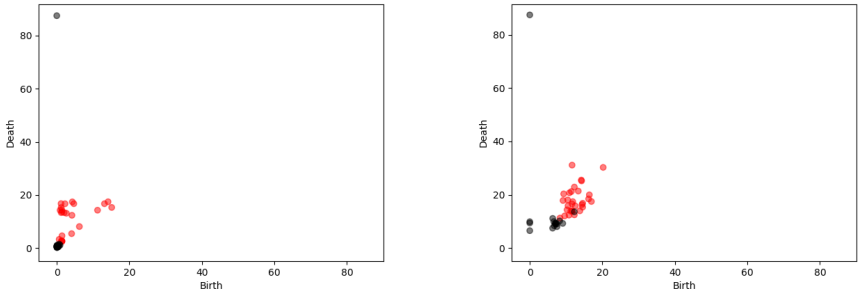


FIGURE 4.25: The two persistence diagrams of the spherical component. Highlighted are the regions of the persistence diagram demonstrating features that are still alive at $\delta = 0.6$. This suggests the single Rips complex at this value of δ will have a spherical topology. The energy function of points on this sphere can be defined, and critical points calculated.

The next step is to define a function on each of the simplices. As opposed to the free energy seen earlier, instead the single point potential energy (as calculated using the MMFF94 forcefield [150]) is used. This energy function may be poorly behaved, as two conformers with a small RMSD may have different energies - this will result in a noisy persistence diagram as a final result. However, the potential function can be extended to simplices in the same way as Equation 2.4, and persistence calculated. The two persistence diagrams, for the potential energy landscape and its inverted form, can be seen in Figure 4.26. As the sphere is a two-dimensional manifold, again it is the first degree components that correspond to extrema.

The persistence of the energy landscape itself (Figure 4.26(a)) details topological features corresponding to maxima. The minima are found from the first degree components of Figure 4.26(b). The simplex which closes the topological feature can be related to a single conformer. By projecting the spherical component of the conformational space into three dimensions, the location of these extrema in the conformational space can be illustrated, as seen in Figure 4.27. The found minima tend to be around the non-singular intersection loops. These correspond to the 'saddle' conformations described in [165]. Their relative 'peak' conformations are however not found. Reasons for this may include differences in potential energy function used (the original work utilised the MM3 forcefield [192] implemented in Tinker), or perhaps that these peak conformations are peaks relative to features in the Klein bottle space - this would merit further investigation.



(a) Potential Energy Landscape (b) Inverted Potential Energy Landscape

FIGURE 4.26: Persistence diagrams of the potential energy landscape and its inverted form. Each point corresponds to a different critical value of the energy function. Black and red correspond to zeroth and first degree homology respectively.

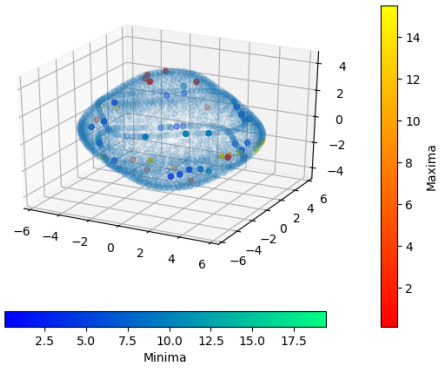
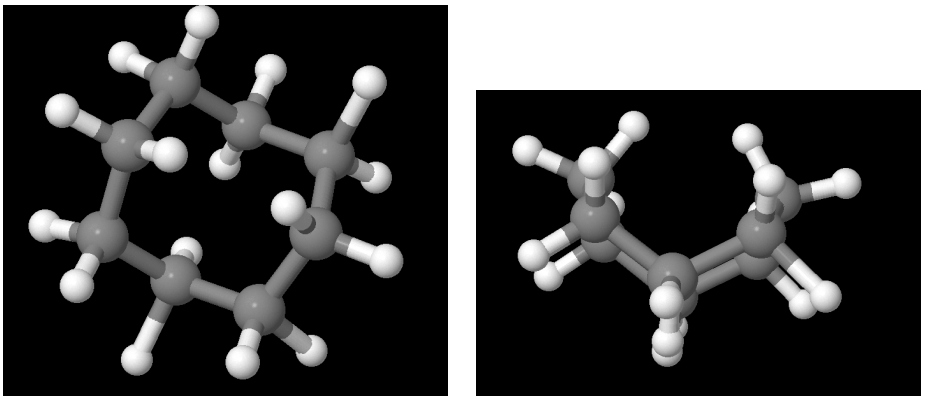


FIGURE 4.27: Three dimensional PCA projection of the spherical component of the conformational space. Highlighted points are the extrema found by persistent homology, which are coloured by their persistence values. Critical points are found in the correct region of the conformational space, as hypothesised by Martin *et al.*

The maxima that *are* found tend to be located near the poles (but are not the poles themselves). These correspond to the twisted-chair-chair conformations of cyclooctane, which are relative maxima to the chair-chair conformations. These tend to have a low persistence, implying that the energy gaps between these conformations are small. The more persistent maxima are found near the equator, and these correspond to the boat-boat conformations. The most persistent maximum - i.e. the conformer found to have the largest gap to its closest minimum, is illustrated in Figure 4.28.

4.7 Conclusions and Future Directions

This chapter shows the efficacy of persistent homology methods in the analysis of conformational spaces and energy landscapes. Applied to alanine dipeptide, persistent homology was indeed able to verify the existence of the underlying toroidal conformational space. When compared to pentane, it was shown that different representations of



(a) Position 1 (b) Position 2

FIGURE 4.28: Two views of the most persistent maximum found from the persistent homology of the energy landscape of cyclooctane. This is a boat-boat conformation of the molecule, as suggest by Martin*et al.*

the conformational space could indeed lead to different results, and therefore the RMSD matrix should be used to ensure that the correct conformational space is found. Furthermore, the analysis of pentane showed that the persistent homology method was able to take molecular symmetry into account, in particular those of the molecular graph. For the more complicated cyclooctane molecule, persistent homology, alongside local dimension estimation, was able to verify the existence of spherical and Klein bottle components to the underlying conformational space.

Persistent homology of real-valued functions on simplicial complexes allowed the analysis of energy landscapes. With alanine dipeptide, a free energy surface on the two torsional degrees of freedom was calculated. The free energy function was extended from points in torsion space onto a simplicial complex with a toroidal topology, before having critical points found using persistence. For cyclooctane, a Rips complex with spherical topology was found, and a single-point energy function calculated. Again, critical points were found, and were shown to match those found with other methods.

The persistent homology methods used in this work could be extended to other molecules. In particular, the conformational spaces and energy landscapes of more complicated molecules could be found, such as fused rings, or other interesting geometries. It is also of interest to gain understanding of the conformational spaces of combinations of molecules, such as butylcyclooctane, and how they can be related to the conformational spaces of the separate molecules.

Chapter 5

Persistent Homology of Water Networks

5.1 Introduction

This work began as an investigation into water solubility. In particular, it was believed that studying the perturbation of water networks due to the influence of a solute would enable better understanding of the solubility process, and lead to improved informatics models. However, this work only uses persistence to understand the behaviour of bulk water systems, as this became a difficult undertaking. This chapter therefore focuses on understanding the intermolecular structure of pure water using persistent homology.

Firstly, the problem at hand is described, including a brief literature review. Then, the use of molecular simulation is explained, including description of the various water models used. The general use of persistence to understand materials is then introduced, explaining what type of persistent homology is calculated on what object. Some introductory results, focusing on single snapshots of simulation are presented, and persistence is shown to be a converged descriptor. Following that new methods, regarding different normalisations of persistence images are tested. It is shown that l_1 normalisation leads to the most size-agnostic descriptor, which can therefore be used to understand the equilibrium properties of bulk water. This method is then tested on a range of systems, investigating the effect of temperature and choice of atomistic model. As a more difficult task, comparisons are drawn between atomistic water models, and the more general coarse-grained Stillinger-Weber model for water. Finally, the use of persistence landscapes as an analysis tool for water networks is briefly presented, before conclusions are drawn about the use of persistence in general for this task.

5.1.1 The Water Network Problem

As mentioned, the study of water networks, and in particular their perturbation with the presence of a solute, is fundamental to the solubility prediction task. Furthermore, water is famous for its anomalous properties, such as the increased density on melting, and high surface tension - attributed to water's ability to form 4 hydrogen bonds [193]. Simulation has regularly been used to understand this behaviour through the structural properties of water. Mark and Nilsson studied the radial distribution function (RDF) [194] of different 3-site atomistic water models. It was shown that the slight differences in these models, namely in Lennard-Jones and Coulombic terms, led to pronounced differences in the heights of peaks in the O-O radial distribution function g_{OO} (Figure 5.1). In

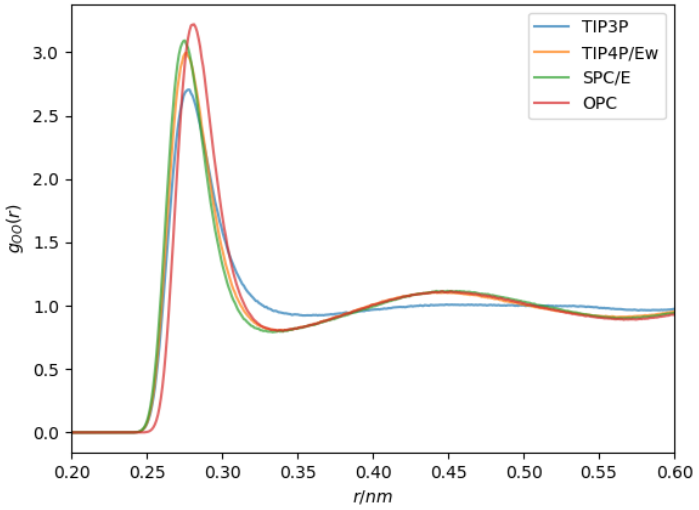


FIGURE 5.1: O-O radial distribution functions for studied water models at 300K.

their 2002 review, comparing water structures obtained from scattering experiments and simulation, Head-Gordon and Hura discuss the 'tetrahedrality' of water networks [195]. Using the following relationship for coordination number [196]:

$$N_c = 4\pi\rho \int_0^{r_{min}} r^2 g_{OO}(r) \mathrm{d}r$$

(where r_{min} is the value of r at the minimum of g_{OO}), it was shown that for water $N_c < 5$. This, combined with the characteristic second peak found in g_{OO}, implied that water had a tetrahedral structure, matching the crystal structure of ice (I). The main difference being that the tetrahedral lattice would have slightly distorted hydrogen bonds [197, 198, 199]. Furthermore, approximately 15% of the hydrogen bonds present in ice would be expected to break during the melting process, which would lead to interstitial sites [200].

As well as the RDF approach to water, methods such as the spatial distribution function (SDF) are used. In this case, the spatial degrees of freedom are not integrated out, giving a 3-dimensional picture of the coordination of molecules [201]. This allowed the observations of what was termed as two different motifs in SPC/E water. Firstly, a purely tetrahedral water, found at all temperatures. Secondly, a temperature dependent non-tetrahedral water. It is therefore unsurprising that there have been discussions about whether water networks are truly tetrahedral, or if this is a misinterpretation of the data. In fact, based on x-ray absorption results from 2004 [202], it was suggested that the tetrahedral structure be replaced by a chain of water molecules, with each having only 2 hydrogen bonds on average. This 'tetrahedral vs chains' discussion is partially thought to be a matter of instanteous vs average properties, with instantaneous measurements ([202]) leading to a chain picture, and average measurements (such as quantities derived from simulation) corresponding to tetrahedral conclusions [203]. Furthermore, the Stillinger-Weber potential contains a parameter tuning the strength of these tetrahedral parameters (discussed in more detail later). Recent work has found that the tetrahedral parameter does indeed have a strong influence on the phase diagram of the system, and a particularly strong tetrahedral parameter leads to the removal of water's well known density anomaly [204].

Alternatively to both radial and spatial descriptions of water networks, graph-theoretical approaches have been used. In particular, Clark *et al* have constructed graphs to represent correlation between water molecules [205, 206]. The main strength of such approaches is the ease of comparison between different environments, for example proximity to a solute molecule. Furthermore, the abundance of graph-theoretical techniques such as Google's PageRank algorithm [207] have led to a wide range of derived descriptors for both local and global structural properties. However, their approach necessitates the use of a heuristic as to where two water molecules are defined to be correlated, such as distance and orientation requirements. This does lead to the idea of creating a filtration of molecular networks using persistent homology, as studied in this work.

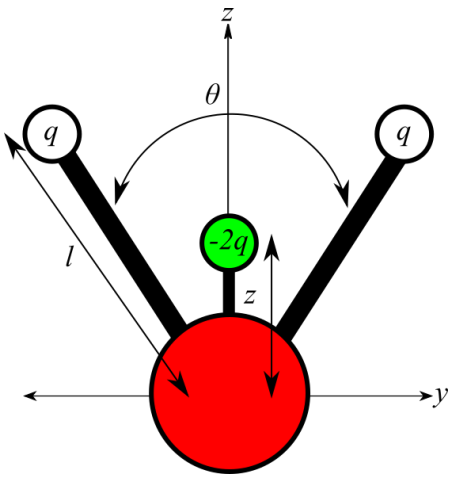
5.2 Simulation Details

5.2.1 Atomistic Water Models

There are a wide range of water models that have been designed to be used in molecular dynamics simulations, and no such model is able to accurately recreate all of water's famous anomalies. A wide range of these models are investigated in this work, as seen in Table 5.1. These models were chosen as they are commonly used within simulation, and previous studies have been performed comparing their structure and dynamics [194]. The Transferable Intermolecular Potential with 3 Points (TIP3P) [208] and Extended Single Point Charge (SPC/E) [209] models are both 3-site, designed to match the 3 atoms

of water. 4-site models, such as TIP4P [208], were created to better replicate physical properties of water, such as its dipole moment. This was achieved by the addition of a massless point charge in the molecular plane - thus altering its dipole. In this work, the TIP4P/Ew model is used, where the original TIP4P model has been slightly adjusted to improve the accuracy of its bulk properties [210]. The Optimal Point Charge (OPC) model [211] is the most recent of the tested models, and was constructed to best match the electrostatics of the water molecule (as opposed to its geometry, or derived physical properties. OPC, like TIP4P/Ew, is a 4-site model, and has a larger computational cost, with more interactions needing to be calculated.

The models chosen in this work are by no means an exhaustive list. For example, the TIPnP series of models has been extended to TIP5P [212] and TIP6P [213], designed to better represent the density of water over a range of temperatures, and the behaviour at the water/ice transition respectively. Also, the previously discussed models have fixed geometries, and are non-polarisable. This makes them unsuitable for use in spectroscopic property prediction, or heterogeneous environments.



Model	q/e	l/Å	z/Å	$\theta/^\circ$	σ_{LJ}/Å	ϵ_{LJ}/kJmol^{-1}
TIP3P	0.4170	0.9572	N/A	104.52	3.15061	0.636
TIP4P/Ew	0.5242	0.9572	0.1250	104.52	3.16435	0.681
SPC/E	0.4238	1.0000	N/A	109.47	3.16600	0.890
OPC	0.6791	0.8724	0.1594	103.60	3.16655	0.89036

TABLE 5.1: The parameters of the various water models used in this study, and their physical meaning. σ_{LJ} and ϵ_{LJ} are Lennard-Jones parameters for non-bonded interactions.

5.2.1.1 Simulation Methodology

A series of simulations were performed on a range of systems of different water models, at different temperatures. The AMBER molecular mechanics program [214], along with

a PME electrostatic approximation were employed for the calculation, and the SHAKE [215] algorithm used to fix the water molecules in a rigid geometry.

The systems were first initialised using the LEAP program in the AMBER suite [216]. A water box of dimension $25 \times 25 \times 25$ Å^3 was created using the `solvatebox` utility. Each simulation cell was then prepared by heating a periodic cubic system to its appropriate temperature, over a period of 0.1ns. Pressure was then allowed to equilibrate for 0.5ns to achieve atmospheric conditions. After preparation, the simulation underwent its production run in the NVT ensemble. In general, this phase would last for 4ns, and snapshots taken every 2ps for a total of 2000 snapshots, and this work specifies when this is not the case.

The `solvatebox` utility results in slightly different numbers of water molecules depending on model choices. The number of molecules for each model can be seen in Table 5.2. The slight differences in water numbers, and the effect in the resulting persistent homology, is investigated and characterised in Chapter 5.6. Simulations were created by designing master input files for AMBER and its utilities, and using bash shell scripting to generate specific instances. These master files can be seen in Appendix B.

Model	N_{water}
TIP3P	4287
TIP4P/Ew	4254
SPC/E	4287
OPC	4302

TABLE 5.2: The number of water molecules for each atomistic model studied in this work.

5.2.2 The Stillinger-Weber Potential

The Stillinger-Weber (SW) potential, in contrast to those previously discussed, is a coarse-grained potential, Originally parameterised for Silicon in 1983 [217], the SW potential has been shown to be incredibly versatile. Its most general functional form is:

$$U = \sum_{i,j} U_2(\mathbf{r_{ij}}) + \lambda \sum_{i,j,k} U_3(\mathbf{r_{ij}}, \mathbf{r_{jk}}) \tag{5.1}$$

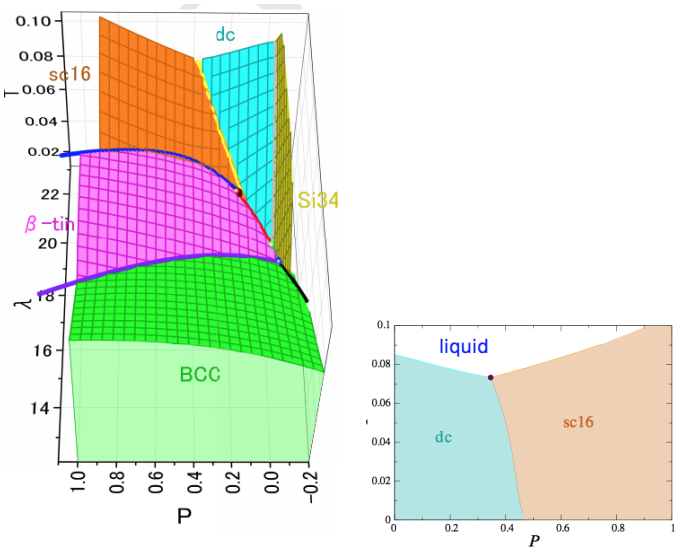
Where the λ parameter allows the tuning of the relative strength of the 3-body interaction. The 2-body interaction U_2 models a steep repulsion at short distances, as well as a potential well:

$$U_2(r) = A\epsilon \left[B \left(\frac{\sigma}{r}\right)^p - \left(\frac{\sigma}{r}\right)^q \right] \exp\left(\frac{\sigma}{r - a\sigma}\right) \tag{5.2}$$

The 3-body interaction can be considered to contain intermolecular angle bending as well as a distance factor term:

$$U_3(r_{ij}, r_{kj}) = \epsilon[\cos\theta_{ijk} - \cos\theta_0]^2 \times \exp\left(\frac{\gamma\sigma}{r_{ij} - a\sigma}\right) \exp\left(\frac{\gamma\sigma}{r_{kj} - a\sigma}\right) \tag{5.3}$$

The hyperparameters used for the studies with the SW potential in this work are: $A = 0.7049556277$, $B = 0.6011145584$, $p = 4$, $q = 0$, $\cos\theta_0 = \frac{1}{3}$, $\gamma = 1.2$, and $a = 1.8$ (all unitless). ϵ and σ determine energy and length scales respectively. λ itself was originally set to 21.0, for silicon, with germanium and carbon described by values of 20.0 and 26.2 respectively. The value of λ for water is 23.15, described as such for matching its density profile over a range of temperatures [204]. The (λ, P, T) phase diagram of the Stillinger-Weber potential can be seen in Figure 5.2. In this work, simulations were performed at various λ at the melting transition (300K and associated pressure) , and were performed by John Russo at the University of Bristol using specifically designed software. Simulations contained 512 water molecules with up to 500 configurations.



(a) λ, P, T phase diagram (b) P, T phase diagram at $\lambda = 22.75$

FIGURE 5.2: Phase diagrams of the Stillinger-Weber potential. Reproduced from the supporting information of [204]. The λ parameter clearly affects the phase of the model.

5.3 Persistence Methodology

The method used for calculating the persistent homology of a pure water simulation is found in Figure 5.3. Persistence is calculated for each frame of simulation individually, and a persistence diagram at time t will be denoted $PD(t)$ The local orientation of water is considered to be determined by hydrogen bond configurations. This implies that the relative location of neighbouring oxygen atoms contains all of the relevant hydrogen bonding information. All of the hydrogen atoms from the simulation are therefore

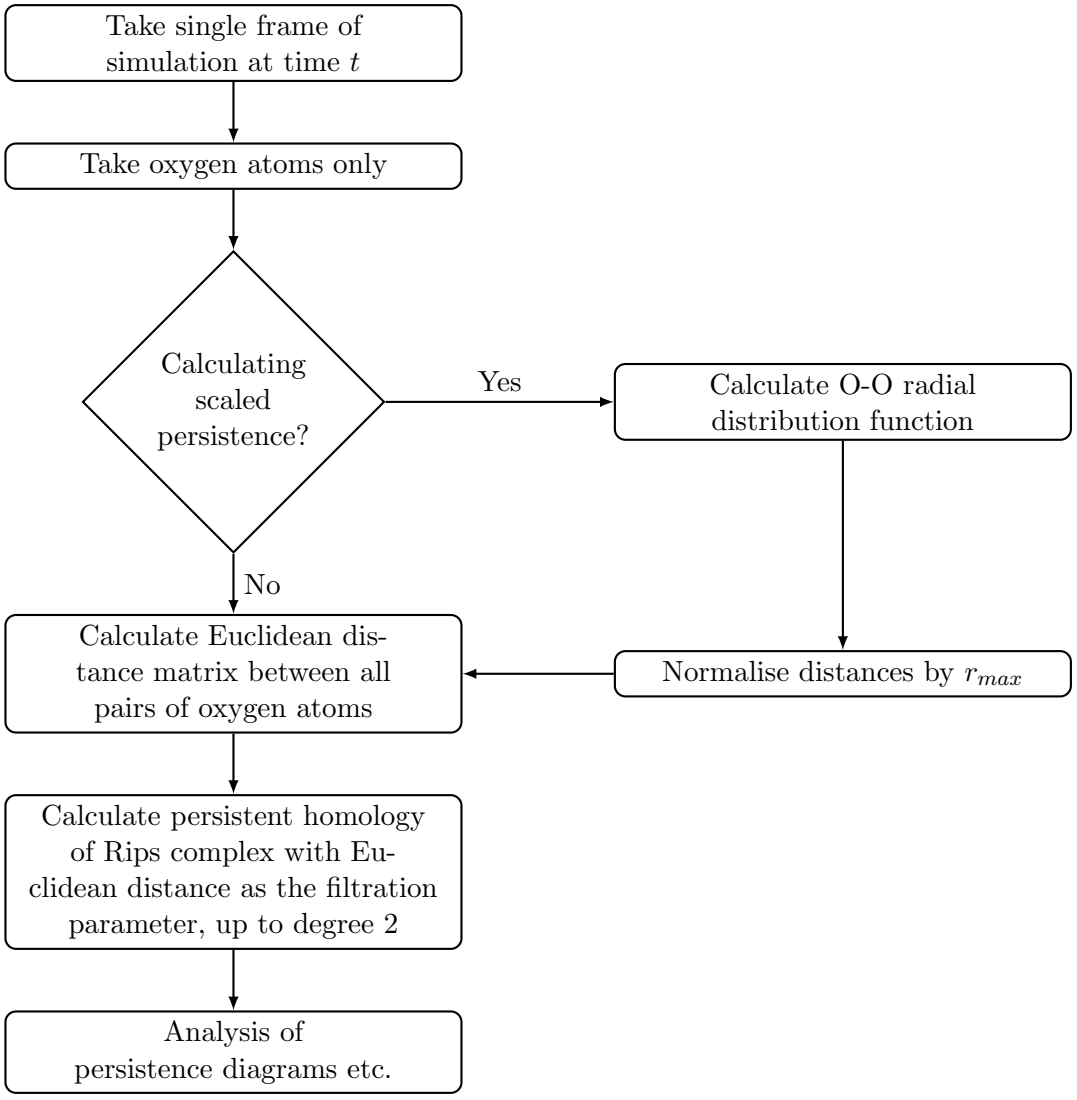


FIGURE 5.3: The general procedure used for calculating the persistent homology of a simulation of pure water. r_{max} denotes the value of r at the maximum of the radial distribution function.

removed, which greatly increases the speed of the persistent homology algorithm. This restricts the number of 1-simplices between any two oxygen atoms to 1, rather than the 9 that would be permitted if persistence was calculated in an all-atom system. The total size (and memory requirement) of the resulting simplicial complex is therefore reduced. Within this chapter, this Rips persistent homology constructed on oxygen atoms will be referred to as *the* persistent homology, as other complexes are not considered here. The main parameter of the procedure determines if the simulation will be scaled. If the system is scaled, the program enters a subroutine to calculate the maximum of the O-O radial distribution function, before normalising all distances by this number. In principle, this allows models for entirely different systems to be compared, by placing their persistent homology on the same scale. For example, this would allow comparison of different materials, such as liquid crystals.

The last set of parameters are for the persistent homology calculation itself. In particular, for the persistent homology of the Rips complex, a maximum length and dimension must be defined. As water exists in Euclidean 3-space, it is reasonable to calculate up to second degree homology. For the maximum length parameter, it is necessary to ensure that any features of note have died. It was found this seemed to occur after approximately 3 nearest neighbour distances. Therefore, for scaled calculations $l_{max} = 3$. The scaled calculations are not discussed further in this work, and are presented purely to show the difference in methodology. For the unscaled calculations, this parameter was set to 8Å. This value was chosen as it is approximately three nearest neighbour distances (as seen in the RDFs in Figure 5.1). In general, the value of 3 nearest neighbour distances was chosen as the Rips complex would have trivial topology, in a uniformly sampled three dimensional grid.

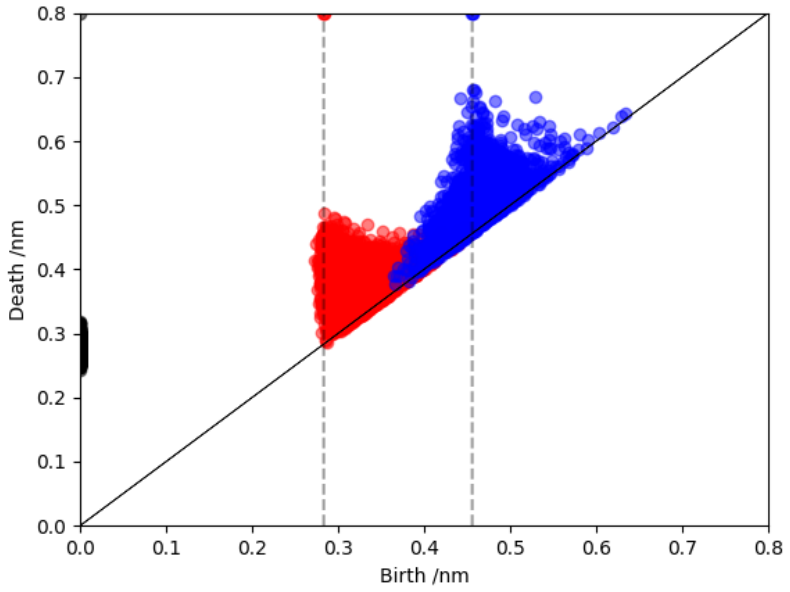
5.4 Persistence Diagrams of Single Frames

The persistence diagram for the Rips complex constructed on the oxygen atoms of a single snapshot of the simulations of pure water boxes at 300K can be seen in Figure 5.4.

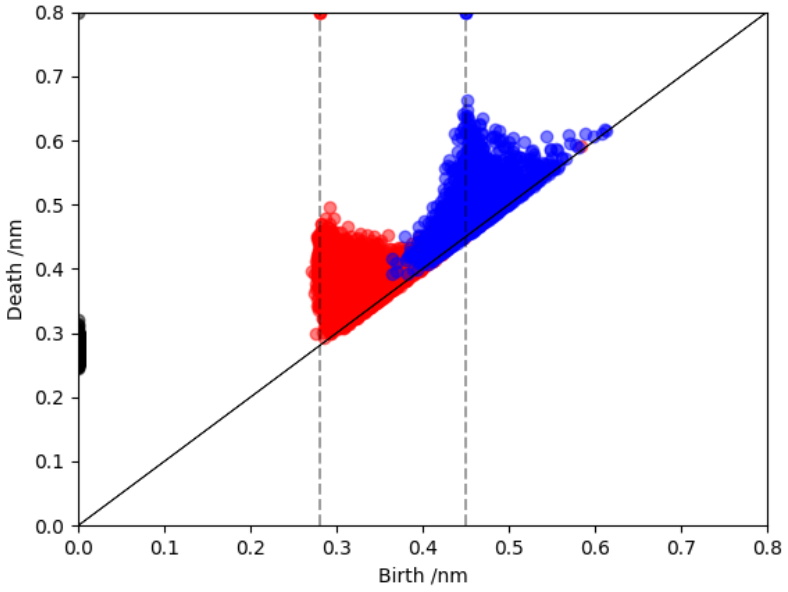
It is difficult to determine features that are able to distinguish between models from these persistence diagrams. However, there are several patterns in common between models. For example, first degree features do not begin to appear until 0.28nm, and second degree features do not begin to appear until 0.35nm. This reflects distances between next- and next-next- nearest neighbours respectively.

Each persistence diagram also contains some long-lived features of each degree. Further analysis of these features yields multiplicities of 1, 3, 3 for degrees 0, 1, 2 respectively. Rather than reflecting the water network connectivity, these features are actually induced by the periodic boundary conditions (PBCs) of the system. The topology induced by PBCs is easiest understood when considering the 1 and 2 dimensional analogues. In 1-D, PBCs can be seen as identifying the two endpoints of a line as equivalent. therefore a map $I \to S^1$, where I is homeomorphic to the unit interval, and S^1 being the topological circle. In 2-D, the PBCs can be considered to be identifying opposite edges of a square as equivalent (and matching orientation). This is therefore a map $I \times I \to S^1 \times S^1 \equiv T^2$.

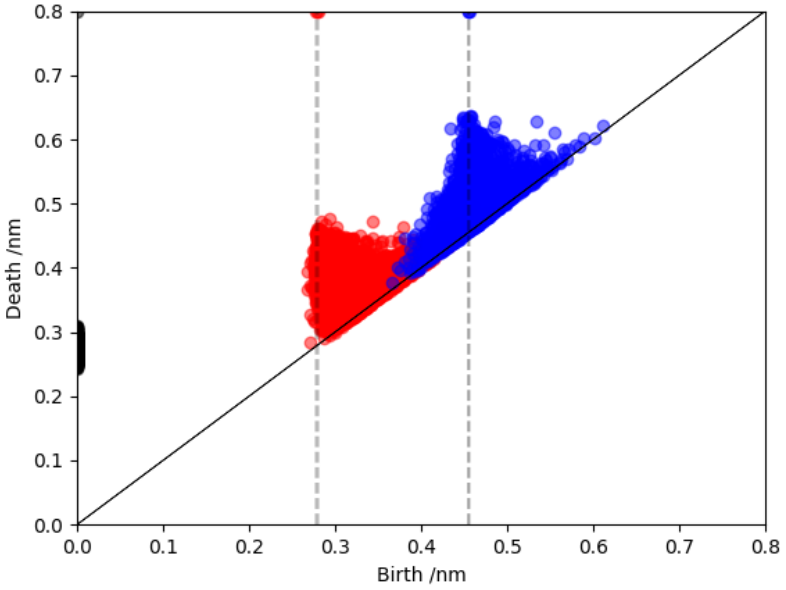
Extending this argument to 3 dimensions, it is recognised that opposite faces of a cube are now identified (again, preserving orientation). This leads to a map $I \times I \times I \to S^1 \times S^1 \times S^1 \equiv T^3$. The homology groups of this space can be calculated (for example with the Künneth theorem), and lead to the Betti numbers $(1, 3, 3, 1)$, matching the multiplicities of the long-lived features in our persistence diagrams (although β_3 is not calculated in this work). The implications of these features are actually quite simple. In particular, the features imply that the water molecules densely sample the space defined



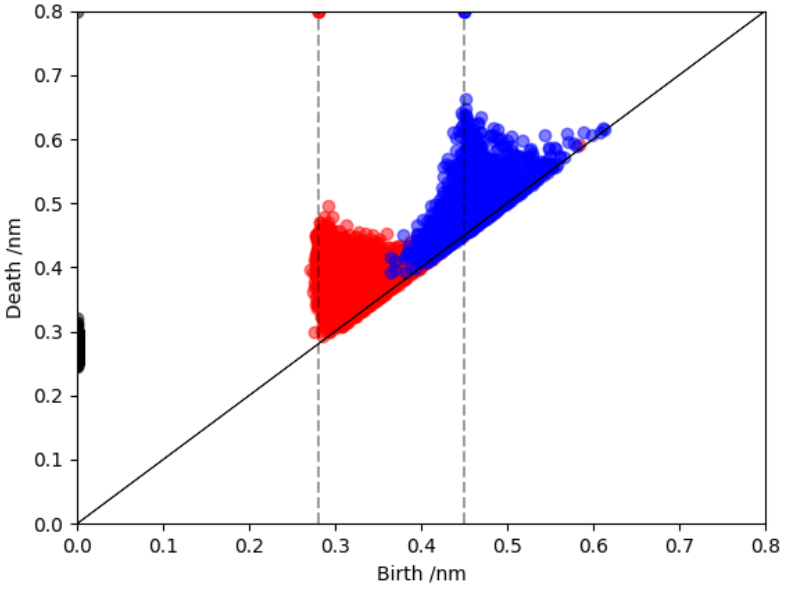
(a) TIP3P



(b) TIP4P



(c) SPC



(d) OPC

FIGURE 5.4: Persistence diagrams for single frames of simulations of pure water boxes at 300K. Black, red and blue points correspond to zeroth, first and second degree homology features respectively. Dashed lines indicate a feature persists to infinity. Black, red and blue correspond to zeroth, first and second degree homology respectively. The persistence diagrams are difficult to distinguish between models for single frames - a more statistical method must be used.

by the box dimensions, and there are no large holes in this system. If the system were to have a vacuum bubble present, due to poor equilibration, a greater number of long lived second degree features would be seen.

One potential use of these features is in the study of surface systems using molecular dynamics. Persistent homology of a sequence of Rips complexes constructed on surface atoms with periodic boundary conditions should have the topology of two disconnected 2-tori. Any long-lived features should therefore have multiplicity $(2, 4, 2)$. Instead, if the surfaces are not far enough apart (either from each other, or from the cell boundaries), different persistent Betti numbers would be found. Depending on the object of interest being studied, kernel based Rips complexes, such as those described in [86] could find value here in particular.

5.5 Persistence as a Descriptor

Before persistence can be used to understand the average properties of water networks, it should first be shown that persistence is a converged descriptor. In particular, it ought to be shown that the persistent homology of a system does not show large fluctuations. Within molecular simulation, the convergence of properties is often described using time autocorrelation functions:

$$C_f(t) = \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau f(t_0)f(t_0+t)\mathrm{d}t_0 = \langle f(t_0)f(t_0+t)\rangle \tag{5.4}$$

Where $f(t)$ is the value of property f at time t, and C_f is its autocorrelation. The autocorrelation at t describes how similar a time signal is with a t delayed version of itself. The no-delay limit $C_f(0)$ can be seen to be equal to $\langle f^2\rangle$. For a converged property, a time signal should become uncorrelated to itself, i.e. $\lim_{t\to\infty} C(t) = \langle f\rangle^2$. Using a standard identity:

$$\langle X^2\rangle - \langle X\rangle^2 = \sigma_X^2$$

(where σ is the standard deviation) it is clear that the autocorrelation function decreases for a converged property, and eventually reaches a limiting value.

It is less obvious as to how to describe the convergence of persistence. The persistent homology could be treated as a time signal, and an appropriate product analogous to $f(t_0)f(t_0+t)$ defined. For example, persistence landscapes could be created, with a possible integral such as the following used for autocorrelation surrogate:

$$C_\Lambda(t) = \int_0^\infty \Lambda(t_0)\Lambda(t_0+t))\mathrm{d}t_0$$

where

$$\Lambda(t_0)\Lambda(t_0+t) = \sum_{k=1}^{\infty} \lambda_k(t_0)\lambda_k(t_0+t)$$

Provided persistent features are removed, this quantity is bounded. Further, a single landscape contains the same information as a single persistence diagram, and as explained in Chapter 2, the inverse transformation is well-defined. However, the product as defined above is an unusual mathematical object. In general, the reason this work has avoided this potential route is that it is felt better to work with persistence diagrams as the fundamental object whenever possible - any statement we can make with diagrams ought to apply to objects derived from them. The opposite is potentially untrue. However, the difficulty of defining an autocorrelation-like property for persistence diagrams lies in determining the appropriate quantity of which to find the average time delay.

Usefully, persistent homology is endowed with various metrics between persistence diagrams, which quantify dissimilarity (Chapter 2.1.9). For a set of persistence diagrams, the following quantity is defined:

$$C_{PD}(t) = \left\langle d_B\left(PD(t_0), PD(t_0+t)\right) \right\rangle \tag{5.5}$$

Where $d_B(PD_1, PD_2)$ is the bottleneck distance (Equation 2.7). Clearly, $C_{PD}(0) = 0$, as the bottleneck distance is a metric. The long-time behaviour of $C_{PD}(t)$ can be understood by considering periodic and non-periodic systems. For a periodic system, one would expect to see oscillations in $C_{PD}(t)$, with a frequency matching the system's recurrent behaviour. In contrast, a non-periodic system would lead to $C_{PD}(t)$ reaching a fixed value, which would remain unchanged, similarly to traditional autocorrelation functions. This definition also does not require the same treatment of persistent features as the method defined for landscapes above - it would only be infinite if there are different numbers of persistent features, which could be considered useful information. One final strength of the use of the bottleneck distance for measuring auto-correlation of persistence is that the it has an associated stability theorem, implying that small changes in point clouds (as expected in short timescales of simulation) do not lead to large changes in persistence.

The bottleneck distance approach to understanding the convergence of persistence was applied to simulated TIP3P water at 300K, and can be seen in Figures 5.5, 5.6, and 5.7 for simulations lasting for 4ps, 40ps and 4ns respectively. For all degrees of homology, it is clear that the bottleneck distance between frames has equilibrated after 2fs. In fact, for zeroth degree homology, the bottleneck correlation has reached a plateau after < 0.1ps. In contrast, second degree homology does not plateau until after approximately 1ps. These times are shorter than the orientational correlation time of approximately 5ps found in [218], suggesting that persistence does not undergo large fluctuations over long timescales.

Furthermore, the bottleneck distance correlation does not display any extrema before converging. From this, it can be inferred that persistent homology of these systems does not contain any recurrent behaviour. This reflects the chaotic nature of the bulk water dynamics.

From these analyses, it is seen that persistence can be used as a descriptor for understanding the behaviour of these water systems. It has been shown that it converges after under 5ps for all degrees of homology of interest. Persistence has also been shown to not contain any spurious recurrent behaviour. It is now possible to move on to discuss average properties of persistence and begin to analyse water networks in a more thorough fashion.

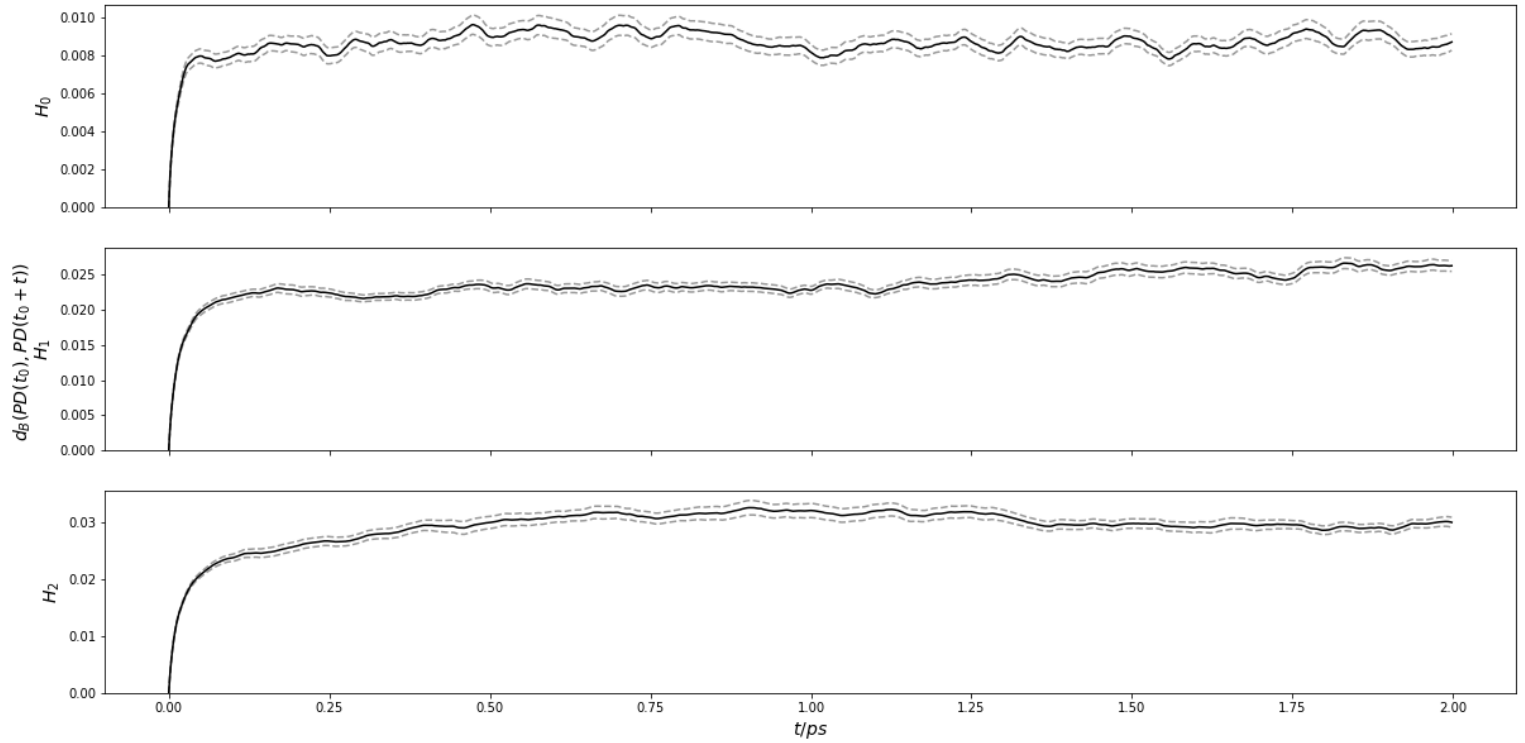


FIGURE 5.5: The bottleneck distance correlation function for TIP3P water. Solid lines are the average value, and dashed lines represent 3 standard errors. The simulation was run for 4ps, with snapshots taken every 2fs. Convergence is clearly achieved quickly, therefore persistence is a reasonable descriptor.

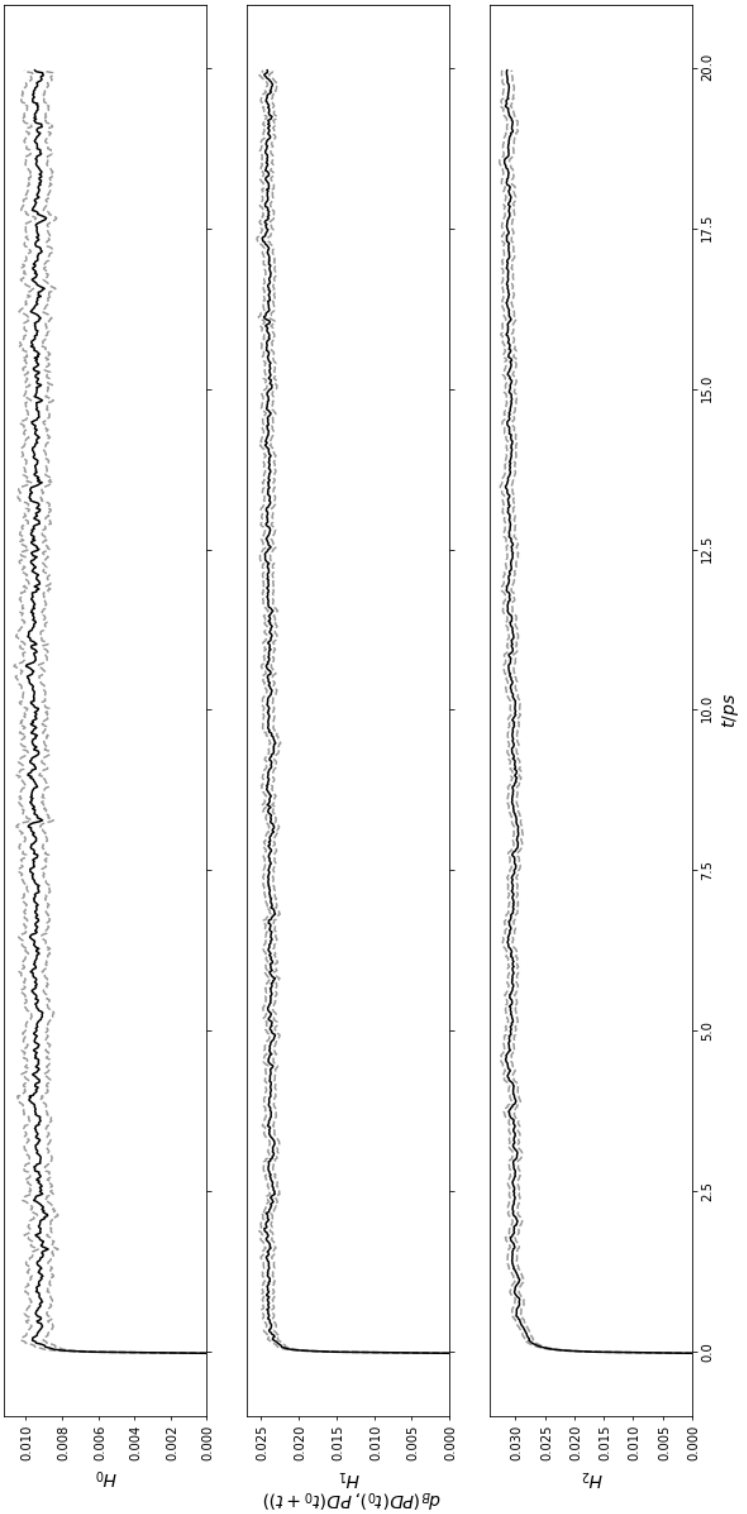


FIGURE 5.6: The bottleneck distance correlation function for TIP3P water. Solid lines are the average value, and dashed lines represent 3 standard errors. The simulation was run for 40ps, with snapshots taken every 20fs. Convergence is clearly achieved quickly, therefore persistence is a reasonable descriptor.

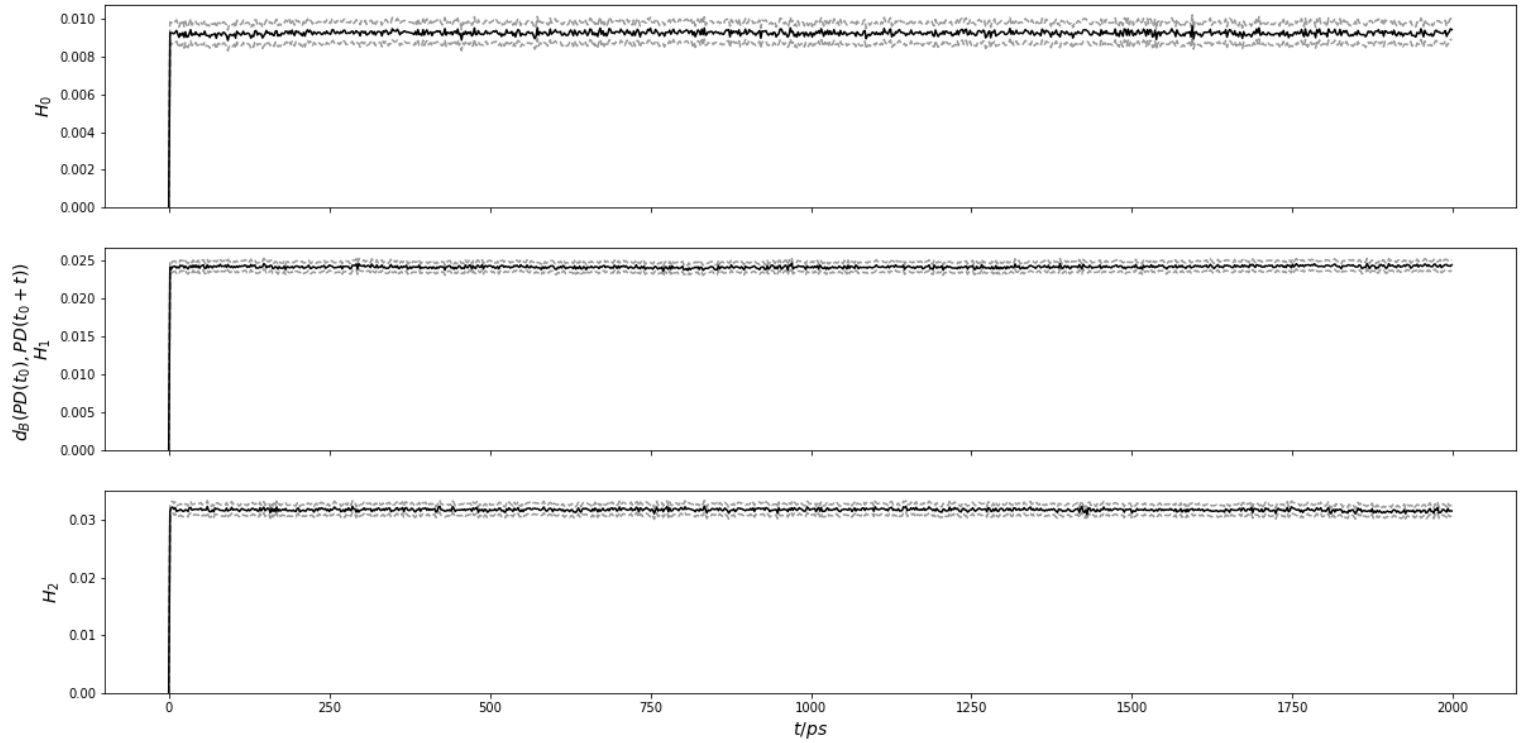


FIGURE 5.7: The bottleneck distance correlation function for TIP3P water. Solid lines are the average value, and dashed lines represent 3 standard errors. The simulation was run for 4ns, with snapshots taken every 2ps. Convergence is clearly achieved quickly, therefore persistence is a reasonable descriptor.

5.6 Is Persistence Intensive or Extensive?

Properties of materials are often labelled as either intensive or extensive, depending on how the property changes as a function of system size. Intensive properties, such as density or temperature, are independent of the size of the system. In contrast, extensive properties, such as mass or volume are additive for subsystems.

Persistent homology is clearly more complicated than this. Taking the total number of (non-trivial) features of a given degree ($n_{features}$) as a basic property, it is obvious that it is not an intensive variable for any degree. This can be seen when considering the limit as the number of oxygen atoms (N_O) tends to zero - the number of features for any degree must also be zero. However, if $N_O > (d+1)$, some features of degree d might form - but the exact value of $n_{features}$ depends on the exact distance relationships between oxygen atoms.

For zeroth degree persistent homology, the behaviour of the number of features is easily understood. By definition, at $\delta = 0$ there are N_O features. No other features can be born at any time. There are therefore N_O total features in the zeroth degree persistent homology - it is an extensive property. Generally, analysis of zeroth degree homology is avoided in this work, as it reduces to a hierarchical clustering on the oxygen atoms. The systems being studied here are of uniform density, and therefore zeroth degree homology is largely restricted to containing information about the distribution of nearest neighbour distances.

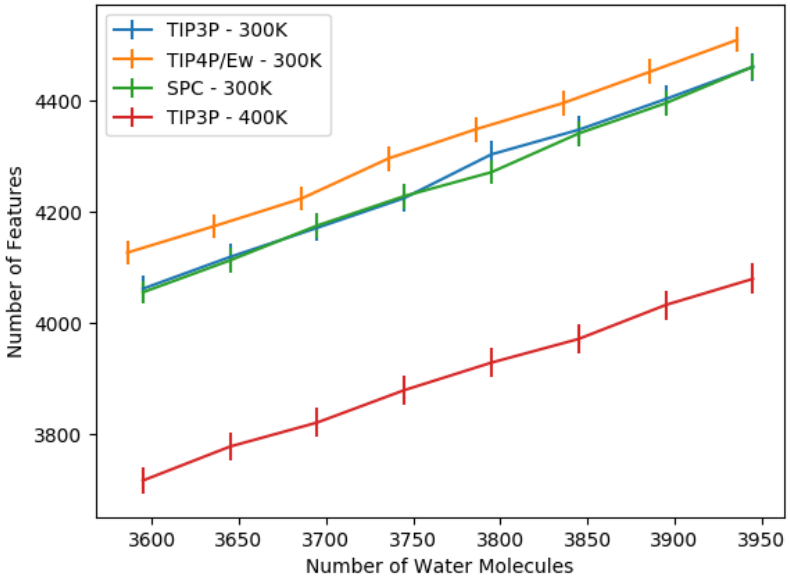
For first and second degree homology, the behaviour of $n_{features}$ as a function of N_O is studied in more detail. From Table 5.2 it can be seen that there are different numbers of water used for different models. For completeness, a simulation was performed at a higher temperature. For first and second degree homology, graphs of $n_{features}$ vs N_O can be seen in Figure 5.8.

For both degrees of homology, it is clear that the temperature difference leads to a more significant change in $n_{features}$ than changing N_O. This is encouraging for two reasons. Firstly, it was earlier assumed that the difference in N_O between models does not vastly alter persistence when studying the single frames. Secondly, all of the models being discussed purport to model the behaviour of water. If the differences between them were more significant than altering the temperature, it would suggest that these models are not remotely alike.

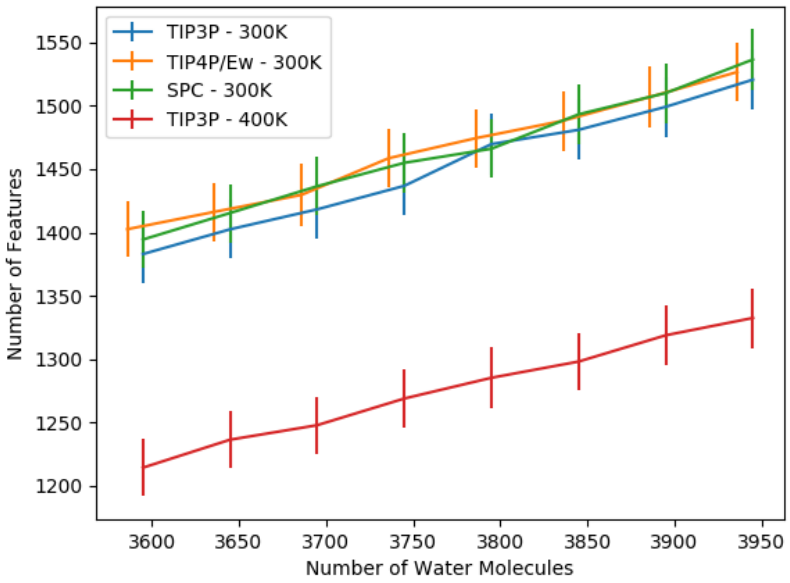
Over the range of N_O studied, there appears to be a linear relationship between N_O and $n_{features}$. Linear models were therefore constructed, of the form:

$$n_{features} = \alpha N_O + \beta$$

with the resulting parameters, associated errors and p-values found in Table 5.3. It is



(a) First degree homology



(b) Second degree homology

FIGURE 5.8: The number of first degree features as a function of the number of water molecules (oxygen atoms) for a range of models and temperatures, and their associated error bars. Lines are used to indicate crossings, and are not directly measured. Persistent homology is clearly not size-independent.

Model	T/K	Degree	α	σ_α	p_α	β	σ_β	p_β
TIP3P	300	1	1.15	0.02	0.00	-66	77	0.42
TIP4P/Ew	300	1	1.10	0.02	0.00	154	69	0.07
SPC/E	300	1	1.14	0.02	0.00	-50	73	0.52
TIP3P	400	1	1.02	0.01	0.00	27	55	0.62
TIP3P	300	2	0.40	0.08	0.00	140	120	0.29
TIP4P/Ew	300	2	0.36	0.09	0.00	-265	127	0.08
SPC/E	300	2	0.39	0.08	0.00	63	109	0.58
TIP3P	400	2	0.33	0.07	0.00	-26	86	0.77

TABLE 5.3: Gradient and intercept of linear models constructed for the relationship $n_{features} = \alpha N_O + \beta$. p-values correspond to testing the null-hypothesis that the variable is equal to 0.

noted that the p-values for testing whether $\beta = 0$ is *not rejected* for any model tested. As previously discussed, this is precisely what is expected, if $N_O = 0$, $n_{features} = 0$. Therefore, a set of models are constructed for the following form:

$$n_{features} = \alpha' N_O$$

with Table 5.4 containing the obtained parameters. From these results, we can see

Model	T/K	Degree	α'	$\sigma_{\alpha'}$	$p_{\alpha'}$
TIP3P	300	1	1.13	0.00	0.00
TIP4P/Ew	300	1	1.14	0.00	0.00
SPC/E	300	1	1.13	0.00	0.00
TIP3P	400	1	1.03	0.00	0.00
TIP3P	300	2	0.38	0.00	0.00
TIP4P/Ew	300	2	0.39	0.00	0.00
SPC/E	300	2	0.39	0.00	0.00
TIP3P	400	2	0.34	0.00	0.00

TABLE 5.4: Gradient and intercept of linear models constructed for the relationship $n_{features} = \alpha' N_O$. p-values correspond to testing the null-hypothesis that the variable is equal to 0. Values are reported to two decimal places, but are in general not equal to 0.

that first and second degree homology are neither extensive nor intensive variables. In particular, they are not additive - doubling N_O does not double $n_{features}$. However, a useful descriptor should be easily transferable between systems of different sizes. A different approach is therefore needed.

5.7 Development of a size-agnostic descriptor

As persistence diagrams are unwieldy mathematical objects, it is only natural to consider one of the other persistence representations with which to build a descriptor. Persistence

landscapes might appear fruitful, as discussed in Chapter 2. A quantity such as

$$Y[\lambda] = ||\Lambda_1||$$

appears reasonable at first glance. However, this would not be size-independent - the number of non-zero landscape functions is equal to $n_{features}$ for zeroth degree homology, and Y would be dependent on this number. It might therefore make sense to define $Y'[\Lambda] = \frac{||\lambda||_1}{n_{features}}$. However, if $n_{features}$ includes only non-trivial features, this functional is non-linear. How could $Y'[\Lambda_a + \Lambda_b] = Y'[\Lambda_a] + Y'[\Lambda_b]$?

This linear functional restriction is not present in the case of persistence images, which are more common mathematical objects. Provided the images are constructed in the same way (for example, the kernel or weight function), the SVM procedure defined in Appendix A even provides a reasonable analysis route. Therefore, this seems a sensible path to follow in descriptor creation.

Focussing on the set of TIP3P model, 300K simulations of different system sizes, a persistence image was created for each frame of simulation. Each set of 2000 frames was then randomly allocated into training and test sets, of 1500 and 500 frames respectively. The SVM classifier was then trained on the training set, before its performance on the test set determined using confusion matrices, seen in Figure 5.9.

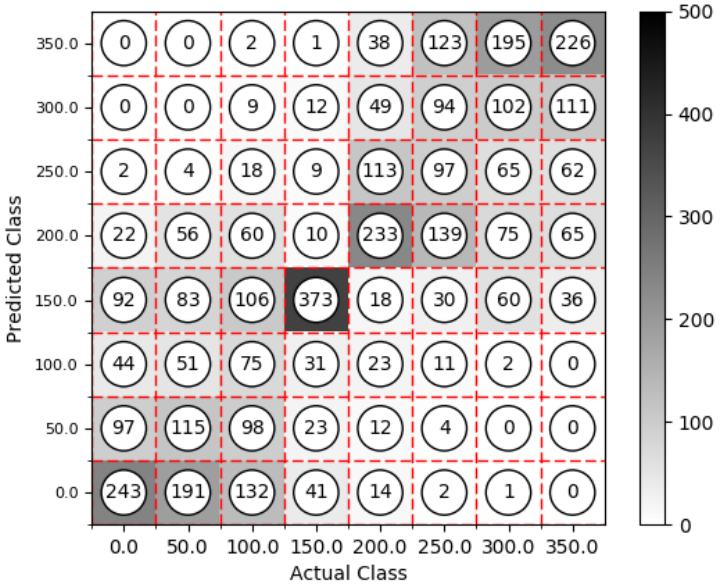
When studying the confusion matrices, it is seen that the classifier can in general distinguish between systems with a large difference in N_O, and performs worse at intermediate values. This is moderately promising - the descriptor being designed is supposed to be size-agnostic, and therefore not be able to distinguish between systems of different sizes. In this case, systems of similar sizes are indistinguishable, but the descriptor is able to distinguish between systems with a reasonable ($\sim 10\%$) difference in N_O.

The most obvious next step to take is to consider different ways of normalising persistence images. The two methods chosen were as follows:

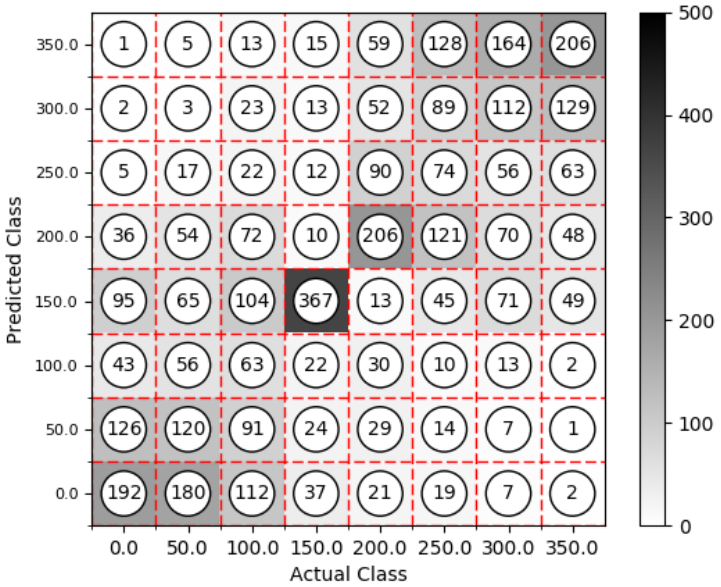
- l_∞ normalisation: The magnitude of the maximum pixel is equal to 1
- l_1 normalisation: The overall integral of a persistence image is equal to 1

The two methods will be referred to as l_∞ and l_1 normalisation respectively. When necessary, an image which has had no normalisation applied will be referred to as *nonorm*.

The classification accuracy for the three different methods, applied to the TIP3P water simulations at 300K are found in Table 5.5. Results are similar for different models and temperatures. It is clear that the nonorm method leads to the strongest classifier, which as discussed previously is able to classify between systems of appreciable size difference. In contrast, the l_∞ and l_1 methods lead to a test set accuracy approximately as efficient



(a) First degree homology



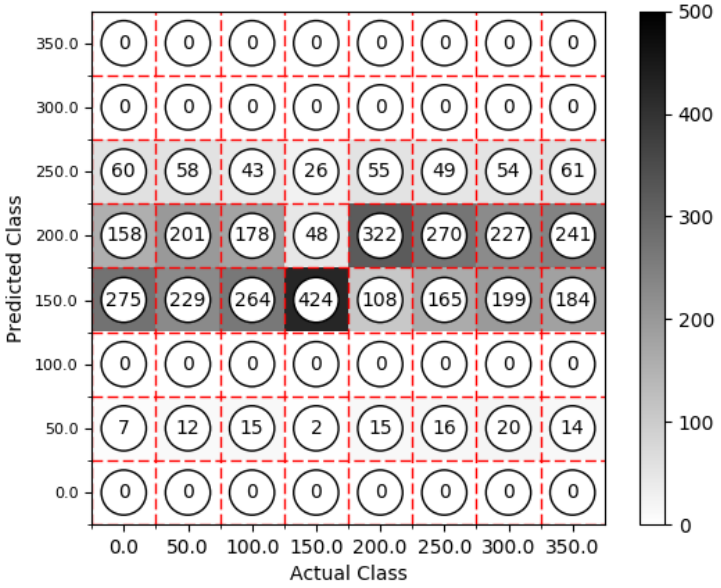
(b) Second degree homology

FIGURE 5.9: Confusion matrices for the support vector machine classifiers of persistence images for TIP3P systems at 300K, with classes defined by the number of water molecules *removed* from the system. The classifiers are able to distinguish systems with different numbers of water molecules.

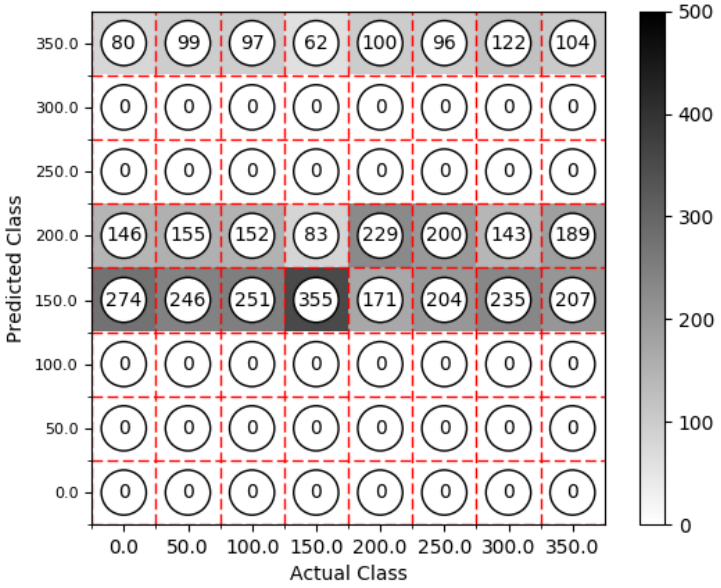
Method	Degree	Training Accuracy	Test Accuracy
nonorm	1	0.530	0.366
	2	0.515	0.335
l_∞	1	0.304	0.234
	2	0.403	0.231
l_1	1	0.204	0.202
	2	0.175	0.172

TABLE 5.5: Classification accuracy for the different normalisation procedures for the set of simulations with different particle numbers. TIP3P model, 300K.

as randomly guessing. For the l_∞ method, the SVM is able to distinguish between systems in the training set, which it is unable to in the test set. This indicates overfitting to the training set, suggesting there are differences that can be detected using the l_∞ procedure. In contrast, the l_1 method performs equally badly on the training and test sets, and as badly as random guessing. This behaviour can again be investigated with confusion matrices (Figure 5.10) . For first and second degree homology, the classifier nearly always predicts one of the two intermediate classes. Interestingly, they seem to predict the system with 150 water molecules removed with a high accuracy, and also seem to be biased towards class 150 when $n_{rem} < 150$, and class 200 otherwise. PCA can be used to help understand this (Figure 5.11). The mean position of l_1-normalised persistence images for the 150 class is clearly an outlier, in both first and second degree homology. This biases the classifier, as it is the easiest to distinguish. The next largest outlier is the class of 200 water molecules removed. The other classes are in general too close to be easily distinguished by the classifier. This leads to the behaviour seen with the confusion matrix. These results are actually a strength of the l_1 method of normalisation. If the method was being strongly influenced by the size of the system, one would expect the 150 and 200 classes to be near each other. This is not the case for the l_1 classifier. It has therefore been shown that, of the methods testing, the l_1 normalised persistence images leads to the most size-agnostic descriptor. This will be the descriptor used going forward, and will be referred to as a L1NPI, for brevity.

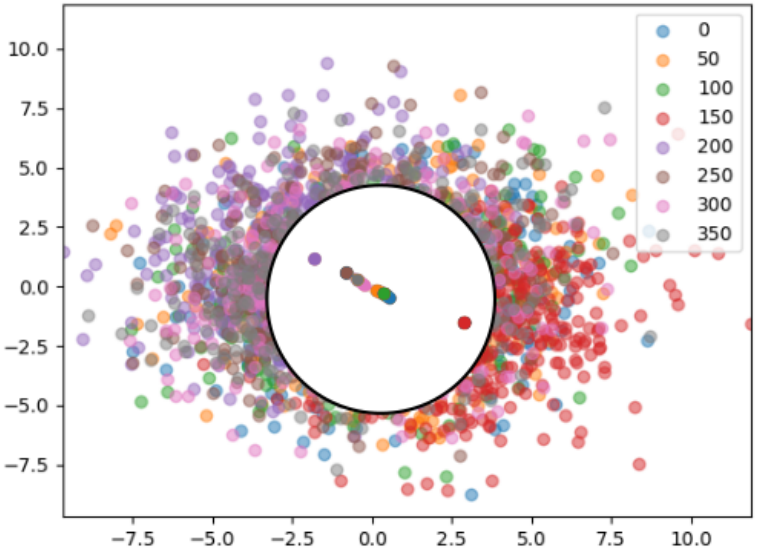


(a) First degree homology

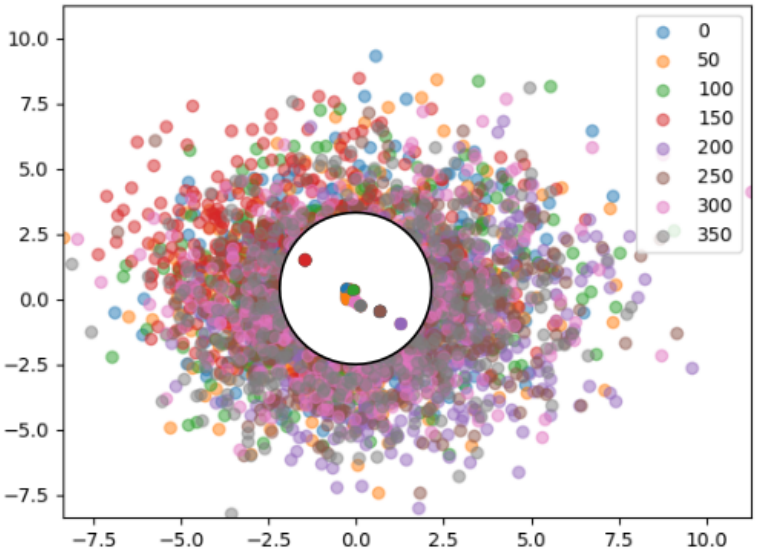


(b) Second degree homology

FIGURE 5.10: Confusion matrices for the support vector machine classifiers of l_1 normalised persistence images for TIP3P systems at 300K. Classes are defined analogously to Figure 5.9. Now, the classifiers are largely unable to distinguish systems of different sizes, suggesting a size-independent descriptor.



(a) First degree homology



(b) Second degree homology

FIGURE 5.11: 2-dimensional principal component space for l_1 normalised persistence images of TIP3P water at 300K. Points are coloured by the number of water molecules removed from the system. Points within the white circle are used to illustrate the mean position for a given class. It is unsurprising that the classifiers defined above are unable to distinguish between these systems.

5.8 L1NPI Analysis

In this work, L1NPIs have been used to study water networks. Tests have been carried out analysing the effect of temperature, using TIP3P water at a range of temperatures between 300 and 400K. Following that, the effect of atomistic model (TIP3P, TIP4P/Ew, SPC/E, OPC) was studied. Lastly a series of simulations of the SW model was used, investigating the differences between these atomistic and coarse-grained models. As the L1NPI is constructed from persistence, which is in turn constructed from distance matrices between oxygen atoms, various comparisons will be made between the L1NPI and the O-O RDF. Analysis will also be performed by looking at the L1NPI themselves, dimensionality reduction of L1NPI space, and the behaviour of linear SVM classifiers.

5.8.1 Effect of Temperature

The effect of temperature on the RDF is well known, and can be seen in Figure 5.12. The height of peaks are reduced, and the value of r at which those peaks occur increases as a function of temperature. This is due to the reduced entropy and increased density respectively. Investigating L1NPIs, it is important to first understand their general

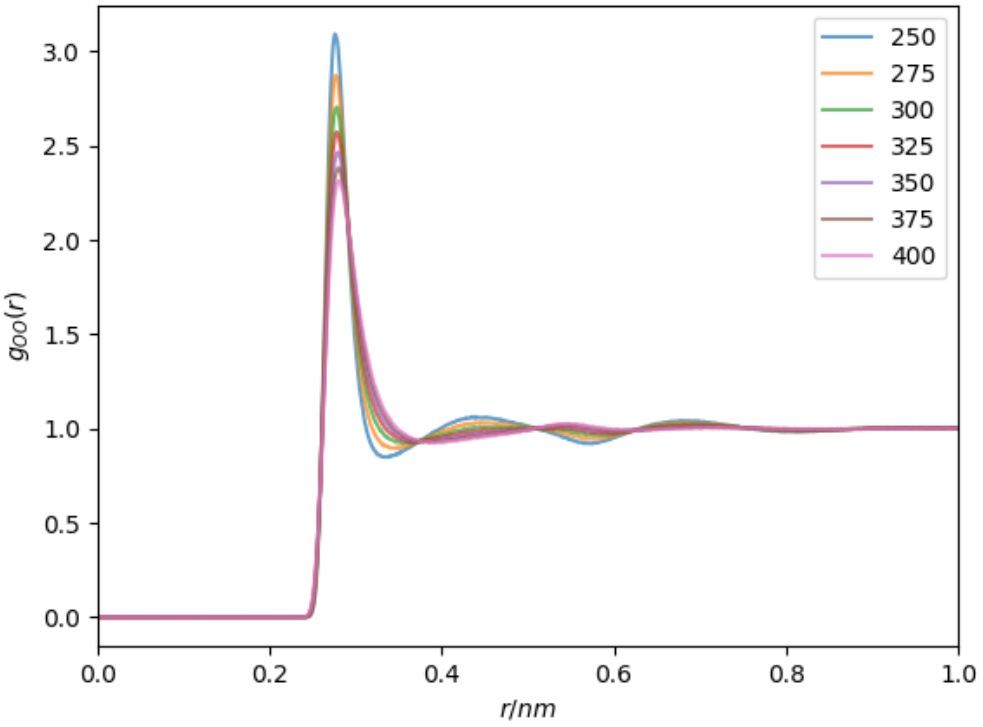
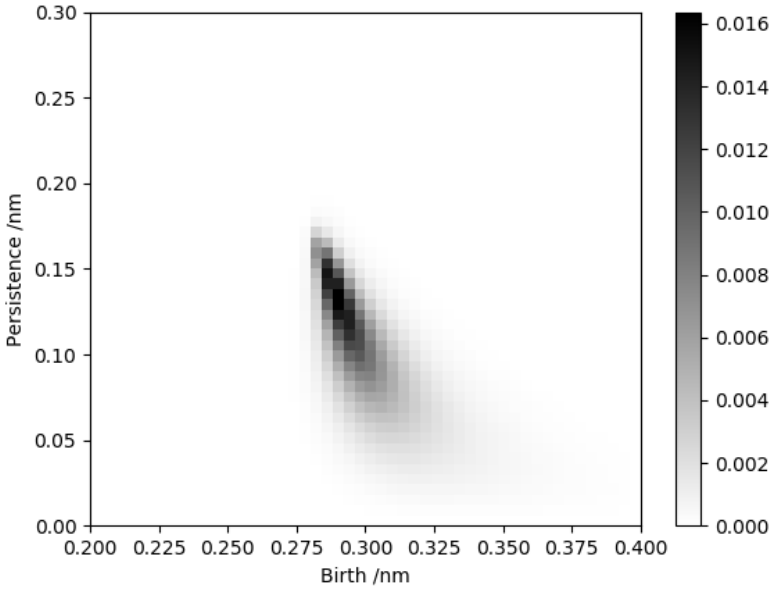
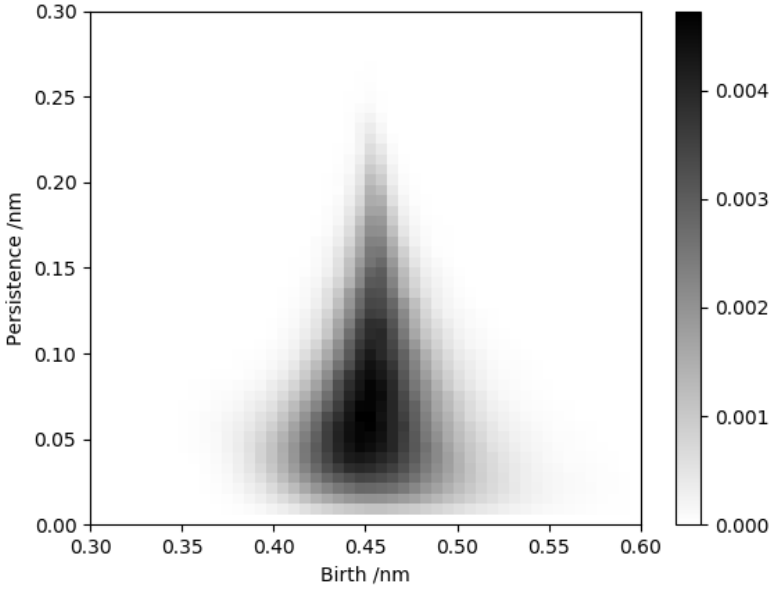


FIGURE 5.12: O-O radial distribution functions for TIP3P at various temperatures

behaviour. For TIP3P water at 300K, mean L1NPIs can be found in Figure 5.13. In first degree homology, features do not begin to appear until $r \approx 0.275$. This reflects the hard-sphere radius, and the sharp rise in the RDF. However, there is more information than this held within the birth time. In particular, the birth time of a first degree feature corresponds to the longest distance between two oxygen atoms within the cycle. These



(a) First degree homology



(b) Second degree homology

FIGURE 5.13: Mean L1NPIs for TIP3P water at 300K. Different regions of the L1NPI can be related to features of the radial distribution function.

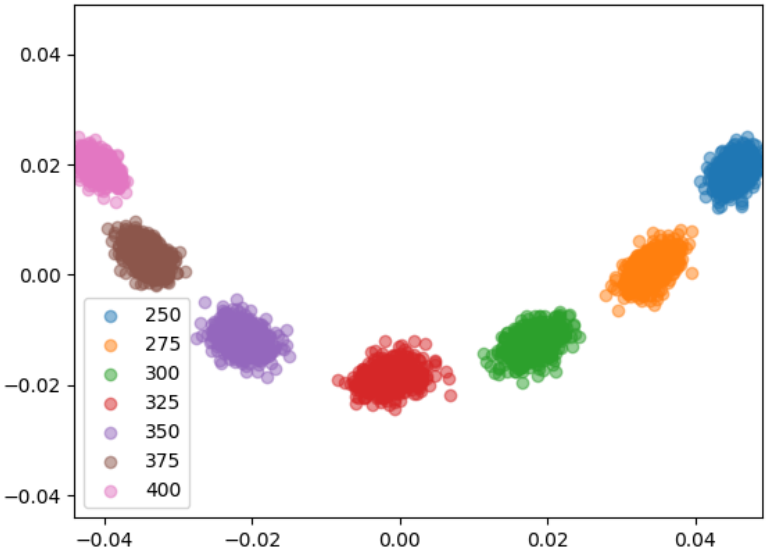
properties can be investigated further by comparing them to critical points in the RDF. The following quantities are defined:

Quantity	Definition
$\arg\max(b(r))$	The argument of the maximum for the distribution of birth times
$\arg\min(b(r))$	The value of r where the distribution of birth times first increases
$\arg\max(g(r))$	The argument of the maximum for the RDF
$\arg\min(g(r))$	The value of r where the RDF first increases

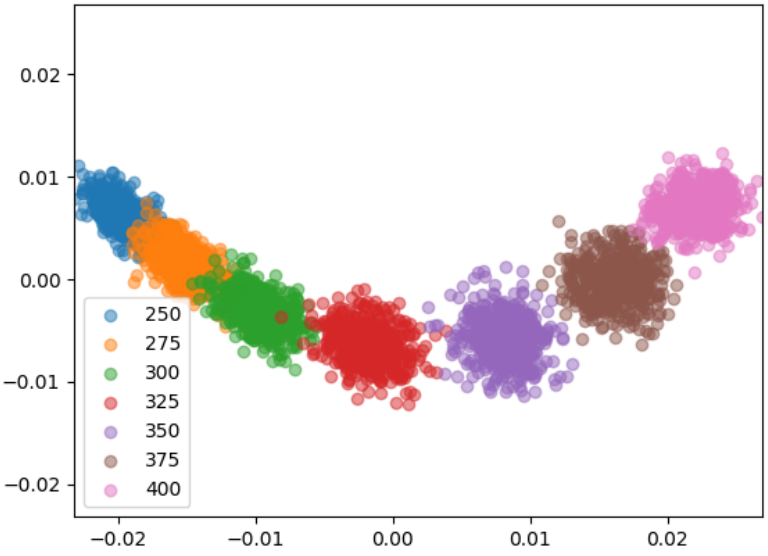
Considering water as tetrahedral, the *natural cycle* that forms are those of adjacent tetrahedral sites, with cycles of 6 oxygen atoms. $\arg\max(b(r))$ can therefore be identified as the longest length found within a cycle, and $\arg\min(b(r))$ identified as the minimum of the longest cycle lengths. This is clearly different to arg max and arg min for the RDF, which contain information as to the average nearest neighbour distance, and the hard-sphere limit respectively. In general, persistent homology (and therefore L1NPIs) contain information about groups of oxygens. First degree homology details the behaviour of nearest neighbours within tetrahedral clusters. This can be contrasted to second degree homology, which for a system such as this details the behaviour of *next-nearest* neighbours within the same clusters.

Returning to L1NPIs, the projection of L1NPI space onto its first two principal components can be seen in Figure 5.14. From these, it can be seen that temperature causes a more pronounced change in first degree homology than second. This can be explained using the density found in persistence images. In second degree homology, a large amount of the density is located near the birth-axis. This is the topological noise. It would be expected that these noisy points would appear regardless of temperature. First degree homology has a density maximum within its topological features, i.e. the hexagonal cycles discussed previously. Small changes to density, as caused by changes in temperature, would alter these cycles, leading to more pronounced shifts in persistence images, and therefore greater separation in the principal components.

From the projection, it can be seen that it is worth investigating the first principal component, as this is where the different simulations are most separable. This principal component can be considered as a vector in L1NPI space, and can be seen in Figure 5.15. This supports the hypothesis that the first principal component corresponds to topological features in first degree homology, but utilises topological noise within second.

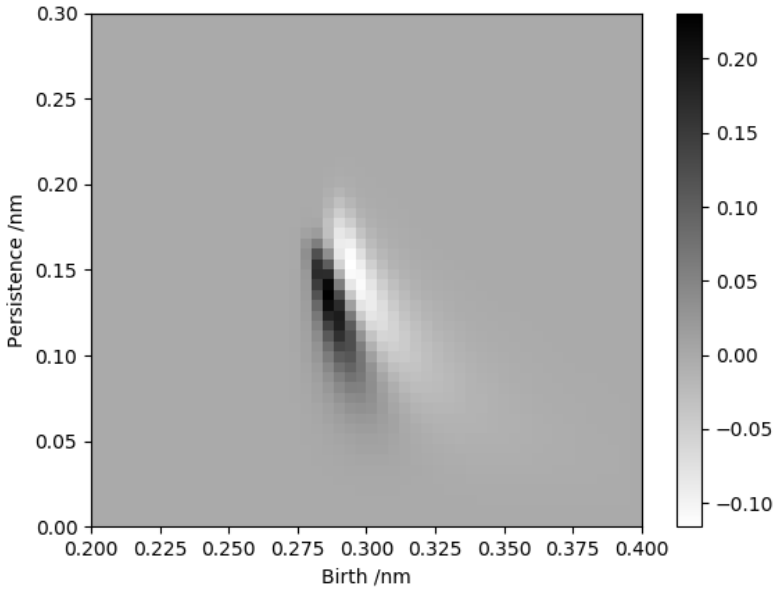


(a) First degree homology

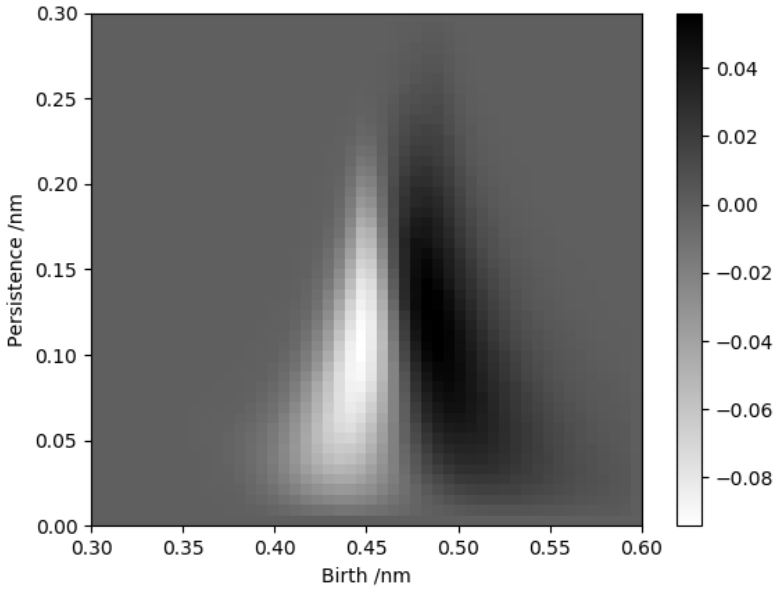


(b) Second degree homology

FIGURE 5.14: 2-dimensional principal component space for l_1 normalised persistence images of various temperatures. Points are coloured by temperature. There is a clear trend in temperature and location, as would be hoped for a chemical descriptor.



(a) First degree homology



(b) Second degree homology

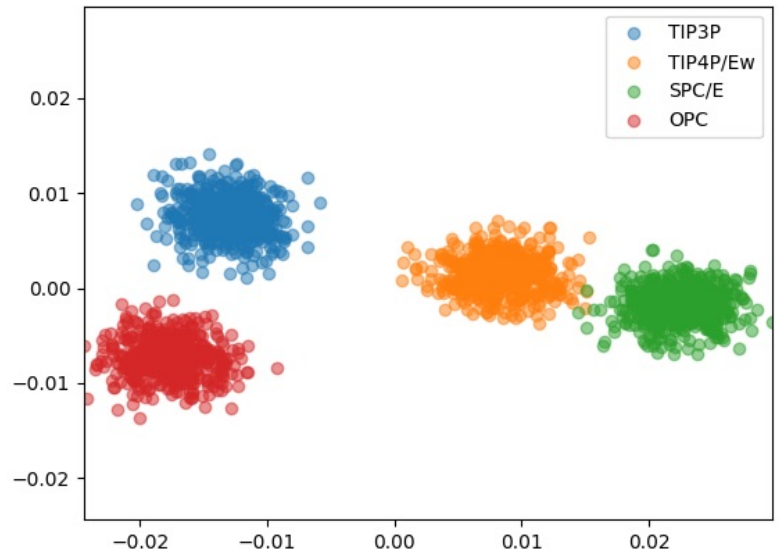
FIGURE 5.15: First principal component of L1NPI space for systems at different temperatures. The principal components illustrate how L1NPIs change as an effect of temperature.

5.8.2 Effect of Atomistic Model

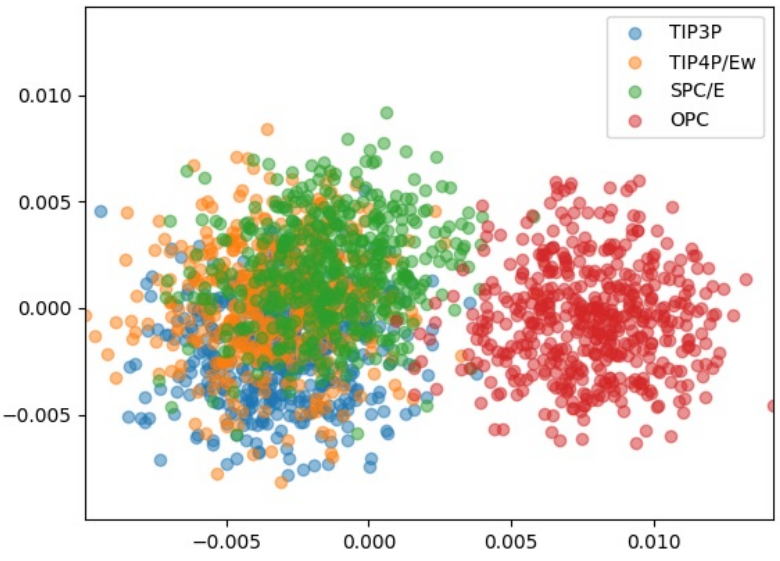
The L1NPI formalism has here been used to compare the differences between atomistic models. PCA on L1NPI space and linear SVM classifier confusion matrices can be found in Figures 5.16 and 5.17 respectively. Also, the PCA space, coloured by predicted class, can be seen in Figure 5.18. First degree homology clearly separates atomistic models. This is reflected in the confusion matrix, which is able to classify all models with a high accuracy. In contrast, the second degree homology is able to distinguish between OPC and other models, but not between TIP3P, TIP4P/Ew and SPC/E.

Looking in first degree homology in more detail, a few results are of interest. TIP4P/Ew and SPC/E are closer to each other than they are to the other models. This similarity is also reflected in the radial distribution functions for the atomistic models studied (Figure 5.1). However, it is not possible to state directly that first degree L1NPI similarity equates to similarity in the RDF, as the RDFs for TIP3P and OPC are not similar. This is discussed in more detail later in the text.

By observation of the principal component space, it might be expected that TIP3P should be identified with 100% accuracy, when this is not the case. This is due to two effects. Firstly, a linear kernel has been used for the SVM classifier. In this low dimensional representation, it is clear that it is not possible to draw a straight line separating TIP3P from all other models. Furthermore, the one-versus-rest multiclass strategy used here compounds this effect. This leads to a classifier being built for every class as a simple 'in/out' problem, and the predicted class chosen based on whichever class it is least likely to be an outlier of. This leads to the predictive behaviour seen in Figure 5.18(a).

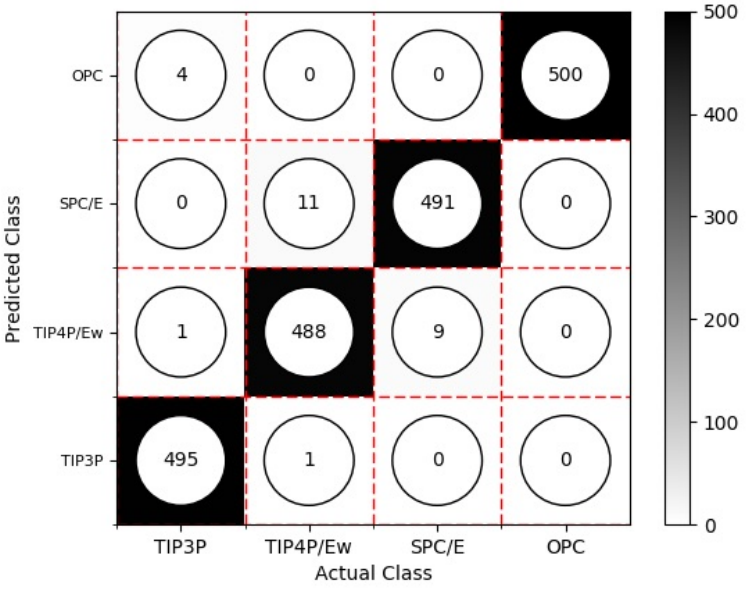


(a) First degree homology

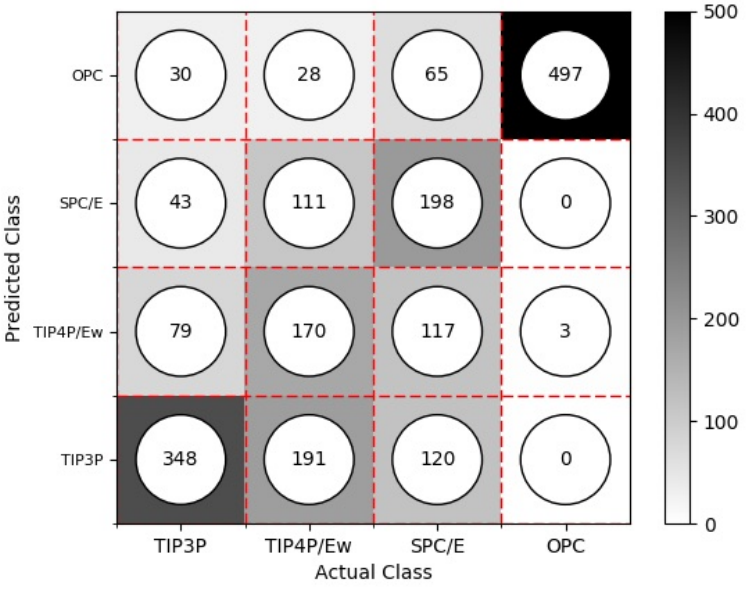


(b) Second degree homology

FIGURE 5.16: 2-dimensional principal component space for l_1 normalised persistence images of various atomistic water models. Points are coloured by atomistic model. All models are distinguishable in first degree homology, whereas only OPC can be distinguished in second degree homology.

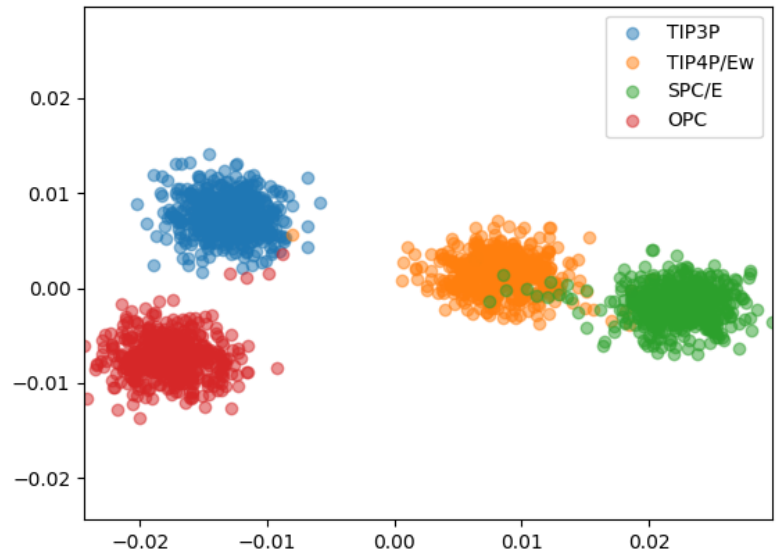


(a) First degree homology

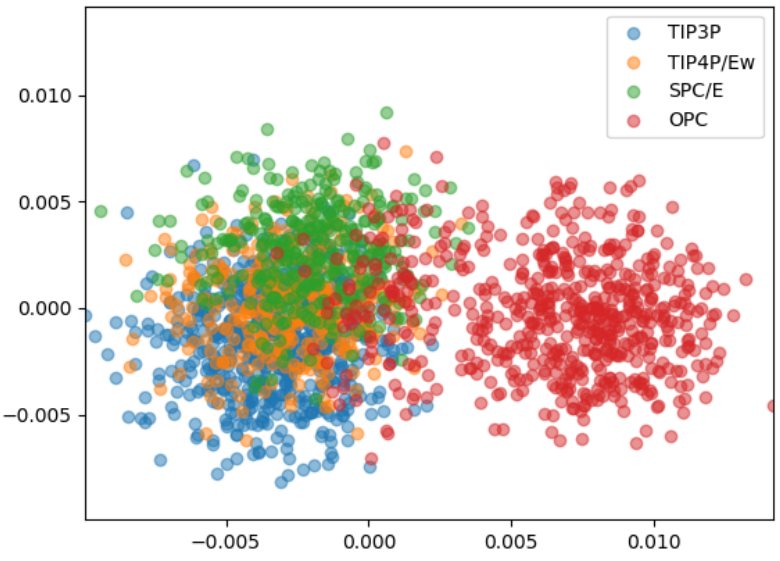


(b) Second degree homology

FIGURE 5.17: Confusion matrices for linear SVM classifiers for l_1 normalised persistence images of various atomistic water models, test set data. The behaviour of these classifiers is as expected from the principal component analysis of L1NPI space.



(a) First degree homology



(b) Second degree homology

FIGURE 5.18: 2-dimensional principal component space for l_1 normalised persistence images of various atomistic water models. Points are coloured by predicted model. The use of the linear kernel as a classifier can lead to unexpected behaviour, such as various models being mistaken as OPC when they are clearly distinguishable by eye.

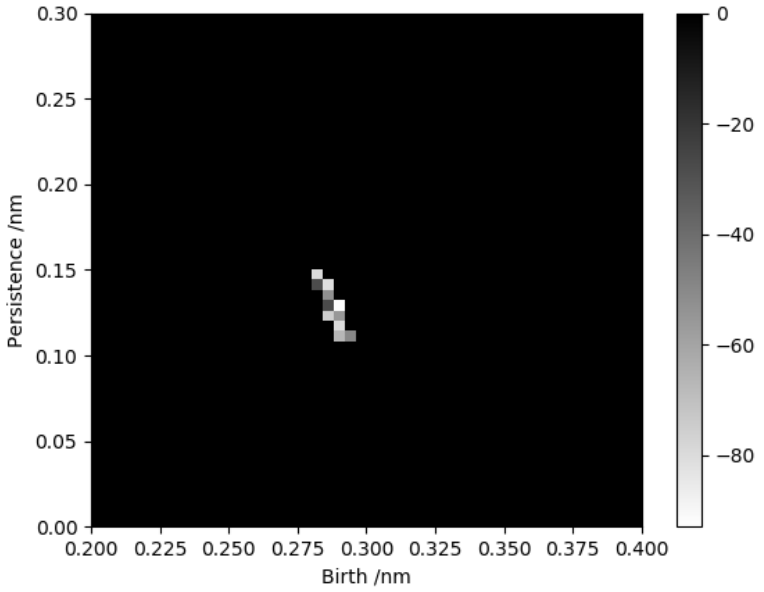
The coefficients of the linear SVM define a hyperplane in L1NPI space, and an associated tangent vector. These vectors can be viewed as differences between L1NPIs, although they are not quite L1NPIs themselves, as the vector can have a negative pixel. This allows the determination of features that separate particular models. As an example, the mean L1NPI and separating hyperplane for TIP3P water in first degree homology can be seen in Figure 5.19.

From the separating hyperplane, it can be seen that the SVM classifier believes that TIP3P is distinguished from the other models by features that are born reasonably early, with a range of death values. These features approximately correspond to the nearest neighbour behaviour of the system. To investigate this further, the mean of the first principal component (i.e the one corresponding to the highest variance within the dataset) was compared to the value of r leading to the highest value of the RDF (Table 5.6) The correlation of these two variables is not good, and certainly not strong enough

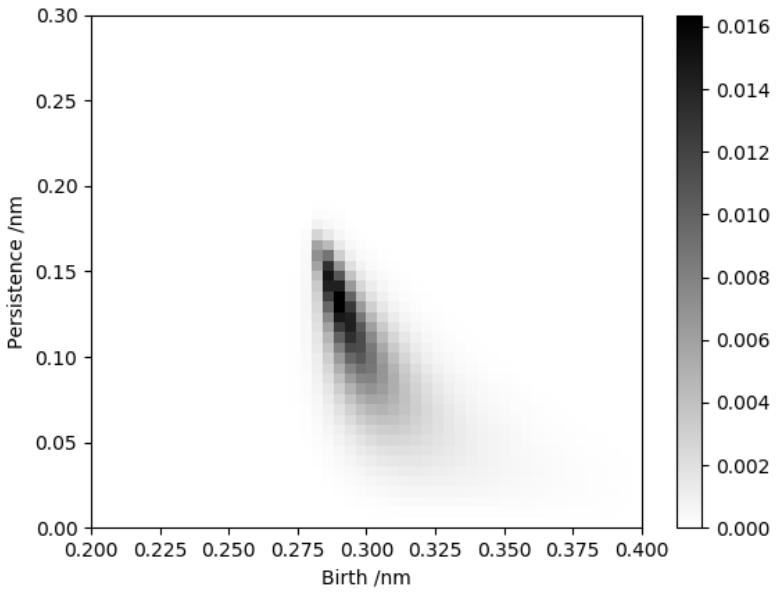
Model	$\langle PC1\rangle/10^{-3}$	$\sigma(PC1)/10^{-3}$	r_{max}/nm
TIP3P	-13.1	2.4	0.2773
TIP4P/Ew	8.3	2.6	0.2762
SPC/E	2.4	2.5	0.2746
OPC	-1.8	2.5	0.2815

TABLE 5.6: Properties of the first principal component and argument of maximum value of radial distribution function for the models studied.

to say for certain that the L1NPI behaviour for first degree homology matches the RDF maximum. However, it is useful to know that the L1NPI contains more information than just the RDF maximum! For example, Figure 5.19(b) suggests that it is a range of features separating TIP3P from other models. In contrast, the same information for OPC (Figure 5.20) shows that OPC is partially separated by a feature at the peak of the persistence image. This feature *must* correspond to the nearest neighbour behaviour, and from the RDF it can be seen that OPC has a longer hard-sphere limit than the other studied points. Therefore, some features can be related directly to the behaviour of the RDF.

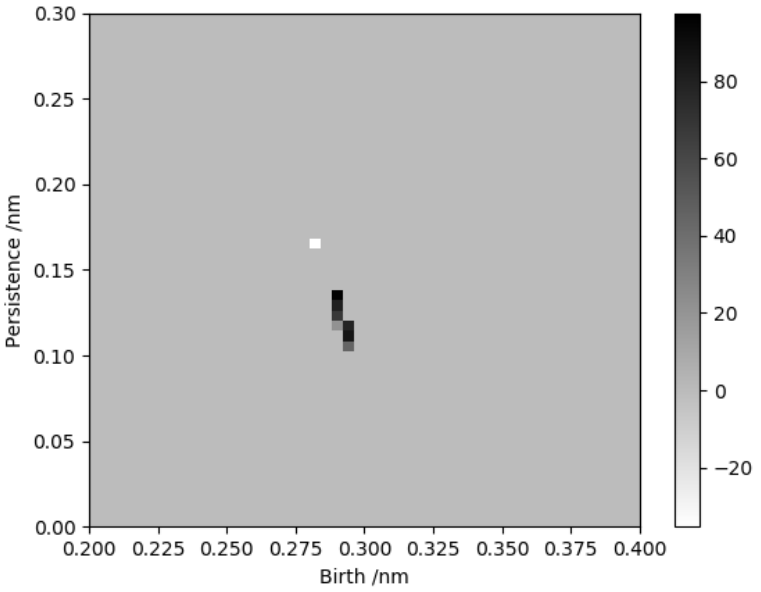


(a) Coefficients of separating hyperplane

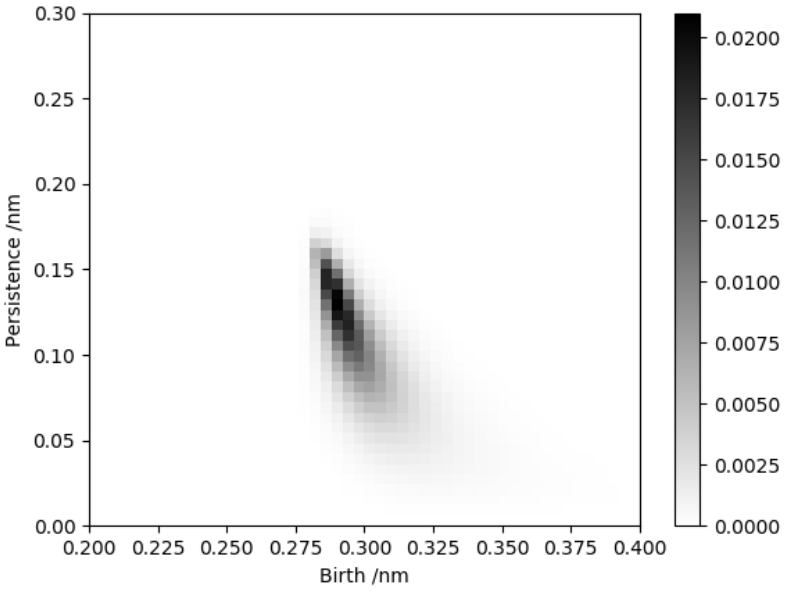


(b) Mean Image

FIGURE 5.19: Separating hyperplane and mean image for TIP3P first degree L1NPI space. It can be learned that it is a difference in points with high persistence that lead to differences between TIP3P and other models in first degree homology.



(a) Coefficients of separating hyperplane

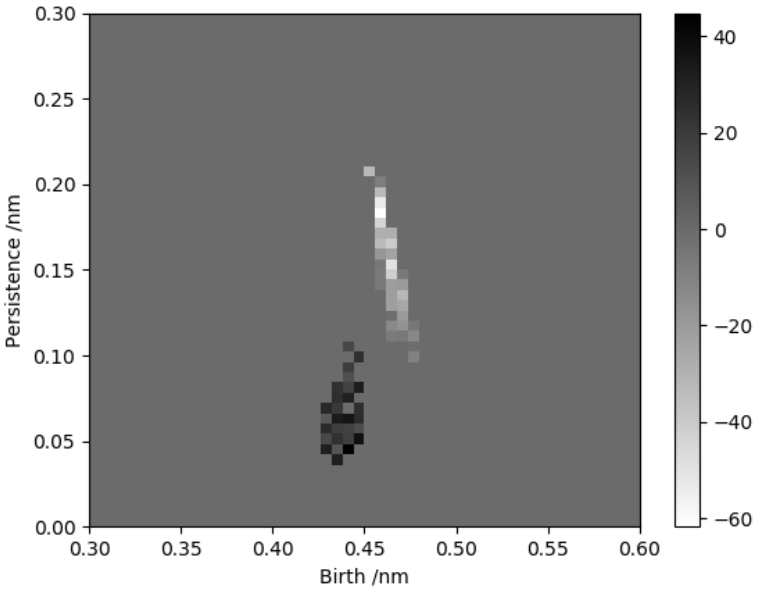


(b) Mean Image

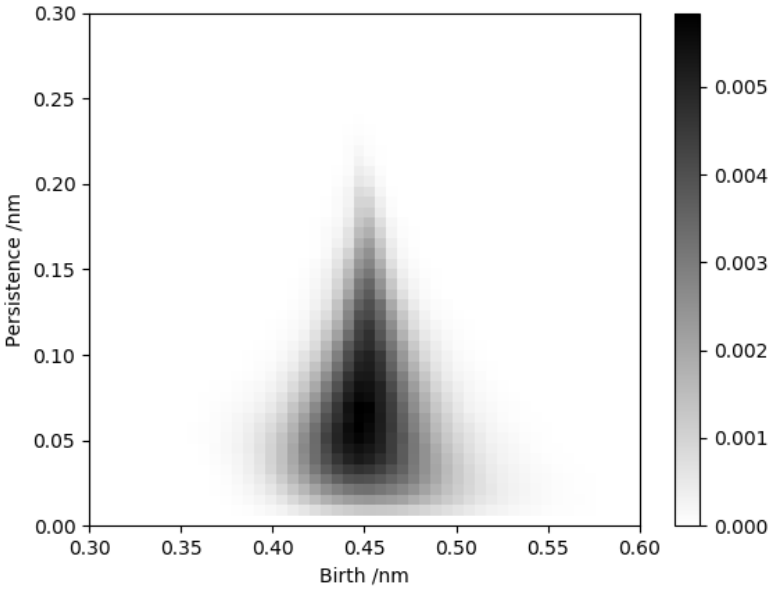
FIGURE 5.20: Separating hyperplane and mean image for OPC first degree L1NPI space. The presence of the single separate point in first degree homology is related to differences at the hard sphere limit of OPC and other models.

Second degree homology can also be studied in more detail. Most obviously, Figure 5.16(b) shows that TIP3P, TIP4P/Ew and SPC/E are largely the same with respect to their second degree homology, and it is not surprising that they cannot be separated. However, OPC can indeed be separated, and the SVM classifier performs reasonably well on both Type I and Type II errors.

Again, by looking at the coefficients it is possible to understand the features that distinguish OPC from other models (Figure 5.21). There are two main regions in this plot. Firstly, the topological noise region, in which it is difficult to make any conclusions. However, there is a clear region which is unlikely to be topological noise, and in particular these are the second degree points that are born late within the persistence cycle. The coefficients determine that OPC has far fewer of these points than the other models. OPC was parameterised differently to the other studied models in this work. In particular, rather than match derived physical quantities, the quantity fitted to was the electrostatic potential. That this has led to such a pronounced difference in second degree homology is certainly worth investigating more in future.



(a) Coefficients of separating hyperplane



(b) Mean Image

FIGURE 5.21: Separating hyperplane and mean image for OPC second degree L1NPI space. OPC contains fewer points born late within persistent homology than the other models, which could be due to differences in the parameterisation methods.

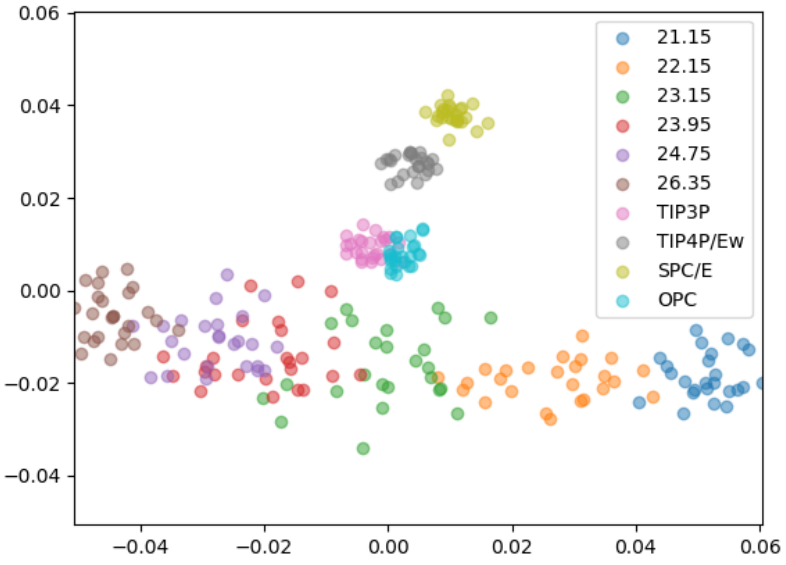
5.8.3 Comparison to Stillinger-Weber Model

As a final discussion with L1NPIs, the SW model is discussed. As mentioned, the λ parameter provides a useful measurement of 'tetrahedrality', tuning the strength of the 3-body interaction.

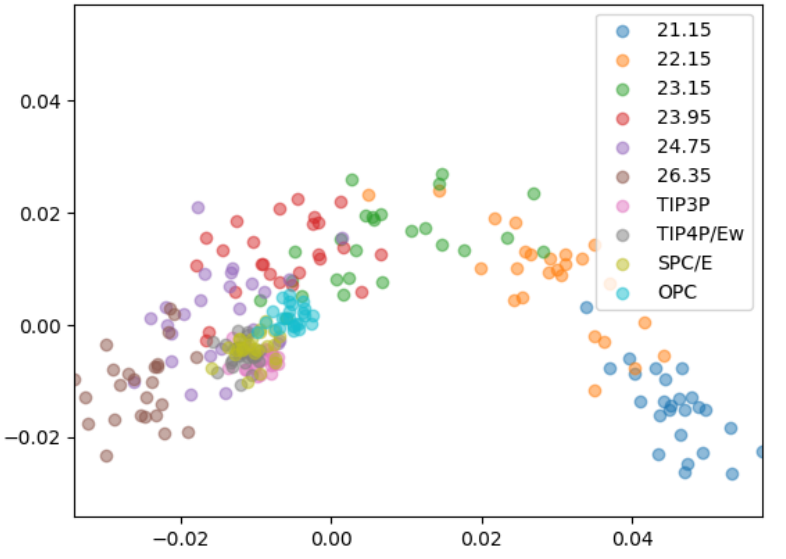
Firstly, the projection of first and second degree L1NPI space is found in Figure 5.22. The most obvious feature is the clear tracking with λ - there is a smooth variation in λ in the PC space, in both first and second degree homology. Also, the density of points for the SW model is much less than for the atomistic model. This suggests that there are larger variations in persistent homology for the SW model. However, rather than being a feature of the model itself, this is actually caused by the fact that there are only 500 water molecules in the simulations of the SW model, rather than in excess of 4000 for the atomistic models studied. The small number of points leads to larger variation in persistence, as there are fewer features. This can be compared to the results in Figure 5.23, showing the principal components of the unnormalised persistence images. It is clear that there is a large separation between the SW and atomistic models, which is due to the large discrepancy in the number of water molecules. In contrast, this does not occur within the L1NPI formalism, with the size-effect being the difference in density. Therefore, the persistence image representation is termed size-agnostic, as opposed to totally size-independent.

Returning to Figure 5.22, it is now possible look at each degree of homology in turn. For first degree homology, it is noted that all of the atomistic models are close to the set of L1NPIs for $\lambda = 23.15$, matching the value of λ for water. TIP3P and OPC are the models closest to the SW models, suggesting that they are the most similar.

With second degree homology, the atomistic models are no longer closest to $\lambda = 23.15$. Instead, the second degree homology is most similar to $\lambda = 23.95$. This separation of different degrees of homology is one of the strengths of the L1NPI analysis, where now it is possible to state that although the atomistic structures share similar loops to those created by the SW potential, they do not match the second degree holes. As mentioned earlier, OPC was parameterised differently to the atomistic models, and it is noted that this appears to make it more similar to the SW water model within the L1NPI analysis.

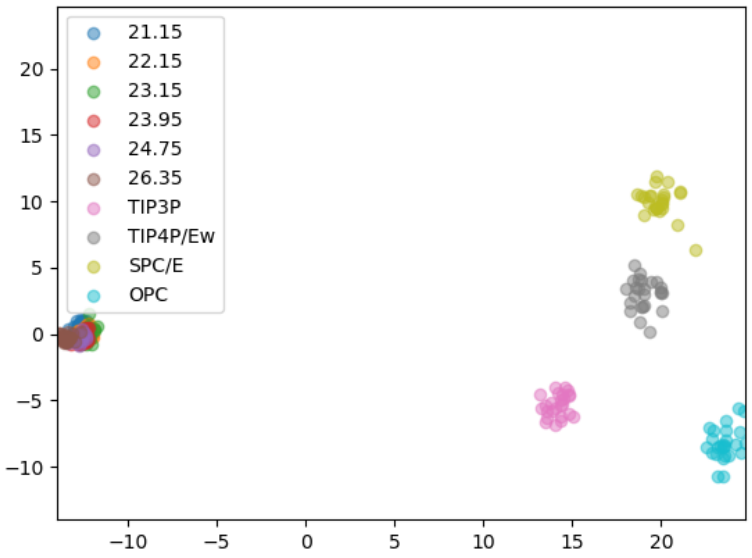


(a) First degree homology

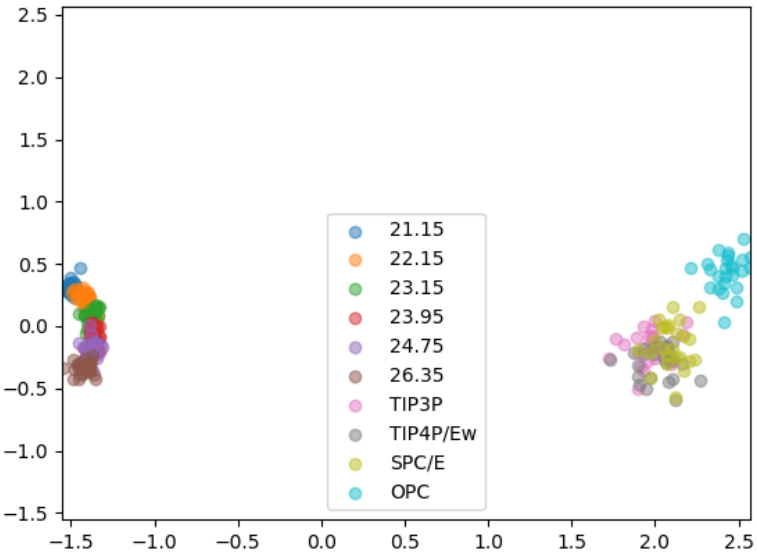


(b) Second degree homology

FIGURE 5.22: 2-dimensional principal component space for l_1 normalised persistence images of a selection of Stillinger-Weber and atomistic models. Points are coloured by model, with numerical values referring to the value of λ. In first degree homology the atomistic and coarse-grained models approximately coincide. However, in second degree homology the L1NPIs at the value of λ for water does not match the L1NPIs of the atomistic model.



(a) First degree homology



(b) Second degree homology

FIGURE 5.23: 2-dimensional principal component space for unnormalised persistence images of a selection of Stillinger-Weber and atomistic models. Points are coloured by model, with numerical values referring to the value of λ. Using the persistence image alone is clearly weighted by the size of the system.

5.9 Conclusions and Future Directions

This chapter showcases the development of a topological descriptor for the analysis of simulated water networks. Firstly, persistent homology has been used to analyse single frames of simulation, and it was shown that this contained information regarding periodic boundary conditions. Following this, the time-autocorrelation of persistent homology was analysed using the bottleneck metric, enabling the discussion of the use of an 'average' persistence, and demonstrating that it was a valid descriptor in this context.

Having shown this, a persistence image based descriptor was developed, in particular aiming to demonstrate size-independence, as it was felt that this is an important property of any descriptor hoping to analyse bulk behaviour. The derived descriptor, the l_1-normalised persistence image (L1NPI) was shown to be the most size-independent of those studied. L1NPIs were then used to analyse a range of simulated water networks, varying in model and temperature. It was shown that the L1NPIs were able to distinguish well between different models in various degrees of homology, and were also able to provide interpretations as to how the studied systems varied. Lastly, L1NPIs were used to compare atomistic and coarse-grained water models, with a large variation in the number of water molecules. This enabled a more full discussion as to the size-independent nature of the L1NPI, as well as demonstrating their utility in the comparison of simulated systems of different materials. It was shown that the L1NPI is not fully size-independent, but instead the effect of size can be easily accounted for in this formalism.

In future, this method could be extended to understand other water models. Polarisable forcefields, such as AMOEBA [219, 220] could be a future area of study. However, as these models are flexible, it might be important to understand the behaviour of the hydrogen atoms of the water molecules. This work only calculates the persistent homology of the oxygen atoms, and therefore a new persistent homology based procedure should be designed if necessary.

Considering the original goal of this work was to aid in solubility prediction, this has to be considered as one of the main future directions of this work. Understanding how the presence of a solute alters the surrounding water network will likely lead to insight in solvation entropy prediction, for example. However, there are several problems that would need to be tackled before progress could be made. In particular, the solute itself would cause a hole in the water network, and this would be detected with persistent homology. Care would need to be taken to ensure that methods are not simply measuring the size of this hole, as more efficient methods exist for this purpose [221]. One potential route in tackling this problem would be the use of Alexander duality - which relates the homology groups of an object to that of its complement. However, at the time of writing,

the only persistent homology extension of this property is through extended persistence [222], and there is no efficient algorithm for its computation.

Chapter 6

Conclusions

The explosion of data within the physical sciences requires a plethora of new tools for its analysis. Although there will always be a place for standard statistical techniques, the development of topological data analysis tools can lead to the answering of new questions, and new methods with which to study all chemical topics. This investigation has attempted to demonstrate three different areas of study for topological data analysis tools within chemistry, and show that topological techniques could be a useful weapon in the arsenal of any chemist.

6.1 Chemical Space

The mapper algorithm has been used to analyse the underlying descriptor space for a data set used in water solubility prediction. Correlations in the mapper network were analysed, and it was shown that there were global correlations for the number of rings and the number of atoms in a molecule. Furthermore, previously unseen correlations led to the conclusion that molecular solubility has a strong dependence on the number of chlorines, but only for those molecules with two cycles. This led to the creation of data-driven models, which were shown to improve the consistency of prediction when compared to ESOL, another widely used solubility model.

Persistent homology has also been used as a descriptor for chemical shape. Using metrics on persistent homology, various maps of chemical shape space have been created, each highlighting different features. It has been shown that the chemical shape space created via persistent homology is strongly linked to the number of atoms in a molecule, and the number of rings - topological features of the molecule as opposed to true chemical features. The effect of different persistent kernels was analysed, and for the simple case studied was shown to have little or no effect. Reasons for this were discussed, as well as possible changes that could be made and what they would likely affect. Finally, the

effect of conformational flexibility on the underlying chemical shape space was studied. It was found that when multiple low-energy conformations were used to create the shape space, it was largely unchanged when compared to the original space created from the minimum energy conformation. The methods used in this chapter are therefore insensitive to this flexibility, although further study can be performed to create shape spaces from persistent homology, that are affected by the existence of multiple low-energy conformations.

6.2 Conformational Space

Persistent homology has been used to characterise the underlying conformational space and energy landscapes of three molecules. Firstly, two commonly used representations of conformational space were defined - a coordinate based representation utilising the Euclidean metric, and a distance matrix using the RMSD metric. It was shown for alanine dipeptide, the difference between these representations was negligible, however for pentane the coordinate representation incorrectly identified the conformational space. This discrepancy was speculated to be due to the difference in alignment procedures used in the creation of the coordinate representation, and it was concluded that the RMSD distance matrix representation should be used in future. With both alanine dipeptide and pentane, the conformational space was found to be a torus, and the energy landscape of alanine dipeptide was analysed and critical points located.

For cyclooctane, the RMSD representation was used to verify the results of Martin *et al*'s landmark paper. Persistent homology showed that the conformational space was indeed non-manifold. The conformational space was then analysed using local PCA, to remove the non-manifold points. A hierarchical clustering algorithm was used to separate the space, before clusters were re-glued back into their manifold components. Persistence was then able to verify the presence of spherical and klein bottle components. The single point energy landscape of the spherical component was then analysed, with critical points located and shown to match those found by other procedures.

6.3 Water Networks

Persistent homology was used to develop a new descriptor for water network structure. Firstly, persistence was shown to be a well-behaved descriptor, with non-recurrent behaviour. Persistence was then shown to be strongly dependant on system size. Using persistence images, a range of descriptors were produced for water network structure, of which one (the L1NPI) was found to be the most size-independent. The L1NPI was then shown to be sensitive to the temperature of simulations. Furthermore, it was found that they could be used for the comparison of different atomistic potentials. Through the use

of support vector machines, differences between water models were analysed through features in their first and second degree homology. Lastly, the L1NPI was used to compare atomistic and coarse-grained water potentials. Through dimensionality reduction techniques, it was possible to demonstrate which of the coarse-grained potentials best matched the atomistic potentials.

Appendix A

Data Science Techniques

This appendix provides a brief introduction to the data science techniques used in the text. These techniques are split into dimensionality reduction, and classification. The dimensionality reduction techniques used in this work are principal component analysis (PCA) and multidimensional scaling (MDS), with the classification technique being a support vector machine (SVM). All of these techniques are fairly common within the world of data science, and indeed more recently chemistry. However, it is worth describing these techniques in more detail.

A.1 Dimensionality Reduction

A data set X of n observations of m variables can be written as a matrix in $\mathbb{R}^{n\times m}$:

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,m} \\ x_{2,1} & x_{2,2} & \dots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,m} \end{pmatrix} \tag{A.1}$$

At its simplest, dimensionality reduction can be understood as a result of the need to visualise this data set in two or three dimensions. This visualisation could allow hidden relationships within the data to be found. However, as dimensionality reduction reduces the degrees of freedom of a data set, their is a strong possibility of some information being lost. Different dimensionality reduction techniques essentially seek to preserve different features of the data set, with their own perspective of what information is thought to be important.

A.1.1 Principal Component Analysis

Principal component analysis (PCA) was originally described by Pearson in 1901 as an analogue to finding the principal axes of rotation of a rigid body [223]. PCA defines the transformation from the original descriptor space onto the set of mutually orthogonal axes with the highest variances.

Firstly, the covariance matrix K of X is calculated, $X^T X$. This matrix is then diagonalised through eigendecomposition. The eigenvector matrix W defines the principal components, with the eigenvalues $\{\alpha_j\}$ corresponding to the relative variance described by the j^{th} eigenvector. The full-dimensional principal component decomposition of X is then calculated as:

$$T = XW \tag{A.2}$$

Which can be projected onto k dimensions by taking the first k columns of T. As a linear transformation, the principal components can be easily computed. Furthermore, if X exists on a k-dimensional subspace of the original m variables, k-dimensional PCA will find that subspace exactly, with no loss of information - similarly if all columns of T are used in the projection there is also no information loss.

Before PCA, the original matrix X is often scaled such that all of the individual descriptors have a mean of zero and a variance of 1. This transformation ensures that the data set is centred on the origin, and also that the principal components are not dominated by descriptors with a larger numerical range.

A.1.2 Multidimensional Scaling

Multidimensional scaling (MDS) seeks to find the low-dimensional representation of a data set which preserves the high-dimensional distances. Therefore, rather than operating on the original data set X, MDS utilises a distance matrix D:

$$D = \begin{pmatrix} d(x_1, x_1) & d(x_1, x_2) & \dots & d(x_1, x_n) \\ d(x_2, x_1) & d(x_2, x_2) & \dots & d(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ d(x_n, x_1) & d(x_n, x_2) & \dots & d(x_n, x_n) \end{pmatrix} \tag{A.3}$$

Where x_i refers to the i^{th} observation of X, i.e. its i^{th} row. The matrix D is therefore an $n \times n$ matrix, which is normally much larger than X, because $n \gg m$.

The function $d(x_i, x_j)$ must satisfy the properties of a metric:

1. $d(x_i, x_i) = 0$
2. $d(x_i, x_j) = d(x_j, x_i)$

3. $d(x_i, x_k) \leq d(x_i, x_j) + d(x_j, x_k)$

but in principle can be any function that satisfies these properties. Therefore, a user is able to use any notion of [dis]similarity to create their own metric for MDS.

MDS seeks to find the matrix D' that minimises the following:

$$s = \left(\frac{\sum_{i,j} (d_{ij} - d'_{ij})^2}{\sum_{i,j} d_{ij}^2} \right)^{\frac{1}{2}} \tag{A.4}$$

Where d_{ij} is a Euclidean distance matrix. The quantity s is often referred to as the *stress* of the dimensionality reduction. Note that, whereas PCA will not lose information if projected into the same number of dimensions as the original space, MDS does not have the same guarantee. In particular, MDS requires $(n-1)$ dimensions to ensure $s = 0$. This is because the original distance metric D does not have to be Euclidean, but in general a metric of n points can be shown to be equivalent to a Euclidean metric in $(n-1)$ dimensions.

A.2 Support Vector Machines

Consider the data set X along with a function $f : X \mapsto \{-1, 1\}$. f assigns a class to all of the points in X. A support vector machine seeks to find the $(m-1)$ dimensional hyperplane that separates the data into their classes. As there are potentially many hyperplanes that can achieve this goal, the support vector machine specifically seeks to find the classifier that minimises the distances between the points and the hyperplane (Figure A.1).

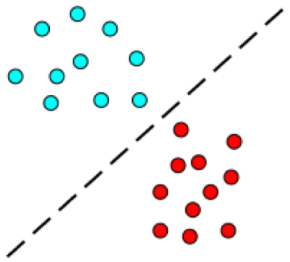


FIGURE A.1: Cartoon illustrating the classification boundary found by a linear SVM

In general, the data classes may not be linearly separable. In this case, the SVM will act to keep errors to a minimum, with a parameter to tune the relative strength of this penalisation.

Within this work, a linear SVM is employed as a classifier. Furthermore, the classifier is used on a multiple class problem. The SVM handles multiple classes by creating a

separate classifier between a class and all other classes (i.e. n_{class} classifiers), and when predicting a new observation choosing the class with the strongest confidence. The distance function used is the l_1 metric, which results in hyperplanes with sparse coefficients. The normal coefficients to the hyperplane are visualised, and used to understand the behaviour of the classifier. However it should be noted that this can be unwise in certain cases (see reference [224] for more information on this topic).

Appendix B

Master Input Files for AMBER

Temperature Equilibration

```
TIP3P: 100ps MD with equilibration T 0-->!TEMPK, cutoff = 10
&cntrl
imin   = 0,
irest  = 0,
ntx    = 1,
ntb    = 1,
cut    = 10.0,
ntc    = 2,
ntf    = 2,
tempi  = 0.0,
temp0  = !TEMP,
ntt    = 3,
gamma_ln   = 10.0,
nstlim = 50000, dt = 0.002,
ntpr = 1000, ntwx = 1000, ntwr = 10000,
ioutfm = 1, iwrap = 1, ig = -1,
nmropt = 1,
/
&ewald
vdwmeth = 0
/
&wt TYPE=TEMP0, ISTEP1=1, ISTEP2=40000, VALUE1=0, VALUE2=!TEMP
/
&wt TYPE=TEMP0, ISTEP1=40001, ISTEP2=50000, VALUE1=!TEMP, VALUE2=!TEMP
/
&wt TYPE=END
/
/
```

Pressure Equilibration

```
TIP3P: 500ps MD, cutoff = 10
&cntrl
imin   = 0,
irest  = 1,
ntx    = 5,
ntb    = 2,
ntp    = 1,
pres0  = 1,
cut    = 10.0,
ntc    = 2,
ntf    = 2,
tempi  = !TEMP,
temp0  = !TEMP,
ntt    = 3,
gamma_ln   = 10.0,  taup = 2,
nstlim = 250000, dt = 0.002,
ntpr = 1000, ntwx = 1000, ntwr = 10000,
ioutfm = 1, iwrap = 1, ig = -1,
/
&ewald
vdwmeth = 0
/
```

Production Run

```
TIP3P: 4ns MD cube, cutoff = 10
&cntrl
imin   = 0,
irest  = 1,
ntx    = 5,
ntb    = 1,
cut    = 10.0,
ntc    = 2,
ntf    = 2,
tempi  = !TEMP,
temp0  = !TEMP,
ntt    = 3,
ene_avg_sampling = 1,
gamma_ln   = 10.0,
nstlim = 2000000, dt = 0.002,
ntpr = 1000, ntwx = 1000, ntwr = 10000,
ioutfm = 1, iwrap = 1, ig = -1,
/
&ewald
vdwmeth = 1
/
```

Appendix C

On Rings and Fields

Within abstract algebra, rings and fields are basic structures, similar to the notion of a set. However, whereas a set has no inherent operations, rings and fields are given mathematical operations that lead to enhanced richness. This appendix will detail these operations, illustrating the differences between rings and fields, which should lead the reader to understand why fields are chosen to be used in this work, rather than the more flexible rings.

A ring R is an algebraic structure of a set $\{a, b, c, \ldots\}$ with two binary operations (denoted $+$ and $\times$, in the sense that they generalise familiar addition and multiplication). These operations obey the *ring axioms*:

1. R is an abelian group under $+$:
   (a) $(a + b) + c = a + (b + c)$: $+$ is associative
   (b) $a + b = b + a$: $+$ is commutative
   (c) There is an element $0 \in R$ such that $a + 0 = a$ for all elements in R: $+$ has an identity
   (d) For each $a \in R$ there exists an $a' \in R$ such that $a + a' = 0$: $+$ has an inverse
2. The properties of $\times$ are as follows:
   (a) $(a \times b) \times c = a \times (b \times c)$: $\times$ is associative
   (b) There is an element $1 \in R$ such that $a \times 1 = 1 \times a = a$ for all a in R: $\times$ has an identity
3. $\times$ is distributive with respect to $+$
   (a) $a \times (b + c) = (a \times b) + (a \times c)$
   (b) $(b + c) \times a = (b \times a) + (c \times a)$

Examples of rings include the integers $\mathbb{Z}$, and the set of 2×2 real matrices (both equipped with addition and multiplication as usually defined).

Fields $\mathbb{F}$ can be considered to be rings with further axioms - they are more restrictive. These axioms are:

1. $a \times b = b \times a$: $\times$ is commutative
2. For each $a \neq 0 \in R$ there exists an $a'' \in R$ such that $a \times a'' = 1$: $\times$ has an inverse

Common examples of fields are the rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$, and the set of 2×2 orthogonal matrices (again all equipped with their usual definitions of addition and multiplication).

An interesting set of algebraic structures are the set of integers modulo a natural number, $\mathbb{Z}_n = \{0, 1, 2, \ldots, n\}$. It can be shown that the multiplicative inverse only exists if n is a prime number. For example, take $\mathbb{Z}_3$. The inverse of 1 is obviously also 1. To find the inverse of 2, the equation to be solved is $2 \times x = 1$. It can be seen that $2 \times 2 = 4$, and $4\%3 = 1$, so the inverse of 2 is 2 in mod 3 arithmetic.

In contrast take the set $\mathbb{Z}_4 = \{0, 1, 2, 3\}$. $2 \times 1 = 2$, $2 \times 2 = 0$, $2 \times 3 = 2$ in mod 4 arithmetic. Therefore, 2 does not have a multiplicative inverse in this set. Summarising, all of the sets $\mathbb{Z}_n$ are rings, with the sets $\mathbb{Z}_p$ (p is prime) are fields, when equipped with traditional multiplication and addition.

Appendix D

The Homology Groups of the Klein Bottle and Real Projective Plane

In the following, $\mathbb{F}$ denotes an arbitrary field.

D.1 Klein Bottle

To calculate the homology groups of the Klein bottle, first a combinatorial representation is needed. Rather than use a simplicial complex, here the CW-complex is used. This simplifies the calculation - and as both representations are homeomorphic to the Klein bottle the same result is obtained. The CW-complex can be seen in Figure D.1. The complex consists of a single vertex v, two directed lines a and b, as well as the face F. For those unfamilar, this construction is completed by 'gluing' together simplices with the same label, such that orientation is preserved. For example, the CW-complex for the torus differs from that for the Klein bottle with the lines b pointing in the same direction.

The sequence of chain complexes for this space is written as follows:

$$0 \xrightarrow{\partial_3} C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0 \tag{D.1}$$

Where C_n is the free abelian group generated by the n-simplices, for example C_0 is the free abelian group generated by v, which is isomorphic to $\mathbb{F}$. For the homology groups, it is necessary to calculate the operation of ∂_p on the p-chains $\in C_p$, in particular their basis elements:

$$\partial_2 F = b + a + b - a = 2b \tag{D.2}$$

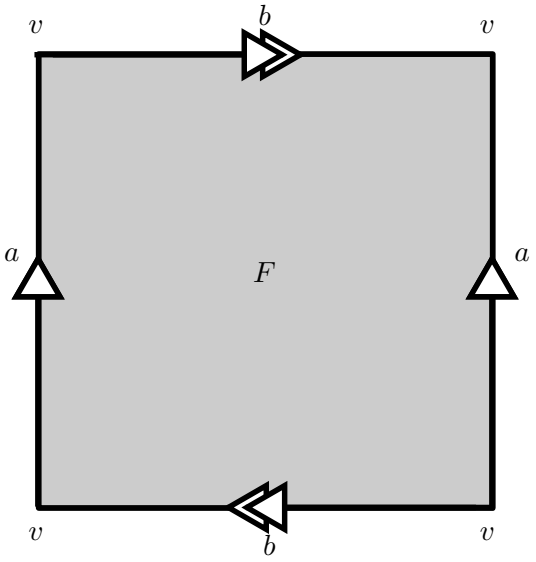


FIGURE D.1: The CW-complex of the Klein Bottle

$$\partial_1 a = v - v = 0 = \partial_1 b \tag{D.3}$$

$$\partial_0 v = 0 \tag{D.4}$$

Therefore, the following is true for any field:

$$\begin{aligned} \partial_2 : \mathbb{F} &\to \mathbb{F} \\ 1 &\mapsto 2 \end{aligned} \tag{D.5}$$

$$\begin{aligned} \partial_1 : \mathbb{F} \oplus \mathbb{F} &\to \mathbb{F} \\ (1,1) &\mapsto 0 \end{aligned} \tag{D.6}$$

$$\begin{aligned} \partial_0 : \mathbb{F} &\to \mathbb{F} \\ 1 &\mapsto 0 \end{aligned} \tag{D.7}$$

Thus enabling the determination of the kernels and images of ∂_n. For ∂_2:

$$\begin{aligned} \ker \partial_2 &= \begin{cases} \mathbb{F} & \text{if } \mathbb{F} \simeq \mathbb{Z}_2 \\ 0 & \text{otherwise} \end{cases} \\ \operatorname{im} \partial_2 &= \begin{cases} 0 & \text{if } \mathbb{F} \simeq \mathbb{Z}_2 \\ \mathbb{F} & \text{otherwise} \end{cases} \end{aligned} \tag{D.8}$$

This is because for any $\alpha \in \mathbb{Z}_2$, $2\alpha = 0$, whereas $2\mathbb{F} \simeq \mathbb{F}$ if $\mathbb{F} \not\simeq \mathbb{Z}_2$. For ∂_1:

$$\begin{aligned} \ker \partial_1 &= \mathbb{F} \oplus \mathbb{F} \\ \operatorname{im} \partial_1 &= 0 \end{aligned}$$

and for ∂_0, trivially:

$$\ker \partial_0 = \mathbb{F}$$
$$\operatorname{im} \partial_0 = 0$$

The homology groups $H_p \equiv \frac{\ker \partial_p}{\operatorname{im} \partial_{p+1}}$ are therefore:

$$H_0 = \mathbb{F}$$
$$H_1 = \begin{cases} \mathbb{F} \oplus \mathbb{F} & \text{if } \mathbb{F} \simeq \mathbb{Z}_2 \\ \mathbb{F} & \text{otherwise} \end{cases}$$
$$H_2 = \begin{cases} \mathbb{F} & \text{if } \mathbb{F} \simeq \mathbb{Z}_2 \\ 0 & \text{otherwise} \end{cases}$$

And the Betti numbers $\beta_p = \dim_{\mathbb{F}} H_p$:

$$\beta_0 = 1$$
$$\beta_1 = \begin{cases} 2 & \text{if } \mathbb{F} \simeq \mathbb{Z}_2 \\ 1 & \text{otherwise} \end{cases}$$
$$\beta_2 = \begin{cases} 1 & \text{if } \mathbb{F} \simeq \mathbb{Z}_2 \\ 0 & \text{otherwise} \end{cases}$$

It is noted that the results obtained here are different than those found in standard texts. This is because traditionally coefficients are taken in the ring of integers $\mathbb{Z}$ rather than an arbitrary field $\mathbb{F}$. For completeness, if $\mathbb{Z}$ was used for coefficients, the Betti numbers would match those of $\mathbb{F} \not\simeq \mathbb{Z}_2$, although the homology group $H_1 = \mathbb{Z} \oplus \mathbb{Z}_2$.

D.2 Real Projective Plane

A CW-complex for the real projective plane $\mathbb{R}P^2$ can be found in Figure D.2. The sequence of chain complexes is the same as above, as the simplices are of the same dimension as for the Klein bottle.

$$0 \xrightarrow{\partial_3} C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0 \tag{D.9}$$

The operation of ∂_p on the elements of C_p is written in terms of the basis elements:

$$\partial_2 F = 2a \tag{D.10}$$

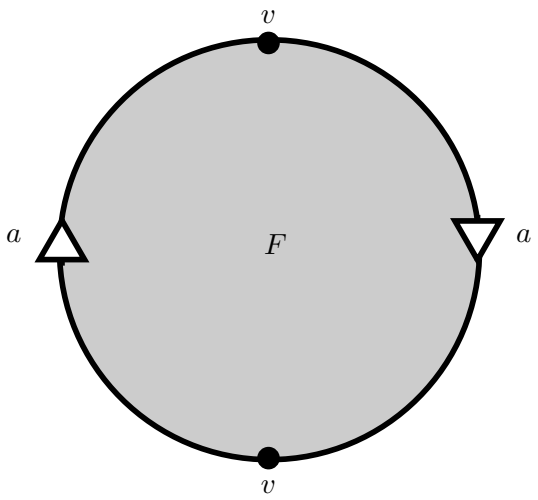


FIGURE D.2: The CW-complex of the Real Projective Plane

$$\partial_1 a = v - v = 0 \tag{D.11}$$

$$\partial_0 v = 0 \tag{D.12}$$

Therefore, for any field:

$$\begin{aligned} \partial_2 : \mathbb{F} &\to \mathbb{F} \\ 1 &\mapsto 2 \end{aligned} \tag{D.13}$$

$$\begin{aligned} \partial_1 : \mathbb{F} &\to \mathbb{F} \\ 1 &\mapsto 0 \end{aligned} \tag{D.14}$$

$$\begin{aligned} \partial_0 : \mathbb{F} &\to \mathbb{F} \\ 1 &\mapsto 0 \end{aligned} \tag{D.15}$$

The kernels and images of ∂_p are as follows:

$$\begin{aligned} \ker \partial_2 &= \begin{cases} \mathbb{F} & \text{if } \mathbb{F} \simeq \mathbb{Z}_2 \\ 0 & \text{otherwise} \end{cases} \\ \operatorname{im} \partial_2 &= \begin{cases} 0 & \text{if } \mathbb{F} \simeq \mathbb{Z}_2 \\ \mathbb{F} & \text{otherwise} \end{cases} \end{aligned} \tag{D.16}$$

$$\begin{aligned} \ker \partial_1 &= \mathbb{F} \\ \operatorname{im} \partial_1 &= 0 \end{aligned}$$

$$\begin{aligned} \ker \partial_0 &= \mathbb{F} \\ \operatorname{im} \partial_0 &= 0 \end{aligned}$$

The homology groups $H_p \equiv \frac{\ker \partial_p}{\mathrm{im}\, \partial_{p+1}}$ are therefore:

$$H_0 = \mathbb{F}$$

$$H_1 = \begin{cases} \mathbb{F} & \text{if } \mathbb{F} \simeq \mathbb{Z}_2 \\ 0 & \text{otherwise} \end{cases}$$

$$H_2 = \begin{cases} \mathbb{F} & \text{if } \mathbb{F} \simeq \mathbb{Z}_2 \\ 0 & \text{otherwise} \end{cases}$$

And the Betti numbers $\beta_p = \dim_{\mathbb{F}} H_p$:

$$\beta_0 = 1$$

$$\beta_1 = \begin{cases} 1 & \text{if } \mathbb{F} \simeq \mathbb{Z}_2 \\ 0 & \text{otherwise} \end{cases}$$

$$\beta_2 = \begin{cases} 1 & \text{if } \mathbb{F} \simeq \mathbb{Z}_2 \\ 0 & \text{otherwise} \end{cases}$$

Appendix E

Proof that repeated application of boundary map is zero

Here, coefficients are taken in $\mathbb{Z}_2$. The proof for coefficients in the integers is found in Hatcher [32]. The definition of the boundary map on a p-simplex σ is:

$$\partial_p(\sigma) = \sum_{i=0}^{p} [v_0, \ldots, \hat{v}_i, \ldots, v_p] \tag{E.1}$$

Where $\hat{v}_j$ is used to denote the removal of the j^{th} vertex. This appendix demonstrates the standard result that $\operatorname{im} \partial_{\mathrm{p}} \subseteq \ker \partial_{\mathrm{p}-1}$, or equivalently $\partial_{p-1} \circ \partial_p = 0$. Making vertices explicit, σ can be written as:

$$[v_0, \ldots, v_n]$$

It follows that a face $\sigma' \subset \sigma$ can be written as:

$$\sigma'_i = [v_0, \ldots, \hat{v}_i, \ldots, v_p]$$

Or that Equation E.1 can be written as:

$$\partial_p(\sigma) = \sum_{i=0}^{p} \sigma'_i \tag{E.2}$$

i.e. the vertex which has been removed from σ becomes explicit. Applying Equation E.1, the operation of ∂_{p-1} on σ' can be written as:

$$\partial_{p-1}\left(\partial_p(\sigma)\right) = \partial_{p-1}\left(\sum_{i=0}^{p} \sigma'_i\right) = \sum_{i=0}^{p} \partial_{p-1}\left(\sigma'_i\right) \tag{E.3}$$

as the boundary operator is linear. The term inside the sum can be found:

$$\partial_{p-1}\left(\sigma_i'\right) = \sum_{j=0}^{i-1}[v_0, \ldots, \hat{v}_j, \ldots, \hat{v}_i, \ldots, v_p] + \sum_{j=i+1}^{p}[v_0, \ldots, \hat{v}_i, \ldots, \hat{v}_j, \ldots, v_p]$$

The first sum denotes the removal of vertices up to v_i, the second denotes the removal of vertices following v_i:

$$\partial_{p-1}\sigma' = \sum_{j<i}[v_0, \ldots, \hat{v}_j, \ldots, \hat{v}_i, \ldots, v_p] + \sum_{j>i}[v_0, \ldots, \hat{v}_i, \ldots, \hat{v}_j, \ldots, v_p]$$

Defining $[v_0, \ldots, \hat{v}_i, \ldots, \hat{v}_j, \ldots, v_p]$ as σ''_{ij}, the above can be rewritten as:

$$\partial_{p-1}\sigma' = \sum_{j<i}\sigma''_{ij} + \sum_{j>i}\sigma''_{ji}$$

Which can be put back into the sum:

$$\partial_{p-1}\left(\sigma_i'\right) = \sum_{i=0}^{p}\left(\sum_{j<i}\sigma''_{ij} + \sum_{j>i}\sigma''_{ji}\right)$$

The outer sum can be factored in, as it is independent from the inner sums:

$$\partial_{p-1}\left(\partial_p(\sigma)\right) = \sum_{j<i}\sum_{i=0}^{p}\sigma''_{ij} + \sum_{j>i}\sum_{i=0}^{p}\sigma''_{ji}$$

The two terms are equal, as the indices i and j may be swapped as they are arbitrary.

$$\partial_{p-1}\left(\partial_p(\sigma)\right) = 2\sum_{j<i}\sum_{i=0}^{p}\sigma''_{ij}$$

Which can be seen to be 0 due to the fact that $2 = 0$ for $\mathbb{Z}_2$.

Bibliography

[1] Glenn T. Seaborg. THE PERIODIC TABLE: Tortuous path to man-made elements. *Chemical and Engineering News*, 57:46–52, may 1979.

[2] Eric Scerri. Looking Backwards and Forwards at the Development of the Periodic Table. *Chemistry International*, 41(1):16–20, jan 2019.

[3] Christopher Southan. Caveat Usor: Assessing Differences between Major Chemistry Databases. *ChemMedChem*, 13(6):470–481, feb 2018.

[4] H. M. Berman. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[5] Colin R. Groom, Ian J. Bruno, Matthew P. Lightfoot, and Suzanna C. Ward. The Cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 72(2):171–179, apr 2016.

[6] Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H. Bryant. PubChem Substance and Compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213, jan 2016.

[7] Anna Gaulton, Anne Hersey, Micha L. Nowotka, A. Patricia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J. Bellis, Elena Cibrian-Uhalte, Mark Davies, Nathan Dedman, Anneli Karlsson, Maia Paula Magarinos, John P. Overington, George Papadatos, Ines Smit, and Andrew R. Leach. The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954, jan 2017.

[8] Teague Sterling and John J. Irwin. ZINC 15 - Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, nov 2015.

[9] Lars Ruddigkeit, Ruud Van Deursen, Lorenz C. Blum, and Jean Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 2012.

[10] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1, aug 2014.

[11] F. J. Anscombe. Graphs in Statistical Analysis. *The American Statistician*, 27(1):17, 1973.

[12] Justin Matejka and George Fitzmaurice. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing.

[13] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, feb 2016.

[14] Larry Wasserman. Topological Data Analysis. *Annu. Rev. Stat. Appl*, 5:501–32, 2018.

[15] Alejandro Robles, Mustafa Hajij, and Paul Rosen. The Shape of an Image: A Study of Mapper on Images. *arXiv:1710.09008*, 2017.

[16] Jorge Galvez, Vincent M. Villar, María Galvez-Llompart, and Jose M. Amigo. Chemistry Explained by Topology: An Alternative Approach. *Combinatorial Chemistry & High Throughput Screening*, 14(4):279–283, nov 2011.

[17] Harry Wiener. Structural Determination of Paraffin Boiling Points. *Journal of the American Chemical Society*, 69(1):17–20, 1947.

[18] Ivan Gutman and Tamás Körtvelyesi. Wiener Indices and Molecular Surfaces. *Zeitschrift fur Naturforschung - Section A Journal of Physical Sciences*, 50(7):669–671, jul 1995.

[19] Erica Flapan and Maia. Averett. *Knots, molecules, and the universe : an introduction to topology.* American Mathematical Society (AMS), 2016.

[20] Jonathan Simon. Topological chirality of certain molecules. *Topology*, 25(2):229–235, 1986.

[21] Jean-Claude Chambron and Dennis K. Mitchell. Chemical Topology: The Ins and Outs of Molecular Structure. *Journal of Chemical Education*, 72(12):1059, dec 1995.

[22] Pawel Dabrowski-Tumanski, Andrzej Stasiak, and Joanna I. Sulkowska. In search of functional advantages of knots in proteins. *PLoS ONE*, 11(11), nov 2016.

[23] Andrew Mugler, Sander J. Tans, and Alireza Mashaghi. Circuit topology of self-interacting chains: Implications for folding and unfolding dynamics. *Physical Chemistry Chemical Physics*, 16(41):22537–22544, oct 2014.

[24] Michael R. Peterson, Imre G. Csizmadia, and Richard W. Sharpe. Topological properties of conformational potential energy surfaces. *Journal of Molecular Structure*, 94:127–135, jan 1983.

[25] 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, 1997.

[26] Jonathan P. K. Doye. Network Topology of a Potential Energy Landscape: A Static Scale-Free Network. *Physical Review Letters*, 88(23):238701, may 2002.

[27] Michelle Francl. Stretching topology. *Nature Chemistry*, 1(5):334–335, aug 2009.

[28] Sergei M Mirkin. DNA Topology: Fundamentals. Technical report, University of Illinois, 2001.

[29] 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(54), dec 2018.

[30] Ingrid Membrillo-Solis, Mariam Pirashvili, Lee Steinberg, Jacek Brodzki, and Jeremy G. Frey. Topology and geometry of molecular conformational spaces and energy landscapes. *arXiv:1907.07770*, jul 2019.

[31] Lee Steinberg, John Russo, and Jeremy Frey. A new topological descriptor for water network structure. *Journal of Cheminformatics*, 11(1), dec 2019.

[32] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, Cambridge, Cambridge, 2002.

[33] Gunnar Carlsson, Afra Zomorodian, Anne Collins, and Leonidas Guibas. Persistence barcodes for shapes. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing - SGP '04*, page 124, New York, New York, USA, 2004. ACM Press.

[34] Afra Zomorodian. *Topology For Computing*. Cambridge University Press, Cambridge, Cambridge, 2005.

[35] Vidit Nanda and Radmila Sazdanović. Simplicial Models and Topological Inference in Biological Systems. In *Discrete and Topological Models in Molecular Biology*, pages 109–141. Springer, Berlin, Heidelberg, 2014.

[36] Ann E. Sizemore, Jennifer E. Phillips-Cremins, Robert Ghrist, and Danielle S. Bassett. The importance of the whole: Topological data analysis for the network neuroscientist. *Network Neuroscience*, 3(3):656–673, jan 2019.

[37] Charles G. Cullen. *Matrices and Linear Transformations*. Dover Publications, 2 edition, 1990.

[38] L. Vietoris. Uber den hoheren Zusammenhang kompakter Raume und eine Klasse von zusammenhangstreuen Abbildungen. *Mathematische Annalen*, 97(1):454–472, dec 1927.

[39] Afra Zomorodian. Fast construction of the Vietoris-Rips complex. *Computers & Graphics*, 34(3):263–271, 2010.

[40] Donald R. Sheehy. Linear-Size Approximations to the VietorisRips Filtration. *Discrete & Computational Geometry*, 49(4):778–796, jun 2013.

[41] Lee Steinberg. Rips complex persistence on hexagon, 2019.

[42] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of Persistence Diagrams *. *Discrete Comput Geom*, 37:103–120, 2007.

[43] Peter Bubenik. Statistical Topological Data Analysis using Persistence Landscapes. *Journal of Machine Learning Research*, 16:77–102, 2015.

[44] Henry Adams, Sofya Chepushtanova, Tegan Emerson, Eric Hanson, Michael Kirby, Francis Motta, Rachel Neville, Chris Peterson, Patrick Shipman, and Lori Ziegelmeier. Persistence Images: A Stable Vector Representation of Persistent Homology. *Journal of Machine Learning Research*, 18:1–35, 2017.

[45] Matteo Rucco, Rocio Gonzalez-Diaz, Maria-Jose Jimenez, Nieves Atienza, Cristina Cristalli, Enrico Concettoni, Andrea Ferrante, and Emanuela Merelli. A new topological entropy-based approach for measuring similarities among piecewise linear functions. *Signal Processing*, 134:130–138, 2017.

[46] Emanuela Merelli, Matteo Rucco, Peter Sloot, and Luca Tesei. Topological Characterization of Complex Systems: Using Persistent Entropy. *Entropy*, 17(10):6872–6892, oct 2015.

[47] Nieves Atienza, Rocio Gonzalez-Diaz, and Matteo Rucco. Persistent Entropy for Separating Topological Features from Noise in Vietoris-Rips Complexes. *Journal of Intelligent Information Systems*, 52(3):637–655, 2019.

[48] David Cohen-Steiner, Herbert Edelsbrunner, and Dmitriy Morozov. Vines and vineyards by updating persistence in linear time. In *Proceedings of the twenty-second annual symposium on Computational geometry - SCG '06*, page 119, New York, New York, USA, 2006. ACM Press.

[49] Peter Bubenik and Paweł Dłotko. A persistence landscapes toolbox for topological statistics. *Journal of Symbolic Computation*, 78:91–114, 2017.

[50] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: an extension of MNIST to handwritten letters. *arXiv:1702.05373*, feb 2017.

[51] Brittany Terese Fasy, Jisu Kim, Fabrizio Lecci, and Clément Maria. Introduction to the R package TDA. *arXiv:1411.1830*, 2014.

[52] Herbert Edelsbrunner. Smooth surfaces for multi-scale shape representation. In *Lecture Notes in Computer Science*, pages 391–412. Springer, Berlin, Heidelberg, dec 1995.

[53] Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, dec 2014.

[54] Ulrich Bauer. Ripser: a lean C++ code for the computation of Vietoris–Rips persistence barcodes. *Software available at https://github. com/Ripser/ripser*, 2017.

[55] Clément Maria, Jean-Daniel Boissonnat, Marc Glisse, and Mariette Yvinec. The Gudhi Library: Simplicial Complexes and Persistent Homology. In *International Congress on Mathematical Software*, pages 167–174. Springer, Berlin, Heidelberg, 2014.

[56] Jean-Daniel Boissonnat, Clément Maria, and Maria Clément. The Simplex Tree: An Efficient Data Structure for General Simplicial Complexes. *Algorithmica*, 70(3):406–427, 2014.

[57] Robert Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45(01):61–76, oct 2007.

[58] Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. *Eurographics Symposium on Point-Based Graphics*, 2007.

[59] Mathieu Carrière and Steve Oudot. Structure and Stability of the One-Dimensional Mapper. *Foundations of Computational Mathematics*, pages 1–64, oct 2017.

[60] Francisco Belchi, Jacek Brodzki, Matthew Burfitt, and Mahesan Niranjan. A numerical measure of the instability of Mapper-type algorithms. *arXiv:1906.01507*, 2019.

[61] Fabio Strazzeri and Rubén J. Sánchez-García. Morse Theory and an Impossibility Theorem for Graph Clustering. *arXiv:1806.06142*, jun 2018.

[62] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, and G. Carlsson. Extracting insights from the shape of complex data using topology. *Scientific Reports*, 3(1):1236, dec 2013.

[63] Gunnar Carlsson. Topology and Data. *Bulletin of the American Mathematical Society*, 46(209):255–308, 2009.

[64] Daniel Mullner and Aravindaskshan Babu. Python Mapper: An open-source toolchain for data exploration, analysis and visualization, 2013.

[65] Hendrik Jacob van Veen and Nathaniel Saul. KeplerMapper. http://doi.org/10.5281/zenodo.1054444, jan 2019.

[66] Nathaniel Saul and Chris Tralie. Scikit-TDA: Topological Data Analysis for Python, 2019.

[67] Yongjin Lee, Senja D. Barthel, Paweł Dłotko, S. Mohamad Moosavi, Kathryn Hess, and Berend Smit. Pore-geometry recognition: on the importance of quantifying similarity in nanoporous materials. *arXiv:1701.06953*, jan 2017.

[68] Yongjin Lee, Senja D. Barthel, Pawel Dlotko, S. 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, jul 2018.

[69] Abraham Gutierrez, Mickaël Buchet, and Sylvain Clair. Persistent homology to quantify the quality of surfacesupported covalent networks. *ChemPhysChem*, page cphc.201900257, 2019.

[70] Kelin Xia, Xin Feng, Yiying Tong, and Guo-Wei Wei. Persistent Homology for The Quantitative Prediction of Fullerene Stability. *Journal of Computational Chemistry*, 36(6):408–422, dec 2014.

[71] Irene Donato, Matteo Gori, Marco Pettini, Giovanni Petri, Sarah De Nigris, Roberto Franzosi, and Francesco Vaccarino. Persistent homology analysis of phase transitions. *Physical Review E*, 93:052138, 2016.

[72] Cássio M.M. Pereira and Rodrigo F. de Mello. Persistent homology for time series and spatial data clustering. *Expert Systems with Applications*, 42(15-16):6026–6038, sep 2015.

[73] Jose Perea and John Harer. Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis. *Foundations of Computational Mathematics*, 15(3):799–838, jul 2015.

[74] 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):304001, 2015.

[75] Yasuaki Hiraoka, Takenobu Nakamura, Akihiko Hirata, Emerson G Escolar, Kaname Matsue, Yasumasa Nishiura, and Giorgio Parisi. Hierarchical structures

of amorphous solids characterized by persistent homology. *Proceedings of the National Academy of Sciences of the United States of America*, 2016.

[76] Takashi Ichinomiya, Ippei Obayashi, and Yasuaki Hiraoka. Persistent homology analysis of craze formation. *Physical Review E*, 95(1):012504, jan 2017.

[77] M. Saadatfar, H. Takeuchi, V. Robins, N. Francois, and Y. Hiraoka. Pore configuration landscape of granular crystallization. *Nature Communications*, 8:15082, may 2017.

[78] Mickaël Buchet, Yasuaki Hiraoka, and Ippei Obayashi. Persistent Homology and Materials Informatics. In *Nanoinformatics*, pages 75–95. Springer Singapore, Singapore, 2018.

[79] Herbert Edelsbrunner and John Harer. Persistent Homology a Survey. In *Contemp. Math.*, pages 257–282. American Mathematical Society (AMS), 2007.

[80] Kelin Xia, Kristopher Opron, and Guo-Wei Wei. Multiscale multiphysics and multidomain modelsFlexibility and rigidity. *Journal of Chemical Physics*, 139, 2013.

[81] Kelin Xia, Zhiming Li, and Lin Mu. Multiscale Persistent Functions for Biomolecular Structure Characterization. *Bulletin of Mathematical Biology*, 80(1):1–31, jan 2018.

[82] Zixuan Cang and Guo-Wei Wei. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLOS Computational Biology*, 13(7):1–27, jul 2017.

[83] Zixuan Cang and Guo-Wei Wei. Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics*, 33(22):3549–3557, mar 2017.

[84] Zixuan Cang and Guo-Wei Wei. Topological fingerprints reveal protein-ligand binding mechanism. *arXiv:1703.10982*, mar 2017.

[85] Zixuan Cang, Lin Mu, Kedi Wu, Kristopher Opron, Kelin Xia, and Guo-Wei Wei. A topological approach for protein classification. *Mol. Based. Math. Biol.*, 3:140–162, 2015.

[86] Zixuan Cang, Lin Mu, and Guo-Wei Wei. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLOS Computational Biology*, 14(1):e1005929, jan 2018.

[87] Zixuan Cang and Guo-Wei Wei. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International Journal for Numerical Methods in Biomedical Engineering*, pages 1–17, aug 2017.

[88] Kevin Emmett, Benjamin Schweinhart, and Raul Rabadan. Multiscale Topology of Chromatin Folding. In Junichi Suzuki and Tadashi Nakano, editors, *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS)*, pages 177–180, New York City, United States, 2016. ACM.

[89] Takashi Ichinomiya, Ippei Obayashi, and Yasuaki Hiraoka. Persistent homology analysis of protein folding. *arXiv:1905.05942*, may 2019.

[90] Marcio Gameiro, Yasuaki Hiraoka, Shunsuke Izumi, Miroslav Kramar, Konstantin Mischaikow, Vidit Nanda, M Gameiro, Y Hiraoka, S Izumi, M Kramar, K Mischaikow, and V Nanda. A topological measurement of protein compressibility. *Japan J. Indust. Appl. Math.*, 32:1–17, 2015.

[91] Yuan Yao, Jian Sun, Xuhui Huang, Gregory R Bowman, Gurjeet Singh, Michael Lesnick, Leonidas J Guibas, Vijay S Pande, and Gunnar Carlsson. Topological methods for exploring low-density states in biomolecular folding pathways. *The Journal of Chemical Physics*, 130:144115, 2009.

[92] Bryn Keller, Michael Lesnick, Theodore L Willke, Suny Albany, and Ted Willke. PHoS: Persistent Homology for Virtual Screening. *chemrXiv:6969260*, 2018.

[93] Kelin Xia and Guo-Wei Wei. Multidimensional persistence in biomolecular data. *Journal of Computational Chemistry*, 36(20):1502–1520, jul 2015.

[94] Violeta Kovacev-Nikolic, Peter Bubenik, Dragan Nikolić, and Giseon Heo. Using persistent homology and dynamical distances to analyze protein binding. *Statistical Applications in Genetics and Molecular Biology*, 15(1):19–38, jan 2016.

[95] Dragan Nikolić and Violeta Kovačev-Nikolić. Dynamical persistence of active sites identified in maltose-binding protein. *Journal of Molecular Modeling*, 23(5):167, may 2017.

[96] Tamal K. Dey, Jian Sun, and Yusu Wang. Approximating loops in a shortest homology basis from point data. In *Proceedings of the 2010 annual symposium on Computational geometry - SoCG '10*, page 166, New York, New York, USA, 2010. ACM Press.

[97] Nurit Haspel, Dong Luo, and Eduardo González. Detecting intermediate protein conformations using algebraic topology. *Bioinformatics*, 18:13–15, 2017.

[98] Kelin Xia. Persistent homology analysis of ion aggregations and hydrogen-bonding networks. *Physical Chemistry Chemical Physics*, 20(19):13448–13460, may 2018.

[99] Kelin Xia, Vijay Anand, Shikhar Saxena, and Yuguang Mu. Persistent homology analysis of osmolyte molecular aggregation and their hydrogen-bonding networks. *Physical Chemistry Chemical Physics*, 2019.

[100] Zhenyu Meng, D Vijay Anand, Yunpeng Lu, Jie Wu, and Kelin Xia. Weighted persistent homology for biomolecular data analysis. *arXiv:1903.02890*, mar 2019.

[101] D Vijay Anand, Kelin Xia, and Yuguang Mu. Weighted persistent homology for osmolyte molecular aggregation and hydrogen-bonding network analysis. *arXiv:1907.06171*, 2019.

[102] Henry Adams, Atanas Atanasov, and Gunnar Carlsson. Nudged Elastic Band in Topological Data Analysis. *arXiv:1112.1993*, dec 2011.

[103] Jean-Louis Reymond, Ruud van Deursen, Lorenz C. Blum, and Lars Ruddigkeit. Chemical space as a source for new drugs. *MedChemComm*, 1(1):30, jul 2010.

[104] Christopher Lipinski and Andrew Hopkins. Navigating chemical space for biology and medicine. *Nature*, 432:855–861, 2004.

[105] Jean-Louis Reymond. The Chemical Space Project. *Acc. Chem. Res.*, 48(3):722–730, 2015.

[106] Junmei Wang, Tingjun Hou, and Xiaojie Xu. Aqueous Solubility Prediction Based on Weighted Atom Type Counts and Solvent Accessible Surface Areas. *Journal of Chemical Information and Modeling*, 49(3):571–581, 2009.

[107] John S. Delaney. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *Journal of Chemical Information and Computer Sciences*, 44(3):1000–1005, may 2004.

[108] Jarmo Huuskonen. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *Journal of Chemical Information and Modeling*, 40(3):773–777, 2000.

[109] T J Hou, K Xia, W Zhang, and X J Xu. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J. Chem. Inf. Comput. Sci*, 44:266–275, 2004.

[110] Neera Jain and Samuel H. Yalkowsky. Estimation of the aqueous solubility I: Application to organic nonelectrolytes. *Journal of Pharmaceutical Sciences*, 90(2):234–252, feb 2001.

[111] Stephen R. Heller, editor. *The Beilstein Online Database*, volume 436 of *ACS Symposium Series*. American Chemical Society, Washington, DC, aug 1990.

[112] Antonio Llinàs, Robert C. Glen, and Jonathan M. Goodman. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *Journal of Chemical Information and Modeling*, 48(7):1289–1303, jul 2008.

[113] Sheila Ash, Malcolm A. Cline, R. Webster Homer, Tad Hurst, and Gregory B. Smith. SYBYL Line Notation (SLN): A versatile language for chemical structure representation. *Journal of Chemical Information and Computer Sciences*, 37(1):71–79, 1997.

[114] Lars Ruddigkeit, Mahendra Awale, and Jean Louis Reymond. Expanding the fragrance chemical space for virtual screening. *Journal of Cheminformatics*, 2014.

[115] Mathias Dunkel, Ulrike Schmidt, Swantje Struck, Lena Berger, Bjoern Gruening, Julia Hossbach, Ines S. Jaeger, Uta Effmert, Birgit Piechulla, Roger Eriksson, Jette Knudsen, and Robert Preissner. SuperScent - A database of flavors and scents. *Nucleic Acids Research*, 2009.

[116] Antonio Zamora. An Algorithm for Finding the Smallest Set of Smallest Rings. *Journal of Chemical Information and Computer Sciences*, 1976.

[117] Greg Landrum. RDKit, 2018.

[118] Andrea Mauri, Viviana Consonni, Manuela Pavan, and Roberto Todeschini. Dragon Software: An Easy Approach to Molecular Descriptor Calculations. *Match Communications In Mathematical And In Computer Chemistry*, 56:237–248, 2006.

[119] Tony Kennedy. Managing the drug discovery/development interface. *Drug Discovery Today*, 2(10):436–444, oct 1997.

[120] Jarmo Huuskonen, Salo Marja, and Jyrki Taskinen. Aqueous Solubility Prediction of Drugs Based on Molecular Topology and Neural Network Modeling. *J. Chem. Inf. Comput. Sci*, 38(3):450–456, 1998.

[121] S H Yalkowsky and S C Valvani. Solubility and partitioning I: Solubility of nonelectrolytes in water. *Journal of pharmaceutical sciences*, 69(8):912–922, aug 1980.

[122] W L Jorgensen and E M Duffy. Prediction of drug solubility from Monte Carlo simulations. *Bioorganic & medicinal chemistry letters*, 10(11):1155–8, jun 2000.

[123] Yingqing Ran and Samuel H. Yalkowsky. Prediction of Drug Solubility by the General Solubility Equation (GSE). *Journal of Chemical Information and Computer Sciences*, 41(2):354–357, mar 2001.

[124] R. D. Wauchope, T. M. Buttler, A. G. Hornsby, P. W. M. Augustijn-Beckers, and J. P. Burt. The SCS/ARS/CES Pesticide Properties Database for Environmental Decision-Making. In *Reviews of Environmental Contamination and Toxicology*, pages 1–155. Springer, New York, NY, 1992.

[125] Rose-Marie Dannenfelser and Samuel H. Yalkowsky. Data base of aqueous solubility for organic non-electrolytes. *Science of The Total Environment*, 109-110:625–628, dec 1991.

[126] Anton J Hopfinger, Emilio Xavier Esposito, A Llinàs, R C Glen, and J M Goodman. Findings of the Challenge To Predict Aqueous Solubility. *Journal of Chemical Information and Modeling*, 49:1–5, 2009.

[127] Antonio Llinàs, Karl J. Box, Jonathan C. Burley, Robert C. Glen, Jonathan M. Goodman, IUCr, Pearson J., and Taylor R. A new method for the reproducible generation of polymorphs: two forms of sulindac with very different solubilities. *Journal of Applied Crystallography*, 40(2):379–381, apr 2007.

[128] M. Hewitt, M. T. D. Cronin, S. J. Enoch, J. C. Madden, D. W. Roberts, and J. C. Dearden. In Silico Prediction of Aqueous Solubility: The Solubility Challenge. *Journal of Chemical Information and Modeling*, 49(11):2572–2587, nov 2009.

[129] Alessandro Lusci, Gianluca Pollastri, and Pierre Baldi. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *Journal of chemical information and modeling*, 53(7):1563–75, jul 2013.

[130] David S Palmer and John B O Mitchell. Is Experimental Data Quality the Limiting Factor in Predicting the Aqueous Solubility of Druglike Molecules? *Molecular Pharmaceutics*, 11:2962–2972, 2014.

[131] Samuel Boobier, Anne Osbourn, and John B. O. Mitchell. Can human experts predict solubility better than computers? *Journal of Cheminformatics*, 9(1):63, dec 2017.

[132] Christel A.S. Bergström and Per Larsson. Computational prediction of drug solubility in water-based systems: Qualitative and quantitative approaches used in the current drug discovery and development setting. *International Journal of Pharmaceutics*, 540(1-2):185–193, apr 2018.

[133] E. Galia, E. Nicolaides, D. Hörter, R. Löbenberg, C. Reppas, and J. B. Dressman. Evaluation of Various Dissolution Media for Predicting In Vivo Performance of Class I and II Drugs. *Pharmaceutical Research*, 15(5):698–705, 1998.

[134] Antonio Llinas and Alex Avdeef. Solubility Challenge Revisited after Ten Years, with Multilab Shake-Flask Data, Using Tight (SD 0.17 log) and Loose (SD 0.62 log) Test Sets. *Journal of Chemical Information and Modeling*, 2019.

[135] David S Palmer, Antonio Llinàs, Iñaki Morao, Graeme M Day, Jonathan M Goodman, Robert C Glen, and John B O Mitchell. Predicting Intrinsic Aqueous Solubility by a Thermodynamic Cycle. *Molecular Pharmaceutics*, 5(2):266–279, 2007.

[136] Kai Luder, Lennart Lindfors, Jan Westergren, Sture Nordholm, and Roland Kjellander. In Silico Prediction of Drug Solubility. 3. Free Energy of Solvation in Pure Amorphous Matter. *J. Phys. Chem. B*, 111:7303–7311, 2007.

[137] Kai Luder, Lennart Lindfors, Jan Westergren, Sture Nordholm, and Roland Kjellander. In Silico Prediction of Drug Solubility: 2. Free Energy of Solvation in Pure Melts. *J. Phys. Chem. B*, 111:1883–1892, 2007.

[138] Jan Westergren, Lennart Lindfors, Tobias Ho, Kai Luder, Sture Nordholm, and Roland Kjellander. In Silico Prediction of Drug Solubility: 1. Free Energy of Hydration. *J. Phys. Chem. B*, 111:1872–1882, 2007.

[139] Wei Guo and Ashis G. Banerjee. Identification of key features using topological data analysis for accurate prediction of manufacturing system outputs. *Journal of Manufacturing Systems*, 43:225–234, apr 2017.

[140] Leo Carlsson, Gunnar Carlsson, and Mikael Vejdemo-Johansson. Fibres of Failure: Classifying errors in predictive processes. *arXiv:1803.00384*, 2018.

[141] David J. W. Grant and Takeru Higuchi. *Solubility Behavior of Organic Compounds*. Wiley, 1990.

[142] Samuel H. Yalkowsky, Yan He, and Parijat Jain. *Handbook of aqueous solubility data*. CRC Press, 2010.

[143] Ashutosh Kumar and Kam Y. J. Zhang. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Frontiers in Chemistry*, 6:315, jul 2018.

[144] Randy J Zauhar, Guillermo Moyna, Lifeng Tian, Zhijian Li, and William J Welsh. Shape Signatures: A New Approach to Computer-Aided Ligand-and Receptor-Based Drug Design. *J. Med. Chem.*, 46:5674–5690, 2003.

[145] Sandhya Kortagere, Matthew D Krasowski, and Sean Ekins. The importance of discerning shape in molecular pharmacology. *Trends in pharmacological sciences*, 30(3):138–47, mar 2009.

[146] James A Haigh, Barry T Pickup, J Andrew Grant, and Anthony Nicholls. Small Molecule Shape-Fingerprints. *J. Chem. Inf. Model.*, 45:673–684, 2005.

[147] Jeremy L Jenkins, Meir Glick, and John W Davies. A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes. *J. Med. Chem.*, 47:6144–6159, 2004.

[148] Thomas S Rush III, J Andrew Grant, Lidia Mosyak, and Anthony Nicholls. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. *J. Med. Chem.*, 48:1489–1495, 2005.

[149] Pedro J Ballester and W. Graham Richards. Ultrafast shape recognition for similarity search in molecular databases. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463(2081):1307–1321, may 2007.

[150] Thomas A. Halgren. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5-6):490–519, apr 1996.

[151] P. W. Fowler and D. E. Manolopoulos. *An atlas of fullerenes*. Dover Publications, 2006.

[152] David Bramer and Guo-Wei Wei. Atom-specific persistent homology and its application to protein flexibility analysis. *arXiv:1903.11037*, 2019.

[153] Jean-Louis Reymond and Mahendra Awale. Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database. *ACS Chem. Neurosci*, 3:649–657, 2012.

[154] H C Longuet-Higgins. The symmetry groups of non-rigid molecules. *Molecular Physics*, 6(5):445–460, 1963.

[155] Evangelos A. Coutsias and Michael J. Wester. RMSD and Symmetry. *Journal of Computational Chemistry*, 40(15):1496–1508, jun 2019.

[156] Timothy F Havel. Distance Geometry: Theory, Algorithms, and Chemical Applications. *Encyclopedia of Computational Chemistry*, 1998.

[157] Josep M. Porta, Lluís Ros, Federico Thomas, Francesc Corcho, Josep Cantó, and Juan Jesús Pérez. Complete maps of molecular-loop conformational spaces. *Journal of Computational Chemistry*, 28(13):2170–2189, oct 2007.

[158] Jean-Paul Ebejer, Garrett M Morris, and Charlotte M Deane. Freely Available Conformer Generation Methods: How Good Are They? *Journal of Chemical Information and Modeling*, 52:1146–1158, 2012.

[159] Patrick J. Ropp, Jacob O. Spiegel, Jennifer L. Walker, Harrison Green, Guillermo A. Morales, Katherine A. Milliken, John J. Ringe, and Jacob D. Durrant. Gypsum-DL: an open-source program for preparing small-molecule libraries for structure-based virtual screening. *Journal of Cheminformatics*, 11(1), dec 2019.

[160] Imre Jákli, Svend J. Knak Jensen, Imre G. Csizmadia, and András Perczel. Variation of conformational properties at a glance. True graphical visualization of the Ramachandran surface topology as a periodic potential energy surface. *Chemical Physics Letters*, 547:82–88, sep 2012.

[161] Stephanie R. Hare, Lars A. Bratholm, David R. Glowacki, and Barry K. Carpenter. Low dimensional representations along intrinsic reaction coordinates and molecular dynamics trajectories using interatomic distance matrices. *Chemical Science*, 2019.

[162] Behrooz Hashemian, Daniel Millán, and Marino Arroyo. Charting molecular free-energy landscapes with an atlas of collective variables. *The Journal of Chemical Physics*, 145(17):174109, nov 2016.

[163] Gordon M. Crippen. Exploring the conformation space of cycloalkanes by linearized embedding. *Journal of Computational Chemistry*, 13(3):351–361, 1992.

[164] Josep M. Porta and Léonard Jaillet. Exploring the energy landscapes of flexible molecular loops using higher-dimensional continuation. *Journal of Computational Chemistry*, 34(3):234–244, jan 2013.

[165] Shawn Martin, Aidan Thompson, Evangelos A Coutsias, and Jean-Paul Watson. Topology of cyclo-octane energy landscape. *The Journal of Chemical Physics*, 132:234115, 2010.

[166] W Michael Brown, Shawn Martin, Sara N Pollock, Evangelos A Coutsias, and Jean-Paul Watson. Algorithmic dimensionality reduction for molecular structure analysis. *The Journal of Chemical Physics*, 129(10):234115–174109, 2008.

[167] M J Frisch, G W Trucks, H B Schlegel, G E Scuseria, M A Robb, J R Cheeseman, G Scalmani, V Barone, B Mennucci, G A Petersson, H Nakatsuji, M Caricato, X Li, H P Hratchian, A F Izmaylov, J Bloino, G Zheng, J L Sonnenberg, M Hada, M Ehara, K Toyota, R Fukuda, J Hasegawa, M Ishida, T Nakajima, Y Honda, O Kitao, H Nakai, T Vreven, J A Montgomery Jr., J E Peralta, F Ogliaro, M Bearpark, J J Heyd, E Brothers, K N Kudin, V N Staroverov, R Kobayashi, J Normand, K Raghavachari, A Rendell, J C Burant, S S Iyengar, J Tomasi, M Cossi, N Rega, J M Millam, M Klene, J E Knox, J B Cross, V Bakken, C Adamo, J Jaramillo, R Gomperts, R E Stratmann, O Yazyev, A J Austin, R Cammi, C Pomelli, J W Ochterski, R L Martin, K Morokuma, V G Zakrzewski, G A Voth, P Salvador, J J Dannenberg, S Dapprich, A D Daniels, Ö Farkas, J B Foresman, J V Ortiz, J Cioslowski, and D J Fox. Gaussian 09 Revision E.01, 2009.

[168] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785–789, 1988.

[169] Dmitriy S Chekmarev, Tateki Ishida, and Ronald M Levy. Long-Time Conformational Transitions of Alanine Dipeptide in Aqueous Solution: Continuous and Discrete-State Kinetic Models. *J. Phys. Chem. B*, 108:19487–19495, 2004.

[170] Vaclav Parchansky, Josef Kapitan, Jakub Kaminsky, Kaminsky Sebestik, and Petr Bour. Ramachandran Plot for Alanine Dipeptide as Determined from Raman Optical Activity. *J. Phys. Chem. Lett*, 17:56, 2018.

[171] Michael Feig. Is Alanine Dipeptide a Good Model for Representing the Torsional Preferences of Protein Backbones? *Journal of Chemical Theory and Computation*, 4:1555–1564, 2008.

[172] Jií Vymětal and Jií Vondrášek. Metadynamics As a Tool for Mapping the Conformational and Free-Energy Space of Peptides - The Alanine Dipeptide Case Study. *J. Phys. Chem. B*, 114:5632–5642, 2010.

[173] Rubicelia Vargas, Jorge Garza, Benjamin P Hay, and David A Dixon. Conformational Study of the Alanine Dipeptide at the MP2 and DFT Levels. *J. Phys. Chem. A*, 106:3213–3218, 2002.

[174] Behrooz Hashemian, Daniel Millán, and Marino Arroyo. Modeling and enhanced sampling of molecular systems with smooth and nonlinear data-driven collective variables. *The Journal of Chemical Physics*, 139:214101, 2013.

[175] Behrooz Hashemian and Marino Arroyo. Topological obstructions in the way of data-driven collective variables. *The Journal of Chemical Physics*, 142(4):044102, jan 2015.

[176] Behrooz Hashemian. *Machine Learning in Multiscale Modeling and Simulation of Molecular Systems*. PhD thesis, Universitat Politècnica de Catalunya, 2015.

[177] Christophe Dugave and Luc Demange. Cis-trans isomerization of organic molecules and biomolecules: Implications and applications, 2003.

[178] Nobuhiro Go and Harold A Scheraga. Ring Closure and Local Conformational Deformations of Chain Molecules1'2. *Macromolecules*, 3(2):178–187, 1970.

[179] Evangelos A. Coutsias, Chaok Seok, Michael J. Wester, and Ken A. Dill. Resultants and loop closure. *International Journal of Quantum Chemistry*, 106(1):176–189, jan 2006.

[180] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000.

[181] James B. Hendrickson. Molecular Geometry. V. Evaluation of Functions and Conformations of Medium Rings. *Journal of the American Chemical Society*, 89(26):7036–7043, 1967.

[182] D. G. Evans, J. C. A. Boeyens, and IUCr. Mapping the conformation of eight-membered rings. *Acta Crystallographica Section B*, 44(6):663–671, dec 1988.

[183] D. Cremer and J. A. Pople. General definition of ring puckering coordinates. *Journal of the American Chemical Society*, 97(6):1354–1358, mar 1975.

[184] Bernadette J Stolz, Jared Tanner, Heather A Harrington, and Vidit Nanda. Geometric anomaly detection in data. *arXiv:1908.09397*, aug 2019.

[185] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12562–6, oct 2002.

[186] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindah. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 2015.

[187] Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *arXiv:1310.0980*, 2013.

[188] Max Bonomi and Carlo Camilloni. Plumed: Cambridge Tutorial, 2015.

[189] Vojtěch Spiwok and Blanka Králová. Metadynamics in the conformational space nonlinearly dimensionally reduced by Isomap. *The Journal of Chemical Physics*, 135(22):224504, dec 2011.

[190] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226—-231, 1996.

[191] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-Based Clustering Based on Hierarchical Density Estimates. In *Lecture Notes in Computer Science*, pages 160–172. Springer, Berlin, Heidelberg, 2013.

[192] Norman L. Allinger, Young H. Yuh, and Jenn Huei Lii. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 1. *Journal of the American Chemical Society*, 111(23):8551–8566, 1989.

[193] Anders Nilsson and Lars G. M. Pettersson. The structural origin of anomalous properties of liquid water. *Nature Communications*, 6(1):8998, dec 2015.

[194] Pekka Mark and Lennart Nilsson. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A*, 105(43):9954–9960, 2001.

[195] Teresa Head-Gordon and Greg Hura. Water Structure from Scattering Experiments and Simulation. *Chem. Rev.*, 102(8):2651–2670, 2002.

[196] G Kresse and J Hafner. Ab. initio molecular dynamics for liquid metals. *Phys. Rev. B*, 47(1):558–560, 1993.

[197] D. M. Dennison. The Crystal Structure of Ice. *Physical Review*, 17(1):20–22, jan 1921.

[198] Aneesur Rahman and Frank H. Stillinger. Molecular Dynamics Study of Liquid Water. *The Journal of Chemical Physics*, 55(7):3336–3359, oct 1971.

[199] Aneesur Rahman and Frank H Stillinger. Hydrogen-Bond Patterns in Liquid Water. *Journal of the American Chemical Society*, 95(24):7943–7948, 1973.

[200] P E Mason and J W Brady. Tetrahedrality and the Relationship between Collective Structure and Radial Distribution Functions in Liquid Water. *J. Phys. Chem. B*, 111(20):5669–5679, 2007.

[201] I M Svishchev and P G Kusalik. Structure in liquid water: A study of spatial distribution functions. *The Journal of Chemical Physics*, 99(10):24516–515, 1993.

[202] P. Wernet, D. Nordlund, U. Bergmann, M. Cavalleri, M. Odelius, H Ogasawara, L. Å. Näslund, T. K. Hirsch, L. Ojamäe, P. Glatzel, L. G. M. Pettersson, and A. Nilsson. The Structure of the First Coordination Shell in Liquid Water. *Science*, 304(5673):995–999, 2004.

[203] Teresa Head-Gordon and Margaret E Johnson. Tetrahedral structure or chains for liquid water. *Proceedings of the National Academy of Sciences of the United States of America*, 103(21):7973–7, may 2006.

[204] John Russo, Kenji Akahane, and Hajime Tanaka. Water-like anomalies as a function of tetrahedrality. *Proceedings of the National Academy of Sciences of the United States of America*, 115(15), mar 2018.

[205] Barbara Logan Mooney, L.René Corrales, and Aurora E. Clark. MoleculaRnetworks: An integrated graph theoretic and data mining tool to explore solvent organization in molecular simulation. *Journal of Computational Chemistry*, 33(8):853–860, mar 2012.

[206] Abdullah Ozkanlar and Aurora E. Clark. ChemNetworks: A complex network analysis tool for chemical systems. *Journal of Computational Chemistry*, 35(6):495–505, mar 2014.

[207] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, apr 1998.

[208] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(52):926–10089, 1983.

[209] H. Berendsen, J. Grigera, and T. Straatsma. The Missing Term in Effective Pair Potentials. *J. Phys. Chem*, 91:6269–6271, 1987.

[210] Hans W. Horn, William C. Swope, Jed W. Pitera, Jeffry D. Madura, Thomas J. Dick, Greg L. Hura, and Teresa Head-Gordon. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *The Journal of Chemical Physics*, 120(20):9665–9678, may 2004.

[211] Saeed Izadi, Ramu Anandakrishnan, and Alexey V Onufriev. Building Water Models: A Different Approach. *Journal of Physical Chemistry Letters*, 5:3863–3671, 2014.

[212] Michael W. Mahoney and William L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *The Journal of Chemical Physics*, 112(20):8910, may 2000.

[213] Hiroki Nada and Jan P. J. M. van der Eerden. An intermolecular potential model for the simulation of ice and water near the melting point: A six-site model of H2O. *The Journal of Chemical Physics*, 118(16):7401, apr 2003.

[214] David A. Pearlman, David A. Case, James W. Caldwell, Wilson S. Ross, Thomas E. Cheatham, Steve DeBolt, David Ferguson, George Seibel, and Peter Kollman. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1-3):1–41, sep 1995.

[215] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, mar 1977.

[216] David A. Case, Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, dec 2005.

[217] Frank H. Stillinger and Thomas A. Weber. Computer simulation of local order in condensed phases of silicon. *Physical Review B*, 31(8):5262–5271, apr 1985.

[218] Pradeep Kumar, Giancarlo Franzese, Sergey V Buldyrev, and H Eugene Stanley. Molecular dynamics study of orientational cooperativity in water. *Phys Rev E*, 73(041505):1, 2006.

[219] Jay W. Ponder, Chuanjie Wu, Pengyu Ren, Vijay S. Pande, John D. Chodera, Michael J. Schnieders, Imran Haque, David L. Mobley, Daniel S. Lambrecht, Robert A. Distasio, Martin Head-Gordon, Gary N.I. Clark, Margaret E. Johnson, and Teresa Head-Gordon. Current status of the AMOEBA polarizable force field. *Journal of Physical Chemistry B*, 114(8):2549–2564, mar 2010.

[220] Marie L. Laury, Lee Ping Wang, Vijay S. Pande, Teresa Head-Gordon, and Jay W. Ponder. Revised Parameters for the AMOEBA Polarizable Atomic Multipole Water Model. *Journal of Physical Chemistry B*, 119(29):9423–9437, jul 2015.

[221] Timothy J. Richmond. Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect. *Journal of Molecular Biology*, 178(1):63–89, sep 1984.

[222] Herbert Edelsbrunner and Michael Kerber. Alexander Duality for Functions: the Persistent Behavior of Land and Water and Shore. *arXiv:1109.5052*, 2011.

[223] Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space . *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, nov 1901.

[224] Isabelle Guyon and André Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.