The University Of Sheffield.

The application of spectral geometry to 3D molecular shape comparison

By:

Matthew Seddon

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

This work was sponsored by

AstraZeneca

and

BBSRC
bioscience for the future

The University of Sheffield
Faculty of Social Sciences
Information School

September 2017

Acknowledgements

First and foremost, I would like to thank my supervisor Val Gillet of the University of Sheffield for her support and guidance and for giving me the opportunity to develop as a scientist and chemoinformatician. I am grateful for all her help and input over the course of the project and her support in allowing me to grow has been invaluable. In compiling this thesis, I read over some of my earliest notes and I have realised just how much I have learnt since. Special thanks must also go to Martin Packer of AstraZeneca for providing industrial supervision, whose knowledge of industrial practice and the scope of computational chemistry has given the project a practical edge. In addition, I would like to thank David Cosgrove who acted as an industrial supervisor for the first half of the project and who continued to provide support and thought provoking comments until its completion.

Next, I would like to acknowledge the past and present members of the chemoinformatics research group for their interesting discussions and companionship: Antonio de la Vega de Leon, Edmund (Ed) Duesbury, Christina Maria Founti, Gianmarco Ghiandoni, John Holliday, Lucy Mazalan, Nor S Sani, Jorge Valencia, James Wallace, and Peter Willett. Special thanks must go to Eleanor Gardiner who made my first year so welcoming.

Special thanks should also go to the helpful people at the Information School who quietly and efficiently run the department. In particular, I would like to thank Matt Jones who is the fount of all knowledge of the Information School and who somehow managed to keep me on top of my paperwork.

I would also like to thank my family for their support; my mother and father who are extremely proud, if not baffled, by what I do; my brother and sister, James and Emma, who are always there when I need them; and *mis suegros* Isidro and María Antonia who have looked after me from afar.

Finally, I would like to thank Belinda, whose love and support has kept me going for the last four years. Her relentless enthusiasm for science inspired me to start this project in the first place and her meticulous approach to science is the gold standard that I am always trying to meet. I cannot imagine life without her. *¡Muchísimas Gracias!*

I would like to dedicate this work to my late grandmother, Phyllis, who from early childhood always kept a corner for me to read and think, even if she never understood why I wanted to stay at school. I owe much to those times in her house and I know she would be proud of what I have achieved.

This work was made possible with the financial support of the BBRSC and AstraZeneca.

Abstract

Molecular shape has long been recognised as a key determinant of molecular interactions (Nicholls et al., 2010). Several methods have been developed to represent the shapes of molecules, however, their performance is inadequate for large-scale virtual screening both in effectiveness and throughput (Maggiora, Vogt, Stumpfe, & Bajorath, 2014). In general, this has been attributed to computational complexity with regards to finding the optimal alignment and bioactive conformation of the small molecules. This work addresses this problem by applying spectral geometry to three-dimensional molecular shape representation.

Spectral geometry has been developed in the field of computer vision for information retrieval of flexible three-dimensional shapes. Typical applications include identifying a shape, such as a human form, in a variety of poses. Of particular interest is the ability to produce 3D shape descriptors that are alignment invariant and capture some notion of flexibility. The main contribution of this thesis is the application of spectral geometry to the domain of 3D molecular shape and the derivation of descriptors suitable for large scale virtual screening. The spectral geometry descriptors are compared to existing shape comparison methods to evaluate their performance for virtual screening. The result is an efficient descriptor that outperforms existing descriptor methods and performs as well as a Gaussian alignment-based approach on some measures.

Contents

List of figures

List of tables

1 Introduction

The discovery of novel drugs by the pharmaceutical industry depends upon the ability to identify drug-like molecules that can be used to target key biological pathways. As such, the industry is based on an ability to identify and synthesize new molecules that have desirable biological functions (Willett, 2008). A number of strategies have been developed to discover new compounds that may eventually make it to market as approved drugs.

The development of High Throughput Screening (HTS) and combinatorial chemistry in the 1990s led to an explosion in the amount of structural and biological data relating to chemicals (Willett, 2008). HTS has allowed pharmaceutical companies to produce large screening libraries of compounds but even the largest known libraries, with approximately 10^6 compounds, are far smaller than the potential compound space of drug sized molecules, which has been estimated to have more than 10^{60} possible compounds (Baker, 2013; Erlanson, 2012). Subsequently computational models offer the opportunity to explore chemical space efficiently and cost effectively. The field that has grown up to address these practical considerations is *chemoinformatics*.

Chemoinformatics has an established history of developing molecular representations for the use of automated computational tasks from compound record retrieval to statistical models and explorations of chemical space (Willett, 2003). The most common molecule representations in use today are 2D topological representations that, while computationally efficient, were originally developed for substructure searching and perform best with structural analogues (Leach & Gillet, 2007). On the other hand, richer 3D geometry representations are subject to significant computational complexity costs that stifle industry-wide adoption.

This thesis aims to overcome the computational complexity problem that has held back wide spread use of 3D molecular shape representations through exploring a fragment-based method and by exploiting *spectral geometry*, a recent technique introduced for deformable shape retrieval in computer vision (Levy, 2006; Ovsjanikov, Bronstein, Bronstein, & Guibas, 2009; Reuter, Wolter, & Peinecke, 2006).

Chapter 2 provides an overview of the field of chemoinformatics and in particular discusses the notions of molecular representation and similarity. With this framework established, a detailed review of the literature of 3D shape comparison is presented, identifying the need for a 3D shape method that reduces the computational complexity that arises from alignment procedures and conformational flexibility.

The first attempt to manage computational complexity is to reduce the 3D shape problem to small rigid bodies known as fragments. Fragment-based drug discovery constructs ligands at the target site by connecting small active fragments. However, the activity of fragments in a 3D setting is difficult to measure. Chapter 3 explores the use of crystallographic structures of ligands bound to the same target site to elucidate bioisosteric fragments for the purposes of deriving a test set to assess a new 3D fragment similarity method. The efficacy of the data set is evaluated in the context of a 2D and 3D similarity search and the difficulties of generalising rules of bioisosteric activity for fragments using these empirical methods is discussed.

As an alternative to the fragment-based approach, Chapter 4 introduces spectral geometry as a framework for developing an alignment invariant 3D shape descriptor. Spectral geometry is a technique that has previously been developed for 3D deformable shape matching in computer vision. First the material is introduced in an intuitive way to give an idea of the underlying concepts before a more formal definition of spectral geometry is described. The spectral geometry framework takes a mesh representation of a surface as an input and produces a matrix of local geometry descriptors that describe the intrinsic geometry around a point on the surface. There are a number of forms that this local geometry descriptor may take and two are explored in this chapter.

Local geometry descriptors cannot be compared directly to give whole molecule similarity comparisons. Therefore, Chapter 5 expands upon the work of Chapter 4 by introducing the concept of the global descriptor as a transformation of the local geometry descriptor. The optimal parameters of the local geometry descriptor are then evaluated in the context of a virtual screening experiment using a global geometry descriptor called the covariance descriptor. In turn, the global descriptor is then evaluated against an established Gaussian shape comparison method and an implementation of the Ultra-Fast Shape Recognition descriptor which is alignment-free.

Chapter 6 then investigates the performance of a second global geometry descriptor that aggregates the local descriptors over the surface with respect to an independent feature codebook. The methodology, called Bag of Features, produces a significantly more compact representation of the global geometry compared to the covariance descriptor. As with Chapter 5, the optimal parameters of this descriptor are evaluated using a virtual screening framework and the final descriptors are tested against two benchmarks in a large scale virtual screening experiment.

Spectral geometry presents an opportunity to investigate the flexibility of molecules using a geometry descriptor. In principle, the spectral geometry descriptors are invariant to shape flexibility for a certain class of deformation. Chapter 7 investigates whether this class of deformation is appropriate for small molecules and presents a framework for quantitatively evaluating how the 3D descriptors deal with flexibility. This discussion concludes with a suggestion of training a conformation variation independent descriptor in a machine learning context. Finally, Chapter 8 summarises the main results of the thesis and presents suggestions for future work.

2 Molecular representation and similarity searching

2.1 Introduction

Chemists have always required ways of communicating with other chemists about the chemical substances they are working on. Historically, this has led to a myriad of different techniques from naming systems to chemical formulae and structure diagrams, each of which conveying a level of information and detail required for a specific task. For example, the information captured in the chemical name 4-(Acetylamino)phenol is more specific than its common name paracetamol but would not necessarily be understood by a patient requiring the treatment. In a similar vein, the chemical formula $C_8H_9NO_2$ lists the elements that compose the compound but gives no information about their configuration in the way that a structure diagram would (Figure 2-1).

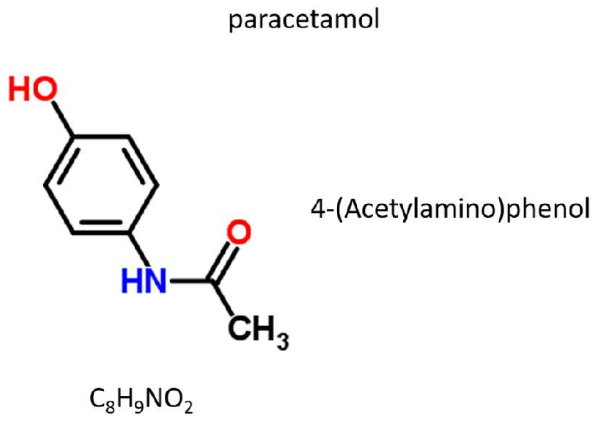


Figure 2-1. Structure diagram, chemical names, and chemical formula of paracetamol.

While the chemical representations in Figure 2-1 are readily interpretable and understood by professional chemists, there is nothing inherent in their form that makes them applicable to computational systems. Therefore, the representation of chemical compounds in a machine readable format is perhaps the most fundamental problem in chemoinformatics. The problem can be formulated as the encoding of information about a chemical compound in such a way as to facilitate the use of computational methods to perform specific tasks. Subsequently, all further chemoinformatics applications and research are founded on this representation problem. While the challenge of representing chemical information may be easy to state, the nature of chemical compounds makes the solutions far more difficult. All representations are simplifications; their use is to encode the appropriate information for the task required. For example, the very meaning of a chemical compound changes depending on the chemical model in use, from the classical Lewis Structure ball and stick model to complex quantum mechanical systems. Subsequently, it is important to remember that the choice of a particular chemical abstraction requires the choice of a particular model (James, Weininger, & Delany, 2011). Many models of chemicals have been produced, such as the valence model or quantum mechanical theory, and the choice of a representation depends on the purpose.

Early work in chemoinformatics focussed on techniques to retrieve records of chemical compounds from large industrial databases, whereas more recent applications have focussed on the use of data mining or machine learning techniques to extrapolate predictive models of molecular activity (Willett, 2003, 2008). The specification of these different tasks requires very different models of information. For the earlier task, the goal is to identify specific compounds whereas the latter task requires a framework that enables a concept of distance between two different compounds. The remainder of this chapter will review the literature of chemoinformatics paying particular attention to representation techniques for use in searching databases for similar compounds. First a review of 2D and topological methods will be presented followed by a detailed review of representations of molecular geometric information in 3D.

2.2 Machine readable formats and identifiers

The canonical representation of a molecule is a two-dimensional structure diagram. These diagrams are so pervasive in chemistry that they can be considered the lingua franca of chemists (Barnard, 2003). The information held in a two-dimensional structure diagram can be represented as a chemical graph. Chemical graphs represent molecules by defining a mathematical graph such that the vertices of the graph are atoms and the edges of the graph represent bonds connecting two atoms. Therefore, the chemical graph represents the topological configuration of inter-atomic bonding in a chemical compound. As the rules for connecting two atoms are derived from the chemical structures themselves and must follow rules of valence, a remarkable amount of information can be encoded in a chemical graph. Additional properties of the molecules, such as atomic weight, charge, and aromaticity can also be represented within this model. Concepts of mathematical graphs such as sub-graphs and graph isomorphism can then be exploited to search and analyse databases of compounds.

A simple application of chemical graphs is to determine if two graphs represent the same chemical structure. This is mathematically equivalent to finding out whether two graphs are isomorphic. A straight forward visual comparison is not feasible given the large number of compounds in a database and even experts have been shown to be biased by different representations such as orientation (Franco, Porta, Holliday, & Willett, 2014). On the other hand, a machine-based approach, while orientation invariant, must test all of the edges in the two graphs against each other. With large structures the number of comparisons grows exponentially and illustrates the combinatorial explosion that can often occur in chemoinformatics.

The most well-known algorithmic procedure to identify graph isomorphism is the Morgan Algorithm (Morgan, 1965). In this method, the “connectivity values” for each vertex in the graph are iteratively calculated. To illustrate with a chemical graph, each atom is initially given a connectivity value equal to the number of atoms it is connected to. In subsequent steps, the connectivity values of each atom are determined by summing the connectivity values of the nearest neighbours. This process continues recursively until all atoms can be maximally distinguished, that is to say that further iterations would not result in more disambiguation. A canonical numbering scheme is then implemented using the final values. For example, the atom with the highest score is listed as the first atom in the connection table, then its nearest neighbours are ordered with respect to their values, etc. Importantly, the canonical numbering scheme of any molecule is unique, thus graph isomorphism is simplified to a comparison of these numbers.

2.3 Chemical graph retrieval

The history of chemoinformatics dates back to early attempts at substructure searching that applied graph theory to the representation of molecules (Willett, 2003). Early applications of graph theory used these mathematical ideas to find chemical structures in a database that had a particular substructure in common (Ray & Kirsch, 1957). Willett (2003) regards this as the first use of modern chemoinformatics to carry out a systematic search of a database to find similar compounds.

2.3.1 Substructure search

Substructure searching is a common approach used to find compounds of interest in a database (Leach & Gillet, 2007). The approach requires finding all compounds in a database that contain a specified substructure. For each compound in the database, the goal is to identify whether the chemical graph representation entirely contains a given subgraph, which is called a sub-graph isomorphism. A sub-graph isomorphism exists if all the vertices and edges of one graph map to a subset of vertices and edges of another graph in such a way that the labels on the vertices and edges are preserved. In other words, one chemical graph is “contained” within another graph. Finding a sub-graph isomorphism is a problem that belongs to the class of NP-complete problems, which means that the worst-case time for evaluation rises exponentially with input. However, methods have been derived to reduce the average time of computation (Barnard, 1993). First is to use faster computers as the development of computer hardware has been an important factor in improving the feasibility of intensive computations, especially when coupled with programming techniques such as parallel programming. Second is to use heuristic methods to identify quickly good candidates or reject candidates that cannot be identified without exhaustive testing.

Alternatively, the database can be pre-processed in such a way to carry out many time consuming operations independent of the query structure.

One successful sub-graph isomorphism search is Ullmann's Algorithm, which uses adjacency matrices to represent the chemical graphs. An adjacency matrix for a molecule, M, is a binary matrix that represents the connectivity of the atoms in the molecule. The rows and columns correspond the vertices, in the graph and the value of the (i, j) element represents the edge connecting the to vertices i and j, with 1 representing a bond between two atoms (i, j) and 0 otherwise. Suppose that in addition to M there is a substructure query S then we define the matching matrix A as a Boolean matrix where $a_{i,j}$ takes a value of 1 if there is a match between the corresponding pair of atoms in the graphs M and S. If a sub-graph isomorphism exists between a molecule M and a sub-graph S then there is a mapping between the two graphs that satisfies the condition $A(AM)^T = S$ (Leach & Gillet, 2007).

The Ullmann Algorithm is generally implemented by combining a back-tracking technique with a relaxation procedure (Barnard, 1993). In a back-tracking technique, an arbitrary vertex is selected in the query molecule and checked for a mapping, it then proceeds to check each neighbouring vertex for a mapping and continues recursively until either the N^{th} vertex completes the sub-graph mapping and an isomorphism is found, or the mapping breaks, in which case the algorithm back-tracks to the last successful vertex and tries again. The relaxation procedure is applied each time a query vertex is selected and attempts to rearrange the matching matrix to produce a row of 0s. If such a row can be produced, then a vertex in the query molecule can have no corresponding vertices in the substructure and the algorithm may back-track without checking any further vertices in the path.

Another important idea when searching databases of chemical structures is the maximal common substructure (MCS), which generalises sub-graph isomorphism by finding the largest sub-graph in common between two chemical graphs. Efficient methods that are exact or approximate MCS searches have been developed to make large scale MCS searches feasible, for example, by using clique detection techniques (Leach & Gillet, 2007).

2.3.2 2D Fingerprints

Graph theoretic methods are computationally expensive. In order to facilitate substructure search, fingerprints were developed to provide a fast screening step before a thorough graph-based substructure search. Fingerprints are derived in a number of ways but they all perform the same task, which is to represent the molecule in a binary format in order to perform fast substructure searches. This representation is usually produced using either a fragment dictionary or a hash key, which also determines the type of information encoded (Leach & Gillet, 2007; Riniker & Landrum, 2013). A fragment dictionary fingerprint is a bitstring of 1s and 0s, where each location in the string corresponds to an entry in an external dictionary of fragments and a value of 1 indicates the fragment is present in the molecule, as illustrated in Figure 2-2 (a). Therefore, the computer can efficiently scan the fingerprint of a molecule to search for particular fragments and eliminate molecules that do not contain all of the fragments in the query molecule from further consideration. Importantly, the choice of the fragment dictionary can significantly affect performance and general fragment dictionaries are avoided in favour of those targeted for the types of molecules expected to compose a database (Leach & Gillet, 2007). A typical example of fragment dictionary fingerprints is the public Molecular ACCess System (MACCS) structural keys (*MACCS structural keys*, 2011), which consist of 166 SMARTS strings as predefined substructures forming the dictionary.

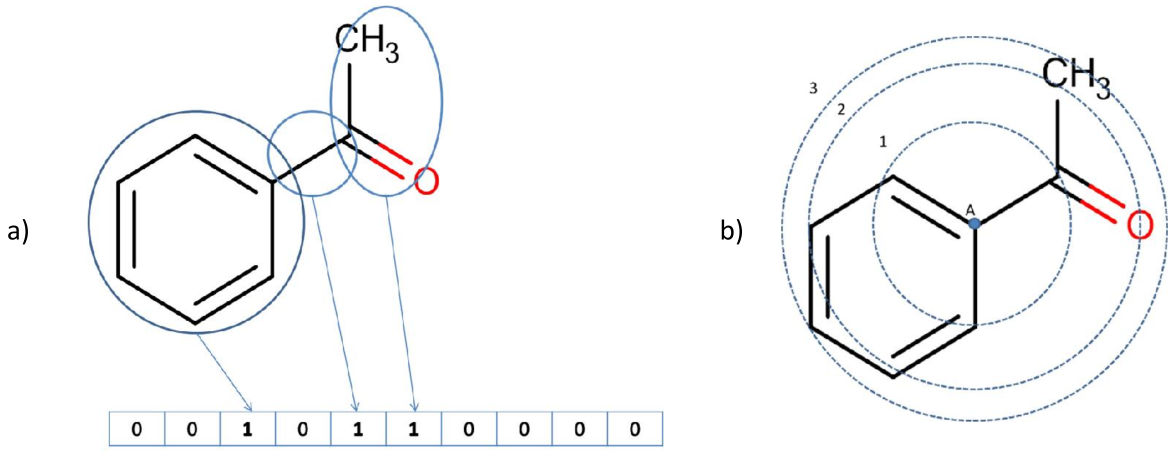


Figure 2-2. Schematic illustrations of fingerprint methods, (a) shows a dictionary fingerprint and (b) shows the rings of the Morgan algorithm from atom A.

In the case of fragment dictionaries, there is a one-to-one correspondence between a bit in the bitstring and a fragment in a molecule. The alternative approach is referred to as a hashed fingerprint or a path-based fingerprint, in which there is a many-to-one relationship between bits in the fingerprint. Path-based fingerprints encode atom types and paths between atom types. They are generated algorithmically without reference to an external dictionary. All paths up to a specified length are traced within a molecule and each path is mapped to a number of bits in the fingerprint using hashing. The hashing step can lead to the same bit being set by different paths. This means that it is not possible to map back from a bit in a fingerprint to a particular molecular fragment. Path-based fingerprints are found in the Daylight software package as well as in open source applications, such as the RDK5 fingerprints in the RDKit software (James et al., 2011; Landrum, n.d.).

2.3.3 Linear notations

An alternative approach to encode the chemical graph in a machine readable format is to use a symbolic language to encode the graph in a textual linear format. Therefore, the retrieval of a compound from a database can be carried out by a text search. The most used example is the Simplified Molecular Input Line Entry System, SMILES (Weininger, 1988). Rather than represent all aspects of the graph in a single data entry, the SMILES symbolic language has a vocabulary to represent objects of chemical information and grammatical rules for combination of terms. SMILES representations have been successful in the chemoinformatics field in part due to their compact nature; a typical SMILES representation will take 50% to 70% less space than a connection table (James et al., 2011). As a result, their use is widespread, especially for tasks such as keys for databases, input of chemical data, or exchange of chemical information. An introduction to SMILES can be found in Leach & Gillet (2007), an overview of their rules and use with the Daylight Toolkit can be found in James et al. (2011), and a detailed description of the SMILES language can be found in the original Weininger article (1988).

SMILES describes the simple structure of the molecule by enumerating the atoms as one “walks” along the structure. Rings are “broken” into linear form and notation is used to denote the start and finish of a ring cycle. Further notational conventions describe aspects of the molecule such as aliphatic atoms, which are written with upper case letters, whereas, aromatic atoms can take lower case letters. Such specification is not always necessary as there are algorithms for calculating the aromatic atoms in a SMILES string (James et al., 2011; Leach & Gillet, 2007). Single bonds are implicit, double bonds are denoted by = and triple bonds by #. As the notation is a representation of the valence model of chemistry, hydrogen atoms need not be specified unless necessary and are described implicitly using normal valence assumptions. Finally, one molecule may have many SMILES representations and so algorithms for developing canonical SMILES for a given molecule have been developed (Leach & Gillet, 2007).

In addition to SMILES, the SMARTS language has been created to represent molecular patterns to be used in substructure searching (James et al., 2011). While a SMILES string represents a molecular structure, or graph, a SMARTS string represents a molecular pattern, or sub-graph. Consequently, it can be used to represent functional groups for sub-graph searching but can also be used to represent individual features such as rotatable bonds. As SMARTS may represent any sub-graph of a molecule, all SMILES representations are SMARTS representations, although not all SMARTS representations are SMILES representations. Further characteristics that are reminiscent of regular expressions are included within the language, such as * denoting the wild card symbol that matches any atom. SMARTS also includes features for logical operators, such as ! for “not”, and & for “and”, and the ability to recursively define chemical environments with the $ symbol. Finally, whereas SMILES represents a molecule and SMARTS a pattern, there are some different semantics. In SMILES, O means an oxygen with zero charge and two hydrogens, or water, via normal valence assumptions, but the SMARTS expression O matches any aliphatic oxygen. The SMARTS expression for water would be [OH2], with all hydrogen atoms enumerated, which is also a valid SMILES expression (James et al., 2011).

2.4 A general model of similarity searching

In an influential book, Johnson and Maggiora (1990) coined the concept of the similarity property principle. This principle states that similar compounds should have similar properties; the most useful property in question for drug development is biological activity. While this appears simple enough, in practice, the problem of defining similarity is non-trivial as compounds that appear structurally dissimilar can have similar biological activity, and structures that are very similar can have vastly different activity profiles. These structure-activity relationships (SAR) are the subject of much investigation and the anomalies described above provide a discontinuity in SAR space that represents a limitation on the effectiveness of the similarity property principle (Bajorath et al., 2009; Maggiora et al., 2014; Peltason & Bajorath, 2007).

Importantly, similarity can be used to define an order relation on a database that allows the molecules in a database to be ranked by their similarity to a query molecule. Using this ranking, the similar property principle would then imply that compounds with similar SAR will be promoted to the top of the list. Searching a database in this way for similar compounds is known as a similarity searching. Unlike substructure searching, which is able to provide exact relationships between the topological structures of two molecules, molecular similarity is more subjective. For example, finding a lead compound with an improved ADMET profile requires finding a compound with a similar biological function, whereas, asking whether two compounds are similar in terms of intellectual property is defined by the rules of similarity in the law. In the first case, the similarity measure is on two compounds in biological or activity space and in the latter the measure is often in structure space. Consequently, the choice of representation plays a fundamental role in similarity analysis, as does how the features are weighted and how the similarity is calculated.

In general, there are three steps to comparing the similarity of two chemical compounds (Maggiora et al., 2014): first, a chemical compound is mapped to an appropriate representation; then weightings are applied to the representation; and finally an appropriate scoring function measures some notion of *closeness* in the descriptor space. Therefore, the similar property principle is adapted to the hypothesis that two compounds are similar if their descriptors are close in a descriptor space. The outcome of the similarity calculation is a measure of molecular similarity, typically a number in the interval [0, 1]. The rest of this section will look at three categories of descriptors (binary vectors, real valued vectors and functions), their spaces, and the appropriate measures. Relative weighting systems are rarely implemented and therefore are not discussed here.

2.4.1 Binary vectors

The fingerprint descriptors introduced for substructure search and described above are binary vectors and can be considered as sets (Maggiora & Shanmugasundaram, 2011). The values indicate the presence, or absence of features. As the values are either one or zero, the full descriptor space is a n-dimensional hypercube. Typically, their similarity is measured using set similarity scores. For example, given two fingerprints A and B, we measure similarity by comparing how many dictionary elements they have in common and normalize the measure by the total number of elements in the dictionary that are in either A or B.

Generally, set similarity can be derived from the Jaccard Index,

$$S_{Jac} = \frac{|A \cap B|}{|A \cup B|}$$

Equation 2-1

which can be represented by the following equation,

$$S_{Jac} = \frac{|A \cap B|}{|A| + |B| - \ |A \cap B|}$$

Equation 2-2

From Equation 2-2 the popular Tanimoto coefficient is derived. Thus, for two molecules, A and B, that are represented as vectors of bitstrings, **a** and **b** respectively, the intersection of the two sets can be replaced by the inner product between the two molecules and the cardinality of the individual sets can be replaced by their inner product. This derives the Tanimoto similarity,

$$S_{Tan} = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}| + |\mathbf{b}| - \boldsymbol{a} \cdot \boldsymbol{b}}$$

Equation 2-3

which can also be written with a simpler notation for bitstrings

$$S_{Tan}(A, B) = \frac{c}{(a + b - c)}$$

Equation 2-4

where *c* is the number of “on” bits in common, and *a* and *b* are the number of bits “on” bits in the bitstring representations of molecules A and B respectively.

An alternative set based similarity coefficient is the Dice coefficient, which for two bitstrings, represented as vectors, is

$$S_{Dice} = \frac{2\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}| + |\mathbf{b}|},$$

Equation 2-5

or alternatively with the simplified bitstring notation,

$$S_{Dice}(A,B) = \frac{2c}{a+b}. \qquad \text{Equation 2-6}$$

Both the Dice coefficient and the Tanimoto coefficient are known as symmetrical similarity coefficients as they give equal importance to the fragments of both molecules (Willett, Barnard, & Downs, 1998). However, there are some circumstances when it is preferable to give the different molecules a greater or lower importance. The following asymmetrical similarity coefficient, known as the Tversky coefficient, can be used in this context,

$$S_{Tv} = \frac{c}{\alpha(a-c)+\beta(b-c)+c} \qquad \text{Equation 2-7}$$

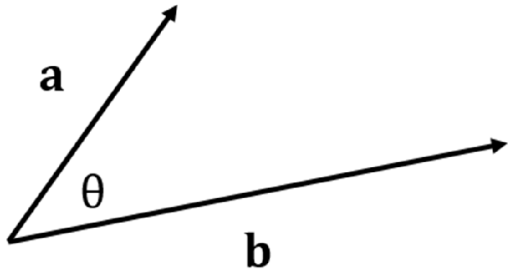
for two arbitrary constants α and β. Notice that for the values $\alpha = \beta = \frac{1}{2}$ and $\alpha = \beta = 1$ the Tversky coefficient decomposes to the Dice coefficient and the Tanimoto coefficient respectively. This is a proof that Dice and Tanimoto coefficients are monotonic for dichotomous bitstrings, which can also be shown to hold generally (Willett et al., 1998). Generally the formula is implemented with $\beta = (1 - \alpha)$ for $\alpha \in [0,1]$ (see for example, (Horvath, Marcou, & Varnek, 2013)). Varying the parameter α allows more weight to be given to smaller query molecules, for example, as larger molecules will have a higher probability of having fragments in common due to the combinatoric nature of their size. In an evaluation of similarity metrics on data derived from the ChEMBL database (Gaulton et al., 2012), a weighted Tversky search was shown to outperform Tanimoto for use in similarity based virtual screening (Horvath et al., 2013).

Several other fingerprints have been introduced for similarity searching, in addition to the fingerprints developed for substructure search. For example, circular molecular fingerprints, such as Extended Connectivity Fingerprints (ECFP) (Rogers & Hahn, 2010), use the Morgan Algorithm (Morgan, 1965) to encode circular atom environments up to a specified radius away from the central atom. The ECFP fingerprint, for example, first assigns each atom in the molecule an identifier. The algorithm then iteratively updates the identifier to take into account the nearest neighbours of the atom by collecting their identifiers, after which a hash function reduces the array back to a single integer. This process is repeated for a specified number of times and duplicate identifiers are removed. The remaining set of integer identifiers define the ECFP as illustrated in Figure 2-2 (b). This process may also be used with functional class identifiers (FCFP) to represent atoms by properties rather than elements (Rogers & Hahn, 2010). The RDKit Morgan fingerprints are an open source equivalent (Landrum, n.d.).

Recent topological fingerprints have been introduced to either improve the chemical information of the fingerprints, such as by including atom typing in traditional fingerprints (Bender, Mussa, Glen, & Reiling, 2004b, 2004a), or by adding additional 3D information (Axen et al., 2017). Bender et al. (2004b, 2004a) introduced atom typing to ECFP fingerprints by assigning SYBYL atom types (Ash, Cline, Homer, Hurst, & Smith, 1997) and then constructing count vectors for each atom type based on topological distance, which were hashed in the same way as ECFP fingerprints. In a similar fashion Axen et al. (2017) extended the Morgan algorithm to encode molecular shape with 3D environments. Instead of increasing topological rings they used spheres in 3D space that encodes the orientation and connectivity of the atom with its neighbours which is hashed and stored as a bit fingerprint.

2.4.2 Real-valued vector similarity

Alternatively, molecular descriptors are frequently represented as continuous valued vectors, for example, a vector of properties such as logP, molecular weight, molar refractivity, or structural properties such as topological indices and kappa shape indices (Leach & Gillet, 2007). Alternatively, a dimensionality reduction operation such as PCA can be used to transform high dimensional descriptors to a lower dimensional representation, which is represented as a vector.



$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos\theta$$

$$\cos\theta = \frac{\boldsymbol{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$$

Figure 2-3. A depiction of the cosine similarity between two vectors.

The descriptor space for continuous valued vectors is a vector space. Typically in chemoinformatics this space is assumed to be the n-dimensional Euclidean space with a Euclidean metric and similarity is measured in terms of cosine distance (Figure 2-3). The cosine distance,

$$S_{cos} = \frac{\mathbf{a} \cdot \mathbf{b}}{\sqrt{\|\boldsymbol{a}\| \|\boldsymbol{b}\|}}, \qquad \text{Equation 2-8}$$

where $\mathbf{a} \cdot \mathbf{b}$ is the inner, or dot, product of the two vectors that is normalised by the magnitude of the two vectors. This has an elegant geometrical interpretation as it can be represented as the cosine of the angle subtended by the two vectors (Figure 2-3). However, it is important to mention that while the vector space model of molecular descriptors is frequently used to measure molecular similarity, vector, or affine, spaces have some specific requirements that are not always properly met (Maggiora & Shanmugasundaram, 2011). For example, the vector space axioms are violated when the sum of two vectors lie in the same space, something which may not be true for property descriptors as this may produce values that are not feasible. Furthermore, the use of a Euclidean distance metric assumes that the variables are orthogonal to one another and that there exists an appropriate basis, which cannot be assumed. A full vector space model of descriptors was constructed to take advantage of vector space properties, such as using the kernel trick as a way of directly evaluating the inner-product of two vectors without explicitly enumerating their representations (Raghavendra & Maggiora, 2007). This was achieved by using an appropriate molecular basis set and was shown to demonstrate consistent similarity behaviour over chemical space.

2.4.3 Function similarity

Typically function descriptors are molecular field based, which arise from the quantum mechanics model of atoms and molecules. A number of publications have dealt with the formal derivation of quantum similarity and its applications (Bultinck, Gironés, & Carbó-Dorca, 2005; Carbó-Dorca, 2000, 2013). In this context, the molecular descriptor is electron density, denoted ρ, and the descriptor space of electron density functions is a Hilbert Space, which is a generalisation of vector spaces to encompass functions as the basic object. The electron density can be considered as the probability distribution of finding an election at a given point in space r. From a given density function at r, a similarity measure can be derived; the ground work for which was described in a seminal work by Carbó-Dorca et al. (1980). First, a quantum molecular similarity measure Z_{AB} for two densities A and B is computed that is equivalent to the inner product between two electron densities,

$$Z_{AB} = \rho_A \cdot \rho_B = \int \rho_A(r)\rho_B(r)dr. \qquad \text{Equation 2-9}$$

Second, self-similarity measures are computed to represent some notion of function magnitude,

$$Z_{AA} = \|\rho_A\|^2 = \int \rho_A(r)\rho_A(r)dr. \qquad \text{Equation 2-10}$$

Then a generalisation of cosine similarity for functions can be used to measure the similarity of two electron densities A and B, which is called the Carbo index,

$$S_{Carbo} = \frac{Z_{AB}}{\sqrt{Z_{AA}Z_{BB}}}. \qquad \text{Equation 2-11}$$

The value represents the overlap of the two molecules with respect to a relative position and orientation around an arbitrary common origin.

2.5 Virtual screening

Screening a database of chemical compounds using computational methods is a technique that has roots in the mid-1970s (Beddell, Goodford, Norrington, Wilkinson, & Wootton, 1976; Cohen, 1977). Such *in silico* methods presented the opportunity to enable a 'structure-based design' as an alternative to costly high throughput screening of empirical incremental substrate-based design (Shoichet, 2004). Structure-based methods are used when the structure of the receptor site is known and the binding mode of the structure at the site is investigated using simulations. However, these methods are time consuming for large databases of chemical compounds and therefore are unable to sample a large amount of chemical space. Additionally, the receptor sites may be too complex to simulate or the crystal structure unknown. In these situations, ligand-based virtual screening methods are used as an alternative. Ligand-based methods are composed of: similarity searches, when one or more active compound is known; pharmacophore methods, when multiple active compounds are known; and machine learning methods, when both active and inactive compounds are known (Leach & Gillet, 2007).

The most common form of similarity searching uses 2D fingerprints with a Tanimoto similarity coefficient (Maggiora et al., 2014) but these methods are restricted to information encoded in the topology of the structure diagram. In particular, these methods were developed for the purpose of substructure searching and therefore perform well at finding structural analogues but are restricted in the span of chemical space that they cover. The rest of this section will focus on the application of 3D similarity methods to virtual screening. In particular, the focus will be on 3D shape similarity.

2.6 3D shape similarity

Molecules are active in 3D, so 3D shape comparison provides the opportunity to get better returns from a similarity screen when compared to 2D methods. When compared to other *in silico* methods, such as molecular docking that simulates the binding of a ligand at a receptor site, 3D shape screening methods offer substantial improvements in computation time and do not require the structure of the receptor site. Recently, 3D shape methods have been shown to be efficient in aiding virtual screening for docking methods (L. Wang et al., 2015).

Two factors have held back the widespread adoption of 3D shape similarity methods as an alternative to 2D similarity methods: computation complexity and conformational flexibility. Unlike the topological descriptors, 3D shapes are necessarily fixed in a coordinate system meaning that in order to compare the 3D shape of two molecules, the coordinate systems of each shape must be transformed to the same space. Typically, this is performed by either aligning the two molecules in 3D space or by mapping the shape to a descriptor space for a vector space comparison, as above.

This section presents an overview of the early approaches to 3D shape comparison before identifying two strands of comparison – direct and indirect – which are reviewed separately. Next, a brief review of the related field of pharmacophore methods is presented.

2.6.1 Early approaches to 3D shape representation

The early approaches to representing 3D shapes of molecules were based on calculations of molecular volume using intersecting hard spheres (Connolly, 1985; Masek, Merchant, & Matthew, 1993). Hard spheres are atom-centred spheres with appropriate radii that represent the extent of the electron density. Molecular shape is then calculated by the intersections of the atom-centred spheres to represent the iso-surface contour of electron density of the whole molecule.

While hard sphere molecular shape representation was the dominant method of the early approaches, a number of other interesting ideas use mathematical properties of shapes in order to efficiently and accurately characterize 3D molecular shape. One such method used the idea of homology groups of algebraic topology (Arteca & Mezey, 1988; Mezey, 1987). This approach represents a volume overlap of hard spheres by computing a hierarchy of shape groups from the original surface. Rather than representing the entire surface, a purely algebraic characterization of a classification of surface points with respect to two or more intersecting spheres is all that is needed to recreate the shape.

Another early approach used Fourier descriptors to represent 3D shape (Leicester, Finney, & Bywater, 1994). Fourier descriptors represent a periodic function using a Fourier series expansion. In the case of molecular shape, the periodic function with respect to the polar coordinates may be chosen to represent the contours of the molecule around a given origin. With the selection of an appropriate basis function, a list of Fourier expansion coefficients may be used to reconstruct the molecular shape.

Sometimes it is not necessary to specify the whole shape of a molecule and it is sufficient to use only a representation of the surface. Often a ligand interacts with its target only over a region of its surface. This is equivalent to a partial shape matching problem that matches patches of the surface to the binding pocket (Finn & Morris, 2013). Two approaches that characterize the surface of the molecule are molecular skin and gnomonic projections. Molecular skin, which is a surface with a thin thickness, is approximated by a thin volume at the surface of the van der Waals volume or using a grid-based method to improve efficiency (Masek et al., 1993; Perkins, Mills, & Dean, 1995). This method may also be extended to look at hydrogen-bond donators or electrostatic potential. Alternatively, gnomonic projections, first described by Chau and Dean (1987), construct a sphere, or a convex polyhedral shape, around the molecule and project the features onto the surface from the centre of the molecule. Hence, a value is where a projection of a property cuts the surface of the volume. The molecule may then be represented as the 2D feature-space created by this projection (Leach & Gillet, 2007). There are some drawbacks with this method, for example, the molecule should be weakly convex, otherwise projections from the centre to the surface may pass through the molecule twice.

From this early body of work, two themes have emerged: alignment-dependent direct shape comparisons and alignment-independent shape descriptor comparisons. The former method directly compares molecular shape in 3D, whereas, the latter maps the shapes to a descriptor space for fast comparisons. The next two sections will look at these approaches in turn.

2.6.2 Direct shape comparisons

Grant & Pickup (1995) demonstrated that 3D molecule shape can be represented as a shape function, that defines the shape density, $H(r)$,

$$\begin{aligned} H(r) = \sum_{i} h_i(r) - \sum_{i<j} h_i(r)h_j(r) + \sum_{i<j<k} h_i(r)h_j(r)h_k(r) \\ - \sum_{i<j<k<l} h_i(r)h_j(r)h_k(r)h_l(r) + \cdots \end{aligned}$$

Equation 2-12

The first term in the expression is summed over all atoms, then all two atom overlaps are subtracted in the second term, the third term adds all three atom overlaps, the fourth subtracts all four atom overlaps and so on until the order of n atoms.

$$h_i(r) = \begin{cases} 1, & (|r - r_i| \ \leq \sigma_i) \\ 0, & otherwise \end{cases},$$

Equation 2-13

where σ_i is the van der Waals radius of atom i and r_i is the position of atom i. The volume of the molecule is then the integral of the shape function,

$$\begin{aligned} V &= \int H(r)dr \\ &= \sum_{i} v_i - \sum_{i<j} v_{ij} + \sum_{i<j<k} v_{ijk} - \sum_{i<j<k<l} v_{ijlk} + \cdots, \end{aligned}$$

Equation 2-14

where v_{ij} is the overlap volume of all two atom overlaps, in a similar fashion to Equation 2-12.

The general methodology of the direct approach to shape comparison can be summarised in two steps. First, the two molecules are aligned, or superimposed, then a scoring function uses this alignment to compute the volume overlap. A schematic example is given in Figure 2-4. Two molecules are represented by their volume, V_A and V_B respectively. In the first step, (1), the molecules are superimposed. The common volume that they both occupy is given by the intersection in green, $V_A \cap V_B$, and the total volume occupied by both molecules is the union in red, $V_A \cup V_B$. In step two, the scoring function computes their shape similarity in the most common manner, which is a volume based variant of the Tanimoto similarity, Equation 2-1.

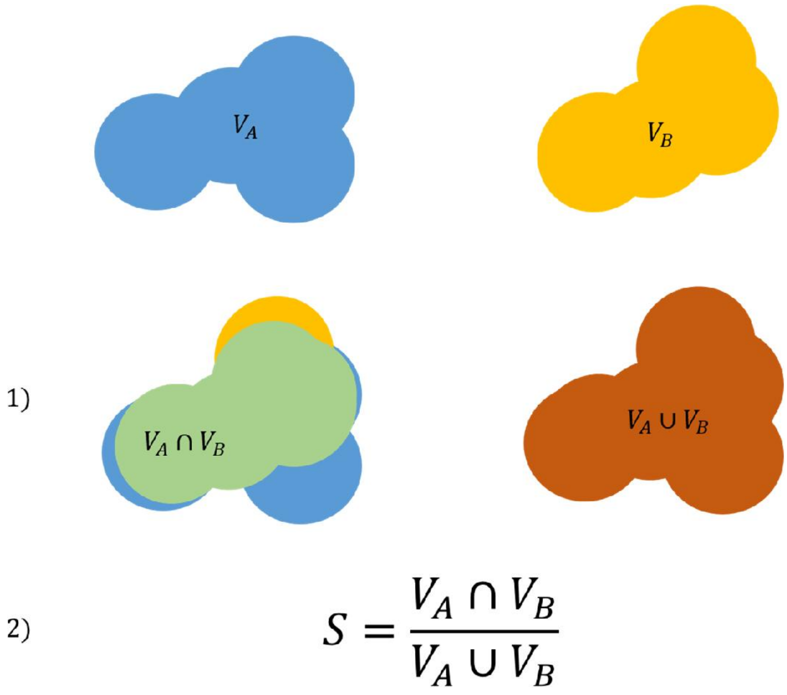


Figure 2-4. A schematic illustration of the direct 3D shape comparison method. Step (1) computes the intersection and overlap of the volumes and step (2) scores the relationship using a typical scoring function.

If $V_{AB} = V_A \cap V_B$ then Equation 2-1 can be rewritten as,

$$S = \frac{V_{AB}}{V_A + V_B - V_{AB}}, \qquad \text{Equation 2-15}$$

which is the typical scoring function used in volume shape comparison.

Unfortunately, hard sphere representations of the molecules present significant computational problems. Therefore, subsequent research has focussed on improving the methodology by introducing effective approximations. The most popular approach has been the application of

Gaussian functions to represent the shape in place of the hard sphere functions in Equation 2-12. Gaussian functions have many practical advantages: first, Gaussian functions have the attractive algebraic property that linear combinations of Gaussian functions are Gaussian functions, thus allowing simple analytical as well as efficient computational solutions to the hard sphere problem; second, the overlap of two Gaussian functions increases as their maxima approach each other (Grant, Gallardo, & Pickup, 1996; Grant & Pickup, 1995).

The most popular application of Gaussian functions is the Rapid Overlay of Chemical Structures (ROCS) program (Rush, Grant, Mosyak, & Nicholls, 2005) in which the properties of Gaussian functions are exploited to achieve a very fast measure of 3D shape similarity. The use of Gaussians in this respect means that the program is able to identify similar shapes very quickly based on three-dimensional volume overlap of optimally pre-aligned molecules. Subsequently, the measure is almost independent of atom types and bonding patterns, which enables the program to identity novel similar compounds that a simple 2D similarity search would not discover. In the ROCS software package, the Gaussian shape comparison is extended to use atom type colouring, known as ROCS *color* (*OpenEye ROCS*, n.d.). The colours can be user-specified and are used to drive the alignment step as well as the similarity score resulting in a shape comparison method that uses chemistry as well as shape.

While popular, Gaussian shape comparison programs such as ROCS have some performance issues. First, as the molecules are required to be aligned, there is a computation cost for computing this alignment. Additionally, this step typically aligns the molecules first by centre of mass meaning that the procedure does not account for size (Hamza, Wei, & Zhan, 2012). Additionally, programs often truncate Equation 2-14 to include only the first term resulting in an over-estimation of the volume overlap (Yan et al., 2013).

Recently publications have addressed these concerns to improve performance. Hamza, Wei, & Zhan (2012) introduced a new overlapping procedure where the molecules are initially overlapped by the centre of mass of a pruned candidate structure that has had some functional groups removed and then aligned using the principal moments of inertia. They introduced a new scoring function, called HWZ, after the authors surnames,

$$HWZ = \sum_{k=1}^{N} \left[a_k \left(\frac{V_{AB}}{V_A} \right)^k + b_k \left(\frac{V_{AB}}{V_B} \right)^k + c_k \left(\frac{V_{AB}}{V_A + V_B} \right)^k \right], \qquad \text{Equation 2-16}$$

where N sets of coefficients a_k, b_k, and c_k are parameters that weight the contributions of the molecules and have been trained and tested against targets in the Binding Database (X. Chen, Liu, & Gilson, 2001). The authors further enhanced the model by introducing a weighted Gaussian function that decomposes the molecule into functional groups and weights their contribution (Hamza, Wei, Hao, Xiu, & Zhan, 2013). The workflow has two steps: first optimal weights were learnt for each target using known actives, before scoring molecules of unknown activity using shape similarity. Finally, in order to obtain a similar alignment to what might be expected at the binding site, pharmacophore features are used to generate a consensus molecular shape pattern to guide target specific overlaps before the scoring function is applied (Wei & Hamza, 2014).

Yan et al. (2013) addressed the over-estimation of volume overlap in ROCS. The original papers showed that sixth-order approximations of the Gaussian system are sufficient for accurate approximations of the Hard Sphere overlap (Grant et al., 1996; 1995), however, the ROCS programme uses a highly simplified approximation method to reduce the computation complexity (Nicholls, MacCuish, & MacCuish, 2004). This is carried out by truncating Equation 2-14 to using the first term in the summation sequence. Yan et al. developed the Weighted Gaussian Algorithm (WEGA) which is designed to create a correction to this volume over-estimation. In their method, the shape density function is modified using a weighting system to reflect how closely packed the atoms are to one another.

One factor contributing to the Gaussian over-estimation of shape overlap is that the Gaussian functions are defined over the whole 3D space, meaning that there are contributions from atoms that are not in close proximity to one another. PhaseShape (Sastry, Dixon, & Sherman, 2011) addressed this by refining the superimposition step. Rather than compare volume overlap, the authors focussed on individual atom overlaps. Using pairwise atom overlaps they find atoms of similar environments using a radial distribution of distances to other atoms. The procedure identifies a triad of atoms in each molecule that are used to align the molecules in a least squares manner that superimposes common structural motifs. This procedure can incorporation additional atoms if required to produce an optimal or near optimal alignment.

Alternatively, SHaeP (Vainio, Puranen, & Johnson, 2009) uses electrostatic information to aid the molecule superimposition and similarity evaluation. The electrostatic information is incorporated by constructing a field graph. Vertices in the field graph were added around each atom using a variety of heuristics from atom hybridization to geometric properties. Each vertex is then assigned two properties: the electrostatic potential and a local shape descriptor. The local shape descriptor is a vector that represents a histogram of distances of atoms in the molecule from a plane tangent to the shape density. The vertices are then connected to form a fully connected graph whereby all vertices are connected by edges which are labelled by the Euclidean distance between the vertices. Subgraph isomorphisms between field graphs representing two molecules are then computed to identify matching points for use in a least squares superimposition. The superimposition is then scored using a weighted average of the overlap of the shape densities and the field graphs to produce a similarity score.

Similarly, SHAFTS (X. Liu, Jiang, & Li, 2011; Lu et al., 2011) uses a hybrid measure that includes shape and colour information to improve 3D shape similarity screening. The method computes a set of pharmacophores for the target molecule that are compared against a given conformation of the query. The pharmacophore was used to aid conformer selection. Using six pharmacophoric feature groups, all feature triplets are enumerated between the query and the target conformations. These triplets are then stored and used to guide the alignment with the best scoring conformation being retained for the shape comparison. Shape similarity and pharmacophore similarities are computed separately and then combined in a weighted measure of both scores. The shape score is computed using a Gaussian framework with the score being the cosine similarity of the volume overlap,

$$S = \frac{V_{AB}}{\sqrt{V_A V_B}}. \qquad \text{Equation 2-17}$$

Similarly, the pharmacophore score is computed using a cosine pharmacophore feature overlap,

$$F = \frac{F_{AB}}{\sqrt{F_A F_B}}, \qquad \text{Equation 2-18}$$

where F_{AB} is the overlap of pharmacophore feature points. The final score was calculated as,

$$SHAFTS = S + wF, \qquad \text{Equation 2-19}$$

where w is a weighting factor.

Direct shape comparisons offer an intuitive and easily interpretable notion of shape similarity. The concepts of alignment and volume overlap fit with the intuitive notion of shape similarity of a chemist. Furthermore, as the similarity is alignment based it can be directly visualised so that any similarity comparison can be recreated and inspected *post hoc* to visualise how the shapes are similar. However, direct shape comparison methods face significant problems for applications to large databases of compounds. Computational complexity is a persistent problem with direct shape overlap comparisons. Although efforts have been made to make the computations quicker, such as using GPUs to make the computation more efficient (Yan, Li, Gu, & Xu, 2014), the optimal superimposition and volume computation are computationally intensive tasks for fast screening of large databases. Additionally, direct shape comparisons are based on the notion of the molecular shape as being a rigid object. The framework within which these direct comparisons operate explicitly requires each conformation to be considered a separate shape. Therefore, the multiple conformer problem is a significant problem for direct comparison methods.

Conformational flexibility subsequently presents a further computational cost to direct shape comparison as typically samples of conformations must be computed or stored in databases to represent the conformer space of a molecule (Nicholls et al., 2004).

2.6.3 Descriptors

An alternative to the direct comparison approach is to map the 3D geometric information of molecules to a descriptor space. A molecule is then represented by a descriptor, typically a vector of values, which can then be compared against other molecules quickly and efficiently. Therefore, 3D descriptors represent an exciting opportunity to perform large-scale virtual screens on big databases. However, mapping 3D geometric features to a descriptor space is not a trivial task and the application to molecules includes some confounding problems. For example, unlike fields in which a similar operation is carried out, such as biomedical imaging models of cortical structures, there is no canonical orientation of a molecular shape. Therefore, an alignment-independent descriptor should produce the same descriptor independent of the original orientation of the molecule. This section reviews the notable efforts to develop 3D shape descriptors for molecular shape similarity.

Recently, Ultra-Fast Shape Recognition (UFSR) has emerged as the canonical alignment free 3D shape descriptor for molecules (Ballester, 2011; Ballester & Richards, 2007). The descriptor is computed by taking the statistical moments of the inter-atomic distances in a given molecule. For each molecule, the inter-atomic distances from four reference points are computed: the molecular centroid; the closest atom to that centroid; the farthest atom from the centroid; and the farthest atom from the atom farthest from the centroid. Therefore, for a given atom, the method produces four distributions that describe the 3D geometry. Rather than use histograms, which can introduce rounding errors related to choosing the bin size, the shape is described by the first three moments of each of these distributions: the mean, the variance, and the skew.

These moments are then collected into a twelve element vector and compared using the Manhattan distance.

Subsequent research in UFSR-related descriptors has focussed on introducing additional chemical information to improve performance in virtual screening. One such approach is Ultra-Fast Shape Recognition with Atom Types (UFSRAT) (Shave, 2010; Shave et al., 2015) used pharmacophore information to compute the inter-atomic distances for all atoms and an additional three pharmacophore features: hydrophobic, acceptor, and donor. First the atoms were labelled according to the pharmacophore type, then the inter-atomic distances were computed for all atoms that were labelled a specific feature. This results in a twelve element vector for each of the four distributions. The Manhattan distance-based scoring function is then computed to assess similarity. The descriptors have been implemented as part of the LIDAEUS virtual screening platform (Taylor et al., 2009). An extension of the UFSRAT methodology, Ultra-fast Shape Recognition with Credo Atom Types (Schreyer & Blundell, 2012) has also been applied to use the atom types from the Credo database (Schreyer & Blundell, 2009).

One problem with using only distances to encode geometric features is that distances do not code information on direction. Therefore, UFSR is not able to recognise chirality features of enantiomers. This was addressed in CSR, a chirality specific descriptor that adapted the UFSR method to include direction information (Armstrong, Morris, Finn, Sharma, & Richards, 2009). The authors recognised that the first method ignored direction because the Euclidean distances are based on the vector dot product. They adapted the methodology by changing the allocation of the centroids as follows: the molecular centroid, the farthest atom from the centroid, the farthest atom from the atom farthest from the centroid and a new fourth centroid that was chosen using the cross product. The cross product has the useful property that it flips the sign under reflection. Therefore, the first three centroids would give the same distance distributions for a molecule and its enantiomer but the fourth centroid would be positioned differently and therefore have a different distribution.

ElectroShape was developed in recognition that a point in 3D space with an associated charge is a 4-dimensional coordinate in the joint Euclidean-Electrostatic charge space (Armstrong et al., 2010). Therefore, the distance between any two atoms with charge is simply the 4-dimensional Euclidean distance. However, as the Euclidean space and the charge space are in different units, the distance measure has to be weighted. In order to ensure that the centroids used are not collinear, a fifth centroid is required to ensure that the points do not lie on a lower dimensional subspace. The centroids used are the first three from CSR – the molecular centroid, the farthest atom from the centroid, the farthest atom from the atom farthest from the centroid – and two additional centroids that use the cross product to ensure chirality along with the points of highest and lowest charge in the charge space. In virtual screening experiments, these descriptors doubled the enrichment factor when compared to the original UFSR descriptor. In the same vein, the method was further extended to take an additional fifth dimension into account, that of lipophilicity (Armstrong, Finn, Morris, & Richards, 2011).

An alternative way to describe the shape of a molecule adapts the Fourier descriptor methods to define the 3D shape as a function and uses the moments of that function as a descriptor. In mathematics and statistics, a moment is used to describe the shape of a function. More formally, a moment is defined as a projection of the function that describes the object as a set of characteristic functions that can be used as a basis to describe the object. The benefit of such an approach is that useful properties of the functions may be exploited to analyse the objects efficiently in a new functional space. Novotni and Klein (2003) define three desirable properties of a descriptor based on moments as invariance, orthonormality, and completeness. The invariance property allows the moments to be transformed without changing them, for example, a molecule may be rotated without changing its description. The orthonormality and completeness properties allow a set of basis functions to be combined as building blocks to form other descriptors of molecules over the domain of the function. The two most common applications of function moments to molecular descriptors are Spherical Harmonics and 3D Zernike descriptors (Kihara, Sael, Chikhi, & Esquivel-Rodriguez, 2011; Mavridis, Hudson, & Ritchie, 2007; Nisius & Gohlke, 2012; Ritchie & Kemp, 1999; Venkatraman, Sael, & Kihara, 2009).

Spherical Harmonics have been used to provide rotation invariant 3D shape descriptors for similarity searching (Mavridis et al., 2007; Peréz-Nueno et al., 2008; Pérez-Nueno, Venkatraman, Mavridis, Clark, & Ritchie, 2011; Ritchie & Kemp, 1999; Ritchie & Pérez-Nueno, 2013). Spherical Harmonics form the solutions to the Laplacian differential equation using spherical coordinates. They are used to form a complete set of orthonormal basis functions that provide a functional description of shape. The descriptor used for the molecule is the coefficients of the spherical harmonic decomposition of the shape function up to an order of L,

$$f(\theta, \phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} c_l^m Y_l^m(\theta, \phi), \qquad \text{Equation 2-20}$$

Where (θ, ϕ) are spherical coordinates, Y_l^m is the spherical harmonic of order l and degree m, and c_l^m are complex coefficients. The coefficients change with changes in coordinates, therefore techniques have been adopted to make the descriptor invariant to rotation around the z-axis by using a canonical alignment along the first principal component with the z-axis and providing the norms of the coefficients. The descriptors are computed by placing the molecule in a unit sphere and finding the optimal coefficients to fit the mesh to the Van der Waals radius at each order. The fit improves with increasing orders (Figure 2-5) and the series is truncated at a value L for the descriptor generation. Once these coefficients have been found, the underlying scaffold can be discarded because the spherical harmonic approximation of the surface can be reconstructed using these coefficients. In practice it has been shown that $L = 9$ is sufficient for mesh reconstruction (Q. Wang et al., 2011). In this respect, the spherical harmonics coefficients act as a descriptor that can be efficiently stored. Shape comparison is then carried out by finding the optimal rotation that minimises the squared distance between the two surfaces. This distance is then used to form the basis of a similarity comparison. Therefore, while spherical harmonics are descriptors in the sense that they are a compact vector of values that describe the surface of a molecule, they must still be rotated in order to compare two molecules. Consequently, they occupy a middle ground between the compact alignment invariant descriptors like UFSR and alignment-dependent shape comparisons like ROCS.

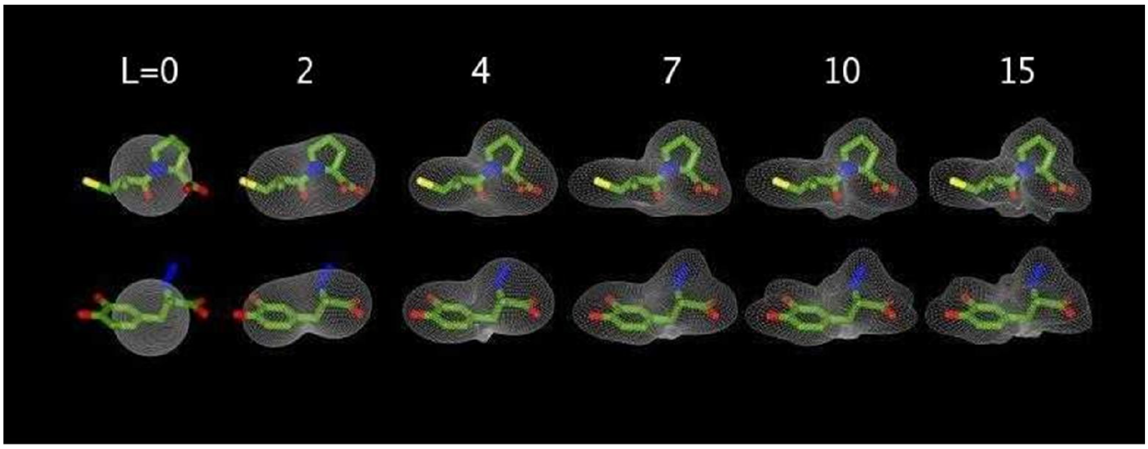


Figure 2-5. The expansion of spherical harmonic bases for an order L to approximate the 3D shape of a molecule. Image taken from (Ritchie & Pérez-Nueno, 2013).

A limitation of the spherical harmonic method is that it requires the shapes to be star-like and that the molecules are required to be placed in a common frame of reference (Venkatraman, Sael, et al., 2009). An alternative formulation used Zernike functions to encode shape in the same way without the need to align the molecules in a common framework first which performed well against a set of benchmark data sets (Venkatraman, Chakravarthy, & Kihara, 2009). The representation is derived from Equation 2-20 and includes a radial function that enables the 3D shapes to be modelled more precisely than spherical harmonics alone and also removes the requirement that the shapes are star-like. The final descriptor is a vector of coefficients that represent the moments of the Zernike representation. The comparison between two shapes can then be carried out using a suitable Euclidean vector space similarity measure.

2.6.4 Pharmacophores

In a related field, 3D pharmacophores have been developed as an approach to describing the 3D configurations of chemical features. A 3D pharmacophore is an abstract collection of chemical features and their orientations in three-dimensional space with reference to a biological target (Leach & Gillet, 2007). Pharmacophores can be seen conceptually as an extension of the concept of maximal common substructure to three-dimensional space with the addition of chemically relevant information. Such features may include hydrogen-bond donors or acceptors and thus give a model of common chemical functionality, which captures the concept of bioisosterism in conformation space (Wolber, Dornhofer, & Langer, 2006).

As mentioned in section 2.6.3, 3D vector based fingerprints have been developed to be alignment invariant. One such application is Pharmacophore Derived Queries (PDQ), which was initially developed for diversity analysis (Pickett, Mason, & McLay, 1996). The approach uses three-point pharmacophore keys that are triples of pharmacophore features with associated distances. For example, one such three-point key would be an acid 3Å from the centre of an aromatic ring and 4Å from a hydrogen-bond donor, with the remaining distance being 5Å thus denoting a 90° valence between the features. Each unique combination of six feature types combined with six distance bins gave rise to 5916 geometrically valid queries, which can be used as a dictionary for a binary bitstring. Recently mRAISE 3D pharmacophore descriptors have been developed for ligand based virtual screening (von Behren, Bietz, Nittinger, & Rarey, 2016; von Behren & Rarey, 2017). Pharmacophore features were placed on heavy atoms rather than potential hydrogen donor sites, or hydrophobic bonds. Then, a sample of conformations for each ligand is taken and indexed. Finally, the descriptors are searched to match pharmacophoric features, which provides an initial alignment invariant screen, then matching descriptors are scored using a weighted pharmacophore feature and Gaussian shape method.

2.6.5 Additional 3D shape considerations

As well as having several different low energy conformers, a molecule may also have a number of different tautomers and protonation states in solution. Tautomers refer to the intra-molecule exchange of a proton from one polar atom to another (Martin, 2009), whereas protonation states refer to the molecule gaining or losing protons from the local environment (Forli, 2015). Collectively, the different protonation and tautomeric states of a molecule are called the protonation states. The protonation state of a ligand can influence the predicted conformation, as well as the binding mode and binding affinity of ligands to proteins. For example, a change in tautomeris state can have a dramatic change on the hydrogen bond properties and tautomers that open heterocyclic rings change the shape of the molecule (Forli, 2015; Martin, 2009). This is of particular interest when investigating molecular interactions such as through simulated docking.

Therefore, one additional issue to be dealt with is how a given molecular representation handles conformation and protonation states. As the tautomers of a molecule have different molecular structures, they will be encoded differently in the bitmaps and fingerprints typically stored in a database, whereas these representations will be invariant to different conformations (Martin, 2009). With regards to 3D similarity screening and docking, the typical approach has been to take ensembles of different protonation states (Guasch et al., 2016; Park, Gao, & Stern, 2011). There are a number of different programs that enumerate tautomeric forms and protonation states, such as Protoss (Bietz, Urbaczek, Schulz, & Rarey, 2014) and UNICON (Sommer et al., 2016). Once a collection of different conformers and protonation states are sampled for a particular molecule, they can then be used as an ensemble in a similarity search or docking simulation.

2.6.6 Evaluating 3D similarity methods

As there is no commonly accepted notion of 3D shape, testing the quality of a 3D shape similarity method is not a trivial task. For example, there is no ground truth against which 3D molecular shape similarity methods can be benchmarked (Ballester & Richards, 2007). This is due to a couple of factors: first, from physical principles, molecules are fuzzy quantum systems so any 3D shape definition is an approximation; second, it is not clear in the literature how 3D shape is defined with relation to conformational flexibility. For example, on the one hand, different conformations of a single molecule are treated as different shapes, which is largely due to most of the 3D shape similarity methods being based on rigid body geometry. On the other hand, conformational variation is often given as a principal reason for the failure of 3D shape methods to perform well in virtual screening tasks due to the bioactive conformation not being included in the screen.

In the absence of a ground truth for 3D molecular shape, virtual screening experiments are used to evaluate the performance of 3D shape similarity methods. Interestingly, this does not necessarily mean that the methods are being evaluated on their ability to describe the 3D geometry of molecular shape but rather their ability to rank active molecules higher for a given target. Implicitly, the underlying belief is that a better descriptor of 3D shape would correspond to a better virtual screening performance. However, it leaves questions about the definition of 3D molecular shape unanswered.

Publically available data sets have been developed to run virtual screening experiments on 3D shape similarity. The most popular data set being the Directory of Useful Decoys (DUD) (N. Huang, Shoichet, & Irwin, 2006) and the recent development of the data set, the Directory of Useful Decoys Enhanced (DUD-E). The DUD data set was originally intended as a benchmarking data set for evaluating docking methods. The data set consists of a set of known actives for 44 targets each with a large number of structurally similar small molecules as decoys. The original purpose was to provide a benchmark for docking that did not bias large molecules, which naturally achieve high scores in docking methods (Verdonk et al., 2004). The DUD-E data set enhanced the original by increasing the number of targets to 102 and reducing the number of false decoys in the set. The data set is designed to be hard for 3D docking procedures and some of the targets are highly flexible or contain both orthosteric or allosteric binding sites. Furthermore, as the decoys are pruned using 2D similarity, the data set is inappropriate for testing 2D methods as there is a strong enrichment bias towards them.

2.6.7 Evaluation of virtual screening

Evaluating the performance of virtual screening methods is an important chemoinformatics problem related to evaluating the performance of classification models or ranking methods in machine learning (Leach & Gillet, 2007). The goal of a virtual screening experiment is to find new bioactive compounds. Therefore, similarity searching methods are evaluated retrospectively against a database of known active and inactive compounds by the ability to predict the active compounds at a given target protein given a single known active molecule. Typically, this experiment is carried out on a data set of $N = N_A + N_D$ molecules where N_A is the number of known actives at a target and N_D is the number of decoys, that is molecules that are known or presumed to be not active at the target site. The goal of a similarity search is to either rank the molecules so that the active molecules are ranked at the top of the list or to classify the molecules at a given threshold such that the molecules higher than the threshold are the active molecules.

Compounds that are correctly classified as active or decoy are known as true positives, TP, and true negatives, TN, respectively. Similarly, compounds that are incorrectly classified as active or decoy are known as false positives, FP, and false negatives, FN, respectively. Therefore, the goal of a similarity search method is to maximise the number of true positives and true negatives as well as minimising the number of false positives and false negatives. Various metrics have been created to interpret the performance of these methods in this framework.

The sensitivity and specificity of a classification method is the proportion of true positives and true negatives respectively,

$$Sensitivity = \frac{TP}{N_A},$$

Equation 2-21

and

$$Specificity \ = \frac{TN}{N_D}. \qquad \text{Equation 2-22}$$

Therefore, for a given target, a method that is sensitive is one that is good at correctly assigning high similarity values to compounds with the same activity. Whereas a method that is specific is one that is good at assigning low similarity values to compounds with different activity (Fawcett, 2006).

For evaluating ranking methods, this idea is generalised to the true positive Rate, TPR, and the false positive rate, FPR. For a given rank, $i \in N$, the true positive rate and the false positive rate are the numbers of true positives and false positives at that position,

$$TPR_i = \frac{TP_i}{N_A}, \qquad \text{Equation 2-23}$$

and

$$FPR = \frac{FP}{N_D}. \qquad \text{Equation 2-24}$$

A ROC curve is a graphical representation of the TPR and the FPR that plots the performance of the ranking method over the data set. A useful performance statistic derived from the ROC curve is the area under the curve metric (AUC) that is used to provide a value to allow comparisons across different methods on the same data set. This statistic is computed as the area under the ROC curve, hence its name, and has some attractive properties: an AUC value of 1.0 is a perfect ranking, and an AUC value of 0.5 is the expected performance of a random sampling (Fawcett, 2006).

While AUC gives a characterisation of the overall ranking of a virtual screen, it does not necessarily meet the criteria required for a chemoinformatics workflow. Typically, data sets in chemoinformatics contain a large number of compounds, only a small proportion of which are likely to be active when applied prospectively for *in silico* drug discovery. Therefore, there is an increased priority placed on the early discovery of actives (Truchon & Bayly, 2007). The enrichment factor of a virtual screen describes how many more actives are within a certain percentage of the ranked list than would have been expected from a random draw. Therefore, the enrichment factor at 1% provides the number of actives in the top 1% of the ranked list and is typically expressed as,

$$EF(X\%) = \frac{TP_x}{\frac{X}{100} \cdot N_A}. \qquad \text{Equation 2-25}$$

However, the enrichment factor has some significant weaknesses. Firstly, it is blind to the ranking within the cut-off, so that a method that has all the actives at the top will give the same enrichment factor as a method that has all the actives uniformly distributed. Secondly, the metric is highly dependent on the ratio of actives to decoys, therefore, while the measure can provide a good comparison of methods on the same target, it is unreliable when averaged over targets with different active-decoy ratios (Kirchmair, Markt, Distinto, Wolber, & Langer, 2008; Truchon & Bayly, 2007).

Subsequently, a large body of recent work has provided solutions to the early screening problem. These publications have focussed on providing an AUC-like score that is enhanced to promote early retrieval. An early method is the Robust Initial Enhancement (RIE) (Sheridan, Singh, Fluder, & Kearsley, 2001), that relates the relative rank of the i^{th} active molecule against the exponential scaled expected value of finding all the actives from a uniform distribution,

$$RIE = \frac{\sum_{i=1}^{N_A} \exp(-\alpha x_i)}{\mathbb{E}(\sum_{i=1}^{N_A} \exp(-\alpha x_i))}, \qquad \text{Equation 2-26}$$

where $x_i = r_i/N$ is the relative rank of the i^{th} active compound and $\mathbb{E}$ denotes the expected value drawn from a unform distribution. Originally the denominator was computed using a Monte Carlo simulation (Sheridan et al., 2001) but an analytical solution was provided by Truchon & Bayly (2007).

A significant disadvantage of RIE is that it is still sensitive to the ratio of actives to decoys and is dependent on a value of α that must be parameterised, which must comply with the condition $\alpha \frac{N_A}{N} \ll 1$ (Truchon & Bayly, 2007). The BEDROC measure is an adaptation of RIE that is invariant of the active-decoy ratio. The value represents the probability that a randomly selected active compound would be ranked higher than a randomly selected inactive compound from a hypothetical probability distribution that has been parameterised using an early retrieval parameter α (Truchon & Bayly, 2007). Thus, BEDROC weights the RIE measure by placing it in the hypothetical range of values,

$$BEDROC = \frac{RIE - RIE_{min}}{RIE_{max} - RIE_{min}}. \qquad \text{Equation 2-27}$$

While the metric is invariant to the relative number to actives and decoys, it is still dependent on a parameter α. The higher the value α the more strongly it weights the early retrieval (Equation 2-26). The authors recommended choosing $\alpha = 20$, which corresponds to the first 8% of the relative ranks contributing to 80% of the BEDROC value.

Clark & Webster-Clark (2008) argued that in addition to early retrieval, virtual screening metrics should also take structural diversity into account. For example, if the top active compounds are all of a similar structure then the actives are restricted to a certain class, whereas a method that produces more actives with a diversity of structures would be preferred. This is exacerbated if virtual screening models are trained on molecules from a similar structural series that happen to be active at a given target (Good & Oprea, 2008). First, they proposed an alternative to the early weighting scheme by log transforming the false positive rate,

$$pROCAUC = \frac{1}{N_A} \sum_{i=1}^{N_A} \log_{10} \frac{1}{FPR_i}. \qquad \text{Equation 2-28}$$

However, the method does not have the same interpretability as the AUC curve. The upper bound is not known and a value of 0.434 is equivalent to the 0.5 value of the AUC. Additionally, the authors augmented the ROC value so that if the molecules are from a known set of classes then the ROC curve can be weighted to maximise class diversity. The authors suggested a harmonic weighting that assigned more weight to higher ranked actives within the class so that the first occurrence of a particular class would receive more weight than the second occurrence of a different class given the same ranking.

Zhao et al. (2009) cast the previous work in a rigorous statistical framework. They used bootstrapping methods to compute theoretical null hypothesis distributions for all of the above metrics and investigated their properties. They found that if a metric is too sensitive to early recognition then it overemphasises the result and a few active compounds have a disproportionately large effect that reduces the power of the statistic to detect true early recognition. This effect was seen in BEDROC with $\alpha = 20$. On the other hand, AUC was found to be the most powerful test but as it equally weights all actives it does not reward early recognition. They concluded that the number of actives affected all null hypothesis distributions and the parameters used in one study may not be appropriate for all other studies.

Finally, the semi-log transformation of the pROC method, also known as logAUC, was generalised to produce the CROC curve (Swamidass, Azencott, Daily, & Baldi, 2010). In this method, either axis may be transformed using a functional mapping that ensures the value falls in the interval $[0,1]$. With special attention paid to the early retrieval problem, the authors chose a functional transformation that applied a relevance weighting to early actives that decayed as a function of the relative rank.

In conclusion, it is clear that there are no standards of publishing virtual screening metrics within the virtual screening literature. A review of virtual screening methods reveals that AUC is still frequently reported as the metric or that average enrichment factors over different targets are reported. Additionally, there is little publication of metrics with clear experimental frameworks, such as publishing the reference molecules that were used in the screen, or reporting metrics with error bars (Truchon & Bayly, 2007). This is likely to be due to a number of factors. Nicholls (2008) suggested that a good metric should have the following properties: independence of extensive variables; robustness; straightforward assessment of error bounds; no free parameters; and easily understood and interpretable. While enrichment factors, for example, fail the first two properties, the alternatives do not have well understood interpretations of the final values and require a more sophisticated understanding to implement and interpret. The value of an AUC of 0.8 has a well understood meaning in the literature whereas a BEDROC value of 0.3 with α of 20 does not. Furthermore, the underlying framework of the advanced measures often requires a sophisticated statistical understanding that will not necessarily be held by many non-computational chemists reading and evaluating the different virtual screening methods in order to introduce them into a full drug development workflow.

2.6.8 Limitations of 3D virtual screening

Despite their potential to describe the geometric properties of a ligand binding at the receptor, 3D methods have in general failed to replace 2D methods as the preferred descriptor type for a large scale virtual screen (Maggiora et al., 2014). This is likely to be because chemists are well trained in topological diagrams and therefore find those methods easier to understand. Additionally, the biologically active conformation of a given molecule is unknown meaning conformational variation must be taken into account somehow. Typically, this is done by using an ensemble of conformers for each molecule in a virtual screen which increases the computational complexity of the methods. Additionally, there is a trade-off between the computational cost of molecule alignment for direct comparison against a loss of geometric information when mapping to a descriptor. These reasons may explain the poor performance of 3D methods when used on standard data sets. For example, when evaluated against the DUD data set, ROCS returns an AUC < 0.5 for a number of targets, which is worse than a hypothetical random selection (Jahn, Hinselmann, Fechner, & Zell, 2009).

A number of studies have been carried out to compare the performance of the 2D and 3D approaches. Nettles et al. (2006) evaluated 2D and 3D fingerprints in a standard virtual screening task on a custom dataset. The 2D fingerprints used were MDL Structural Keys and ECFP6 and the 3D shape fingerprints were computed using pharmacophore fingerprints. They found that 2D methods performed best, especially when close structural analogues were required. However, they found that 3D methods were superior at returning diverse scaffolds and performed best below a certain similarity threshold. Later, Tresadern, Bemporad, and Howe (2009) compared the performance of 2D and 3D similarity methods on an in-house dataset obtained from a High Throughput Screen for CRF1 antagonism. They found that 3D similarity methods using ROCS performed the best and retrieved more new active scaffolds than 2D fingerprints(Tresadern et al., 2009).

Two further studies have compared 2D and 3D methods against the directory of useful decoys, DUD (Hu et al., 2012; Venkatraman, Pérez-Nueno, Mavridis, & Ritchie, 2010). In both cases the authors found that 2D methods performed significantly better than 3D methods. In particular, both methods produced similar AUC scores, yet 2D methods performed significantly better than 3D methods at early retrieval of actives. Venkatraman et al. (2010) suggested this may be due to only using a single conformer for the 3D search and (Hu et al., 2012) included a sample of conformation space to address this. However, neither paper reported the inherent enrichment bias of the DUD dataset to topological similarity, as described in 2.6.5, that makes the dataset inappropriate for evaluating a 2D topological screen.

In conclusion, while 3D methods promise to capture the shape properties crucial to the binding process, it is clear that there is room for improvement. Unless 3D methods establish demonstrable proof that they are superior to 2D methods in some use cases then it is likely that the field will remain sceptical to their promise. In practice, 3D methods are hampered by the assumption of the molecules as rigid bodies that do not capture the flexibility of conformational variation and in some cases are further restricted by the alignment requirement. Nevertheless, as 3D methods are not structure based, it has been established that they perform better at finding actives with novel scaffolds and therefore have access to a broader sample of chemical space. Therefore, rather than asserting the dominance of one method over another, it is likely that the best approach will be to combine 2D methods and 3D methods into a single workflow.

2.7 Similarity of chemical activity

Occasionally, it is necessary to consider the similarity of two compounds with respect to their activity rather than their structures. *Bioisosterism*, the term given to similarity of biological activity, has been defined as "Groups of molecules which have chemical or physical similarities producing broadly similar biological properties." (Thornber, 1979, p. 563). As novel drug compounds are typically desired to target a specific biological process, often specific proteins, similarity of biological function is determined by the chemical biology of how a compound interacts at the protein target site or the proteome in general. Therefore, rather than looking at structural or physicochemical similarities, bioisosterism relates to a pairwise comparison of "biological signatures" and how these affect activity profiles (Maggiora et al., 2014; Petrone et al., 2012).

2.8 Fragment-based drug discovery

As mentioned in the introduction, chemical space is vast and searching chemical space to identify drug compounds is an enormously difficult task. However, the size of potential compound space falls exponentially with a decrease in molecule size, therefore it has been suggested that a more efficient approach to drug discovery would be to screen collections of small molecules, or fragments, and combine them to produce novel applications (Erlanson, 2012; Erlanson, McDowell, & O'Brien, 2004). Additionally, as the number of atoms in a molecule falls, a higher proportion of the atoms in the fragment are likely to be directly involved in binding with the target, which should improve binding efficiency (Rees, Congreve, Murray, & Carr, 2004). The identification of small active molecules and their development into drug compounds is called fragment-based drug design (FBDD).

A recent review from AstraZeneca noted that the disadvantage of experimental screening of fragments is that the fragments themselves often bind to the target with a low affinity, meaning that physical screening of fragment libraries needs high concentrations and requires large amounts of materials (Joseph-McCarthy, Campbell, Kern, & Moustakas, 2014). The authors then proposed that computational approaches were of particular importance in building models that integrate data from different sources, such as biophysical, biostructural, and biochemical approaches in lead identification (Joseph-McCarthy et al., 2014, p. 693). The remainder of this literature review will cover developments in chemoinformatics methods for fragment-based drug discovery.

2.8.1 Fragment generation

Fragment generation is the automatic partition of a ligand into substructures to form fragments. Many fragment generation methods have been described in the literature and have been recently reviewed (Boyd, Turnbull, & Walse, 2012; Lounkine, Batista, & Bajorath, 2008; Sheng & Zhang, 2013). In general, fragment generation methods can be classified into four categories: substructure methods, hierarchical methods, retrosynthetic methods, and stochastic methods. Each of which will be reviewed in the remainder of this section.

Substructure methods are applied to generate fragments when there is no explicit requirement for the use of a chemical basis for the fragmentation. Kennewell et al. (2006) created an overlapping set of fragments per molecule in order to identify bioisosteric fragment pairs. All single bonds were broken unless they were ring bonds or terminal bonds. The method was applied recursively so that after each fragmentation, the resulting fragments were in turn fragmented using the same rules. Finally, the fragment set was filtered to remove single atoms.

Hierarchical methods of fragmentation are derived from early work on scaffold hopping methods by Bemis and Murcko (1996, 1999). They defined a hierarchy of ring systems, linkers, and side chains that characterize the 2D graphs of drug-like molecules and whose union is the framework of the molecule. Using these definitions, they decomposed the Comprehensive Medicinal Chemistry (CMC) database (Accelrys, Inc, n.d.), in order to analyse the diversity of shapes. The set of molecular fragment shapes generated could subsequently be used to assemble novel molecules to aid de novo drug design. They found that half the known drugs in the database could be classified by only 32 types of shape meaning that a diverse number of molecules with different properties, such as polarity or conformation, share the same topology.

Retrosynthetic fragmentation methods are designed to provide a chemical basis for fragment generation by identifying fragments that can be resynthesized using well known chemical reactions (Lewell, Judd, Watson, & Hann, 1998). The most commonly used methods in practice are the retrosynthetic combinatorial analysis procedure, RECAP (Lewell et al., 1998), and breaking of retrosynthetically interesting chemical substructures, BRICS (Degen, Wegscheid-Gerlach, Zaliani, & Rarey, 2008). RECAP uses 11 common chemical reactions, stored as SMARTS, to identify the bonds to fragment. In doing so, each fragment is tagged to represent the class of the bond so that *in silico* methods may be used to resynthesize the fragments using known chemistries at a later stage. BRICS updated the RECAP bond rules to incorporate medicinal chemistry concepts not covered by the original method. Degen et al. (2008) found that when compared with RECAP, BRICS was able to cleave about 10% more molecules and produced more fragments with multiple connection points, leading to greater branching possibilities. Recently Kalliokoski, Olsson, and Vulpetti (2013) used the BRICS algorithm implementation in RDKit (Landrum, n.d.) to create fragment alignments in order to identify sub-pockets of shared pharmacophore features and fragment binding to measure sub-pocket similarity

Finally, stochastic fragmentation methods have been applied by Bajorath's group to analyse similarity relationships in a manner that aims to be descriptor independent and to classify a hierarchy of fragments associated with an activity class (Lounkine et al., 2008). The random fragmentation approach uses a program called MolBlaster, whereby fragments are generated by random deletion of rows from connectivity tables of 2D hydrogen-suppressed graphs representations. The fragment populations that are generated are dependent upon two parameters, firstly the maximum number of bond deletions per step, and secondly the number of fragmentation iterations.

2.8.2 Three-dimensional fragment informatics

Thus far, the FBDD methods have regarded fragments as small 2D topological substructures. Three-dimensional fragments are expected to have different substitution vectors and should generate alternative pharmacophoric relationships when compared to two-dimensional representations (Morley et al., 2013). Within 3D space, fragment conformations can be classified as rods, discs, or balls. The shape diversity of a fragment database can be evaluated by assessing the distribution of fragment shapes between these three poles. Morley et al. used Principal Moments of Inertia to calculate the diversity of the ZINC database that had been decomposed using RECAP. They found that most fragment shapes lie on the interval between disc-like and rod-like. They suggested that combining these plots with maximum similarity plots will allow a medicinal chemist to compose a fragment library based on their needs, with more diverse libraries used for general screening and those with a greater level of internal similarity best used when key binding targets have already been identified.

A number of studies have applied fragment-based drug discovery to three-dimensional data sets. One of the early approaches, SPLICE, developed an algorithm for overlapping fragments in order to generate novel candidate active molecules (Ho & Marshall, 1993). SPLICE was adapted in BREED (Pierce, Rao, & Bemis, 2004), which later became the influence for the 3D fragment shuffling workflow (Nisius & Rester, 2009). The aim was to produce a method for incrementally constructing novel ligands using a tree search then allow for recombination of fragments using retrosynthetic principles and 3D distance measures. The method first aligned protein ligand complexes and gave a score to each atom based on the contribution to binding. Then three fragmentation schemes were applied to give a hierarchical set of fragments, the last scheme being RECAP which generated "anchors" for later recombination. Finally fragment scores were given based on the initial atom scores. Ligand design could then be carried out using this hierarchical data structure by combining fragments and optimizing using 3D distance. Other investigations include work focused on crystallographic data (Fechner & Schneider, 2006, 2007; Kennewell et al., 2006) or have used novel indexing methods to handle the combinatorial explosion (Maass, Schulz-Gasch, Stahl, & Rarey, 2007).

Finally, two approaches have used a quantum mechanical approach. The BROOD tool from OpenEye ('BROOD', 2006) evaluates a query fragment against a database of fragments with regards to 3D shape, electrostatics and available chemical descriptors. Shape comparisons are made using Gaussian overlays and appropriate replacements are suggested. Another interesting approach, ParaFrag, has used spherical harmonics to produce atom-independent shape descriptors, which are then combined with field calculations to derive topology independent quantum mechanical property descriptors of fragments (Jakobi, Mauser, & Clark, 2008).

2.9 Conclusions

In conclusion, there is a strong history of using chemoinformatics methods to aid the drug development process. In particular, methods have been developed to represent chemical compounds for use in computational applications. In general, these representations in conjunction with an appropriate similarity measurement can be used to search a database to find compounds that have similar representations. Guided by the similar property principle, these similarity searches can be used to prospectively find bioactive compounds when compared to a known active. While the most common application of similarity searching is to use binary fingerprints with Tanimoto similarity, 3D shape methods represent an opportunity to improve on 2D methods for two reasons. The first is that 2D methods encode topological features of chemical graphs and were originally developed for the purpose of substructure searching. Subsequently they are restricted to structural analogues in terms of similarity. Furthermore, 3D methods encode the geometry in a way that is representative of the ligand at the receptor site. Nevertheless, wide spread adoption of 3D shape methods has been hampered by computational complexity and conformational flexibility. Therefore, there is a need for further research in 3D molecular shape descriptors. In particular, there is a gap in the literature for a compact 3D shape descriptor that is computationally efficient, alignment invariant, and is aware of conformational flexibility, research into which forms the core of this thesis.

3 Generation of test set for fragment-based similarity

3.1 Introduction

Although similarity measures are now well established for comparing whole molecules, there has been less progress in the development of methods that are suitable for comparing fragments. The typical two-dimensional fingerprint methods are biased towards finding fragments that are structurally similar. Furthermore, the standard similarity coefficients such as the Tanimoto coefficient do not perform well for small molecules due to the size bias that is inherent in the method. Three-dimensional similarity methods, on the other hand, offer an opportunity to find compounds that have similar activities but are structurally diverse. However, one obstacle to developing effective fragment-based methods is the lack of established data sets for evaluating the methods.

The aim of fragment-based similarity searching is generally to identify bioisosteric fragments, that is, pairs of fragments that can be exchanged in order to alter some of the properties of a molecule, such as its solubility or metabolic stability, while maintaining its activity. A number of data sources of bioisosteric pairs have been compiled from the literature. These include the manually compiled BIOSTER database (Ujváry, 1997; Ujváry & Hayward, 2012) and the SwissBioisostere database (Wirth, Zoete, Michielin, & Sauer, 2012) which was constructed by identifying the matched molecular pairs (MMPs) in experimental assay data in ChEMBL. An alternative approach was developed by Kennewell et al. (2006) who identified target specific bioisosteres based on crystallographic data. The method finds pairs of fragments of ligands that are active in the same parts of a target site and, assuming that they have an equivalent role in biological activity, identifies them as bioisosteres. A more recent approach based on crystallographic data is the sc-PDB-Frag database which combines the structural similarity of fragments with their interaction fingerprints that describe the interactions the fragments make with the protein. A bioisostere is defined if a fragment pair has a low structural similarity and a high interaction pattern similarity (Desaphy & Rognan, 2014).

Kennewell's method is attractive for developing a set of bioisosteric pairs that could be used as a test set to evaluate a three-dimensional similarity method. This is because the bioisosteres are defined based on their shape similarity and their common location in a target binding site. Since Kennewell's work, the number of structures in the Protein Data Bank (PDB) has increased significantly, thus providing a much greater number of ligands and targets that can be analysed for bioisosteric pairs and, therefore, the potential to generate a richer test set.

The aims of this chapter are to identify bioisosteric pairs from a larger set of crystallographic data than has been used previously, based on the Kennewell methodology. The analysis is applied to a data set of high quality crystallographic ligand overlays developed for pharmacophore validation, referred to as the pharmacophore validation data set herein, (Giangreco, Cosgrove, & Packer, 2013). Kennewell's method was limited to finding target specific bioisosteres, however, given the much larger number of targets, the extent to which the approach can be used to identify bioisosteres that can be generalised across targets is also investigated. Overall, the principal aim of this chapter is to use these target-specific methods over a large number of targets to collect empirical observations of bioisosteric fragments that occur frequently enough to be general bioisosteres. If enough of these bioisosteric pairs are found, then they would form the basis of a novel 3D fragment bioisosteric test set.

Section 3.2 provides an overview of the method previously published by Kennewell et al. (2006). Then details of the pharmacophore validation data set used to derive bioisosteres are presented before a discussion on the types of fragmentation method considered here. Finally, a new approach adapted from the Kennewell method is presented along with details on the scoring function used to identify bioisosteric pairs. Section 3.3 presents the results of applying the methodology to the pharmacophore validation data set. A 2D similarity search method is then evaluated on its effectiveness to identify bioisosteric pairs, first using a similarity threshold and then using a ranking test. These results are subsequently compared to a 3D similarity search. Section 3.4 discusses the results highlighting key themes, including the definition of bioisosterism, the performance of the 2D and 3D searches, and issues that arise from the choice of data.

3.2 Methods

Crystallographic-based approaches to finding bioisosteric fragments provide the potential to derive precise experimentally validated bioisosteric fragment pairs. The driving assumption of this approach is that two fragments located at the same place in the binding site will have the same function. Subsequently, the manner in which fragments are identified from individual ligands, the fragmentation process, and the process by which they are identified as being in the same position, the scoring function, will determine whether two fragments are deemed to be bioisosteric.

Whereas potential bioisosteric descriptors have traditionally been evaluated on curated databases, it was thought that the pharmacophore validation data set presented an opportunity to derive a bioisostere validation set from known crystallographic structure data. In this way, a set of bioisosteres could be obtained that had experimental validation and at the same time would be able to give novel fragment pairs that may not be in a curated database, thus giving a robust validation set for exploring novel fragment space.

Once pairs of bioisosteric fragments have been obtained from a target site, they carry additional, target-specific, information that can be exploited to obtain a stronger model of bioisosterism. For example, each pair comes with target site location information as well as target information. Further work was undertaken to try and use these different sources of information to build a greater insight into bioisosterism. Firstly, at the target site pairs can be grouped into sets of bioisosteric pairs that may be interchangeable. An algorithm was developed to find groups of pairs in the target site and its effect was investigated. Secondly, target information may be used to find bioisosteric pairs that are common across targets. Ultimately, bioisosteric pairs that are active in a single target may not be useful outside of that small domain, so an ability to find general bioisosteres across targets has a practical application which, in turn, could be adapted to produce a set of general bioisosteres appropriate for testing novel bioisosteric similarity methods. Therefore, the extent to which bioisosteric pairs can be generalised across targets was also investigated.

3.2.1 Overview of previously published method

An overview of the Kennewell et al. (2006) method is provided here as a reference for the rest of the chapter. As mentioned earlier, the authors set out to identify target-specific bioisosteres using crystallographic data. Given a set of protein-ligand complexes representing a single target, the complexes were superimposed based on the protein binding site and the superimposed ligands were extracted. The ligands were then fragmented and the fragments compared by their volume overlap. More specifically, each ligand in the data set was made the reference ligand in turn and compared against all other ligands in the set. Each reference ligand was split into what the authors called "sections", which are non-overlapping fragments of the ligand. In contrast, each query ligand was fragmented into a set of overlapping fragments. Each query fragment was then compared against each section of the reference ligand by calculating the volume overlap using a simplified atom-centered Gaussian. Fragments with a high degree of overlap were assumed to have a similar role in their interactions with the target protein. Finally, the set of overlaps were filtered using some chemical criteria for bioisosteric pairs. The method was tested on 3D crystallographic data for 12 targets taken from the Protein Data Bank (PDB) (Bernstein et al., 1977) with differing numbers of binding ligands and an array of structurally diverse fragment pairs was produced.

3.2.2 Choice of data

Since the original Kennewell (2006) paper was published, considerably more crystal structures have become available. Data sets such as the pharmacophore validation overlays now give the opportunity to apply the original method to a large and diverse set of targets and their binding ligands. For this project, the data were taken from the AstraZeneca molecular overlays for pharmacophore validation (Giangreco et al., 2013). This data set comprises 121 overlays of high-quality crystallographic structures publicly available to download from the Cambridge Crystallographic Data Centre ('The Cambridge Crystallographic Data Centre (CCDC)', n.d.). The data set were curated to ensure sensible charge and tautomeric states, check to allosteric binding sites, and to maximise 2D diversity. Each overlay contains a set of ligands for a particular target site that have been pre-aligned based on the protein active sites meaning that conformer generation and ligand alignment were not necessary for this protocol. The Macrophage metalloelastase (P39900) overlays were selected for development purposes as the aligned ligands had the lowest RMSD published in the original article, which would ensure close overlays in the test data. Figure 3-1 shows the overlay for P39900. The image shows all 17 ligands in their three-dimensional orientation at the binding site.

Table 3-3 gives a summary of the pharmacophore validation data set (Giangreco et al., 2013) giving the Uniprot ID of the protein along with the target name and the number of ligands in the target overlay that were used in the experiment. The table also reports the number of unique fragments and the results of the bioisosteric experiment that will be described in the following sections. The targets are ordered alphabetically by Uniprot ID. The number of ligands in each set ranges from 39 in cathepsin B (P07688) to 4 ligands in cathepsin S (P25774). It is important to highlight that the number of ligands reported in the table is not the same as the number of the ligands in each target in the data set because the RDKit chemistry package was not able to import all ligands due to strict *sanitization* procedures. Consequently, the number of ligands reported are those that were used in the experiment.

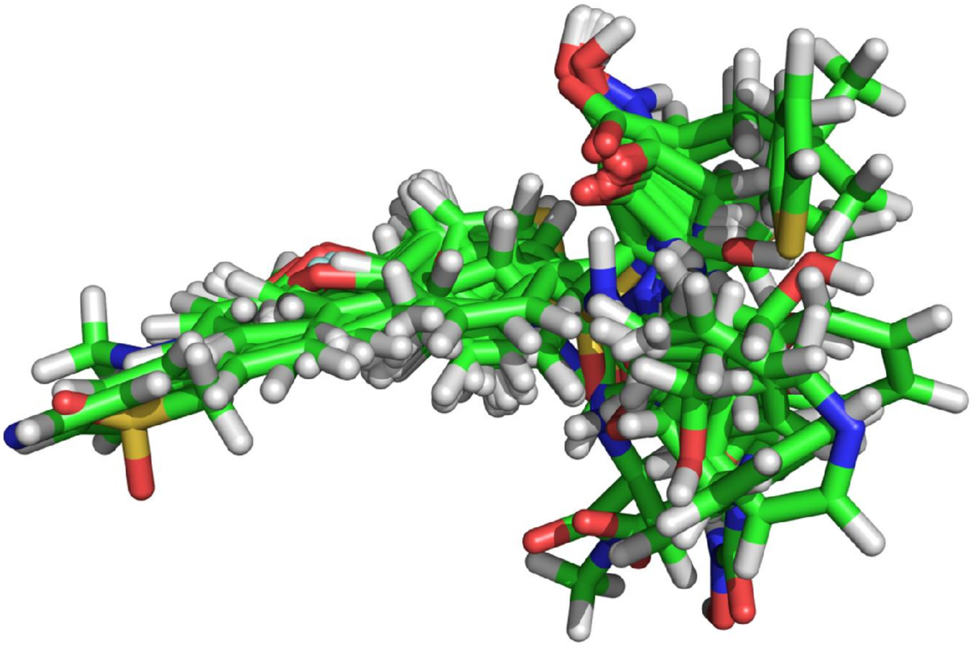


Figure 3-1. 3D visual representation of the Macrophage metalloelastase (P39900) overlay.

3.2.3 Choice of fragmentation process

Initially, three different fragmentation schemes were investigated to find the most appropriate scheme for the later volume overlay. The schemes are: an implementation of the fragmentation scheme from Kennewell et al (2006) with both overlapping and non-overlapping fragments; a modification in which fragments were limited to non-overlapping fragment sets; and the retrosynthetic fragmentation scheme BRICS (Degen et al., 2008) with non-overlapping fragments. The original method from Kennewell et al. produces a set of overlapping fragments by cutting single, non-ring, non-terminal bonds. In the implementation developed here, a SMARTS representation of rotatable bonds[1] was used to search for the bonds to be broken. The bonds were then broken to form either a set of non-overlapping or overlapping fragments. In the case of the non-overlapping fragmentation, each molecule had all the identified bonds broken simultaneously to produce a disjoint set of fragments that constituted the original ligand, an illustration of which can be seen in Figure 3-2. In order to produce a set of overlapping fragments, one bond was broken at a time to produce two fragments. The process was repeated recursively until there were no remaining bonds to break. The collection of all these fragments produced a set of all the overlapping fragments for each ligand and an illustration is also given in Figure 3-2.

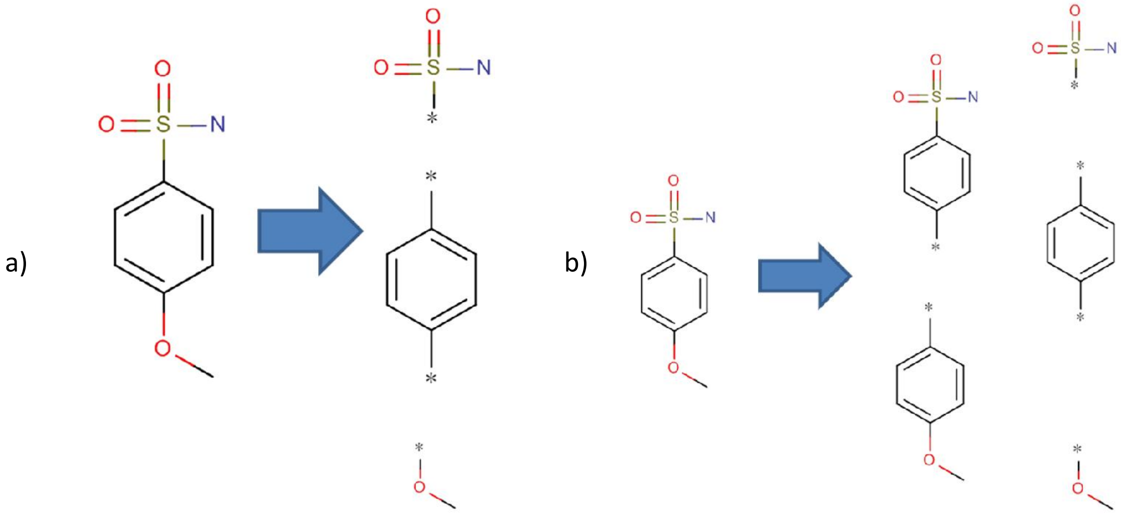


Figure 3-2. Molecule from P39900 that has been fragmented according to the Kennewell fragmentation scheme where (a) shows the non-overlapping fragments are illustrated and (b) shows the full set of overlapping fragments.

BRICS fragmentation builds on the popular RECAP fragmentation method, which uses rules that fragment a molecule based on retrosynthesis (Lewell et al., 1998), with the aim of producing synthetically relevant fragments. The BRICS fragmentation scheme uses modified rules in order to obtain a more diverse fragment space (Degen et al., 2008). For this work, the RDKit implementation of the BRICS scheme was used with the 13 rules encoded as SMARTS expressions

[1] The SMARTS used was '[!$([NH]!@C(=O))&!D1&!$(*#*)]-&!@[!$([NH]!@C(=O))&!D1&!$(*#*)]’

that fragment the molecule to form non-overlapping fragments. An illustration of BRICS fragmentation using the same molecule as Figure 3-2 is shown in Figure 3-3 where it can be seen that the BRICS scheme has resulted in two fragments, whereas the scheme from Kennewell produces either three or five fragments.

The protocol for identifying bioisosteric fragments described later in this section was developed using all fragmentation methods. The overlapping method produced a larger number of fragments, which corresponded to bioisosteric fragment pairs that were more structurally diverse than the non-overlapping pairs. However, it was thought that the fragments did not represent the notion of a fragment as used in lead optimisation. Moreover, a retrosynthetic scheme produces fragments using chemically relevant information and is thus more likely have a chemically intuitive meaning. For example, the top two fragments in Figure 3-2 (a) are not split by the BRICS fragmentation, which suggests that this would not be an optimal fragmentation method for retrosynthesis. Therefore, the non-overlapping BRICS fragmentation method was adopted for the experiments reported in this chapter.

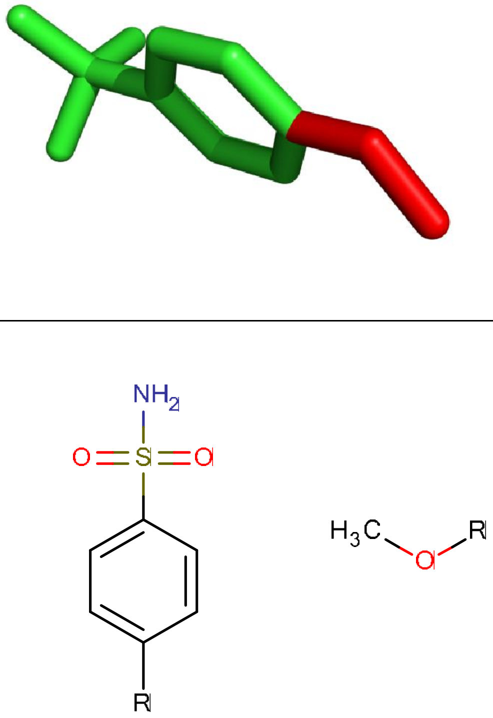


Figure 3-3. Molecule from P39900 that has been fragmented according to the BRICS fragmentation scheme. In the image, each fragment is represented by a different colour. Additionally, the non-overlapping fragments are illustrated.

3.2.4 The algorithm

An overview of the algorithm used to generate bioisosteric pairs and groups from a single file of overlaid ligands for a given target site is illustrated in Figure 3-4. Suppose a ligand overlay contains three ligands, *A, B,* and *C*, whose conformations and orientations at the active site are known. First, ligand A is taken as the reference ligand and fragmented into a set of three fragments {a1, a2, a3}. The remaining ligands in the overlay are treated as the query ligands. For example, let *B* be the first query ligand, which is fragmented to produce a set of three fragments {b1, b2, b3}. The algorithm then compares the fragments of ligand *A* with the fragments of ligand *B*. To start, all the fragments in *B*, are scored by their volume overlap with the first fragment in ligand *A*, a1. If the overlap with fragment from *B*, b1 for example, is greater than a given threshold then {a1,b1} forms a bioisosteric pair, as defined by Kennewell et al. (2006). The algorithm then compares the remaining fragments of *B* against *A* and then moves on to the fragments of *C*. In Figure 3-4 the algorithm has already moved to molecule *C* and found two bioisosteric pairs {a1,b1} and {a1,c2}.

Once a pair has been found, it is then tested to see whether it can be added to a bioisosteric group. A bioisosteric group is the set of fragments that occupy the same volume in the receptor site. As this collection of fragments are found in the same volume of the active site in the receptor it is assumed that they all play an equivalent role and are thus labelled as bioisosteres. For example, the bioisosteric group of a1 would be the combination of all of the bioisosteric pairs of a1. An example of one such bioisosteric group is illustrated later in Figure 3-7. In Figure 3-4, as both b1 and c2 have an overlap with a1, they are included in the group of a1. Finally, the algorithm is repeated using *B* and *C* as the reference ligands to obtain all of the bioisosteric groups from the ligand overlay file. Additionally, the section groups are filtered to remove fragments that have an identical 2D structure and merged to include all groups that occupy the same space. For example, if two section groups contain fragment a1, then they are combined in order to remove order dependence.

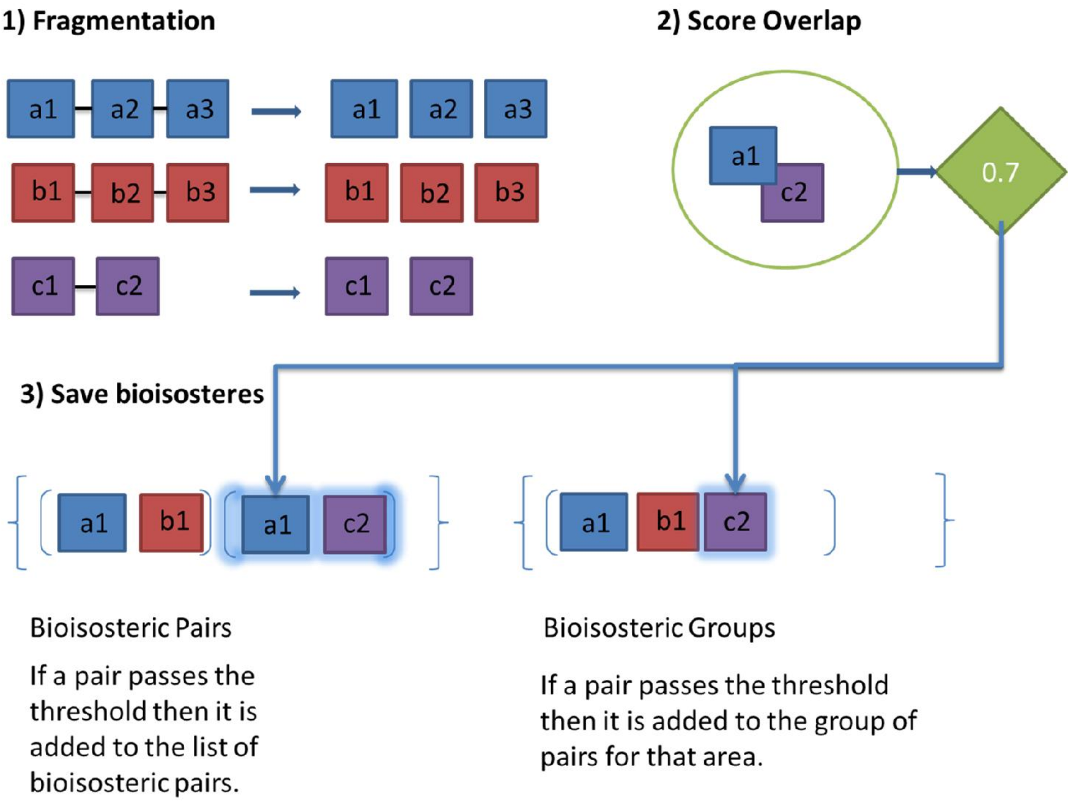

1) Fragmentation
a1
a2
a3
a1
a2
a3
b1
b2
b3
b1
b2
b3
c1
c2
c1
c2
2) Score Overlap
a1
c2
0.7
3) Save bioisosteres
a1
b1
a1
c2
a1
b1
c2
Bioisosteric Pairs
If a pair passes the threshold then it is added to the list of bioisosteric pairs.
Bioisosteric Groups
If a pair passes the threshold then it is added to the group of pairs for that area.

Figure 3-4. An illustration of the algorithm.

3.2.5 Identification of fragment pairs

The volume overlap was determined for each fragment pair using a simplified Gaussian function, following Kennewell et al. (2006). For each atom in the reference fragment, the Euclidean distances were measured to all atoms in the query fragment and a score was assigned for the given distance using the scoring function below. The scores for each query atom were then summed to give a score for the atom in the reference fragment. Then summing these scores over the atoms in the reference fragment produced a score for each pair of fragments. Again, following the published protocol, the average score of the overlaid pair was calculated to account for size bias in the formula,

$$score = \frac{2}{m+n}\sum_{j=1}^{m}\sum_{i=1}^{n} \mathrm{e}^{-d_{i,j}^{2}}, \qquad \text{Equation 3-1}$$

where *m* and *n* are the numbers of atoms in the reference and query fragments, respectively, and a pair of fragments with a score greater than 0.7 was kept as a candidate bioisostere.

3.3 Results

Number of targets	Number of ligands	Number of fragments	Number of unique fragment pairs	Number of unique bioisosteric pairs	Number of bioisosteric groups
121	1445	6586	43341	3551	1003

Table 3-1. A summary of results from the test set. The number of unique fragment pairs refers to all combinations of fragments irrespective of which target they came from, which have been filtered to remove 2D duplicates. Number of unique bioisosteric pairs is the number of bioisosteric pairs derived from the targets and filtered to remove 2D duplicates.

The methods detailed in Section 3.2 were carried out on the pharmacophore validation ligand set to find all the bioisosteric pairs and groups. Table 3-1 shows a summary of these results where it can be seen that in all 121 targets in the pharmacophore validation data set there are 1445 ligands, which produced 6586 fragments in total. As with the import of the ligands, there was an additional problem of fragments failing to pass the RDKit sanitization tests, so that the total number of fragments used is less than the total number generated. For example, in Table 3-3 the target elastase (P00772) had fewer fragments generated than total ligands. From the 6586 fragments, duplicate pairs were removed and a set of 43341 unique fragment pairs was derived. Table 3-3 shows a summary of the results for each target including: the number of fragments in the target site created using the BRICS fragmentation scheme; and the number of bioisosteric pairs that the adapted Kennewell methodology produced. In order to distinguish between the adaption of the Kennewell method for this project and the original method, the adapted implementation will be referred to as the BRICS-fragmentation method herein. For example, the target carbonic anhydrase II, P00918, resulted in 96 fragments produced from 14 ligands. These fragments produced 225 pairs, from which 21 groups were found. An example pair from P00918 can be seen in Figure 3-5.

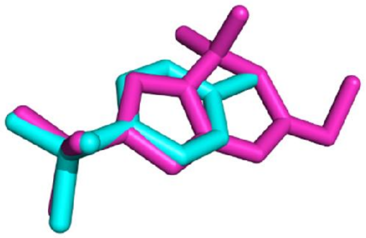


Figure 3-5. An example bioisosteric pair from target P00918.

3.3.1.1 Common pairs across targets

While the original motivation for the Kennewell method was to identify target-specific bioisosteric pairs, it was thought that if a pair was present in more than one target in the test set, then the pair might be less sensitive to target specificity. Additionally, it was hoped that when searched over a diverse set of ligands, a collection of general bioisosteres could be found. These general bioisosteric pairs would be useful for a bioisosteric test set as they would be the minimal requirement for a bioisosteric similarity method to discover. On the other hand, fragment pairs that were only observed in one target would not necessarily be helpful as they may be applicable in a limited number of scenarios only. Therefore, a bioisosteric similarity methodology that returned these would not necessarily produce any meaningful transformations that could be used in novel protein targets or chemical environments.

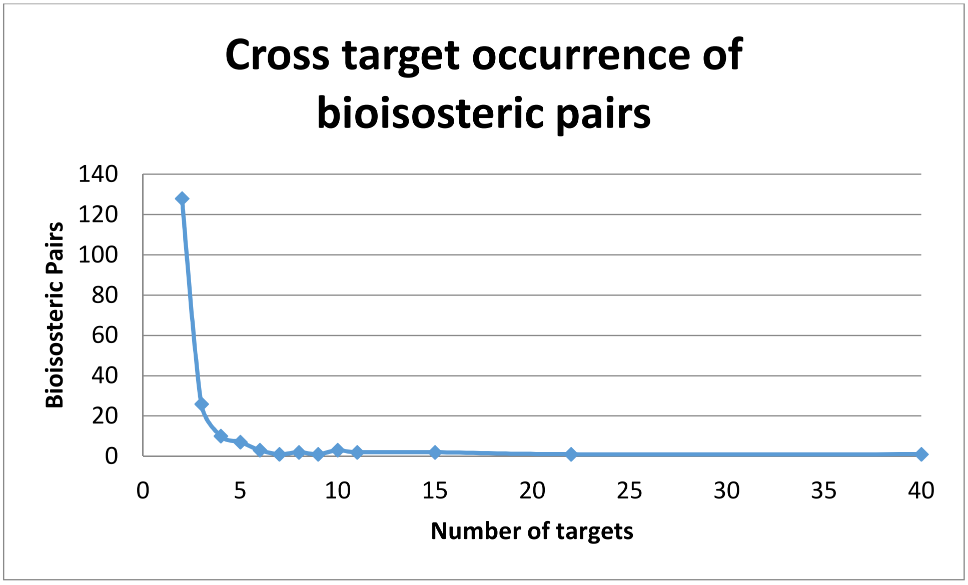


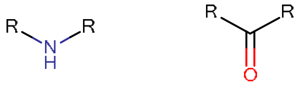
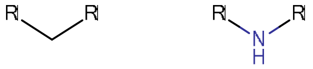
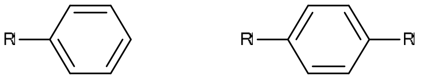
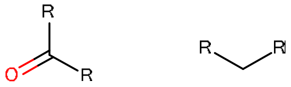
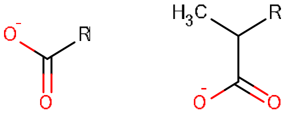
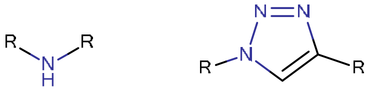
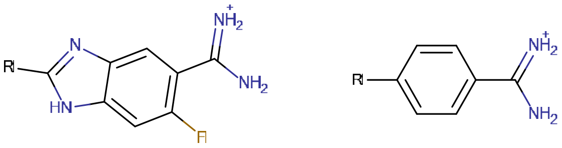
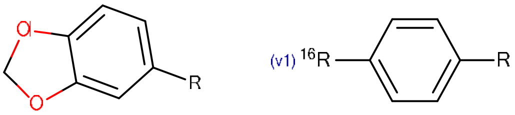
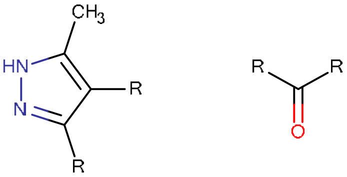
Figure 3-6. Graph showing the cross target occurrence of the bioisosteric pairs in the validation overlays data set. The graph shows the frequency of pairs found for different numbers of targets.

One aspect of the pharmacophore validation data set is that it has been gathered from a large number of activity classes. While this may mean that the bound ligands themselves would be diverse and not necessarily applicable across different types of targets, it was thought that a fragment-based approach would have some independence of this restriction. Being fragments, they would be more likely to represent functional groups or pharmacophoric points, so some generality over targets would be expected; especially as functional group based descriptors have been shown to be successful in the past (Holliday, Jelfs, Willett, & Gedeck, 2003).

Once all unique bioisosteric pairs were found from the validation overlays, the frequency of their occurrence across targets was counted. Pairs from two targets were considered the same if they had identical topological structures. A summary of the results is presented in Figure 3-6. The vertical axis of the graph is the frequency of bioisosteric pairs that were found in a given number of targets and the horizontal axis is the number of targets. As an example, 128 bioisosteric pairs were found in two targets. However, this number declines rapidly to just above 20 found in three targets; ten found in four targets; and seven found in five targets. The number of pairs found in more than five targets declines rapidly and is between one and three, with only one pair being found in 22 and 40 targets. In total, nine bioisosteric pairs were found in 10 or more targets. In fact, the graph can be interpreted as a measure of the specificity or generalisability of the bioisosteres discovered by the BRICS-fragmentation method. The horizontal axis can be considered a measurement of increasing generality of the bioisosteres and the curve that connects the points can be used to visualise the extent to which the bioisosteric pairs discovered cannot be generalised across a diverse number of targets. The graph therefore shows that there is little evidence to suggest that fragment bioisosteric pairs are generalizable between targets.

Table 3-2 shows some example cross-target bioisosteric pairs along with the number of targets in which they were found with decreasing generality down the table. These bioisosteric pairs have been obtained directly from aligned bioactive structures without using chemical information. Consequently, the chemistry of the bioisosteric pairs can be investigated *post hoc* to see whether they would be expected to share common chemical properties. Immediately a couple of striking observations come to light. Firstly, the most common pair is a surprising observation as the amine and the ketone are likely to have very different chemistry thanks to their different polarities. Yet they may have similar roles as hydrogen bond acceptors in some environments. This is also the case with the second pair being an amine with an alkane. Secondly, the third and penultimate pair have a different number of attachment points, suggesting that one is a terminal fragment, whereas the second may still have an extra functional group. Finally, the last pair have an unusual size difference. Nevertheless, the pairs with different expected chemical properties are not singletons and are observed across different targets. Therefore, they present an opportunity to explore novel bioisosteric interactions. However, in these cases, it is important to return to the original structures along with the receptor and investigate the surrounding chemical environment.

Table 3-2. Examples of bioisosteric pairs and the number of targets in which they were found.

Pair	Number of targets in the test set where the bioisosteric pair was found.
	40
	22
	11
	8
	6
	5
	3
	3
	2

3.3.2 Identified bioisosteric groups target test set

The next step was to look at the bioisosteric groups that were found in the data set. Table 3-3 shows the number of groups produced for each target. To return to the example of target P00918, Table 3-3 shows that from the 225 bioisosteric pairs, 21 bioisosteric groups were formed. An example of a bioisosteric group taken from P00918 is shown in Figure 3-7. Thus, Figure 3-7 shows the group of fragments that occupy the same three-dimensional space at the binding site in P00918 and which are assumed to have the same function. In P00918, the average pairwise similarity using 2D fingerprints and the Tanimoto similarity coefficient of the groups was 0.265 which suggests that the groups in the target are reasonably diverse, on average.

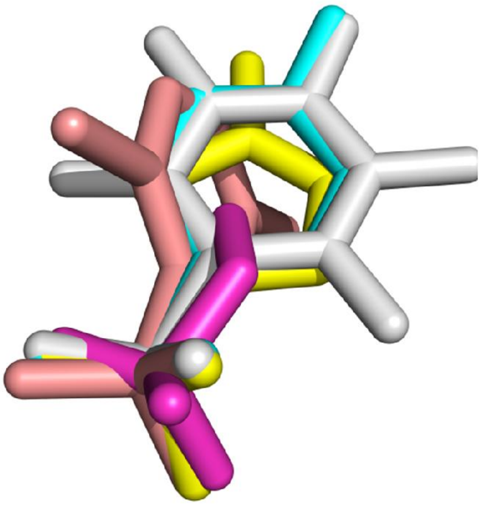


Figure 3-7. An illustration of a bioisosteric group taken from target P00918.

Bioisosteric groups hold the potential for implicitly encoding more chemically relevant information than bioisosteric pairs. This is due to the fact that they not only encode the target from which they came but also information relating to the precise location of the fragments at the binding site. For example, they might be related to specific chemotypes or pharmacophoric features that the fragments have in common.

3.3.2.1 Diversity of bioisosteric groups

The group diversity is analysed in more depth in Figure 3-8, which shows a histogram distribution of the average pairwise similarity using 2D fingerprints and the Tanimoto similarity coefficient of all groups derived from the target test set along with a percentage cumulative frequency curve. It is seen that on average the bioisosteric groups have a high diversity, if measured by 2D similarity.

The modal average similarity value is 0.175 and the percentage cumulative frequency curve shows that half of the bioisosteric groups have an average similarity or 0.219 or less.

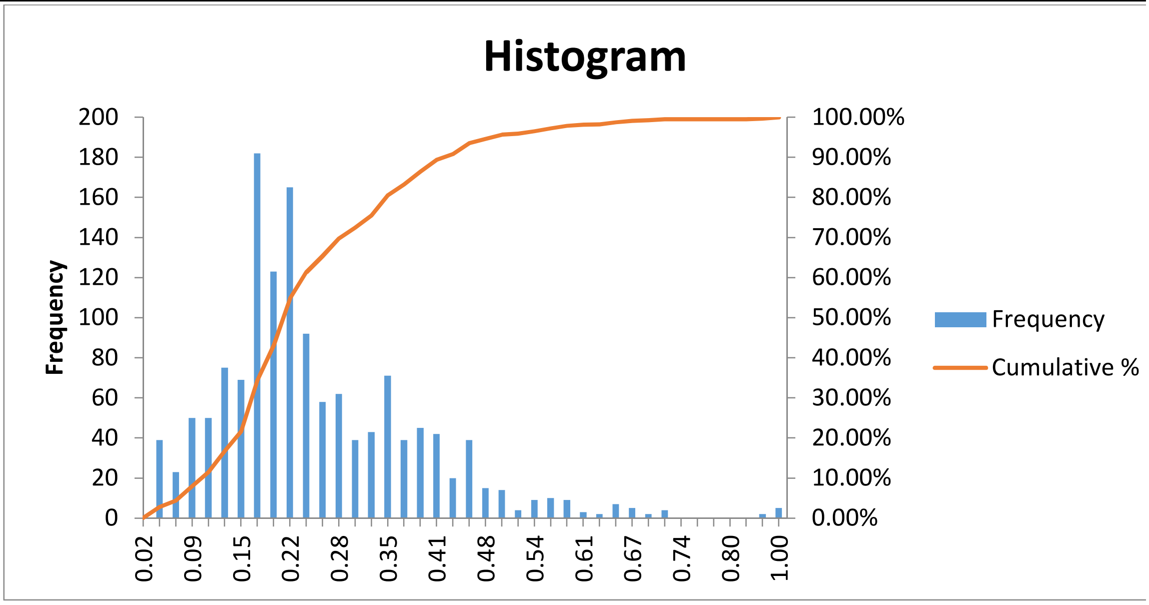


Figure 3-8. A histogram showing the average similarity of all bioisosteric groups derived from the target test set.

3.3.3 Evaluation of 2D similarity search using data set

In order to see whether a simple 2D similarity search could produce bioisosteric pairs that could be verified by the above data, a 2D similarity search was carried out on the pharmacophore validation data set to find non-identical similar pairs. As the bioisosteric pairs had been generated using the BRICS-fragmentation method, they were bioisosteres from particular target sites. In order to use the bioisosteric pairs as a test set, it was necessary to define a pair as being a known bioisostere if it was a bioisostere in at least one of the target sites. Therefore, a similarity search was said to have correctly predicted a bioisosteric transformation if a pair of fragments with a high similarity score was also a pair of fragments produced as a bioisostere in at least one of the targets using the method described in Section 3.2. The 2D Tanimoto similarity coefficient was calculated for all 43341 pairs of fragments for each target using Morgan extended circular fingerprints with a radius of 2 from the RDKit (Landrum, n.d.). (RDKit Morgan fingerprints with radius 2 are the RDKit equivalent of ECFP_4 fingerprints.) If the similarity was greater than 0.65 then the pairs were kept. This threshold was selected because work by the Bajorath group suggested that below a similarity of 0.64, the Tanimoto similarity coefficient is not good at distinguishing between molecules (Lounkine et al., 2008). Once these pairs had been collected, they were tested to see whether they were also found in the bioisosteric pairs derived using the BRICS-fragmentation method. Of the 43341 unique pairs from the data set, 82 pairs had a similarity more than 0.65. Furthermore, of those 82 pairs, only 69 were known to be bioisosteres from the Kennewell methodology. So, of a potential 3551 bioisosteric pairs, a 2D similarity search with a threshold identified 69 with 11 false positives, which is a false positive rate of 16%. An illustration of a correctly predicted active bioisosteric pair that was retrieved by this method is shown in Figure 3-9.

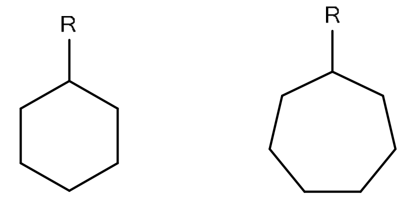
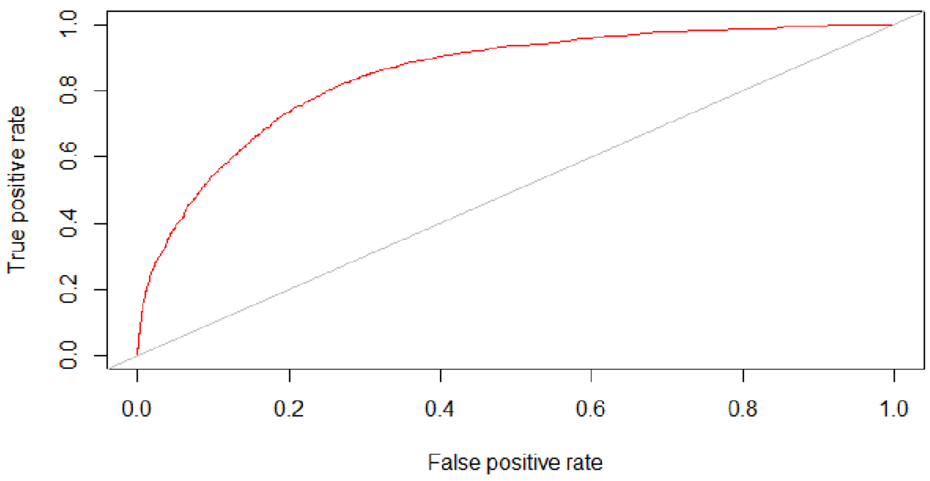


Figure 3-9. Correctly predicted bioisosteric pair from P43235 found using a 2D Tanimoto similarity search with a 0.65 cut off.

3.3.3.1 ROC analysis

While the use of an arbitrary threshold for a similarity search performed poorly, it is well known that 2D similarity for small molecules is necessarily small due to the number of atoms to be compared. As they have few atoms, the number of bits set in a fingerprint tends to be lower than for larger molecules and thus tend to have smaller similarity measurements using Tanimoto (Leach & Gillet, 2007). Subsequently the performance of a similarity search was assessed by evaluating the relative position of known bioisosteres in a ranked list of pairs. The data set had 43341 total pairs of which 3551 were classified as active bioisosteres (Table 3-1). A similarity search using RDKit Morgan circular fingerprints then ranked all the pairs and a ROC curve was plotted to reflect the ranking of those active bioisosteres. AUC and the BEDROC statistic, $\alpha = 20$, which is weighted to reflect early classification in ranked lists, were calculated as well as enrichment factors at the 1% and 5% level.

The bioisosteres and fragment pairs were pooled over all targets in order to produce more robust estimates of performance. It was thought that using only the ligands from a single target would be biased from the likelihood that most ligands within a target were likely to share some structural properties and would thus have a higher topological similarity. It should also be noted that it was assumed there was a low occurrence of bioisosteric fragments that were not classified as bioisosteres in our test set as a result of the fragments being present in different targets and hence not having the opportunity to be aligned using the above method. The ROC curve along with summary statistics are presented below in Figure 3-10.



	AUC	BEDROC	Enrichment factor 1%	Enrichment factor 5%
2D Similarity Search	0.850	0.809	17.7	7.78

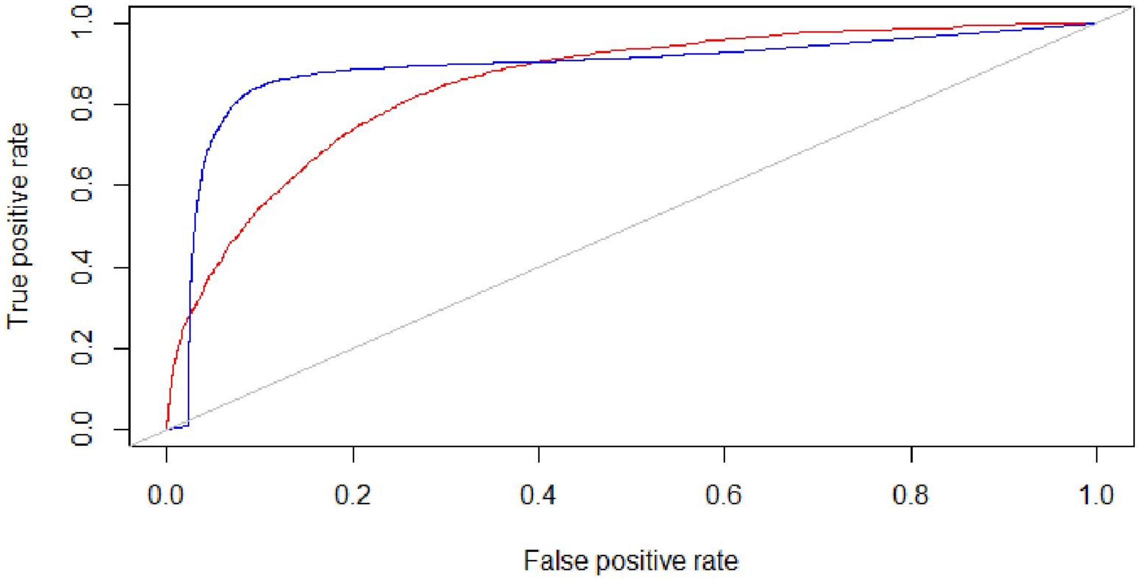
Figure 3-10. ROC curve to show the performance of RDKit Morgan circular fingerprints.

Unlike the simple similarity threshold search, this analysis gives a much greater insight into the performance of the 2D similarity method for use in bioisosteric searches with virtual screening. The AUC statistic shows that, on average, the 2D Tanimoto search will be 85% more likely to rank a randomly chosen positive bioisosteric pair higher than a randomly chosen non-bioisosteric pair. The BEDROC statistic is biased to promote methods that rank positive pairs early (Zhao et al., 2009). Here the BEDROC statistic shows that there is an 81% probability that a randomly selected positive bioisosteric pair will be ranked higher than a randomly selected negative pair with an exponential distribution. The enrichment factors show the extra number of positive pairs that would have been selected compared to a random selection from the database in the top 1% and 5% of the ranked compounds.

3.3.3.2 Comparison with 3D search method

The above 2D search was further compared with a method to find 3D shape similarity. As the 3D approach explicitly takes conformational information into account, it was expected to perform better than a topological 2D search. The software used was Shape-it (*Shape-it*^{TM}, n.d.), an open source implementation of the Gaussian Rapid Overlay of Chemical Structures (Grant et al., 1996) written in C++.

Shape-it first aligns the molecules and then calculates the shape overlap using Gaussian volumes. Shape-it aligns the molecules in 3D space by first calculating the centres of mass in order to centre the fragments which are then aligned by their principal axes. A gradient ascent algorithm is then used to find the optimal rotation. Additionally, simulated annealing is carried out to test whether the algorithm has been caught in a local optimum. The volume overlap score is then calculated using either a Tanimoto or Tversky similarity coefficient.



	AUC	BEDROC, $\alpha = 20$	Enrichment factor 1%	Enrichment factor 5%
2D Similarity Search	0.850	0.809	17.7	7.78
3D Similarity Search	0.890	0.872	0.268	14.3

Figure 3-11. ROC curve comparing the performance for discovering bioisosteric pairs of both 2D and 3D similarity searches. The red curve represents 2D similarity using circular fingerprints and Tanimoto coefficient and the blue curve represents the performance of the 3D similarity using a Gaussian overlap and Shape-it. The table shows summary statistics for the two approaches including area under curve (AUC), BEDROC, and Enrichment factors at 1% and 5% levels.

Following the experiment for the 2D fingerprint scheme above, the 3D Tanimoto score for volume overlap of the two fragments was calculated for each pair and their ability to classify bioisosteres was evaluated. The comparison of the two methods is shown with the ROC curve and corresponding summary statistics in Figure 3-11.

Intuitively, the graph suggests that after a weak start, the 3D method performs better overall than the 2D method as the true positive rate increases very quickly. This is confirmed by the AUC statistics with the AUC of the 3D method performing very well and better than the 2D method. As the two methods have been tested on the same test set with the same underlying distributions and using the same tuning parameters, the BEDROC test can be used to directly compare their effectiveness at early detection of bioisosteric pairs (Zhao et al., 2009). The BEDROC statistic for the 3D method is also greater than the 2D method, which suggests that the 3D search performs well in comparison to the 2D search when early discovery is factored in. This appears to be supported by the 5% enrichment statistic but there appears to be an anomaly with the 1% enrichment statistic as the 3D search performs particularly badly. This is due to the Shape-it implementation, which occasionally gives a 0 score for two fragments without explanation and requires further investigation.

On the other hand, using this method to assess a 3D similarity search is likely to be inappropriate as it is itself a 3D similarity method based on a simplified Gaussian score function. Subsequently, there will be a high correlation between the two methods. The reason why the classification of the Gaussian method is not perfect is likely to be due to the fact that the Gaussian method uses the centres of gravity of the fragments as the start point for the alignment, whereas the alignment used to identify bioisosteric fragments is based on the protein structures and may not necessarily be centred on their centres of gravity. Secondly, the simplified scoring function does not take atomic radii into account so the volumes will not be equal between the two methods.

3.4 Discussion

This chapter adapted the work of Kennewell et al., with the aim of producing a test set of bioisosteric fragments using high quality crystallographic data. Since the initial publication of the method, a large amount of crystallographic data has been made available. This has given the opportunity to apply the methodology to a considerably larger number of targets than the original paper. Whereas, the Kennewell publication produced pairs of bioisosteric fragments, here the concept of bioisosteric groups was introduced. When the method was applied to the pharmacophore validation data set (Giangreco et al., 2013), a number of interesting target specific groups were identified, which were found to be diverse with respect to 2D structural similarity.

3.4.1 Definition of bioisosterism

A recurrent theme in this analysis is the definition of bioisosterism. When searching for bioisosteric pairs, the bioisosteric relation is defined such that two pairs are bioisosteric when they occupy the same three-dimensional space within a given target. When this is extended to bioisosteric groups the assumption of transitivity is added to the relation so that if A and B are bioisosteric within a target X, and B and C are bioisosteric within the same target X, then A and C are bioisosteric within target X. Nevertheless, when used in practice, the definition of bioisosterism is often desired to be invariant to a specific target and a specific location. When pairs or groups are combined over targets, they necessarily lose this target or location specific information.

3.4.2 Performance of Kennewell methodology

The application of the methodology to a much larger data set also enabled the target-specific methodology of the original paper to be evaluated. First, in relation to the ability to discover bioisosteric pairs, the results showed that very few bioisosteres found in this manner are general over the data set, i.e., found in more than one target. Thus, the method is highly sensitive to target information and performs poorly when looking for general bioisosteres. Given that of the 3551 bioisosteric pairs identified, only 187 were present in more than one target, it suggests that this is not a generally applicable method for finding bioisosteric pairs. In fact, even the occurrence of pairs in two targets was low compared to the total number of pairs discovered. Although, the data set is diverse with respect to the protein types, there were a number of targets from the same protein family, thus it would have been thought that if this information was domain specific, there would at least be a high occurrence of bioisosteric pairs found in two or three targets. This suggests that even with an increase in crystallographic data for many targets, it is unlikely that the method would have a use in anything other than a highly specific target based setting. In practise, once a binding site has sufficient research interest to require a number of known binding ligands to be crystallised with the protein and their structural information elucidated, it is not clear what practical use this method would bring.

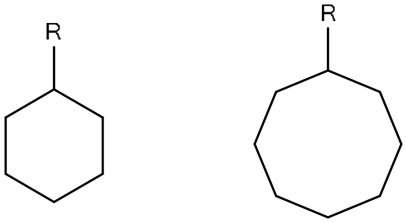


Figure 3-12. A pair from P00760 that was not classified as a bioisostere but in assay data was attached to aliphatic linkers in 74% of cases.

Additionally, while the methodology produces a consistent method for identifying a bioisosteric relation, it is unlikely to give a complete set of bioisosteres for a given set of binding ligands. For example, a bioisosteric pair may be missed simply by not being overlaid at the binding site because the data are not extensive enough, which could lead to false negatives in the data set. For example, the pair from target P00760 shown in Figure 3-12 was not identified as a bioisosteric pair. Yet, their 2D similarity scored higher than the threshold in Section 3.3.3. However, according to the SwissBioisostere database (Wirth et al., 2012), this transformation is bioisosteric in 74% of assays when attached to an aliphatic linker.

In summary, this chapter provides evidence against using target-specific methods for producing generalizable bioisosteric pairs. In addition, it also suggests that the use of the Kennewell methodology is unlikely to have a practical use even with a growth of high quality crystallographic data.

3.4.3 Data

There may also be some restrictions with the data set used to generate the validation bioisosteric pairs. First, our sample may be subject to selection bias based on past design decisions. In order to find a set of ligands that bind to the target, it is likely that rather than search all chemical space, close 2D analogues would have been selected to trial. Subsequently, our data may be biased towards molecules that have a high 2D similarity. This is somewhat negated by the design of the AstraZeneca data set, which deliberately selected a diverse set of ligands from those available in the PDB.

Additionally, when collating bioisosteric pairs over a number of different targets, there may be a number of pairs within the set that are bioisosteric but are not classified as such. One reasonable explanation for this is that there may be two bioisosteric fragments that were not present in the same target ligand set and so were not able to be compared using the BRICS-fragmentation method, thus giving an incorrect classification. Therefore, in order to use this data set as a validation set, we must assume that these occurrences are low.

The data may also be subject to bias from the nature of crystallisation. There may be properties that make molecules easy to crystallise that bias the properties of the training set. This is more likely to influence the protein choice for the target rather than the ligand as that is the molecule that is most likely to determine crystallisation. It would be interesting to see whether there was any correlation between bioisosteres discovered by crystallisation and other methods for finding the 3D coordinates of ligands in a target site such as NMR.

3.5 Conclusions

The original motivation of the work in this chapter was to create a test set of bioisosteric fragment pairs that were based on experimentally validated crystallographic data that could be used to evaluate a 3D fragment-based similarity search method. However, there are some inherent problems with the methodology that are summarised here. Firstly, there is a positive signal problem as a result of the way the data is collected. That is to say that a ligand will only be in the data set if it can bind to the target site. Thus, there are no negative data points and therefore the methodology does not produce decoys which can be used in a virtual screening evaluation.

The bioisosteric assumption can be described as *“if two fragments occupy the same space in the binding site then they are bioisosteric”*. However, this mistakes a necessary condition for a sufficient condition. It may well be the case that two fragments that overlap in free space and have no role in the activity of the ligand are classified as bioisosteric. In other words, while it is necessary for two fragments to be overlaid at the target site in order to be bioisosteric, it is not a sufficient condition; a given target site would then have a number of false positives.

Additionally, an unwritten non-bioisosteric assumption can be hypothesised: *“if two fragments do not lie in the same volume at the target site then they are not bioisosteric”*. As the bioisosteric pairs are generalised over a number of different targets, these false negatives will be exacerbated. This demonstrates the other consequence of the positive signal problem mentioned above. As the relation is necessary but not sufficient, the data only show positive relations, it is impossible to say anything about non-actives in the data set.

In general, these empirical fragment-based methods present a generalisability problem. On the one hand, the bioisosteric fragments most useful for a 3D similarity method are those that can be generalised as active over many targets. On the other hand, these bioisosteric pairs are the least likely to exhibit some novel or interesting chemotype or activity. In contrast, bioisosteric fragment pairs that are rare over targets are those that are most likely to have high information value and exhibit interesting activity profiles, yet these are more likely to be target-specific and not of value to a fragment-based similarity scheme. However, further research could use a probabilistic model to exploit the general and target-specific information in developing fragment-based drug development workflows.

For the reasons stated above, this suggests that the methodology described is not appropriate for the construction of a test set without further work on the false positives and the false negatives. Consequently, it will prove difficult to construct an adequate test set for the development of a 3D fragment-based similarity method. Therefore, the following chapters focus on the development of a 3D similarity method for comparing whole molecules for which there are established data sets that can be used to evaluate the method.

Table 3-3. Table giving the ligand data set and the number of fragments produced for each set along with the number of bioisosteric pairs identified and the number of collected groups of bioisosteres.

Target	Target name	Number of ligands	Number of fragments	Number of pairs	Number of groups
A9JQL9	dehydrosqualene synthase	8	26	2	2
O14757	serine/threonine-protein kinase Chk1	12	143	414	18
O14965	serine/threonine-protein kinase 6	9	80	63	9
O15530	3-phosphoinositide dependent protein kinase-1	9	17	1	1
O15530	3-phosphoinositide dependent protein kinase-2	5	57	78	8
O60674	tyrosine-protein kinase JAK2	7	51	59	4
O60885	human BRD4	8	31	14	4
O76074	cGMP-specific 3′,5′-cyclic phosphodiesterase	9	53	35	6
O76290	Pteridine reductase	9	30	21	3
P00374	dihydrofolate reductase	9	69	118	10
P00469	thymidylate synthase	11	13	3	1
P00489	protein (glycogen phosphorylase)	10	73	243	8
P00509	aspartate aminotransferase	7	30	9	5
P00517	cAMP-dependent protein kinase, alpha-catalytic subunit	9	90	173	9
P00520	proto-oncogene tyrosine-protein kinase ABL	11	27	4	3
P00523	proto-oncogene tyrosine-protein kinase Src	10	66	30	4
P00730	carboxypeptidase A	8	23	8	3
P00734	alpha thrombin	8	201	526	41
P00742	coagulation factor XA	9	231	810	50
P00749	protein (urokinase-type plasminogen activator)	7	107	228	19
P00760	trypsin	6	97	317	21
P00772	elastase	37	25	3	1

P00797	renin	27	25	12	2
P00808	beta-lactamase	7	31	10	3
P00811	beta-lactamase	8	91	32	13
P00918	carbonic anhydrase II	14	96	225	21
P00929	tryptophan synthase	6	35	21	3
P02829	HSP82	6	37	37	3
P03372	estrogen receptor	9	80	176	17
P04035	protein (HMG-COA reductase)	5	34	23	6
P04058	acetylcholinesterase	28	23	3	3
P04642	L-lactate dehydrogenase A chain	8	43	31	9
P05326	isopenicillin n synthase	7	44	31	4
P06239	LCK kinase	8	58	42	3
P06401	progesterone receptor	8	38	12	3
P07688	cathepsin B	39	52	35	8
P07900	HSP 90-alpha	27	103	216	21
P08069	insulin-like growth factor 1 receptor precursor	11	49	11	4
P08235	mineralocorticoid receptor	14	13	21	2
P08254	stromelysin-1	8	55	23	7
P08581	hepatocyte growth factor receptor	7	59	69	11
P08709	coagulation factor VII	11	38	15	6
P09467	fructose-1,6-bisphosphatase 1	22	26	12	5
P09955	procarboxypeptidase B	5	42	27	7
P09960	leukotriene A-4 hydrolase	14	96	328	22
P0A017	dihydrofolate reductase	7	52	95	9
P0A5J2	methionine aminopeptidase	9	31	6	3
P0ABP9	purine nucleoside phosphorylase	22	25	32	4
P0AD64	beta-lactamase SHV-1	13	14	1	1
P0AE18	methionine aminopeptidase	6	76	82	11
P0C5C1	beta-lactamase	16	42	43	6
P10275	androgen receptor	12	47	39	5
P11309	proto-oncogene serine/threonine-protein kinase Pim-1	8	91	271	11

P11509	cytochrome P450, family 2, subfamily A, polypeptide 6	21	20	8	4
P11838	endothiapepsin	6	40	16	8
P12758	uridine phosphorylase	8	29	21	8
P14174	macrophage migration inhibitory factor	30	46	33	4
P14324	farnesyl pyrophosphate synthetase	10	15	20	2
P14324	farnesyl pyrophosphate synthetase	5	25	10	4
P15090	fatty acid-binding protein, adipocyte	23	26	9	2
P15121	aldose reductase	23	102	230	14
P16184	dihydrofolate reductase	31	39	19	3
P17612	cAMP-dependent protein kinase	19	52	40	9
P18031	protein (protein-tyrosine phosphatase 1b)	14	160	360	28
P22906	dihydrofolate reductase	13	32	25	3
P23470	receptor-type tyrosine-protein phosphatase gamma	8	29	23	7
P24182	biotin carboxylase	7	43	36	7
P24627	lactotransferrin	8	46	8	4
P24941	cyclin-dependent kinase 2	18	119	316	16
P25440	bromodomain-containing protein 2	9	35	32	7
P25774	cathepsin S	4	104	171	15
P25779	cruzain	8	43	13	6
P27487	dipeptidyl peptidase IV soluble form	10	177	413	26
P28482	mitogen-activated protein kinase 1	11	36	8	6
P28523	casein kinase II	7	53	46	7
P28845	corticosteroid 11-beta-dehydrogenase isozyme 1	13	45	5	4
P30291	wee1-like protein kinase	5	25	35	4
P30405	peptidyl-prolyl cis–trans isomerase F, mitochondrial	9	17	11	2
P35557	glucokinase isoform 2	27	43	32	8
P35968	vascular endothelial growth factor receptor 2	12	63	31	9

P36897	TGF-beta receptor type I	15	17	15	3
P39900	macrophage metalloelastase	16	78	117	12
P41148	endoplasmin	7	28	13	4
P42330	aldo-keto reductase family 1 member C3	6	36	10	4
P42574	caspase-3	8	36	2	2
P43235	cathepsin K	5	68	51	9
P45452	collagenase 3	5	94	178	15
P47811	mitogen-activated protein kinase 14	6	69	49	15
P48736	phosphatidylinositol-4,5-bisphosphate 3-kinase	5	13	5	2
P49841	glycogen synthase kinase-3 beta	24	47	49	5
P50579	protein (methionine aminopeptidase)	5	46	8	6
P51857	3-oxo-5-beta-steroid 4-dehydrogenase	14	12	8	2
P51955	serine/threonine-protein kinase NEK2	13	68	69	6
P52700	metallo-beta-lactamase L1	10	26	7	3
P53779	mitogen-activated protein kinase 10	15	89	90	16
P54760	ephrin type-B receptor 4	10	58	42	5
P56658	adenosine deaminase	12	46	35	7
P56817	beta-secretase 1	16	83	50	12
P59071	phospholipase A2	14	50	4	4
P61823	pancreatic ribonuclease A	9	29	26	6
P68400	casein kinase II	11	41	36	7
P78536	ADAM 17	7	83	91	16
P80457	xanthine dehydrogenase	5	17	23	4
Q00511	**uricase**	**6**	**10**	**18**	**2**
Q02127	dihydroorotate dehydrogenase, mitochondrial	5	53	55	9
Q04771	activin receptor type-1	27	27	19	4
Q07343	cAMP-specific 3′,5′-cyclic phosphodiesterase 4B	14	84	64	13
Q08499	cAMP-specific 3′,5′-cyclic phosphodiesterase 4D	18	82	134	18
Q10714	angiotensin converting enzyme	10	29	19	5

Q13526	peptidyl-prolyl cis–trans isomerase NIMA-Interacting 1	8	110	284	26
Q16539	p38 MAP kinase	5	157	280	26
Q3JRA0	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	22	17	3	1
Q57834	Tyrosyl-tRNA synthetase	5	21	9	2
Q581W1	pteridine reductase 1	12	21	37	3
Q92731	estrogen receptor beta	21	37	170	4
Q9BJF5	calmodulin-domain protein kinase 1	7	53	107	9
Q9BZP6	acidic mammalian chitinase	6	18	8	4
Q9L5C8	beta-lactamase CTX-M-9	13	61	66	14
Q9QYJ6	phosphodiesterase-10A	17	57	36	10
Q9T0N8	cytokinin dehydrogenase 1	23	30	40	5
Q9Y233	cAMP and cAMP-inhibited cGMP 3′, 5′-cyclic phosphodiesterase 10A	19	35	7	3

4 Spectral geometry of molecular shape

4.1 Introduction

Three-dimensional molecular similarity searching involves comparing the 3D geometry of a reference molecule against a database of molecules in order to find those that are likely to have the same properties (Leach & Gillet, 2007). In principle, as compounds are active in three-dimensional space, their 3D shape ought to have greater information content when compared to conventional 2D methods. Yet in practice the application of 3D similarity searches to large databases of molecules presents a number of obstacles. In particular, 3D molecular similarity methods face two problems of computational complexity that have limited the success of large scale 3D virtual screening methods to date: generating the optimal 3D alignment of molecules and handling conformation variation.

In order to rank a database of molecules against a reference molecule using 3D shape it is necessary to map the shapes to a space where some notion of distance can be used to measure how close, or similar, the 3D shape of each molecule in the database is to the 3D shape of the reference molecule. Given two 3D shapes, there are two main ways of measuring this distance. The first is to place the shapes into the same 3D space and measure their common volume overlap. The second is to map both shapes to a descriptor that captures 3D geometry and measure how close the descriptors are in descriptor space. In both cases, the space in which the molecules are compared can be called the comparison space. Importantly, the geometric properties that are to be compared may be sensitive to how they are mapped to the comparison space. For example, if rotating a molecule results in a different location in the comparison space then this will affect the similarity score between two molecules. Thus, a similarity score using volume overlap will change if one molecule is rotated out of the optimal alignment. If a rotation of a molecule does not change its location in the comparison space then the mapping, or descriptor, is said to be rotation invariant. Likewise, if a molecule is translated along a vector and the point in comparison space remains unchanged then this mapping is invariant to translations. Rotation and translation are the only transformations that can be applied to a shape in Euclidean space without bending, tearing, or creating holes. Hence, they are known as rigid transformations. As an alignment of two molecules is carried out using rotations and translations, finding the optimal alignment is the same as finding the rigid transformation that maximises the volume overlap between two molecules. Conversely, a mapping that is not changed by rigid transformations is said to be alignment invariant. Given the computational cost of finding an optimal alignment a descriptor that encodes a rich amount of 3D geometric information and is invariant to alignment is desirable.

This chapter describes the application of recent alignment invariant descriptors developed in the field of computer vision to molecular shape similarity. Spectral and diffusion geometry apply concepts of computational geometry and algebraic topology to the computational analysis of shape (Coifman & Lafon, 2006; Reuter et al., 2006; Sun, Ovsjanikov, & Guibas, 2009). The intuition behind the techniques is to treat a 3D shape as a surface and extract geometric information of the shape from analysis of physical properties of the surface. At the heart of spectral and diffusion geometry is the Laplace-Beltrami Operator, which can be considered the spatial component of partial differential equations over the surface. Geometric information extracted from this operator is intrinsic, meaning that it is defined with respect to the shape itself rather than an external embedding space. Therefore, the descriptors derived from this approach are invariant to transformations in Euclidean space, meaning the descriptors are the same irrespective of alignment. They are also invariant to a specific class of deformation, thus allowing them to capture some notion of flexibility. For example, in the computer vision field, a typical application has been to recognise common objects such as animals or people in different orientations and different poses such as standing and sitting. These properties of alignment independence and deformation – or pose – invariance are desirable in a molecular shape comparison method to allow the comparison of fixed conformers of molecules and for handling conformational flexibility of molecules, respectively.

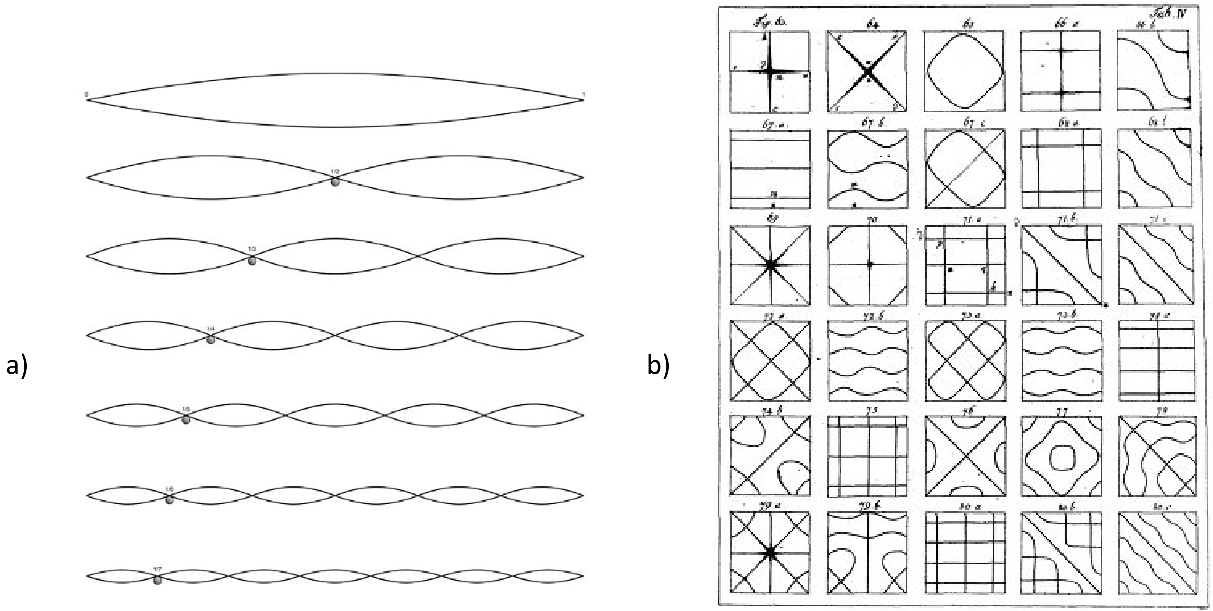
As the following material includes many concepts that may not be well known to the reader, the chapter starts with an intuitive discussion. Once the general concepts have been introduced, a literature review is provided. Spectral geometry is then used as a framework for developing local geometry descriptors for molecules. The main focus in this chapter is on considering molecules as rigid shapes since effectiveness at this level is a pre-requisite for investigating the more complex issue of conformational flexibility but there will be some discussion of the conformation invariance of the local geometry descriptors. The properties of the descriptors are investigated by a qualitative evaluation of local geometry descriptor diversity and a visual demonstration of how well they are preserved under conformation variation. Following chapters then take these local geometry descriptors as the starting point to develop descriptors suitable for virtual screening.

4.2 Spectral geometry and local geometry descriptors

This section gives an intuitive introduction to the ideas underlying spectral geometry as a non-technical overview of the concepts before providing a formal definition of spectral geometry.

4.2.1 Background

Music is heard as a set of frequencies emanating from a vibrating body. For example, when a violin string is rubbed by a bow, it causes the string to vibrate at certain frequencies. As the string is fixed at both ends, there are two waves travelling in opposite directions. At specific frequencies, the waves travelling in both directions fit into each other symmetrically so that they have the same frequency. In music these are the pure tones. Figure 4-1 (a) shows that when this happens there are certain points on the string that do not vibrate. In other words, they are stationary. Of particular interest here are the frequencies that induce this behaviour. The values of these frequencies depend on the length of the string and are known as the resonant frequencies of the string.



Public domain image from Wikimedia Commons (2007)
https://commons.wikimedia.org/wiki/File:Overtone.jpg

Chladni's original sketches. Public domain image from Wikimedia Commons (2006)
https://commons.wikimedia.org/wiki/File:Clafig1.jpg

Figure 4-1. (a) Normal node resonances of a vibrating string and (b) Chladni's original sketches of the patterns formed by sand on a vibrating plate.

As reported by Levy (2006), in 1787, Ernst Chladni carried out a famous series of experiments where he applied a violin bow to a metal plate covered with a thin layer of sand. He noticed that at certain frequencies the sand formed well defined complex geometric patterns, which can be seen in Figure 1 (b). These patterns are formed by sand collecting on the plate where there is no vibration. When the plate vibrates at certain frequencies, the vibrations are distributed over the plate. In some areas there is more vibration than in others and the sand moves away from these areas. At specific frequencies, the vibrations in all directions over the plate have the same frequency and there are regions of the plate that have no vibration at all. This can be considered as the two dimensional case of the string above, where the patterns are equivalent to the points on the string where there is no vibration. In the two dimensional case, the area is one factor that determines the frequencies required to produce these patterns. These frequencies and the patterns they induce are known generally as the resonance and normal nodes of the surface (Levy, 2006).

The frequencies and patterns produced by the above experiment have a precise and elegant mathematical formulation. The vibration of a surface is described by a wave function and the normal nodes are stationary waves. In other words, for a given surface, S, with an area, A, the task is to find the vibrations over that surface, f, such that the vibrations in all spatial directions vibrate with a single frequency, λ. Let the movement of the vibrations in all spatial directions over the surface be described by the operator, Δ, then the behaviour of the vibrations in all directions over the surface is represented as, Δf. If all vibrations over the surface have a single frequency then they can be described as, λf, meaning that the state of a surface when it is vibrating at a resonance node can be articulated mathematically by the following equation, where λ is negative by convention,

$$\Delta f = -\lambda f. \qquad \text{Equation 4-1}$$

The operator in Equation 4-1 is known as the Laplace operator of the surface.

Vibrations propagate over surfaces as waves, which provide a functional representation for f. All that is left to characterise the resonance nodes is to choose the frequencies such that when the vibrations propagate over the surface they do so with the same frequency, λ, in all directions. Generally, the analysis of these stable properties of an operator over a surface is part of a family of problems known as eigenvalue problems that look at the invariance properties of operators on spaces. The solutions to eigenvalue problems are the functions, or vectors in a discrete setting, with the corresponding eigenvalues such that the above equation holds. With vibrating surfaces these are the resonances and the normal nodes. In general, the eigenvalues and their corresponding eigenfunctions, or eigenvectors, are called the spectrum of the operator over the space.

In one dimension, the frequency depends only on the length of the string so that the shape of a one-dimensional line – its length – is characterised entirely by the normal nodes. In two dimensions, the area of a surface and the distances the waves travel in all directions will determine the frequencies required to obtain the resonance of the surface. Intuitively this relates the shape of a surface with the solutions to Equation 4-1. For a two dimensional Cartesian, or flat, surface the Laplace operator is defined as the sum of the second derivatives with respect to the two directional axes, $\Delta = \frac{\partial^2}{\partial^2 x} + \frac{\partial^2}{\partial^2 y}$. Thus, the Laplace operator can be regarded as the spatial component of partial differential equations over a space and encodes how a surface changes with respect to direction. Importantly, the geometric information is captured in the spectrum of the Laplace operator, so that if two shapes are the same then the eigenvalues of their respective Laplace operators are also the same. Furthermore, the spectrum of the Laplace operator can be used to extract geometric information of the shape of the surface, such as the area, the length of its border, and its topological genus, which loosely speaking describes the number of holes in a shape (Levy, 2006). In fact, the geometrical information that can be extracted from the spectrum of the Laplace operator is the foundation of all spectral geometry.

Another intuitive way of considering the spectrum of the Laplacian of the surface is to think of the normal nodes as forming an orthogonal basis. Orthogonal, and orthonormal, bases are the founding principles of many areas of science, from Principal Components Analysis (PCA), to Fourier analysis, and Quantum Mechanics. Normal nodes are the set of vibrations and the associated frequencies that vibrate independently of one another. In other words, normal nodes are orthogonal to each other. While difficult to visualise there are two analogies in related areas of science that have equivalent properties: Fourier analysis, and PCA. Fourier analysis and the normal nodes of a surface are closely related as the waves on a surface are sinusoidal and Fourier analysis shows that a periodic function over an interval, such as a general sound wave, can be decomposed into a linear combination of sinusoidal basis functions. Knowing this, it is possible to approximate any sound wave by adding together a finite number of basis functions and assigning each basis function a weight depending on how much influence it has. PCA treats the covariance function of the data as an operator on the data space and finds the appropriate rotations that give the independent directions of variability in the data. In principle, it assumes that the data points can be constructed from a linear combination of independent variables. These independent variables are directions in the data that do not vary with respect to each other and form the spectrum of the covariance matrix over the data. Once the spectrum of the covariance matrix has been found, the data are then projected into PCA space, giving a low dimensional approximation of the high-dimensional data space. Intuitively, these analogies tell us that the spectrum of an operator over a space contains important information on how the operator behaves, which can be used to reconstruct an approximation of the space.

These concepts can then be generalised from strings and flat surfaces to examine the resonance properties of non-flat surfaces, deformable surfaces, or higher-dimensional shapes, such as volumes.

4.2.2 Shapes as manifolds

In spectral geometry, to enable the concepts described above to be applied to a 3D shape, the shape is defined as a curved 2D surface embedded in 3D space. However, the Laplace operator is not appropriate for curved surfaces. Instead, these geometric properties can be inferred from the Laplace-Beltrami operator, which is the generalised Laplace operator over curved surfaces. The Laplace-Beltrami operator is unique to the surface, with the exception of some rare examples. As the Laplace-Beltrami operator also admits an eigendecomposition, these eigenvalues and eigenfunctions are unique to the Laplace-Beltrami operator and therefore to the shape.

Consider a sheet of paper with two points drawn on it. Any point on the paper can be defined in terms of its x and y coordinates. Additionally, any two points on the sheet of paper can be related to each other by drawing a straight line between them. The length of this straight line is the Euclidean distance between them. Picking the paper up and attaching the two shorter sides to each other creates a 3D shape: a cylinder. However, from the point of view of two points sitting on the surface, it may still be meaningful to describe the distance between them as the straight line along the flat sheet. This brief example has already introduced some interesting concepts: firstly, points on a 2D surface are related to each other by some concept of distance along the surface; secondly, the 2D surface may inhabit a 3D space. In this 3D space, the points on the surface also have a third coordinate, z, to describe their position. Additionally, from the point of view of an observer in the 3D space, the surface has properties that are associated with 3D shape, for example, the cylinder has curvature. This is an example of a 2-manifold *embedded* in 3D space.

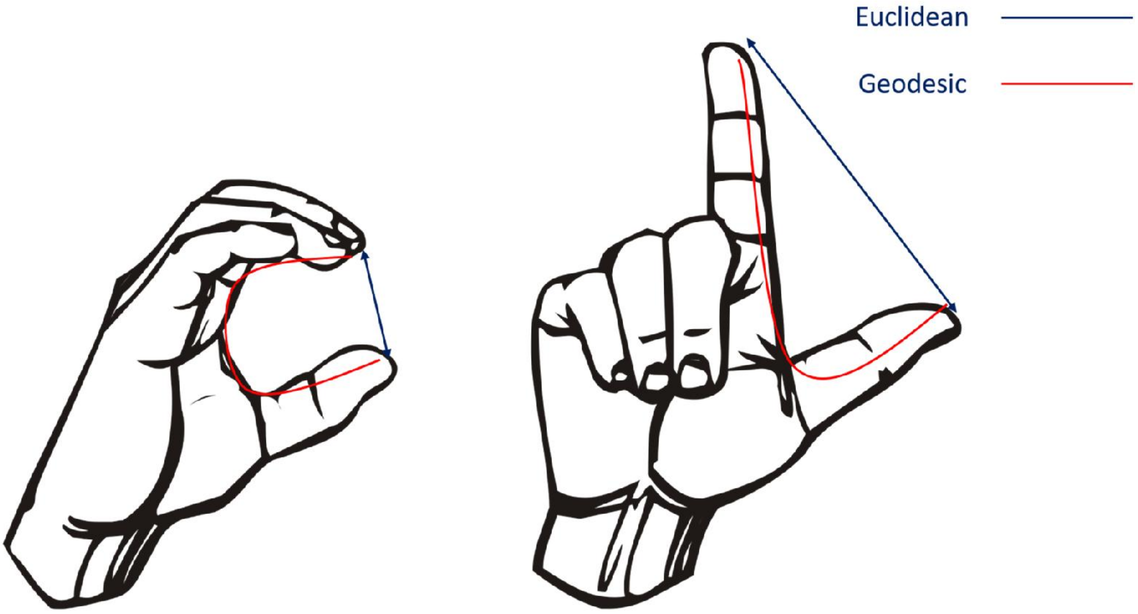
For a further example, the planet Earth can be described as a sphere in 3D space. Nevertheless, all locations on the surface of the Earth are defined by two coordinates: longitude and latitude. From the perspective of people on the surface of the Earth, it makes sense to talk about distances between two cities as being the shortest distance over the surface, which is a straight line on a flat 2D map but a curved line from the perspective of 3D space. The shortest distance between two cities may be a straight line in 3D space – the Euclidean distance – but that may well require a journey through the Earth, which would be useless to someone who wanted to make the trip. In this respect, the shortest distance over the surface of the Earth is called the geodesic distance. An important concept here is that while this distance is a straight line on the 2-manifold surface of the Earth – also known as a map – from the perspective of a satellite or someone floating in space, the distance would be a curved line.

This gives an insight into the geometric information obtained from spectral shape analysis. The shape is treated as a curved 2D surface in 3D space along with a metric that measures the distances between all points along the surface. The Laplace-Beltrami operator encodes the spatial variation, or geometry, of curved surfaces and captures the intrinsic geometry of the shape by describing the rates of change of the properties over the surface in terms of the embedding space. In other words, the fundamental idea is that the surface has a unique measurement of distance that can be used to define the Laplace-Beltrami operator. Two shapes with different metrics will have different spectra of the Laplace-Beltrami operator that will encode different geometric properties.

4.2.2.1 Intrinsic geometry

A very important characteristic of spectral geometry is that the metric that defines distances between points is defined purely in terms of distances over the surface. That is, while the Laplace-Beltrami operator describes curvature and geodesic distances in the embedding space, its properties are defined by the metric over the surface, which is independent of the embedding space. A measurement defined over a manifold independently of the embedding space is called *intrinsic*. Intrinsic geometry has some useful properties that are demonstrated in Figure 4-2. The picture of the hand can be thought of as a surface manifold; the blue line indicates the distance between the thumb and forefinger in the embedding space and the red line shows the geodesic distance along the surface of the hand. The first thing to notice is that regardless of the distance measure used, if either pose of the hand is rotated or translated on the page then the distances are not distorted. Therefore, a geometric property that is derived purely in terms of distances is invariant to rigid Euclidean transformations. This means that the information on the initial position and orientation of the shape is not represented in the geometric property. In practice it means that the geometric properties of two shapes can be compared without the need to align the shapes in the embedding space. In other words, the spectrum is invariant to rigid transformations in the embedding space and is alignment invariant.

Furthermore, it can be seen from the two different poses of the hand in Figure 4-2 that while the distance in the embedded space changes for each different pose, the distance along the surface does not. The transformation of one pose of the hand to the other in Figure 4-2 is a type of non-rigid transformation called isometric deformation. Isometric deformation is a transformation of the surface in the embedding space, such as bending, that maintains the geodesic distances between all points on the surface. Returning to the paper analogy, a sheet of paper is a 2D surface in a 3D space and any transformation that is applied to the sheet of paper without tearing it can be considered an isometric transformation. For example, the paper can be shifted to a new location on the table or rotated from portrait to landscape without tearing the page. These are the rigid Euclidean transformations. Additionally, the paper can be creased and bent without tearing, which are the non-rigid transformations. In this case, descriptors for the two poses that are derived from the Euclidean distances would produce two distinct descriptors that treat the hands as two different shapes. However, descriptors for the two poses that are derived from the geodesic distances would be the same as these are invariant to Euclidean transformations and additionally to isometric-deformations.



Public domain image from Wikimedia Commons (2016).
https://commons.wikimedia.org/wiki/File:Sign_language_L.svg and
https://commons.wikimedia.org/wiki/File:Sign_language_C.svg

Figure 4-2. Distance measurements in embedding space using the Euclidean metric and the Geodesic metric over the surface. The images are the signs for C and L using the American Sign Language respectively. Using geodesic distance, they can be considered as isometric transformations of the same shape.

An important intrinsic geometric feature is Gaussian curvature, which provides a definition of curvature that is independent of the orientation of the shape and is depicted in Figure 4-3. Formally speaking, Gaussian curvature is the product of the principal curvatures at a given point on a manifold. Principal curvatures describe the curvature of a surface around a given point in two principal directions. Using the normal vector as an axis, depicted as the orange arrow at right angles to the surface, a slice is taken in the principal direction and then a second slice in the orthogonal direction. These slices are depicted as the blue planes in Figure 4-3. The curvature the shape makes over these slices is a principal curvature, which is depicted in the planes to the right. Consider Figure 4-3 (a) where the first slice is a negative parabola and the second slice is also a negative parabola, this means that their product is positive and the region has a cupula shape. Alternatively, if the image was rotated 180^{o} both would be positive parabolas and the region will have a bowl shape. Given that it is desirable to have a description of curvature that is intrinsic, meaning that it is independent of any external view or rotational transformation, then both bowls and cupulas ought to have the same type of curvature. In these examples, the two shapes over the slices both have curvatures with the same sign so that their product, the Gaussian curvature, is positive. On the other hand, Figure 4-3 (b) depicts an alternative type of Gaussian curvature where one slice is a positive parabola and the second slice is a negative parabola. The shape would be a saddle or a valley if rotated 180^{o} and the Gaussian curvature would be negative. In general, cupula features will have positive Gaussian curvature and valley features will have negative Gaussian curvature. As the definition of curvature is derived uniquely from properties of the planes intersecting a manifold through the normal vector, it is invariant to the orientation of the shape in the embedding space. Hence, it is an intrinsic measurement of curvature. Later these concepts will provide an insight into the geometric properties that the local geometry descriptors encode.

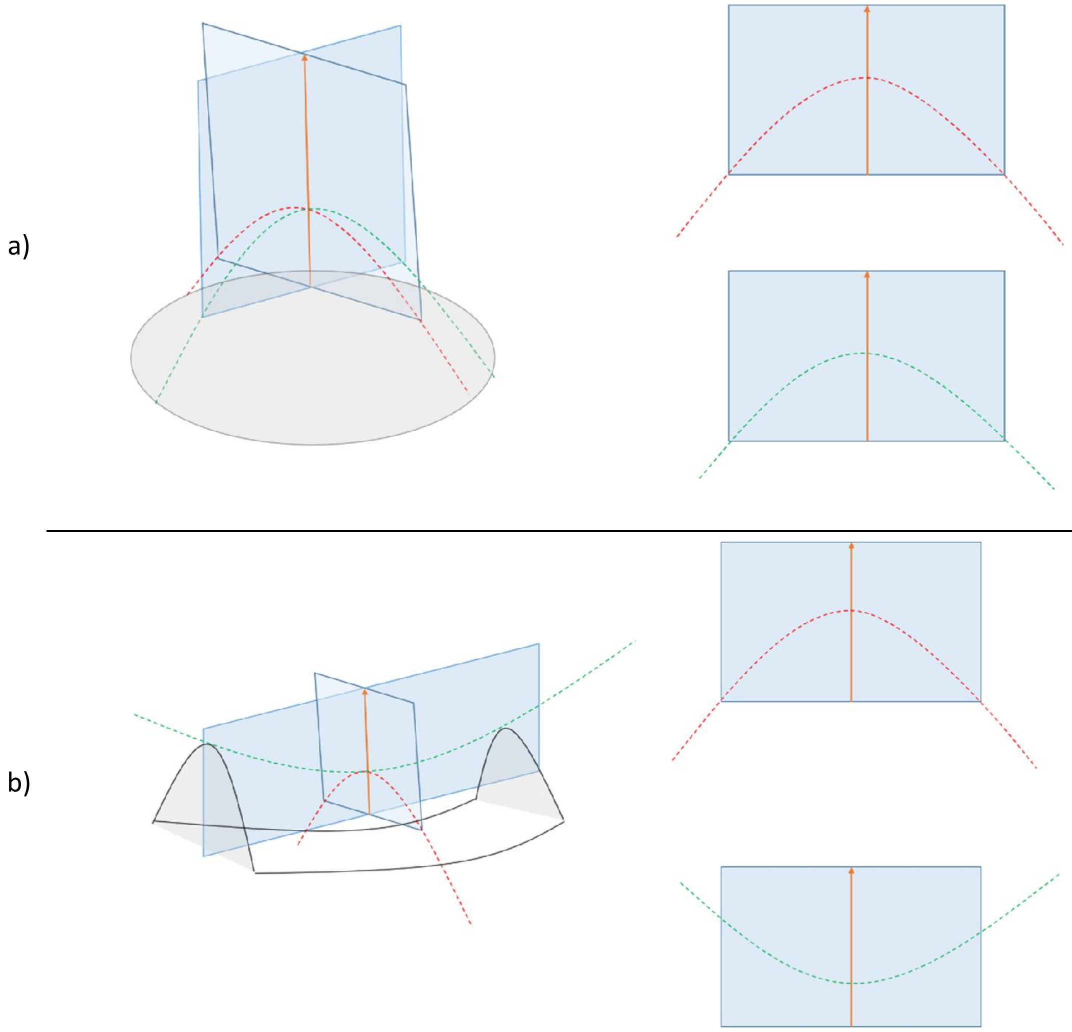


Figure 4-3. A depiction of Gaussian curvature that shows the principal curvatures of (a) positive Gaussian curvature and (b) negative Gaussian curvature.

In summary, the geometric properties that are captured by the spectrum of the Laplace-Beltrami operator are intrinsic. In particular, they are invariant to two classes of transformations: rigid Euclidean transformations, making them alignment invariant; and isometric transformations, making them invariant to a certain class of flexibility. This means that the spectrum of the Laplace-Beltrami will be the same irrespective of the alignment of the input shape and will also give the same result for an input shape that has been isometrically deformed.

4.2.3 Development of spectral geometry shape descriptors

The properties of spectral geometry described above make a convincing case for using the methodology for 3D shape descriptors. As each different shape has a unique Laplace-Beltrami operator and each Laplace-Beltrami operator has a unique eigendecomposition, it can be asserted that two shapes are the same when the eigenvalues of the Laplace-Beltrami operator are the same. In the first published application of spectral geometry to 3D shape matching, Reuter et al. (2006) took the spectrum of the Laplace-Beltrami operator over the mesh of the surface of a 3D shape and proved that the eigenvalues can be used as a descriptor of the isometric geometry of the shape. In other words, it could be viewed as the fundamental identity of a shape in the same way that DNA is used in genomics, hence they called their approach Shape-DNA. First, Equation 4-1 is restated using the traditional eigenvalue decomposition notation,

$$\Delta\phi = \lambda\phi \qquad \text{Equation 4-2}$$

where Δ is now the Laplace-Beltrami operator and λ, ϕ, denote the vector of eigenvalues and the vector of eigenfunctions respectively. Therefore, for a given spectrum given in Equation 4-2, the Shape-DNA is characterised as the k-dimensional vector,

$$SDNA = [\lambda_1, \lambda_2, \dots, \lambda_k] \qquad \text{Equation 4-3}$$

At the same time as Shape-DNA was being developed as a descriptor for deformable shapes, Coifman & Lafon (2006) used diffusion processes of graphs of a data set to provide a description of the geometry of the data set, as a form of dimensionality reduction. In this work, the term geometry is used to refer to the relationships between data points in a data set, whereby two points are connected if they are close in some feature space. The diffusion processes were used to describe a random walk over the data. In this way, a data set can be thought of as having some geometric structure whereby points that are close to each other reveal a structure that is akin to a surface and are considered unconnected from points that are distant. They noted that the spectrum of the Laplace-Beltrami operator can be used to recover this underlying geometry of the data set.

While Shape-DNA provided the intellectual cornerstone to the emerging field, it is limited in its ability to provide rich descriptions of intrinsic geometry. Building on both Reuter et al.’s and Coifman and Lafon’s work, a number of authors used the spectrum of the Laplace-Beltrami operator to compute dense point-wise *local geometry descriptors* of the intrinsic isometric geometry of a shape (Aubry, Schlickewei, & Cremers, 2011; Sun et al., 2009). These are obtained by taking the Laplace-Beltrami spectrum for a single shape and constructing a vector for each point on the surface that maps regions of the spectrum onto the point. The information captured at a single point on the surface describes the local geometry of the point and is known as a local geometric descriptor.

Sun et al. (2009) demonstrated that the eigenvalues and eigenfunctions of the Laplace-Beltrami operator are the key ingredients to the kernel solution of the heat equation, which describes the distribution of heat over a region of space at a given time. Thus, the heat transfer from one point to another over the surface of a shape can be computed using the spectrum of the Laplace-Beltrami operator. They defined the *Heat Kernel Signature*, HKS, a descriptor that describes the geometric properties at all points on a mesh. The HKS originates in the classical heat diffusion problem and describes how heat dissipates over the surface in order to capture geometric features.

The solution to the classical heat diffusion problem from one point, x, to another, y, at a given time point, t, is given as

$$h_t(x, y) = \sum_{k \geq 1} \exp(-\lambda_k t)\phi_k(x)\phi_k(y), \qquad \text{Equation 4-4}$$

where λ_k is the k^{th} eigenvalue from Equation 4-2 and $\phi_k(\cdot)$ is the k^{th} eigenfunction from Equation 4-2 evaluated at points x and y respectively. This demonstrates that the heat diffusion is governed by the spectrum of the Laplace-Beltrami operator. Specifically, the HKS uses the autodiffusive heat kernel which for a given point on the surface, x, and a given point in time, t, is,

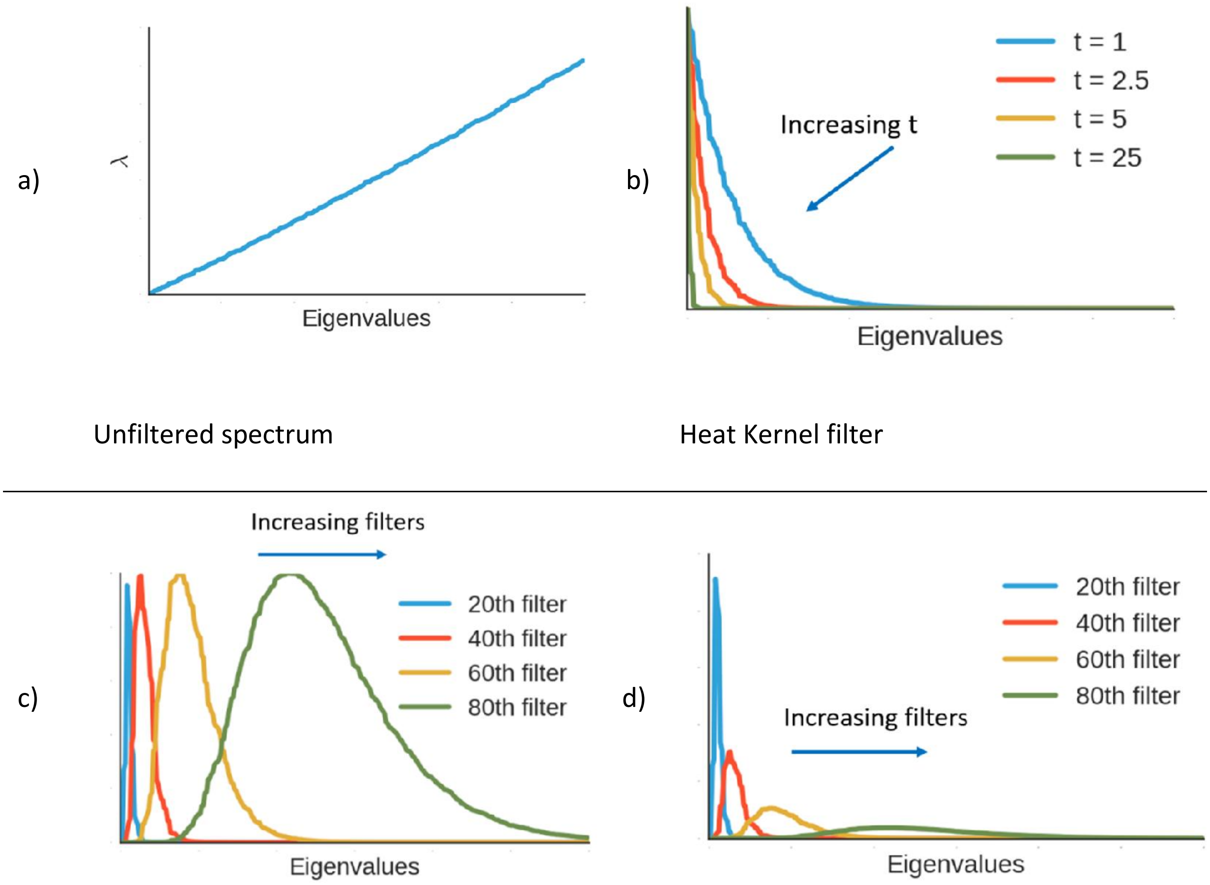
$$HKS(t,x) = \sum_{k\geq 1} \exp(-\lambda_k t)\phi_k(x)^2. \qquad \text{Equation 4-5}$$

The parameter that controls the properties of the HKS is the sample of time values used in the functional form. Each value of t corresponds to an element of the vector assigned to a point. The authors went on to prove that a dense descriptor that assigns a vector of heat kernel values for different time points was informative enough to describe point-to-point correspondences. In other words, the same points on isometrically deformed shapes would be assigned the same point descriptors.

Aubry et al. (2011) showed that the HKS is equivalent to a signal processing filter bank applied to the Laplace-Beltrami spectrum of the general form shown below.

$$f(x) = \sum_{k} \tau(\lambda_k)\phi(x)_k^2 \qquad \text{Equation 4-6}$$

where λ and ϕ are the first k eigenvalues and eigenfunctions of the Laplace-Beltrami operator respectively. The function $\tau(\lambda_k)$ is typically a transfer function that acts upon the eigenvalues and can take various functional forms of which the HKS is one. An illustration of the signal processing approach is shown in Figure 4-4. The eigenvalues have a natural ordering from small to high values. As will be illustrated later, the lower values correspond to global shape variation and higher values to local shape variation. Consequently, the eigenvalues can be plotted as an increasing linear function shown as the unfiltered spectrum (Figure 4-4 (a)). The signal filters then amplify and dampen different parts of the spectrum by transforming the eigenvalues to act as weights on the eigenfunctions. The HKS (Figure 4-4 (b)) is an example of a low-pass filter that amplifies values on the lower end of the spectrum.



Unfiltered spectrum

Heat Kernel filter

Wave Kernel filter

Figure 4-4. The eigenvalue spectrum and the transformed spectrum using the kernels from the local geometry descriptors.

Using an analogy from quantum physics, Aubry et al (2011) proposed the *Wave Kernel Signature* (WKS) as a band-pass filter that has better feature localisation properties. In the case of a discrete signal, a band-pass filter amplifies the signal over an interval and dampens signals outside of that interval. The WKS samples the spectrum by splitting it into a number of intervals that are slices of the spectrum and then a Gaussian function is centred at the middle of interval to amplify the signal around that point, which is akin to sliding the filter function along the spectrum (Figure 4-4 (c)). The earlier filters amplify global signals and the later filters amplify local geometric information. For a given point on the surface, x, and a given sample point in the spectrum, E, which is the mean point of the E^{th} interval around which the band-pass operates, the WKS is

$$WKS(E,x) = c_E \sum_{k=0}^{K} \phi_k(x)^2 f_E(\lambda_k). \quad \text{Equation 4-7}$$

where $f_E(\lambda_k)$ is the functional form of the band pass filter,

$$f_E(\lambda_k) = \exp\left(-\frac{(\log e - \log \lambda_k)^2}{2\sigma^2}\right). \quad \text{Equation 4-8}$$

and c_E is a normalising constant for each sample. In Equation 4-8, e is the mean value in the E^{th} interval so that the nominator is the squared distance of the log of the k^{th} eigenvalue from the log of the middle of the interval. The σ^2 in the denominator is an arbitrary parameter that represents the variance of the log normal distribution. Previous work has established that the value of $\sigma^2 = 7$ gives best performance (Aubry et al., 2011). In order to weight the contributions equally, a normalisation constant, c_E, is applied to give the area under each filter the same value (Figure 4-4 (d)). The result is a signature that describes a point on the surface by its contribution to both global and local intrinsic geometry. The number of intervals used to evaluate the WKS (called evaluations in the original paper), determines the number of elements in the local geometry descriptor assigned to each point.

The fundamental concept in local geometry descriptors is that the propagation of geometric features from a single point on the surface is governed by the filtered spectrum of the Laplace-Beltrami operator. Therefore, local geometry descriptors with different functional forms can be defined using different signal filters.

4.2.4 Further work in local geometry descriptors

Following Sun and Aubrey, there has been a large amount of research that has developed or applied local geometry descriptors. This has included methods to add surface information such as texture to the spectrum in addition to geometry to implicitly include the variation of surface information with the notion of shape (Kovnatsky, Bronstein, Bronstein, & Kimmel, 2012). Additionally, a large amount of research has investigated methods to learn optimal local geometry descriptors using machine learning methods. Such approaches allow the local geometry descriptor to learn non-isometric deformations in shape classes. The first applications of this cast the filter banks of the signal processing interpretation as a set of general functions that could be learnt using metric learning (Litman & Bronstein, 2014; Windheuser, Vestner, Rodolà, Triebel, & Cremers, 2014). Alternatively, a classification approach was applied to learn optimal descriptors by training a random forest algorithm to classify deformations of the shape in a particular class (Emanuele Rodolà, Bulo, Windheuser, Vestner, & Cremers, 2014). Recently, with the rise of deep neural networks, investigation has turned to using the mesh structure of the shape as the input to a deep learning neural network in order to extract intrinsic geometry descriptors. However, the underlying mathematics of the manifold approach do not allow a direct application of convolutional neural nets. This is due to the non-Euclidean metric at the heart of the local geometry descriptors. This has been overcome by either constructing topological discs over the surface and sampling from those (Boscaini et al., 2015) or by using the spectrum of the Laplace-Beltrami operator directly to modify the filters that pass over the surface (Masci, Boscaini, Bronstein, & Vandergheynst, 2015). Finally, recent work has adapted the local filters to incorporate anisotropic kernels, which are sensitive to direction (Boscaini, Masci, Rodolà, Bronstein, & Cremers, 2016) and therefore allow the local geometry descriptor to disambiguate reflection symmetries.

Another theoretical breakthrough recognised that two shapes could be compared by functional maps (Ovsjanikov, Ben-Chen, Chazal, & Guibas, 2013; Ovsjanikov, Ben-Chen, Solomon, Butscher, & Guibas, 2012). The underlying idea is that the local descriptors are functions over the surface. For two shapes that are equipped with a basis, such as the spectrum of the Laplace-Beltrami operator, there exists a mapping between the two shapes that maps the functions rather than the points. The result is a matrix that transforms the basis of one shape into the basis of another. This approach liberated shape correspondence problems from strict point-to-point requirements. Furthermore, if the mapping is the identity matrix, it can be shown that the two shapes are isometric. Therefore, the correspondence can be interpreted as a measure of how isometric two shapes are, indeed, the amount of distortion of the metric required to make them the same can be visualised (Ovsjanikov et al., 2013). These correspondences have then been used to formulate an intrinsic shape difference measure (Rustamov et al., 2013). The isometric requirement is strict, so later papers have relaxed the isometric property by coupling the bases between the shapes (Kovnatsky, Bronstein, Bronstein, Glashoff, & Kimmel, 2013). This method was applied to exploring large databases of 3D shapes (Q. Huang, Wang, & Guibas, 2014). Other research has relaxed the problem to include non-isometric shape correspondences using functional correspondence by matrix completion (Kovnatsky, Bronstein, Bresson, & Vandergheynst, 2015). These approaches mean that intrinsic shape alignment can be carried out. Additionally, sparse coding has been used to learn a permuted correspondence that can use the information in the local geometry descriptor to identify matching regions between two shapes (Pokrass, Bronstein, Bronstein, Sprechmann, & Sapiro, 2016).

Finally, an important problem is to identify coherent regions of the shape using the rich amount of intrinsic geometric information. This was first attempted using the spectrum of the Laplace-Beltrami operator (Reuter, 2009; Reuter, Biasotti, Giorgi, Patanè, & Spagnuolo, 2009). Later work used ideas from topological data analysis to identify topologically stable segments. This work used the idea of the local geometry descriptor as being a function defined over the space of the shape, represented as a manifold. This approach identified the stable regions with the same values, which is akin to finding contour lines, and segmented the shape by finding the regions contained by a contour (Skraba, Ovsjanikov, Chazal, & Guibas, 2010). The most recent approaches have used the functional correspondence matrix and structured the functional map to identify coherent segments between two shapes (E. Rodolà, Cosmo, Bronstein, Torsello, & Cremers, 2016) and within a reference shape (Litany, Rodolà, Bronstein, Bronstein, & Cremers, 2016).

4.3 Methods

This section describes the implementation of spectral and diffusion geometry methods for the analysis of 3D molecular shape. The focus in this chapter is on developing and evaluating local geometry descriptors. The aggregation of local geometry descriptors to give a global descriptor of molecular shape that can be used for virtual screening is described in Chapters 5 and 6. An overview of the workflow is given in Figure 4-5. The first step is to find a suitable representation of the shape of a 3D molecule. The second is to obtain a discrete representation of the surface using a triangular mesh. Then the spectrum of the Laplace-Beltrami operator is computed and finally, this spectrum is used to compute the local geometry descriptors.

The concepts of spectral shape are founded in the continuous world of smooth manifolds and linear operators. However, in order to apply these concepts in a computer, it is necessary to translate those principles to a discrete representation. In this respect, continuous surfaces become triangulated meshes, linear operators become matrices, and functions become vectors. This is visualised in Figure 4-5 where matrices are denoted by bold letters with their dimensions given underneath. In brief, the workflow takes a shape and obtains a triangulated mesh representation of the surface. The mesh is represented by a set of N vertices $\mathcal{V}$ (with x,y,z coordinates), and a set of faces, $\mathcal{F}$. The next step is to solve for the spectrum of Laplace-Beltrami operator over this mesh. Once the problem is represented in matrix form, the geometric properties of the Laplace-Beltrami operator can be solved using techniques in linear algebra. This can be achieved in two ways: either directly or indirectly. To compute the spectrum directly, the Laplace-Beltrami operator is estimated as an $N \times N$ matrix, where N is the number of vertices in the mesh. Typically, the estimation is carried out using the cotangent method. On the other hand, the spectrum can be computed indirectly using the finite element method, which makes the spectrum less dependent on the underlying mesh representation. In both cases, the computation is a sparse eigendecomposition that is truncated to provide the first k-eigenvalues. The spectrum obtained is a pair of objects: a k-dimensional vector of eigenvalues and an $N \times k$ matrix representing the eigenfunctions. The eigenvalues and eigenfunctions of the Laplace-Beltrami operator then form the basis of all subsequent spectral geometry analyses of the shape.

The remainder of this section will look at each step in the workflow in detail.

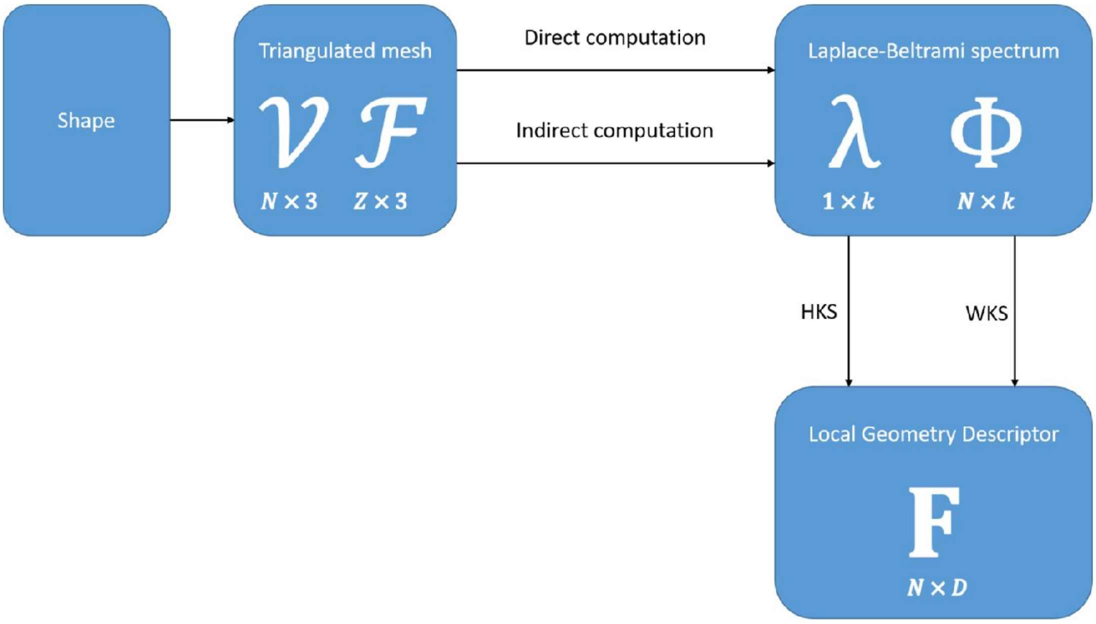

Shape
Triangulated mesh
$\mathcal{V}$ $\mathcal{F}$
$N \times 3$ $Z \times 3$
Direct computation
Indirect computation
Laplace-Beltrami spectrum
λ Φ
$1 \times k$ $N \times k$
HKS
WKS
Local Geometry Descriptor
$\mathbf{F}$
$N \times D$

Figure 4-5. An overview of the process to generate a local descriptor for a shape.

4.3.1 Definition of molecular shape

There are various notions of 3D molecular shape, from volume-centric hard sphere representations to non-volume-centric representations such as a molecular surface. In order to apply spectral geometry to 3D molecular shape a surface is necessary, and the adequacy of the surface representation is determined by two factors: first by the chemistry definition of the surface, and second by the quality of its representation as a triangular mesh.

Molecular surfaces have been predominantly used in chemoinformatics and computational biomedical science for visualisation (M. Chen & Lu, 2011, 2013), and, typically, molecular surface refers to the solvent accessible surface of a molecule. This solvent accessible surface (SAS) is the part of the molecule that a water molecule can see. In practice, it is calculated by rolling a probe, often with the van der Waals radius of a water molecule, over the molecule. Alternatively, the

Gaussian surface of a molecule is the sum of Gaussian kernel functions placed at the location of each atom. With the correct parameters, the Gaussian surface can approximate different surface types (Duncan & Olson, 1993).

In spectral geometry, a continuous surface is usually represented by a discrete mesh for computer processing (Botsch, 2010; Botsch, Pauly, Rossl, Bischoff, & Kobbelt, 2006; Grinspun, Desbrun, Polthier, Schröder, & Stern, 2006). Intuitively, the mesh can be thought of as a finite sample of the manifold to be studied. However, in order for the mesh to be an appropriate sample of the manifold for spectral geometry, some specifications are required.

A mesh is a lattice graph in 3D space composed of vertices and edges. In general, a lattice graph is a graph embedded in $\mathbb{R}^d$ space so that every vertex in the graph has a *d*-dimensional coordinate. Therefore, each vertex in a 3D mesh has (x, y, z) coordinates and a connection between two vertices is called an edge. Additionally, the vertices in a mesh are connected such that each edge forms the boundary of an enclosed region. These regions are polygons and are called faces. An illustration of a triangular mesh is given in Figure 4-6. This figure shows a mesh with six vertices, {A, B, C, D, E, F} connected by edges. The edges enclose triangular regions to form the faces, with the five faces given as {(A,B,C), (A,C,D), (A,D,E), (A,E,F), (A,F,B)}. As all edges must form the boundary of an enclosed face, it means that each edge must be a member of at least one face and that a mesh may be described entirely by its vertices and faces. Subsequently, a 3D mesh is defined by a set of N vertices, $\mathcal{V} = \{v_1, v_2, \dots, v_i, \dots, v_N\}$, and M faces, $\mathcal{F} = \{f_1, f_2, \dots, f_i, \dots, f_M\}$, where each vertex has a 3D coordinate such that $v_i \in \mathbb{R}^3, \forall\, v_i \in \mathcal{V}$ and each face is a list of three connected vertices.

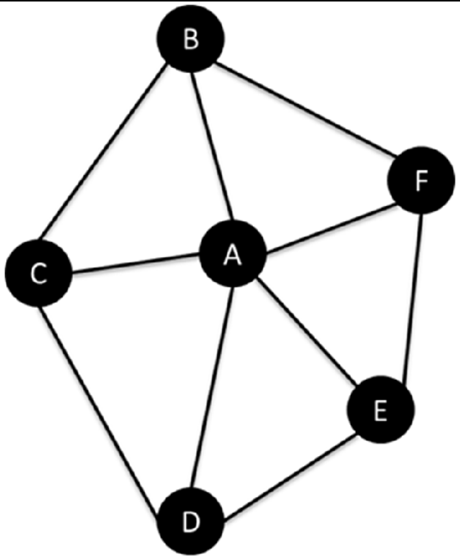


Figure 4-6. A simple triangular mesh with six vertices, {A,B,C,D,E,F} and five faces {(A,B,C), (A,C,D), (A,D,E), (A,E,F), (A,F,B)}.

For spectral geometry, the mesh must be able to represent a manifold, meaning that the mesh ought to have some sense of 'smoothness'. This means that additional requirements must be placed on the mesh. The mesh must be fully connected, meaning it must be possible to trace a path over edges between any two vertices. In principle, this ensures that a notion of distance along the mesh exists for all points on the mesh, in practice, it means that there are no parts of the mesh disconnected from other parts. Additionally, there must be a strictly positive distance between all points in the embedding space. In practice, this means that duplicate vertices cannot exist. Finally, constraints must be put on the mesh to rule out non-manifold vertices and edges. Non-manifold vertices and non-manifold edges cannot be handled by most algorithms as the geodesic behaviour around them is poorly defined (Botsch, 2010). A non-manifold vertex is one where two surfaces meet at a single point, as illustrated in Figure 4-7. A non-manifold edge is a member of more than two faces, creating a self-intersection as illustrated in Figure 4-7. Notice that a mesh may still have a boundary, that is, a collection of edges that only belong to one face. A mesh with no boundary edges is called a closed mesh.

With this in mind, TMSmesh (M. Chen & Lu, 2011; M. Chen, Tu, & Lu, 2012) was chosen to produce the molecular surface. This recent mesh generation programme was created for analytical use of meshes, in particular for solving computational chemistry systems, thus making it suitable for computing the spectrum of the Laplace-Beltrami operator. It uses an atom centred Gaussian that can be parameterised to approximate different molecular surfaces. The result is a smooth representation of the surface that has guaranteed behaviour for computing the spectrum of the Laplace-Beltrami operator. For the purposes of this experiment, the parameters were taken as those that best approximated the solvent accessible surface, with the decay value, $d = 0.4$, and the isovalue, $c = 1.2$ (M. Chen et al., 2012).

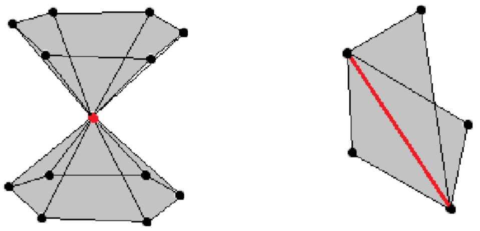


Figure 4-7. An example of a non-manifold vertex and a non-manifold edge in triangulated manifolds, where the non-manifold elements are highlighted in red. Image adapted from (Botsch et al., 2006).

4.3.2 Approximation of the Laplace-Beltrami operator on the mesh

Once a well-defined manifold mesh has been created, the next step is to obtain the spectrum of the Laplace-Beltrami operator. As mentioned above, there are two ways to do this: first, the Laplace-Beltrami operator is estimated directly, for which a number of discrete approximations of the Laplace-Beltrami operator have been proposed (Belkin, Sun, & Wang, 2008; Pinkall & Polthier, 1993; Reuter et al., 2006). The direct approach defines the Laplace-Beltrami operator as an $N \times N$ matrix, $L_{N \times N}$, where N is the number of vertices in the mesh, and weights are assigned to represent the relationship between any two vertices. In the cotangent weighting scheme, first described by Pinkall and Polthier (1993), the elements of the matrix are defined as,

$$L_{i,j} = \begin{cases} 1 & if\ i = j \\ w_{i,j} & if\ i \neq j\ \wedge\ j\ \in R(i) \\ 0 & otherwise \end{cases}$$ Equation 4-9

where i and j are vertices and $R(i)$ is the set of adjacent vertices connected to a vertex i. Weights of $w_{i,j} = 1$ for $i \neq j$ describe the connectivity graph for the mesh that encodes topology. Information on the geometry of the mesh is encoded by assigning weights to adjacent vertices as the average cotangent of the opposite angles, $w_{i,j} = \frac{1}{2}(\cot\theta_1 + \cot\theta_2)$. This is illustrated in Figure 4-8 where the formula is used to calculate the weight between vertices A and C. Notice that for a well-defined manifold mesh, where edges may only be members of two or fewer faces, this will produce a very sparse matrix with most elements being zero.

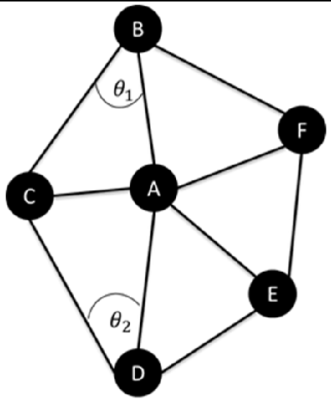


Figure 4-8. An illustration of the cotangent weighting scheme.

Then, in order to obtain the spectrum of the operator, the following system is solved numerically,

$$L\phi = \lambda\phi$$ Equation 4-10

where ϕ are the eigenfunctions with λ the corresponding eigenvalues. As noted above, this is a very sparse system as the vast majority of vertices are not connected. Therefore, the system can be solved using sparse eigendecomposition methods.

An alternative method is to compute the spectrum indirectly using the finite element method (FEM) which computes the spectrum without having to approximate it directly (Reuter et al., 2006). The details of FEM are highly technical and are given in Appendix B. In brief, the geometric properties of the mesh are captured by placing a local matrix at each vertex that encodes geometric relationships with local points through distances and cross-products and which constitutes a basis for the procedure. These local matrices are then collated to form two matrices, A and B that are the input to a generalised eigenvalue decomposition. The benefit of FEM is that it represents a smoother approximation that the direct method as it is less susceptible to noise from the mesh generation.

Both the direct and indirect approaches to evaluating the Laplace-Beltrami operator were implemented and applied to the meshes from TMSmesh. Preliminary results, that are not reported here, demonstrated the superiority of FEM, so for the rest of the work in this thesis, all spectra are computed using FEM.

4.3.3 Computation of the local geometry descriptors

Once the spectrum (eigenvalue vector and the eigenfunction matrix) has been obtained, the final step is to compute the local geometry descriptors.

In practice, the local geometry descriptors are constructed by first applying each transfer function to the eigenvalues. This operation returns a $D \times k$ matrix, T, of transformed eigenvalues where each row corresponds to an individual transfer function. The local geometry descriptor is denoted F and is then computed by a matrix multiplication of this matrix with the $N \times k$ matrix of squared eigenfunctions, ϕ^2,

$$F = \phi^2 T^\top. \quad \text{Equation 4-11}$$

This is relatively straightforward as the operation can be expressed as a matrix multiplication (Equation 4-11).

This matrix multiplication interpretation has two important interpretations that are referred to later to give insights on how spectral geometry encodes molecular shape: row-wise and column-wise. The row-wise interpretation considers the rows of the local geometry descriptor matrix. Each row is a $1 \times D$ vector that corresponds to a point on the mesh and each value in the vector is the value of the filter function evaluated at that point on the surface. Therefore, for a point x evaluated at D filter functions, $f_i(x)$ for $i \in 1 \dots D$, the row-wise local geometry descriptor of a point is

$$\boldsymbol{r}(x) = \big(f_1(x), \dots, f_D(x)\big). \quad \text{Equation 4-12}$$

In the Heat Kernel case, the values describe the heat diffusion over increasing time points, whereas in the Wave Kernel, the values describe the values at that point as the filter slides along the spectrum. This can be intuitively thought of as the descriptor of the local geometry around a specific point on the surface. Subsequently, and as will be explored in the next chapter, global descriptors that take the rows as the basic input are related to the aggregated point-wise geometry over the surface.

Conversely, the column-wise interpretation takes the columns of the local geometry descriptor matrix. Each column is a $1 \times N$ vector that corresponds to an individual filter function evaluated over the entire surface and each value of the vector is the value of that specific filter function over all points. Therefore, for the i^{th} filter function evaluated at all N points, $f_i(x_j)$ for $j \in 1 \dots N$, the column-wise geometry descriptor of a filter over the whole shape, X, is,

$$\boldsymbol{c}(X) = \big(f_i(x_1), \dots, f_i(x_N)\big). \quad \text{Equation 4-13}$$

In the Heat Kernel case, the values describe the heat diffusion over the whole surface at a specific time point, whereas in the Wave Kernel case, the values describe the values of the band pass filter over a specific part of the spectrum. Subsequently, global descriptors that take the columns as the basic input are related to local geometry variation over the entire surface.

4.3.4 Programming details

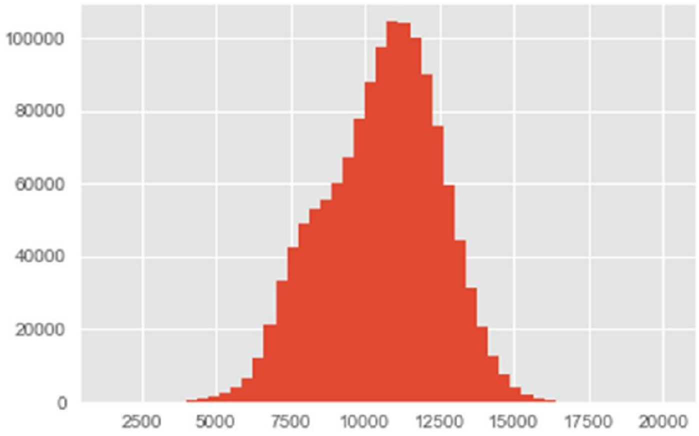
Laplace-Beltrami spectra for molecular surfaces were computed using the following workflow: first a molecular surface was calculated using TMSmesh and the mesh was converted to a list of vertices and faces that could be stored as a NumPy file. Python implementations of the cotangent method and the finite elements methods were written using SciPy and NumPy to extract the spectrum of the Laplace-Beltrami operator. In particular, the sparse generalised eigendecomposition was carried out using the underlying ARPACK routines found in LAPACK. The routines used the Implicitly Restarted Lanczos Method (Calvetti, Reichel, & Sorensen, 1994). The local geometry descriptors were computed with linear algebra operations in NumPy, which rely on underlying implementations of the BLAS and LAPACK numerical computing libraries. All code was written as Python modules for ease of use and portability. The surfaces were visualised in MayAvi, a Python library that plots 3D shapes and can also plot scalar values on the surface. Online implementations in MATLAB from the original Shape-DNA paper were used as a reference (Reuter et al., 2006).

4.4 Results

To explore the properties of spectral geometry and the local geometry descriptors when applied to 3D molecular shape, the spectra were computed for a number of molecules. The results section presents an analysis of the properties of the spectrum of the Laplace-Beltrami operator and local geometry descriptors, the HKS and the WKS. This is carried out first by visualising the properties of the spectrum and the local geometry descriptor filters on an example molecule. Then an investigation of the effect of the parameters on the properties of the descriptors is undertaken to investigate the diversity of the local geometry descriptors and how well they are preserved under conformation deformation.

4.4.1 Mesh generation of DUD-E targets

The first task was to generate the meshes for the 102 DUD-E targets, which was carried out using TMSMesh. The parameters used were given in section 4.3.2 to best approximate the solvent accessible surface decay value, $d = 0.4$, and the isovalue, $c = 1.2$ (M. Chen et al., 2012). In total 1,357,144 molecules were processed with the majority of meshes having between 10,000 and 12,000 vertices (Figure 4-9). Figure 4-9 shows the summary of the number of vertices over the data set with the mean number of vertices being 10,544, and the inter-quartile range being between 9,185 and 11,941 vertices. The smallest mesh had 1,367 vertices while the largest had 20,130 vertices. Furthermore, Table 4-1 gives a breakdown of the mean number of vertices for the meshes per target in the DUD-E data set with average molecular weight of the target data set. The data in the table are summarised in Figure 4-10 that shows average molecular weight and average number of vertices are positively correlated.



	Number of vertices
Mean	10544.47
Min	1367
25%	9185
50%	10721
75%	11941
Max	20130

Figure 4-9. Distribution and summary statistics of the number of vertices in the DUD-E data set.

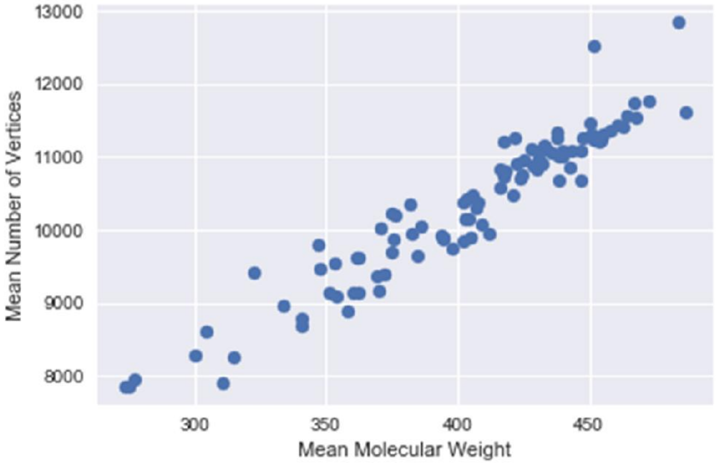


Figure 4-10. A scatter plot of the mean molecular weight and mean number of vertices for the DUD-E data set.

Table 4-1. Mean molecular weight and mean number of vertices per target in the DUD-E data set.

Target	Mean molecular weight	Mean number of vertices	Number of actives	Number of decoys
AA2AR	417.5	10728.9	845	11152
ACE	402.1	9842.6	809	6543
ACES	415.8	10846.6	665	26374
ADA	322.4	9425.5	263	5473
ADA17	451.2	11304.9	960	36648
ADRB1	417.3	11206.7	459	15959
ADRB2	421.4	11276.7	448	15256
AKT1	422.5	10907.3	424	16577
AKT2	424.4	10771.1	191	6953
ALDR	340.8	8686.4	221	9137
AMPC	295.7	1524.1	63	2903
ANDR	357.8	8876.4	524	14504
AOFB	276.9	7958.3	169	6932
BACE1	466.8	11734.5	486	18222
BRAF	438.6	11006.3	252	10099
CAH2	382.7	9951.2	836	31711
CASP3	437.8	11263.5	351	10823
CDK2	386.0	10055.4	799	28329
COMT	300.5	8287.3	87	3927
CP2C9	407.2	10315.8	184	7575
CP3A4	428.6	10885.3	364	11941
CSF1R	423.9	10708.4	287	12435
CXCR4	368.4	9996.0	123	3415
DEF	374.6	10237.8	162	5739
DHI1	384.3	9638.6	520	19624
DPP4	362.1	9622.3	1080	41374

DRD3	405.6	10482.3	878	34189
DYR	361.7	9622.7	567	17385
EGFR	433.7	11125.4	833	35443
ESR1	404.1	10142.5	628	20819
ESR2	394.2	9908.3	596	20314
FA10	467.3	11550.4	793	20418
FA7	428.2	11104.4	186	6303
FABP4	393.3	9913.8	58	2856
FAK1	432.7	11156.7	115	5403
FGFR1	451.3	12535.8	243	8700
FKB1A	430.5	11077.4	274	5833
FNTA	454.9	11308.7	1693	52050
FPPS	311.0	7893.4	214	9016
GCR	425.2	10202.0	564	15186
GLCM	347.2	9802.0	314	3838
GRIA2	353.9	9085.2	298	12062
GRIK1	315.1	8253.5	153	6618
HDAC2	382.0	10355.2	239	10367
HDAC8	376.1	10206.8	235	10515
HIVINT	372.0	9399.5	212	6757
HIVPR	472.6	11772.4	1396	36279
HMDH	446.5	11078.8	300	8885
HS90A	418.2	10809.1	126	4943
HXK4	416.3	10582.1	128	4804
IGF1R	464.4	11576.4	227	9408
INHA	347.6	9464.1	72	2319
ITAL	486.2	11623.1	234	8691
JAK2	407.3	10387.7	154	6591
KIF11	394.6	9877.0	198	6913
KIT	440.1	11088.4	253	10610
KITH	402.1	10375.9	133	2867
KPCB	438.1	10671.2	249	8845
LCK	442.9	11093.4	684	27857
LKHA4	370.8	10034.9	245	9478
MAPK2	362.2	9149.8	207	6245
MCR	404.7	9901.8	194	5241
MET	454.3	11237.1	245	11434
MK01	402.7	10161.2	140	4629
MK10	403.3	10419.7	187	6715
MK14	430.0	10824.1	916	36433
MMP13	450.0	11278.6	1039	38009
MP2K1	435.6	11068.5	243	8242
NOS1	304.3	8611.5	235	8074
NRAM	333.7	8969.4	223	6228
PA2GA	430.7	10981.6	128	5217
PARP1	350.7	9144.2	743	30430
PDE5A	439.7	11018.2	707	27827

PGH1	340.6	8797.9	252	10943
PGH2	369.2	9371.1	532	23406
PLK1	447.3	11276.2	156	6880
PNPH	273.8	7853.7	234	7017
PPARA	460.7	11434.1	545	19832
PPARD	462.7	11406.7	289	13233
PPARG	451.7	11237.2	724	25868
PRGR	360.4	9140.0	445	15815
PTN1	446.7	10691.8	226	7434
PUR2	420.8	10483.8	202	2726
PYGM	398.0	9737.2	115	4046
PYRD	369.9	9155.6	135	6649
RENI	483.9	12847.0	388	6985
ROCK1	353.4	9540.6	204	6378
RXRA	411.8	9947.3	163	7708
SAHH	275.3	7849.0	191	3484
SRC	457.6	11363.7	832	34960
TGFR1	374.6	9704.8	282	8678
THB	442.6	10848.5	169	7654
THRB	437.8	11342.9	862	27322
TRY1	425.0	10973.0	759	26220
TRYB1	450.2	11460.7	172	7714
TYSY	409.0	10077.3	312	6884
UROK	375.7	9863.6	307	9934
VGFR2	431.9	10919.6	621	25281
WEE1	453.4	11212.6	138	6235

4.4.2 Laplace-Beltrami spectra of molecular shape

The signal processing interpretation of the local geometry descriptor relies on the fact that the spectrum of the Laplace-Beltrami operator has a readily interpretable structure, with increasing eigenfunctions encoding more localised geometric features. To illustrate this structure, the spectrum was computed for a single molecule, shown in Figure 4-11, and the eigenfunctions were plotted on the surface (Figure 4-12). The colours show the variation in the values for the chosen eigenfunction with colder blue colours corresponding to the low values and warmer red colours corresponding to high values. An interesting observation is that the eigenvalues, λ_i, are ordered and are increasing in size and that, with increasing magnitude, the geometric information represented in the corresponding eigenfunctions changes. Figure 4-12 (a) shows the first eigenfunction where it is interesting to note that the colours are aligned along the longest part of the molecule, which can be thought of as the x-axis. This can be considered analogous to the first component in PCA, showing the direction of largest variation. As the eigenvalues increase in size, the corresponding eigenfunctions show smaller directions of variation. The 5th eigenfunction shows global shape variation in two principal directions. On the other hand, Figure 4-12 (c) and Figure 4-12 (d) show the 10th and 250th eigenfunctions. These show more local variation over small sections of the surface of the molecule. Therefore, the first eigenfunctions encode the largest global variations of shape over the molecule, whereas, the later functions encode more local variations of the shape. In general, the smaller eigenvalues correspond to global intrinsic geometry, with the larger eigenvalues corresponding to local geometry.

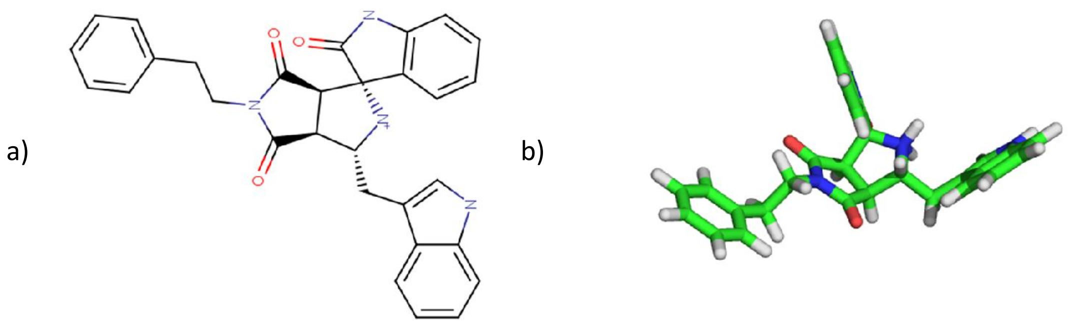
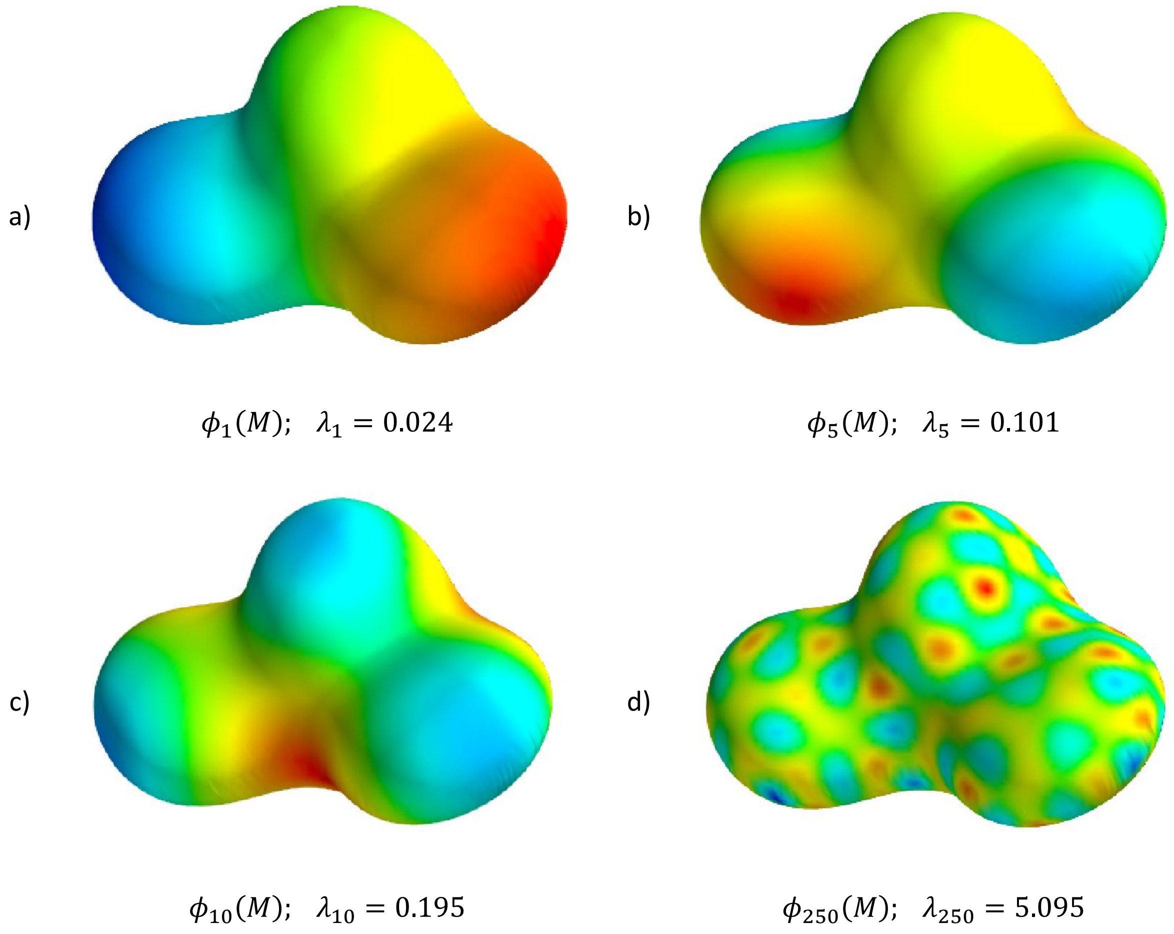


Figure 4-11. Test molecule used for visualising the properties of the spectrum and the local geometry descriptors.



$\phi_1(M);\quad \lambda_1 = 0.024$

$\phi_5(M);\quad \lambda_5 = 0.101$

$\phi_{10}(M);\quad \lambda_{10} = 0.195$

$\phi_{250}(M);\quad \lambda_{250} = 5.095$

Figure 4-12. A sample molecule from the DUD-E data set with eigenfunctions plotted over the surface along with their corresponding eigenvalues.

4.4.3 The Heat Kernel Signature (HKS) for molecules

Once a spectrum has been obtained for a molecule, the local geometry descriptors can be computed. To give an insight into how heat diffusion is related to intrinsic geometry, an illustration of the heat kernel can be seen in Figure 4-13, which plots the heat diffusion from a single point highlighted in red to the rest of the surface at time, $t = 5$. The heat transfer from one point to all others on a curved surface is determined by the spectrum of the Laplace-Beltrami operator (Equation 4-4). In particular, the heat dissipates with the curvature of the surface showing that the only geometric information comes through the spectrum of the Laplace-Beltrami operator.

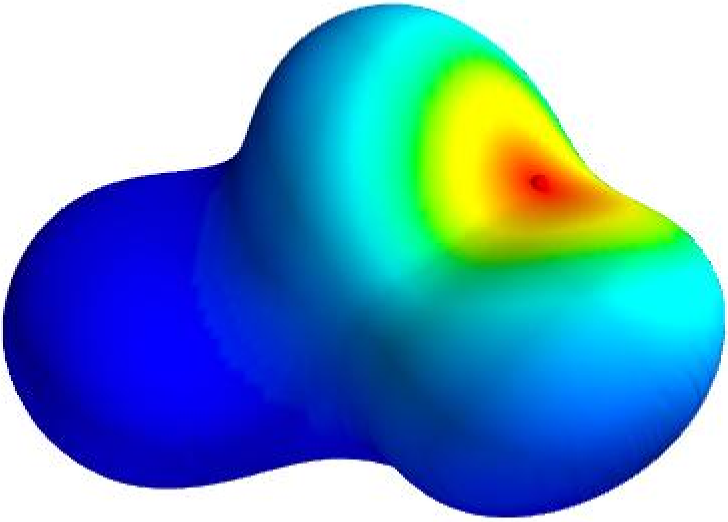
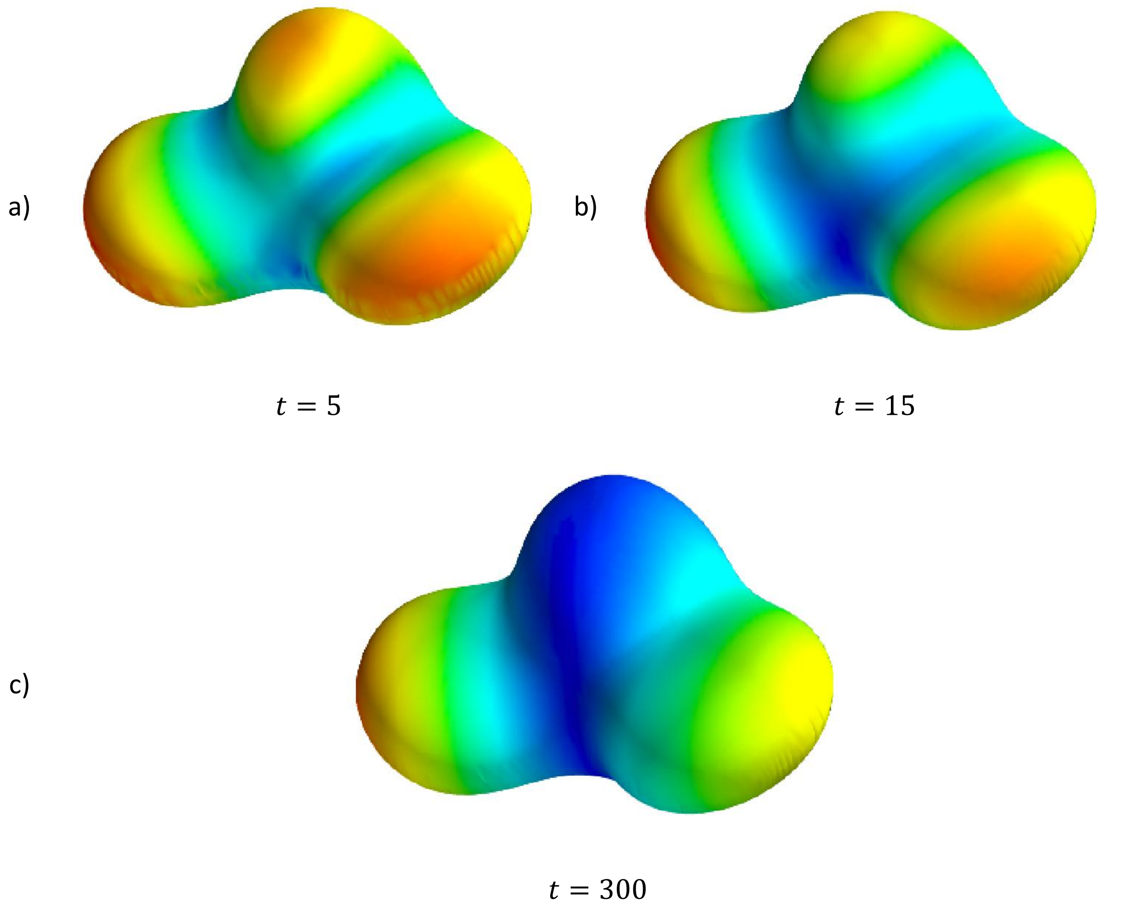


Figure 4-13. Heat transfer from one point on the surface to the rest of the shape.

The functional form of the HKS used to generate local geometry descriptors is the autodiffusive heat kernel (Equation 4-5), which measures the heat remaining at a particular point once that heat has been applied. A sample of the HKS on a molecular surface is demonstrated in Figure 4-14. Each image depicts the autodiffusive heat kernel at a single point in time. In the column-wise framework, each time point in the figure is a $1 \times N$ vector that assigns each point with the value of its autodiffusive heat kernel (Equation 4-13). Figure 4-14 (a) is for $t = 5$, which shows that at very small time points, the HKS has picked up noise from surface rendering as well as a small amount of local curvature information, with rounded cupula shapes, that is, positive Gaussian curvature, corresponding to warmer colours and valley shapes, that is, negative Gaussian curvature, corresponding to colder colours. However, noise is smoothed out at higher time points such that at time point $t = 15$ the values show a smoothed approximation of local Gaussian curvature. As time values increase, the colouring appears to encode features that are increasingly global. Figure 4-14 (c) colours the two rounded features on the left and right as warm and the colder colours extend from the valley between with a blue band in the middle. However, Figure 4-14 also shows that while at low values of t Gaussian curvature is encoded, the encoding of these geometric features is less clear at higher values of t. This behaviour can been seen in Figure 4-4 that shows the low-pass filter giving more weight to the smaller eigenvalues. In conjunction with Figure 4-12, which shows the corresponding eigenfunctions of small eigenvalues have increasingly global shape properties, it can be seen that the HKS emphasises global features. Furthermore, increasing values of the time parameter, t, result in pulling the function in towards the origin, which is shown in Figure 4-4 (b). Subsequently, with increasing values of t more global features are emphasised.



$t = 5$ $t = 15$

$t = 300$

Figure 4-14. The autodiffusive heat kernel for all vertices at three different time points.

To construct the local geometry descriptor, D time points are evaluated and collected to form the local geometry descriptor matrix. A row in this matrix corresponds to the local geometry descriptor of a point on the surface and is best interpreted as a sample of the heat dissipation at that point over time. As heat dissipation is determined by the intrinsic geometry of the surface, this forms a descriptor of the local geometry of the point. On the other hand, the local geometry properties are also defined by the autodiffusive function, which is a column-wise operation that maps the autodiffusive function over the entire surface. Therefore, the final descriptor for the shape is constructed by D time points, which are used as the columns to create a full descriptor F, Figure 4-15.

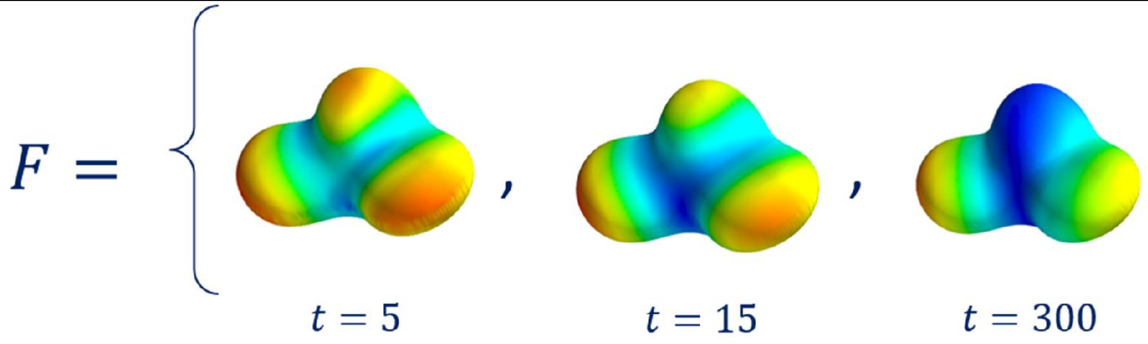
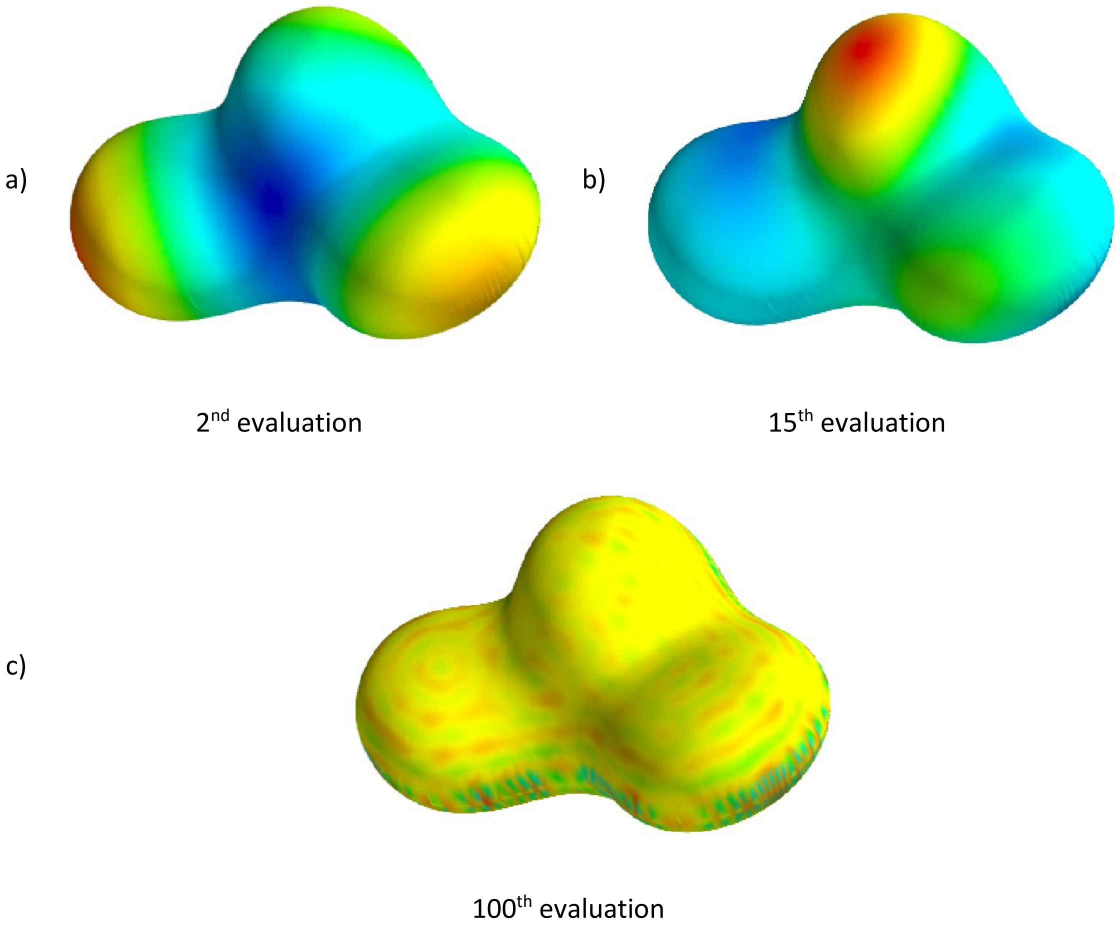


Figure 4-15. Constructing a HKS descriptor at three time points.

4.4.4 The Wave Kernel Signature (WKS) for molecules

The WKS samples the spectrum in a transparent way using the band filter approach. Figure 4-16 gives a visual illustration of the structure of the WKS. Figure 4-16 (a) shows the second filter in the WKS, that appears to encode global curvature such as Gaussian curvature and has a similar interpretation to Figure 4-14 (b). Then as the evaluations increase, more local features are encoded in Figure 4-16 (b) until the last evaluation Figure 4-16 (c) encodes local features to the extent that the variations are explained by noise and general artefacts of the mesh generation process. Again, when viewed in the light of Figure 4-4 (c) and (d) in conjunction with Figure 4-12, the way in which the band-pass filter encodes levels of features can be seen. The early evaluations in Figure 4-4 (c) and (d) amplify the lower eigenvalues whose corresponding eigenfunctions in Figure 4-12 encode global shape information whereas larger evaluations amplify larger eigenvalues that correspond to local shape features in Figure 4-12. Subsequently, with increasing filter functions in the WKS the descriptor describes increasingly local shape features.



2nd evaluation

15th evaluation

100th evaluation

Figure 4-16. The WKS filter at three different evaluations for all vertices at three different time points.

In order to construct the local geometry descriptor, the spectrum is split into D evaluations. As the whole range of the spectrum is sampled, this parameter determines how narrow in range each of the band filters will be, which is equivalent to the granularity of the sample. The D-evaluations of the WKS then form the columns of the WKS local descriptor as shown in Figure 4-17. In this case, the columns correspond to the filter mapped over the whole surface and the rows describe the local geometry around each point by recording the contribution of each point to each filter.

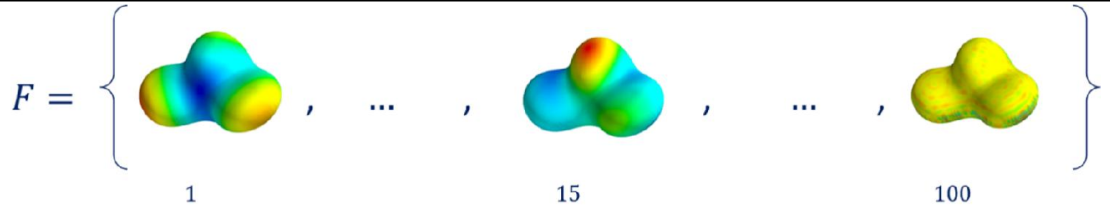


Figure 4-17. The construction of the WKS from the 1st, 15th, and 100th evaluations.

4.4.5 Filter spectrum and dimensionality

Recall the filter can be thought of as a weighting scheme that weights the contribution of the eigenfunctions in either a row-wise view (Equation 4-12), by weighting the contribution of the eigenfunctions of a particular point on the mesh, or a column-wise view (Equation 4-13), by weighting the contributions of the eigenfunctions over the surface. This section will look in detail at the properties of the filters and investigate the subsequent properties of the local geometry descriptor they create.

As the final local geometry descriptor is composed by collecting the filters over the surface as the columns of the local geometry descriptor, the choice of these D filters is an important aspect of the local geometry descriptor design. Recall from Figure 4-4 (b) that the HKS filter amplifies the lower parts of the spectrum and that increasing values of t have the effect of pulling the values in towards the origin. This would explain the behaviour of the increasing global features observed in Figure 4-14. However, the way in which the HKS encodes the geometric information in the spectrum is not clear; it only increasingly weights the lower end of the spectrum by pulling the values towards the origin. In contrast, the increasing dimensions in the WKS slide along the x-axis and separate the frequency bands (Figure 4-4 (c) and (d)), which illustrates how the filter samples the spectrum in a more transparent manner. The effect of this increasing dimensionality explains the properties of the WKS presented in Figure 4-17.

From a row-wise point of view, one important task of the local descriptor is to be able to differentiate the geometric properties of different points on the surface. The overall effect of the different filters on the local geometry descriptor for a given vertex can be seen in Figure 4-18, which shows a molecule with two points highlighted in red and green. Figure 4-18 (a) depicts the local geometry descriptor using the HKS sampled at 100 time points, $t = [1, 2, \dots, 100]$, for both the red and the green vertex respectively. Likewise, Figure 4-18 (b) depicts the WKS with 100 evaluations for the red and green vertices respectively. Therefore, both descriptors are of the same dimension, $D = 100$. The most striking difference between the two signatures is that the red and green descriptors for the HKS are very similar; this emphasises the property of the HKS that encodes global features. On the other hand, there is a large variation between the red and the green signatures of the WKS, demonstrating that the feature separation is more discriminative between the descriptors of the two vertices. Therefore, the WKS is more specific to local geometry, meaning that it will be more likely to correctly classify two vertices as being different shape features. However, the HKS will be more sensitive to local geometry meaning that it will be more likely to correctly conclude that two points have the same shape features. This sensitivity-specificity trade-off is at the heart of deciding which local geometry descriptor to use for the encoding local geometry features of a shape and is managed by the way in which the filters manage the geometric information of the spectrum.

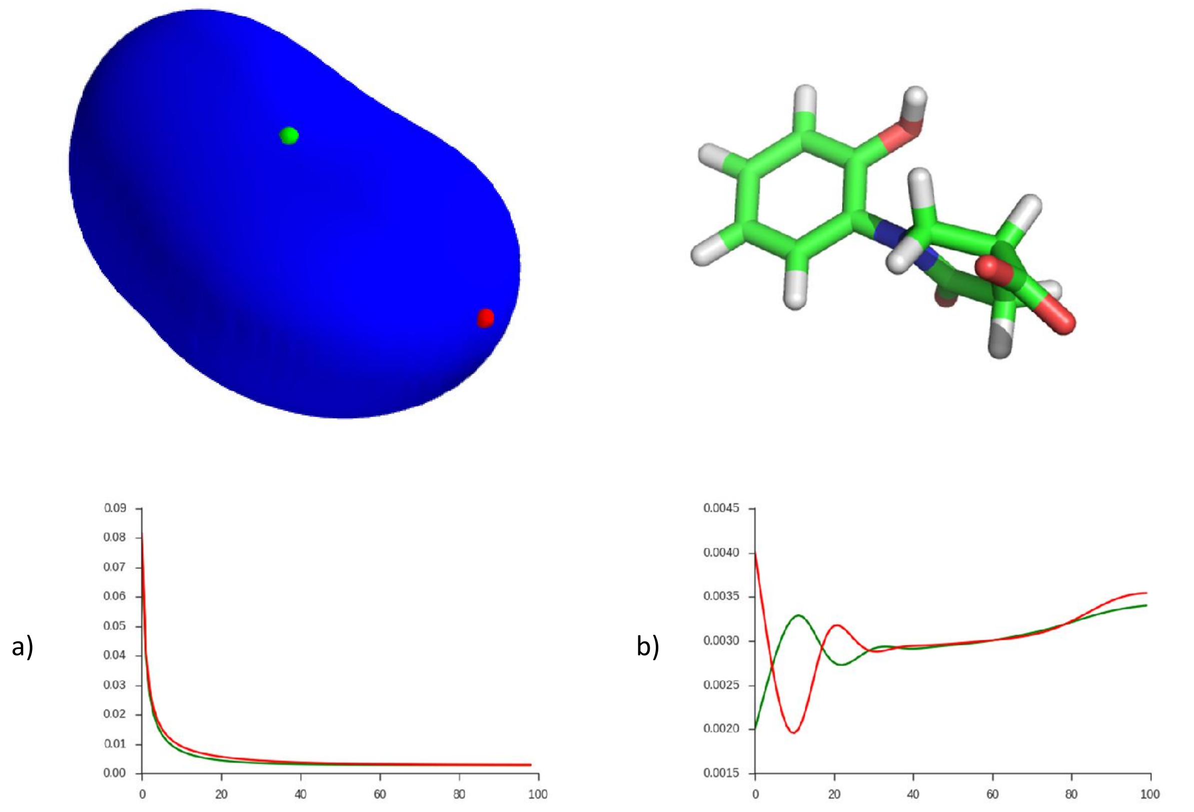


Figure 4-18. The local geometry descriptors for two vertices. (a) depicts the HKS of the points and (b) the WKS of the points, where the red and green lines correspond to the descriptors for the red and green vertices.

4.4.6 Evaluation of local geometry descriptors of molecular shape

In order to evaluate the performance of local geometry descriptors for describing molecular shapes, experiments must be designed to allow the comparison of the descriptors of individual points. Ideally, an experiment would show whether a method was sensitive to finding similar points on shapes with similar local geometries and specific enough to discriminate between points on the surface with very different local geometries. However, obtaining these points for a number of different meshes is not a trivial problem to solve. Traditionally in the field of computer vision there are data sets of shapes in different poses and that have labelled surface features such as the SHREC data set for dense correspondence finding (Bronstein et al., 2010). In chemoinformatics these features cannot be labelled automatically, nor is there a simple surface feature taxonomy, such as with faces, for example, where the nose, eyes, and ears are distinct surface features.

In lieu of being able to construct a data set of labelled points on a molecular surface, two alternative approaches are proposed (the mapping of local descriptors to global descriptors that represent whole shapes is the subject of the next chapter). First, using a row-wise interpretation, the distributions of the pairwise distances of the local descriptors for the molecule in Figure 4-11 are presented. These distributions give an insight into the diversity of the descriptors for a molecule. In principle, if the distribution is skewed towards similar values then the descriptor is not effective at discriminating between different points on the surface. Second, a qualitative approach is used whereby the quality of deformation invariance is analysed by visually comparing the values of the descriptors on different conformations of the same molecule. In this approach, a single reference point is selected on the surface of each conformer and the distance in descriptor space from all the other descriptors on the surface is plotted. This enables the visualisation of those points that have similar descriptors to the reference in the context their local geometries.

4.4.7 Distribution of pairwise distances

The distance distributions of the rows of the local geometry descriptor of a molecule were used to give a qualitative insight the specificity of the local geometry descriptors with different parameters. In this experiment, the molecule from Figure 4-11 was taken and its spectrum was computed using 300 eigenvalues. Then the local geometry descriptor matrix was computed with different parameters specifying the filter functions. Once computed, the cosine distance of the rows of the matrix was used to evaluate an $N \times N$ matrix of distance values, with lower distances corresponding to more similar descriptors. Finally, the distribution of the distance matrix was inspected. The aim was to provide a distribution with a bell-shape as this would suggest some points were very similar and some were very different with a mean distance in the centre of the distribution. This in turn would have suggested that the local descriptor was able to discriminate different points on the surface of the shape.

4.4.7.1 Heat Kernel Signature (HKS)

The key parameter of the HKS is the time points at which the kernel is evaluated. However, as shown above, while the HKS approximates Gaussian curvature at low time points, it is not clear how increasing the time range corresponds to the feature description of the spectrum. For this section, six time ranges were used which are shown in Table 4-2. The first, T_0, consists of six sample points that were found to be the optimum for deformable human shape data (Ovsjanikov et al., 2009). However, visual inspection of the values of the HKS suggested that there was little or no variation at higher time points, reflecting that molecular shape has little local geometry variation in comparison to more complex deformable shapes such as human models, therefore, smaller time ranges were also selected. Times $T_1 - T_3$ also have six elements but sample the time space up to 1500, 2500, and 700, respectively. To investigate whether performance would be substantially improved by sampling more data points, ranges up to 700 and 1000 were sampled at 1000 equally spaced points in time samples T_4 and T_5.

Table 4-2. Range of time points for sampling the HKS local geometry descriptor.

Time number	Time range
T_0	[1024, 1351, 1783, 2353, 3104, 4096]
T_1	[20, 70, 300, 500, 900, 1500]
T_2	[50, 100, 500, 1000, 2500]
T_3	[20, 70, 150, 275, 400, 700]
T_4	[1, 1.7, 2.1 , 2.8, …, 698.6, 699.3, 700]
T_5	[1, 2, 3, …, 999, 1000]

The distribution of the distance values is given in Figure 4-19. As this is a cosine distance, all values fall in the interval [0,1] where a cosine distance of 0 indicates that two vectors are the same and a value of one indicates they are orthogonal. It is important to highlight two observations: first is that the bottom axis is of a different scale for each time sample. In the case of the T_0, taken from the literature, the distances range from 0 to 7×10^{-16}, whereas in the case of T_2, the distances range from 0 to 0.035. The second observation is that all of the local descriptors have a high degree of similarity. In fact, the highest distance is 0.035, which is equivalent to similarity score of 0.965. Additionally, it can be seen that all of the distributions are highly skewed to the left, meaning that the vast majority of values are clustered around 0. Finally, the increase in the number of time samples over a range does not have a large impact on the shape of the distributions: the shape of the distribution is very similar irrespective of whether the same range has been sampled six times, as in the case of T_2 and T_3, or 1000 times, as in T_4 and T_5 respectively. In fact, the distance values are more diverse for the local descriptors with six elements, which can be seen in the range of values in the x-axis.

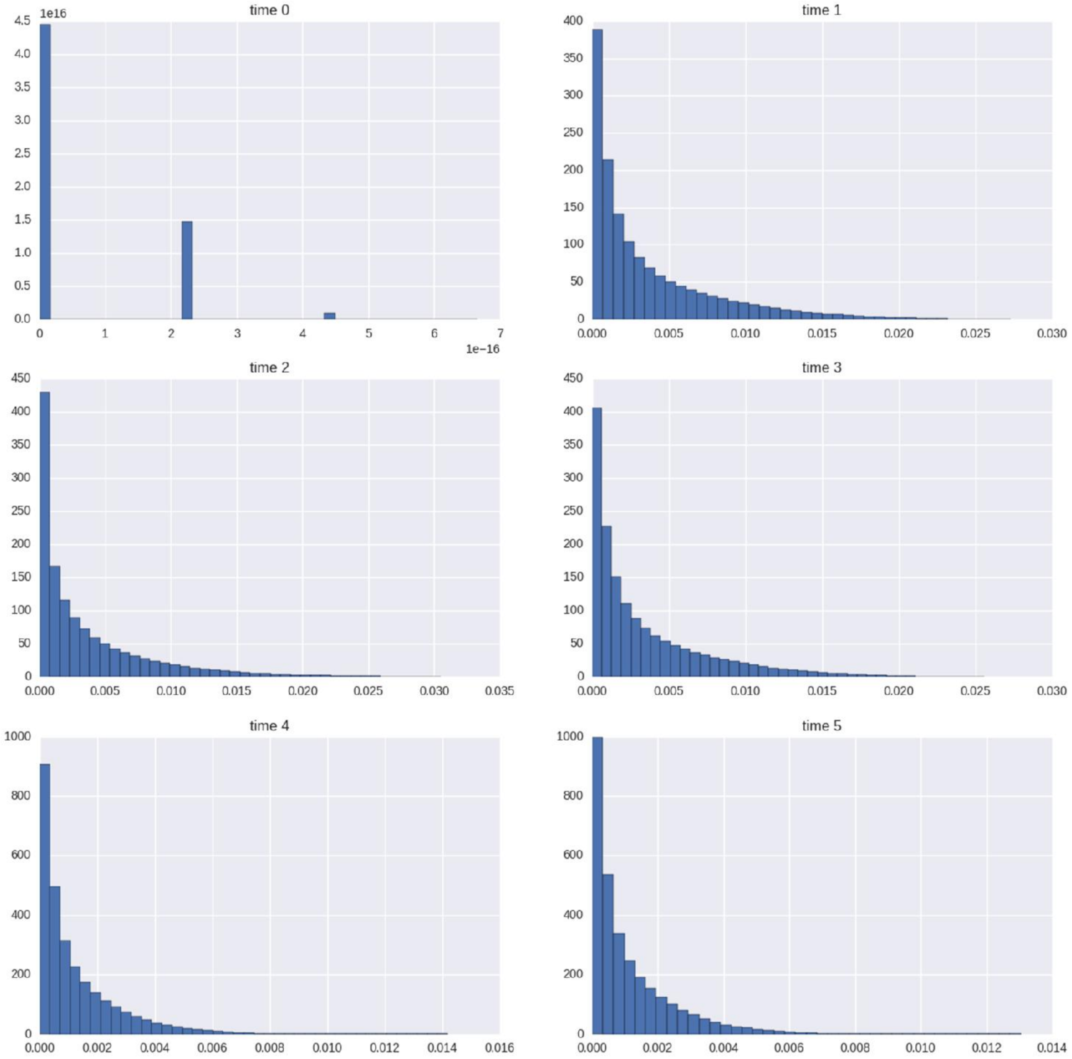


Figure 4-19. Pairwise similarity values for HKS using six different time samples for a representative molecule.

In light of the discussion above, the distributions presented in Figure 4-19 are consistent with the idea that the HKS prioritises global features rather than localisation. Subsequently, as suggested in Figure 4-18, the individual descriptors of different vertices are assigned very similar values in all filter functions, which suggests that if the descriptor were to be used to rank all the points on the surface there would be many false negatives and so it is not specific.

4.4.7.2 Wave Kernel Signature (WKS)

The key parameter for the WKS is the number of evaluations of the spectrum. The higher the number of evaluations, the more times the spectrum is sampled. As the WKS samples over the same range then choosing a higher number of D is equivalent to sampling over narrower channels so that the descriptor is more granular. In principle, this would make the local geometry descriptors more localised with respect to the individual vertices. A set of 12 values was used to test the effect of the number of evaluations, which are shown in Table 4-3. The number of evaluations directly corresponds to the dimension of the local geometry descriptor at each vertex. The intra-molecule pairwise distance distributions using the same representative molecule as Figure 4-11 are shown in Figure 4-20.

Table 4-3. Different evaluations used for WKS testing.

Parameter	Parameters tested
Evals	16, 32, 64, 100, 150, 200, 250, 300, 400, 500, 750, 1000

Again, as above, the bottom axis is of a different scale for each parameter. Also, like the HKS, the distance values are clustered around zero for the smaller number of evaluations. However, unlike the HKS, two phenomena can be observed. The first is that with an increase in the number of evaluations, the range of cosine distance values increases to 0.7 for 1000 evaluations. Also, the skew of the distribution of distance values shifts to the right with increasing evaluations, which appears to converge to a normal distribution. Therefore, at 1000 evaluations there are some local geometry descriptors that are very similar and others that are very dissimilar, the majority of the distance values are around 0.3. This suggests there is a much wider diversity of local geometry descriptors with increasing numbers of evaluations, which in turn suggests that the WKS has a higher specificity for describing the local geometry of individual vertices.

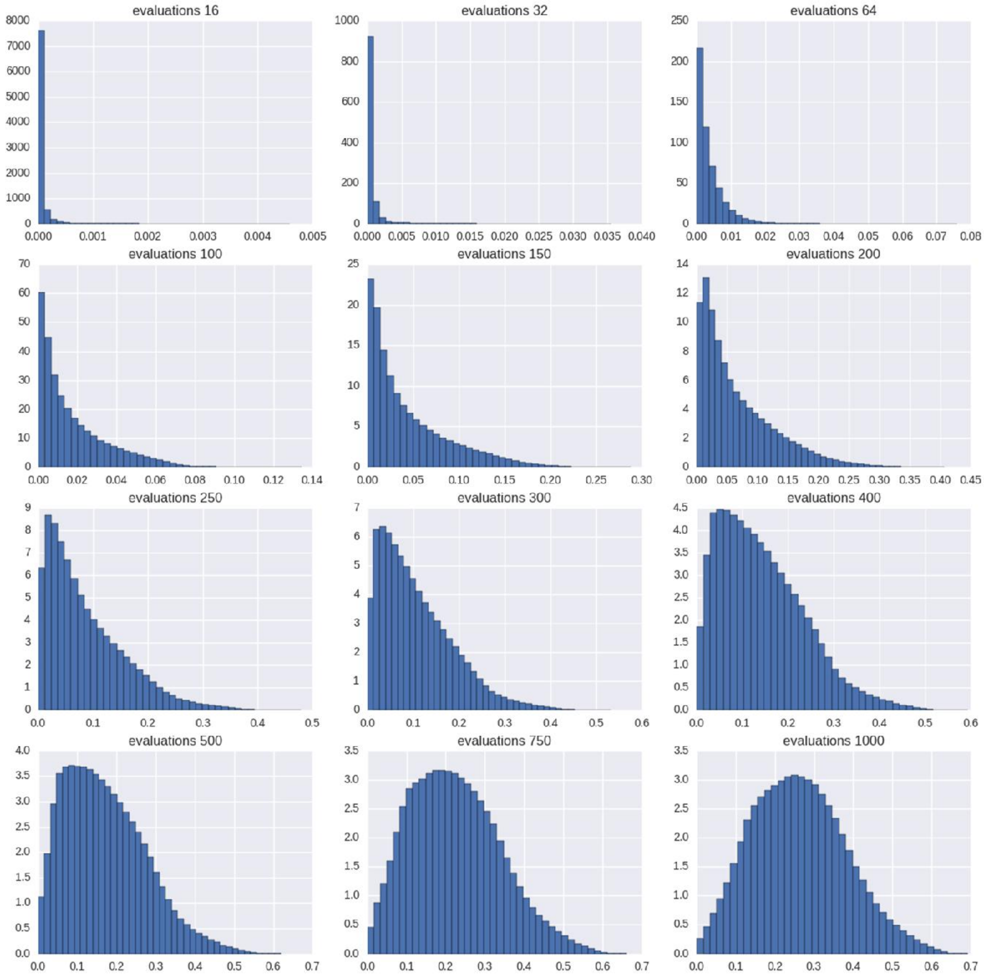
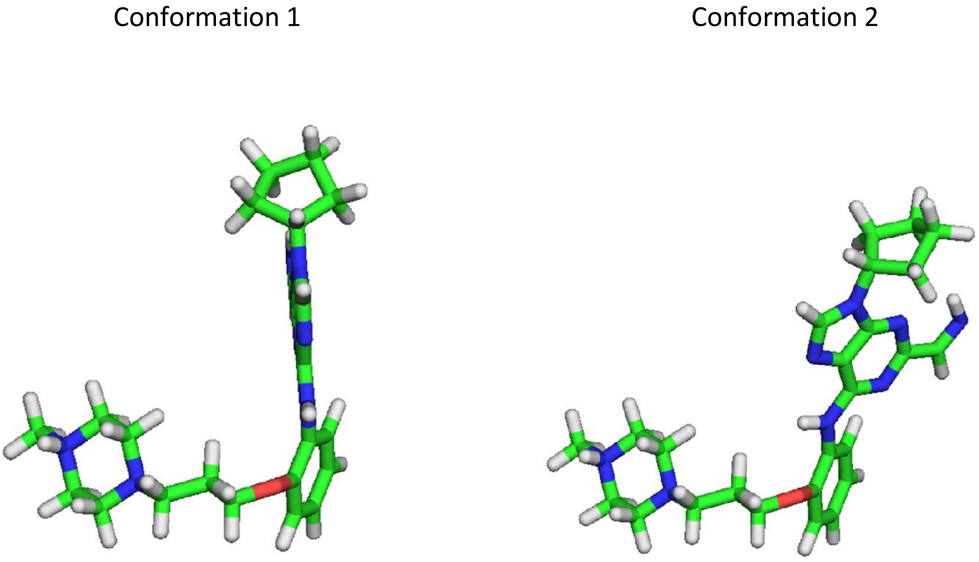


Figure 4-20. Pairwise similarity values for WKS using six different time samples for a representative molecule.

4.4.7.3 Deformation invariance

While the pairwise distance experiments give an insight into the specificity of the different descriptors, too much localisation may result in descriptors that are not sensitive to points that are similar. In other words, a similarity search of geometric features may be so localised that truly similar points on a shape may be falsely categorised as dissimilar. In order to investigate whether points on a shape with similar geometry are assigned similar descriptors, a visual inspection of the descriptors with respect to a reference point was carried out. The aim was to gain an insight into two types of sensitivity: local and global. In the local case, the cosine distances from a reference point to all other points on the surface of a molecule were plotted. The best performing descriptor was the one that assigned similar descriptors to points of similar curvature on the surface. To give a global comparison, the descriptors were computed for different conformations of the same molecule, the same reference was selected in each and the descriptor distances computed as above. Rather than look at the distance values of individual points, this visualisation gives an insight into how the local descriptors are preserved between different conformations of a single molecule. If the parameters of a local descriptor result in a descriptor that is very specific, the descriptors for points that are similar on one conformation may not be preserved on the other.

For this, a ligand (1u9x_lig_IHJ) in the P39900 target in the pharmacophore validation set was selected and 20 low energy conformations were computed. Of these, two conformations were selected, which can be seen in Figure 4-21. The difference between the two conformations is the rotation of the fused ring fragment, which is a single point of flexibility that makes the visualisation of the change between conformations clear. The surface was computed for both conformations and a vertex was selected on each to act as the reference vertex. This vertex represents the same point on the surface of both conformations. Local descriptors were computed for each conformation using four different parameters for both the HKS and the WKS. Once the local descriptors had been obtained, the cosine distance between all descriptors and the reference vertex was computed. The surfaces were then coloured based on relative surface value. In other words, the colours represent the spatial distribution of the distances of the descriptors over the surface, rather than an absolute value that is based on a consistent scale across all local descriptors. As before, the colouring is based on cosine distance with the colder colours representing descriptors that are more similar.



Conformation 1

Conformation 2

Figure 4-21. The two conformations of the ligand 1u9x_lig_IHJ from the target P39900 used for the testing of the local geometry descriptor pairwise similarity distribution.

The results are presented in Figure 4-22. Figure 4-22 (a) shows the distances for the HKS calculated using the time samples T_2, T_3, T_4, and T_5, from Table 4-2 . The reference point is the black dot on the top of the molecule, indicated by an arrow, which is in an area of positive Gaussian curvature. From Table 4-2 it can be seen that the parameters for T_2 and T_3 have only six elements, whereas, the parameters for T_4 and T_5 have 1000 elements. In general, the descriptors with six elements have a more diverse distribution of the local descriptors than the descriptors with 1000 elements, which suggests that the smaller dimension descriptors are better at discriminating between geometric features over the whole shape. The parameters for T_2 and T_3 assign similar descriptors for rounded cupula features which are coloured in blue with valley-like features coloured in warmer colours. In contrast, the parameters for T_4 and T_5 assign similar descriptors for almost all points on the surface. Secondly, there appears to be a good conservation of the spatial distribution of local descriptor distance values between the two conformations, which is more evidence for the intuition that the HKS assigns similar local geometry descriptors to all the vertices on the mesh.

Figure 4-22 (b) shows the distances for the WKS evaluated using 32, 64, 100, 500, and 1000 evaluations. In general, it can be seen that at low numbers of evaluations, such as 32, the descriptors appear to encode artefacts and noise from the mesh generation software. With respect to the local similarity properties, there is a higher diversity of local geometry descriptors with increasing number of evaluations. In fact, at 1000 evaluations, the most similar points on the surface are those that are in the immediate vicinity of the reference vertex. Furthermore, the colouring does not appear to fall into a simple Gaussian curvature explanation. This may be because other factors, such as the scale of the geometric properties around the point, or the relationship of local points to different geometries is encoded. When comparing the descriptors across the two different conformations, the colouring is preserved to a reasonable degree but less so than for the HKS. This suggests that the descriptor is more sensitive to perturbations in the global shape as this is likely to have a greater effect on local geometry features, which is in keeping with the analysis of the filters above.

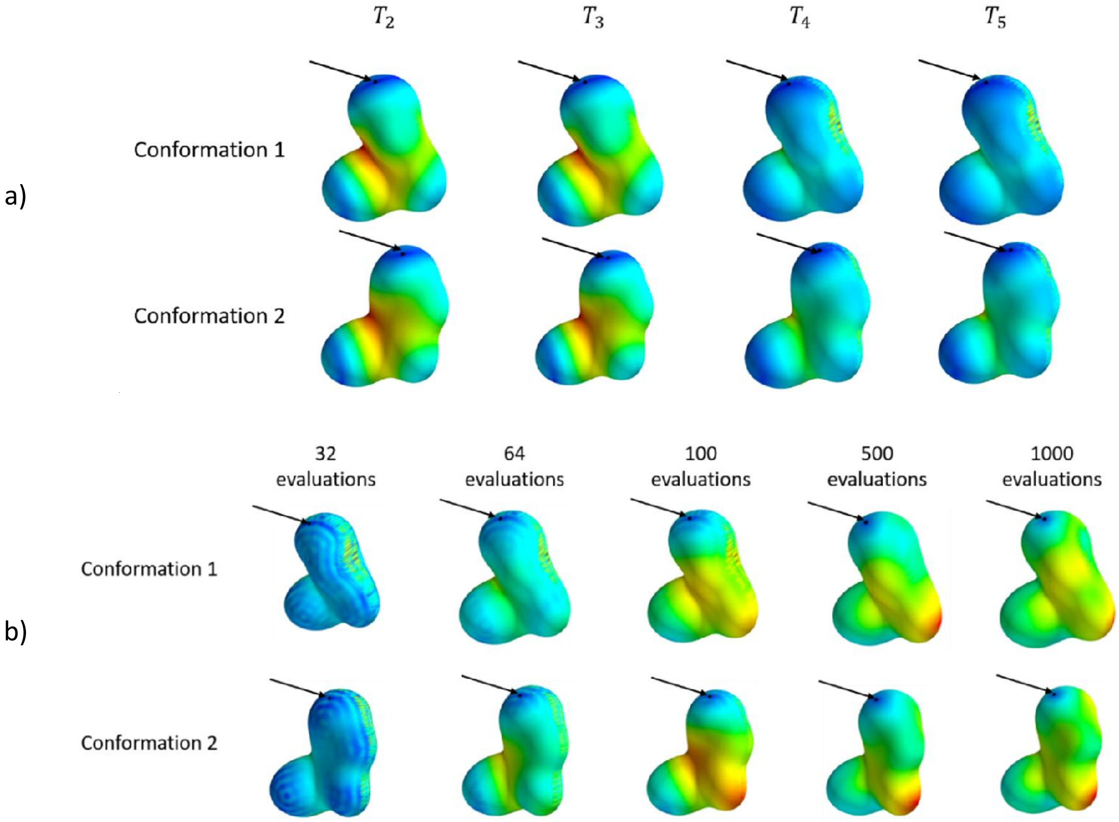


Figure 4-22. Cosine distance from a reference local geometry descriptor over the surface.

In conclusion, the distribution and the distance plotting have confirmed the behaviour of the two local geometry descriptors that was developed in the previous section. In general, optimal parameters can be selected for both methods. For the HKS, the parameters of T_2 and T_3 give the best diversity in the pairwise distances distribution analysis, which is also observed in the local similarity investigation in the distance plots. Secondly, the distance plots confirm that these global similarity properties are preserved for different conformers of the same molecule. Whereas the parameters with more values T_4 and T_5 exhibit less diversity. Therefore, parameters T_2 and T_3 are best for the HKS. In the case of the WKS, increasing the number of evaluations give the best performance in terms of specificity; the distributions are the most normal and their spatial distribution when plotted on the mesh confirms that the most similar points are the ones in the immediate vicinity. However, this is at a cost of global sensitivity where the similarities are not necessarily carried over between conformations. Therefore, there needs to be a trade-off between the higher specificity of more evaluations and higher sensitivity of lower evaluations. Qualitatively speaking, 100 evaluations appears to provide this balance. However, the optimal number may well be task specific depending on the intrinsic shape properties under investigation.

4.5 Conclusions

This chapter has presented a novel method for the description of 3D molecular shape. The central theme has been defining a 3D shape as a 2D manifold embedded in a 3D space and deriving a point-wise descriptor of local geometry on the surface. This framework implies that the shape is defined independently of the embedding space and that the manifold may have a number of poses realised in 3D space. In using this framework, the notion of 3D molecular shape has been decoupled from a rigid body conformation.

Central to the reformulation has been the spectrum of the Laplace-Beltrami Operator, which encodes the intrinsic geometric properties of the manifold. In other words, the spectrum of the Laplace-Beltrami operator is the same irrespective of the initial orientation or alignment of the 3D shape, which is a highly desirable property for 3D shape comparisons of molecules. Using these properties of the spectrum of the Laplace-Beltrami operator means that alignment invariant, local geometry descriptors can be constructed using the spectrum. In particular, local geometry descriptors use the spectrum to describe a rich amount of intrinsic geometry. The underlying framework being a bank of filter functions that modify the spectrum to amplify desired properties of the geometry. These descriptors have an elegant linear algebra representation as a matrix multiplication that allows their properties to be interpreted in terms of rows and columns.

For the purposes of this chapter, two existing local geometry descriptors have been analysed: the HKS and the WKS. The HKS uses heat diffusion as an analogy but, in essence, is a low-pass filter on the spectrum that amplifies global features. Consequently, the properties of the descriptor mean that there is poor localisation of individual descriptors, whereas the WKS, which implements a band pass filter, delivers much better feature localisation. The increasing dimensions of the WKS sample the spectrum from a global to an increasingly local sense and are weighted to balance the relative contributions.

In two qualitative experiments, the optimal parameters of the local geometry descriptors were investigated. These showed that the HKS is not diverse when applied to small molecules which would likely result in low specificity. However, the descriptor was able to describe similar points both locally and across conformations suggesting that it successfully describes points with similar geometry with the optimal parameters being T_2 and T_3 from Table 4-2. Nevertheless, the properties of the low pass filter mean that almost all points were assigned highly similar descriptors and exhibit poor feature localisation. In contrast, the WKS exhibited more diversity of descriptors suggesting that the descriptor is considerably more specific than the HKS. However, the descriptors gave good feature localisation at the cost of global features. The optimal performance was seen at 100 evaluations but the best granularity was seen with 1000 evaluations. The sensitivity-specificity trade-off is therefore a vital choice to make in the design workflow for a descriptor and will depend upon the desired properties of the task at hand. The strength of the descriptors presented in this chapter is that this trade-off can be investigated clearly and the properties are well understood. Chapter 5 will investigate the task-specific parameters for the optimal local geometry descriptors.

5 Parameterisation of the local geometry descriptor for virtual screening

5.1 Introduction

As discussed in the previous chapter, there are a number of important parameters to choose when designing a local geometry descriptor for molecular shape comparison. These are the number of k-eigenvalues, the choice of the functional form of the local geometry descriptor, and the corresponding parameter for the D-filters. Additionally, there is no canonical labelling system of the molecular surface. Furthermore, the properties of the filter banks and the sensitivity-specificity trade-off mean that the optimal parameters are likely to be task specific.

Virtual screening experiments provide a good opportunity to evaluate quantitatively how the local descriptors perform for a global similarity task. While the previous chapter evaluated the local geometry descriptors qualitatively for sensitivity and specificity of the individual vertex descriptors, the goal of this chapter is to evaluate the parameters of the local geometry descriptors for the purpose of global shape matching tasks, such as virtual screening. A further complication is that the local geometry descriptors of two shapes themselves cannot be compared directly. Mathematically, this is a non-trivial task as the local geometry descriptor matrices represent either rows of point descriptors or columns of filter functions. However, even if it were possible to directly compare the two matrices there are a couple of problems: first, there is a different number of vertices in each mesh representation, and second, there is not a canonical ordering of vertices that would allow the comparison of point descriptors for the *same* vertices on each shape. In order to perform such a comparison, there must be a framework imposed that would allow the comparison of the matrices to be feasible.

Global geometry descriptors are mappings of the local geometry descriptors to a global descriptor space in such a way that they can be compared using a similarity metric. The simplest way of mapping from local geometry descriptors is to calculate the covariance matrix over the columns. This requires minimal parameterisation and therefore allows the focus to be upon the optimisation of the local geometry descriptors for the use in global shape matching tasks.

It is not necessarily clear how the sensitivity and specificity of the local geometry descriptors will affect the performance of global geometry descriptors. On the one hand, a local descriptor that is more sensitive will assign similar local geometry vector descriptors to the majority of vertices, which when aggregated might improve global similarity performance. On the other hand, local geometry descriptors that are more specific may perform better when aggregated as global geometry descriptors as they would emphasise local geometry variation over the shape and provide better discrimination between shapes that have different local structures. The important theme is how the mapping to a global descriptor manages these sensitivity and specificity properties.

The next section introduces the covariance descriptors and evaluates the best local geometry descriptor parameters for virtual screening performance. In particular there is an investigation into the optimal number of truncated eigenvalues of the spectrum, the performance of the two different functional forms of the local geometry descriptor (i.e., the Heat Kernel Signature (HKS) and the Wave Kernel Signature (WKS)) and the corresponding optimal choices of the D-filters. The performance of the local geometry descriptors is then evaluated using a virtual screening experiment assessed by three metrics: AUC, BEDROC and enrichment factor 0.5%.

5.2 Covariance descriptors

The task of developing a global geometry descriptor for a molecule can be described as mapping the local geometry descriptors with N vertices and D dimensions from their native space $\mathbb{R}^{N \times D}$ to a global descriptor space. Once in the global descriptor space, the similarity of two global geometry descriptors can be evaluated using a suitable distance metric.

One way of mapping the local geometric descriptors to a common space is to use the covariance of the features over patches on the shape (Tuzel, Porikli, & Meer, 2006). For a local geometry descriptor $X \in \mathbb{R}^{N\times D}$, the columns of the descriptor are frequency channels of a filtered signal over the shape, which form the features of the shape descriptor. Therefore, using the terminology of Chapter 4, the covariance descriptor uses the column-wise interpretation of the local geometry descriptor to map to a global descriptor. A patch is a subsection of the object to be described. In an image processing framework, a patch is a collection of pixels. In the case of the mesh representation of the molecular surface, a patch is a vertex and its nearest neighbours to form a connected subset of the mesh. Intuitively as the descriptor describes the variation over the frequency channels, it encodes some information on spatial variation over the shape.

The covariance descriptor then takes a patch of the shape and computes how the features in the patch vary with respect to each other. For a local geometry descriptor with D-columns and a patch that is a connected subset of the mesh, $m \subseteq M$, the covariance descriptor is a $D \times D$ matrix calculated as,

$$C_m = \int_m (m-\mu)(m-\mu)^T dm, \qquad \text{Equation 5-1}$$

where μ is a vector of means of the features, $\mu = \int m\, dm$. In other words, the local geometry descriptor will always project on to the same size global descriptor irrespective of the number of rows.

To see the significance of the projection, let the $\overline{m} = (m-\mu)$ be the mean centred patch, so that $\overline{m_i} = (m_i - \mu_i)$ then,

$$C_m = \begin{pmatrix} \langle \overline{m}_1, \overline{m}_1 \rangle & \langle \overline{m}_1, \overline{m}_2 \rangle & \cdots & \langle \overline{m}_1, \overline{m}_D \rangle \\ \langle \overline{m}_2, \overline{m}_1 \rangle & \langle \overline{m}_2, \overline{m}_2 \rangle & \cdots & \langle \overline{m}_2, \overline{m}_D \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \overline{m}_D, \overline{m}_1 \rangle & \langle \overline{m}_D, \overline{m}_2 \rangle & \cdots & \langle \overline{m}_D, \overline{m}_D \rangle \end{pmatrix}, \qquad \text{Equation 5-2}$$

where m_j is the j^{th} feature column in the patch m.

The covariance descriptors are then computed for M patches over the shape and the final descriptor of the shape is then the average of all the covariance patches,

$$C = \frac{1}{M}\sum_{m=1}^{M} C_m. \qquad \text{Equation 5-3}$$

Thus, the global geometry descriptor can be computed as the weighted sum of the patches around each vertex. Larger patches can also be used, for example, the 2-ring, 3-ring, up to the N-ring. Ultimately the largest patch is the one where a vertex is connected to all vertices in the mesh, which is equivalent to the entire shape.

The advantage of the patch-approach is that for suitably large patches, there will be a balance of information from the segments, or fragments of the whole shape. However, a small patch is likely to have small variation that is unlikely to be informative, especially in a dense mesh, as the local descriptors around a small patch are likely to be very similar. Selecting larger patches increases computation time as it requires the computation of connected subgraphs in the mesh. Therefore, there is a trade-off between patch size and computational efficiency, with the exception of using a single patch: the whole shape.

For the purpose of the global molecular shape descriptors described here, the whole shape is input as a single patch.

5.2.1 Similarity of covariance descriptors

The strength of the covariance descriptor is that it maps all D-dimension local geometry descriptors to the same space: the space of $D \times D$ covariance matrices. However, covariance matrices cannot be compared directly using traditional descriptor comparison metrics such as the cosine distance. This is due to them belonging to the group of symmetric positive semi-definite matrices that lie on a Riemannian manifold such that they cannot be compared using a Euclidean metric. An alternative approach is to stack the columns of the covariance matrices and compare the D^2 dimension vectors (Masci et al., 2015). From virtual screening tests not presented in this thesis on the AMPC target, it was found that the best performance for this method of comparison came from using the Bray-Curtis metric,

$$d(C_1, C_2) = \frac{\sum_i |c_{1,i} - c_{2,i}|}{\sum_i |c_{1,i} + c_{2,i}|}, \qquad \text{Equation 5-4}$$

which is likely to be because it is an element by element comparison, as opposed to a dot-product based vector distance approach. The important information comparison is how one covariant in a covariance matrix differs to its corresponding covariate in the other matrix.

5.3 Parameters to be tested

Figure 5-1 shows the workflow to produce a global geometry descriptor from a local geometry descriptor with the optimisation steps highlighted in red. The first decision is to choose the optimal number of eigenvalues. The choice of the truncation point of the spectrum determines the properties of the spectrum that underlies all further work with the local geometry descriptors. The more eigenvalues that are included in the descriptor, the more granularity can be encoded into the descriptor. In theory, there are infinite numbers of eigenvalues as the Laplace-Beltrami operator is a linear operator. However, in practice it is limited to the number of vertices in the mesh. Nevertheless, even this value is too high as the majority of the geometric information is captured at the beginning of the spectrum. Recall from the previous chapter that the spectrum has a structure that shows the most global information at the beginning with the variation becoming increasingly local as k increases so the contribution of geometric information has diminishing returns. Therefore, beyond some number of eigenvalues the additional information on the geometry will be so local as to add little or no value to the descriptor. This step is also crucial for the computational efficiency of calculating the spectrum. In particular, the algorithm is quadratic with respect to the number of vertices. Therefore, the best value of k is a trade-off between the best virtual screening performance and the lowest computation time.

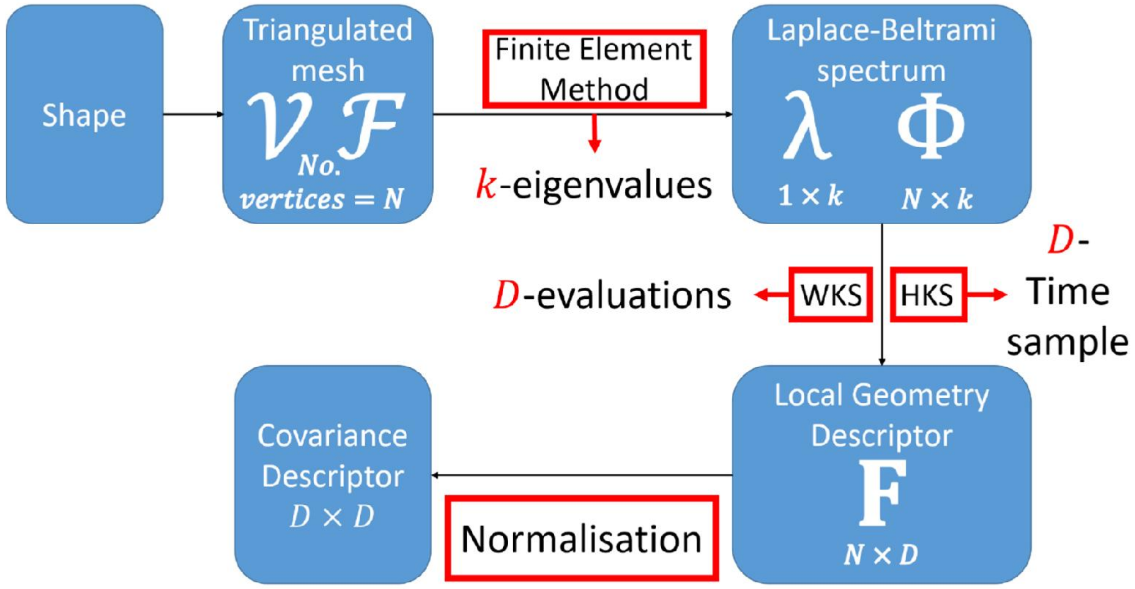


Figure 5-1. Workflow to produce a local geometry descriptor for virtual screening that require parameterisation.

The next optimisation step depends upon the choice of the functional form of the local geometry descriptor. First, the best performing parameters are found for each of the HKS and the WKS and their relative performance is compared. The parameters investigated for HKS are those in the previous chapter, which are summarised in Table 5-1. In the qualitative assessment of the sensitivity and specify of Chapter 4 it was observed that the larger dimensions (T_4 and T_5) had less specificity and encoded most vertices as being similar. However, in a global framework the additional information of more time points to sample might give further information that may allow the global descriptor to distinguish better between whole shapes. If this is the case, then they will have better performance in a virtual screening experiment.

Table 5-1. HKS and WKS parameters to be tested.

Time Sample	Time range
T_0	[1024, 1351, 1783, 2353, 3104, 4096]
T_1	[20, 70, 300, 500, 900, 1500]
T_2	[50, 100, 500, 1000, 2500]
T_3	[20, 70, 150, 275, 400, 700]
T_4	[1, 1.7, 2.4, 3.1, …, 999.3, 700]
T_5	[1, 2, 3, …, 999, 1000]

Likewise, the parameters of the WKS that are tested here are the same as those in the previous chapter, which are summarised in Table 5-2. In contrast to the HKS, where all the individual time points need to be enumerated, all that is required in the WKS is to specify the number of evaluations, which in turn equates to the number of evenly spaced band filters on the spectrum. The qualitative assessment of local vertex similarity in the previous chapter suggested that with increasing dimensions the WKS was increasingly localised, so that the preservation of descriptors for similar points on the same shape or the same point on a different pose of the same shape was compromised. Therefore, it would be expected that this would have a detrimental impact on global shape similarity as increasing dimensions promote local features at the expense of global features.

Table 5-2. Parameters to be tested for the WKS.

Parameter	Parameters tested
D	32, 64, 100, 150, 200, 250, 300, 400, 500, 750, 1000

A final experiment was carried out to investigate the effect of row-wise normalisation of the local geometry descriptor. In particular, ℓ^2 and ℓ^2 norm weightings were applied to the rows of the local geometry descriptor before computing the covariance matrix. These weightings are defined in detail in 6.2.4. When comparing two vectors using a dot product based metric such as the cosine distance, the length of the vector may have an unwanted influence. A pre-processing step can then be used to ensure that the vectors are of the same unit length. The choice of vector length depends upon the norm metric used, for the case of this chapter, only ℓ^1 and ℓ^2 metrics were used.

5.4 Virtual screening experimental set up

The DUD-E data set was used to test the performance of the descriptors for virtual screening. Initially the experiments were carried out on a subset of the data set with targets being selected from the diverse subset taken from the website ('DUD-E diverse subset', n.d.). An overview of the targets used for profiling is presented in Table 5-3. Twenty active molecules were selected at random for each target and virtual screening experiments carried out using each compound as a query. The AUC and BEDROC, $\alpha = 20$, values are reported for each target. For target-specific comparisons, enrichment factors at the 0.5% level are also reported. The virtual screening results are presented as point graphs with lines indicating the 95% confidence intervals to give an illustration of the variation of the results. In general, the confidence intervals are large due to the comparatively small sample sizes, $N = 20$, used in the experiments.

Table 5-3. Profiling targets taken from the DUD-E data set.

Target	Number of molecules	Number of actives	Percentage of actives
AMPC	2964	62	2.1%
CXCR4	3536	122	3.5%
GCR	15183	563	3.7%

The parameterisation experiments presented here are preliminary results as a proof-of-concept of the descriptors. Due to the long computation time that is compounded by the large number of molecules in the data set, it was not feasible to carry out a more detailed set of experiments. Consequently, the parameterisation experiments are carried out on three targets. The computation time is further compounded for the eigenvalue decomposition with a large number of eigenvalues, subsequently the parameterisation of the number of eigenvalues, k, uses two targets. Therefore, there is a risk that the results presented are not generalisable.

5.5 Results

5.5.1 Parameterisation of the covariance descriptor

Before testing the parameters of the local geometry descriptor, a virtual screening experiment was carried out to determine the best patch size for the covariance descriptor. The experiment was carried out on the AMPC target of the DUD-E data set. The results for the first five patches are presented in Table 5-4, where a ring size of ALL refers to the whole molecule being taken as a single patch.

Table 5-4. Impact of ring size on virtual screen of AMPC for covariance descriptor.

Ring size	AUC	enrichment 0.5%	enrichment 1%	enrichment 5%
1	0.61	4.94	3.28	1.71
2	0.61	5.67	3.56	1.68
3	0.62	6.50	4.24	1.68
4	0.62	6.71	4.37	1.85
5	0.62	6.71	3.98	1.96
All	0.62	6.56	3.98	1.79

Table 5-4 shows that enrichment performance increases with ring size for all measurements, with the exception of ring size 5 whereby performance dips for enrichment factor at 1%. These results are consistent with increasing ring sizes having increased region description. Interestingly the whole molecule patch delivers a good performance on its own, with only ring sizes of 3, 4 and 5 outperforming it. This observation suggests that there is sufficient informative geometric information being captured across the filter channel of the entire shape. Additionally, finding the rings on a mesh is a computationally expensive operation that requires a search for the k-nearest neighbours in a graph. Therefore, there is a trade-off of virtual screening performance with computational efficiency. A covariance descriptor calculation for the whole molecule took approximately 0.8 seconds to compute, whereas the calculation using a ring size of 5 took 25 minutes of computation time. Consequently, the whole molecule patch for covariance descriptor was chosen for the rest of the chapter.

5.5.2 Parameterisation of the Local Geometry Descriptors for Virtual Screening

5.5.2.1 Number of eigenvalues

To determine the optimal number of eigenvalues to form the truncated spectrum a virtual screening experiment was carried out on the AMPC and CXCR4 targets with the WKS as the benchmark local geometry descriptor. The number of evaluations was $D = 100$ and there was no additional normalisation. Figure 5-2 shows the results of the virtual screening experiment for increasing values of k. Figure 5-2 (a) – (c) show the target level performance for AMPC and CXCR4. The results are presented as point plots with the 95% confidence intervals that have been computed using 10,000 bootstrap iterations. The value k has little or no effect for AMPC; the average AUC performance increases narrowly with increasing k and the average enrichment factor is marginally best at $k = 100$. The BEDROC score, which can be interpreted as a weighted statistic of these two measures, shows that $k = 100$ produces marginally better performance on average but this is inconclusive given the 95% confidence intervals. On the other hand, there is a large variation in performance for CXCR4. Figure 5-2 (a) shows that the AUC values of $k = 50$ and $k = 100$ are significantly better than at higher values and (b) shows that $k = 100$ produces enrichment factors that are significantly higher than the other values at the 95% confidence level. The weighted BEDROC results shows that $k \leq 100$ performed significantly better than higher values of k at the 95% confidence interval with $k = 100$ performing better on average. Finally, Figure 5-2 (d) and (e) show the combined results for both targets using AUC and BEDROC respectively. The results show that across both targets, $k \leq 100$ performed significantly better than higher values of k at the 95% confidence interval with $k = 50$ performing best on average for AUC and $k = 100$ performing better on average for BEDROC indicating that there is no performance gain for using higher numbers of eigenvalues. This may be due in increased encoding of noise and mesh generation artefacts from using increasingly localised features at higher eigenvalues. In order to give best performance for early enrichment, $k = 100$ was used for all future experiments.

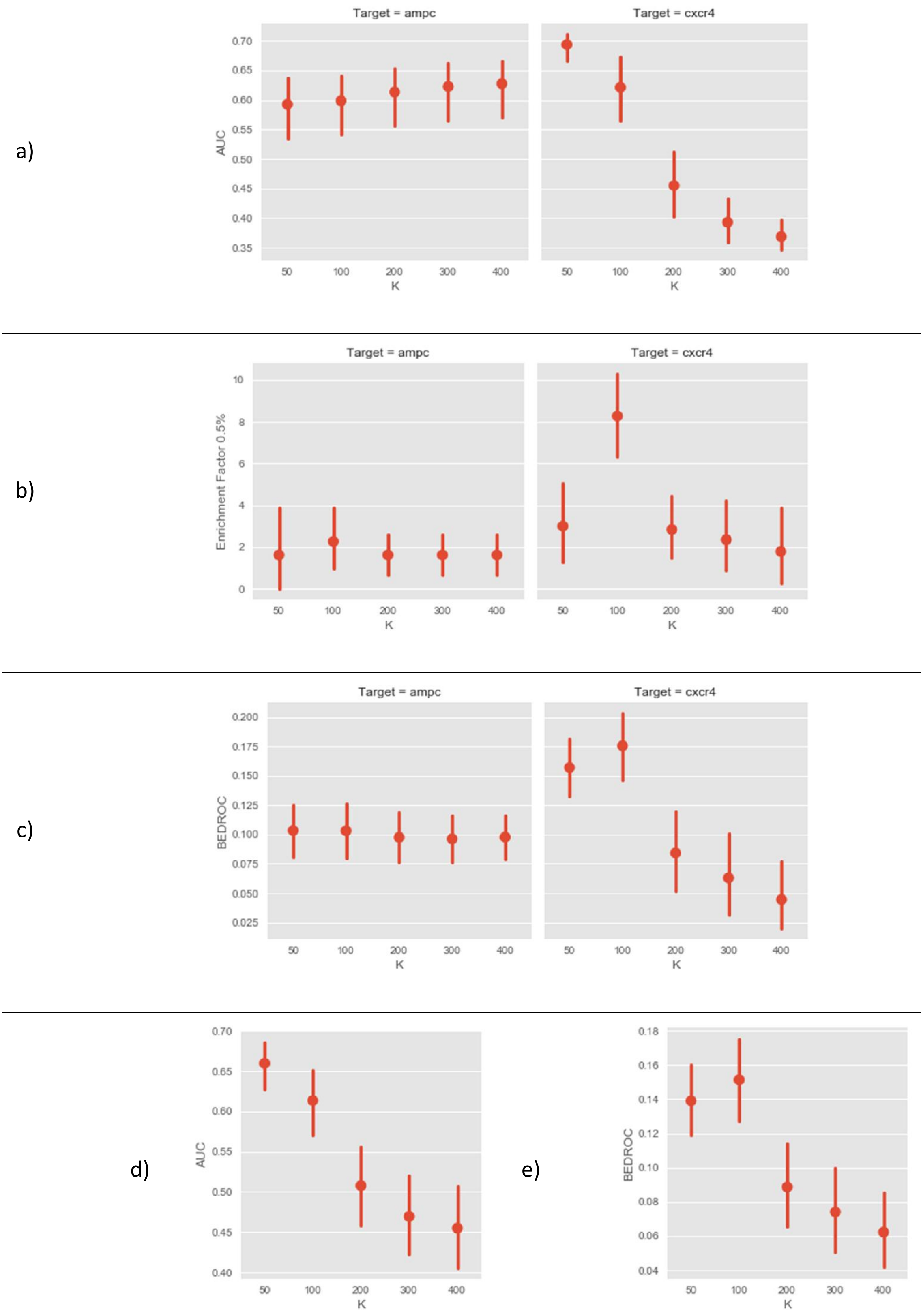


Figure 5-2. Virtual screening performance of the covariance descriptors for a number of eigenvalues.

The computational cost of the eigendecomposition is illustrated in Figure 5-3. Figure 5-3 (a) shows the computation time of a single molecule mesh with 5985 vertices at increasing numbers of eigenvalues. The figure suggests that there is an increasing, non-linear relationship between the number of eigenvalues to be computed in the spectrum and the time taken. The computation time for 100 eigenvalues is around 1 second, which increases to 5 seconds for 300 eigenvalues and finally 16 seconds for 500 eigenvalues. Additionally, Figure 5-3 shows the time taken to compute 300 eigenvalues for a random sample of 250 molecules. The number of vertices for the sample molecule used in Figure 5-3 (a) is on the lower end of the distribution with the majority being between 6,000 and 10,000. Subsequently the majority of the computation times are between 5 and 12 seconds, with the longest computation time taking 15 seconds, suggesting that in terms of performance and computation time, the choice of $k = 100$ values is acceptable.

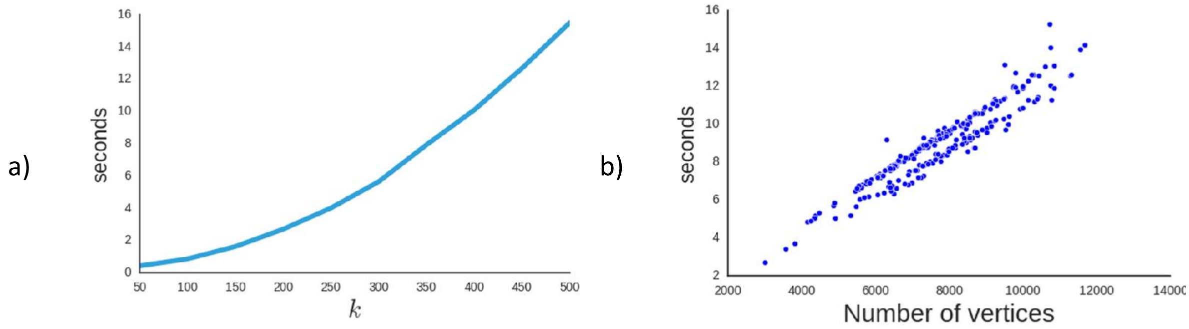


Figure 5-3. (a) shows the computation time for k-eigenvalues of a sample mesh and (b) shows the time taken to compute 300 eigenvalues with respect to the number of vertices for a sample of 250 molecules.

5.5.2.2 *Choice of time set for HKS*

The sample time points for use in the HKS were tested on the AMPC, CXCR4, and GCR data sets using the $k = 100$ number of eigenvalues obtained above. In all cases, the parameters from the general deformable shape descriptors reported in the computer vision literature, T_0, performs the worst, showing that molecular shape has domain specific features that require their own parameterisation (Figure 5-4). In particular, T_0 performs significantly worse than all other parameters at the 95% confidence interval across all targets (Figure 5-4 (d) and (e)). In the AMPC target, the local geometry descriptors with smaller dimension, T_1 to T_3, have a higher enrichment factor on average than those with a higher dimension, T_4 and T_5 (Figure 5-4 (b)). Conversely, the descriptors with the higher dimension in CXCR4 and GCR perform better than the lower dimension descriptors on all measures. Overall, between the two dimensions, the descriptors with time ranges up to 750, T_3 and T_4, perform better than those with upper bounds of 1000 or 2500. However, no set of parameters is significantly better than the others at the 95% confidence level. Therefore, the time samples T_3 and T_4 were selected for screening on a larger data set.

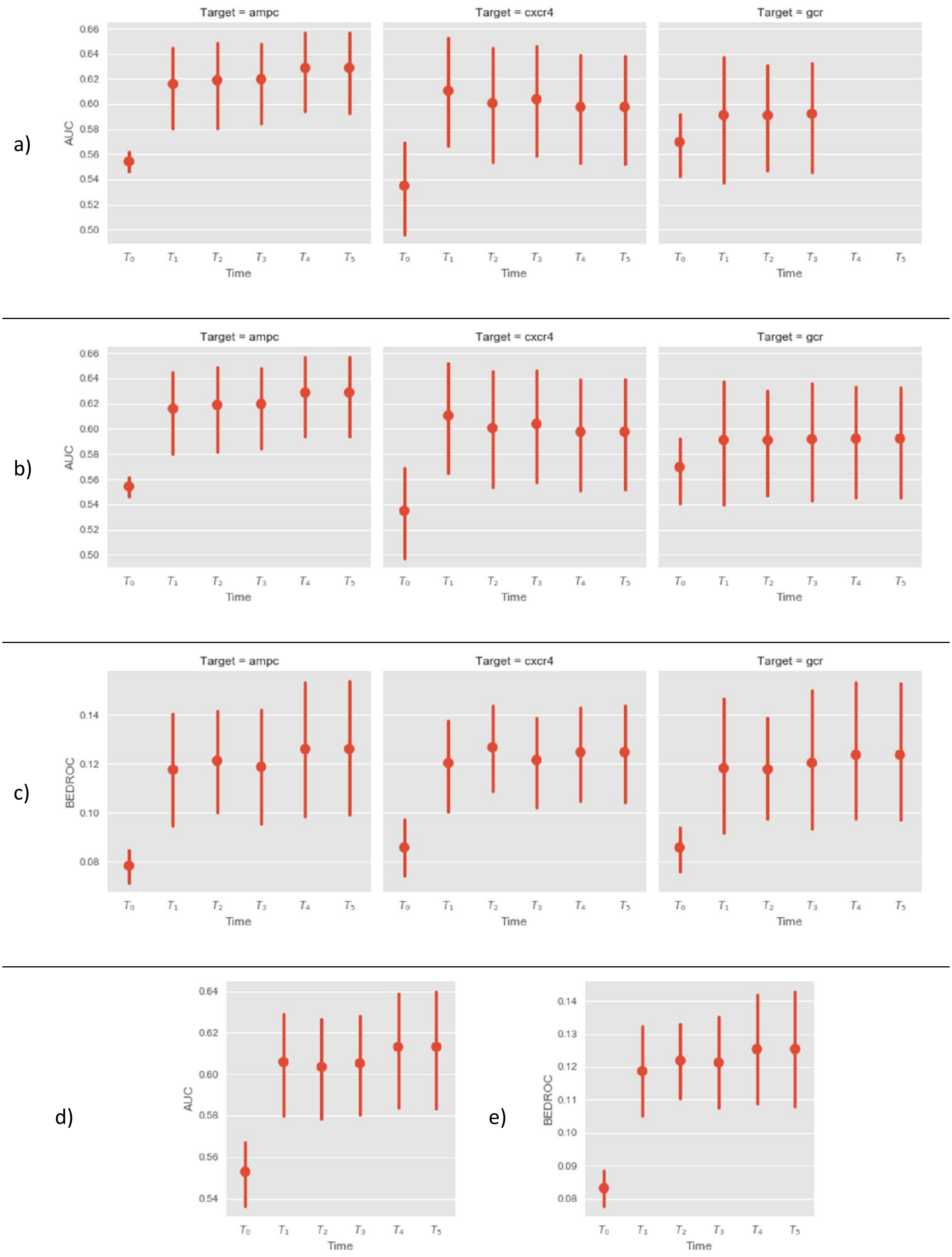


Figure 5-4. HKS parameters evaluated on three profiling targets.

5.5.2.3 Choice of number of dimensions of the WKS

The optimal number of evaluations for the WKS was investigated for the AMPC, CXCR4, and GCR targets using the $k = 100$ number of eigenvalues obtained above. The results are presented in Figure 5-5. This virtual screening experiment shows that no single parameter performs significantly better than the others. However, on average, $D = 32$ performs worse than all other parameters on average. Figure 5-5 (a) shows there is not a large variation in AUC performance with a small peak occurring at $D = 64$ and $D = 100$. On the other hand, the early enrichment performance improves with increasing D (Figure 5-5 (b) and (c)). On balance, when averaged over the three targets, $D = 64$ and $D = 100$ gives marginally better AUC scores whereas BEDROC performance improves with higher values of D.

In conclusion, the best performing WKS parameters are either the most specific and localised parameters, $D = 1000$, or a mid-point that is also sensitive to global features, $D = 100$. Therefore, the parameters taken forward for the larger screen of the DUD-E data set are $D = 100$ and $D = 1000$.

5.5.2.4 Choice of normalisation of the descriptor

A final virtual screening experiment was carried out to investigate the effect of additional row-wise normalisation. Figure 5-6 presents the results for the three targets with no normalisation, ℓ^1 normalisation and ℓ^2 normalisation. The results show that normalisation is likely to be target sensitive with the most noticeable improvement occurring in CXCR4. On the whole, across all targets, ℓ^2 performed marginally better than ℓ^1 so that normalisation is used for further experiments on the full DUD-E data set.

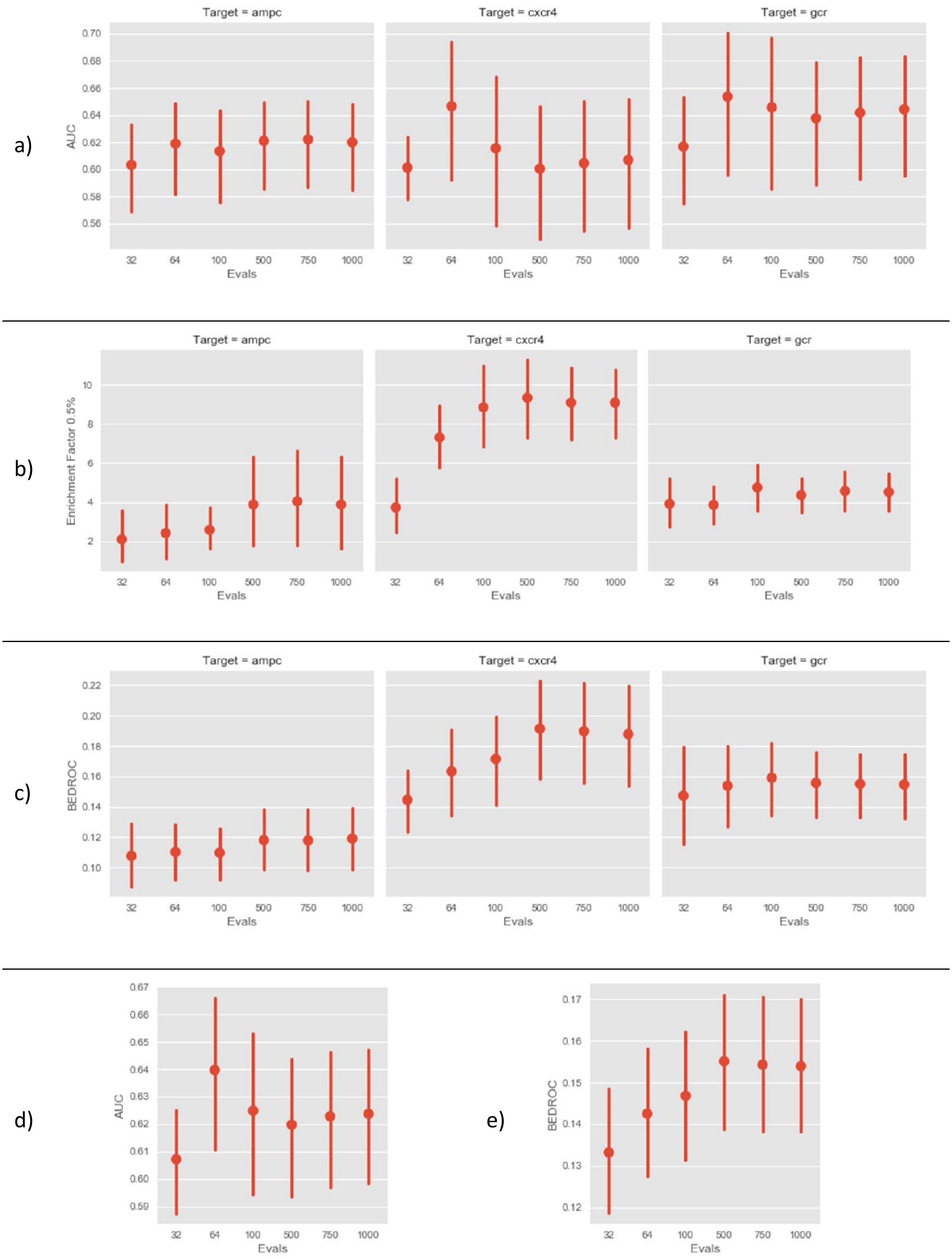


Figure 5-5. WKS parameters evaluated on three profiling targets.

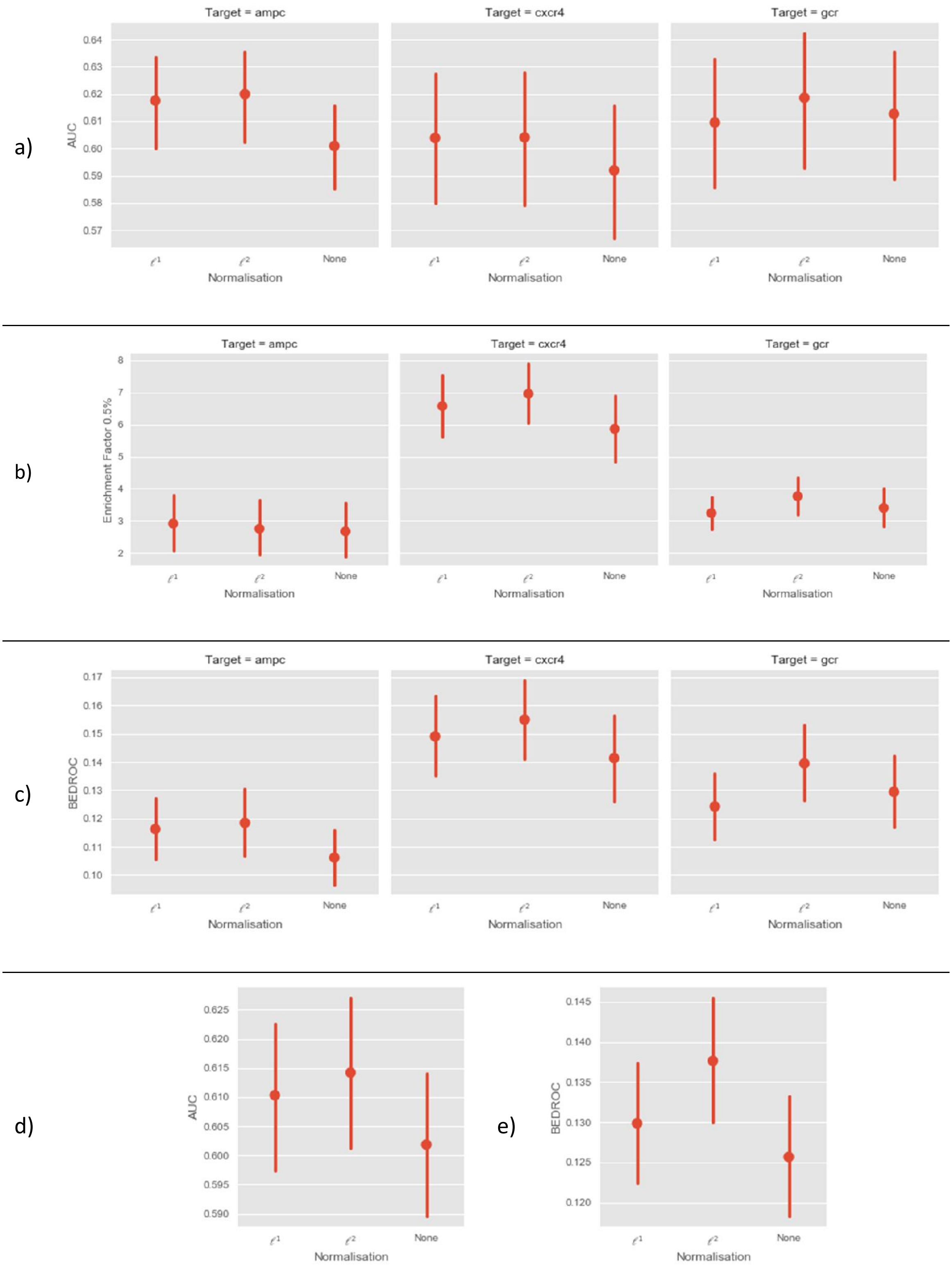


Figure 5-6. Normalisation of descriptors evaluated on three profiling targets.

5.6 Comparison against established shape similarity methods

The optimal parameters identified above were carried forward to a series of virtual screening experiment using the full DUD-E data set consisting of 102 targets, reported in 4.4.1. The spectral geometry results are compared with established shape similarity searching methods, namely Shape-it, which is an open source implementation of the Grant and Pickup (1995) Gaussian shape comparison method, and CDK Shape Moments, which is an open source implementation of UFSR implemented by CDK. The results across all targets are summarised in Figure 5-7. Comparing the two functional forms of the spectral geometry methods, the results show that the WKS performs significantly better than the HKS at the 95% confidence level using both AUC and BEDROC. The WKS, $D = 100$ descriptor performed better than the WKS, $D = 1000$ descriptor on the AUC metric but the higher dimensional descriptor performed better for the early retrieval problem.

When comparing against the alignment-free established method, both WKS descriptors (D=100 and D=1000) perform significantly better than the CDK Shape Moments descriptor using the AUC metric. WKS ($D = 1000$) also performs significantly better than CDK Shape Moments using BEDROC, and although WKS ($D = 100$) performs markedly better on average than CDK Shape Moments, it is not possible to reject the null hypothesis of no difference at the 95% confidence level. Most interestingly, the WKS descriptors produce a comparable performance with the much more computationally demanding alignment-based shape method based on AUC, with WKS ($D = 100$) performing markedly better on average than Shape-it, although it is not possible to reject the null hypothesis of no difference at the 95% confidence level. Nevertheless, Shape-it remains the best performing method using BEDROC which represents the early retrieval problem. The enrichment factor is not used here as there are different ratios of actives and decoys between the targets.

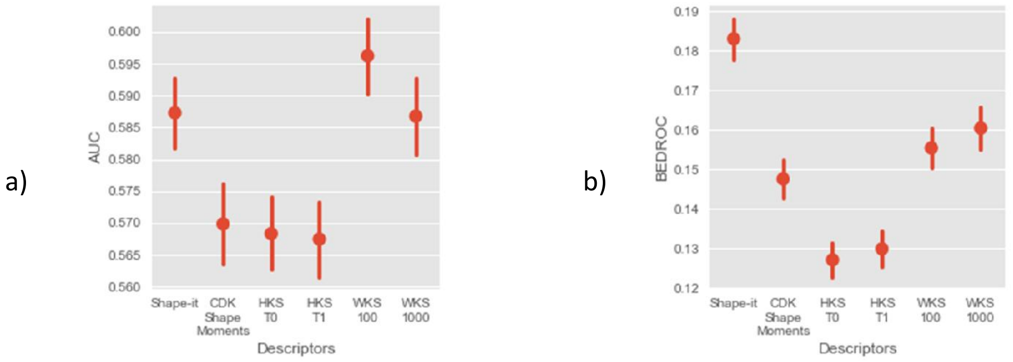


Figure 5-7. Comparison against benchmark shape comparisons.

Figure 5-8 and Figure 5-9 show the pairwise comparisons of the different shape methods across all targets in DUD-E using the AUC and BEDROC statistics, respectively. In each figure, the leading diagonal represents the distribution of values for a given shape method and the off-diagonal figures are scatter plots of the values obtained using different methods. The solid line represents equal performance between the two methods so that points above the line show better performance for the method on the y-axis of the scatter plot and points below the line show better performance for the method on the x-axis of the scatter plot.

The histograms in Figure 5-8 show that on the whole, all methods have the majority of AUC values below 0.6, with all methods having some values lower than 0.5, which confirms the difficulty of the DUD-E data set described by Jahn et al. (2009). Additionally, in general, Shape-it performs worse than other methods when both methods return high AUC values. The best performing targets in both methods are found in the top right quadrant and the solid line denotes the relative performance. This is demonstrated in the Shape-it row where most points in the top right quadrant are below the solid line meaning that Shape-it performs relatively worse when both perform well. This is even the case in the HKS descriptors that perform the worst overall. Furthermore, the WKS shows good AUC performance against all other methods, with an even distribution of points above the solid line in the respective rows. However, when compared against Shape-it, it appears that there is a cluster of mid-range AUC values where Shape-it performs better, while on the whole the WKS descriptors have better performance in the lower and higher ranges of AUC values. In other words, the WKS descriptors perform better than Shape-it for targets in which both methods perform comparatively poorly and well.

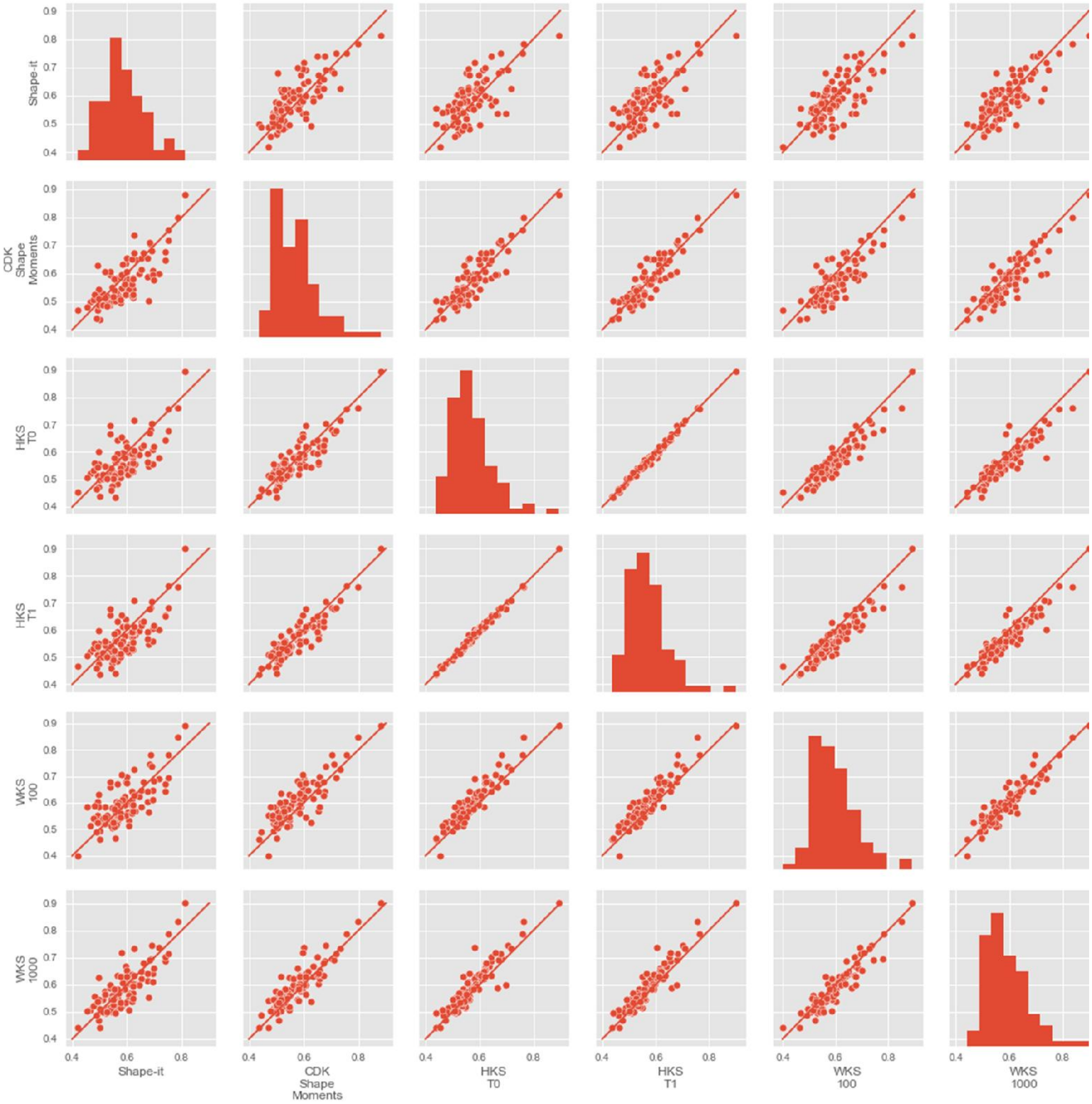


Figure 5-8. Benchmark shape comparison AUC values comparison.

Figure 5-9 shows that all methods have a distribution that is skewed towards the lower end of the BEDROC values. When compared against the other methods, Shape-it performs best across the board, although interestingly, when both methods perform well, Shape-it generally has a lower BEDROC value than the other method, a phenomena also observed with AUC values in Figure 5-8.

Furthermore, when compared against the WKS descriptors, the CDK Shape Moments descriptor performs worse in nearly all targets. The relative closeness of the BEDROC values in comparison to the AUC values in Figure 5-7 is therefore due to many of those targets being clustered around the solid line.

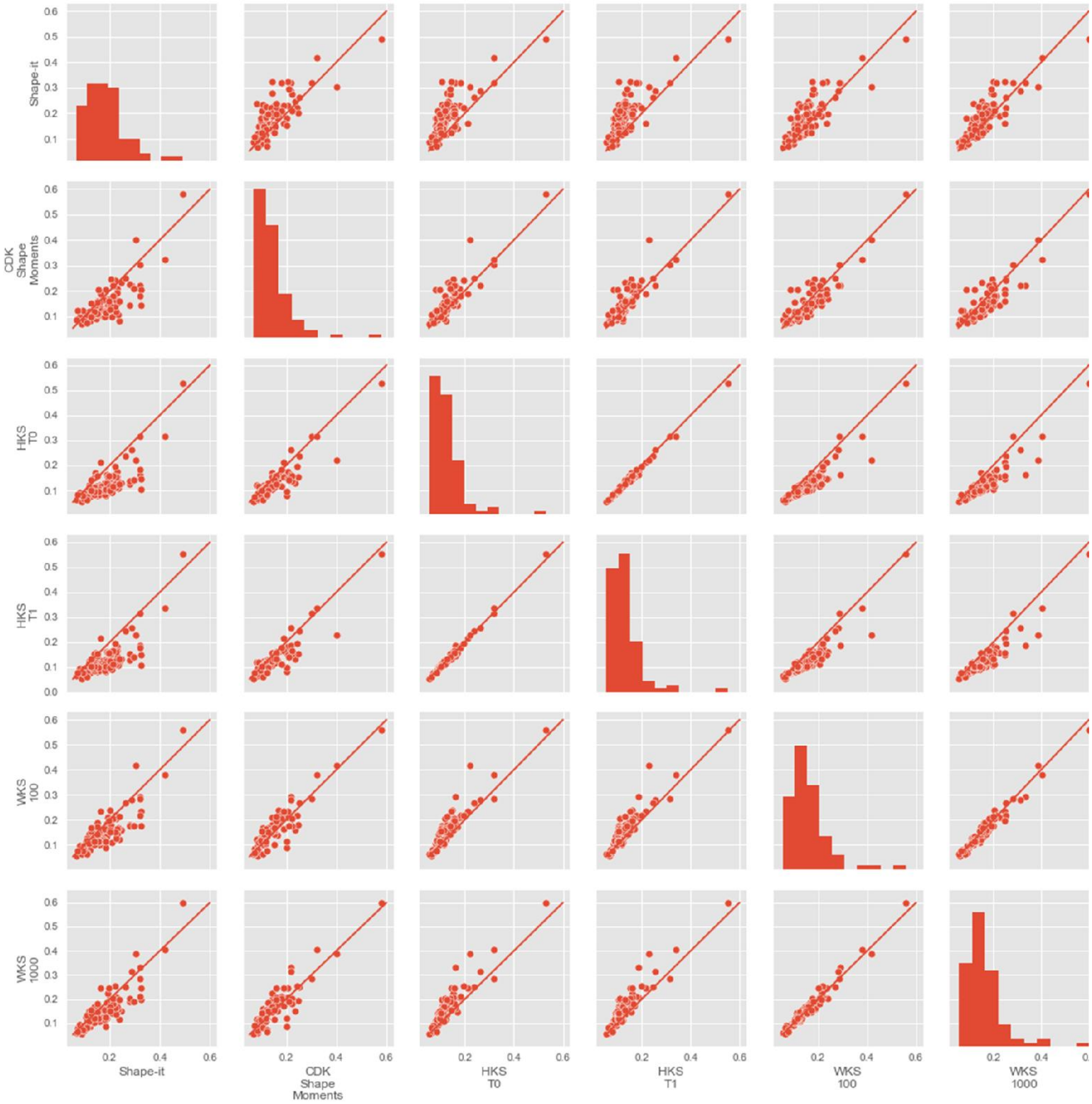


Figure 5-9. Benchmark shape comparison BEDROC, $\alpha = 20$, values comparison.

5.7 Discussion

This chapter has carried out a series of quantitative experiments to determine the best parameters of the local geometry descriptor for the use in a global geometry descriptor for virtual screening. The results show that the local geometry descriptor performs best as a global descriptor when the first $k = 100$ eigenvalues are used. With respect to the functional forms of the local geometry descriptors, the time samples T_3 and T_4 are used for the HKS, $D = 100$ and $D = 1000$ are used for the WKS. Tests on the use of different normalisation techniques applied to the local geometry descriptors suggest that there is no significant improvement for using normalisation as a pre-processing step but that ℓ^2 performs best overall. Furthermore, the covariance descriptor performs well as a descriptor for 3D shape screening of a standard test set against some benchmark comparison methods.

The parameter tests for the two functional forms of the local geometry descriptors can be viewed in the light of the specificity and sensitivity discussion in the Chapter 4. The best performing HKS descriptors are those that sample time up to 700, suggesting that the added global properties from higher time values do not improve the encoding of geometric features over the whole shape. The best performing WKS descriptors are either the most specific and localised parameters, $D = 1000$, or a mid-point that is also sensitive to global features, $D = 100$. When the two functional forms are compared for their overall performance, the WKS outperforms the HKS, suggesting that the specificity of the WKS results in a better framework for a global shape similarity task. As a corollary, it can be hypothesised that the best global shape descriptors are derived from local geometry descriptors that perform best at distinguishing the geometry of individual points, rather than local geometry descriptors that emphasise global geometric properties.

Promisingly, the spectral geometry descriptors perform favourably on a large scale virtual screening experiment when compared to current 3D shape descriptors. When compared against an implementation of another alignment-free shape descriptor, CDK Shape Moments, the WKS performs significantly better on both AUC and BEDROC. This suggests that more geometric information is encoded in the spectral geometry descriptors than in the CDK Shape Moments which are based on interatomic distances. The result is especially promising as the covariance descriptor requires minimal parameterisation suggesting that global descriptors that can be parameterised for the domain may perform even better. Most interestingly, the covariance descriptors also perform comparatively well against Shape-it, a much more computationally demanding Gaussian-based method, using AUC as a metric.

The results have shown the optimal choice of local geometry descriptor parameters for the purpose of global similarity tasks. However, the results are based upon an assumption that these properties will be consistent for different choices of global geometry descriptors. The strength of the covariance descriptor method introduced in this chapter is that it requires minimal optimisation. It is a direct translation of a local geometry descriptor to a global descriptor space and does not require additional artefacts, such as a codebook for the bag of words descriptors, which will be described in Chapter 6. To this extent, its performance is impressive as there are no global parameters to optimise. However, it is worth highlighting two potential limitations. The first is that the covariance descriptor is a column-wise mapping of the local geometry descriptor, so it is unclear how these optimal parameters would be consistent when applied to row-wise global descriptors. Secondly, as the covariance descriptors are D^2 in size, there is a practical concern for storing large databases of molecules for very large choices of D.

A surprising result is the poor performance seen for large numbers of eigenvalues, for example, for $k > 100$. Figure 5-2 shows that at values larger than $k = 100$, the performance of the global descriptors falls markedly. This may be due to the mesh generation process, which naturally introduces noise into the descriptor. While the mesh generation software is designed to be as smooth as possible, there are still some elements of noise introduced, such as creases around the edges of high curvature or flat sections that have no curvature. Given that the higher eigenvalues correspond to increasingly local shape variation, the higher values may encode more of these noisy artefacts, which may be detrimental to the performance of the local geometry descriptors for a global similarity problem. However, the effect may be dominated by one target and overall might be due to a small sample size. Future work should test this on more targets to see if the result can be generalised.

5.8 Conclusions

In summary the performance of the local geometry descriptor for a global similarity task was evaluated at the optimal parameters were $D = 100$ and $D = 1000$ for the WKS, and T_3, and T_4 for the HKS. In particular, the results suggest that the best global shape descriptors are those that are based on local geometry descriptors that are best at discriminating between the geometry of local points, rather than those that emphasises global geometry in the local geometry descriptors. When evaluated against standard 3D shape descriptors on virtual screening of the DUD-E data set, the WKS descriptors outperform the CDK Shape Moments descriptor and is competitive with the Gaussian shape comparison method, which implies that the spectral geometry approach is very promising for developing as a descriptor for 3D molecular shape. In order to develop a global geometry descriptor for efficient large scale virtual screening of molecular shape, the next chapter will investigate if performance can be further improved using global geometry descriptors that have more parameters.

6 Bag of Features global geometry descriptors

6.1 Introduction

The work carried out in previous chapters has developed local spectral geometry descriptors for molecules and investigated their behaviour as global shape descriptors using the covariance matrix method (Chapter 5). While the purpose of the covariance descriptor was to evaluate the descriptive properties of the local geometry descriptors at the global shape level, it also introduced the concept of mapping local shape descriptors to a global descriptor space for shape comparison. However, for a given local geometry descriptor matrix, there are two ways in which the matrix can be mapped to a local descriptor: column-wise, which maps the value of each filter over the entire space, or row-wise, which maps the local geometry properties of each point on the surface. The covariance descriptor is a column-wise mapping and describes global shape according to how the filters co-vary with each other over the entire manifold of the shape. The work presented in this chapter is a row-wise approach, which aggregates the values of the local point descriptors into a space of spectral geometry features and describes how each shape relates to these features on average.

Bag of Features descriptors are the most common form of global geometry descriptor for spectral geometry in computer vision (Bronstein, Bronstein, Guibas, & Ovsjanikov, 2011) but have a longer history in the field of image processing and signal compression and originate as descriptors used for text retrieval (Salton, Wong, & Yang, 1975). The method uses a *codebook* that represents key geometric features in feature space and aggregates the frequency of the occurrence of the features over the object to be described. The benefits compared to the covariance descriptor are two-fold. First, the descriptor is computed row-wise, that is to say it is a summary of the geometric features of points on the surface rather than filter functions mapped over the surface. Second the descriptor is more compact: for a codebook with c codewords a $N \times D$ descriptor will be mapped to a $1 \times c$ vector rather than the D^2matrix generated using the covariance method. These benefits suggest that the Bag of Features descriptors may offer a performance improvement over the covariance descriptors for use as a global spectral geometry descriptor of molecular shape. However, there are a number of parameters that need to be optimised in order for the descriptor to be optimal for the domain and the parameters appropriate for representing molecular shape are not known. Therefore, building on the work of Chapter 5, this chapter will introduce the Bag of Features descriptors and find the optimal parameters for their use as a global descriptor for virtual screening of 3D molecular shape.

6.2 Methods

The Bag of Features descriptor can be considered intuitively as an attempt to give a semantic foundation to global spectral geometry descriptors. At the heart of this interpretation is the codebook, which takes a large sample of rows from different local geometry descriptors and finds the prominent features in the local geometry descriptor row space. In doing so, each point is endowed with some semantic meaning by describing how close it is to specific features. For example, suppose a codebook had a codeword that represented a local geometry descriptor vertex in a cupula-like shape in the region. Then a given local geometry descriptor vector on a shape can be characterised on how cupula-like it is by determining how close the local geometry descriptor vector is to the cupula codeword. In practice, the codewords are abstract points in the spectral geometry feature space that serve to give a semantic grounding to the local descriptors and do not necessarily have nameable geometric properties. The step that assigns a meaning to a point on a surface with respect to the words in a given codebook is called *encoding*. Finally, these encoded points are aggregated and normalised. In the example above, this is akin to saying how cupula-like the points on a surface are on average. The rest of this section looks at these design choices one by one.

An overview of the workflow is presented in Figure 6-1. For a shape X defined as a manifold surface, let $\mathscr{f}(x_i)$ be a set of spectral geometry filter functions evaluated at point $x_i \in X$. This mapping is of a point on the surface to the row-space of the local geometry descriptor so that $\mathscr{f}(x_i)$ is a $1 \times D$ vector where D is determined by the number of filter functions presented in Chapter 4. Therefore, if $\mathbf{F}$ is an $N \times D$ local geometry descriptor matrix of X then $\mathscr{f}(x_i)$ corresponds to the row $\mathbf{f}_i$. The workflow then has three important steps, the first is to compute a codebook, V, that represents important features in the descriptor space. The codebook can be generated using a number of parameters that will be explored in the next section. In Figure 6-1, the codebook, V, has three codewords, v_1, v_2, and v_3. The second step then uses the local geometry descriptor row vectors along with the codebook to encode each row descriptor with respect to how similar it is to the codewords in the codebook. The result is an encoded vertex $\theta(x_i)$ for all $x_i \in X$. The encoding step also has a number of parameters that will be explored further below. When the encoding step is complete, each vertex is related to a codebook that is common to all shapes. These encodings are then aggregated over the shape to produce a frequency histogram. Finally, the histogram is normalised to give the global shape descriptor. The different normalisation steps that can be undertaken are also explored in detail below.

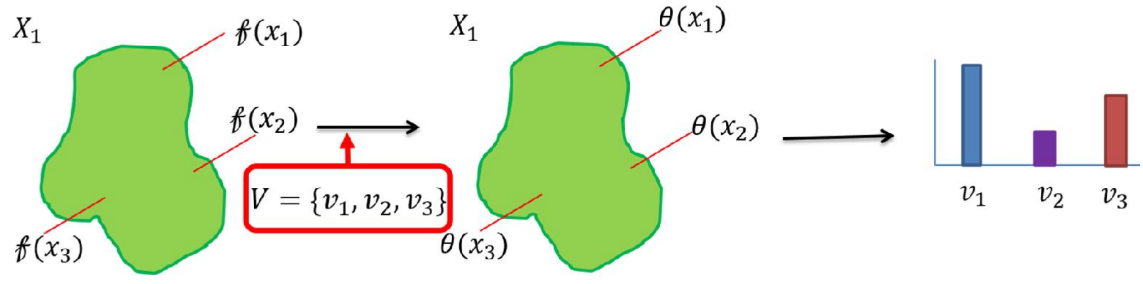


Figure 6-1. An overview of the Bag of Features descriptor workflow.

6.2.1 Codebook generation methods

The purpose of the codebook is to determine a set of features that are discriminative enough in the feature space to be used to describe the local geometry of a shape. In an illustrative example, suppose that all features on the surface of a non-rigid shape are either peaks or valleys depending on the Gaussian curvature. The goal would then be to find two coordinates in the local descriptor space that best represent peaks and valleys. Once this has been achieved, these representative coordinates can be used to determine whether each point on the shape is a peak or a valley. However, in practical applications, the non-rigid features on a shape are unlikely to be that clear cut. The approach then is to find a method of producing a collection of representative coordinates for the K most discriminate features. This section addresses the topic of producing the best codebook from a collection of molecular shapes.

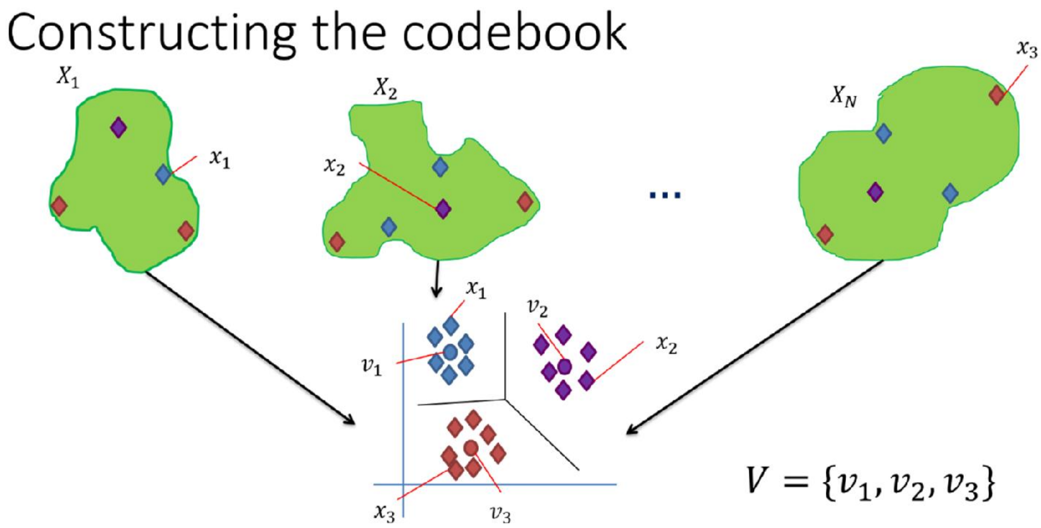


Figure 6-2. An overview of the codebook generation process.

Figure 6-2 illustrates the codebook generation workflow, which takes a sample of N molecular shapes and computes their local geometry descriptors with respect to a set of parameters from Chapter 4. These local geometry descriptors are collated to give a large data set of rows that represents the space of local geometry descriptor vectors of the points on the surfaces of all the shapes. In Figure 6-2, the dimension of the local geometry descriptor is $D = 2$ so that the local geometry descriptors derived from all the molecules can be plotted in two-dimensions, as shown at the bottom of the figure. In this example, there are three features on the shapes, one that represents *valley* like features in blue, one that represents *cupula* like features in red, and flat purple features. While these are located at different points over the three example shapes, in the feature space, they form three clear clusters, suggesting that there are three distinct geometric features in the descriptor space. The centroid of each cluster is taken as a representative of that cluster and the three centroids are collected together to form the codebook $V = \{v_1, v_2, v_3\}$.

The important variables for finding the best codebook representation are the sample of local descriptors used to train the codebook and the algorithm used to cluster the descriptors and identify the centroid for each cluster. The remainder of this section will describe the different sampling approaches and introduce the different models to find a good codebook to represent local descriptor feature space.

6.2.1.1 Sampling

The codewords in the codebook are learnt from a sample of the local descriptor space. This sample can be obtained using a number of different methods and may have an effect on the semantic features of the resulting codebook. For example, one design choice would be to decide whether it is necessary to learn a codebook for each target, or whether a single codebook can be discriminative enough for all shapes. For the purposes of virtual screening, it would be ideal to have a codebook that would be sufficiently discriminative across all targets as it would require just one global shape descriptor per molecule to carry out the screening of an entire database regardless of the target. As the bag of features descriptor is a row-wise descriptor there are two sources of sampling variation that can be considered. The sample can be taken as a subset of rows from a given matrix (surface sampling) as well as a sample of shapes from a data set. Both aspects were tested as part of the parameterisation of the codebook generation.

Surface sampling is used when the number of vertices in the mesh is such that using the whole local geometry descriptor is impractical. Such an application is typically used in the spatially sensitive expressions descriptor below. The farthest point sampling methodology (Eldar, Lindenbaum, Porat, & Zeevi, 1997) can be used to obtain the most representative sample of the surface.

Farthest Point Sampling is a sampling strategy that aims to find the most diverse sample of points based on distance. The algorithm partitions the surface or space into Voronoi cells and takes a representative member of each cell as a sample point. A Voronoi cell is a region on a plane such that all points within the region are closer to a particular central point than to any other. The Voronoi partition is determined using the distances between points on the surface taken from a distance matrix and there is flexibility in the choice of the distance metric. In the case of the local geometry descriptors there are two options: 3D Euclidean distance, which partitions the molecule's surface into Voronoi volume cells, and the diffusion distance, which uses the non-Euclidean geometry of the manifold to partition the surface into cells of equal distance over the surface. The Euclidean distance is more efficient in terms of computing, whereas the diffusion distance best incorporates the local surface geometry. The diffusion distance is defined as,

$$d(x, y) = \sqrt{2h_t(x, y) - h_t(x, x) - h_t(y, y)}, \qquad \text{Equation 6-1}$$

where $h_t(\cdot,\cdot)$ is the heat kernel between two points at time, t.

While the above method identifies the most diverse sample of vertices of a single shape, the diversity of local geometry descriptors will also depend on a diverse sample of the molecular shapes themselves. In this respect, choosing a diverse set of molecules is desired to be able to capture local geometry descriptor space fully for the purpose of codebook generation. In principle, the larger the sample size the better. However, this is offset by the amount of data required to be held in memory during the learning phase. For the parameterisation carried out in this chapter, a diverse range of shapes was obtained by randomly sampling the DUD-E data set.

6.2.1.2 K-means

K-means clustering is the typical codebook generation technique for Bag of Features descriptors (Ovsjanikov et al., 2009) and is used in a wide range of signal processing applications for signal compression and reconstruction (Murphy, 2012, p. 356). K-means clustering is a technique that aims to partition the data space into k regions (Murphy, 2012) centred around cluster centroids, whose coordinates are the means of the data points in the cluster. K-means is an example of prototype learning, whereby ideal representative data points are inferred from the data, and each prototype point is the mean coordinate in the cluster. In the context of Bag of Features, a centroid then becomes a *word* in the codebook so that all local geometry descriptor vectors can be evaluated by how close they are to this centroid.

The algorithm starts with K initial seeds labelled as centroids and the remaining local geometry vectors are assigned to the nearest seed. The centroid of each cluster is then recomputed and the local geometry vectors are reassigned to the nearest centroid, which is repeated until there is no change in the centroids or a user-defined upper limit of iterations is reached. Once completed, the centroids of the K clusters become the *codewords* of the codebook.

The traditional K-Means algorithm has some performance issues under very large data sets. For example, when a large data set of local geometry descriptors was run on a high-performance computing cluster with 256GB of allocated RAM, the K-Means algorithm failed to converge. In order to address this issue for large data sets the Mini-batch K-Means algorithm was used (Sculley, 2010). Mini-batch K-Means uses a subsampling strategy to provide fast training convergence yet still uses the same K-Means objective function. A result of the subsampling process is that there is reduced quality in the clusters, however Sculley (2010) showed that the quality is improved when compared to other algorithm optimisation techniques such as stochastic gradient descent.

6.2.2 Encoding Methods

The encoding step uses the codebook to give some notion of semantic meaning to the local geometry of a vertex. It does this by associating the local geometry descriptor with the codewords in the codebook. For example, and as discussed in section 4.2, suppose the codebook has two codewords that have geometric properties related to Gaussian curvature whereby one codeword describes the local geometry of a peak and the other describes the local geometry of a valley. The encoding step can then be used to see how close the local descriptor of a given vertex is to a peak or a valley. In the simplest case the method would encode all vertices on the mesh as being peaks or valleys. In the more general case, the method would encode each point according to how closely the local geometry describes a peak or valley feature. In other words, the encoding method describes the local geometry in terms of the codewords of the codebook. In the example, this is akin to describing the local geometry as either "peak-ness" or "valley-ness".

An illustration of the different encoding methods is presented in Figure 6-3. In this example, the codebook has been taken from the one obtained in Figure 6-2 and an example local geometry descriptor vector $f(x_i)$ has been selected, with coordinates that position it as the grey diamond in the descriptor space. There are a number of ways of encoding the semantic properties of the local geometry descriptor vector, of which three are selected for this chapter: hard vector quantisation (HQ), soft allocation vector quantisation (SA) and K-nearest-neighbours vector quantisation (KNN). An overview can be seen in Figure 6-3 and more detail is given in this section.

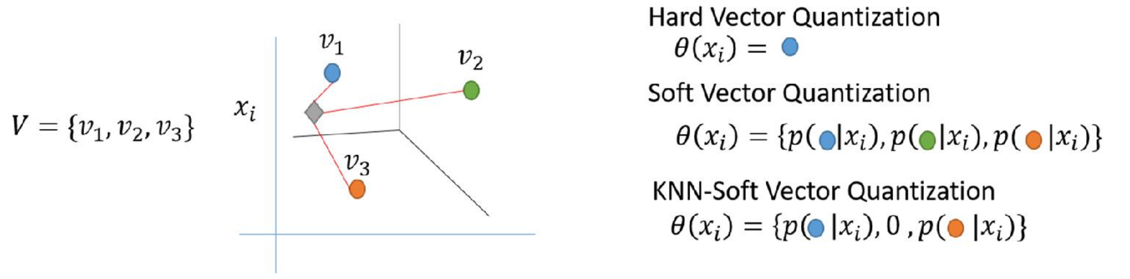

$V = \{v_1, v_2, v_3\}$
x_i
v_1
v_2
v_3
Hard Vector Quantization
$\theta(x_i) =$ ●
Soft Vector Quantization
$\theta(x_i) = \{p(● | x_i), p(● | x_i), p(● | x_i)\}$
KNN-Soft Vector Quantization
$\theta(x_i) = \{p(● | x_i), 0, p(● | x_i)\}$

Figure 6-3. Illustrative overview of the different encoding methods.

6.2.2.1 Hard Vector Quantisation (HQ)

Let $\theta(x)$ denote the encoding of a vertex, $x \in X$. Hard vector quantisation (HQ) is the simplest encoding method. Each vertex, $x \in X$ is assigned the closest codeword in the codebook based on its local descriptor $g(x)$,

$$\theta(x) = \underset{v_i \in \mathcal{V}}{\operatorname{argmin}}\{d(g(x), v_i)\} \qquad \text{Equation 6-2}$$

for codewords $\mathscr{v}_i \in \mathcal{V}$.

This method has the highest amount of information loss, for example, the relationship of the grey diamond to the purple and orange circles representing codewords has been discarded and the vertex has been encoded by a single codeword.

6.2.2.2 *Soft Allocation Vector Quantisation (SA)*

Soft allocation vector quantisation (SA) attempts to reduce the amount of information lost in the transformation by assigning a vector of probabilities to each vertex. Each vertex, $x \in X$, is assigned a vector of size $1 \times K$, where K is the number of codewords in the codebook. Then the i^{th} element of the vector represents the probability that the local geometry of the vertex is close to the i^{th} codeword in the codebook. The probability scores are determined using the softmax formula (Bronstein et al., 2011),

$$p(v_i|x) = c(x)\exp\left(\frac{\|\mathscr{f}(x) - \mathscr{v}_i\|_2^2}{2\sigma^2}\right)$$

Equation 6-3

where $c(x)$ is a normalisation constant that ensures $\|\theta(x)\|_1 = 1$. The resulting encoding is then a vector of K dimensions where each element corresponds to the probability that the local geometry of the vertex is close to a given codeword feature,

$$\theta(x) = \{\, p(\mathscr{v}_1|x), \dots, p(\mathscr{v}_K|x)\}.$$

Equation 6-4

This quantisation method is referred to as *soft* as it allows the local geometry of a vertex to be encoded as a mixture of features. In doing so, more information of the local geometric properties is preserved. Nevertheless, in assigning a non-zero probability to features that are far away, this encoding method may introduce additional noise that may be amplified when pooling over all the vertices in a shape mesh. For example, in Figure 6-3 although the grey diamond is a long way from the purple dot it is assigned a non-zero probability of being a member of the purple dot cluster. When aggregated over the shape, this noise will be cumulative resulting in a global descriptor assigning shape properties to a shape that are not present.

6.2.2.3 K-Nearest Neighbour Vector Quantisation (KNN)

K-nearest neighbour vector quantisation (KNN) is an attempt to balance the trade-off between the information loss of HQ with the increased noise of SA. This encoding method assigns the softmax probability to the k-nearest codewords to each local descriptor. The encoding can therefore be defined as,

$$\theta(x) = \begin{cases} c(x)\exp\left(\frac{\|\mathcal{f}(x) - \mathcal{v}_i\|_2^2}{2\sigma^2}\right) & \forall \mathcal{v}_i \in KNN \\ 0 & otherwise \end{cases}, \qquad \text{Equation 6-5}$$

where KNN is the set of the k-nearest codewords to the local descriptor $\mathcal{g}(x)$. In Figure 6-3, with K=2 nearest neighbours, the descriptor will assign a stronger membership of the blue dot than the orange dot, which reflects its closer proximity to blue, and will assign a zero value to the purple feature as this feature is not a near neighbour.

6.2.3 Pooling methods

The pooling step involves collating the encoded vertices over the entire shape to produce the global descriptor. A simple pooling technique would sum the occurrence of each feature over the shape to give a histogram of the codewords that provides a global descriptor of feature occurrence. In this approach, however, the spatial relationship between the features is necessarily lost. For example, consider two shapes that have the same frequencies of features of the codewords; in one case the features may be clustered at opposite ends of the shape, whereas, in the other they may be evenly distributed over the shape, yet both shapes are given the same descriptor. While spatial pooling methods do exist (Li & Hamza, 2013), in the case of non-rigid shapes, any spatial pooling method must take into account the inherent non-Euclidean geometry of the local descriptor. In this section, two pooling methods are presented. The first is a simple summation over the vertices and the second is a spatial pooling method that describes the pairwise distribution of features.

6.2.3.1 Summation pooling

For a non-rigid shape X, let a global shape descriptor be denoted as $\bar{\theta}(X)$. The most straightforward transformation of the encoded vertex descriptors to a global descriptor is a summation over all vertices,

$$\bar{\theta}(X) = \int_X \theta(x)dx. \qquad \text{Equation 6-6}$$

In the discrete setting of a mesh sampled over a shape, this becomes,

$$\bar{\theta}(X) = \sum_{x \in X} \theta(x), \qquad \text{Equation 6-7}$$

where $x \in X$ is a vertex of a mesh over shape X and $\theta(x)$ denotes its respective local encoded descriptor. As $\theta(x)$ is a $1 \times K$ dimension vector, where K denotes the number of features in the codebook, $\theta(X)$ is also a $1 \times K$ dimension vector. In the case of hard vector quantisation, the i^{th} element of $\theta(X)$ is the sum of the number of vertices encoded as the i^{th} codeword. In other words, this becomes a simple frequency histogram of the codeword features on the shape. When normalised by the total number of vertices in the mesh, the descriptor describes the relative frequency of the codewords in the shape. Suppose the two codeword codebook is taken from the example above, a shape encoded using the hard vector quantisation coding would produce a descriptor that describes how may valley features there are relative to how many peak features there are.

Furthermore, the semantic meaning of the descriptors is extended to other encoding methods. For a mesh encoded with the soft allocation vector quantisation, a normalised histogram would describe the average probability that a vertex is near a feature in feature space. Likewise, with k-nearest neighbour descriptors, the histogram represents the average probability that a vertex will lie near one of the codewords as its k-nearest neighbour. This demonstrates an attractive feature of this pooling method: despite the shape and local geometry descriptors being abstract, the use of a codebook and encoding method produces a descriptor that has some semantic meaning, thus allowing it to be compared meaningfully across different shapes.

6.2.3.2 Spatially Sensitive Expression pooling

In order to preserve information regarding the spatial distribution of the codewords over the shape, Bronstein et al. (2011) defined the spatially sensitive expression descriptor. The motivation for this descriptor is to give a notion of spatial relationships by describing the co-occurrence of codewords. These pairwise occurrences are then weighted by their diffusion distance in order to take the non-Euclidean geometry into account. Consequently, the spatially sensitive expression descriptor is defined as,

$$\bar{\theta}(X) = \int_{X\times X} \theta(x)\theta(y)^T k_t(x,y) d\mu(x) d\mu(y), \qquad \text{Equation 6-8}$$

for all $x, y \in X$ where $k_t(x,y)$ denotes the value of the heat kernel evaluated on vertices x and y at time t. The two crucial components of the above equation are $\theta(x)\theta(y)^T$ and $k_t(x,y)$. The first of which is the outer product between two encodings that defines a $K \times K$ matrix showing all pairwise combinations of the elements in the codebook. Therefore, it is a 2D histogram that gives the frequencies of pairwise occurrences in the hard vector quantisation encoding and the joint probabilities that a vector is near both codewords in the soft allocation vector quantisation encoding. The second crucial component is the heat kernel $k_t(x,y)$ that is the amount of heat transferred over the surface of the shape between vertices x and y at time t. As this is a measure of heat transfer, it is best thought of as a measure of diffusion distance between the two points. When used as a weight in the above equation, it weights the outer product by a non-Euclidean measure of distance over the surface. Thus it promotes near features and demotes far features in a manner that explicitly takes in to account the underlying non-rigid geometry of the shape descriptor.

Due to the large number of vertices in the meshes used to represent molecules in the DUD-E data set, the spatially sensitive expression descriptors are computed on a sample of the surface points selected using farthest point sampling.

Finally, spatially sensitive expression descriptors are matrices that are compared using the Frobenius norm,

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left|a_{ij}\right|^2}, \qquad \text{Equation 6-9}$$

which is the matrix equivalent of the Euclidean norm for a vector. The distance between two matrices is then computed as,

$$d(A,B) = \|A - B\|_F. \qquad \text{Equation 6-10}$$

Importantly, this is a distance measurement whereas the other metrics used in this chapter are similarity measurements so the rankings have to be changed accordingly to compare between distance and similarity methods.

6.2.4 Normalisation

Finally, once the vertices have been encoded and pooled over the molecule, one final task remains, which is to normalise the histograms. Normalisation can be carried out in a number of ways. One of which is to divide by the number of vertices. In the case of the hard vector quantisation this means that each element in the vector is the proportion of the molecule encoded in a specific feature. In other words, it is the probability that a given vertex will be encoded as a codeword. On the other hand, the soft allocation vector quantisation encoding, the element represents the average probability that a vertex lies near a codeword using the softmax distance.

Other options include weighting each histogram so that all histograms lie in the unit circle of a given metric. This means dividing the histogram by the sum of the lengths using some metric space, normally ℓ_1 or ℓ_2. To put it simply, the histograms are divided by the sum of the entries, in the case of ℓ_1 or the squared entries in the case of ℓ_2. Notice that when the hard vector quantisation is used, this is the same as the ℓ_1 weighting as the sum of all the entries is equal to the number of vertices.

For a histogram $\theta(X)$ for a given molecule X, the $\ell_1$weights are,

$$\frac{1}{\sum_i \theta_i(X)}, \qquad \text{Equation 6-11}$$

and the $\ell_2$weights are,

$$\frac{1}{\sum_i \theta_i(X)^2}. \qquad \text{Equation 6-12}$$

These normalisation steps are crucial if a dot-product similarity metric is used as it ensures that the descriptor vectors have the same magnitude.

6.2.4.1 Term Frequency – Inverse Document Frequency (tf-idf)

Finally, a weighting scheme that was originally derived in text retrieval can be used. This weighting scheme is an attempt to include the significance of a codeword. To use an example from text retrieval, a codebook that includes words like "the” would be common to all documents but would not give much, if any, semantic meaning to the document. Therefore, codewords that are highly common would not be able to discriminate very well between molecules. The term frequency-inverse document frequency (tf-idf) method weights the descriptors in order to promote the occurrence of discriminative codewords. It is composed of two parts: the first computes the term frequency to find the frequency of codewords in the shape, and the second computes the inverse document frequency to find the global occurrence of the codewords in the corpus, which in this case would be the database of molecules. The final weighting is then the product of the term frequency and the inverse document frequency.

The term frequency can be taken as the histograms themselves, $\theta(X)$, and the i^{th} inverse document frequency is computed as,

$$idf_i = \log \frac{N+1}{DF(v_i)+1},$$ Equation 6-13

where $DF(v_i)$ is the frequency of the codeword v_i over all N molecules. As the descriptors are sometimes sparse, such as the hard vector quantisation encoding, then it is necessary to add one to the denominator in order to guarantee smoothness by ruling out dividing by zero. This would occur when a codeword is assigned 0 over all the molecules in the database. The weighted i^{th} entry in the histogram, denoted $\bar{\theta}_i$, then becomes,

$$\bar{\theta}_i(X) = \theta(x_i) \cdot \log \frac{N+1}{DF(v_i)+1}.$$ Equation 6-14

Alternatively, a probabilistic formulation can be used,

$$\bar{\theta}_i(X) = \theta(x_i) \cdot \log \frac{N - DF(v_i)}{DF(v_i)}.$$ Equation 6-15

6.2.5 Experimental

Profiling was carried out to find optimal parameters for virtual screening using the AMPC, CXCR4 and GCR targets from the DUD-E data sets, as in the previous chapter, and Wave Kernel Signature (WKS), $D = 100$ was used as the local geometry descriptor for efficiency purposes, unless otherwise stated. Each target had similarity searches performed on 50 randomly selected active molecules from that target and the results were collated to provide average performance statistics with confidence intervals. The virtual screening statistics reported are AUC for full ranking performance and enrichment factor 0.5% for early retrieval performance. To balance the two, the BEDROC statistic with $\alpha = 20$ is also reported. Virtual screening results are depicted using point plots representing the mean of 50 reference molecules with 95% confidence intervals calculated using bootstrapping with 10,000 iterations. All code was written in Python using Python scientific computing libraries NumPy, Pandas, SciPy, and Sci-kit Learn. The figures were plotted using the Python plotting library seaborn.

The parameterisation experiments presented here are preliminary results as a proof-of-concept of the descriptors. Due to the long computation time that is compounded by the large number of molecules in the data set, it was not feasible to carry out a more detailed set of experiments. Consequently, and as discussed in Chapter 5, the parameterisation experiments are carried out on three targets. Therefore, there is a risk that the results presented are not generalisable.

6.3 Results

The parameterisation steps are broken down into parameters for codebook generation and parameters for encoding and producing the final histogram descriptor. The codebook parameters tested are the number of molecules in a sample used to train the K-Means codebook and the number of codewords used. The encoding and production of the histogram parameters are: the encoding method; the method by which the results are pooled over the molecule; and the normalisation applied at the end. This is followed by a visual investigation of different codebook encodings by mapping features onto the surface of a molecule, and an investigation of the sparsity of the histograms. Finally, the selected parameters are tested on a WKS local geometry descriptor of higher dimensions than used in the previous chapter. Finally, the best parameters are tested in virtual screening on the whole DUD-E data set.

6.3.1 Codebook parameterisation

As mentioned above, the codewords included in the codebook are chosen from a sample of molecules using a K-Means clustering algorithm. To test the effect of the number of molecules in the sample, codebooks were computed using 50 codewords with WKS, $D = 100$ local geometry descriptors, using a range of sample sizes from 100 to 1000 molecules at intervals of 100. The parameters were tested using virtual screening experiments on three targets, APMC, GCR, and CXCR4, and 50 randomly selected reference molecules for each target. The AUC, BEDROC and Enrichment factor at 0.5% values are summarised in Figure 6-4. In all cases in this section, the final histograms are generated using the summation method followed by normalising by dividing by the number of vertices. The results are presented as point plots with the 95% confidence intervals that have been computed using 10,000 bootstrap iterations. The results in Figure 6-4 are inconclusive and exhibit a large amount of variation. This variation can be explained by target-specific variation in performance.

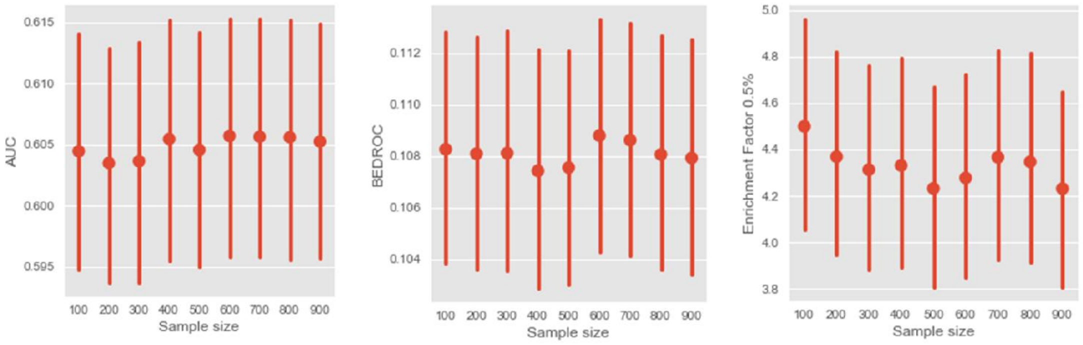


Figure 6-4. Virtual screening statistics for 50 reference molecules by increasing the number of molecules in the codebook training sample.

Figure 6-5 plots the results by target. The results indicate that there is no significant performance benefit for increasing the number of molecules in the sample, suggesting that sufficient information is captured in a smaller sample. In comparison to Figure 6-4, it becomes clear that there is a large amount of variation in the performance of different targets. Furthermore, the confidence intervals suggest that the variation within targets is target specific. Over the three measures, CXCR4 has the most variation whereas, with the exception of AUC, GCR has the smallest variation. Additionally, CXCR4 performs significantly better than the other two targets.

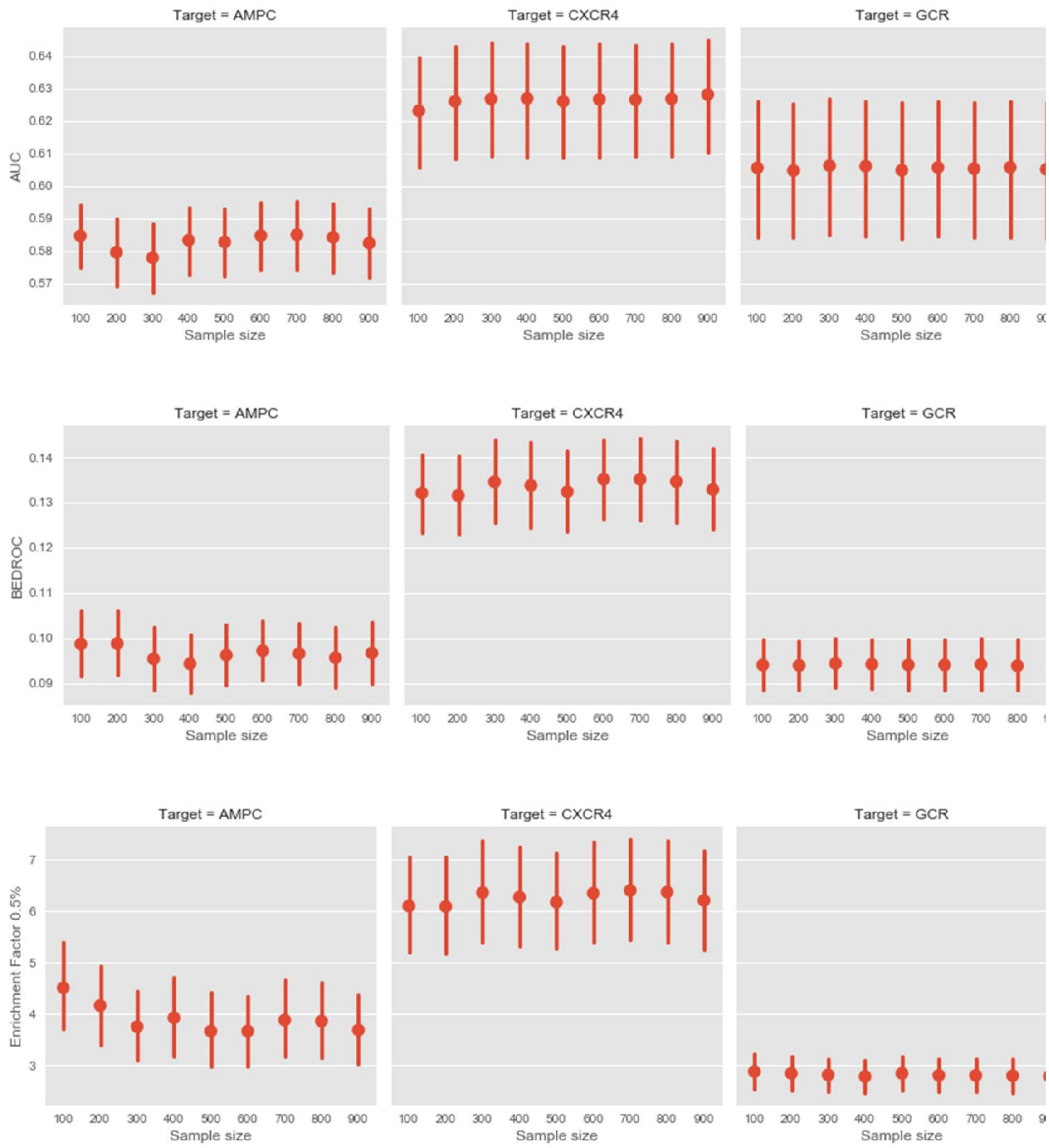


Figure 6-5. Target specific virtual screening statistics for 50 reference molecules by increasing the number of molecules in the codebook training sample.

The previous experiment was repeated using the Mini-batch K-Means algorithm to investigate whether the improved computational performance had a significant virtual screening penalty. The results are presented in Figure 6-6 with 95% confidence intervals calculated as above with the Mini-batch algorithm presented in blue. The results suggest that there is no penalty in using the Mini-batch algorithm. The Mini-batch results closely match those from K-Means with the benefit of significantly improved computation times, which is illustrated in Figure 6-6 showing the computation times for increasing sample sizes of WKS, $D = 100$ local geometry descriptors. Subsequently, for the remainder of this chapter the codebook will be computed using a sample size of 600 and the Mini-batch K-means algorithm.

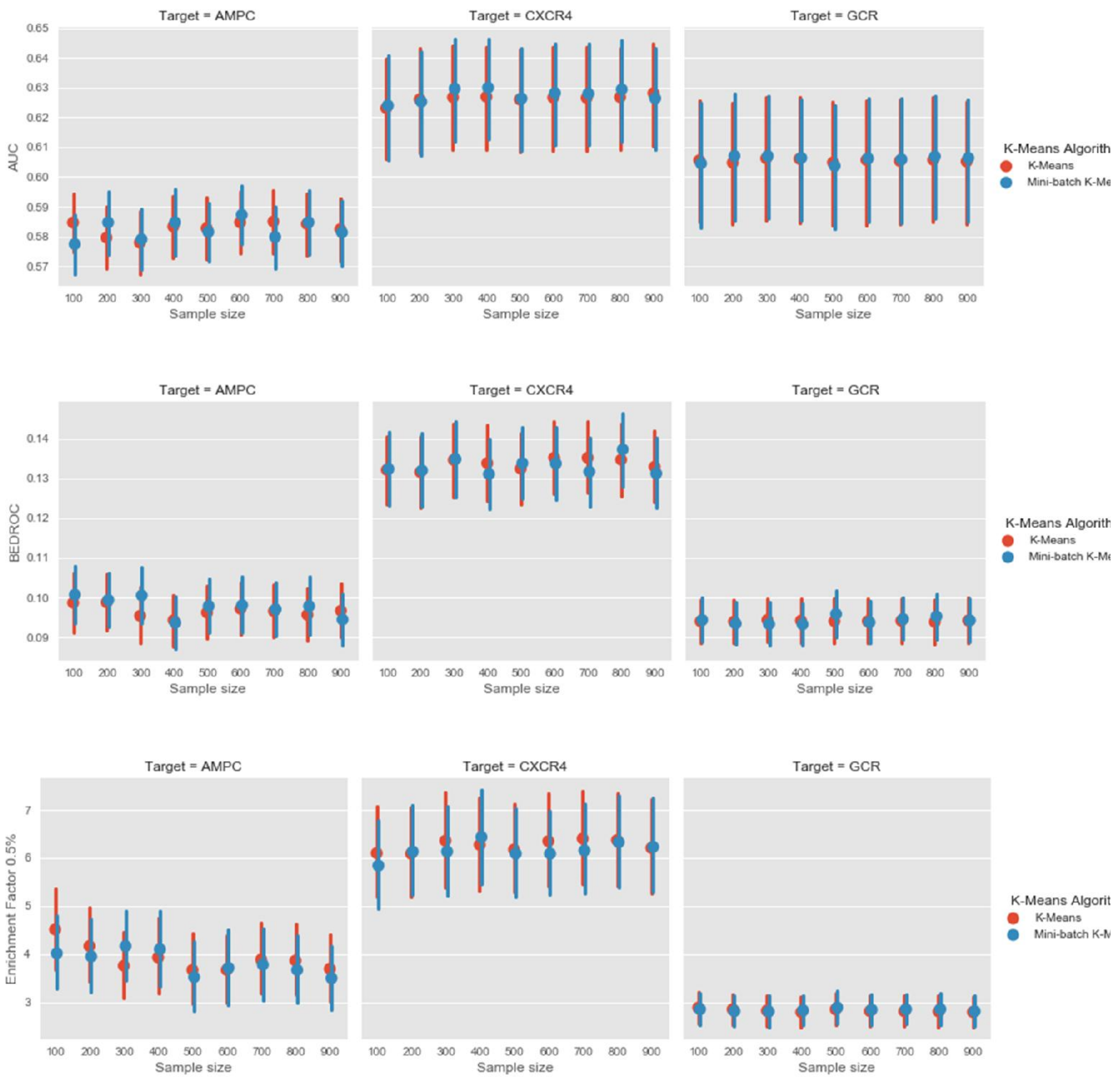


Figure 6-6. Target specific virtual screening statistics for 50 reference molecules by increasing the number of molecules in the codebook training sample for K-Means and the Mini-batch K-Means algorithm.

To find the optimal number of codewords in the histogram, a virtual screen of fifty reference molecules was run on the three profile targets using both the hard vector quantisation and soft allocation vector quantisation encodings. The results are presented in Figure 6-7, which shows the virtual screening performance using AUC, BEDROC, and Enrichment factor 0.5% with 95%

confidence intervals computed using bootstrapping 10,000 iterations. Figure 6-7 shows that the virtual screening performance for the AMPC target is unaffected by the number of words in the codebook. However, for the CXCR4 and GCR targets there is a conflict between the AUC results and the Enrichment factor, with the AUC performance decreasing slightly with increasing number of codewords and stabilising at around 500 words, whereas the Enrichment factor increases with the number of codewords and stabilises at around 700 words. The BEDROC statistics provide a balance of the AUC statistic and the early discovery that is captured in the Enrichment factor and shows that there is no significant increase in performance for CXCR4 and GCR with an increase in the number of codewords in the codebook.

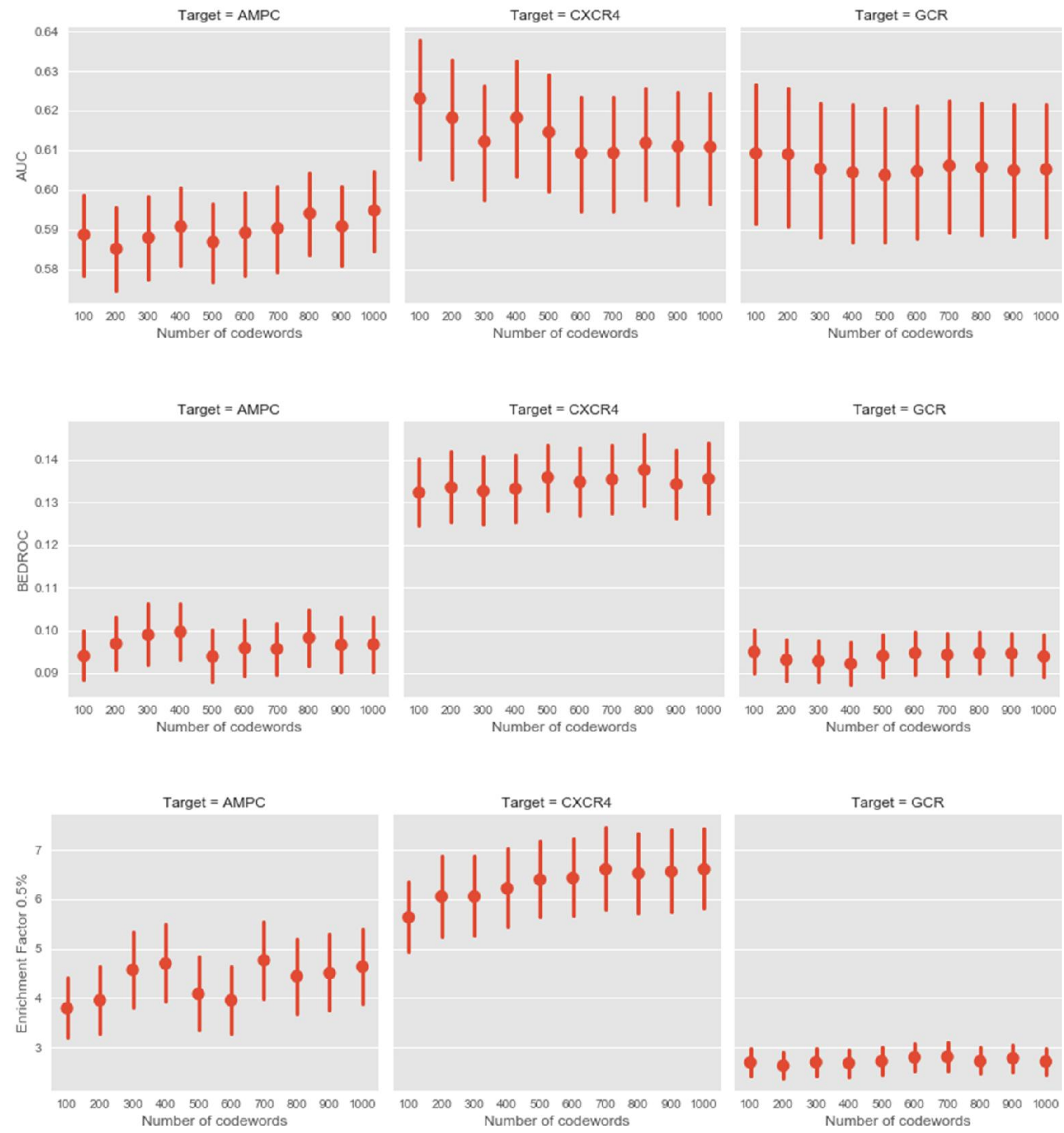


Figure 6-7. Target specific virtual screening statistics for 50 reference molecules by increasing the number codewords in the codebook.

6.3.2 Histogram encoding and production

In order to investigate the findings in Figure 6-7 further and to obtain an insight into the performance of individual encoding methods, the data from Figure 6-7 are plotted in Figure 6-8 and separated by encoding type. This shows that the soft allocation vector quantisation encoding exhibits consistently poorer performance than the hard vector quantisation encoding, and suggests that it is the soft allocation vector quantisation encoding that is pulling down the average performance in Figure 6-8. The figure also shows that the AUC statistic is noisier. In fact, the soft allocation vector quantisation encoding consistently gives higher AUCs for the AMPC and GCR target. In contrast, the Enrichment factor is more decisive where the hard vector quantisation performs significantly better for all targets. For example, in CXCR4 the hard quantisation encoding performs significantly better than the soft allocation vector quantisation encoding for codebook sizes greater than 300. This is balanced out in the BEDROC statistic, which shows the hard vector quantisation descriptor is still significantly better than the soft allocation vector quantisation encoding in the CXCR4 target for codebooks of greater than 300 words, and in all cases in the GCR target.

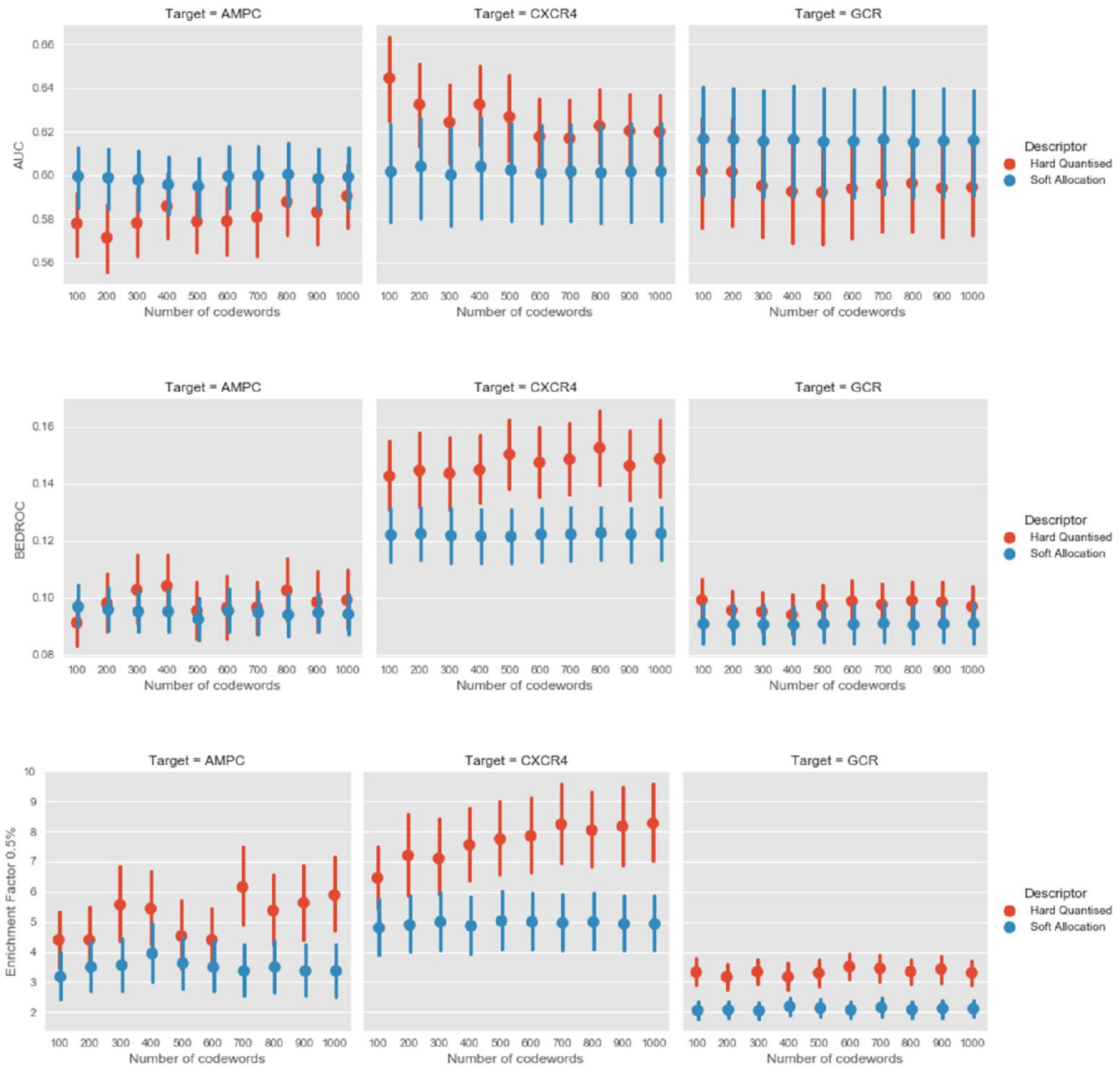


Figure 6-8. Comparison of hard quantisation and soft allocation encoding methods using target specific virtual screening statistics for 50 reference molecules by increasing the number codewords in the codebook.

The K-nearest neighbours encoding is proposed as a balance to the information loss of the hard vector quantisation and the noise of the soft allocation vector quantisation encoding. Figure 6-9 shows the performance of the K-nearest neighbour encoding with an increasing sample of K-neighbours. Figure 6-9 shows that there is an improvement in average AUC performance in the CXCR4 target, which would be consistent with the encoding becoming closer to a soft allocation vector quantisation encoding with increasing K. However, there is no observable change in the AMPC and the GCR target. In a similar way, there is a decrease in average Enrichment factor as K increases reflecting that the descriptor is becoming less like the hard vector quantisation encoding. The BEDROC statistics show a balance between these two observations, with a small decrease in average BEDROC performance for an increase in K. However, it is clear that there is a larger amount of variance in the underlying performance for all the targets making it hard to discern a clear performance pattern.

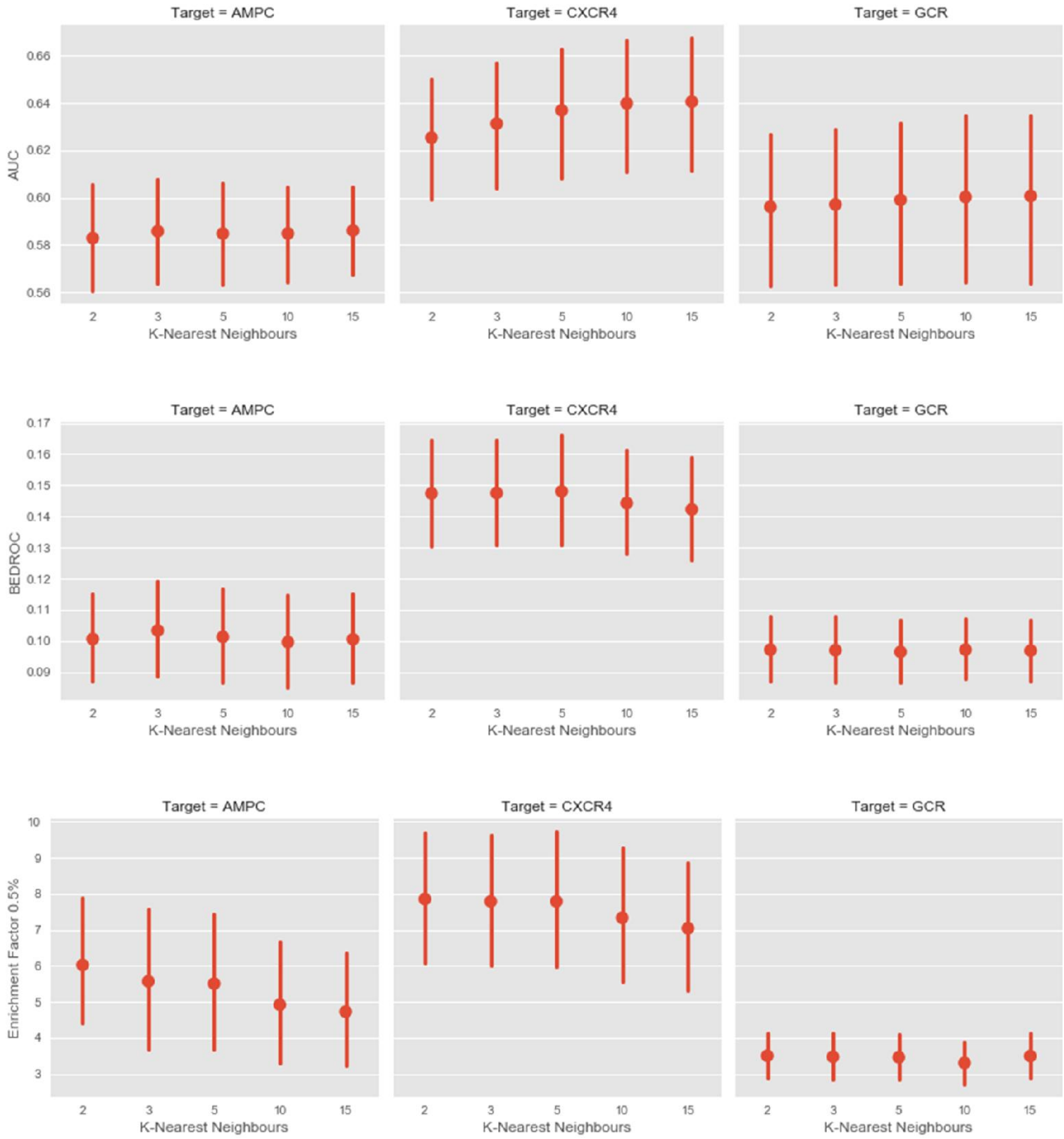


Figure 6-9. Target specific virtual screening statistics for 50 reference molecules using a K-nearest neighbours encoding method with increasing numbers of K-nearest neighbours.

Figure 6-10 summarises the performance of the different encoding types across all codebook sizes and K-nearest neighbours. It shows that virtual screening performance is dependent upon the choice of statistic. Soft allocation vector quantisation performs significantly better than hard vector quantisation in AMPC and GCR using AUC as a metric. Whereas hard vector quantisation performs significantly better than soft allocation vector quantisation in all targets using the enrichment factor, suggesting that it is an improved early retrieval rate. Finally, these statistics are weighted in the BEDROC statistic which shows the hard vector quantisation has better performance in all three targets on average. However, in the AMPC and GCR targets the confidence interval bars overlap suggesting that the null hypothesis that there is no difference in performance between the two encoding types cannot be rejected. The K-nearest neighbours is intended to provide a balance between information loss and noise and Figure 6-10 shows that on average it does have an improved performance over the hard vector quantised encoding in AUC as well as Enrichment factor in AMPC and GCR. Overall the K-nearest neighbours encoding has the best average performance in the BEDROC statistic for AMPC and GCR, and is on a par with hard vector quantisation in CXCR4. Nevertheless, the improvement on average comes at the cost of an increase in the variance, which is clear in Figure 6-10. Consequently, it is not possible to make statistically significant judgements about overall performance in comparison to hard vector quantisation, whereas both hard vector quantisation and K-nearest neighbours are both statistically significantly better than the soft allocation vector quantisation descriptor. Therefore, it can be hypothesised that the hard vector quantisation captures the main features such that shapes that share the dominant features are promoted leading to early retrieval. However, there is a large amount of information loss in the hard vector quantisation encoding that restricts capturing active shapes with a number of smaller, less prominent, features in common and thereby reduces performance in the AUC, which measures the overall ranking.

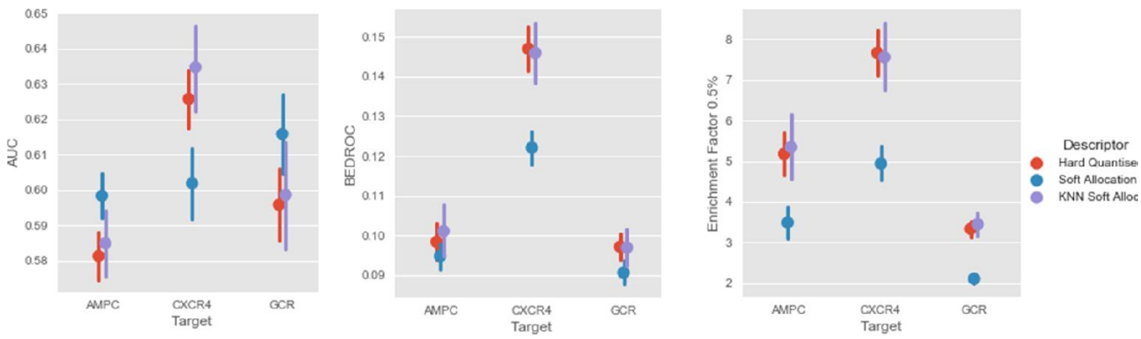


Figure 6-10. Collated results to compare the hard vector quantisation, Soft allocation vector quantisation, and K-nearest neighbour descriptors.

6.3.3 Histogram pooling

In all results reported so far, the histograms were pooled by summing the frequency of the codewords, and normalised by dividing by the number of vertices. Spatially sensitive pooling was tested using the same virtual screening experiments to see whether the spatial information could be encoded in a meaningful way. For efficiency reasons, the surface was sampled using the farthest point sampling method using the Euclidean distance matrix. Figure 6-11 shows the results for the three targets. The results in the AMPC target are unusually worse than those for GCR, which is normally the more difficult target. The SSE descriptor performs best based on the AUC measure, and exhibits the highest AUC in CXCR4 and GCR of the descriptors tested so far but performs poorly on AMPC. However, the early enrichment statistic is surprisingly poor in all targets, which drags down the overall BEDROC score, especially in AMPC. This may be due to the properties of the shapes in GCR that have well separated features over the surface meaning the spatial encoding is effective at distinguishing actives and decoys over the whole data set but is not specific enough for promoting actives to the top of the rankings for the enrichment factor. However, further investigation is required to evaluate this hypothesis.

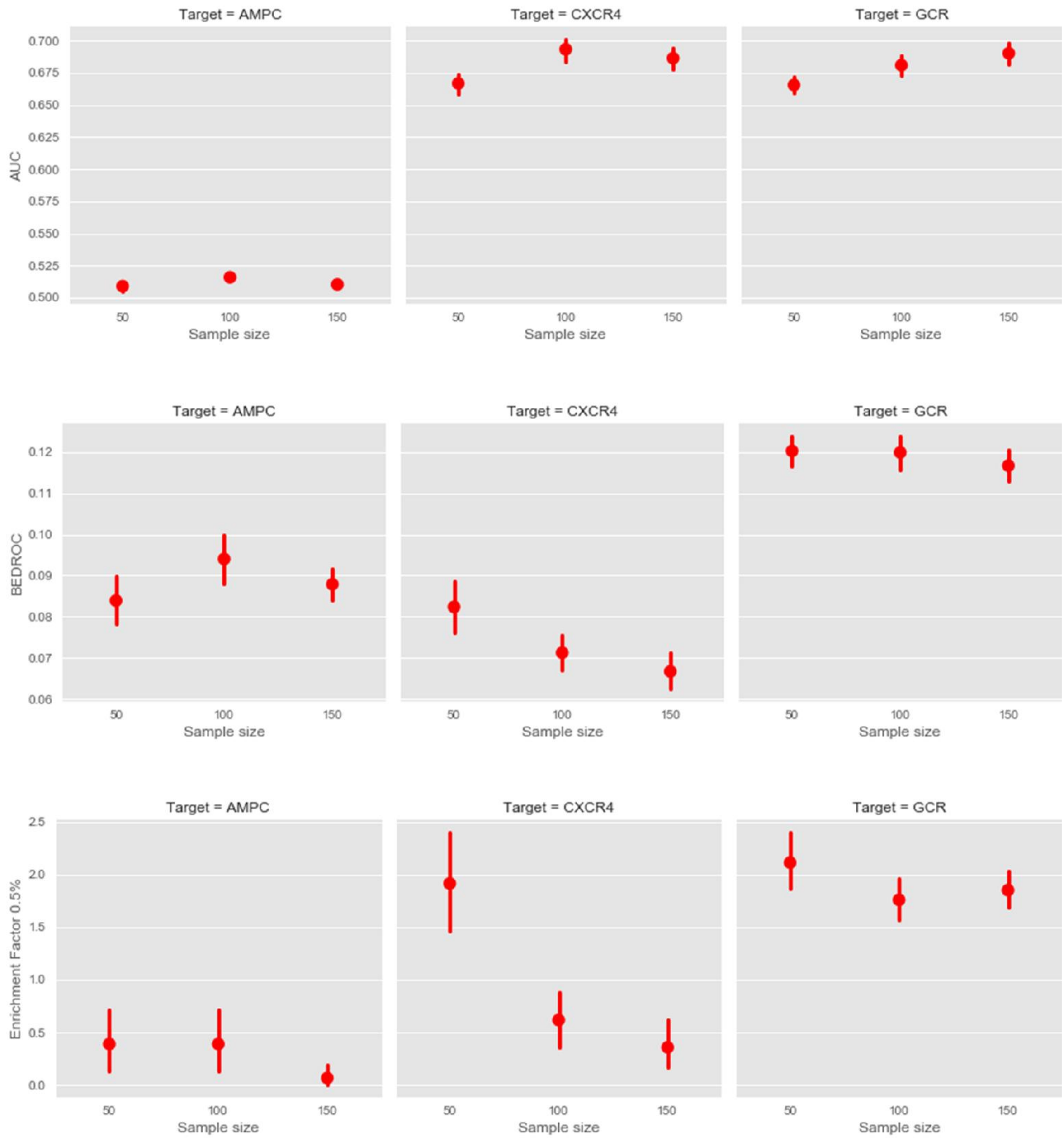


Figure 6-11. Target specific virtual screening statistics for 50 reference molecules using spatially sensitive encoding method with increasing numbers of points sampled on the surface.

When carrying out the spatially sensitive pooling, the Euclidean distance was used for the farthest point sample as the diffusion distance was found to be computationally expensive. As an illustration, the computation time for computing the diffusion distance matrix was found to be quadratic with respect to the number of vertices in the mesh with a maximum time of around 100 seconds to compute the distance matrix (Figure 6-12). In comparison, the computation time took a maximum of 1.2 seconds for the Euclidean metric, suggesting that the worst case time is in the order of 100 times worse for the diffusion metric. This is due to the Euclidean distance matrix being computed using low level numerical computing libraries, such as BLAS and MKL, that are heavily optimised.

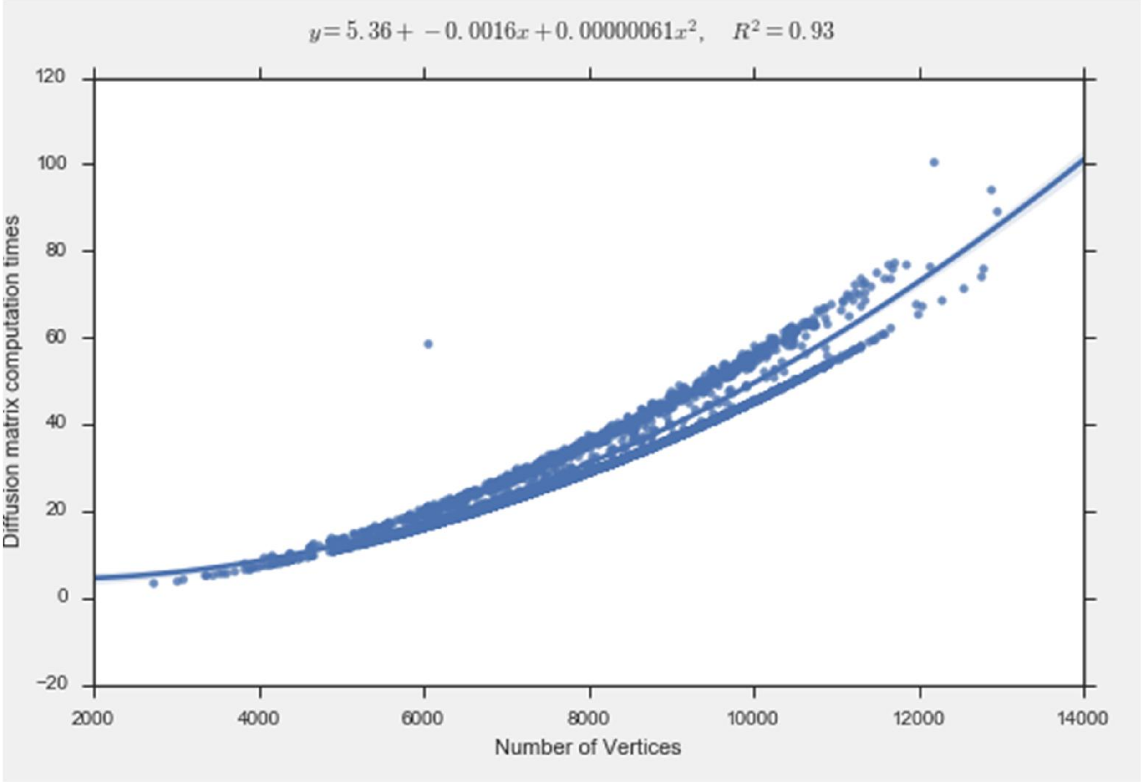


Figure 6-12. Computation time of the diffusion metric distance matrix by the number of vertices in the mesh.

6.3.4 Histogram normalisation

The term frequency inverse document frequency (tf-idf) weighting method was tested to see whether it made a difference to the overall performance of the descriptors using a K-Means codebook with different numbers of codewords. The results are presented in Figure 6-13. It is interesting to note that in general it did not improve the performance of the descriptors considering the early enrichment measures. In fact, for the majority of the descriptors, the encoding performed identically when compared to normalising by the number of vertices. One possible explanation for this is to recall again that the feature space of the local geometry descriptors is not particularly rich. In which case, the molecules are likely to have a large number of the important features of the shape and codewords that occur less frequently are more likely to be noise or artefacts of an encoding method, such as the hard vector quantisation. Visual inspection of the descriptors showed that the weightings altered the dominant features but appeared to have the same effect on all histograms leaving their rankings relatively unchanged.

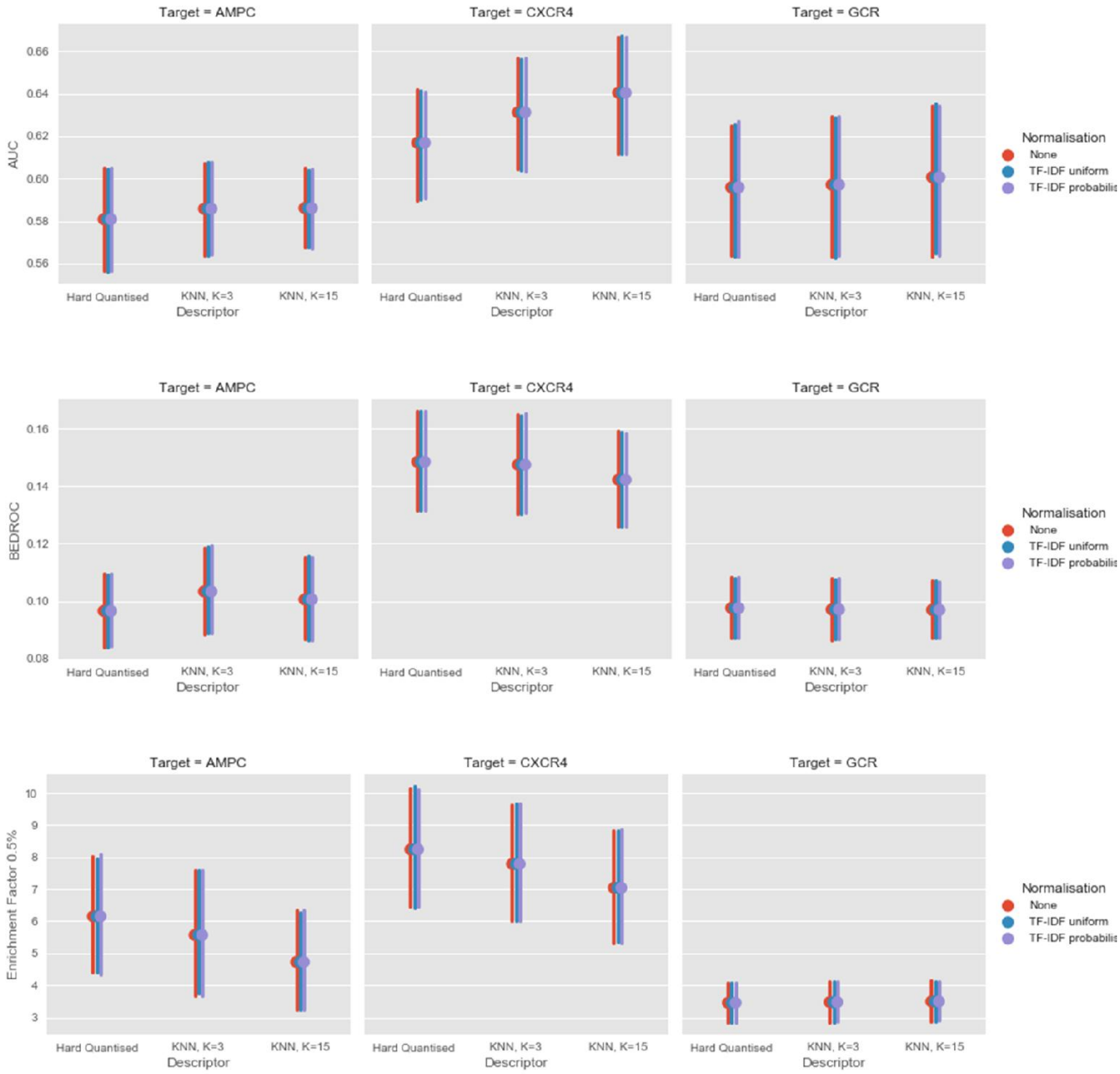


Figure 6-13. Comparison of tf-idf normalisation weights for virtual screening of 50 compounds.

6.3.5 Encoding visualisation

To develop a clearer understanding of the relationship between the codebook, the local geometry descriptor, and the encoding schemes, a series of plots are presented. Figure 6-14 shows the hard vector quantisation encodings plotted on the surface of a sample molecule where each colour represents a different codeword. Figure 6-14 (a) shows the encoding of a codebook with 100 codewords using a WKS local geometry descriptor with dimension, $D = 100$, whereas Figure 6-14 (b) shows an encoding using a codebook of the same size but with a WKS local geometry descriptor with dimension, $D = 1000$. In both cases it is striking that relatively few colours are plotted, which suggests that the vast majority of shape features are captured in relatively few codewords. In Figure 6-14 (a) large patches of the surface are represented by a handful of colours. While the reverse of the molecule cannot be seen in the image, it can be interpreted that the WKS, D=1000 local geometry descriptor uses more of the 100 features in the codebook because there is a greater variety of colours, including yellows and greens. Furthermore, the patches that are coloured are smaller, although this is hard to see because a large part of the image is coloured with very similar greens. This supports the theory and findings in chapter 4 that the higher dimension WKS descriptors exhibit a greater amount of feature localisation.

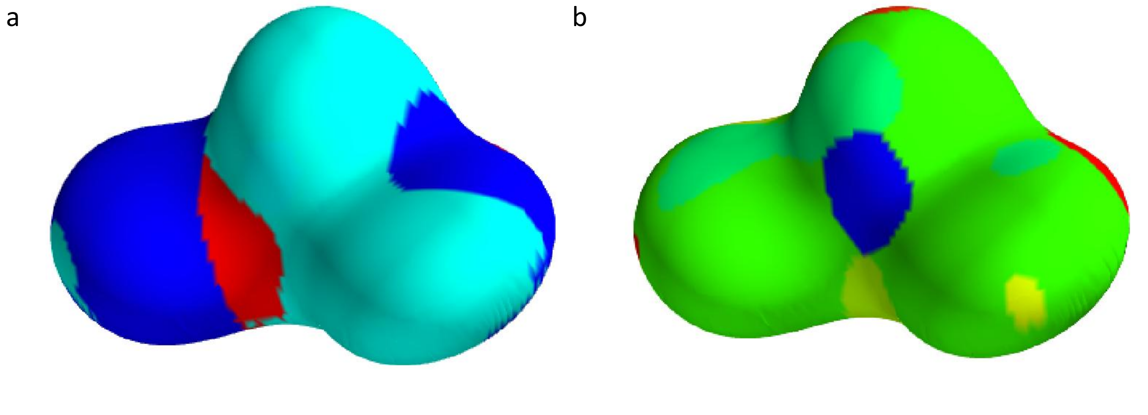


Figure 6-14. Hard vector quantisation encoding on the same molecule using a) WKS, $D = 1\ 00$and b) WKS, $D = 1\ 000$

To get a greater insight into the results in Figure 6-14, Figure 6-15 presents the same molecule with the surfaces coloured with respect to the softmax distance from a given codeword in the codebook. The codewords have been deliberately selected so that they appear to be encoding similar features. Figure 6-15 (a) and (b) plots the softmax distance from two different codewords for WKS, $D = 100$ and WKS, $D = 1000$ respectively. The subplots (i) and (ii) are codewords that have been selected to approximate the same features. In general, the softmax distances for WKS, $D = 100$ are more uniform over the shape than WKS, $D = 1000$. Therefore, the encodings from WKS, D=100 would be expected to be more global than D=1000, which are more local.

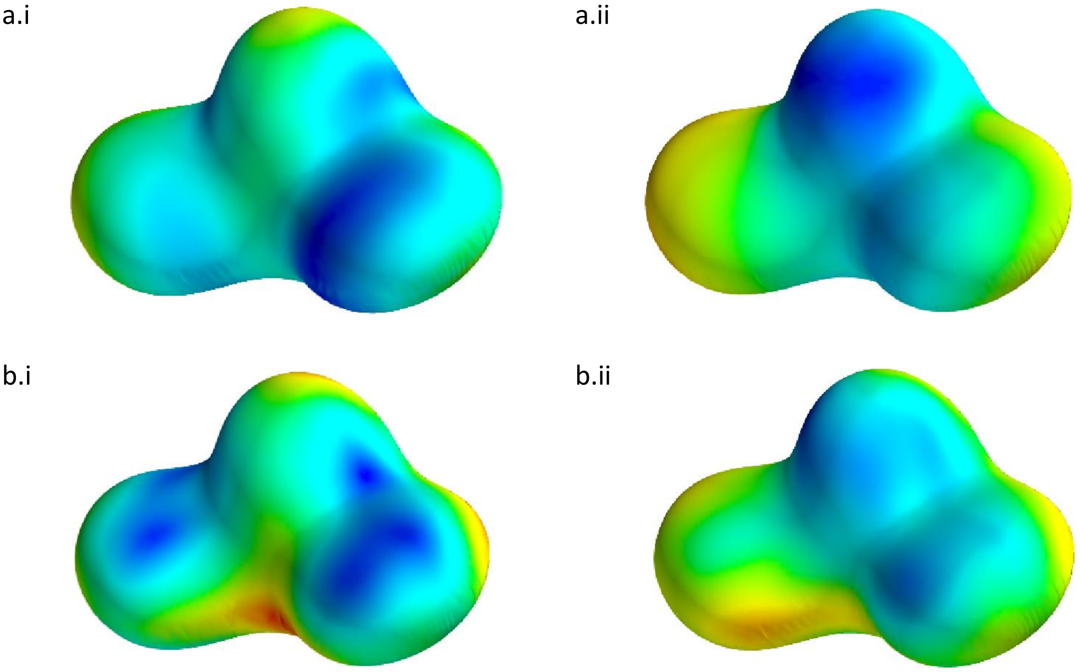


Figure 6-15. Distance to codewords for equivalent codewords – i and ii – using a) the WKS, $D = 1\ 00$and b) and b) the WKS, $D = 1\ 000$local geometry descriptor.

These observations are best observed in the resulting histograms (Figure 6-16). The hard vector quantisation descriptor of Figure 6-16 (a) is dominated by two codewords, the 28th and 47th, with only three other codewords having a small proportion of the descriptor, the 44th, 60th and 93rd codewords. However, in the soft allocation vector quantisation descriptor, the relative value of the 47th codeword is significantly smaller, and almost indistinguishable from other codewords in the codebook (Figure 6-16 (b)). Finally, Figure 6-16 (c) shows the k-nearest neighbours histogram for k=3. In this descriptor, the prominence of the 47th codeword is still strong and other codewords that were also in the hard vector quantisation descriptor, such as the 60th and 44th are amplified. In addition, some codewords that had zero values in the hard vector quantisation descriptor are assigned non-zero values, for example the 24th codeword. This indicates that there are relatively few features that are important in describing the molecular shape. Additionally, the soft allocation vector quantisation descriptor appears to add a significant amount of noise.

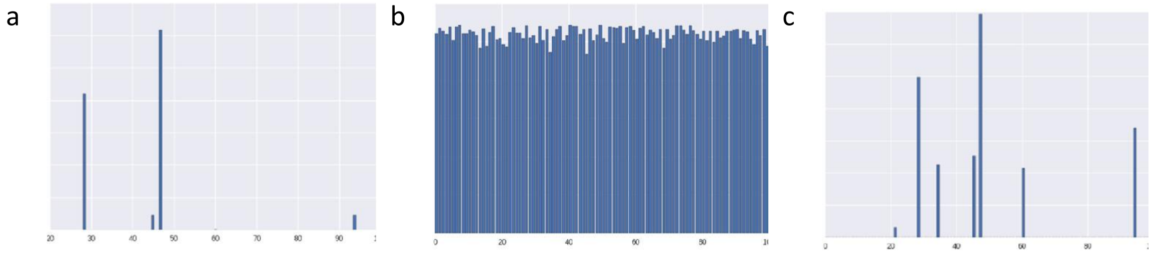


Figure 6-16. An illustration of the final descriptors using a) hard quantised encoding, b) soft allocation encoding and c) K-nearest neighbours encoding, $k = 3$.

6.3.6 Higher local geometry descriptor dimensions

Figure 6-14 provides evidence to suggest that higher dimensions of the WKS local geometry descriptor give more granular and localised point-wise descriptors than lower dimensions. Therefore, given that the Bag of Features descriptor describes the shape in a row-wise manner, it was hypothesised that this improved granularity would improve the performance of the descriptor. Furthermore, as the size of the descriptor depends on the size of the codebook, rather than the local geometry descriptor, then there would not be an additional space penalty in storing Bag of Features descriptors derived from higher order local geometry descriptors. Therefore, the previous virtual screening experiment was performed using WKS local geometry descriptors with dimensions, $D = 500$, $D = 700$, and $D = 1000$ using hard vector quantisation encoding. The results using codebooks of size 100, 500, and 700 have been combined so that each point represents the average results using three different codebooks. Interestingly, the results indicate that there is no improvement in virtual screening performance for an increase in local geometry descriptor size (Figure 6-17). In fact, the higher dimension local geometry descriptors performed worse on all performances measurements, which is contrary to the covariance descriptors in Chapter 5 (Figure 5-5). This may be due to the retrieval performance being dependent upon more global shape features.

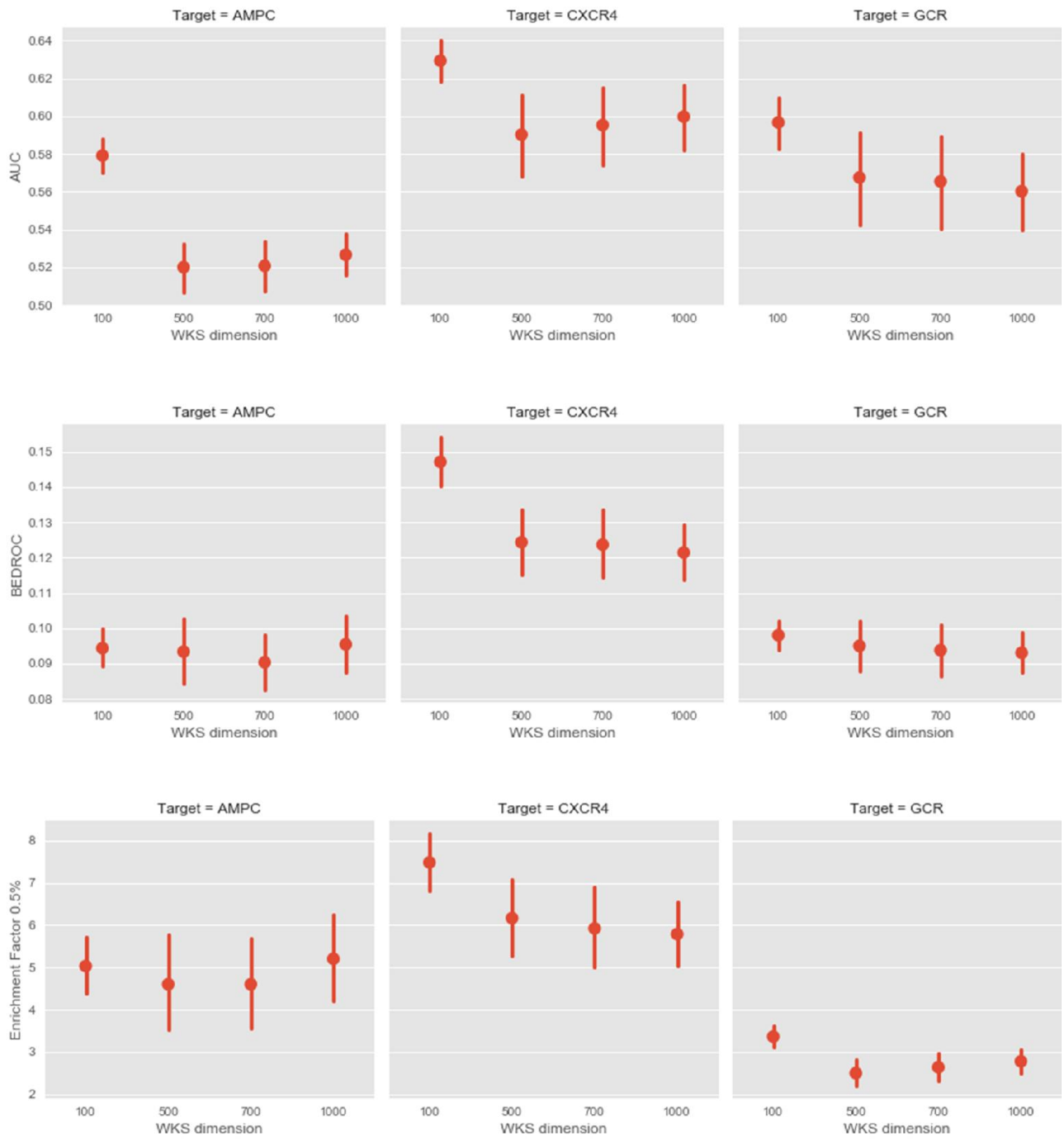


Figure 6-17. Comparison of the hard vector quantisation encoding descriptors using increasing dimensions of the WKS descriptor.

6.3.7 Virtual screening on the full DUD-E data set

The optimal parameters identified above were carried forward to a series of virtual screening experiment using the full DUD-E data set consisting of 102 targets, reported in section 4.4.1. The experimental set up was the same as reported in Chapter 5.6 and the Bag of Features descriptors were tested against Shape-it, CDK Shape Moments, and the covariance matrix descriptors for

WKS, $D = 100$ and $D = 1000$. The bag of features descriptors used were computing using a 700 word codebook with the hard quantised (HQ), and k-nearest neighbours descriptors with $K = 3$ and $K = 10$ encodings. The results across all targets are summarised in Figure 6-18.

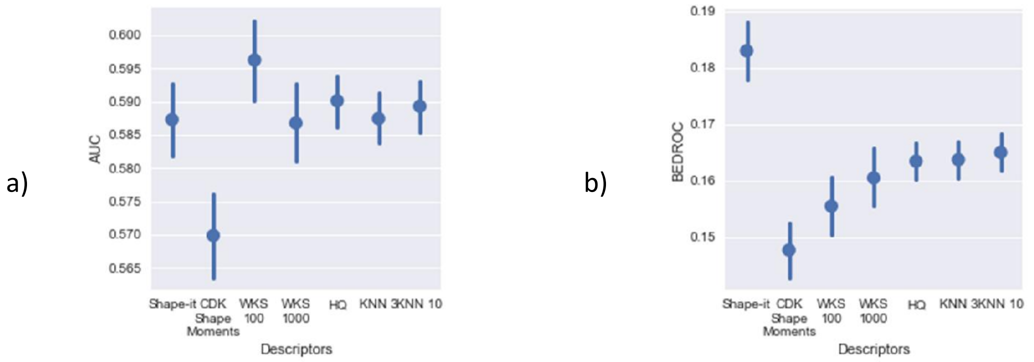


Figure 6-18. Bag of Feature descriptor comparison against benchmark shape comparisons on the DUD-E data set.

When comparing against the alignment-free established method, all Bag of Features methods perform significantly better than the CDK Shape Moments descriptor using the both AUC and BEDROC, $\alpha = 20$. Interestingly, the Bag of Features descriptors have a similar performance to Shape-it using the AUC metric, with hard vector quantisation and K-nearest neighbour, $K = 10$, performing better on average, although it is not possible to reject the null hypothesis of no difference at the 95% confidence level. Additionally, the bag of features descriptors performed similarly to the covariance descriptor for WKS, $D = 1000$. With respect to early retrieval, the Bag of Features descriptors perform better on average than all the covariance descriptors, although it is not possible to reject the null hypothesis of no difference at the 95% confidence level. Nevertheless, Shape-it remains the best performing method using BEDROC.

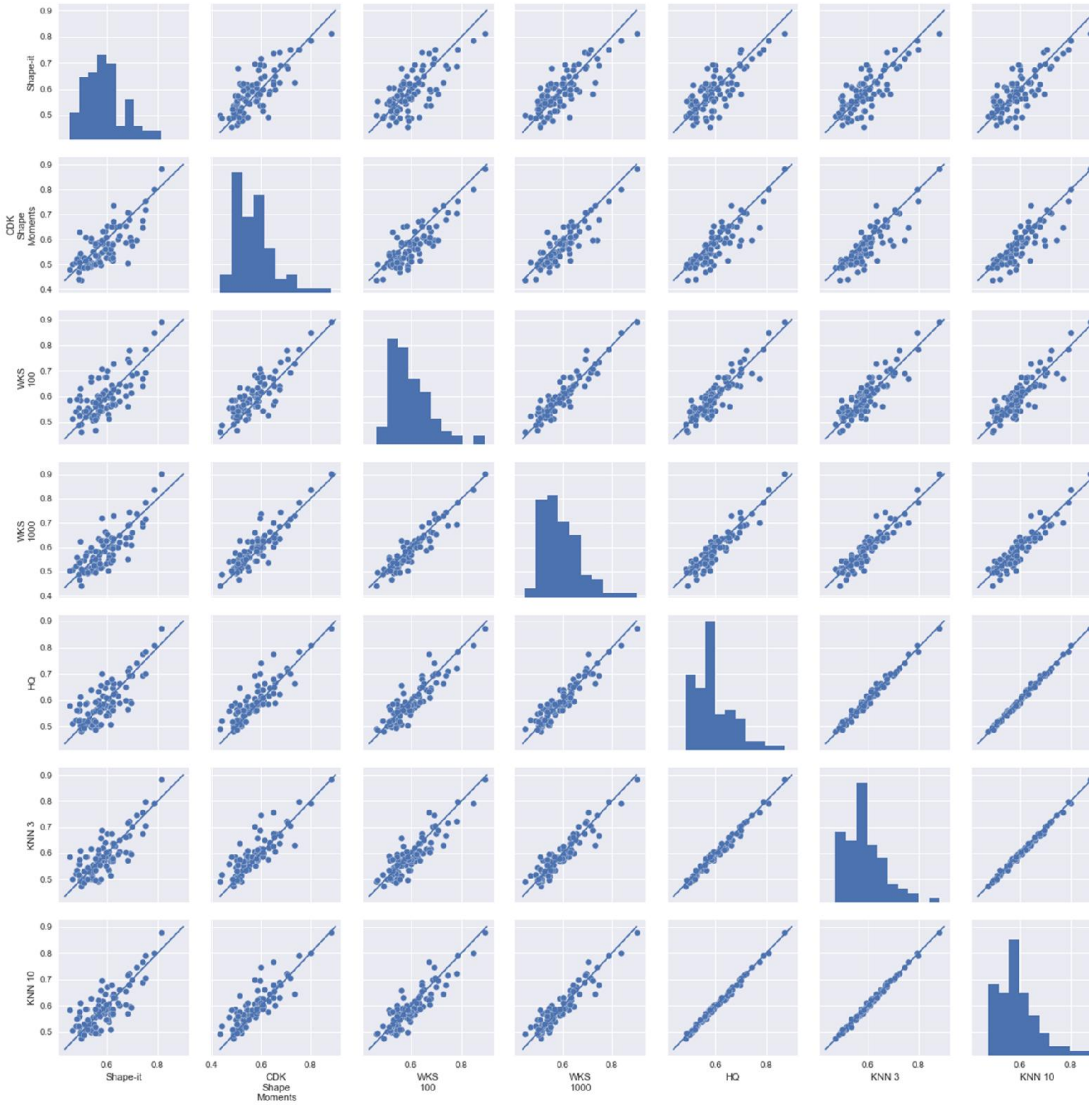


Figure 6-19 Pairwise AUC comparison of the virtual screening results for the full DUD-E data set.

Figure 6-19 and Figure 6-20 show the pairwise comparisons of the different shape methods across all targets in DUD-E using the AUC and BEDROC statistics, respectively. As in Chapter 5.6, the leading diagonal represents the distribution of values for a given shape method and the off-diagonal figures are scatter plots of the values obtained using different methods. The solid line represents equal performance between the two methods so that points above the line show better performance for the method on the y-axis of the scatter plot and points below the line show better performance for the method on the x-axis of the scatter plot.

The histograms in Figure 6-19 show that on the whole, all methods have the majority of AUC values below 0.6, with all methods having some values lower than 0.5. As with Chapter 5.6, Shape-it generally performs worse than other methods when both methods return high AUC values. This is demonstrated in the Shape-it row where most points in the top right quadrant are below the solid line. In general, the Bag of Features descriptors all performed very similarly, with the points being clustered around the diagonal. However, when compared against the covariance descriptors, it appears that there is a cluster of mid-range AUC values above the diagonal where the Bag of Features descriptors performs better. In other words, as observed in Chapter 5, the WKS descriptors perform better than most descriptors for targets in which both methods perform comparatively poorly and well.

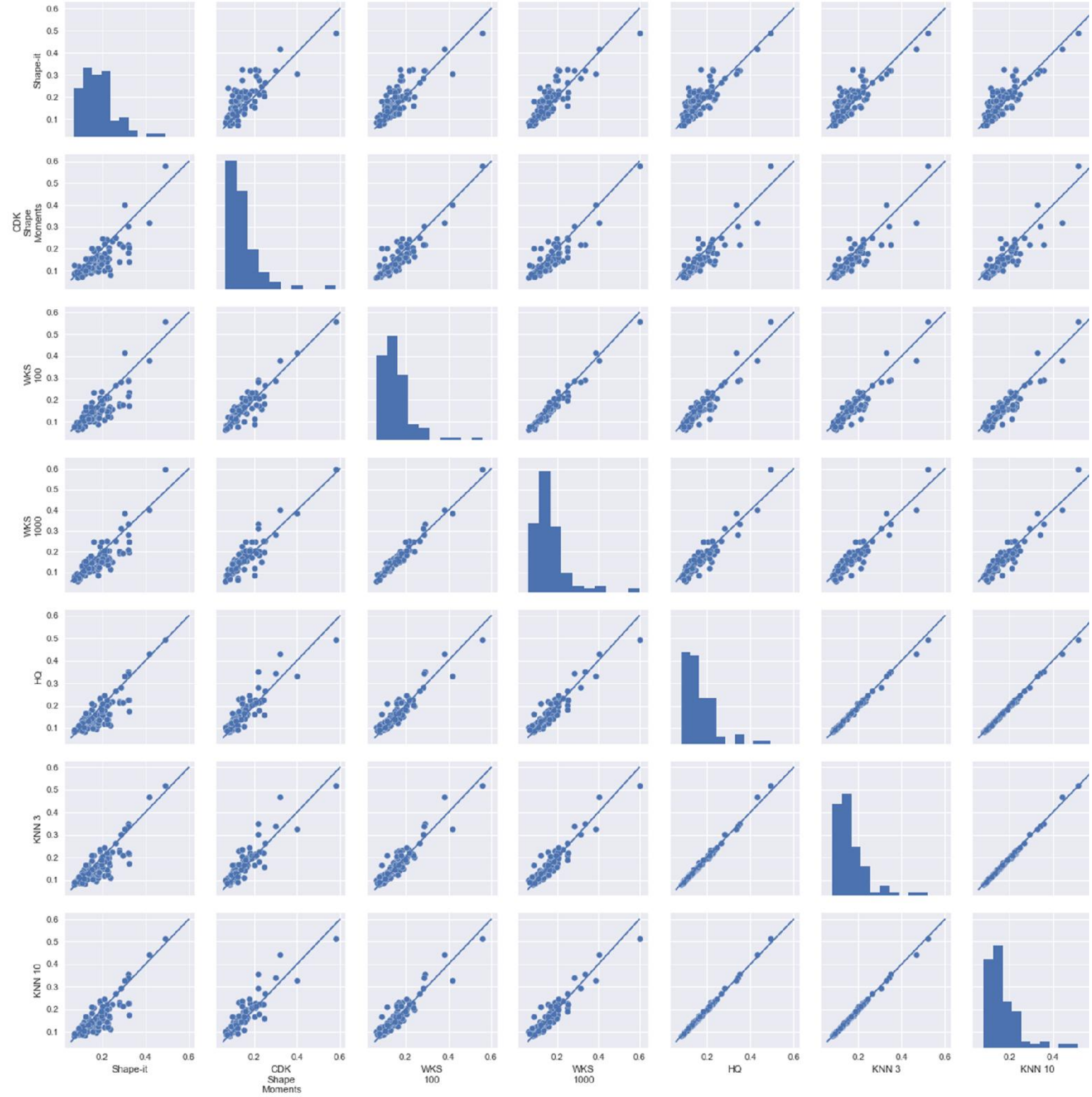
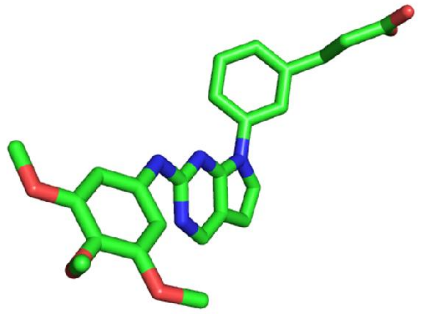


Figure 6-20. Pairwise BEDROC comparison of the virtual screening results for the full DUD-E data set .

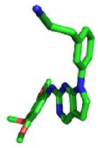
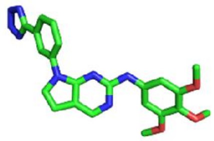
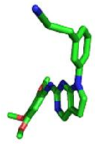
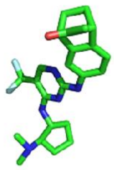
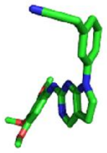
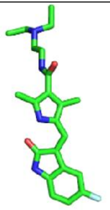
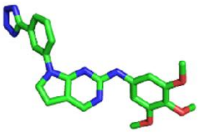
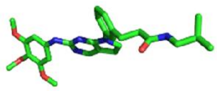
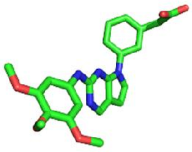
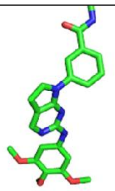
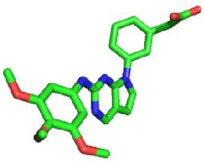
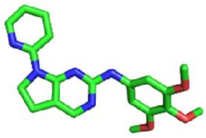
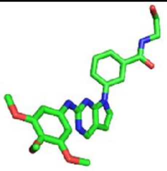
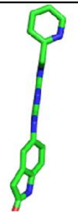
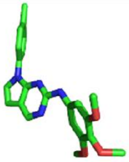
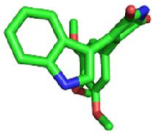
Figure 6-20 shows that all methods have a distribution that is skewed towards the lower end of the BEDROC values. When compared against the other methods, Shape-it performs best across the board, although, when both methods perform well, Shape-it generally has a lower BEDROC value than the other method. Furthermore, when compared against the spectral geometry descriptors, the CDK Shape Moments descriptor performs worse in nearly all targets. As with the AUC figures, the Bag of Features methods are all highly correlated to each other. The covariance descriptors are also closely correlated with the Bag of Features descriptors, albeit to a lesser extent.

6.3.8 Comparison of molecules retrieved

The above experiments provide results on the overall information retrieval performance of the different methods, however, they do not distinguish between the molecules have been retrieved. Figure 6-21 and Figure 6-22 show examples from the top retrieved actives in the FAK1 and LCK target for four different methods with the reference molecule given above. This gives an opportunity to inspect how the different methods define shape similarity and how they may be combined to produce a complimentary set of candidate drug molecules. Figure 6-21 illustrates the results for FAK1, in which the Shape-it method out-performed the other methods. The top Shape-it actives all have a common substructure with the fused ring and attached to a ring with three carbonyl groups by an amine linker. This is the same for the actives retrieved using the WKS, $\mathrm{D} = 100$ covariance matrix, suggesting that the WKS filter over the surface emphasises similar shape features to the Shape-it volume. On the other hand, while the CDK Shape Moments and the K-Nearest Neighbours methods also return molecules with the same substructural feature, they also rank more diverse structures highly. This is especially the case in the K-Nearest Neighbours descriptor with molecules 1 and 4.



FAK1 reference molecule Active id: 88

	Shape-it	CDK Shape moments	WKS 100	KNN 3
1	 Active id: 63	 Active id: 61	 Active id: 68	 Active id: 109
2	 Active id: 63	 Active id: 70	 Active id: 61	 Active id: 107
3	 Active id: 87	 Active id: 75	 Active id: 87	 Active id: 102
4	 Active id: 83	 Active id: 55	 Active id: 48	 Active id: 73

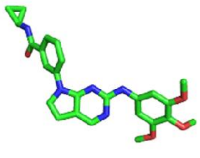
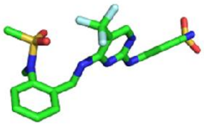
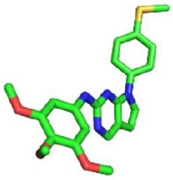
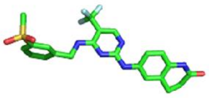
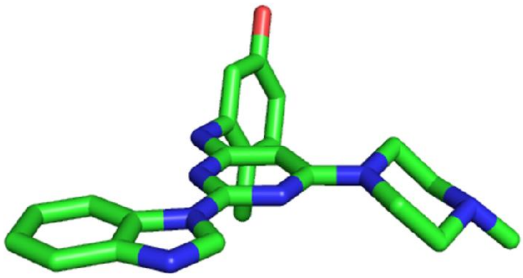
5				
	Active id: 59	Active id: 9	Active id: 92	Active id: 80

Figure 6-21. Comparison of top retrieved actives for 4 different methods using an example reference molecule from FAK1

Figure 6-22 shows examples of the top retrieved actives from an example reference molecule in the LCK target and is an example of a molecule in which the K-Nearest Neighbours descriptor exhibited different behaviour to the other methods. The top retrieved molecules are shared for Shape-it, CDK Shape Moments, and the covariance descriptor using WKS, $D = 100$. Actives 216, 235, and 19 are the top two retrieved molecules for these three methods. The third molecule differs in each and while the common volume can be viewed with the Shape-it method, it is not necessarily clear how the distribution of inter-atomic distances and the covariance of the filter functions are similar to the reference molecule from visual inspection. The differences in the 3D structures of Actives 216, 235, and 19 are closely related as they only cause small perturbations in the shape definition for Shape-it, which is defined by volume, and CDK Shape moments, which is defined by inter-atomic distance. The shape definition of the WKS $D = 100$ covariance descriptor, which relates to the filters mapped over the surfaces, has also preserved these features. Conversely, the top ranked actives from the K-Nearest Neighbour descriptor are starkly different. In this respect it can be hypothesised that the K-Nearest Neighbours histogram acts in a similar way to a dictionary fingerprint where similarity is governed by the presence of local geometry features and results in a high scaffold diversity.



Example LCK reference molecule Active id: 296

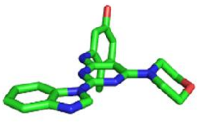
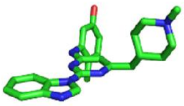
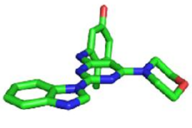
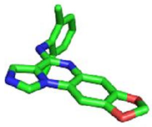
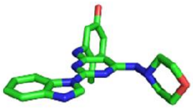
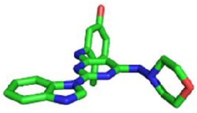
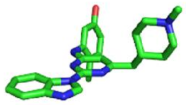
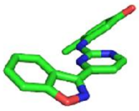
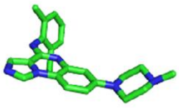
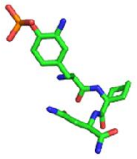
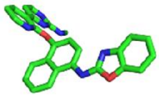
	Shape-it	CDK Shape Moments	WKS 100	KNN 3
1	 Actives id:216	 Actives id: 235	 Actives id: 216	 Active id: 132
2	 Actives id: 19	 Actives id: 19	 Actives id: 235	 Actives id: 452
3	 Actives id: 278	 Actives id: 74	 Actives id: 421	 Actives id: 670

Figure 6-22. Comparison of top retrieved actives for 4 different methods using an example reference molecule from LCK

Overall, the illustration of the top retrieved actives shows how the descriptors determine shape. Although these two reference molecules are not sufficient to draw concrete conclusions and more research is required, it appears that the Shape-it methodology and the covariance descriptor emphasise similar shape features. In the case of Shape-it, this is the volume occupied by the atoms, whereas the covariance descriptor may be emphasising similar global properties due to the nature of comparing filter functions that have been mapped over the whole surface. On the other hand, the figures suggest that the CDK Shape moments and K-Nearest Neighbours descriptors can be used as complementary methodologies. In particular, K-Nearest Neighbours descriptors appear best placed to pick out local shape features that the molecules have in common, this is likely to be due to the sparsity of the fingerprint and the diversity of the codebook. Nevertheless, it suggests that a 3D virtual screening workflow could be improved by fusing the complementary methodologies to improve the quality of 3D similarity searches.

6.3.9 Speed comparison

To test the speed of the 3D shape comparisons, the virtual screen of the AMPC target was timed for each of the 20 reference molecules used in Chapter 5. All experiments were carried out on a local desktop computer with a 3.40GHz Intel Core i7 processor (i7-3770), 32 GB of RAM, running Fedora Linux (version 25). The average time per reference molecule on 2964 target molecules is presented in Table 6-1. The slowest method is the WKS, $D = 1000$ covariance matrix followed by Shape-it, and the WKS, $D = 100$ covaraince matrix. Shape-it requires an alignment step that slows the comparison significantly, whereas the covariance descriptors are slowed by the number of necessary comparisons. For example, $D = 100$ and $D = 1000$ will require $10{,}000$ and $1{,}000{,}000$ comparisons respectively. The Bag of Features spectral geometry descriptors are significantly faster than the other methods. This is because the Bag of Features descriptors do not require an alignment step and 700 length Bag of Features descriptors are significantly smaller than the covariance matrix descriptors.

Table 6-1. Average time in seconds of a screen of a single reference compound on the AMPC target.

Shape-it	WKS, D=100	WKS, D=1000	HQ	KNN, K=3
15.439	0.390	26.738	0.005	0.005

6.4 Discussion

The results suggest that the best performance for the Bag of Features descriptors were for a codebook of size 700 using either hard vector quantisation encodings or K-nearest neighbour encodings with K=3 and K=10. Figure 6-10shows that while there is enough variation in the results to make it hard to determine which the best performing Bag of Features descriptors, it is clear that the hard quantised and k-nearest neighbours descriptors outperform the soft allocation encoding. This is likely to be due to additional noise generated by the soft allocation encoding method.

When considering the virtual screening experiments on the full DUD-E data set, the Bag of Features descriptors performed significantly better than the CDK Shape Moments descriptor on both metrics and gave a comparable performance to Shape-it on average AUC performance. In addition, the time to screen a single reference molecule is significantly faster on average than the alignment-based method. In comparison to the covariance matrix descriptors, the Bag of Features have a similar AUC performance when compared with the WSK, $D = 1000$ covariance matrix method, showing that the same amount of information can be significantly reduced to a low dimensional representation, which in turn significantly speeds up the comparison time.

The Bag of Features descriptors result in a low dimensional representation of the geometric features in the local geometry descriptor with the aid of a codebook. This low dimensional representation is a form of signal compression and therefore, at the heart of the chapter, is the issue of dealing with the necessary information loss while amplifying the optimal features. When a local geometry descriptor is converted into a global descriptor, information loss occurs at two stages: first is a loss of geometric information when encoding each vertex in relation to a word in the codebook, the second is a loss of the spatial relationship between the encoded features when pooling the encoded points over the whole shape.

Figure 6-16 showed that the hard vector quantisation step reduces the descriptor to a sparse representation with few codewords turned on. Nevertheless, the additional noise added by the soft allocation vector quantisation encoding is evidently detrimental to performance (Figure 6-8). Subsequently the results indicate that molecular shape is characterised by few spectral geometry features giving sparse but effective representations in hard vector quantisation and K-nearest neighbours encodings where K is small. Interestingly the noise in the soft allocation vector quantisation descriptor is reasonably evenly distributed over the codebook suggesting that the codewords themselves are relatively close to each other in local geometry vector descriptor space. It was hypothesised that the prevalence of a few features used in the description would have given a good opportunity to use tf-idf normalisation as this would remove the importance of features that were prevalent throughout the data set and enhance less common features. Surprisingly, tf-idf normalisations could only perform as well as the final descriptors (Figure 6-13), which suggests that the relative frequency of the highly prevalent features over the data set is crucial in virtual screening of Bag of Features descriptors for molecular shape.

The second channel of loss is through the loss of spatial relationships of the features, however, this is not a trivial problem to solve for molecular shape as molecules have no natural orientation and therefore lack a canonical frame of reference. The alternative option is to use spatially sensitive expressions, however, these performed poorly compared to the bag of features histograms with the exception of promising AUC results for AMPC and GCR (Figure 6-11). This could be explained by the distribution of the features over the shape in the two targets being of particular importance in the overall ranking. However, of all the descriptors tested, these performed worst with early retrieval tests of enrichment factor 0.5%. Therefore, they are evidently not sensitive enough to promote actives to the top of the rankings.

The choice of virtual screening metric also had an effect on how the descriptors were evaluated. It appeared that occasionally the AUC results contradicted those of early enrichment, see Figure

6-8, for example. Therefore, while some descriptors performed well on ranking the entire test set, such as soft allocation vector quantisation encoding in AMPC (Figure 6-8), they performed poorly at early retrieval. Using the BEDROC statistic as a balance between these two options was helpful and was observed to equalise the performance in a way that normalised total ranking against early retrieval.

A final important property of the Bag of Features descriptors with respect to their comparison with covariance descriptors is their size. As the histogram is the size of the codebook, c, whereas the covariance descriptors is the square of the dimension of the local descriptor, D^2, it is always the case that $c \ll D^2$. Therefore, the storage and retrieval of the histogram descriptors is more efficient as is the comparison time. These factors are increasingly important when carrying out virtual screening over large data sets.

6.5 Conclusions

Overall, the Bag of Features descriptors have been parameterised for the purposes of shape-based virtual screening of molecules. The optimal parameters are a codebook of approximately 700 words using a hard vector quantisation encoding or a K-nearest neighbours encoding with $k = 3$. In a comparison against the baseline descriptors of Shape-it and UFSR as well as the covariance descriptors from Chapter 6, they performed better on average with respect to early retrieval and comparatively to WKS, $D = 1000$ with respect to AUC. Furthermore, when the efficiency of comparison time is taken into account, they represent promising candidates for 3D shape descriptors for large databases of molecules.

Future work can use machine learning to learn the codebooks in a supervised manner that would make the descriptors invariant to conformation deformation. Additionally, the descriptors could be weighted using information on the known actives in a given target using a supervised Bayesian method in order to derive target-specific descriptors.

7 Conformation and spectral geometry

7.1 Introduction

The handling of conformation is thought to be the main reason why 3D virtual screening and molecular docking methods perform below expectations (Maggiora et al., 2014). Therefore, a descriptor that captures sufficient 3D conformational variation of a single molecule in a concise way to include the bioactive conformation would appear to be the ideal 3D molecular shape descriptor. On the other hand, conformational variation can be drastic enough to require two very different conformations of the same molecule as different shapes. Therefore, there is an inherent conflict in the definition of 3D molecular shape between the notion that the shape captures all possible bioactive conformations of a single molecule, or conversely, that each conformation represents a different shape. Intuitively, it is probably more desirable to regard large conformational deformations as different shapes and small perturbations of a conformation as the same shape. In practice, 3D virtual screening is carried out with a sample of conformations to represent the conformational space of the molecule. The fundamental problem is a lack of a framework to provide a vocabulary for 3D molecular shape. However, spectral geometry provides an opportunity to analyse these matters systematically.

The fundamental geometric framework for dealing with the conflict of shape definition for molecular shape is not fully understood. Liu et al. (2011) developed flexible shape descriptors for protein shape and argued that deformable shape descriptors based on intrinsic geometry, such as the spectrum of the Laplace-Beltrami operator, are not suitable for 3D shape descriptors of chemical compounds because conformational deformation is not isometric. The authors subsequently proposed using alternative metrics such as the inner distance. A further approach to providing a deformable shape descriptor for 3D molecular shape of proteins uses diffusion distances (Axenopoulos, Rafailidis, Papadopoulos, Houstis, & Daras, 2016). This method enumerates all the diffusion distances between vertices and uses a singular valued decomposition of the diffusion distance matrix to produce a compact global shape descriptor. This is combined with local descriptors computed at selected points as histograms of geometric information in the local area to return a hybrid descriptor for similarity searching. This method was intended to be applied to proteins but can also be used on small molecules. These methods are very recent approaches that address the issue of capturing conformational variation into a single descriptor, however, it is important to note that large macromolecular structures pose different challenges compared to small molecules with respect to the flexibility of the shape. For example, proteins are prone to articulated movements, which are not relevant for the majority of small molecules that are more likely to be subject to smaller range torsion and rotation deformations. Subsequently, the surface of a protein undergoing an articulated deformation would have a greater stretch over the surface at those articulation points compared to small molecules whose surfaces would be subject to less extreme flexibility.

The spectral geometry approach described in Chapter 4, takes the spectrum of the Laplace-Beltrami operator, which is invariant to isometric transformations, and applies signal processing filters to vary the extent to which local and global geometric features are captured. The result is a local geometry descriptor that is near-isometric, rather than purely isometric, and allows for minor perturbations in conformation. This chapter uses the framework of spectral geometry to determine how conformational variation affects the spectral geometry descriptors and analyses the extent to which conformational variation is isometric. First, the individual spectra are investigated without any filters applied to characterise the isometric variation between conformers of the same molecule. Then filtered descriptors are compared using heat plots to determine how spectral geometry descriptors vary with respect to conformational deformation. Finally, a functional correspondence analysis is carried out in order to identify the causes of non-isometric deformation in different conformers.

7.2 Methods

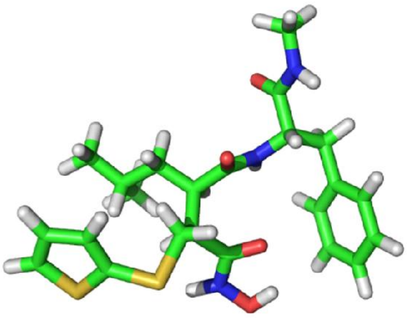
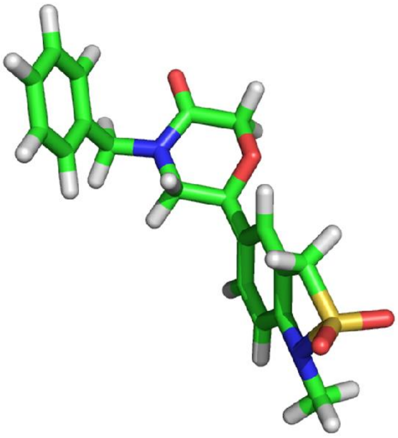
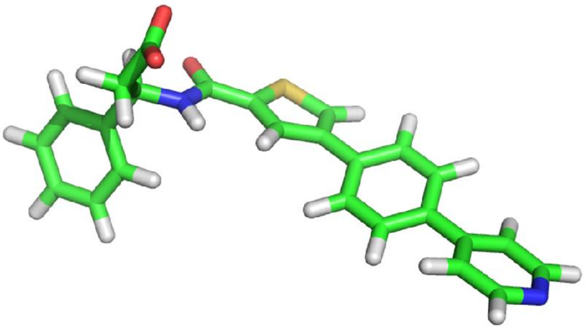
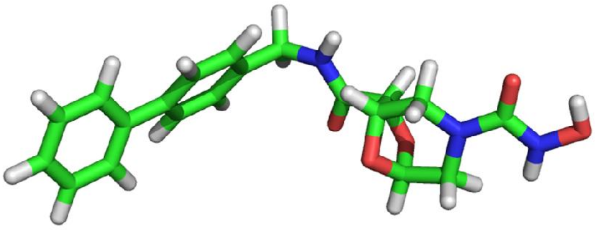
7.2.1 Data set

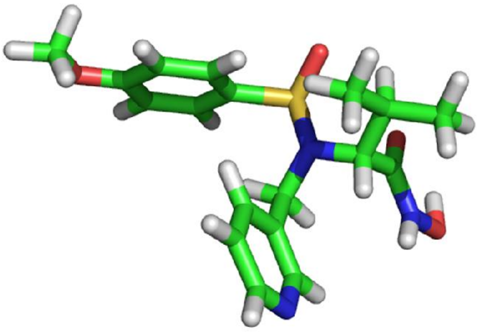
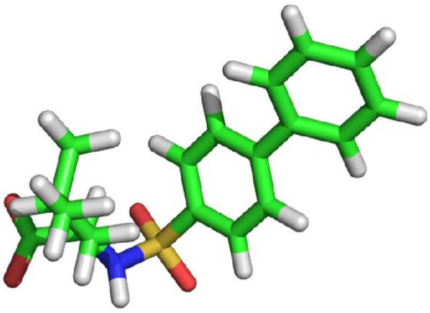
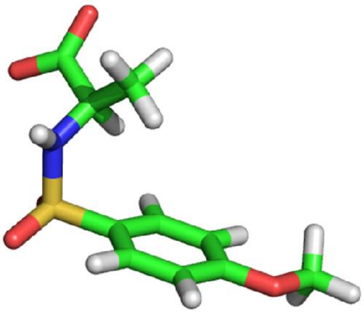
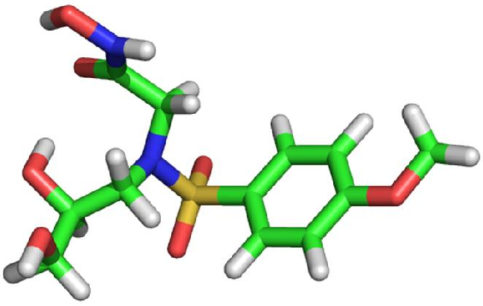
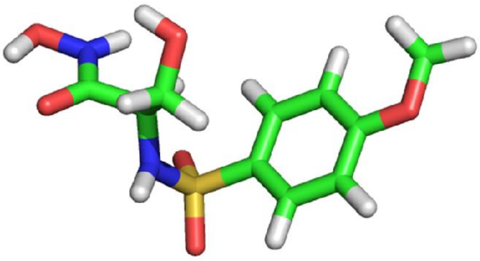
The data set for the conformational analysis was taken from the AstraZeneca molecular overlays for pharmacophore validation (Giangreco et al., 2013). This data set was introduced in Chapter 3 and comprises of 121 overlays of high-quality crystallographic structures publicly available to download from the Cambridge Crystallographic Data Centre. Table 7-1 shows the three targets selected for the analysis. The first twenty low energy conformations were generated for each ligand in each target using the OpenEye OMEGA software (Hawkins & Nicholls, 2012; Hawkins, Skillman, Warren, Ellingson, & Stahl, 2010). In the cases where OMEGA found fewer than twenty optimal low energy conformations then all the conformations were used in the analysis. Consequently, the total number of conformations generated for a target is not twenty times the number of molecules.

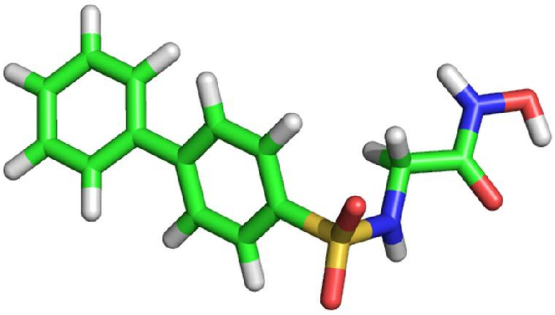
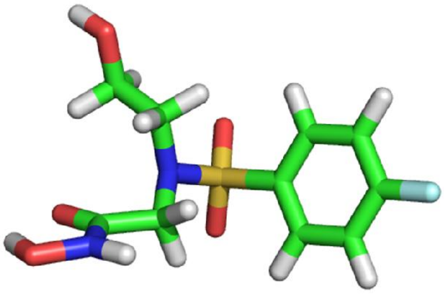
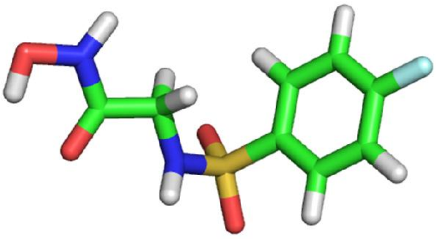
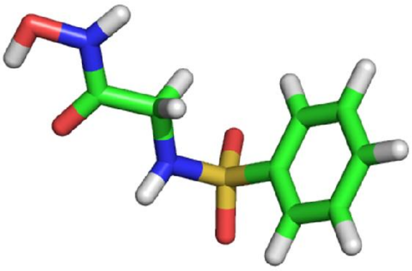
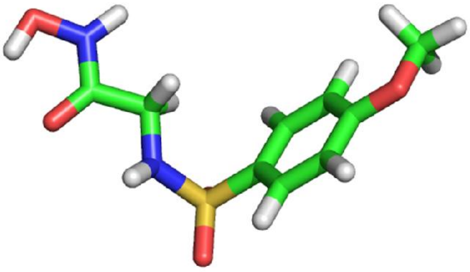
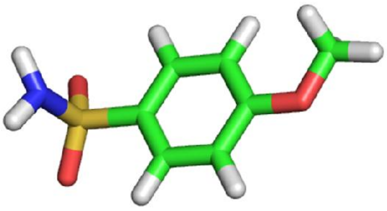
Table 7-1. The number of ligands and conformers per target used for conformation analysis.

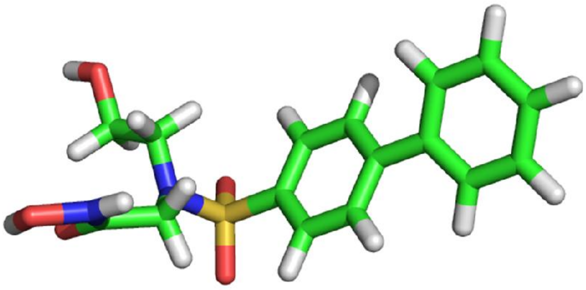
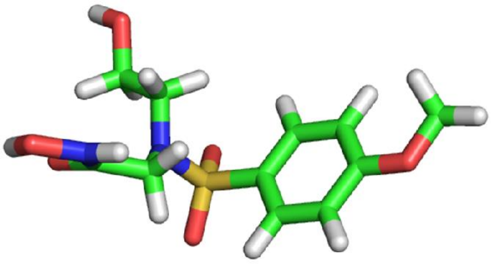
Target	Number of ligands	Number of conformers
P39900	17	324
P04058	8	81
P61823	7	134

Table 7-2. The first conformation of the 17 ligands used for the analysis in the P39900 target along with the number of conformers generated for that ligand.

		Ligand name	Number of conformations
1		1jk3_lig_BAT	20
2		1utt_lig_CP8	20
3		1utz_lig_PF3	20
4		2hu6_lig_37A	20

5		2w0d_lig_CGS	20
6		3ehx_lig_BDL	20
7		3ehy_lig_TBL	20
8		3f15_lig_HS1	20
9		3f16_lig_HS3	20

10		3f17_lig_HS4	20
11		3f18_lig_HS5	20
12		3f19_lig_HS6	20
13		3f1a_lig_HS7	20
14		3lk8_lig_Z79	20
15		3lka_lig_M4S	4

16		3n2v_lig_JT5	20
17		3nx7_lig_NHK	20

7.2.2 Isometry and PCA plots

The unfiltered spectra of the conformations of a set of molecules were investigated visually by plotting the spectra and by an additional PCA plot. The PCA plots show the first two components of a PCA decomposition of the eigenvalues of each conformation. The purpose of these plots was to investigate whether the conformations of particular molecules grouped together. The line plots give a notion of where the spectra lie with respect to one another and the PCA plots give a notion of how similar the spectra are with respect to one another. For the conformers of a single molecule to be considered the same shape in spectral geometry then they need to have similar isometries. In this respect, similar spectra would be plotted near each other in the line plots and the points would be clustered together on a PCA plot. On the other hand, for the molecules to be considered different shapes then the clusters of the conformers need to be separable.

7.2.3 Heat map testing

Similarity heat maps were used to test how the spectral geometry descriptors vary with respect to conformation deformations. The ideal heat map would have a block structure with blocks of similar descriptors for conformations of the same molecule along the diagonal that are separated from blocks of descriptors for conformations of different molecules on the rows. Furthermore, the off-diagonal elements should be similar for different molecules that have conformations with similar shapes.

For each conformation, covariance descriptors were computed as described in Chapter 5 and were compared using Bray-Curtis similarity, which is 1 – Bray Curtis distance. The heat maps were plotted using a Python plotting library.

7.2.4 Specificity and sensitivity testing

While the heat maps give an intuitive sense of the sensitivity and specificity of the descriptors, the framework of the conformation variation provides an opportunity to evaluate the sensitivity and specificity quantitatively. Chapter 2, defined the sensitivity of a classifier as the proportion of correctly classified examples of the object of interest, or the true positive rate (TPR), and the specificity of a classifier as the proportion of correctly classified examples different to the object of interest, or the true negative rate (TNR).

For the purpose of this chapter a descriptor of a particular conformation is classified as being the same as a reference conformation if the Bray-Curtis similarity is greater than a threshold. Therefore, the sensitivity of the descriptor is the proportion of correctly identified conformations of the same molecule and the specificity of the descriptor is the proportion of correctly identified conformations of a different molecule.

7.3 Results

7.3.1 Conformation variation and the eigenvalues of the Laplace-Beltrami operator

To establish the isometric variation in conformers, twenty low energy conformers for the sixth, seventh, and twelfth molecules in Table 7-2 were generated and their spectra computed. Figure 7-1 (a) shows the first 100 eigenvalues for the 20 conformations of each molecule with a different colour used for the conformers of each molecule. Interestingly, the spectra are clustered into three groups corresponding to the three different molecules, with the spectra of two molecules being closer than the third. Figure 7-1 (b) demonstrates this clustering further using a plot of the first two PCA components of the spectrum. The PCA clusters of the three molecules are distinct and separable suggesting that the spectra for each molecule are more similar to each other than they are to other molecules. Additionally, the two clusters that are closer together are those that have similar shapes when visually inspected. This supports the claim that isometry can be used to describe the shape of molecules and that their conformations have similar shapes with respect to spectral geometry.

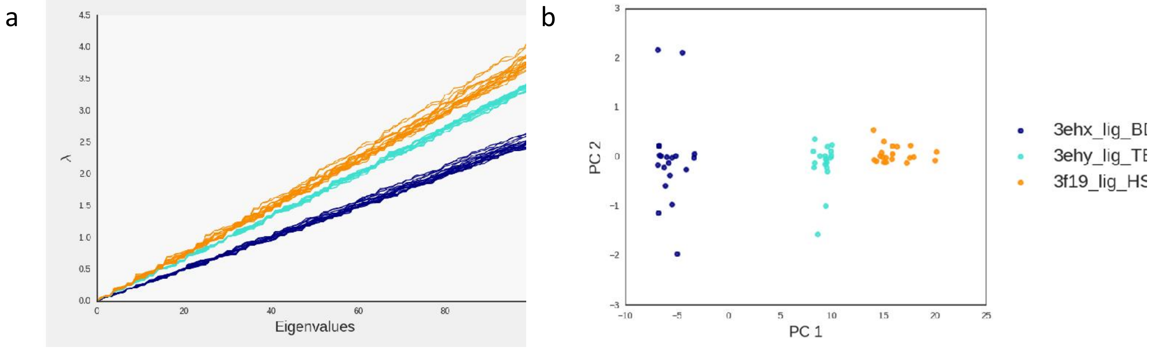
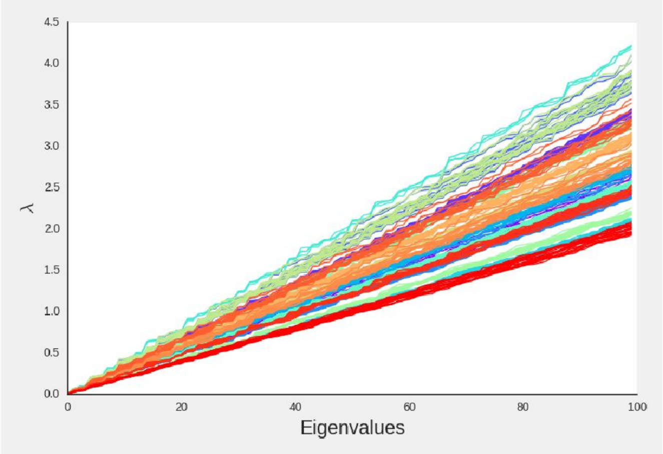


Figure 7-1 (a) plot of the spectra for 20 conformations of three ligands from P39900 and b) PCA plot of the spectra for three ligands in P39900.

When extended to look at all of the conformations generated for the P39900 target it is clear that not all conformations are separable to the same extent. Figure 7-2 shows the spectra for all 324 conformations of the 17 ligands in the P39900 target (Table 7-2) along with the PCA plot of the first two principal components. Figure 7-2 (a) shows the spread of the spectra for the conformations of the molecules. In general, the conformations appear to be grouped together according to the molecule, however, there is a large amount of overlap. Figure 7-2 (b) shows the corresponding PCA plot. Of particular interest is that the conformations of a molecule are vertically stacked meaning that the first principal component can identify the molecule. On the other hand, they are spread vertically by the second principal component, which suggests that the second component captures conformational variation within a molecule. Additionally, some molecules, such as 3f19_lig_HS6 and 3f1a_lig_HS7 appear to have almost identical spectra for the different conformations suggesting they have very similar shape and conformation variation. A visual inspection of the two molecules in Table 7-2 confirms that the two molecules have a very similar 3D shape.

a)



b)

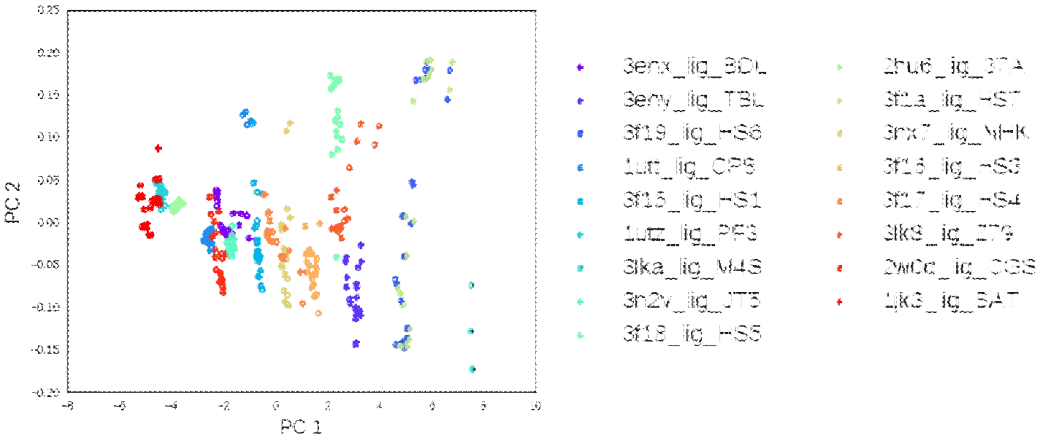


Figure 7-2. (a) plot of the spectra of all conformations for all ligands from P39900 and (b) PCA plot of the spectra for all ligands in P39900.

Once the conformations for a particular target have been generated, the next question is to find out whether active molecules at a particular target have similar spectra when compared to the conformations of molecules that are active in another target. Figure 7-3 repeats the previous plots for three targets in the pharmacophore validation data set. The spectra from P39900 dominate Figure 7-3 (a) showing that there are more conformers in that target but additionally that the spectra span most of the shape region in the diagram. A number of conformers from P04058 are at the top end of the range but others are at the bottom, on the other hand, the conformers from P61823 are hidden below the conformers from P39900, in the middle of the range. These patterns are better observed in the PCA plot of Figure 7-3 (b) whereby the conformations of P39900 are scattered across the range of the plot with the conformations from P04058 being separated either side and P61823 concentrated in the middle, suggesting that the spectra for active molecules from different targets are not separable in spectrum space. Additionally, it is interesting to notice that the structure of the PCA plot from Figure 7-2 is preserved with the conformations being stacked upon each other in the first component and spread in the second component.

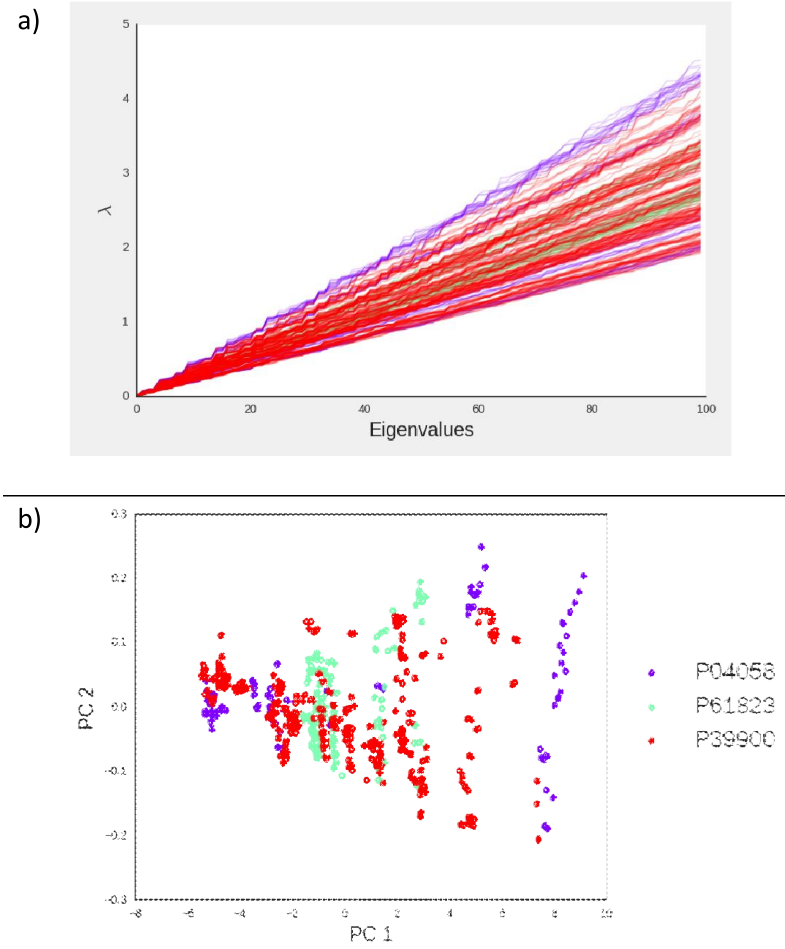
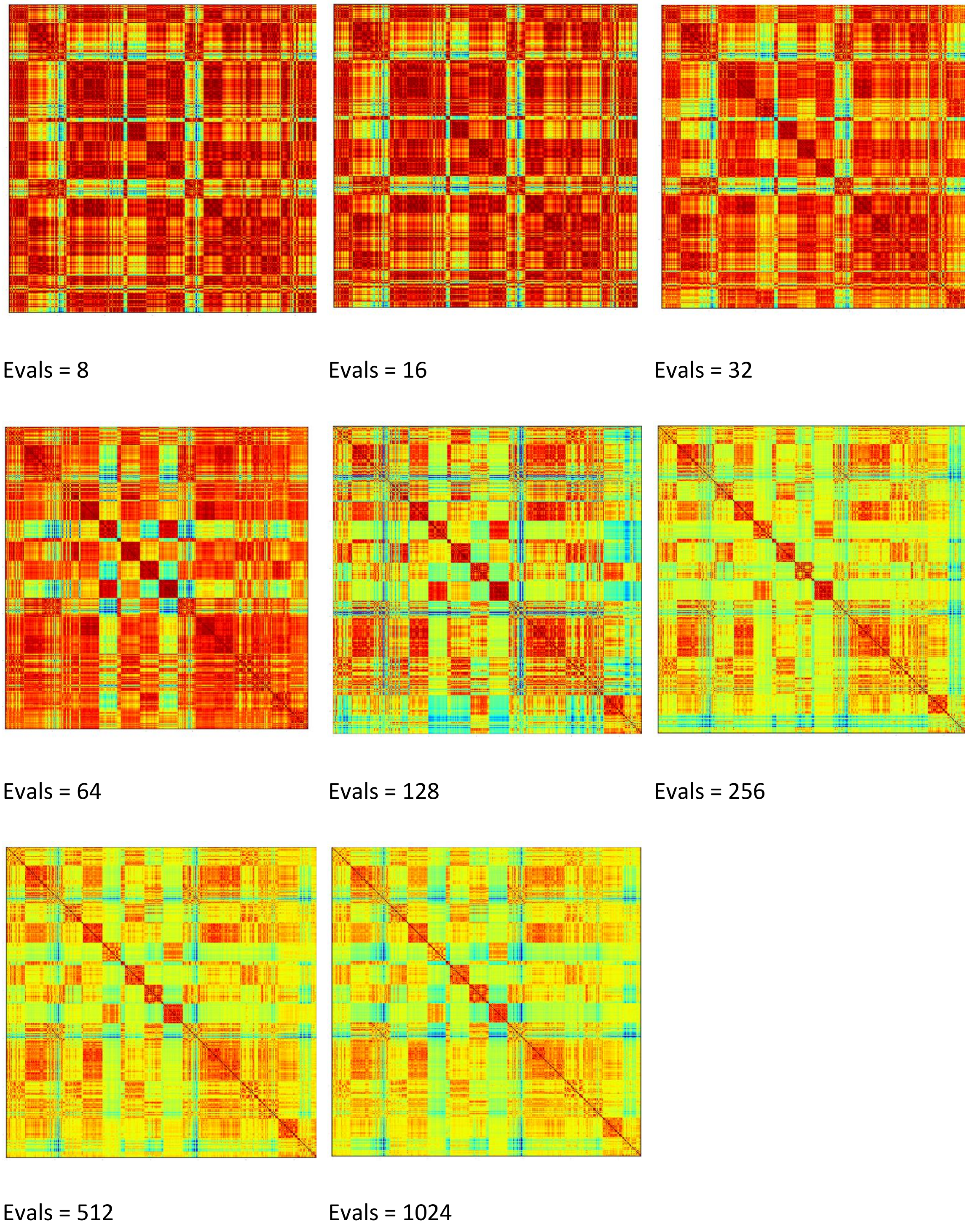


Figure 7-3. (a) plot of the spectra of all conformations for all ligands from P04058, P61823, and P39900 and (b) PCA plot of the spectra for all ligands in P04058, P61823, and P39900.

The line diagrams and the PCA plots of the eigenvalues show that conformation deformation is not isometric, which confirms the observation of Yu-Shen Liu et al. (2011). However, the PCA plots indicate that a large amount of information about conformational variation is encoded in the spectrum of the Laplace-Beltrami operator. This is demonstrated by the stacking of the conformations of the same ligand in the first principal component, which suggests there is intrinsic geometric information that is specific to all conformations of a single molecule and that spectral geometry methods are appropriate for small molecules.

7.3.2 Conformational variation and global descriptors

The relationship between conformational variation and global shape descriptors was investigated using heat maps. Whereas the previous section (7.3.1) looked at the eigenvalues of the Laplace-Beltrami operator for each conformation, this section investigates the similarity of the global descriptors based on those eigenvalues using heat maps (Figure 7-4). The heat map is a 324×324 matrix that shows the similarity values for the pairwise Bray-Curtis similarity values of the covariance descriptors for all 324 conformations of the 17 ligands in P39900 using the Wave Kernel Signature (WKS). In Figure 7-4, each heat map is plotted for a particular number of evaluations which increases from 8 to 1024. The immediate observation is that there is an increasing dissimilarity with the number of evaluations as the heat maps become cooler (move from red to green), on the whole. The heat maps for evaluations 8 through to 64 are largely warmer colours indicating that all the conformations have similar descriptors, however, there is an increasing block pattern emerging on the diagonal. The block diagonal pattern is most prominent for evaluations 128 and 256. Nevertheless, there are a couple of molecules in these plots that have weak inter-conformation similarity, which suggests that there is a large shape variation in the conformations for these molecules. Finally, the most diverse plots (greatest variation in colour) are for evaluations 512 and 1024.



Evals = 8

Evals = 16

Evals = 32

Evals = 64

Evals = 128

Evals = 256

Evals = 512

Evals = 1024

Figure 7-4. Heat maps of the covariance descriptors for the WKS for all conformations in the P39900 target with their corresponding number of evaluations.

Overall, the heat maps in Figure 7-4 indicate that there is high sensitivity for lower evaluations of the WKS, shown by the high similarity values between conformations of the same ligand, and high specificity for higher values of the WKS, indicated by the low similarity values across the ligands.

Interestingly some molecules have conformations that produce drastically different descriptors such that their heat maps demonstrate very little inter-conformer spectral shape similarity. This demonstrates that conformation variation in some molecules corresponds to a diverse number of shapes with respect to their intrinsic geometry. Additionally, the clusters of the full set show that some clusters of molecules are separable whereas others are not (Figure 7-2). This gives an interesting conclusion that spectral geometry provides a vocabulary for talking precisely about 3D flexible shape. However, it also presents a dilemma: on the one hand, shape can be defined purely in terms of isometry, thus giving a fundamental way of describing shapes through the metrics and the distortion. On the other hand, the shape space of a molecule may be characterised by a small subset of fundamentally distinct shapes with respect to their intrinsic geometry.

7.3.3 Sensitivity and specificity of conformation variation for the Wave Kernel Signature

The descriptors for different conformations of the ligands were used to quantitatively test their sensitivity and the specificity in the context of classifying descriptors as being part of the same molecule. In this experiment, the descriptors were computed and two descriptors were predicted to be the same conformation if the Bray-Curtis similarity was greater than a threshold. Table 7-3 shows the sensitivity, True Positive Rate, and specificity, True Negative Rate, for the conformations of all ligands in the P39900 target using a threshold of 0.8. These values are calculated from a classification experiment using each conformation as a reference molecule and the actives are defined as the conformations of the same molecule with the decoys being conformations of a different molecule. The average sensitivity and specificity values are reported for each ligand. A typical ligand is 3f15_lig_HS1 which has a sensitivity of 1 and a specificity of 0.42 at the lowest number of evaluations, evals = 8, and at the other end of the table, evals = 1014, has a sensitivity of 0.55 and a specificity of 0.99. However, some molecules exhibit unusual behaviour. For example, 2w0d_lig_CGS has a sensitivity of 0.55 at 8 evaluations which improves to 0.85 for 256 evaluations and eventually returns to a level below the initial value at 0.35 for 1024 evaluations, which suggests that there is too much shape variation in the conformations and the descriptor struggles to identify true positives at high numbers of evaluations.

These patterns are further illustrated in Figure 7-5, which shows the average sensitivity and specificity values for increasing evaluations at three different thresholds: 0.75, 0.8, 0.85. The curves show that on average the WKS becomes less sensitive (the TPR decreases) with increasing evaluations whereas its specificity (TNR) increases with increasing evaluations. Additionally, sensitivity and specificity behaviour is closely related to the choice of the threshold value. For a threshold value of 0.8, the True Positive Rate and the True Negative Rate cross at around 100 evaluations, whereas for a threshold value of 0.85, the True Positive Rate and the True Negative Rate cross at around 50 evaluations. This suggests that at the higher threshold of 0.85 the specificity of the WKS is quick to identify all True Negatives meaning the True Negative Rate quickly dominates. Conversely, at a threshold of 0.75 the curves converge at 1024 evaluations. This suggests that at this lower threshold the True Negative Rate is less dominant as the WKS is less specific, which, in turn, means that fewer True Negatives are immediately identified with the trade-off being improved sensitivity. Interestingly, the sensitivity-specificity behaviour is captured in this small range of thresholds.

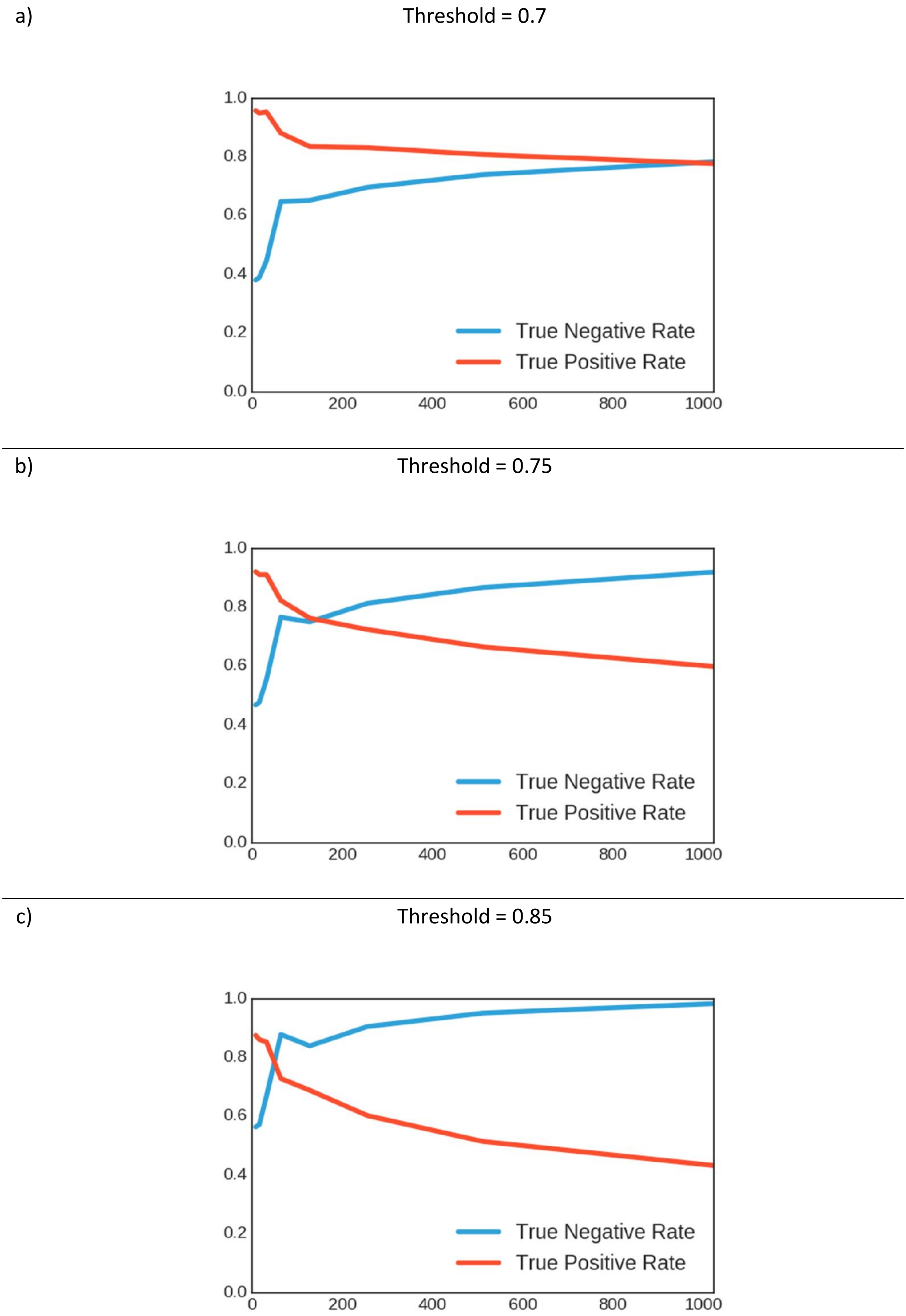


Figure 7-5. The average sensitivity and specificity plots for the conformation classification experiment using three different thresholds, (a) threshold = 0.7, (b) threshold = 0.75, and (c) threshold = 0.85.

7.4 Discussion

The results presented in this chapter suggest that there is a relationship between the shape variation of the different conformers of a molecule and spectral geometry descriptors. In particular, there is sufficient information in the first principal component following a PCA of the eigenvalues of the spectrum to differentiate the conformers of different molecules (Figure 7-2). This demonstrates that there is information captured by the spectrum that is closely linked to the conformational variation.

The heat maps demonstrate an intrinsic conflict in the definition of 3D molecular shape for the purposes of representation and similarity searching. In particular, some molecules have substantially different conformers that could not be captured by a global descriptor (Figure 7-4). Nevertheless, the conflict between capturing all conformations in a single global descriptor and treating distinctly different conformers as different shapes can be captured in a sensitivity specificity trade-off (Section 7.3.3). This framework can subsequently be exploited to evaluate conformational diversity and to train descriptors that are invariant to conformation specific deformation.

The concept of shape in drug discovery is rooted in the notion of a 3D rigid configuration in Euclidean space. For example, one important task in drug discovery is to find common shapes between molecules to identify bioactive conformations. The framework presented by spectral geometry necessarily abstracts the Euclidean model to allow for a notion of flexibility. Consequently, there is no method to identify a common conformation between two molecules given a single spectral geometry descriptor for each. On the other hand, a high similarity between two descriptors suggests that a there is a high probability that one such shape exists, which cannot necessarily be said for rigid body methods. Further work is needed to investigate precisely how the conformational flexibility of these descriptors is handled. Rather than use these global descriptors directly to identify flexible common shapes between two molecules, a new frame work is required based on partial matching. Partial matching of local spectral geometry descriptors is computed using the functional correspondence methodology (E. Rodolà et al., 2016). This method identifies the common patches of the local geometry descriptors that map to each other. These can be visualised to identify the areas on the surface that the molecules have in common, which by extension identify the putative bioactive conformations of two molecules.

7.5 Conclusions

In conclusion, this chapter has investigated the applicability of spectral geometry to small molecules with respect to how the descriptors preserve geometric information under conformational deformation. In particular, the PCA plots suggest that a large amount of intrinsic geometric information encoded in the spectrum of the Laplace-Beltrami operator is preserved for different conformations of the same molecule. Furthermore, the heat maps demonstrate that the signal processing filters can be used to create near-isometric invariant descriptors whose parameters can be adjusted to alter how sensitive and specific the descriptors are to conformation deformation. Finally, the sensitivity and specificity plots quantify the effect of changing the local geometry descriptor parameters on conformation invariance.

Ultimately the chapter is not able to resolve the dilemma posed in the introduction but instead lays the ground work for future analysis on shape and conformation. For example, the sensitivity and specificity data can be used to learn a weighting scheme on the global geometry descriptors can be trained to optimise the sensitivity / specificity trade-off. The, result would be a pseudo-isometric descriptor that is invariant to conformation deformation specifically, rather than purely isometric deformation. Finally, while not explored here, the recent work on functional correspondence reported in the literature (Ovsjanikov et al., 2013) could be used to further quantify the behaviour of the metric of the shape manifold to identify the points the molecular surface that deform the most under conformational variation. A thorough investigation would then demonstrate and isolate chemically meaningful areas of distortion that a pure shape method would not be able to interpret.

Table 7-3. Sensitivity (TPR) and specificity (TNR) for covariance descriptors of the WKS of different numbers of evaluations.

		evals=8		evals=16		evals=32		evals=64		evals=128		evals=256		evals=512		evals=1024	
Ligand	Number of conformations	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR
3ehx_lig_BDL	20	0.45	0.71	0.47	0.70	0.58	0.67	0.84	0.36	0.81	0.39	0.88	0.30	0.95	0.23	0.99	0.19
3ehy_lig_TBL	20	0.63	0.97	0.63	0.96	0.68	0.95	0.82	0.87	0.72	0.82	0.81	0.83	0.90	0.58	0.96	0.39
3f19_lig_HS6	20	0.78	0.63	0.78	0.62	0.81	0.57	0.90	0.39	0.83	0.36	0.89	0.34	0.93	0.24	0.97	0.18
1utt_lig_CP8	20	0.47	0.87	0.48	0.86	0.60	0.68	0.86	0.59	0.86	0.55	0.93	0.50	0.96	0.42	0.99	0.32
3f15_lig_HS1	20	0.42	1.00	0.43	1.00	0.53	1.00	0.84	1.00	0.72	0.98	0.85	0.83	0.93	0.74	0.97	0.55
1utz_lig_PF3	20	0.57	0.71	0.58	0.68	0.89	0.83	0.94	0.81	0.94	0.91	0.97	0.57	1.00	0.35	1.00	0.21
3lka_lig_M4S	4	0.86	1.00	0.85	1.00	0.85	1.00	0.94	1.00	0.81	1.00	0.88	1.00	0.96	1.00	0.99	0.88
3n2v_lig_JT5	20	0.61	0.98	0.65	0.97	0.74	1.00	0.89	0.98	0.85	0.93	0.93	0.74	0.97	0.57	1.00	0.47
3f18_lig_HS5	20	0.44	0.99	0.45	0.99	0.61	1.00	0.95	0.95	0.91	0.77	0.94	0.51	0.97	0.49	0.99	0.43
2hu6_lig_37A	20	0.47	0.92	0.48	0.88	0.66	0.95	0.94	0.97	0.93	1.00	0.97	0.93	1.00	0.72	1.00	0.62
3f1a_lig_HS7	20	0.78	0.66	0.78	0.64	0.80	0.59	0.90	0.40	0.84	0.35	0.90	0.32	0.94	0.27	0.97	0.20
3nx7_lig_NHK	20	0.41	1.00	0.42	1.00	0.53	1.00	0.80	0.97	0.70	0.83	0.83	0.66	0.91	0.53	0.96	0.44
3f16_lig_HS3	20	0.54	0.98	0.55	0.97	0.60	0.98	0.78	0.90	0.70	0.83	0.80	0.78	0.90	0.48	0.97	0.25
3f17_lig_HS4	20	0.46	0.76	0.47	0.72	0.56	0.70	0.84	0.40	0.79	0.50	0.88	0.41	0.95	0.32	0.99	0.23
3lk8_lig_Z79	20	0.62	0.92	0.63	0.88	0.65	0.84	0.83	0.52	0.77	0.47	0.89	0.41	0.97	0.32	0.99	0.26
2w0d_lig_CGS	20	0.51	0.55	0.53	0.52	0.63	0.48	0.91	0.42	0.94	0.82	0.98	0.67	1.00	0.51	1.00	0.35
1jk3_lig_BAT	20	0.57	0.78	0.56	0.74	0.80	0.64	0.98	0.36	0.99	0.22	1.00	0.17	1.00	0.16	1.00	0.12

8 Conclusions and future work

The main body of this thesis has addressed the application of the concepts of spectral geometry to develop a new alignment independent descriptor for molecular shape comparison. Spectral geometry provides an exciting opportunity to consider the shape of a molecule as a 2D surface that may have a number of 3D poses rather than a rigid configuration that must be aligned in 3D space. In doing so, the resulting descriptors are alignment-invariant and also invariant to isometric deformation. However, in order to produce final shape descriptors of molecules, a fair amount of parameterisation must be carried out first. This parameterisation and application to virtual screening formed the fundamental contribution of this work, which can be generalised as being composed of two steps: generation of the local geometry descriptor, and parameterisation of the global geometry descriptor for virtual screening. When tested against two benchmark shape comparison methods, the spectral geometry descriptors performed better than CDK Shape Moments, a benchmark alignment-invariant shape descriptor, and had comparable AUC values to Shape-it, a benchmark alignment-dependent method.

This work also investigated the application of 3D shape descriptors to fragment-based drug development. Chapter 3 provided the first experimental results of the thesis by applying an empirical method for finding 3D bioisosteric pairs to crystallographic data in order to derive a 3D bioisosteric test set. However, this approach faces a fundamental issue about the nature of bioisosterism for fragments. While a promising approach to aid *in silico* drug discovery, there is an issue with the generalisability of the bioisosteric fragment pairs. Very few bioisosteric fragments were found to be generalizable between targets (Figure 3-6). Section 3.3.2 introduced the notion of bioisosteric groups and the notion of transitivity to bioisosteric pairs. However, it was also found that bioisosteric groups were not generalizable between targets. Ultimately it was thought that there was a fundamental issue with the definition of bioisosterism for fragments. This is due to the nature of how common these fragment pairs appeared. Only few generalizable pairs were found, whereas less general pairs may have been endowed with target-specific information that rendered them unsuitable to general applications.

Chapter 4 introduced the concept of spectral geometry for deriving an alignment invariant descriptor of 3D molecular shape. This represents a novel way of describing shape by changing the concept of shape from a fixed rigid conformation to a 2D surface that may have a number of poses. The method is based on obtaining the spectrum of the Laplace-Beltrami operator over the surface and then applying signal processing filters to the spectrum in order to emphasise the desired geometric properties of the spectrum. In particular, two functional forms were investigated: the Heat Kernel Signature (HKS) and the Wave Kernel Signature (WKS). It was found that the two descriptors could be distinguished with respect to the sensitivity and specificity of the local geometry descriptors. However, in the absence of an appropriate data set, these properties could only be investigated visually (Figure 4-22). This visual inspection suggested that the local geometry descriptors of the WKS were more specific.

Chapter 4 also introduced two frameworks for interpreting the local geometry descriptors: a row-wise interpretation that considers each row as a vector that describes the geometry around the corresponding point in the mesh; and a column-wise interpretation that considers each column as a mapping of a particular filter function over the entire surface. The latter interpretation formed the basis of the covariance descriptor that was used to describe the whole shape in Chapter 5 and the former formed the basis of the Bag of Features descriptor that was used to describe the whole shape in Chapter 6.

Chapters 5 developed the covariance descriptor for whole molecule comparisons and used this to form a framework for finding the optimal parameters of the local descriptor for virtual screening. It was found that the WKS was the best performing spectral geometry descriptor using the parameters $D = 100$ and $D = 1000$. When compared to two benchmark 3D shape methods, the spectral geometry descriptors outperformed an implementation of the standard 3D alignment-free shape descriptor, and had comparable results to a Gaussian shape method, demonstrating that the spectral geometry descriptors capture a rich amount of geometric information.

Chapter 6 implemented the Bag of Features framework as an alternative global geometry descriptor. It was found that the best parameters for the Bag of Features descriptor were the hard vector quantisation encoding method and a k-nearest neighbours histogram with $k = 3$ and $k = 5$. The histograms performed best when based on a WKS, $D = 100$. When compared against the benchmark shape comparison methods, the Bag of Features descriptors outperformed the CDK Shape Moments and had a comparable performance to Shape-it with regards to AUC (Figure 6-18). When compared against the covariance matrix method from Chapter 5, the AUC performance was comparable to the WKS, $D = 1000$, and the performance was better on average using BEDROC, $\alpha = 20$. However, when the comparison time was taken in to account, the Bag of Features descriptors were orders of magnitude faster than the covariance matrix methods, showing that the rich geometric information captured in the covariance matrices can be compressed to a 700 dimensional vector without harming performance. In addition, the method was also significantly faster than the alignment-based method Shape-it, suggesting the Bag of Features descriptors are excellent candidates for 3D shape similarity searches of large databases.

Finally, Chapter 7 investigated the relationship between spectral geometry and conformation and found that the eigenvalues of the Laplace-Beltrami operator can be used to preserve geometric information between conformations of small molecules. In particular, Figure 7-1 showed that the first principal component of the eigenvalues captures sufficient information so that the eigenvalues of the same molecule are stacked vertically. Furthermore, when the WKS was used as a filter bank over the spectrum it was demonstrated that the change in parameter changes the sensitivity and specificity of the global geometry descriptor with respect to the conformation (Figure 7-5). This provided a way to describe how 3D shape descriptors manage conformation variation and could also be used as a metric for learning the parameters for a domain specific descriptor.

8.1 Future work

This thesis has introduced the concepts of spectral geometry to 3D molecular shape description and represents an initial implementation of the concepts in that field. Future work can build upon this work in a number of ways. The most natural extension of the work carried out in this thesis would be to improve the spectral geometry descriptors for the purpose of virtual screening. In general the virtual screening workflow presented in Figure 8-1 can have additional domain specific optimisations for computing the local geometry descriptors and the global geometry descriptors.

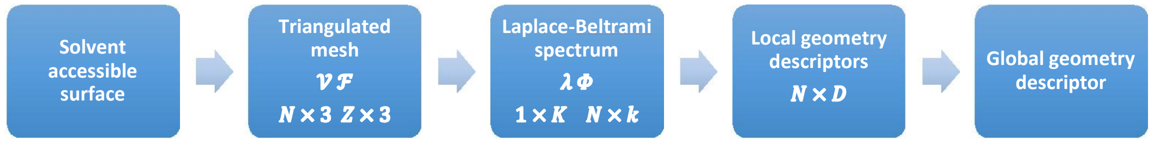

Solvent accessible surface
Triangulated mesh
$\mathcal{V}\ \mathcal{F}$
$N \times 3\ Z \times 3$
Laplace-Beltrami spectrum
$\lambda\ \Phi$
$1 \times K\ \ N \times k$
Local geometry descriptors
$N \times D$
Global geometry descriptor

Figure 8-1. Workflow for developing a global geometry descriptor using spectral geometry.

Domain specific local geometry descriptors have been trained using a convolutional deep neural networks (DNN) to learn the filter banks that best capture non-isometric deformations of shape. Spectral geometry objects cannot be directly fed into a DNN because the underlying metric is non-Euclidean. However, this has been handled by constructing topological discs over surfaces and sampling these (Boscaini et al., 2015) and by using the spectrum of the Laplace-Beltrami operator directly to modify the filters that pass over the surface (Masci et al., 2015). Alternatively, in the final step of mapping the local geometry descriptors to global geometry descriptors, the global Bag of Features descriptors presented in Chapter 6 can be modified by supervised learning of the optimal histograms using sparse coding (Litman, Bronstein, Bronstein, & Castellani, 2014). This would produce global geometry descriptors that are specific to the shape variations found in 3D molecular shape.

There are a number of other opportunities for future work outside of the scope of the global geometry descriptors for virtual screening that take advantage of the underlying mathematical properties of the local geometry descriptors. Functional correspondence analysis presents an exciting opportunity to analyse the deformation properties of conformational variation. Functional correspondence treats the local descriptors as functions over the surface and compares two shapes as functional maps rather than as point-to-point comparisons (Ovsjanikov et al., 2012). A mapping between the shapes can then be calculated that maps the functions rather than the points. The result is a matrix that transforms the basis of one shape into that of another. The functional correspondence can be interpreted as a measure of how isometric two shapes are: if the mapping is the identity matrix, the two shapes are isometric.

These methods present a number of exciting opportunities for shape-based virtual screening. First, an inspection of the metric distortion over the surface of a conformational ensemble would give new insights into how the solvent accessible surface changes as a result of conformational deformation. The degree of distortion will provide a means of clustering the intrinsic geometry of conformations and a novel quantitative framework for determining when two conformations of a molecule are sufficiently different to be considered different shapes. Second, the functional correspondence can be used to derive the best mapping of vertices between two shapes and a visualisation of the alignment based on intrinsic geometry (Ovsjanikov et al., 2013). Therefore, once the most similar molecules have been determined in a virtual screen, the alignment of the molecules can be recovered to explain why molecules are similar; this information is lost for all current vector-based 3D descriptors making their similar properties opaque to the user. In particular, this gives the opportunity of visualising a *flexible alignment* that is not constrained by the requirements of rigid geometry.

In this context, spectral geometry descriptors present the possibility of sampling conformational space based on shape difference which offers two significant advantages over the typical atom-based RMSD approach. First, a protein “sees” a small molecule as a shape (with associated properties on its surface) so that difference in shape is much more relevant to protein-ligand binding than difference in atom positions. Second, it is likely that conformational space can be represented in a much smaller number of shapes than would be required using atom coordinates, thereby significantly reducing the number of comparisons required in virtual screening. Conformers that are isometric will have identical spectral geometry descriptors. This provides a natural way of encoding elements of conformational space into a single descriptor and of identifying when two different molecules may adopt the same shape. This is a novel way of considering conformational flexibility, however, the relationship between conformational space and isometry is complex. Following work in computer vision, machine learning methods will be developed to learn local geometry descriptors that explore this relationship in the context of drug discovery, considering issues such as: when should a conformational change give rise to a new descriptor; when does a change in the descriptor becomes significant; and what is the minimum number of shapes required to represent the conformational space of molecules.

Thus far, all spectral geometry applications have been shape only. However, molecular recognition is driven by complementarity of both shape and electrostatic properties. Therefore, it is no surprise that the inclusion of chemical properties alongside shape tends to improve the performance of virtual screening (Shave et al., 2015). The purely shape-based descriptor developed thus far could be extended to include the spatial distribution of chemical properties directly within the spectrum. A natural option is to use some notion of potential based on a scalar value that is calculated on the surface. For example, the electrostatic potential can be calculated at each mesh point and included in the finite element method as an extra dimension with the vertices on the mesh represented by (x,y,z,p) values. The mesh then moves over the manifold of the shape and the electrostatic potential space and the Laplace-Beltrami operator becomes defined over the joint manifold of the shape and the property space. The challenge then becomes how to weight the property values with respect to the purely shape based features. Other chemical properties could also be investigated, for example, inclusion of hydrogen bonding information on the surface which, although a cruder representation than electrostatics, may more appropriate for some applications such as when a key interaction is known to be important.

Finally, a combination of the above work can be implemented to return to the idea of fragment-based virtual screening. Partial shape matching methods could be developed to analyse accessible space via the fragments themselves, without the need to enumerate the compounds. The functional correspondence work described above provide an efficient approach to partial matching. For example, the functional correspondence matrix has been used to identify coherent segments between two shapes and within a reference shape. Additionally, sparse coding has been used to learn a permuted correspondence based on information in the local geometry descriptor to identify matching regions between shapes. These methods will be adapted for the partial matching of molecules to enable large combinatorial spaces to be searched very efficiently.

Overall, this thesis has applied spectral geometry to the description of 3D molecular shape. It has demonstrated that the Laplace-Beltrami operator can capture the geometric features of molecular surfaces and that application of a bank of signal processing filters can emphasise the desired geometric features for use as virtual screening descriptors. However, as Section 8.1 demonstrates, there are a large number of applications that can be used to improve 3D virtual screening as well as providing a rigorous mathematical treatment of 3D molecular shape and conformation variation. As such, this work can be viewed as the first step towards a larger body of research that applies spectral geometry to 3D molecular shape representation.

A Formal definition of spectral geometry

In this section a brief overview of the technical framework for spectral geometry is given based on summaries provided by Kovnatsky et al and Litman and Bronstein (Kovnatsky, Raviv, Bronstein, Bronstein, & Kimmel, 2013; Litman & Bronstein, 2014). However, there are further references for the interested reader. An excellent tutorial on the mathematical background is given in (Biasotti, Cerri, Bronstein, & Bronstein, 2015); this tutorial constructs the notion of shapes as metric spaces starting from topological spaces and their transformations, which gives a lovely insight into the underlying concepts. A more abstract implementation is demonstrated in (Zobel, Reininghaus, & Hotz, 2011).

A.1 Shapes as manifolds

In the field of spectral geometry, shape is modelled as a compact two-dimensional manifold, M, possibly with a boundary ∂M. The manifold is equipped with a Riemannian metric, g. The Riemannian metric is a local inner product $g(x, x') = \langle x, x' \rangle_g$ that measures the linear distance between points on a plane, T_x, that is tangent to each point on the manifold $x \in M$.

Let f be a smooth scalar field on M, such that $f{:}\, M \rightarrow \mathbb{R}$ then the gradient of f, denoted as grad f, is defined as the vector field that satisfies the following equation, $f(x + dx) = f(x) + \langle \text{grad}\ f(x), dx \rangle_g$ where dx is an infinitesimally small tangent vector. This is a general formulation of the first derivative when the underlying space is a manifold rather than a Euclidean space.

The metric g is then used to define the Lapace-Beltrami operator, Δ_g as,

$$\int f\, \Delta_g\, h\, da = -\int \langle \text{grad}\ f, \text{grad}\ h \rangle_g\, da\,, \qquad \text{Equation A-1}$$

where f and h are smooth scalar fields $f, h : M \rightarrow \mathbb{R}$ and da illustrates integration with respect to the standard area measured on the manifold. This definition using the integral is known as the Stokes identity. Importantly this equation shows that the Laplace operator is uniquely defined by the metric g, thus demonstrating that the operator is an intrinsic property of the manifold M and is independent of all embedding spaces. Additionally, as it is related to the gradient of a scalar field over the manifold, it describes the spatial variation of the manifold, which would give an intuitive notion of shape. For simplicity, the notion used in the body of the thesis dropped the metric g however, Equation A-1 shows that it is a defining characteristic of the shape as a manifold.

The spectrum of the Laplace-Beltrami operator,

$$\Delta_g \phi = -\lambda \phi, \qquad \text{Equation A-2}$$

is composed of the eigenvalues, λ, and eigenfunctions, ϕ, of the Laplace-Beltrami operator that define the spectral geometry of the shape. That is, it describes a geometry that is intrinsic and invariant to transformations of the shape that do not change the underlying metric g. When the metric g is the Euclidean metric, then Δ_g is the standard Laplace operator. In this case, the shapes are rigid bodies. On the other hand, if the metric is the Riemannian metric for the shape as a 2-manifold, then the spectrum is invariant to non-rigid transformations that include isometric flexibility. The Laplace-Beltrami operator spectrum was first proposed to be used as a shape descriptor by Reuter et al. (2006).

Intuitively, the shape is assumed to be a manifold M that exists outside of any embedding space and in order to view the manifold it must be it must be projected into a viewing space as a pose. Each pose is then one of many potential ways in which the manifold can be represented in the embedding space. Furthermore, each pose is realised as a mapping, called the embedding function, that assigns each point on the manifold to a coordinate in the embedding space in such a way that the metric of the manifold is not distorted. Therefore, if the full space of possible embedding functions is given by $\mathcal{H}$ then a particular pose of a shape is $f(M)$, for $f \in \mathcal{H}$. Now imagine a pose is transformed in some way by the transformation, ω. This may be any transformation, for example, a translation, or a magnification, or even tearing and introducing holes. Then the transformed pose is the composition of the transformation and the pose, $\omega \circ f(M)$. Under the current framework, this transformed pose is regarded as the same shape if the transformed embedding function is a valid embedding function of the manifold. From a practical point of view this is equivalent to saying that a descriptor derived from the observed pose is invariant to a transformation if $\omega \circ f(M) \in \mathcal{H}$. As the space of embedding functions, $\mathcal{H}$, consists of all embedding functions that do not distort the metric then any transformation of the embedding function, ω, that preserves the distances over the surface will also be a valid pose. The set of transformations, Ω, that do not distort the metric are called *isometric* transformations. Naturally these include rigid-transformations but also a class of bending that is isometric. Subsequently any descriptor derived from spectral geometry will be alignment invariant and also invariant to isometric flexibility.

While this may appear to be an overly abstract definition of shape, that of a space whose properties are observed by how embedded points behave under transformations, it has some attractive properties. First is that the invariant properties of the shape can be defined precisely, so that the underlying manifold does not change when we change the embedded points by translation, rotation, or by bending isometrically. Second, is that it engenders a framework for thinking about shape as the behaviour of points projected from a higher dimensional space which, in turn, enables surface properties to be treated as intrinsic geometric properties.

A.2 Local spectral descriptors

The local descriptor assigns a vector of values to each point on the surface of the shape. In this case, it can be thought of as being a vector field over the manifold. The vector of the field at each point describes the local shape around the point. This is most commonly achieved using a functional form that exploits the properties of the Laplace-Beltrami Operator. The choice of functional form varies but the most common are the heat kernel and the wave kernel signatures.

The relationship between heat diffusion and geometry can be observed in the heat equation in Equation A-3. Heat dissipation for a surface with a metric, g, is calculated using a set of partial differential equations,

$$\left(\Delta_g + \frac{\partial}{\partial t}\right) u(x,t) = 0, \qquad \text{Equation A-3}$$

where $u(x,t)$ is the amount of heat at a specific point x and at time t. At the heart of heat equation are differential operators that describe the evolution of the heat values through space, using the Laplace-Beltrami operator Δ_g, and through time.

A solution to Equation A-3 can be found using the heat kernel, $h_t(x,y)$, that describes the amount of heat transferred to point x from a starting distribution placed at point y at time t. Importantly, the heat equation has a spectral decomposition given in Equation A-4,

$$h_t(x,y) = \sum_{k \geq 1} \exp(-\lambda_k t)\phi_k(x)\phi_k(y)$$

Equation A-4

were λ_k and $\phi_k(\cdot)$ are the k$^{\text{th}}$ eigenvalue and eigenfunctions of the Laplace-Beltrami operator respectively. In this form, it can be seen that for a given time point, t, the heat transfer is defined by the spectrum of the Laplace-Beltrami operator. Therefore, the heat diffusion properties of a surface are inherently tied to the isometric geometry of the shape.

A.2.1 Heat Kernel Signature

The heat transfer from a single point x, to itself is called the *autodiffusivity function* (Sharma & Horaud, 2010) and is written as,

$$h_t(x,x) = \sum_{k \geq 1} \exp(-\lambda_k t)\phi_k(x)^2.$$

Equation A-5

Sun et al. (2009) used the information from Equation A-5 to create a descriptor that captured the full information of the metric of the manifold, including Gaussian curvature of the surface and diffusion distance. An expansion of the *autodiffusivity function* is related to Gaussian curvature by Equation A-6 (Litman & Bronstein, 2014),

$$h_t(x,x) = \frac{1}{4\pi t} + \frac{K(x)}{12\pi} + \mathcal{O}(t)$$

Equation A-6

where $K(x)$ is Gaussian curvature and $\mathcal{O}(t)$ is big-O notation that describes the asymptotic behaviour of the expansion.

Two further observations can be made from Equation A-5: first that the heat at point x is an exponential function of the eigenvalue at a point in time. If the heat points are sampled over a range of increasing time points, then this captures some notion of heat decay over time. As heat will diffuse over the surface over time, then the geometric relationship between local points is captured, for small values of t, and further points, for larger values of t. Secondly, Equation A-5 can be interpreted as a functional weighting of the eigenvalues that gives a greater weight to smaller eigenvalues and a smaller weight to higher eigenvalues.

The autodiffusive heat kernel can then be used to construct a local geometry descriptor called the Heat Kernel Signature. The Heat Kernel Signature is computed by associating every point on the surface with a D-dimensional vector of autodiffusive heat kernels for D periods of time,

$$\boldsymbol{p}(x) = \big(h_1(x,x), \dots, h_D(x,x)\big). \qquad \text{Equation A-7}$$

In the discrete case of a triangulated mesh with N vertices, the Heat Kernel Signature descriptor is then a $N \times D$ matrix where each row is a vertex in the mesh and each column is the autodiffusive heat kernel at a given point in time.

A.2.2 Wave Kernel Signature

The Wave Kernel Signature is an alternative formulation of a local geometry descriptor that takes the framework of Quantum Mechanics as inspiration to produce a descriptor that is parameterised by frequency rather than time (Aubry et al., 2011).

The key to the Wave Kernel Descriptor is that it is derived from the Schrodinger equation that describes the evolution of a quantum particle over time where i is the imaginary unit,

$$\frac{\partial \psi}{\partial t} = i\, \Delta\psi(x,t). \qquad \text{Equation A-8}$$

Crucially, the dynamics of this equation are governed by oscillations rather than dissipation. The authors then approximate energy probability distribution, f_E^2 for energy E, to derive the wave function of a particle at a given point at a given time,

$$\psi_E(x,t) = \sum_{k=0}^{\infty} \exp(iE_k t)\phi_k f_E(E_k).$$ Equation A-9

As the time parameter has no direct shape interpretation the authors choose to average the probability over time to produce a Wave Kernel Signature,

$$WKS(E,x) = \sum_{k=0}^{\infty} \phi_k(x)^2 f_E(E_k)^2.$$ Equation A-10

In practice the authors choose a log scale of E and sample values of the signature over that scale.

The Wave Kernel samples the spectrum at a specified number of intervals, called evaluations in the original paper (Aubry et al., 2011). The Wave Kernel splits the spectrum up in to intervals and then amplifies the signal from the spectrum that falls into those intervals. In this respect, the Wave Kernel Signature acts as a band pass filter from signal processing. The functional form is taken from the log power distribution and filters the spectrum by amplifying the distribution around an initial mean value given by the logarithm of the value of the chosen interval, of the spectrum, $\log e$,

$$\tau(\lambda) = \exp\left(-\frac{(\log e - \log\lambda)^2}{2\sigma^2}\right).$$ Equation A-11

A.3 General descriptors

The heat kernel and the wave kernel signatures can be generalised to the functional form given by,

$$f(x) = \sum_k \tau(\lambda_k)\phi(x)_k^2$$ Equation A-12

where λ and ϕ are the first k eigenvalues and eigenfunctions of the Laplace-Beltrami operator respectively. The function $\tau(\lambda_k)$ is typically a transfer function that acts upon the eigenvalues. In this framework, the local descriptor is a general mapping from points on a manifold $x \in M$ to a vector of Q values, $f: M \rightarrow \mathbb{R}^Q$. Transfer functions are used to change the spectrum of the data by amplifying important parts or removing unwanted parts of the spectrum. These local descriptors are isometric by construction because of the way that the spectrum of the Laplace-Beltrami operator is used. The properties of the local descriptor, in terms of the information that is captured, depend upon the influence of the different sections of the transfer functions. The analysis of these signal processing filter properties is useful in understanding the aspects of the shape that the local geometry descriptors are representing.

For a sample of N points on the surface, the final local geometry descriptor is an $N \times D$ matrix, F. The various explanations of the local geometry descriptor presented so far in this section can be categorised in two groups: first a row-wise interpretation where each point on the surface is assigned a vector over a number of filter functions at that point; second a column-wise interpretation where each column is a transfer function evaluated over the surface. In the row-wise view the local geometry of a sample point is encoded by sampling different filters with respect to the eigenfunction of the point, each of which constitutes a row in the final matrix. Whereas in the column-wise view, the filtered geometric properties of the spectrum are projected onto the surface and values are assigned to each sample point, which produces a column in the final matrix.

Therefore, we can combine these two approaches in an elegant linear algebra expression. Let the filter bank be a $D \times k$ matrix filter functions over the eigenvalues, $T = \left(\tau_1(\lambda), \tau_2(\lambda), \dots, \tau_D(\lambda)\right)^\top$, that filter the k eigenvalues λ, so that each row corresponds to k dimensional vector of filtered eigenvalues. The squared eigenfunctions matrix is a $N \times k$ matrix, Φ^2. Therefore, a general formulation of the local geometry descriptor is,

$$F = \Phi^2 T^\top. \qquad \text{Equation A-13}$$

From Equation A-13 the row-wise and column-wise views can be derived. For a point on the surface, $x_i \in M$, the eigenfunction from that point is $\phi(x_i)$ and the point descriptor is the dot product of the filter-bank. Therefore, the i^{th} row in the local geometry descriptor is,

$$F_{i,:} = \left(\tau_1(\lambda) \cdot \phi^2(x_i), \tau_2(\lambda) \cdot \phi^2(x_i), \dots, \tau_D(\lambda) \cdot \phi^2(x_i)\right),$$ Equation A-14

where $\tau_j(\lambda) \cdot \phi^2(x)$ is the dot product of the j^{th} filter bank with the eigenfunction relating to the point on the surface, $x \in M$,

$$\tau_j(\lambda) \cdot \phi^2(x_i) = \sum_{k=1}^{K} \tau_j(\lambda_k)\phi_k^2(x_i).$$ Equation A-15

Note that this is equivalent to Equation A-12.

On the other hand, the column-wise interpretation can be viewed as the inner product of the jth filter bank, $\tau_j(\lambda)$ with all eigenfunctions in Φ,

$$F_{:,j} = \Phi^2 \tau_j(\lambda),$$ Equation A-16

which can be written in an element wise manner as,

$$F_{:,j} = \left(\tau_j(\lambda) \cdot \phi^2(x_1), \tau_j(\lambda) \cdot \phi^2(x_2), \dots, \tau_j(\lambda) \cdot \phi^2(x_N)\right)^{\mathsf{T}}.$$ Equation A-17

Again notice that this is equivalent to Equation A-12.

These two different interpretations are important for the applications of the local geometry descriptors. On the one hand, the goal of the local geometry descriptor is to describe the geometry around a single point – row-wise – yet the properties of the vertex descriptors are derived from the properties of the transfer functions over the whole shape – column wise.

B Finite element method for obtaining the spectrum of the Laplace-Beltrami operator

The indirect approach to computing the eigenfunctions and eigenvalues of the discrete Laplace-Beltrami the uses the finite element method (FEM) which computes the spectrum without having to approximate it directly (Reuter et al., 2006). FEM is a method for estimating solutions to partial differential equations (G. R. Liu & Quek, 2014). Rather than estimating the Laplacian directly, FEM takes the partial differential equation recall from Equation A-2 that

$$\Delta_g f = -\lambda f$$

and assumes that a solution exists using a basis Ψ. This means that the expression can be rewritten as,

$$\langle \Delta_g f, \psi_i \rangle = -\lambda \langle f, \psi_i \rangle, \qquad \forall\, \psi_i \in \Psi \qquad \text{Equation B-1}$$

for any smooth basis Ψ that forms a finite basis that spans the manifold. In particular, this allows the eigenfunction f to be written as a linear combination of the basis, so that supposing the finite basis has q basis functions, $f = u_1\psi_1 + \cdots + u_i\psi_i + \cdots + u_q\psi_q$. Then, by summing over the basis, the equation can be written as,

$$\sum_q u_q \langle \Delta_g \psi_q, \psi_r \rangle = -\lambda \sum_q u_q \langle \psi_q, \psi_r \rangle \qquad \text{Equation B-2}$$

for a given ψ_r. Given that the basis functions are known, the eigenvalues and eigenfunctions can be solved directly in terms of the basis. This transforms the eigendecomposition problem into one of solving the general eigenvalue problem,

$$Au = \lambda Bu \qquad \text{Equation B-3}$$

where $a_{rj} = \langle \Delta_g \psi_j, \psi_r \rangle$ and $b_{rj} = \langle \psi_j, \psi_r \rangle$.

9 References

Accelrys, Inc. (n.d.). Comprehensive Medicinal Chemistry (CMC). Retrieved 29 August 2014, from http://accelrys.com/products/databases/bioactivity/comprehensive-medicinal-chemistry.html

Armstrong, M. S., Finn, P. W., Morris, G. M., & Richards, W. G. (2011). Improving the accuracy of ultrafast ligand-based screening: incorporating lipophilicity into ElectroShape as an extra dimension. *Journal of Computer-Aided Molecular Design*, *25*(8), 785–790. https://doi.org/10.1007/s10822-011-9463-8

Armstrong, M. S., Morris, G. M., Finn, P. W., Sharma, R., Moretti, L., Cooper, R. I., & Richards, W. G. (2010). ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *Journal of Computer-Aided Molecular Design*, *24*(9), 789–801. https://doi.org/10.1007/s10822-010-9374-0

Armstrong, M. S., Morris, G. M., Finn, P. W., Sharma, R., & Richards, W. G. (2009). Molecular similarity including chirality. *Journal of Molecular Graphics and Modelling*, *28*(4), 368–370. https://doi.org/10.1016/j.jmgm.2009.09.002

Arteca, G. A., & Mezey, P. G. (1988). Shape characterization of some molecular model surfaces. *Journal of Computational Chemistry*, *9*(5), 554–563. https://doi.org/10.1002/jcc.540090513

Ash, S., Cline, M. A., Homer, R. W., Hurst, T., & Smith, G. B. (1997). SYBYL line notation (SLN): a versatile language for chemical structure representation. *Journal of Chemical Information and Computer Sciences*, *37*(1), 71–79. https://doi.org/10.1021/ci960109j

Aubry, M., Schlickewei, U., & Cremers, D. (2011). The wave kernel signature: A quantum mechanical approach to shape analysis. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (pp. 1626–1633). https://doi.org/10.1109/ICCVW.2011.6130444

Axen, S. D., Huang, X.-P., Cáceres, E. L., Gendelev, L., Roth, B. L., & Keiser, M. J. (2017). A simple representation of three-dimensional molecular structure. *Journal of Medicinal Chemistry*. https://doi.org/10.1021/acs.jmedchem.7b00696

Axenopoulos, A., Rafailidis, D., Papadopoulos, G., Houstis, E. N., & Daras, P. (2016). Similarity search of flexible 3D molecules combining local and global shape descriptors. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *13*(5), 954–970. https://doi.org/10.1109/TCBB.2015.2498553

Bajorath, J., Peltason, L., Wawer, M., Guha, R., Lajiness, M. S., & Van Drie, J. H. (2009). Navigating structure–activity landscapes. *Drug Discovery Today*, *14*(13–14), 698–705. https://doi.org/10.1016/j.drudis.2009.04.003

Baker, M. (2013). Fragment-based lead discovery grows up. *Nature Reviews Drug Discovery*, *12*(1), 5–7. https://doi.org/10.1038/nrd3926

Ballester, P. J. (2011). Ultrafast shape recognition: method and applications. *Future Medicinal Chemistry*, *3*(1), 65–78. https://doi.org/10.4155/fmc.10.280

Ballester, P. J., & Richards, W. G. (2007). Ultrafast shape recognition to search compound databases for similar molecular shapes. *Journal of Computational Chemistry*, *28*(10), 1711–1723. https://doi.org/10.1002/jcc.20681

Barnard, J. M. (1993). Substructure searching methods: Old and new. *Journal of Chemical Information and Computer Sciences*, *33*(4), 532–538. https://doi.org/10.1021/ci00014a001

Barnard, J. M. (2003). Representation of molecular structures-overview. In J. Gasteiger (Ed.), *Handbook of Chemoinformatics* (pp. 27–50). Wiley-VCH Verlag GmbH. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9783527618279.ch3/summary

Beddell, C. R., Goodford, P. J., Norrington, F. E., Wilkinson, S., & Wootton, R. (1976). Compounds designed to fit a site of known structure in human haemoglobin. *British Journal of Pharmacology*, *57*(2), 201–209. https://doi.org/10.1111/j.1476-5381.1976.tb07468.x

Belkin, M., Sun, J., & Wang, Y. (2008). Discrete Laplace operator on meshed surfaces. In *Proceedings of the Twenty-fourth Annual Symposium on Computational Geometry* (pp. 278–287). New York, NY, USA: ACM. https://doi.org/10.1145/1377676.1377725

Bemis, G. W., & Murcko, M. A. (1996). The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, *39*(15), 2887–2893. https://doi.org/10.1021/jm9602928

Bemis, G. W., & Murcko, M. A. (1999). Properties of known drugs. 2. side chains. *Journal of Medicinal Chemistry*, *42*(25), 5095–5099. https://doi.org/10.1021/jm9903996

Bender, A., Mussa, H. Y., Glen, R. C., & Reiling, S. (2004a). Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier. *Journal of Chemical Information and Computer Sciences*, *44*(1), 170–178. https://doi.org/10.1021/ci034207y

Bender, A., Mussa, H. Y., Glen, R. C., & Reiling, S. (2004b). Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *Journal of Chemical Information and Computer Sciences*, *44*(5), 1708–1718. https://doi.org/10.1021/ci0498719

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., … Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, *112*(3), 535–542.

Bietz, S., Urbaczek, S., Schulz, B., & Rarey, M. (2014). Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *Journal of Cheminformatics*, *6*(1). https://doi.org/10.1186/1758-2946-6-12

Boscaini, D., Masci, J., Melzi, S., Bronstein, M. M., Castellani, U., & Vandergheynst, P. (2015). Learning class-specific descriptors for deformable shapes using localized spectral convolutional networks. *Computer Graphics Forum*, *34*(5), 13–23. https://doi.org/10.1111/cgf.12693

Boscaini, D., Masci, J., Rodolà, E., Bronstein, M. M., & Cremers, D. (2016). Anisotropic diffusion descriptors. *Computer Graphics Forum*, *35*(2), 431–441. https://doi.org/10.1111/cgf.12844

Botsch, M. (Ed.). (2010). *Polygon mesh processing*. Natick, Mass: A K Peters.

Botsch, M., Pauly, M., Rossl, C., Bischoff, S., & Kobbelt, L. (2006). Geometric modeling based on triangle meshes. In *ACM SIGGRAPH 2006 Courses*. New York, NY, USA: ACM. https://doi.org/10.1145/1185657.1185839

Boyd, S. M., Turnbull, A. P., & Walse, B. (2012). Fragment library design considerations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *2*(6), 868–885. https://doi.org/10.1002/wcms.1098

Bronstein, A. M., Bronstein, M. M., Castellani, U., Dubrovina, A., Guibas, L. J., Horaud, R. P., … Sharma, A. (2010). SHREC'10 Track: Correspondence Finding. In *Eurographics Workshop on 3D Object Retrieval*. Retrieved from https://doi.org/10.2312/3DOR/3DOR10/087-091

Bronstein, A. M., Bronstein, M. M., Guibas, L. J., & Ovsjanikov, M. (2011). Shape Google: geometric words and expressions for invariant shape retrieval. *ACM Transactions on Graphics*, *30*(1), 1:1–1:20. https://doi.org/10.1145/1899404.1899405

BROOD. (2006). Retrieved 16 June 2014, from http://www.eyesopen.com/brood

Bultinck, P., Gironés, X., & Carbó-Dorca, R. (2005). Molecular quantum similarity: theory and applications. In K. B. Lipkowitz, R. Larter, & T. R. Cundari (Eds.), *Reviews in Computational Chemistry* (pp. 127–207). John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/0471720895.ch2/summary

Calvetti, D., Reichel, L., & Sorensen, D. C. (1994). An implicitly restarted Lanczos method for large symmetric eigenvalue problems. *Electronic Transactions on Numerical Analysis*, *2*, 1–21.

Carbó-Dorca, R. (2000). *Molecular quantum similarity in QSAR and drug design*. Berlin ; New York: Springer.

Carbó-Dorca, R. (2013). Quantum similarity. In S. K. Ghosh & P. K. Chattaraj (Eds.), *Concepts and Methods in Modern Theoretical Chemistry*. Retrieved from http://www.crcpress.com/product/isbn/9781466505285

Carbó-Dorca, R., Leyda, L., & Arnau, M. (1980). How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *International Journal of Quantum Chemistry*, *17*(6), 1185–1189. https://doi.org/10.1002/qua.560170612

Chau, P. L., & Dean, P. M. (1987). Molecular recognition: 3D surface structure comparison by gnomonic projection. *Journal of Molecular Graphics*, *5*(2), 97–100. https://doi.org/10.1016/0263-7855(87)80007-3

Chen, M., & Lu, B. (2011). TMSmesh: a robust method for molecular surface mesh generation using a trace technique. *Journal of Chemical Theory and Computation*, *7*(1), 203–212. https://doi.org/10.1021/ct100376g

Chen, M., & Lu, B. (2013). Advances in biomolecular surface meshing and its applications to mathematical modeling. *Chinese Science Bulletin*, *58*(16), 1843–1849. https://doi.org/10.1007/s11434-013-5829-8

Chen, M., Tu, B., & Lu, B. (2012). Triangulated manifold meshing method preserving molecular surface topology. *Journal of Molecular Graphics and Modelling*, *38*, 411–418. https://doi.org/10.1016/j.jmgm.2012.09.006

Chen, X., Liu, M., & Gilson, M. (2001). BindingDB: a web-accessible molecular recognition database. *Combinatorial Chemistry & High Throughput Screening*, *4*(8), 719–725. https://doi.org/10.2174/1386207013330670

Clark, R. D., & Webster-Clark, D. J. (2008). Managing bias in ROC curves. *Journal of Computer-Aided Molecular Design*, *22*(3–4), 141–146. https://doi.org/10.1007/s10822-008-9181-z

Cohen, S. (1977). A strategy for the chemotherapy of infectious disease. *Science*, *197*(4302), 431–432. https://doi.org/10.1126/science.195340

Coifman, R. R., & Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, *21*(1), 5–30. https://doi.org/10.1016/j.acha.2006.04.006

Connolly, M. L. (1985). Computation of molecular volume. *Journal of the American Chemical Society*, *107*(5), 1118–1124. https://doi.org/10.1021/ja00291a006

Degen, J., Wegscheid-Gerlach, C., Zaliani, A., & Rarey, M. (2008). On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, *3*(10), 1503–1507. https://doi.org/10.1002/cmdc.200800178

Desaphy, J., & Rognan, D. (2014). sc-PDB-Frag: a database of protein–ligand interaction patterns for bioisosteric replacements. *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/ci500282c

DUD-E diverse subset. (n.d.). Retrieved 29 August 2017, from http://dude.docking.org/subsets/diverse

Duncan, B. S., & Olson, A. J. (1993). Shape analysis of molecular surfaces. *Biopolymers*, *33*(2), 231–238. https://doi.org/10.1002/bip.360330205

Eldar, Y., Lindenbaum, M., Porat, M., & Zeevi, Y. Y. (1997). The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, *6*(9), 1305–1315. https://doi.org/10.1109/83.623193

Erlanson, D. A. (2012). Introduction to fragment-based drug discovery. *Topics in Current Chemistry*, *317*, 1–32. https://doi.org/10.1007/128_2011_180

Erlanson, D. A., McDowell, R. S., & O'Brien, T. (2004). Fragment-based drug discovery. *Journal of Medicinal Chemistry*, *47*(14), 3463–3482. https://doi.org/10.1021/jm040031v

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fechner, U., & Schneider, G. (2006). Flux (1): a virtual synthesis scheme for fragment-based de novo design. *Journal of Chemical Information and Modeling*, *46*(2), 699–707. https://doi.org/10.1021/ci0503560

Fechner, U., & Schneider, G. (2007). Flux (2): comparison of molecular mutation and crossover operators for ligand-based de novo design. *Journal of Chemical Information and Modeling*, *47*(2), 656–667. https://doi.org/10.1021/ci6005307

Finn, P. W., & Morris, G. M. (2013). Shape-based similarity searching in chemical databases. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *3*(3), 226–241. https://doi.org/10.1002/wcms.1128

Forli, S. (2015). Charting a path to success in virtual screening. *Molecules*, *20*(12), 18732–18758. https://doi.org/10.3390/molecules201018732

Franco, P., Porta, N., Holliday, J. D., & Willett, P. (2014). The use of 2D fingerprint methods to support the assessment of structural similarity in orphan drug legislation. *Journal of Cheminformatics*, *6*(1), 5. https://doi.org/10.1186/1758-2946-6-5

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., … Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, *40*(D1), D1100–D1107. https://doi.org/10.1093/nar/gkr777

Giangreco, I., Cosgrove, D. A., & Packer, M. J. (2013). An extensive and diverse set of molecular overlays for the validation of pharmacophore programs. *Journal of Chemical Information and Modeling*, *53*(4), 852–866. https://doi.org/10.1021/ci400020a

Good, A. C., & Oprea, T. I. (2008). Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *Journal of Computer-Aided Molecular Design*, *22*(3–4), 169–178. https://doi.org/10.1007/s10822-007-9167-2

Grant, J. A., Gallardo, M. A., & Pickup, B. T. (1996). A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *Journal of Computational Chemistry*, *17*(14), 1653–1666. https://doi.org/10.1002/(SICI)1096-987X(19961115)17:14<1653::AID-JCC7>3.0.CO;2-K

Grant, J. A., & Pickup, B. T. (1995). A Gaussian description of molecular shape. *Journal of Physical Chemistry*, *99*(11), 3503–3510. https://doi.org/10.1021/j100011a016

Grinspun, E., Desbrun, M., Polthier, K., Schröder, P., & Stern, A. (2006). Discrete differential geometry: an applied introduction. *ACM SIGGRAPH Course*, *7*.

Guasch, L., Yapamudiyansel, W., Peach, M. L., Kelley, J. A., Barchi, J. J., & Nicklaus, M. C. (2016). Experimental and chemoinformatics study of tautomerism in a database of commercially available screening samples. *Journal of Chemical Information and Modeling*, *56*(11), 2149–2161. https://doi.org/10.1021/acs.jcim.6b00338

Hamza, A., Wei, N.-N., Hao, C., Xiu, Z., & Zhan, C.-G. (2013). A novel and efficient ligand-based virtual screening approach using the HWZ scoring function and an enhanced shape-density model. *Journal of Biomolecular Structure and Dynamics*, *31*(11), 1236–1250. https://doi.org/10.1080/07391102.2012.732341

Hamza, A., Wei, N.-N., & Zhan, C.-G. (2012). Ligand-based virtual screening approach using a new scoring function. *Journal of Chemical Information and Modeling*, *52*(4), 963–974. https://doi.org/10.1021/ci200617d

Hawkins, P. C. D., & Nicholls, A. (2012). Conformer generation with OMEGA: learning from the data set and the analysis of failures. *Journal of Chemical Information and Modeling*, *52*(11), 2919–2936. https://doi.org/10.1021/ci300314k

Hawkins, P. C. D., Skillman, A. G., Warren, G. L., Ellingson, B. A., & Stahl, M. T. (2010). Conformer generation with OMEGA: algorithm and validation using high quality structures from the protein databank and Cambridge structural database. *Journal of Chemical Information and Modeling*, *50*(4), 572–584. https://doi.org/10.1021/ci100031x

Ho, C. M. W., & Marshall, G. R. (1993). SPLICE: A program to assemble partial query solutions from three-dimensional database searches into novel ligands. *Journal of Computer-Aided Molecular Design*, *7*(6), 623–647. https://doi.org/10.1007/BF00125322

Holliday, J. D., Jelfs, S. P., Willett, P., & Gedeck, P. (2003). Calculation of intersubstituent similarity using R-group descriptors. *Journal of Chemical Information and Computer Sciences*, *43*(2), 406–411. https://doi.org/10.1021/ci025589v

Horvath, D., Marcou, G., & Varnek, A. (2013). Do not hesitate to use Tversky—and other hints for successful active analogue searches with feature count descriptors. *Journal of Chemical Information and Modeling*, *53*(7), 1543–1562. https://doi.org/10.1021/ci400106g

Hu, G., Kuang, G., Xiao, W., Li, W., Liu, G., & Tang, Y. (2012). Performance evaluation of 2D fingerprint and 3D shape similarity methods in virtual screening. *Journal of Chemical Information and Modeling*, *52*(5), 1103–1113. https://doi.org/10.1021/ci300030u

Huang, N., Shoichet, B. K., & Irwin, J. J. (2006). Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry*, *49*(23), 6789–6801. https://doi.org/10.1021/jm0608356

Huang, Q., Wang, F., & Guibas, L. (2014). Functional map networks for analyzing and exploring large shape collections. *ACM Transactions on Graphics*, *33*(4), 36:1–36:11. https://doi.org/10.1145/2601097.2601111

Jahn, A., Hinselmann, G., Fechner, N., & Zell, A. (2009). Optimal assignment methods for ligand-based virtual screening. *Journal of Cheminformatics*, *1*(1), 14. https://doi.org/10.1186/1758-2946-1-14

Jakobi, A.-J., Mauser, H., & Clark, T. (2008). ParaFrag—an approach for surface-based similarity comparison of molecular fragments. *Journal of Molecular Modeling*, *14*(7), 547–558. https://doi.org/10.1007/s00894-008-0302-3

James, C. A., Weininger, D., & Delany, J. (2011). *Daylight theory manual - Daylight 4.9*. Daylight Chemical Information Systems, Inc. Retrieved from http://www.daylight.com/dayhtml/doc/theory/

Johnson, M. A., & Maggiora, G. M. (1990). *Concepts and applications of molecular similarity*. New York: Wiley.

Joseph-McCarthy, D., Campbell, A. J., Kern, G., & Moustakas, D. (2014). Fragment-based lead discovery and design. *Journal of Chemical Information and Modeling*, *54*(3), 693–704. https://doi.org/10.1021/ci400731w

Kalliokoski, T., Olsson, T. S. G., & Vulpetti, A. (2013). Subpocket analysis method for fragment-based drug discovery. *Journal of Chemical Information and Modeling*, *53*(1), 131–141. https://doi.org/10.1021/ci300523r

Kennewell, E. A., Willett, P., Ducrot, P., & Luttmann, C. (2006). Identification of target-specific bioisosteric fragments from ligand–protein crystallographic data. *Journal of Computer-Aided Molecular Design*, *20*(6), 385–394. https://doi.org/10.1007/s10822-006-9072-0

Kihara, D., Sael, L., Chikhi, R., & Esquivel-Rodriguez, J. (2011). Molecular surface representation using 3D Zernike descriptors for protein shape comparison and docking. *Current Protein & Peptide Science*, *12*(6), 520–530.

Kirchmair, J., Markt, P., Distinto, S., Wolber, G., & Langer, T. (2008). Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—What can we learn from earlier mistakes? *Journal of Computer-Aided Molecular Design*, *22*(3–4), 213–228. https://doi.org/10.1007/s10822-007-9163-6

Kovnatsky, A., Bronstein, M. M., Bresson, X., & Vandergheynst, P. (2015). Functional correspondence by matrix completion. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on* (pp. 905–914). IEEE. https://doi.org/10.1109/CVPR.2015.7298692

Kovnatsky, A., Bronstein, M. M., Bronstein, A. M., Glashoff, K., & Kimmel, R. (2013). Coupled quasi-harmonic bases. *Computer Graphics Forum*, *32*(2pt4), 439–448. https://doi.org/10.1111/cgf.12064

Kovnatsky, A., Bronstein, M. M., Bronstein, A. M., & Kimmel, R. (2012). Photometric heat kernel signatures. In *Proceedings of the Third International Conference on Scale Space and Variational Methods in Computer Vision* (pp. 616–627). Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-642-24785-9_52

Landrum, G. (n.d.). RDKit: Open-source cheminformatics. Retrieved 27 November 2013, from http://www.rdkit.org

Leach, A. R., & Gillet, V. J. (2007). *An introduction to chemoinformatics*. Dordrecht: Springer. Retrieved from http://dx.doi.org/10.1007/978-1-4020-6291-9

Leicester, S., Finney, J., & Bywater, R. (1994). A quantitative representation of molecular surface shape. I: Theory and development of the method. *Journal of Mathematical Chemistry*, *16*(1), 315–341. https://doi.org/10.1007/BF01169216

Levy, B. (2006). Laplace-Beltrami eigenfunctions towards an algorithm that 'understands' geometry. In *IEEE International Conference on Shape Modeling and Applications, 2006. SMI 2006* (pp. 13–13). https://doi.org/10.1109/SMI.2006.21

Lewell, X. Q., Judd, D. B., Watson, S. P., & Hann, M. M. (1998). RECAP Retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of Chemical Information and Computer Sciences*, *38*(3), 511–522. https://doi.org/10.1021/ci970429i

Li, C., & Hamza, A. B. (2013). Intrinsic spatial pyramid matching for deformable 3D shape retrieval. *International Journal of Multimedia Information Retrieval*, *2*(4), 261–271. https://doi.org/10.1007/s13735-013-0041-9

Litany, O., Rodolà, E., Bronstein, A. M., Bronstein, M. M., & Cremers, D. (2016). Non-rigid puzzles. *Computer Graphics Forum*, *35*(5), 135–143. https://doi.org/10.1111/cgf.12970

Litman, R., & Bronstein, A. M. (2014). Learning spectral descriptors for deformable shape correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(1), 171–180. https://doi.org/10.1109/TPAMI.2013.148

Litman, R., Bronstein, A. M., Bronstein, M. M., & Castellani, U. (2014). Supervised learning of bag-of-features shape descriptors using sparse coding. *Computer Graphics Forum*, *33*(5), 127–136. https://doi.org/10.1111/cgf.12438

Liu, G. R., & Quek, S. S. (2014). *The finite element method: a practical course* (Second edition). Amsterdam ; Oxford: Butterworth-Heinemann.

Liu, X., Jiang, H., & Li, H. (2011). SHAFTS: a hybrid approach for 3D molecular similarity calculation. 1. method and assessment of virtual screening. *Journal of Chemical Information and Modeling*, *51*(9), 2372–2385. https://doi.org/10.1021/ci200060s

Liu, Y.-S., Ramani, K., & Liu, M. (2011). Computing the inner distances of volumetric models for articulated shape description with a visibility graph. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(12), 2538–2544. https://doi.org/10.1109/TPAMI.2011.116

Lounkine, E., Batista, J., & Bajorath, J. (2008). Random molecular fragment methods in computational medicinal chemistry. *Current Medicinal Chemistry*, *15*(21), 2108–2121.

Lu, W., Liu, X., Cao, X., Xue, M., Liu, K., Zhao, Z., … Li, H. (2011). SHAFTS: a hybrid approach for 3D molecular similarity calculation. 2. prospective case study in the discovery of diverse p90 ribosomal S6 protein kinase 2 inhibitors to suppress cell migration. *Journal of Medicinal Chemistry*, *54*(10), 3564–3574. https://doi.org/10.1021/jm200139j

Maass, P., Schulz-Gasch, T., Stahl, M., & Rarey, M. (2007). Recore: a fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. *Journal of Chemical Information and Modeling*, *47*(2), 390–399. https://doi.org/10.1021/ci060094h

MACCS structural keys. (2011). Accelrys, San Diego, CA.

Maggiora, G. M., & Shanmugasundaram, V. (2011). Molecular similarity measures. In J. Bajorath (Ed.), *Chemoinformatics and Computational Chemical Biology* (Vol. 672, pp. 39–100). Totowa, NJ: Humana Press. Retrieved from http://www.springerprotocols.com/Abstract/doi/10.1007/978-1-60761-839-3_2

Maggiora, G. M., Vogt, M., Stumpfe, D., & Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, *57*(8), 3186–3204. https://doi.org/10.1021/jm401411z

Martin, Y. C. (2009). Let’s not forget tautomers. *Journal of Computer-Aided Molecular Design*, *23*(10), 693–704. https://doi.org/10.1007/s10822-009-9303-2

Masci, J., Boscaini, D., Bronstein, M. M., & Vandergheynst, P. (2015). ShapeNet: convolutional neural networks on non-Euclidean manifolds. *arXiv:1501.06297 [Cs]*. Retrieved from http://arxiv.org/abs/1501.06297

Masek, B. B., Merchant, A., & Matthew, J. B. (1993). Molecular shape comparison of angiotensin II receptor antagonists. *Journal of Medicinal Chemistry*, *36*(9), 1230–1238. https://doi.org/10.1021/jm00061a014

Mavridis, L., Hudson, B. D., & Ritchie, D. W. (2007). Toward high throughput 3D virtual screening using spherical harmonic surface representations. *Journal of Chemical Information and Modeling*, *47*(5), 1787–1796. https://doi.org/10.1021/ci7001507

Mezey, P. G. (1987). The shape of molecular charge distributions: Group theory without symmetry. *Journal of Computational Chemistry*, *8*(4), 462–469. https://doi.org/10.1002/jcc.540080426

Morgan, H. L. (1965). The generation of a unique machine description for chemical structures-a technique developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, *5*(2), 107–113. https://doi.org/10.1021/c160017a018

Morley, A. D., Pugliese, A., Birchall, K., Bower, J., Brennan, P., Brown, N., … Wyatt, P. G. (2013). Fragment-based hit identification: thinking in 3D. *Drug Discovery Today*, *18*(23–24), 1221–1227. https://doi.org/10.1016/j.drudis.2013.07.011

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. Cambridge, MA: MIT Press.

Nettles, J. H., Jenkins, J. L., Bender, A., Deng, Z., Davies, J. W., & Glick, M. (2006). Bridging chemical and biological space: ‘target fishing’ using 2D and 3D molecular descriptors. *Journal of Medicinal Chemistry*, *49*(23), 6802–6810. https://doi.org/10.1021/jm060902w

Nicholls, A. (2008). What do we know and when do we know it? *Journal of Computer-Aided Molecular Design*, *22*(3–4), 239–255. https://doi.org/10.1007/s10822-008-9170-2

Nicholls, A., MacCuish, N. E., & MacCuish, J. D. (2004). Variable selection and model validation of 2D and 3D molecular descriptors>. *Journal of Computer-Aided Molecular Design*, *18*(7–9), 451–474. https://doi.org/10.1007/s10822-004-5202-8

Nicholls, A., McGaughey, G. B., Sheridan, R. P., Good, A. C., Warren, G., Mathieu, M., … Kelley, B. (2010). Molecular shape and medicinal chemistry: a perspective. *Journal of Medicinal Chemistry*, *53*(10), 3862–3886. https://doi.org/10.1021/jm900818s

Nisius, B., & Gohlke, H. (2012). Alignment-independent comparison of binding sites based on DrugScore potential fields encoded by 3D Zernike descriptors. *Journal of Chemical Information and Modeling*, *52*(9), 2339–2347. https://doi.org/10.1021/ci300244y

Nisius, B., & Rester, U. (2009). Fragment shuffling: an automated workflow for three-dimensional fragment-based ligand design. *Journal of Chemical Information and Modeling*, *49*(5), 1211–1222. https://doi.org/10.1021/ci8004572

Novotni, M., & Klein, R. (2003). 3D Zernike descriptors for content based shape retrieval. *Proceedings of the 8th ACM Symposium on Solid Modeling and Applications*. Retrieved from http://cg.cs.uni-bonn.de/aigaion2root/attachments/novotni-2003-3d.pdf

OpenEye ROCS. (n.d.). OpenEye Scientific Software, Santa Fe, NM, USA. Retrieved from https://www.eyesopen.com/rocs

Ovsjanikov, M., Ben-Chen, M., Chazal, F., & Guibas, L. (2013). Analysis and visualization of maps between shapes: analysis and visualization of maps between shapes. *Computer Graphics Forum*, *32*(6), 135–145. https://doi.org/10.1111/cgf.12076

Ovsjanikov, M., Ben-Chen, M., Solomon, J., Butscher, A., & Guibas, L. (2012). Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics*, *31*(4), 30:1–30:11. https://doi.org/10.1145/2185520.2185526

Ovsjanikov, M., Bronstein, A. M., Bronstein, M. M., & Guibas, L. J. (2009). Shape Google: a computer vision approach to isometry invariant shape retrieval. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops* (pp. 320–327). IEEE. https://doi.org/10.1109/ICCVW.2009.5457682

Park, M.-S., Gao, C., & Stern, H. A. (2011). Estimating binding affinities by docking/scoring methods using variable protonation states. *Proteins: Structure, Function, and Bioinformatics*, *79*(1), 304–314. https://doi.org/10.1002/prot.22883

Peltason, L., & Bajorath, J. (2007). SAR index: quantifying the nature of structure−activity relationships. *Journal of Medicinal Chemistry*, *50*(23), 5571–5578. https://doi.org/10.1021/jm0705713

Peréz-Nueno, V. I., Ritchie, D. W., Rabal, O., Pascual, R., Borrell, J. I., & Teixido, J. (2008). Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand−receptor docking. *Journal of Chemical Information and Modeling*, *48*(3), 509–533. https://doi.org/10.1021/ci700415g

Pérez-Nueno, V. I., Venkatraman, V., Mavridis, L., Clark, T., & Ritchie, D. W. (2011). Using spherical harmonic surface property representations for ligand-based virtual screening. *Molecular Informatics*, *30*(2–3), 151–159. https://doi.org/10.1002/minf.201000149

Perkins, T. D. J., Mills, J. E. J., & Dean, P. M. (1995). Molecular surface-volume and property matching to superpose flexible dissimilar molecules. *Journal of Computer-Aided Molecular Design*, *9*(6), 479–490. https://doi.org/10.1007/BF00124319

Petrone, P. M., Simms, B., Nigsch, F., Lounkine, E., Kutchukian, P., Cornett, A., … Glick, M. (2012). Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS Chemical Biology*, *7*(8), 1399–1409. https://doi.org/10.1021/cb3001028

Pickett, S. D., Mason, J. S., & McLay, I. M. (1996). Diversity profiling and design using 3D pharmacophores: Pharmacophore-Derived Queries (PDQ). *Journal of Chemical Information and Computer Sciences*, *36*(6), 1214–1223. https://doi.org/10.1021/ci960039g

Pierce, A. C., Rao, G., & Bemis, G. W. (2004). BREED: generating novel inhibitors through hybridization of known ligands. Application to CDK2, P38, and HIV Protease. *Journal of Medicinal Chemistry*, *47*(11), 2768–2775. https://doi.org/10.1021/jm030543u

Pinkall, U., & Polthier, K. (1993). Computing discrete minimal surfaces and their conjugates. *Experimental Mathematics*, *2*(1), 15–36. https://doi.org/10.1080/10586458.1993.10504266

Pokrass, J., Bronstein, A. M., Bronstein, M. M., Sprechmann, P., & Sapiro, G. (2016). Sparse models for intrinsic shape correspondence. In M. Breuß, A. Bruckstein, P. Maragos, & S. Wuhrer (Eds.), *Perspectives in Shape Analysis* (pp. 211–230). Cham: Springer International Publishing. Retrieved from http://link.springer.com/10.1007/978-3-319-24726-7_10

Raghavendra, A. S., & Maggiora, G. M. (2007). Molecular basis sets: a general similarity-based approach for representing chemical spaces. *Journal of Chemical Information and Modeling*, *47*(4), 1328–1340. https://doi.org/10.1021/ci600552n

Ray, L. C., & Kirsch, R. A. (1957). Finding chemical records by digital computers. *Science*, *126*(3278), 814–819. https://doi.org/10.1126/science.126.3278.814

Rees, D. C., Congreve, M., Murray, C. W., & Carr, R. (2004). Fragment-based lead discovery. *Nature Reviews Drug Discovery*, *3*(8), 660–672. https://doi.org/10.1038/nrd1467

Reuter, M. (2009). Hierarchical shape segmentation and registration via topological features of Laplace-Beltrami eigenfunctions. *International Journal of Computer Vision*, *89*(2–3), 287–308. https://doi.org/10.1007/s11263-009-0278-1

Reuter, M., Biasotti, S., Giorgi, D., Patanè, G., & Spagnuolo, M. (2009). Discrete Laplace–Beltrami operators for shape analysis and segmentation. *Computers & Graphics*, *33*(3), 381–390. https://doi.org/10.1016/j.cag.2009.03.005

Reuter, M., Wolter, F.-E., & Peinecke, N. (2006). Laplace-Beltrami spectra as ‘shape-DNA’ of surfaces and solids. *Computer-Aided Design*, *38*(4), 342–366. https://doi.org/10.1016/j.cad.2005.10.011

Riniker, S., & Landrum, G. A. (2013). Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. *Journal of Cheminformatics*, *5*(1), 43. https://doi.org/10.1186/1758-2946-5-43

Ritchie, D. W., & Kemp, G. J. L. (1999). Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *Journal of Computational Chemistry*, *20*(4), 383–395. https://doi.org/10.1002/(SICI)1096-987X(199903)20:4<383::AID-JCC1>3.0.CO;2-M

Ritchie, D. W., & Pérez-Nueno, V. I. (2013). Spherical harmonic molecular surfaces (ParaSurf and ParaFit). In N. Brown (Ed.), *Scaffold Hopping in Medicinal Chemistry* (pp. 183–194). Wiley-VCH Verlag GmbH & Co. KGaA. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9783527665143.ch12/summary

Rodolà, E., Cosmo, L., Bronstein, M. M., Torsello, A., & Cremers, D. (2016). Partial functional correspondence: partial functional correspondence. *Computer Graphics Forum*, 222–236. https://doi.org/10.1111/cgf.12797

Rodolà, Emanuele, Bulo, S., Windheuser, T., Vestner, M., & Cremers, D. (2014). Dense non-rigid shape correspondence using random forests. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on* (pp. 4177–4184). IEEE. https://doi.org/10.1109/CVPR.2014.532

Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, *50*(5), 742–754. https://doi.org/10.1021/ci100050t

Rush, T. S., Grant, J. A., Mosyak, L., & Nicholls, A. (2005). A shape-based 3-D scaffold hopping method and its application to a bacterial protein−protein interaction. *Journal of Medicinal Chemistry*, *48*(5), 1489–1495. https://doi.org/10.1021/jm040163o

Rustamov, R. M., Ovsjanikov, M., Azencot, O., Ben-Chen, M., Chazal, F., & Guibas, L. (2013). Map-based exploration of intrinsic shape differences and variability. *ACM Trans. Graph.*, *32*(4), 72:1–72:12. https://doi.org/10.1145/2461912.2461959

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220

Sastry, G. M., Dixon, S. L., & Sherman, W. (2011). Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring.

Journal of Chemical Information and Modeling, *51*(10), 2455–2466. https://doi.org/10.1021/ci2002704

Schreyer, A., & Blundell, T. (2009). CREDO: a protein-ligand interaction database for drug discovery. *Chemical Biology & Drug Design*, *73*(2), 157–167. https://doi.org/10.1111/j.1747-0285.2008.00762.x

Schreyer, A., & Blundell, T. (2012). USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *Journal of Cheminformatics*, *4*(1), 27. https://doi.org/10.1186/1758-2946-4-27

Sculley, D. (2010). Web-scale k-means clustering (p. 1177). ACM Press. https://doi.org/10.1145/1772690.1772862

Shape-it^{TM}. (n.d.). (Version 1.0.1). Silicos-it. Retrieved from http://silicos-it.com/software/shape-it/1.0.1/shape-it.html

Sharma, A., & Horaud, R. (2010). Shape matching based on diffusion embedding and on mutual isometric consistency. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 29–36). https://doi.org/10.1109/CVPRW.2010.5543278

Shave, S. (2010). *Development of high performance structure and ligand based virtual screening techniques*. The University of Edinburgh. Retrieved from http://hdl.handle.net/1842/4333

Shave, S., Blackburn, E. A., Adie, J., Houston, D. R., Auer, M., Webster, S. P., … Walkinshaw, M. D. (2015). UFSRAT: Ultra-Fast Shape Recognition with Atom Types –the discovery of novel bioactive small molecular scaffolds for FKBP12 and 11βHSD1. *PLOS ONE*, *10*(2), e0116570. https://doi.org/10.1371/journal.pone.0116570

Sheng, C., & Zhang, W. (2013). Fragment informatics and computational fragment-based drug design: an overview and update. *Medicinal Research Reviews*, *33*(3), 554–598. https://doi.org/10.1002/med.21255

Sheridan, R. P., Singh, S. B., Fluder, E. M., & Kearsley, S. K. (2001). Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *Journal of Chemical Information and Computer Sciences*, *41*(5), 1395–1406. https://doi.org/10.1021/ci0100144

Shoichet, B. K. (2004). Virtual screening of chemical libraries. *Nature*, *432*(7019), 862–865. https://doi.org/10.1038/nature03197

Skraba, P., Ovsjanikov, M., Chazal, F., & Guibas, L. (2010). Persistence-based segmentation of deformable shapes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 45–52). https://doi.org/10.1109/CVPRW.2010.5543285

Sommer, K., Friedrich, N.-O., Bietz, S., Hilbig, M., Inhester, T., & Rarey, M. (2016). UNICON: A powerful and easy-to-use compound library converter. *Journal of Chemical Information and Modeling*, *56*(6), 1105–1111. https://doi.org/10.1021/acs.jcim.6b00069

Sun, J., Ovsjanikov, M., & Guibas, L. (2009). A concise and provably informative multi-scale signature based on heat diffusion. *Computer Graphics Forum*, *28*(5), 1383–1392. https://doi.org/10.1111/j.1467-8659.2009.01515.x

Swamidass, S. J., Azencott, C.-A., Daily, K., & Baldi, P. (2010). A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics*, *26*(10), 1348–1356. https://doi.org/10.1093/bioinformatics/btq140

Taylor, P., Blackburn, E., Sheng, Y. G., Harding, S., Hsin, K.-Y., Kan, D., … Walkinshaw, M. D. (2009). Ligand discovery and virtual screening using the program LIDAEUS: Ligand discovery and virtual screening. *British Journal of Pharmacology*, *153*(S1), S55–S67. https://doi.org/10.1038/sj.bjp.0707532

The Cambridge Crystallographic Data Centre (CCDC). (n.d.). Retrieved 5 September 2017, from https://www.ccdc.cam.ac.uk/

Thornber, C. W. (1979). Isosterism and molecular modification in drug design. *Chemical Society Reviews*, *8*(4), 563–580. https://doi.org/10.1039/CS9790800563

Tresadern, G., Bemporad, D., & Howe, T. (2009). A comparison of ligand based virtual screening methods and application to corticotropin releasing factor 1 receptor. *Journal of Molecular Graphics and Modelling*, *27*(8), 860–870. https://doi.org/10.1016/j.jmgm.2009.01.003

Truchon, J.-F., & Bayly, C. I. (2007). Evaluating virtual screening methods: good and bad metrics for the ‘early recognition’ problem. *Journal of Chemical Information and Modeling*, *47*(2), 488–508. https://doi.org/10.1021/ci600426e

Tuzel, O., Porikli, F., & Meer, P. (2006). Region covariance: a fast descriptor for detection and classification. In A. Leonardis, H. Bischof, & A. Pinz (Eds.), *Computer Vision – ECCV 2006* (Vol. 3952, pp. 589–600). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://link.springer.com/10.1007/11744047_45

Ujváry, I. (1997). Extended Summary: BIOSTER—a database of structurally analogous compounds. *Pesticide Science*, *51*(1), 92–95. https://doi.org/10.1002/(SICI)1096-9063(199709)51:1<92::AID-PS608>3.0.CO;2-9

Ujváry, I., & Hayward, J. (2012). Bioster: a database of bioisosteres and bioanalogues. In N. Brown (Ed.), *Bioisosteres in Medicinal Chemistry* (pp. 53–74). Wiley-VCH Verlag GmbH & Co. KGaA. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9783527654307.ch4/summary

Vainio, M. J., Puranen, J. S., & Johnson, M. S. (2009). ShaEP: molecular overlay based on shape and electrostatic potential. *Journal of Chemical Information and Modeling*, *49*(2), 492–502. https://doi.org/10.1021/ci800315d

Venkatraman, V., Chakravarthy, P., & Kihara, D. (2009). Application of 3D Zernike descriptors to shape-based ligand similarity searching. *Journal of Cheminformatics*, *1*(1), 19. https://doi.org/10.1186/1758-2946-1-19

Venkatraman, V., Pérez-Nueno, V. I., Mavridis, L., & Ritchie, D. W. (2010). Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods. *Journal of Chemical Information and Modeling*, *50*(12), 2079–2093. https://doi.org/10.1021/ci100263p

Venkatraman, V., Sael, L., & Kihara, D. (2009). Potential for protein surface shape analysis using spherical harmonics and 3D Zernike descriptors. *Cell Biochemistry and Biophysics*, *54*(1–3), 23–32. https://doi.org/10.1007/s12013-009-9051-x

Verdonk, M. L., Berdini, V., Hartshorn, M. J., Mooij, W. T. M., Murray, C. W., Taylor, R. D., & Watson, P. (2004). Virtual screening using protein−ligand docking: avoiding artificial enrichment. *Journal of Chemical Information and Computer Sciences*, *44*(3), 793–806. https://doi.org/10.1021/ci034289q

von Behren, M. M., Bietz, S., Nittinger, E., & Rarey, M. (2016). mRAISE: an alternative algorithmic approach to ligand-based virtual screening. *Journal of Computer-Aided Molecular Design*, *30*(8), 583–594. https://doi.org/10.1007/s10822-016-9940-1

von Behren, M. M., & Rarey, M. (2017). Ligand-based virtual screening under partial shape constraints. *Journal of Computer-Aided Molecular Design*, *31*(4), 335–347. https://doi.org/10.1007/s10822-017-0011-z

Wang, L., Si, P., Sheng, Y., Chen, Y., Wan, P., Shen, X., … Li, W. (2015). Discovery of new non-steroidal farnesoid X receptor modulators through 3D shape similarity search and structure-based virtual screening. *Chemical Biology & Drug Design*, *85*(4), 481–487. https://doi.org/10.1111/cbdd.12432

Wang, Q., Birod, K., Angioni, C., Grösch, S., Geppert, T., Schneider, P., … Schneider, G. (2011). Spherical harmonics coefficients for ligand-based virtual screening of Cyclooxygenase inhibitors. *PLoS ONE*, *6*(7), e21554. https://doi.org/10.1371/journal.pone.0021554

Wei, N.-N., & Hamza, A. (2014). SABRE: ligand/structure-based virtual screening approach using consensus molecular-shape pattern recognition. *Journal of Chemical Information and Modeling*, *54*(1), 338–346. https://doi.org/10.1021/ci4005496

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, *28*(1), 31–36. https://doi.org/10.1021/ci00057a005

Willett, P. (2003). A history of chemoinformatics. In J. Gasteiger (Ed.), *Handbook of Chemoinformatics* (pp. 6–20). Wiley-VCH Verlag GmbH. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9783527618279.ch2/summary

Willett, P. (2008). From chemical documentation to chemoinformatics: 50 years of chemical information science. *Journal of Information Science*, *34*(4), 477–499. https://doi.org/10.1177/0165551507084631

Willett, P., Barnard, J. M., & Downs, G. M. (1998). Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, *38*(6), 983–996. https://doi.org/10.1021/ci9800211

Windheuser, T., Vestner, M., Rodolà, E., Triebel, R., & Cremers, D. (2014). Optimal intrinsic descriptors for non-rigid shape analysis (p. 44.1-44.11). British Machine Vision Association. https://doi.org/10.5244/C.28.44

Wirth, M., Zoete, V., Michielin, O., & Sauer, W. H. B. (2012). SwissBioisostere: a database of molecular replacements for ligand design. *Nucleic Acids Research*, gks1059. https://doi.org/10.1093/nar/gks1059

Wolber, G., Dornhofer, A. A., & Langer, T. (2006). Efficient overlay of small organic molecules using 3D pharmacophores. *Journal of Computer-Aided Molecular Design*, *20*(12), 773–788. https://doi.org/10.1007/s10822-006-9078-7

Yan, X., Li, J., Gu, Q., & Xu, J. (2014). gWEGA: GPU-accelerated WEGA for molecular superposition and shape comparison. *Journal of Computational Chemistry*, *35*(15), 1122–1130. https://doi.org/10.1002/jcc.23603

Yan, X., Li, J., Liu, Z., Zheng, M., Ge, H., & Xu, J. (2013). Enhancing molecular shape comparison by weighted Gaussian functions. *Journal of Chemical Information and Modeling*, *53*(8), 1967–1978. https://doi.org/10.1021/ci300601q

Zhao, W., Hevener, K. E., White, S. W., Lee, R. E., & Boyett, J. M. (2009). A statistical framework to evaluate virtual screening. *BMC Bioinformatics*, *10*(1), 225. https://doi.org/10.1186/1471-2105-10-225