Assessing and developing methods to explore the role of molecular shape in computer-aided drug design

Joanna Maria Zarnecka

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Doctor of Philosophy

This research programme was carried out in collaboration with MedChemica Ltd.

February 2018

My big thesis is that although the world looks messy and chaotic, if you translate it into the world of numbers and shapes, patterns emerge and you start to understand why things are the way they are.

Marcus du Sautoy

Acknowledgments

Firstly, I would like to express my sincere gratitude to my advisor Dr Andrew G. Leach for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr Steven J. Enoch and Prof. Mark T.D. Cronin, for their insightful comments and encouragement.

My sincere thanks also go to Dr Al Dossetter and Dr Ed Griffen from MedChemica for the valuable discussion and all the support throughout the project.

I thank my colleagues: David Ebbrell, Maria Sapounidou, Antonio Cassano, Julia Pletz and the rest of the QSAR lab the stimulating discussions and for all the fun we have had in the last three years. In particular, I am grateful to Dr Iva Lukac and Dr Claire Mellor for support and for a really enjoyable time spent together.

I would like to thank Stephen Messham for providing dataset for solubility predictions.

A very special gratitude goes out to MedChemica Ltd. And Liverpool John Moores University for helping and providing the funding for PhD research.

Last but not the least, I would like to thank my family: my parents and my brother for supporting me spiritually throughout writing this thesis and my life in general.

Table of Contents

Abstract

Shape-based approaches have many potential areas for development in the future for application to *in silico* pharmacology. Further exploration of the role of molecular shape may lead to better understanding of the substrate specificity of enzymes and the possibility to reduce toxic effects that may be caused by ligands binding to undesired target proteins. Methods exploiting molecular shape for activity and toxicity prediction might have a great influence on the drug discovery process.

There are different approaches that might be used for this purpose, e.g. shape fingerprints and shape multipoles. Both methods describe the shape of molecules, discarding any chemical information, using numerical values. Focusing only on shape can lead to identifying novel core structures of molecules, with improved properties.
Molecular fingerprints are binary bit strings that encode the structure or shape of compounds; shape is measured indirectly by alignment to a database of standard molecular shapes – the reference shapes. The Shape Database should represent a wide range of possible molecular shapes to produce accurate results. Therefore, this was the main focus of the investigation.

The shape multipoles method is a fast computational method to describe the shape of molecules by using only numbers and therefore it requires low storage needs and comparison is performed by simple mathematical operations. To describe the shape, it uses only 13 values (3 quadrupole components and 10 octupole components).

The performances of both methods in grouping compounds based on shared biological activity were evaluated using several test sets with slightly better results in case of shape fingerprints. However, the shape multipole approach showed potential in finding differences in shape between enantiomers. Among the possible applications of the shape fingerprints method are solubility prediction (on comparable level as well-established methods) and virtual screening.

List of abbreviations

AChE	Acetylcholinesterase
AHR	Aryl Hydrocarbon Receptor
AKT1	Serine/Threonine-protein Kinase AKT
AMPC	AmpC Beta-lactamase
AUC	Area Under Curve
AV method	Average Value method
BOV	Bit On Value, threshold value used in shape fingerprints generation process
Cav 3.2	Voltage-gated calcium channel subunit alpha Cav3.2
CoMFA	Comparative Molecular Field Analysis
CoMSIA	Comparative Molecular Similarity Index Analysis
CP3A4	Cytochrome P450 3A4
CXCR4	C-X-C Chemokine Receptor type 4
D1	Dopamine D1 receptor
DBD	DNA Binding Domain
DPP-IV	Dipeptidyl peptidase IV
DT	Design Tanimoto, threshold value used in Shape Database generation process
EGFR	Epidermal Growth Factor Receptor
ER	Estrogen Receptor
ERE	Estrogen Response Element
F+I	Fragment and Index
FT	Similarity (Fingerprint) Tanimoto, result of comparing two bit strings
FXR	Farnesoid X Receptor
GCR, GR	Glucocorticoid Receptor
GSE	General Solubility Equation
hERG	Human ether-a-go-go-related gene potassium channel 1

HIVPR	Human Immunodeficiency Virus type 1 Protease
HIVTR	Human Immunodeficiency Virus type 1 Reverse Transcriptase
HRE	Hormone Response Element
KIF11	Kinesin-like protein 1
LBD	Ligand Binding Domain
LXR	Liver X Receptor
LXRE	Liver X Response Element
MCSS	Maximum Common Substructure
MMP	Matched Molecular Pair
MP	Melting Point
MV method	Maximum Value method
NR	Nuclear Receptor
PPAR	Peroxisome Proliferator-Activated Receptor
PR	Progesterone Receptor
PXR	Pregnane X Receptor
QSAR	Quantitative Structure–Activity Relationship
RAR	Retinoic Acid Receptor
ROC	Receiver Operating Characteristic
ROCS	Rapid Overlay of Chemical Structures
RXR	Retinoid X Receptor
SD	Shape Database
ST	Shape Tanimoto, result of overlaying two molecules
THR	Thyroid Hormone Receptor
TR	Thyroid Hormone Receptor
TRE	Thyroid Response Element
VDR	Vitamin D Receptor

Chapter 1

Theoretical Background

1.1.Introduction

By definition, shape is the form of an object or its external boundary, outline, or external surface. Shape is used by everyone on a daily basis as a way of identifying and organizing visual information. Shapes are also the very first thing children learn in early childhood – it is because shape is one of the very noticeable attributes of the world around us.[1] Before learning the name of the object, we use the description of it – its shape and colour, which helps us to visualize things and to help others understand what object we are describing.[1] Although shapes like triangles, squares or circles that are present in the macroworld are not difficult to describe, it is not easy to define more complex shapes, especially those that build the microworld – the shape of molecules.

1.2.Importance of shape in chemistry and biology

There are many different biological processes in all organisms that distinguish molecules based on their shape. The size and spatial features of molecules play a crucial role in activation of G-protein-coupled receptors (which are responsible for regulation of diverse cell functions), opening ligand-gated ion channels (e.g.

participating in neurotransmission), antibody recognition and activation of a variety of enzymes (**Figure 1-1**).[2,3,4]

Figure 1-1. Lock and key model – antibodies memorize pathogens mostly based on their shape. [2]

This ability of proteins to bind ligands with specific shapes was first explained by Emil Fischer[5] in 1894 and later further explored by Linus Pauling[6] during his scientific work. In the proposed "lock-and-key" model a substrate is treated as a key that needs to fit perfectly, by the means of size and shape complementarity, to a conceptual lock, which is the binding site of a target protein (**Figure 1-2**). Many ligand-protein complexes were analysed in the past to examine this theory, only to come to the conclusion that even though the ligand adapts multiple conformations, which are difficult to predict, the lock and key concept is still the working model for designing new compounds in many areas of study.[3] The theory was broadly studied especially for its usage in rational drug design.[7,8,9] A lot of computational methods like molecular docking and pharmacophore modelling were developed based on the hypothesis that a ligand needs to match the active site of the target protein in terms of shape and more besides.

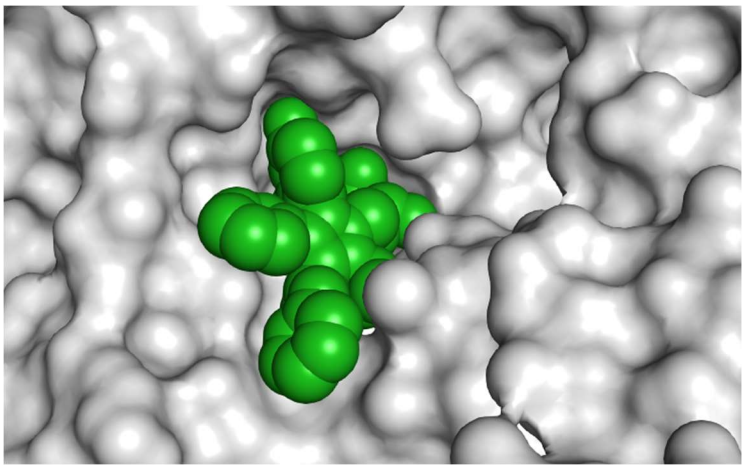


Figure 1-2. Complex of the human HMG-CoA reductase with Atorvastatin as an example of the lock and key model (pdb code: 1HWK).[10]

Nowadays, there are two main approaches in similarity searching: based on chemical or shape similarity. Relying on the chemical structure or shape of known drugs to screen databases in order to find the compounds that might have a desired potency has a background in bioisosterism.[11,12,13,14] The concept was introduced by Friedman[15] and implies that molecules that are similar in size and shape are more likely to show similar activity towards the same target macromolecule.[16] More common are simple isosteric replacements, which involves the substitution of a group of atoms in one compound by others with similar shape or chemistry,[14] e.g. changing OH group with NH_2. Bioisosterism concerns more drastic changes, more important or bigger groups of the ligand, like replacement of the core of the molecule (**Figure 1-3**).[11,14]

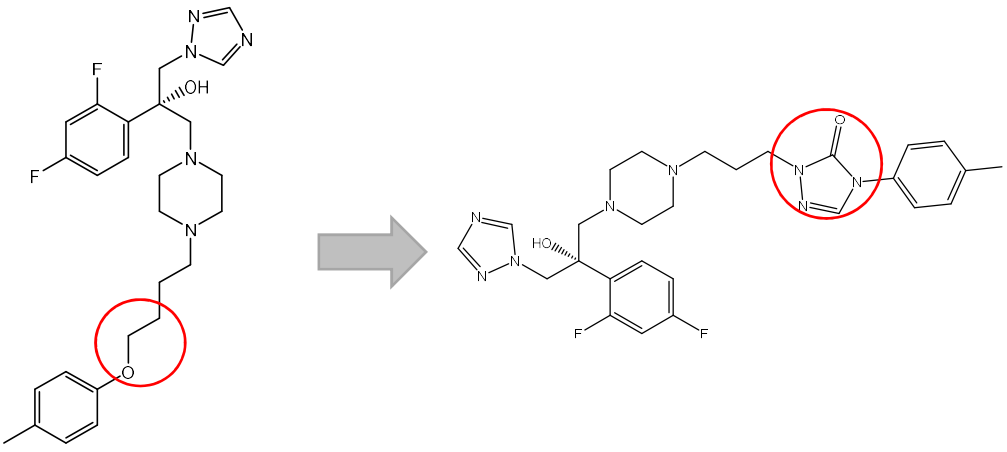


Figure 1-3. Example of bioisostere replacement: oxygen atom is being replaced by triazolone.[17]

The pharmaceutical industry uses this concept to improve potential drugs, which, although they show a desired activity, cause side effects. It may also lead to improvement of potency and changes in physicochemical properties, such as solubility.[14] Thus, a principle motive behind many virtual screening studies is to find compounds similar in shape and size, meaning that they show similar activity, but different enough to have physicochemical properties that may improve drug action and reduce toxicity.[16,18] It can be also applied to find analogues of a drug that cannot be introduced to the market because of intellectual property rights.

1.3.Structure-based vs. ligand-based methods

Implementation of Fischer's theory can be found in many tools used by scientists, mostly focusing on receptor-based approaches.[7] Such methods, e.g. molecular docking,[19] can predict the interactions that occur between the potential drug and its receptor. Molecular docking, since its introduction in 1982,[20] has been proved to be a very successful technique.[19,21,22] Its usefulness in predicting the ligand poses in the binding pocket as well as its ability to estimate the binding energy based on the formed non-covalent interactions, made it the leading technique in drug discovery (**Figure 1-4**).[23,24]

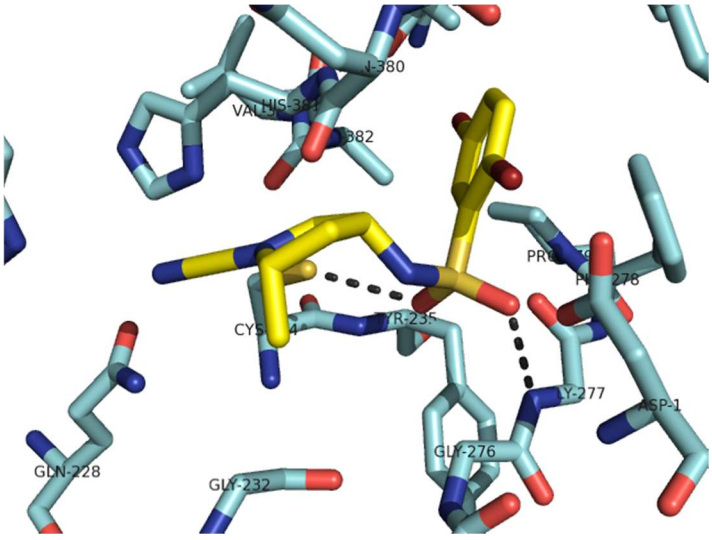


Figure 1-4. 2,5-dibromo-N-{(3R,5S)-1[(Z)-iminomethyl]-5-methylpyrrolidyn-3-yl}benzenesulfonamide docked, using AutoDock, in the active site of cathepsin C (pdb code: 3PDF) with shown interactions.

However, all receptor-based techniques need sufficient data about the structure of the receptor or at least about its binding site.[25] This might be challenging, because current protein structure databases may not provide the necessary structures or structures of a good enough quality and this way it may lead to inaccurate calculations.[26] Additionally, ligand-based approaches are usually much faster than docking methods. Those are major reasons that ligand-based methods (which do not require information about the receptor's structure) can be useful in the early stages of the drug discovery process, especially in lead generation and lead optimization.[27] Screening thousands or tens of thousands of compounds in order to discover a new drug, similar to the known ligand or lead compound, that would bind to the target better or cause less side effects is commonly used in the pharmaceutical industry.[28,29,30]

1.4. 2D methods

When structural data of the receptor is unknown or is of poor quality, the ligand-based drug design is preferred.[26,31] Ligand-based approaches use the information of one or more known ligands of the protein.[32,33] The underlying hypothesis is that molecules with similar features to the already known actives are more likely to have comparable biological activity.[34] Among the most common approaches, there are methods based on chemical similarity and QSAR (quantitative structure–activity relationship) models, [26,31] which predict the activity/properties of molecules based on their chemical structure.[35,36]

Most ligand-based techniques assign descriptors to the molecular structures to later compare and rank the molecules based on the similarity scores. The ligand-based methods used in medicinal chemistry relate to 2D structures of molecules, because of the simplicity and speed of such approaches. The similar molecules that show neighbourhood behaviour have similar representations as SMILES strings[29,37] (Simplified Molecular Input Line Entry Specification)[38] or similar fingerprints. [39]

1.4.1. 2D fingerprints

There are ligand-based methods that encode the 2D structure of a molecule or its features using bit string representations. Such an approach is termed as molecular 2D fingerprints and these are one of the most commonly used methods in drug discovery nowadays.[28,39,40,41,42] Their focus is on chemical similarity searching, therefore that is the information encoded in their bit strings. There are several types of molecular fingerprints with different fragmentation strategies (e.g. linear, radial, dendritic), atom types (e.g. generic, where all atoms and bonds are equivalent, functional, where atoms are distinguished by functional type or atoms recognized as hydrogen bond acceptors or donors), bit scaling rules (e.g. scaling by feature size to molecule size) and similarity indices.[41,43]

Among the commonly used 2D fingerprints are MACCS,[44] LINGO,[45,46] Circular,[47] Path and Tree.[40] The summary of them can be found in **Table 1-1**. Some of the fingerprints generation processes do not assign a specific feature to a bit position and use the hash function to map the information of molecular structure into a fixed size bit string.[43] In such fingerprints it is therefore common to observe bit collisions, where the bit is set on for more than one fragments/feature.[43] These are topological fingerprints, which are often called hashed fingerprints.[43]

Table 1-1. 2D fingerprint methods and descriptions.

FINGERPRINT METHOD	DESCRIPTION
MACCS	166 or 960 bit structural key descriptors based on SMARTS patterns
CIRCULAR	enumerated circular fragments hashed into a fixed-length bit string
LINGO	text-based molecular similarity search method based on fragmentation of canonical isomeric SMILES strings
TREE	enumerated tree fragments hashed into a fixed-length bit string
PATH	enumerated linear fragments hashed into a fixed-length bit string

1.4.1.1. MACCS

MACCS keys, also known as MDL 2D keysets, are structural keys based on SMARTS patterns.[44] This type of fingerprint encodes the presence or absence of a given substructure in a molecule, as shown in the **Figure 1-5**.

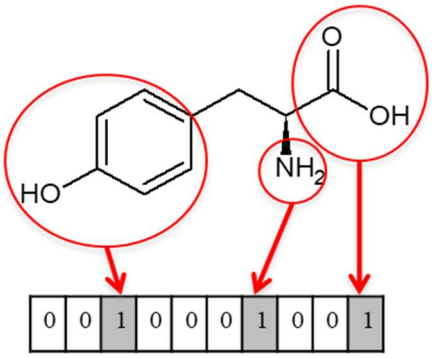


Figure 1-5. Example of the hypothetical 10-bit substructure.[40]

There are two possible lengths of a MACCS fingerprint – 960 or 166 bit.[40,44] The most used is the shorter version, which covers most of the interesting chemical features for drug discovery and virtual screening and it is available in many software packages.[40]

1.4.1.2. Path and Tree fingerprints

Path-based and Tree fingerprints are both topological fingerprints, in which the fragments are analysed either in a tree or a linear way up to a given size (a certain number of bonds) and then all those paths are hashed into a fingerprint.[40] Unlike in MACSS fingerprints, the particular bits cannot be traced back to certain chemical features as some of them might encode more than one feature. These fingerprints are suitable for substructure queries.[40]

The image originally presented here cannot be made freely available via LJMU E-Theses Collection because of copyright restrictions.

The image was sourced at

Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58–63.

Figure 1-6. Example of a hypothetical 10-bit path-based fingerprint.[40]

1.4.1.3. Circular fingerprints

Circular fingerprints are also topological fingerprints.[47] However, as the name suggests, the radius of each heavy atom of a molecule is analysed. ECFPs (Extended-connectivity fingerprints) are circular fingerprints based on the Morgan algorithm.[40] This type of fingerprint cannot be used for substructure queries and are rather used for similarity searches of the whole structure.[40]

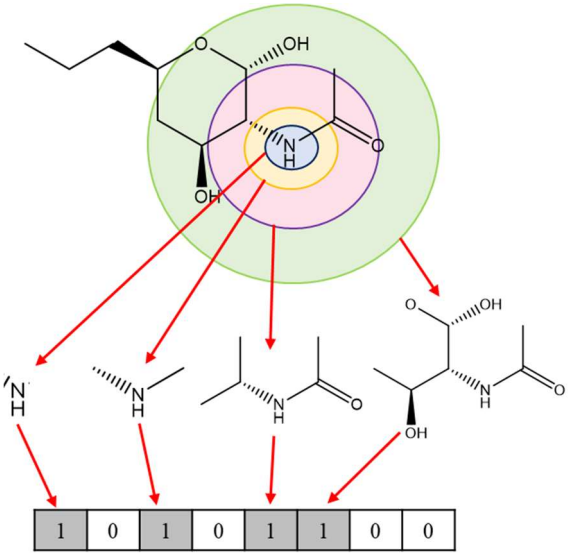


Figure 1-7. Example of hypothetical 8-bit circular fingerprint.

1.4.1.4. LINGO

LINGO fingerprints are text strings[45] instead of bit strings like other fingerprints types. Their string includes different letters, numbers and symbols, which are obtained by decomposing a SMILES string.[45,46] As described by Vidal et al.:[45] *"a q-LINGO is a q-character string, including letters, numbers and symbols, such as "(", ")","[", "]", "#", etc. obtained by stepwise fragmentation of a canonical SMILES molecular representation*". The SMILES string of length n would result in (n-(q-1)) substrings of length q. As shown in the **Figure 1-8**, where q=4, the chlorpromazine is decomposed into 25 substrings, with three of them having two occurrences, which is 28 in total (31-(4-1) = 28).

The image originally presented here cannot be made freely available via LJMU E-Theses Collection because of copyright restrictions. The image was sourced at

Vidal, D.; Thormann, M.; Pons, M. LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarities. *J. Chem. Inf. Model.* **2005**, *45* (2), 386–393.

Figure 1-8. LINGO generation process.[45]

1.4.2. Similarity measures

When comparing two molecular fingerprints, the presence or absence of a structural fragment is represented by setting a bit On or Off. Therefore, similar molecules are the ones with a higher number of bits in common, which indicates possession of similar features or functional groups.

The comparison is performed by alignment of two molecules – the query and the one from the database, or more accurately their bit strings, and counting the number of bits set on in only one of the strings and those in common. The results are scored using one of many similarity coefficients. The most common metrics are the Tanimoto,

Tversky, Cosine, Euclidean and Dice.[40,48] The summary of the popular similarity measures can be seen below in **Table 1-2**.

Table 1-2. Most popular similarity measures.

TYPE	FORMULA	RANGE
TANIMOTO	$\frac{bothAB}{onlyA + onlyB + bothAB}$	[0.0 – 1.0]
TVERSKY	$\frac{bothAB}{\alpha * onlyA + \beta * onlyB + bothAB}$	[0.0 – N]
COSINE	$\frac{bothAB}{\sqrt{(onlyA + bothAB) * (onlyB + bothAB)}}$	[0.0 – 1.0]
EUCLIDEAN	$\sqrt{\frac{bothAB + neitherAB}{onlyA + onlyB + bothAB + neitherAB}}$	[0.0 – 1.0]
DICE	$\frac{2 * bothAB}{onlyA + onlyB + 2 * bothAB}$	[0.0 – 1.0]

In the table above (**Table 1-2**), bothAB represents the number of common features for the molecules A and B, neitherAB is the number of bits set off in common for molecules A and B, while onlyA and onlyB are the numbers of the features present only in the molecule A and B, respectively.

1.4.3. Matched Molecular Pairs

Another, widely used ligand-based 2D approach is Matched Molecular Pair Analysis (MMPA). The key principle of MMPA is that the difference in properties is more easily and accurately predicted than the absolute value of properties.[49,50] The concept was first introduced by Kenny and Sadowski in 2005.[51] The method is used to screen large databases in order to find pairs of molecules with a common structural part with a promising change in properties.[49,51,52] A Matched Molecular Pair (MMP) involves two compounds that have a common core and have a different fragment R as shown in **Figure 1-9**.

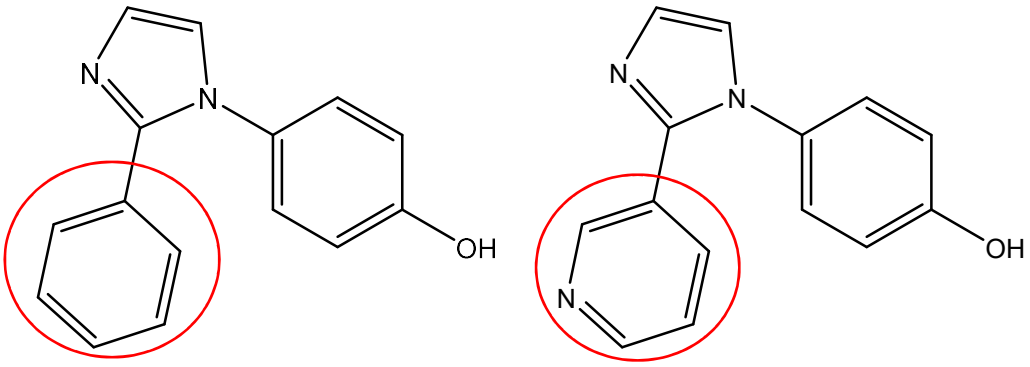


Figure 1-9. An example of pair of MMP with fragment part marked in red circle.

There are two common approaches to finding MMPs: Fragment and Index (F+I)[53] and Maximum Common Substructure (MCSS).[54] In the MCSS approach, as used in the WizePairZ algorithm described by Warner et al.,[54] two molecules are compared and the maximum common substructure shared by them is identified – the fixed part. The remaining part is called the changing part. In order to compare the molecules using this approach, the molecules are converted into graphs, which enables identifying common substructures between compounds.

The outcome of the MCSS approach is governed by the choices made about deciding what to classify as a pair. Warner et al.[54] elected to use a fast graph comparison technique that does not permit disconnected substructures to be generated and chose to limit the changing part of the molecule to be less than 10% of the fixed part.[49] They encoded the output from their pair finding as SMIRKS (reaction transform language)[55] that permits any structural changes identified as being of interest to be able to be applied to molecules to which they may be relevant.[54]

The key limitations of the MCSS approach are that the substructure comparison is slow and that most algorithms for finding the MCSS require all of the atoms to be contiguous and therefore prevent pairs in which linkers change from being found.[49]

The second approach was introduced by Hussain and Rea.[53] The algorithm works by generating fragments of the molecule based on predefined rules and then indexing those fragments. The generated fragments are stored as key – value pairs.[53,56]

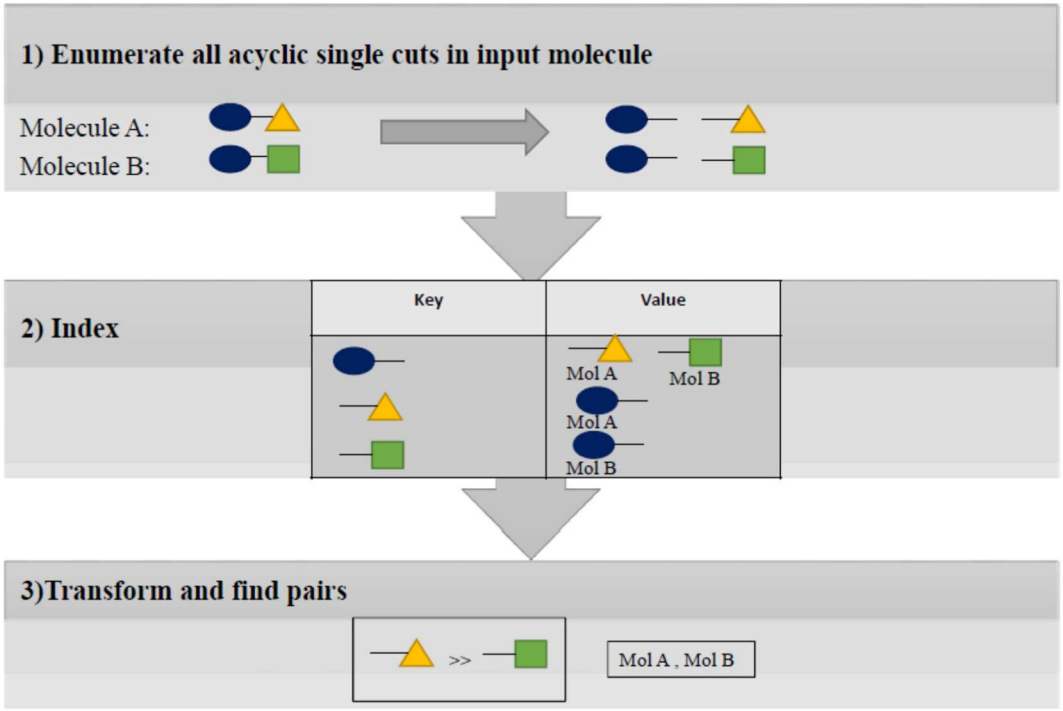


Figure 1-10. Scheme of identifying pairs in the Fragment and Index approach when single cuts are performed.

In the first step, each molecule is fragmented, by breaking selected bonds (**Figure 1-10**). Hussain and Rea[53] achieve this by defining the bonds to be broken using a SMARTS pattern. They aim to have this pattern be specific to acyclic single bonds. Bonds are broken one at a time, and the resulting fragments are then stored as SMILES strings, which can be manipulated as text.[49,53] This is a great advantage of the approach: after initial fragmentation, all subsequent steps toward the identification of matched pairs involve only rapid text processing.[49] Also, once a molecule has been fragmented and added to the database, it is available to any new molecule that is added.

Among the limitations of the fragment and index approach to finding matched pairs are that small changes to rings cannot be identified as pairs, highly substituted core changes are limited by the number of fragmentations considered, and the diversity of the structural changes can be limited by the restrictions that are imposed (usually heavy atom counts or ratio).[49] One specific set of changes that are not readily identified is modifications to macrocyclic rings.

Table 1-3. Advantages and disadvantages of MCSS and F+I approaches.

METHOD	ADVANTAGES	DISADVANTAGES
MCSS	• small changes to rings can be identified	• slow • most algorithms for finding the MCSS require all of the atoms to be contiguous
F+I	• fast • once a molecule has been fragmented and added to the database, it is available to any new molecule that is added	• restrictions in bond breaking - small changes to rings cannot be identified • can lead to fragmentations and grouping into sets of pairs that chemists would not normally consider to be chemically sensible

1.5. Scaffold hopping – chemistry vs. shape

There are many approaches that rely on the chemistry of the molecule, discarding the information about the three-dimensional shape of a molecule. Those approaches rarely lead to compounds that are chemically very distinct and thus they are not well-suited for scaffold hopping.[57,58] Knowing that molecules similar in shape could bind to the same protein even with completely different chemistry, opens new possibilities to find novel, sometimes unexpected compounds.

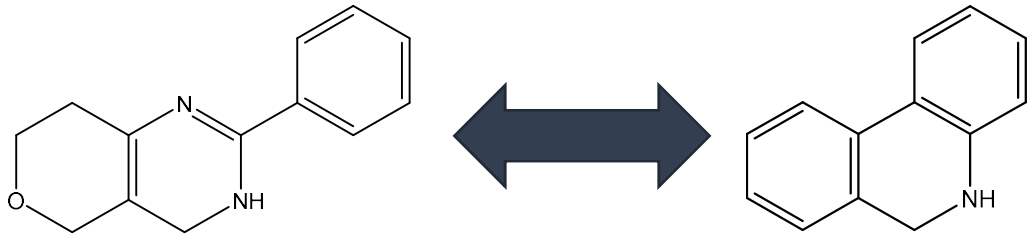


Figure 1-11. Example of scaffold hopping.

Therefore, scaffold hopping gains a lot of attention in medicinal chemistry. The concept was introduced in 1999 by Schneider et al.[59] It is a technique that identifies compounds with different chemistry or central core but with similar shape or electrostatic surface and thus leading to comparable or improved activity.[57,58,60,61] The scaffold hopping can lead to drastic changes in molecular properties, e.g. changes in solubility by replacing a lipophilic structure with a more polar one, changes in the stability of a compound, reduction of toxicity, or improvements in DMPK (drug metabolism and pharmacokinetics).[57]

There are many successful compounds identified using scaffold hopping. One of the examples can be seen in the **Figure 1-12**. The structure of morphine and tramadol is very different but they share positions of the tertiary amine, the aromatic ring, the hydroxyl group and also have some similarities in the overall shape.[61,62]

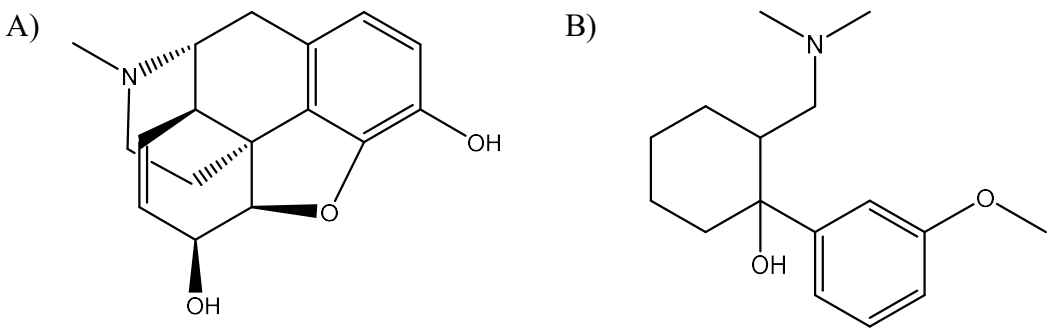


Figure 1-12. Structures of pain killing drugs: Morphine (A) and Tramadol (B).

Molecules with different core structures but similar activity are of high interest in medicinal chemistry mostly due to a desire to improve potency or reduce toxicity.[16] It can be also applied to find analogues that are novel and patentable.

1.6. 3D methods

More and more methods are used for describing molecular shape. Among them, the most common and simple are pharmacophores.[63] Pharmacophores are defined as arrangement of atoms or the features of molecule that are essential in its biological activity.[58,63] It uses the information of hydrogen donors or acceptors, acidic/basic

groups and hydrophobic features in the molecule and their position in space.[63] However, that approach may not accurately indicate the shape because of treating the shape of molecule just as sets of atoms in space and not focusing on its volume and surface (**Figure 1-13**).

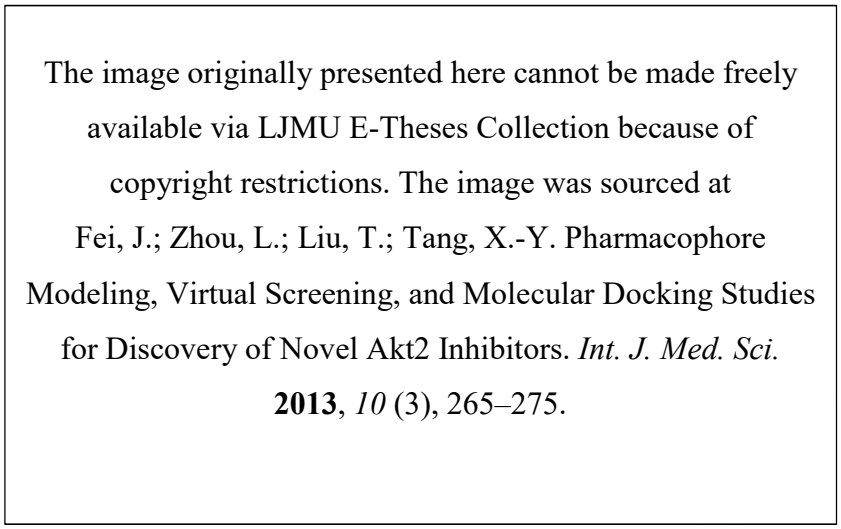


Figure 1-13. The example of a pharmacophore model with three pharmacophoric features: one hydrogen bond acceptor (green), one hydrophobic feature (blue), and an aromatic ring (orange).[64]

CoMFA (Comparative Molecular Field Analysis)[65] is a 3D-QSAR method that links the shape-dependant properties of molecules to their biological activity.[32] The molecules binding to the same receptor and in the same way are selected to develop the models.[32,66,67] Then they are aligned based on their shape and their molecular fields are mapped to the 3D grid.[32,67] Field values in each grid point are calculated corresponding to the potential energy, which is then correlated to biological activity.[32,67] Similarly, CoMSIA (Comparative Molecular Similarity Indices)[66,68] is also a 3D-QSAR model, but includes the hydrophobic, hydrogen-bond donor and acceptor together with steric features.[66]

The other methods often used are shape fingerprints,[69] Gaussian-based methods[70] and moment-based method (e.g. the shape multipole method).[71] These methods do not require chemical information and depend only on the spatial distribution of shape. Therefore, they might be more appropriate for scaffold hopping approaches.

1.6.1. Gaussian-based methods – ROCS

A Gaussian description of molecular shape is implemented in ROCS,[72] which stands for Rapid Overlay of Chemical Structures, is an Openeye's software. Two or more molecules are compared according to their volume overlap. The method, which describes molecules in terms of atom-centred Gaussian functions, performs optimization by rigid translation and rotation of one of the matched molecules with respect to the other.[70] The difference between two shapes is returned as a Tanimoto or Tversky coefficient ranked from the highest to the lowest values.

The ROCS package[72] allows for addition of chemical feature information that can be used together with volume overlap to compare molecules. For these, various scores might be used e.g. TanimotoCombo, TverskyCombo, which includes the chemical information as well as shape and compares the compounds in different ways. Additionally, multi-conformer molecules can be used as both query and database molecules. However, applying these features requires more computational time.

1.6.2. Shape fingerprints

Fingerprint methods are not only used to describe the chemical connectivity of a molecule, e.g. LINGO or Tree fingerprints, but may also be applied to encode the shape of the molecule – via so called shape fingerprints.[69] Haigh et al.[69] introduced the concept of shape fingerprints and established the parameters that can be varied to tune the fingerprint. The shape of a molecule is measured indirectly by alignment to a database of diverse reference shapes, as shown in the example in **Figure 1-14**. Therefore, to produce accurate results, the database with reference shapes is supposed to represent a wide range of possible shapes of molecules to produce accurate results.

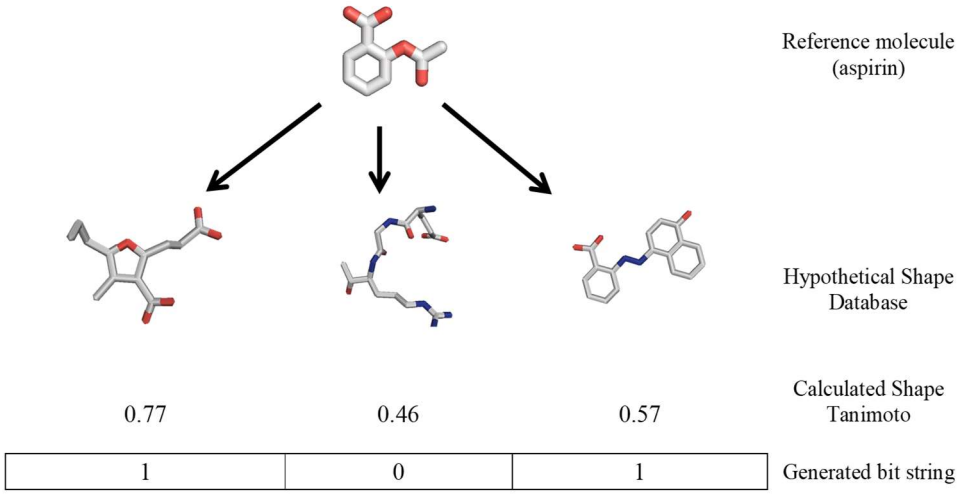


Figure 1-14. The example of shape fingerprint generation process for aspirin with a hypothetical set of reference shapes and BOV of 0.5.

The corresponding bits are turned on or off depending on whether the similarity between the shape of molecules (defined by Shape Tanimoto (ST), which is calculated as shown in **Equation 1-1**) is greater or smaller than the cut-off value.

Equation 1-1. Shape Tanimoto.

$$ST_{AB} = \frac{V_{AB}}{V_{AA} + V_{BB} - V_{AB}}$$

Where V_{AB} is the Gaussian overlap volume of the two molecules (A and B) aligned in such a way as to maximize the overlap. V_{AA} and V_{BB} are self-overlap volumes. Shape Tanimoto can vary from 0 (for the most dissimilar shaped molecules) to 1 (for molecules of identical shape).

The shape similarity measurement is obtained by comparisons of bit strings in the created fingerprints, in the exact same way as described in section 1.4.2 about the most common similarity coefficients used in molecular fingerprint methods.

1.6.3. Moment-based methods – shape multipoles

There are many methods characterizing shape by a set of descriptors, which allow for much faster screening of even large libraries of conformations. One such method is the shape multipole method, developed by A. Grant and B. Pickup.[71] They described the algorithm for calculation of shape multipoles based on Gaussian density, which allows fast comparison of molecular shapes.

Sets of descriptors need to be generated for the comparison phase - centroids and multipoles (monopole, dipoles, quadrupole, octupole moments). These are used to generate a quantitative comparison of the molecules by the sums of differences. It is assumed that a more detailed estimation of shape is obtained when higher order multipoles are calculated and used for comparison.

Reduction of the time needed for comparison of two molecules is achieved by purely describing shape as numbers - thus those approaches are often labelled as numerical methods. The centroids and multipoles required for a similarity search can be computed at negligible cost.

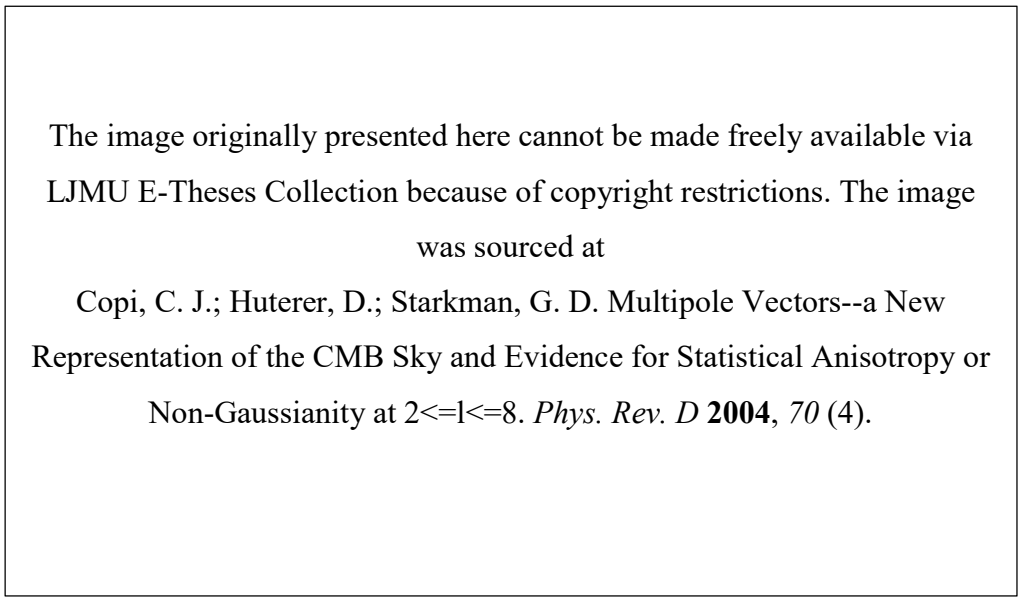


Figure 1-15. The visualisation of l=2 to l=8 multipole moments.[73]

In physics the electric dipole moment (**Figure 1-15**) is used to describe the distribution of charge within a system, as shown in **Equation 1-2** and the analogous equation can be written based on Gaussian density (**Equation 1-3**).[71]

Equation 1-2. Electric dipole moment, defined as the first order of the electric multipole expansion, where ρ^{elec} is electrostatic charge density and r is a Cartesian coordinate of a point.

$$p = \int r\rho^{elec}(r)dr$$

Equation 1-3. The first order term in the shape multipole expansion, where V is defined as the Gaussian volume, r_α, r_β, r_γ, etc. are the Cartesian coordinates of a point and ρ^g is Gaussian density of a molecule.

$$S_\alpha^{(1)} = \frac{1}{V}\int r_\alpha \rho_M^g(r)dr$$

Higher order terms, shape quadrupoles and octupoles, can be defined as in **Equation 1-4** and **Equation 1-5**, respectively, leading to a more and more accurate description of the shape of a molecule.

Equation 1-4. Shape Quadrupole, the second order moment.

$$S_{\alpha\beta}^{(2)} = \frac{1}{V}\int r_\alpha r_\beta \rho_M^g(r)dr$$

Equation 1-5. Shape Octupole, the third order moment.

$$S_{\alpha\beta\gamma}^{(3)} = \frac{1}{V}\int r_\alpha r_\beta r_\gamma \rho_M^g(r)dr$$

Chapter 2

Shape Fingerprints

2.1. Introduction

One commonly used ligand-based approach is molecular fingerprinting in which binary bit strings encode the structure of compounds allowing fast calculations with low storage needs.[40] There are many types of fingerprints and most encode only the atom types and how they are connected to one another and so do not describe the three-dimensional character of molecules.[41,74,39,47,45,40,75] However, these techniques rarely lead to compounds that are chemically very distinct and thus they are not well-suited for scaffold hopping wherein compounds are sought that have a desired activity, but are different enough to have improved ADMET properties (and hopefully are novel and patentable). [57,76]

By contrast, the shape fingerprint method encodes only the shape of compounds and not chemical structural information.[77] The shape of a molecule is measured indirectly by alignment to a database of diverse reference shapes. To be effective, these reference shapes must represent all shapes of molecules that are likely to bind to proteins. With a set of reference shapes in hand, shape similarity can be assessed by comparisons of the bit strings in the created fingerprints. Previous work of Haigh et al.[77] has outlined

the concept of shape fingerprints and established the parameters that can be varied to tune the fingerprint. However, a link between shape fingerprints and biological activity has never been established. This chapter shows how the description of shape via fingerprints for the explanation of biological activity was optimized. It presents the results of such optimization and shows that shape fingerprints are able to group molecules that are similar enough to have shared biological activity. Further, it describes the optimal method for making this link and makes the data needed to perform these calculations available.

2.2.Methods

2.2.1. Creating a database of reference shapes

In the work described in this chapter, sets of reference shapes were generated by implementation of the algorithm previously described by Haigh et al.[77] The algorithm (which uses Openeye's Shape Toolkit [78]) randomly selects a first reference molecule out of the input dataset. The remaining molecules in the dataset are compared to the reference molecule and a Shape Tanimoto (ST) calculated that can be defined as:

$$ST_{AB} = \frac{V_{AB}}{V_{AA} + V_{BB} - V_{AB}}$$

Where V_{AB} is the Gaussian overlap volume of the two molecules (A and B) aligned in such a way as to maximize the overlap. V_{AA} and V_{BB} are self-overlap volumes. Shape Tanimoto can vary from 0 (for the most dissimilar shaped molecules) to 1 (for molecules of identical shape). Molecules with ST greater than a user-selected value (the Design Tanimoto, DT) were discarded. The molecule with the smallest ST was then selected as the next reference molecule and the same process repeated until all molecules have either been selected as a reference shape or discarded (**Figure 2-1**). Thus, Design Tanimoto defines how similar are the shapes that are forming the Shape Database. Each set of reference shapes forms a Shape Database, referred to here as SDx where x is a distinguishing number.

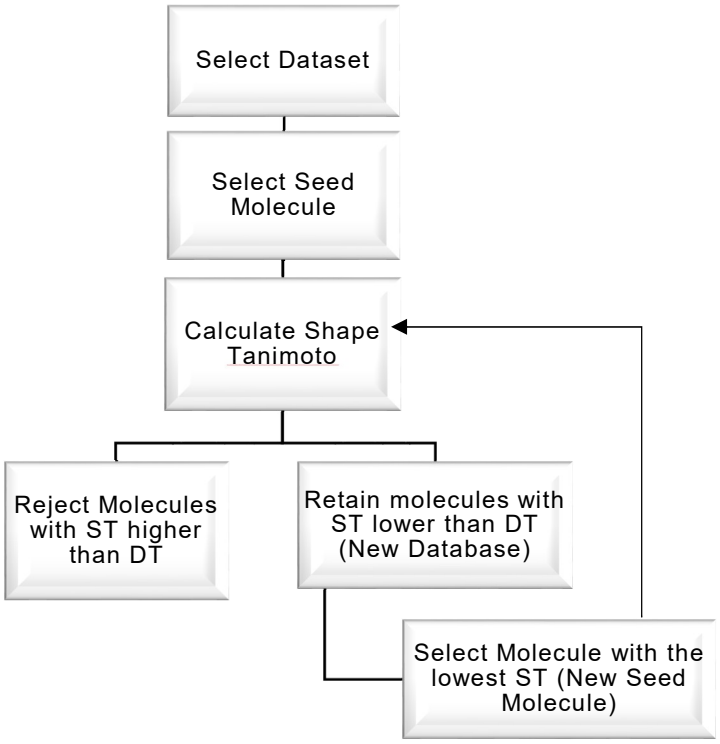


Figure 2-1. The workflow of Shape Database generation process.

2.2.2. Generating shape fingerprints

Shape fingerprints were generated by comparing a query molecule with each reference shape in the Shape Database in turn. As shown in the **Figure 2-2**, for each reference shape, if the ST was above another user-defined value, the Bit-On Value (BOV) then the corresponding bit was set to 1 and if below the BOV, the bit was set to 0, and this way producing the bit string of length equal to number of shapes included in Shape Database.

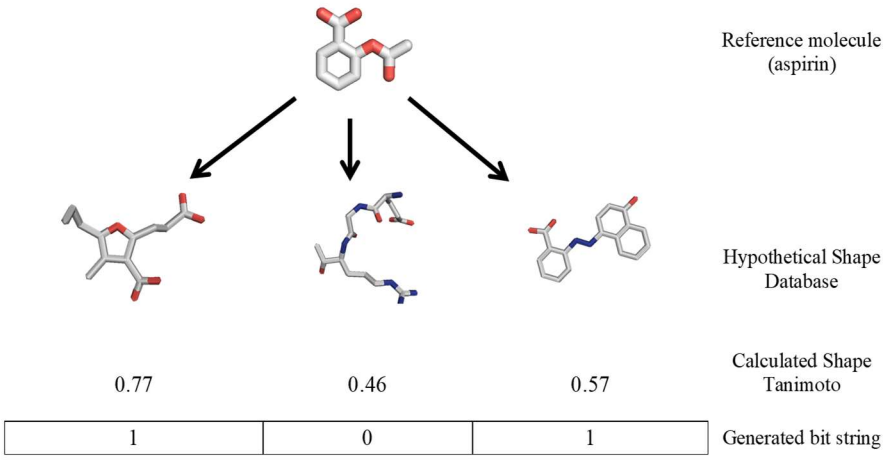


Figure 2-2. The example of the shape fingerprint generation process for aspirin with a hypothetical set of reference shapes and BOV of 0.5.

All the molecules were compared by aligning their bit strings, counting the number of bits set on (at 1) only in one of the strings and those set on in both strings. Bit Strings for molecular shapes A and B were compared using the Fingerprint Tanimoto (FT) as a similarity measure:

$$FT_{AB} = \frac{bothAB}{onlyA + onlyB + bothAB}$$

Where onlyA and onlyB are the numbers of unique bits on in the bit strings for A and B respectively, while bothAB is the number of bits on in common to A and B. Fingerprint Tanimoto similarity values vary from 0 (for dissimilar compounds) to 1 (for the most similar molecules).

2.3.Analysis

The evaluation of the shape fingerprints approach was performed by employing two test sets: 1) a set described by Taylor et al.,[79] which was devised to test pharmacophore models and consists of 87 molecules binding to 10 different proteins as shown in **Table 2-1** 2) a group from the Astex diversity set,[80] which includes 45 molecules binding to 4 selected proteins, shown in **Table 2-2**.

Table 2-1. Test Set 1 - Test set described by R. Taylor et al. [79] used for validation of the shape fingerprints method.

PROTEIN	NUMBER OF COMPLEXES	PDB CODES
Protein kinase 5 (PK5)	2	1v0o, 1v0p
Fatty acid binding protein (FABP)	3	1tou, 1tow, 2hnx
Neprilysin (NEP)	4	1dmt, 1r1h, 1r1j, 1y8j
Dihydrofolate reductase (DHFR)	6	1drf, 1hfr, 1mvt, 1pd9, 1s3v, 2dhf
Checkpoint kinase (Chk1)	16	1nvq, 1nvr, 1nvs, 1zlt, 1zys, 2br1, 2brb, 2brg, 2brh, 2brm, 2bro, 2c3l, 2cgu, 2cgw, 2cgx, 2hog
Neuraminidase (NEU)	11	1a4g, 1a4q, 1b9s, 1b9t, 1b9v, 1inf, 1inv, 1ivb, 1nsc, 1nsd, 1vcj
Carbonic anhydrase (CA)	13	1bn3, 1bn4, 1bnq, 1cim, 1eou, 1if7, 1oq5, 1xpz, 1zgf, 1zh9, 2eu3, 2hoc, 2nng
Adenosine deaminase (ADA)	11	1krm, 1ndv, 1ndw, 1ndy, 1o5r, 1qxl, 1uml, 1v7a, 1v79, 1wxy, 2e1w
Heat shock protein 90 (HSP)	10	1byq, 1uy8, 1yc1, 1yc4, 1yet, 2bsm, 2byi, 2bz5, 2cct, 2uwd
Acetylcholinesterase (AChE)	11	1dx6, 1e66, 1eve, 1gpk, 1gpn, 1h23, 1w4l, 1zgb, 2ack, 2c5g, 2ckm

Table 2-2. Test Set 2 - Test set made from the selected targets from Astex diversity set [80] used for validation of the shape fingerprints method.

PROTEIN	NUMBER OF COMPLEXES	PDB CODES
Chitinase B	8	1w1p, 1w1t, 1w1v, 1w1y, 3wd1, 3wd2, 3wd3, 3wd4
TMK	8	1mrs, 1w2g, 1w2h, 4unn, 4unp, 4unq, 4unr, 4uns
Tryptophan Syntase	13	1k3u, 1k7e, 1k7f, 1qop, 1yjp, 1wbj, 2cle, 2clh, 2clk, 2j9y, 4hpx, 4ht3, 4kkx
VDR	16	1db1, 1ie8, 1ie9, 1s0z, 1s19, 1txi, 2ham, 3auq, 3aur, 3ax8, 3kpz, 3vhw, 3x31, 3x36, 4ite, 5gt4

In order to analyse the results, the ROC curve was used, which is a tool for diagnostic test evaluation.[81] The ROC curves and AUC values were produced in R.[82] Half of the matrix without the diagonal was used in these calculations.

2.4.Conformations

SMILES were generated using Openeye's OEChem Toolkit [78] for all the molecules shown in **Table 2-1** and **Table 2-2**. Some of the generated SMILES needed manual assignment of stereochemistry. Conformations were generated using OMEGA software[83] with the maximum number of generated conformers set to 5. Shape fingerprints were generated for each conformation. When two molecules were compared, all fingerprints of one molecule were compared with all those of the other. Two summary values for this comparison were investigated: 1) the highest value of FT amongst the array arising from comparisons of all conformations of one molecule with all conformations of the other is selected – the MV (maximum value) method or 2) the average of those values is selected – the AV (average value) method.

2.5. Results and Discussion

2.5.1. Defining a set of reference shapes

A shape fingerprint is a binary encoding of the similarity of the shape of any given query molecule to a series of reference shapes. Having an appropriate set of these reference shapes is therefore critical. Haigh et al.[77] used a set of reference shapes generated from the Cambridge structural database of small molecule crystal structures[84] and a set of conformations generated by CORINA for the MDDR database[85] of molecules that have been studied clinically. As there is great interest in protein-ligand interactions, thus instead it was chosen to use the database of ligands studied by X-ray crystallography in complex with a protein – the Ligand Expo dataset.[86] At the time, this contained the experimental coordinates for 1,158,763 non-polymer molecules and non-standard amino acids and nucleotides. Various filtering criteria based on molecular weight were applied to these molecules (**Figure 2-3**), leading to 9 databases of shapes that were considered as input to the algorithm that was used to generate the sets of reference shapes. An alternative filtering based on the number of heavy atoms was also performed and yielded similar effects.

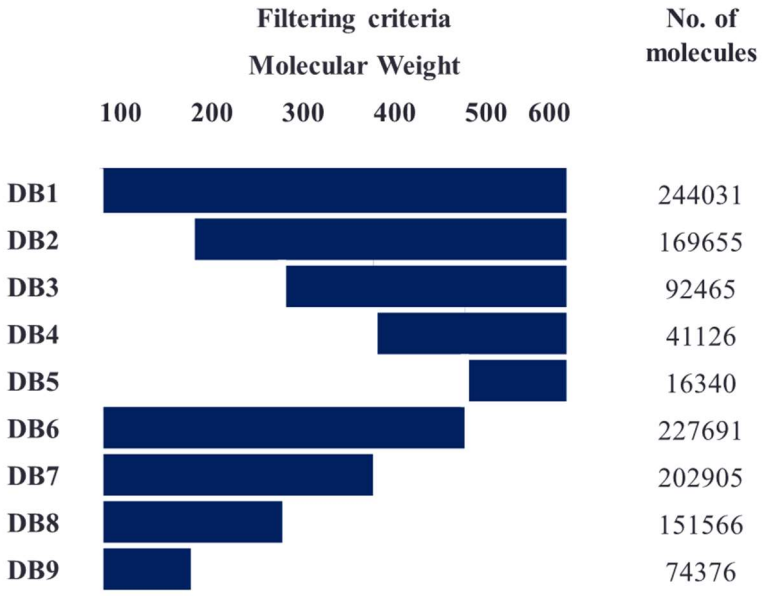


Figure 2-3. The Ligand Expo Dataset [86] was filtered in 9 different ways according to the lower and upper limit of molecular weight shown. The number of structures to pass the filter criteria is shown.

An implementation of the algorithm described by Haigh et al.[77] permitted each set of ligand structures to be clustered in such a way that every shape had a Shape Tanimoto to at least one reference shape that is above a user-selected cut-off called the Design Tanimoto (DT). Shape comparisons were performed with Openeye's Shape Toolkit, [78] as implemented in ROCS software.[72] A randomly selected shape is the first reference shape and its Shape Tanimoto (ST) with every other shape in the input database is computed. All those that have Shape Tanimoto values above DT are rejected from further consideration. After all comparisons have been made, the shape that has the lowest Shape Tanimoto with the starting shape is selected and becomes the next reference shape in the database. The process is repeated until all shapes have either been rejected or selected as a reference shape. During initial investigations, a low (0.5) and a high value of DT (0.7) were investigated; lower values of DT lead to smaller Shape Databases.

2.5.2. Evaluating shape fingerprints

The shape fingerprints were evaluated by computing their ability to correctly group the molecules in two test sets. The first test set comprises a set of 87 molecules each of which is known to bind to one of ten proteins listed in **Table 2-1**.[79] A second test set, shown in **Table 2-2**,was extracted from the Astex diversity set and comprised 45 molecules binding to four different proteins.[80] Both test sets include only molecules with known protein-ligand structures and hence ligand bioactive conformations.

As described above, the choice of DT influences the size and nature of the Shape Database and this was investigated. Shape fingerprints are generated by computing the ST between a query structure and every shape in the Shape Database. When the calculated ST is above a user-defined cut-off, the Bit On value (BOV), the bit is set On (1) otherwise it is set to Off (0). Lower BOVs lead to higher bit densities. In the initial testing of the shape databases, a high and low value (0.7 and 0.5 respectively) for each of DT and BOV were used.

The shape fingerprints for every molecule in both test sets were compared to those for every other molecule in the set. The comparison yielded another Tanimoto, the Fingerprint Tanimoto (FT). Receiver Operating Characteristic curves (ROC curve)

were created that plot the true positive rate against the false positive rate. The computed AUC (Area Under Curve) for these is a measure of accuracy, where 0.9 – 1 represents a perfect test while 0.5 represents a poor one (equivalent to random). The AUCs for Test Set 1 are shown in **Figure 2-4**.

When DT=0.5 and BOV=0.7, the AUC is rarely distant from 0.5 suggesting no discrimination was achieved. The combination of low DT and high BOV leads to a low number of bits set On and so these are unlikely to be able to connect molecules (which requires bits to be set in common). The best AUC values for this test set was obtained for SD05 with DT=0.5, BOV=0.5 (AUC=0.67). The difference between AUC values for all settings was small (excluding the aforementioned settings: DT=0.5, BOV =0.7) for all databases except for SD05 and SD09.

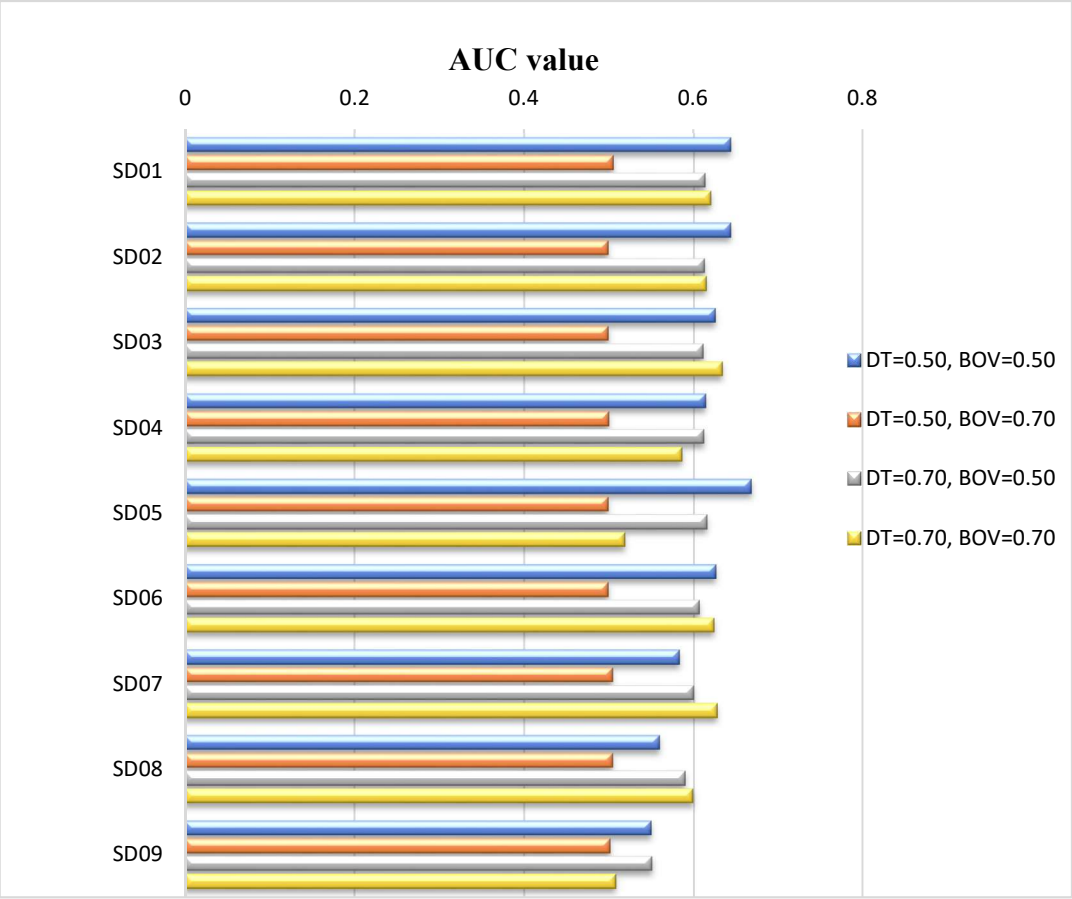


Figure 2-4. The AUC values for Test Set 1 when applying different settings: DT=0.5 with BOV=0.5 and 0.7, and DT=0.7 with BOV=0.5 and 0.7.

The plot looks quite similar for Test Set 2, as shown in **Figure 2-5**. Once again, SD09 performs noticeably worse than the other sets of reference shapes and the combination of DT=0.7 with BOV=0.5 provides poor discrimination. The highest AUC values, exceeding 0.83, are obtained for SD01 and SD04 with DT=0.5, BOV=0.5 and for SD03 and SD06 with DT=0.7 and BOV=0.7. When results for DT=0.5 and BOV=0.7 are excluded, SD03 performs the best with an average AUC of 0.81, followed by SD01 and SD02 with the same average AUC of 0.80.

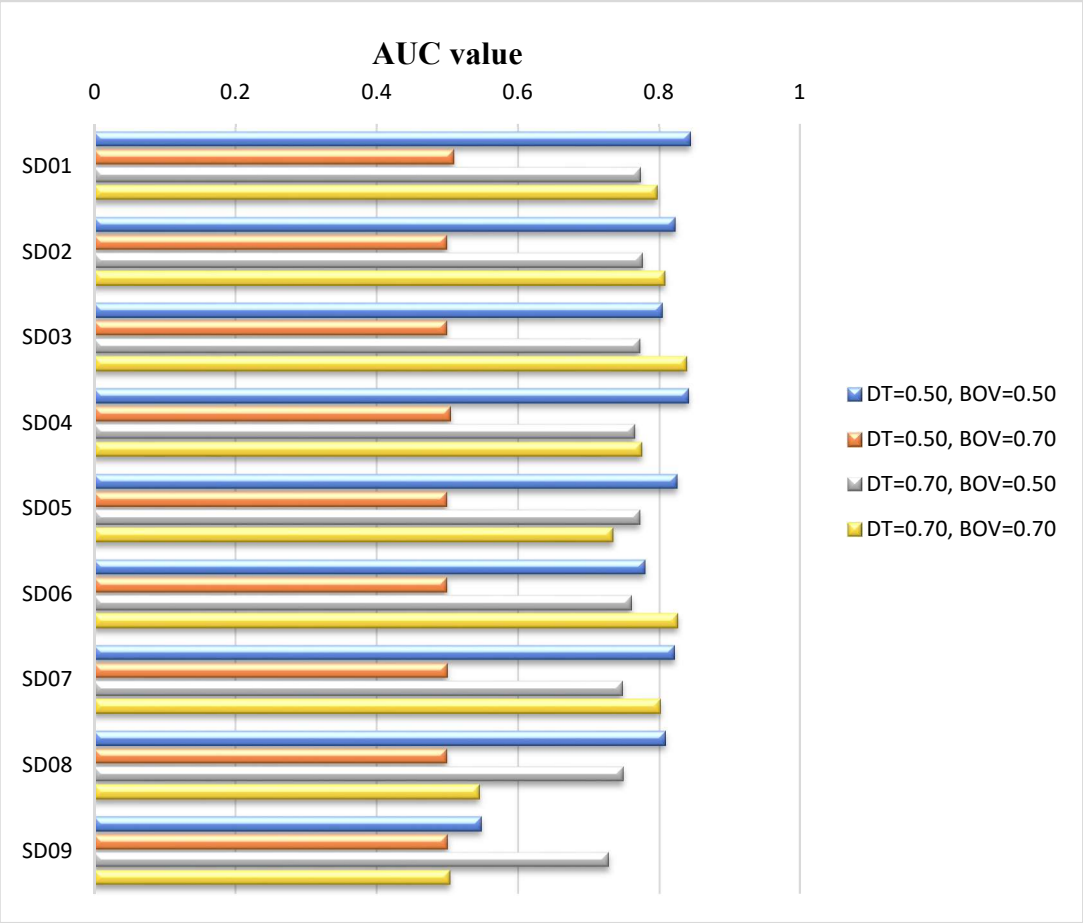


Figure 2-5. The AUC values for Test set 2 with varying settings.

When the average AUC values obtained from both test sets are computed (excluding DT=0.5, BOV=0.7), SD03 and SD06 (with AUCs of 0.71 and 0.70 respectively) stand out as best when SD01, the unfiltered shape database, is excluded. The filtering criteria used to generate SD03 and SD06 were therefore combined to generate Shape Database 10 with molecular weight in the range 300 to 500 with the expectation that this would provide the best balance of accuracy and speed (76125 molecules pass the filters for

consideration in the database generation process, compared to 244031 in SD01, 92465 in SD03 and 227691 in SD06).

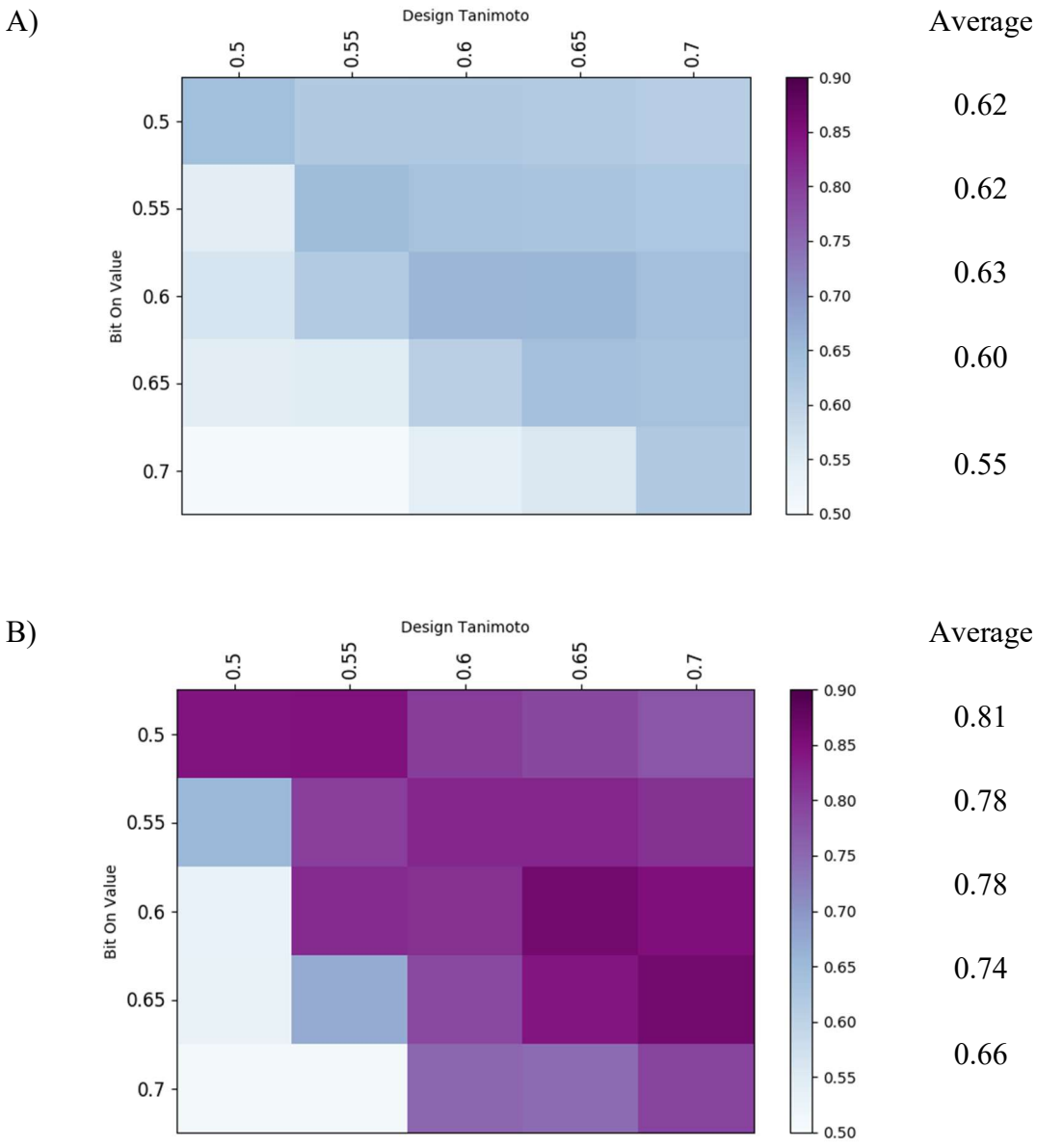


Figure 2-6. The heatmaps generated based on AUC values for Test Set 1 (A) and Test Set 2 (B) when using SD01 with varying DT and BOV.

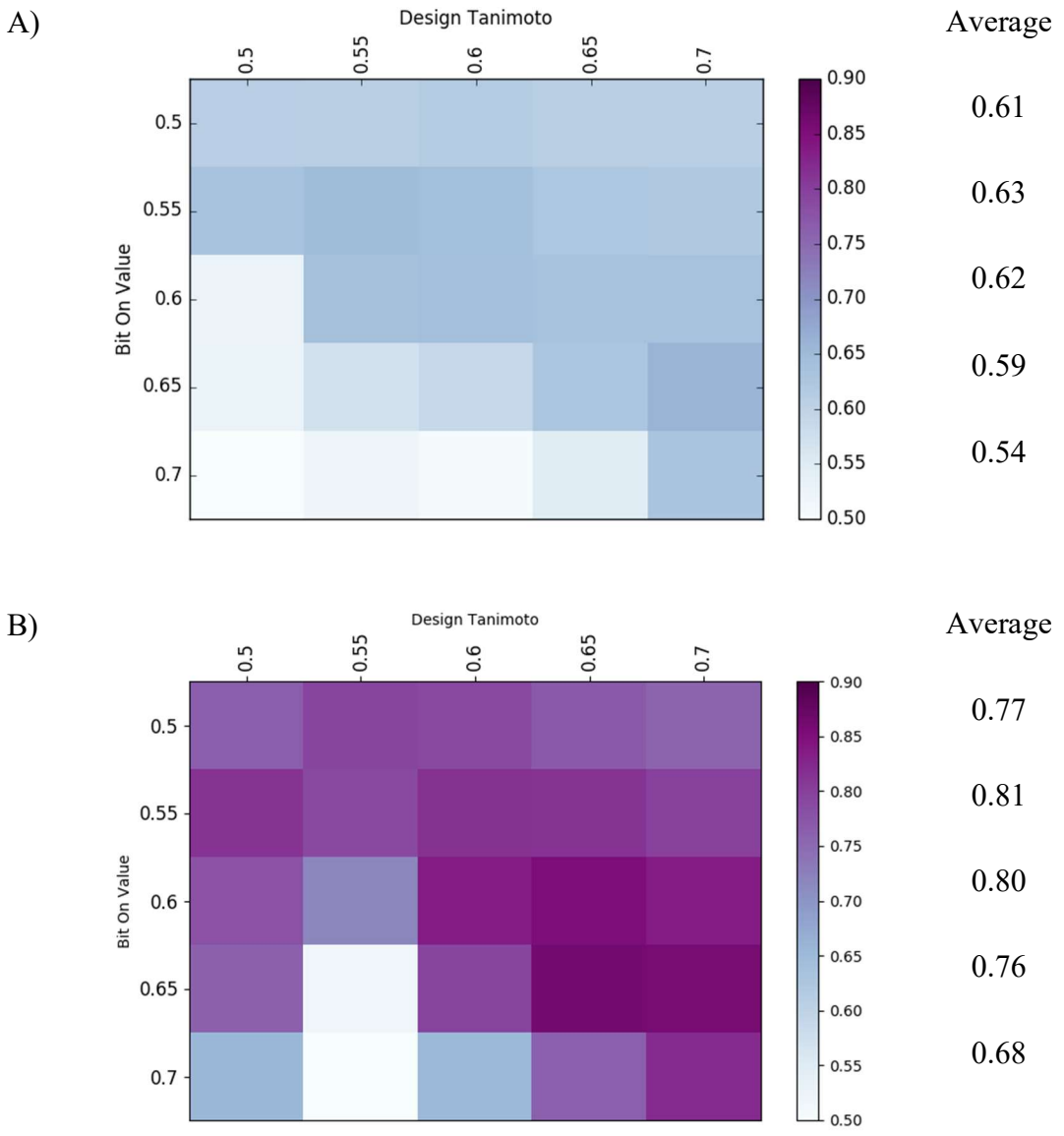


Figure 2-7. The heatmaps generated based on AUC values for Test Set 1 (A) and Test Set 2 (B) set when using SD10 with varying DT and BOV.

The values of DT and BOV were then systematically varied in steps of 0.05 between 0.5 and 0.7. As can be seen in **Figure 2-6** and **Figure 2-7**, the AUC values vary less for Test Set 1 (**Figure 2-6**A, **Figure 2-7**A), than for Test Set 2 (**Figure 2-6**B, **Figure 2-7**B). This might be caused by differences in molecular weight distribution in both sets. [79,80] Experience with other datasets had shown that small molecules (about 200 Da and below) and large molecules (about 800 Da and above) set very few (or no) bits and so cannot be correctly described by these shape fingerprints. The molecular weight ranges for the two test sets used in the present study are shown in **Figure 2-8**. In Test

Set 2, the range is slightly wider than for Test Set 1 and this may facilitate the correct grouping of Test Set 2. The generally good performance suggests that the shape databases obtained are applicable to molecules spanning the molecular weight range ~200 to ~600 and should therefore be useful for most drug-like molecules.

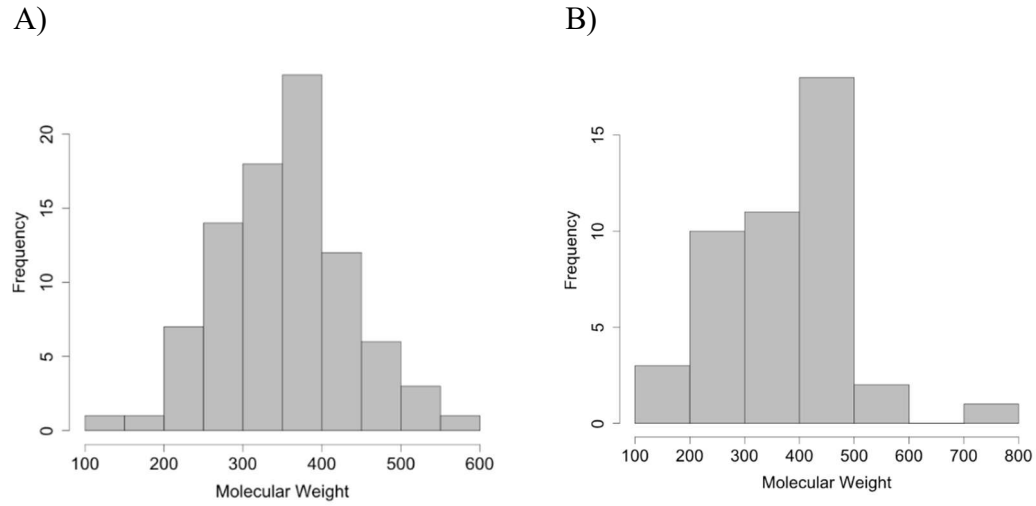


Figure 2-8. The molecular weight distribution of Test Set 1 (A) and Test Set 2 (B).

When the variation caused by changing DT and BOV is considered in more detail, the influence of these settings on the AUC obtained is as shown in **Figure 2-6** when using SD01 and **Figure 2-7** when using SD10. This shows that using BOV=0.55 gives the best results on average and thus the impact of DT was considered while BOV was fixed. The choice of the best DT requires a consideration of the size of the Shape Database, which determines the computational time required to generate each fingerprint. When the average AUC value is viewed as a function of the size of Shape Database (**Figure 2-9**), the difference in average AUC value for the two highest values of DT for both test sets is quite small (ΔAUC is 0.022 and 0.003 for Test Set 1 and Test Set 2 respectively), yet the difference in size of the Shape Databases is significant (1346 reference shapes). Therefore, SD10 with DT = 0.65 is selected as the best performing Shape Database. The results show that for this setting of DT, the optimum BOV is 0.6. The recommended settings are therefore to use DB10 with DT=0.65 and BOV=0.6 when grouping molecules according to their likelihood of binding to the same protein.

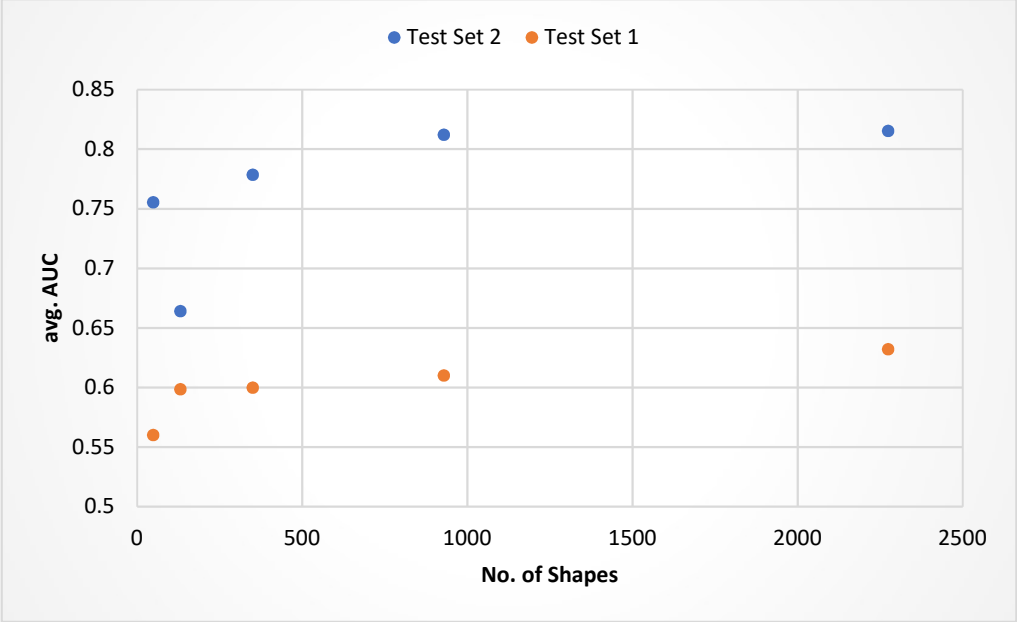
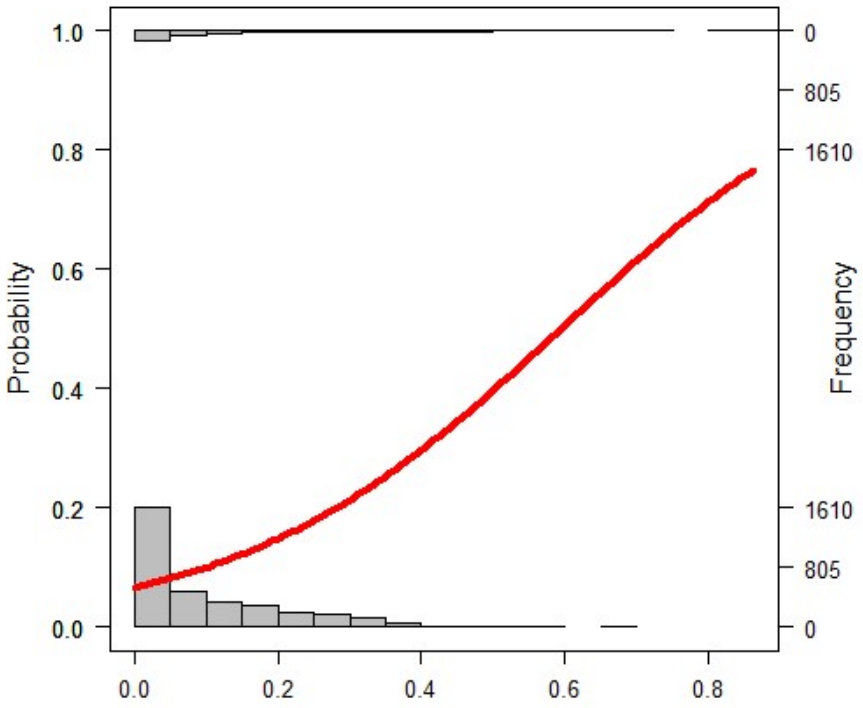


Figure 2-9. The graphs showing average AUC value as a function of size of Shape Database for both test sets when Shape Database 10 was used.

A)



B)

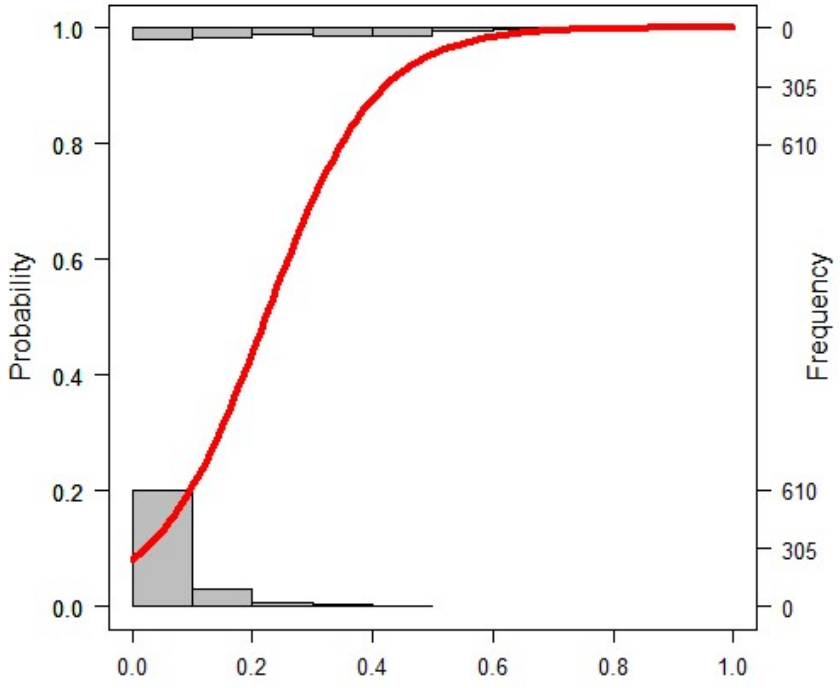
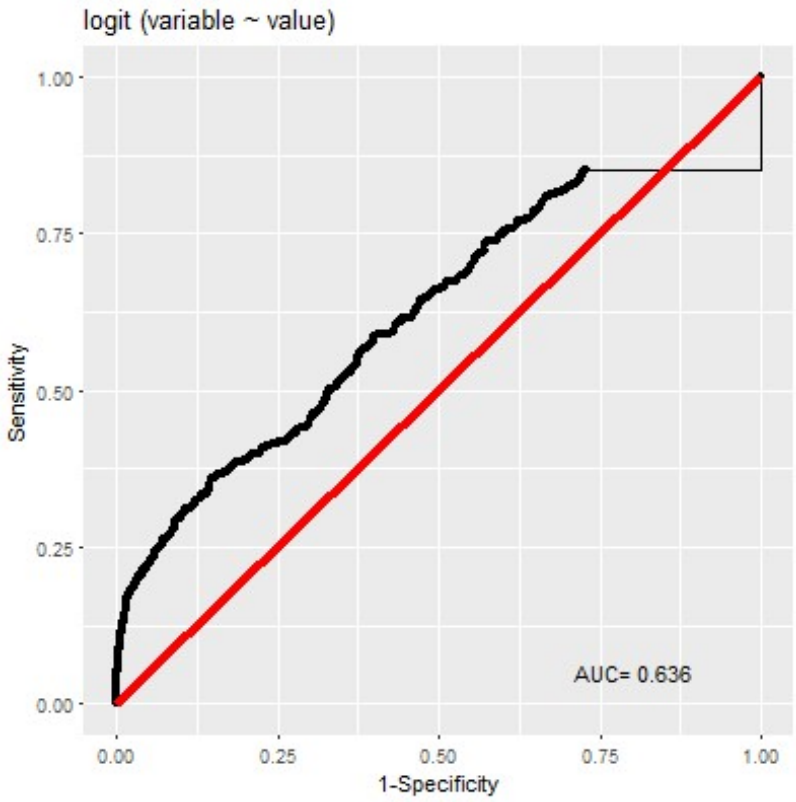


Figure 2-10. Logistic regression plot for Test Set 1 (A) and Test Set 2 (B) when using SD10 with DT=0.65 and BOV=0.60. The plots were created using R.[82]

A)



B)

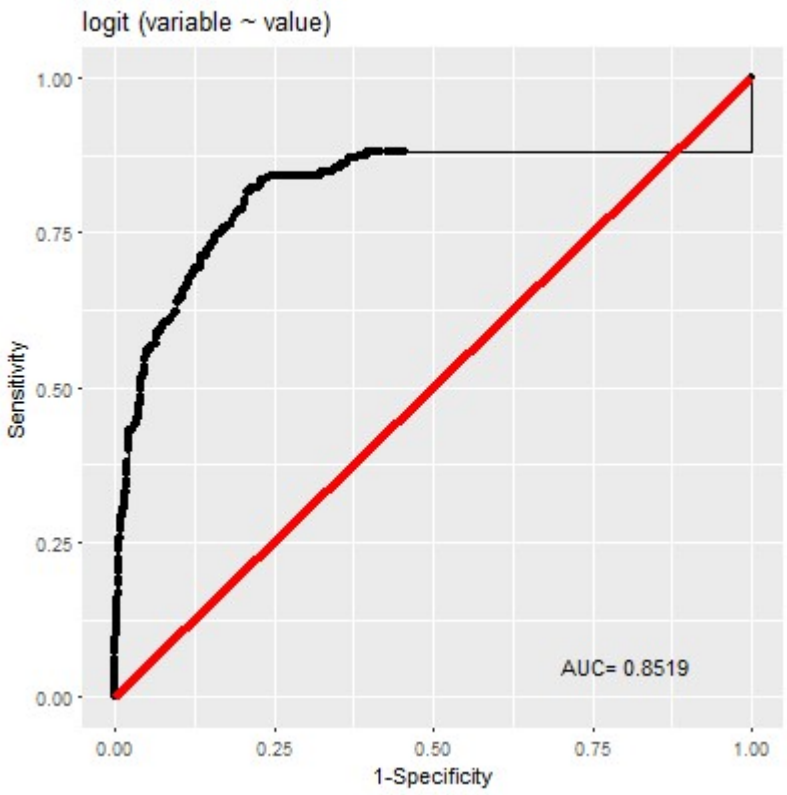


Figure 2-11. ROC curve for Test Set 1 (A) and Test Set 2 (B) when using SD10 with DT=0.65 and BOV=0.60. The plots were created using R. [82]

Having selected the optimum method, it is useful to define a cut-off value of FT above which molecules have a defined likelihood of sharing biological activity. Logistic regression permits a continuous variable, such as FT, to be linked with likelihood of belonging to a particular class and has been performed for the two test sets. This is shown in **Figure 2-10** with the frequency histograms of molecules sharing activity shown at the top and those that do not share activity at the bottom. This reveals that there is some variability between the two test sets such that when FT is above about 0.6 for Test Set 1 or above about 0.25 for Test Set 2 there is a greater than 50% chance of shared biological activity. In situations where trial data on a set of compounds is available, it should be used to calibrate the value of FT that should be used as a cut-off for the purposes of clustering. However, by merging the two datasets (giving an evaluation based on 14 protein targets) and performing logistic regression on the combined test set (**Figure 2-12**), we suggest that a value of FT above 0.45 is a reasonable estimate of when compounds are more likely to share biological activity than not.

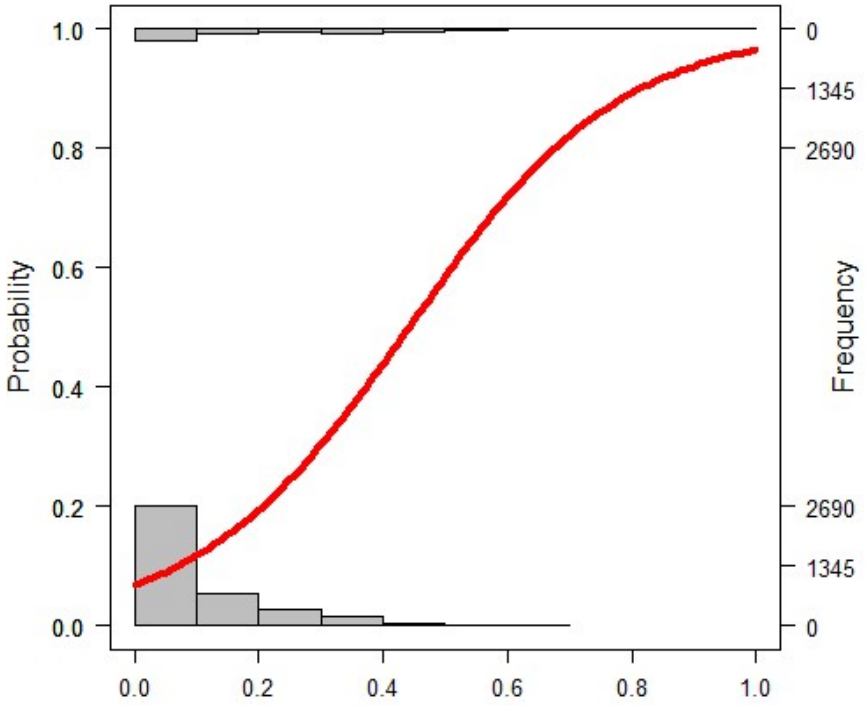


Figure 2-12. Logistic regression plot Test Set 1 and Test Set 2 combined together when using SD10 with DT=0.65 and BOV=0.60.

2.5.3. Resampling SD10

Given that the selection of reference shapes begins with a random choice, it is possible that the results are dependent upon this starting point. Therefore, SD10 was regenerated with DT=0.65 ten more times. Each of the new Shape Databases was used to generate fingerprints for both test sets and the AUC value was recomputed. The results show little variation (standard deviations vary from 0.002 to 0.033 depending on BOV) and can be seen in the **Table 2-3** and **Table 2-4**. On average, with DT=0.65 the best performance is found when using BOV=0.60. The AUC values of 0.64 and 0.84 for Test Set 1 and Test Set 2 respectively, are obtained.

Table 2-3. The AUC values for 10x resampled Shape Database 10 with DT=0.65 for Test Set 1.

		ITERATION									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
BOV	**0.50**	0.61	0.61	0.61	0.61	0.61	0.60	0.61	0.61	0.61	0.61
	0.55	0.63	0.63	0.63	0.62	0.63	0.63	0.63	0.63	0.62	0.63
	0.60	0.65	0.65	0.64	0.63	0.63	0.64	0.64	0.64	0.64	0.64
	0.65	0.65	0.64	0.64	0.63	0.63	0.62	0.66	0.62	0.66	0.62
	0.70	0.60	0.55	0.55	0.56	0.59	0.59	0.57	0.60	0.57	0.56

Table 2-4. The AUC values for 10x resampled Shape Database 10 with DT=0.65 for Test Set 2.

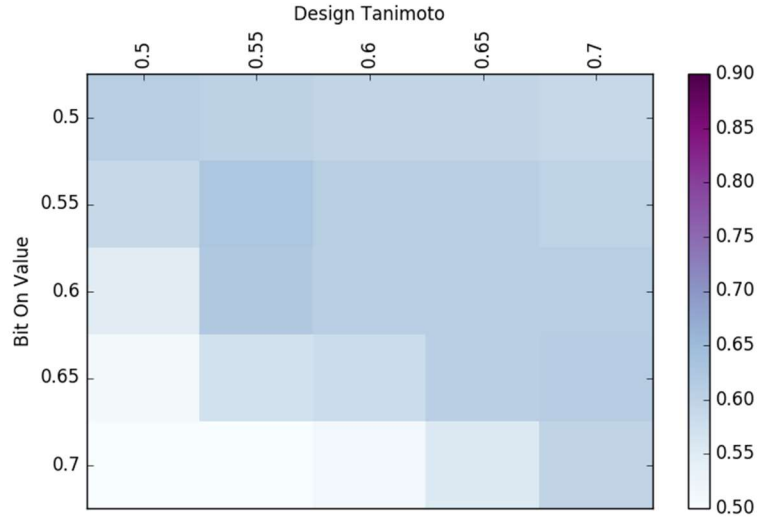
		ITERATION									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
BOV	**0.50**	0.78	0.77	0.77	0.77	0.77	0.78	0.77	0.77	0.77	0.78
	0.55	0.81	0.81	0.81	0.80	0.83	0.81	0.82	0.81	0.80	0.81
	0.60	0.86	0.85	0.85	0.83	0.83	0.86	0.84	0.83	0.82	0.85
	0.65	0.85	0.84	0.81	0.84	0.82	0.84	0.83	0.85	0.82	0.83
	0.70	0.76	0.67	0.76	0.74	0.75	0.74	0.73	0.74	0.68	0.76

2.5.4. Conformations

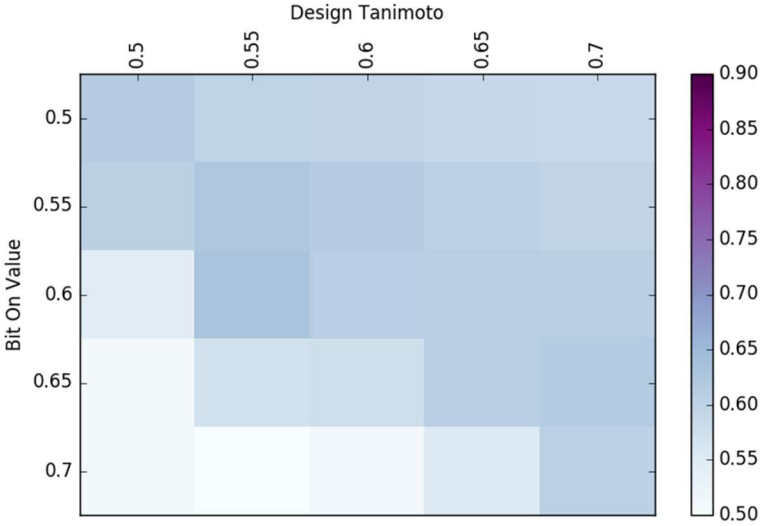
Naturally for most shape comparisons that might be of interest, a protein-ligand crystal structure would not be a useful requirement and would suggest that the activity of the molecule is already known. Therefore, conformations were generated from the SMILES string for each molecule in the test sets using Openeye's OMEGA software,[83] a knowledge-based conformer generator. In this case, relatively limited sets of up to five conformations were generated although some molecules in the set were conformationally restricted and generated less than this.

For all the conformations, shape fingerprints were generated using Shape Database 10. Two approaches for evaluating the comparison of two molecules were investigated: 1) the highest value of FT amongst the array arising from comparisons of all conformations of one molecule with all conformations of the other (MV) or 2) the average of those values (AV). As shown in **Figure 2-13**, there is only a small difference in AUC values between the methods (AV and MV), with the MV being slightly better. This is consistent with molecules requiring only one (reasonable) conformation to be similar in shape in order to share biological activity. Comparing the AUC values obtained for conformations generated from SMILES with those for crystal structures shows only a little deterioration (**Table 2-5**). The logistic regression plots and ROC curves for both test sets (when using AV and MV methods with SD10 and DT=0.65 and BOV=0.60) can be seen in **Figure 2-14** and **Figure 2-15**, respectively. Thus, using conformations generated from SMILES instead of crystal structures does not greatly affect the accuracy of the shape fingerprint method. This shows that the method can be successfully used even when the bioactive conformation of the ligand is not known.

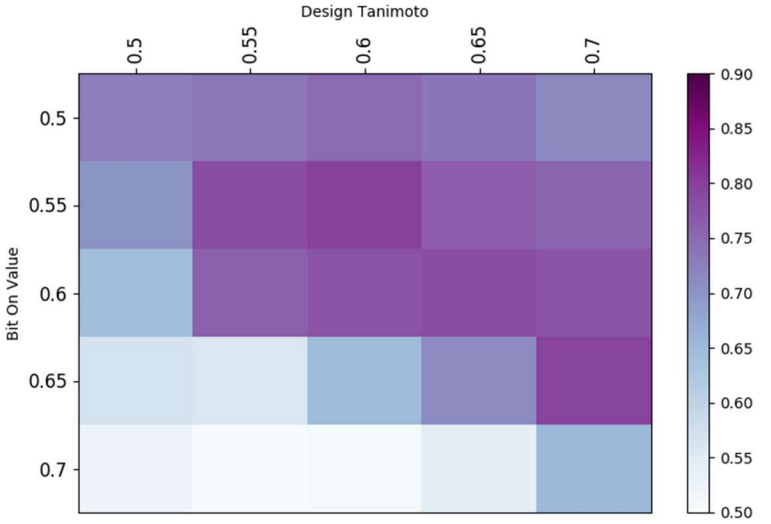
1A)


Design Tanimoto
0.5
0.55
0.6
0.65
0.7
Bit On Value
0.5
0.55
0.6
0.65
0.7
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50

1B)


Design Tanimoto
0.5
0.55
0.6
0.65
0.7
Bit On Value
0.5
0.55
0.6
0.65
0.7
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50

2A)


Design Tanimoto
0.5
0.55
0.6
0.65
0.7
Bit On Value
0.5
0.55
0.6
0.65
0.7
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50

2B)

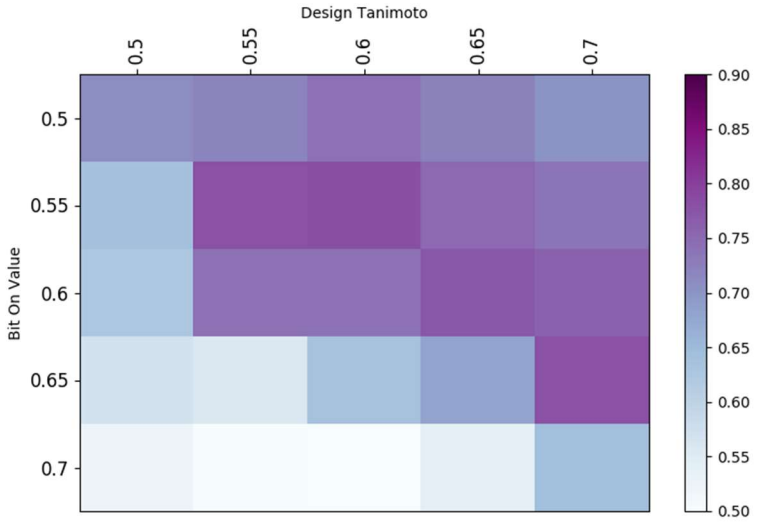
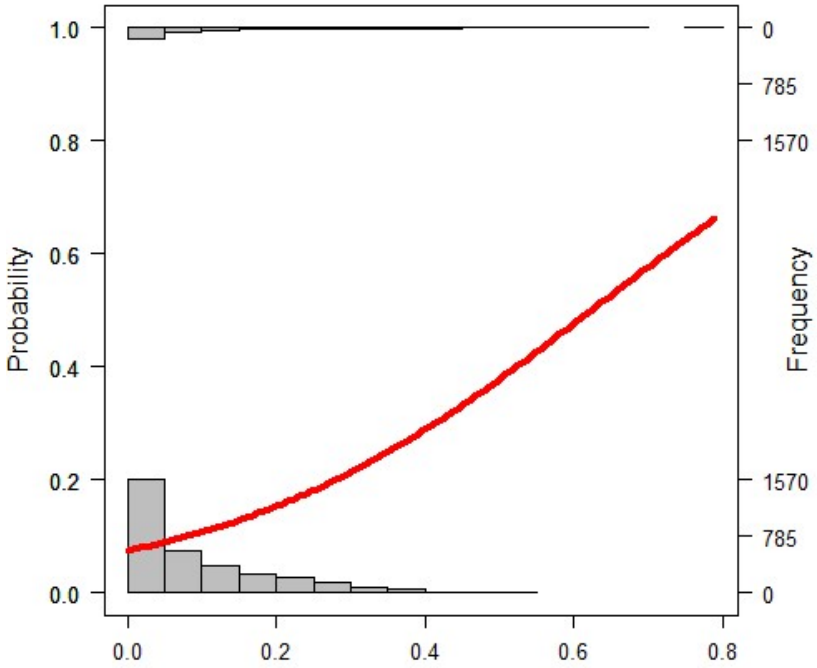
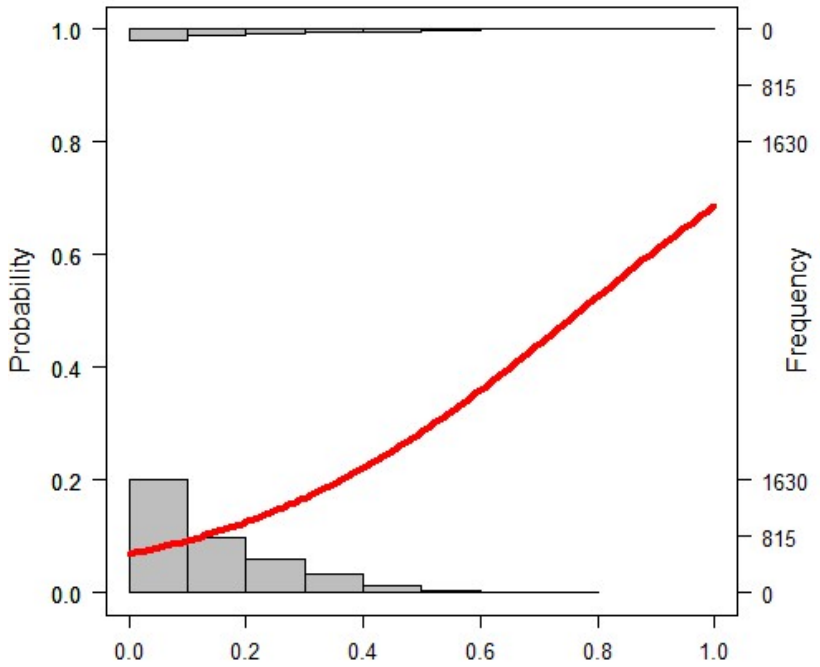


Figure 2-13. Heatmaps with AUC values for MV (A) and AV (B) methods for Test Set 1 (1) and Test Set 2 (2) when using SD10 with various DT and BOV.

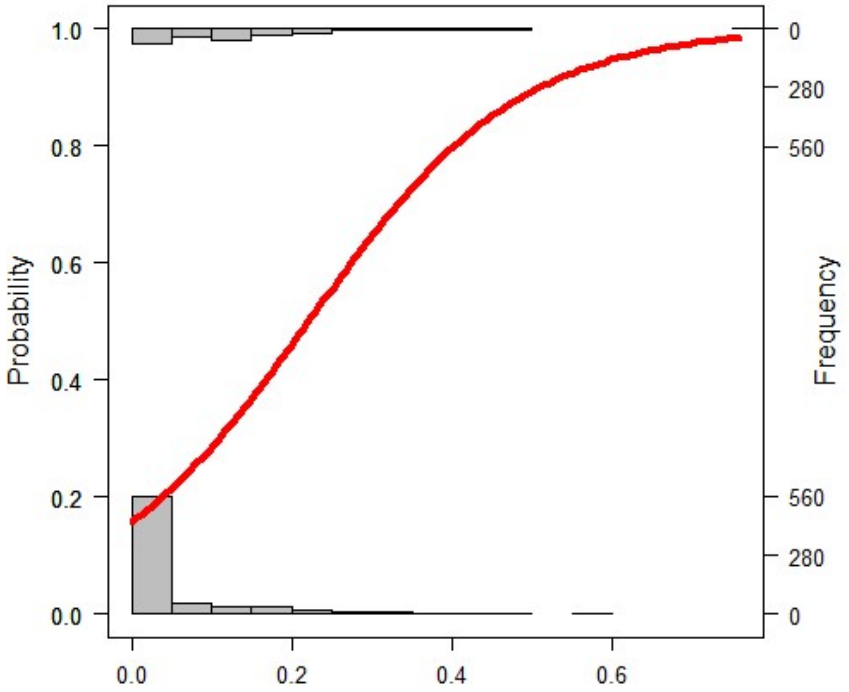
1A)



1B)


Probability
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
0
815
1630
1630
815
0
0.0
0.2
0.4
0.6
0.8
1.0

2A)


Probability
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
0
280
560
560
280
0
0.0
0.2
0.4
0.6

2B)

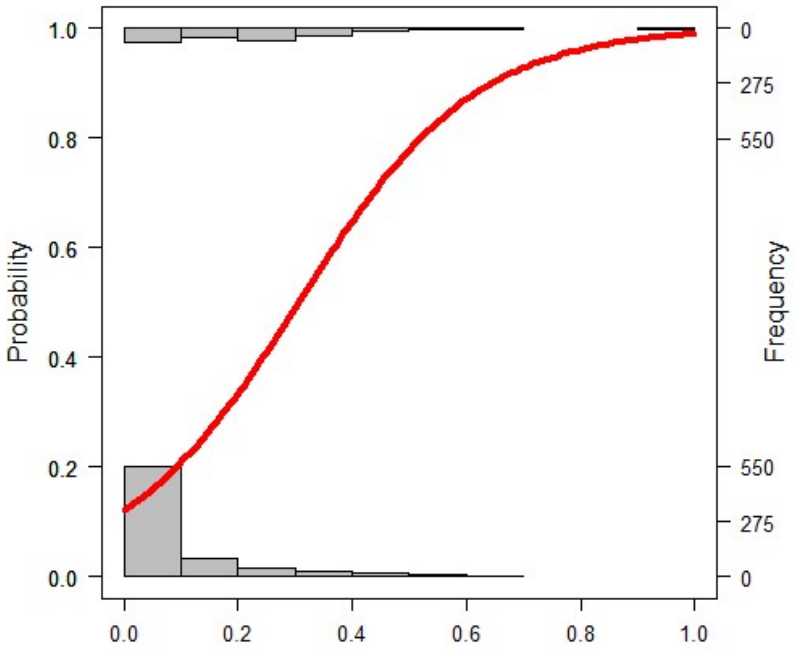
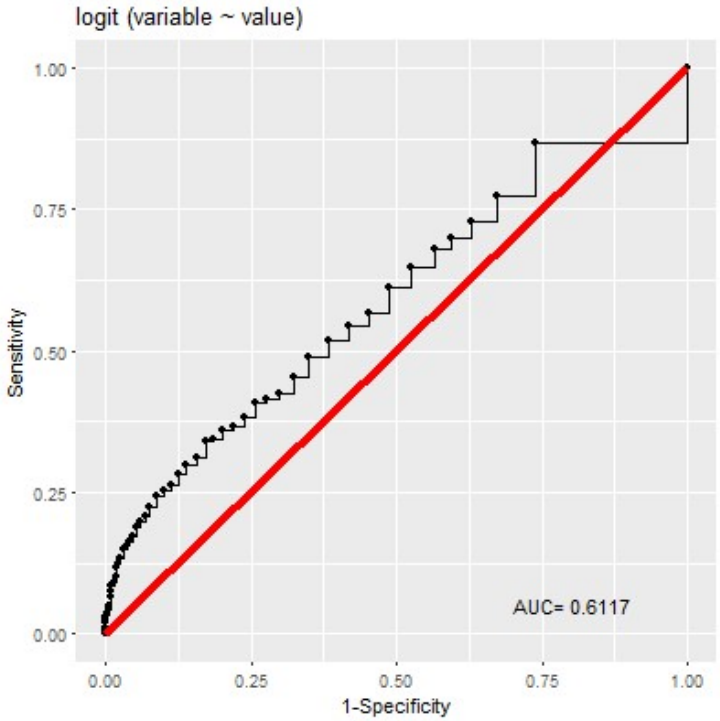
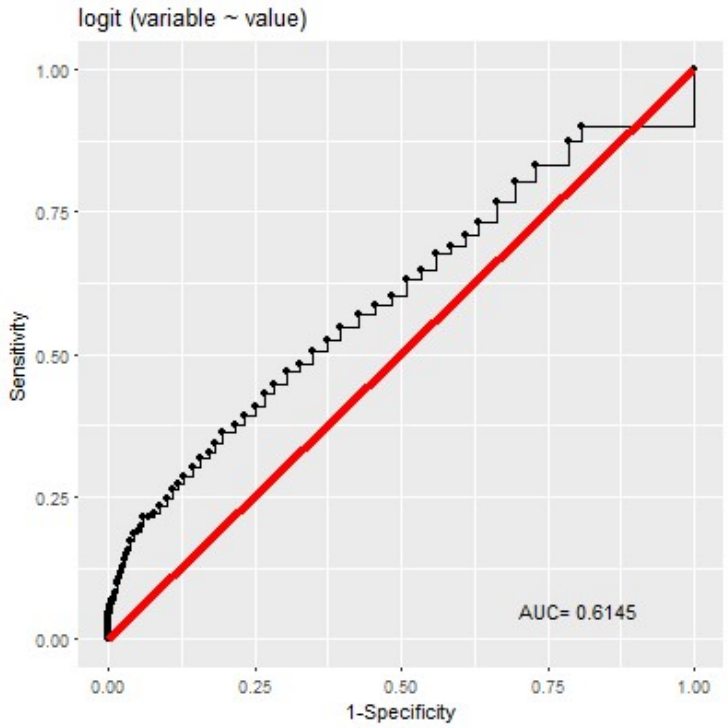


Figure 2-14. Logistic regression plot for Test Set 1 (1) and Test Set 2 (2) using AV (A) and MV (B) methods for SD10 with DT=0.65 and BOV=0.60.

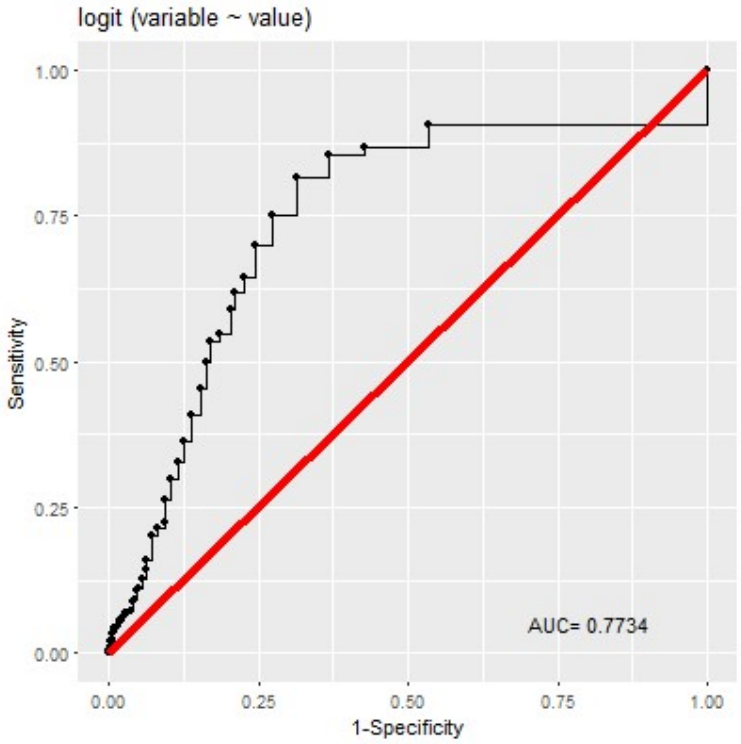
1A)



1B)


logit (variable ~ value)
1.00
0.75
0.50
0.25
0.00
Sensitivity
AUC= 0.6145
0.00
0.25
0.50
0.75
1.00
1-Specificity

2A)


logit (variable ~ value)
1.00
0.75
0.50
0.25
0.00
Sensitivity
AUC= 0.7734
0.00
0.25
0.50
0.75
1.00
1-Specificity

2B)

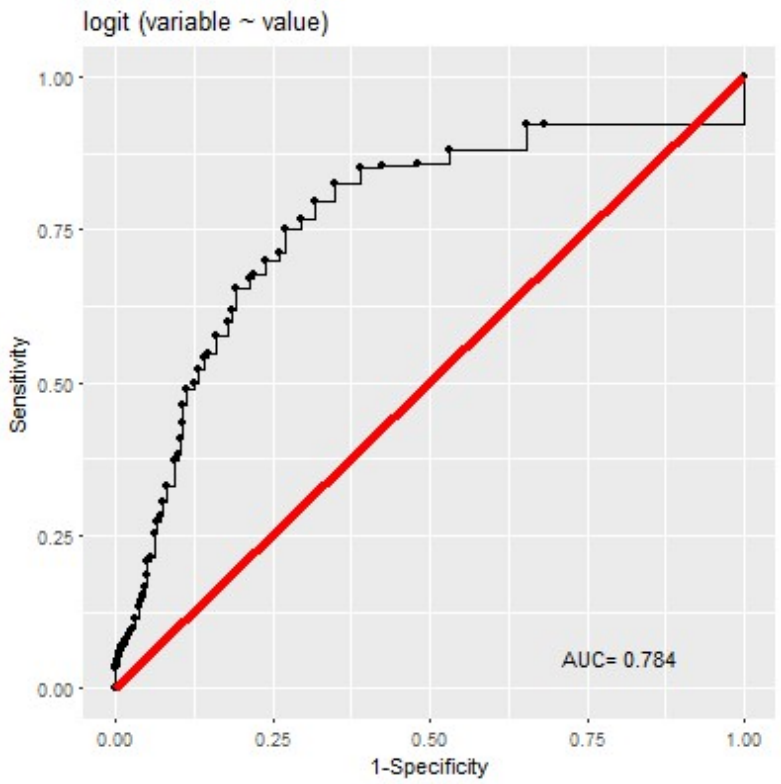


Figure 2-15. The ROC curves for Test Set 1 (1) and Test Set 2 (2) when using AV (A) and MV (B) methods for SD10 with DT=0.65 and BOV=0.60.

Table 2-5. The comparison of the AUC values of both test sets when using conformations generated from SMILES and crystal structures for SD10 with DT=0.65 and BOV=0.60.

	CONFORMATIONS		**CRYSTAL STRUCTURES**
	AV	**MV**	
TEST SET 1	0.61	0.61	0.64
TEST SET 2	0.77	0.78	0.85

2.5.5. Comparison with 2D fingerprints and scaffold hopping

As already mentioned, shape fingerprints neglect the chemical structure of the molecule. Therefore, they should complement the 2D fingerprint methods that are exclusively dependent on the chemical structure. In order to compare and contrast the two approaches, 2D fingerprints for both test sets were generated and compared using a Similarity Tanimoto. The calculated AUC values are shown in **Table 2-6**. The AUC values are higher when using 2D fingerprints for both test sets. However, considering that shape fingerprints do not use any chemical information of the molecules but only their shape, the slightly worse AUC values than for well-established methods is not too surprising.

Table 2-6. Comparison of the AUC values for different fingerprint methods. In the case of shape fingerprints, values obtained for SD10 with DT=0.65 and BOV=0.60 are shown.

	FINGERPRINT METHOD				
	MACCS166	**Path**	**Tree**	**Circular**	**Shape Fingerprints**
TEST SET 1	0.74	0.67	0.69	0.69	0.64
TEST SET 2	0.94	0.94	0.94	0.97	0.85

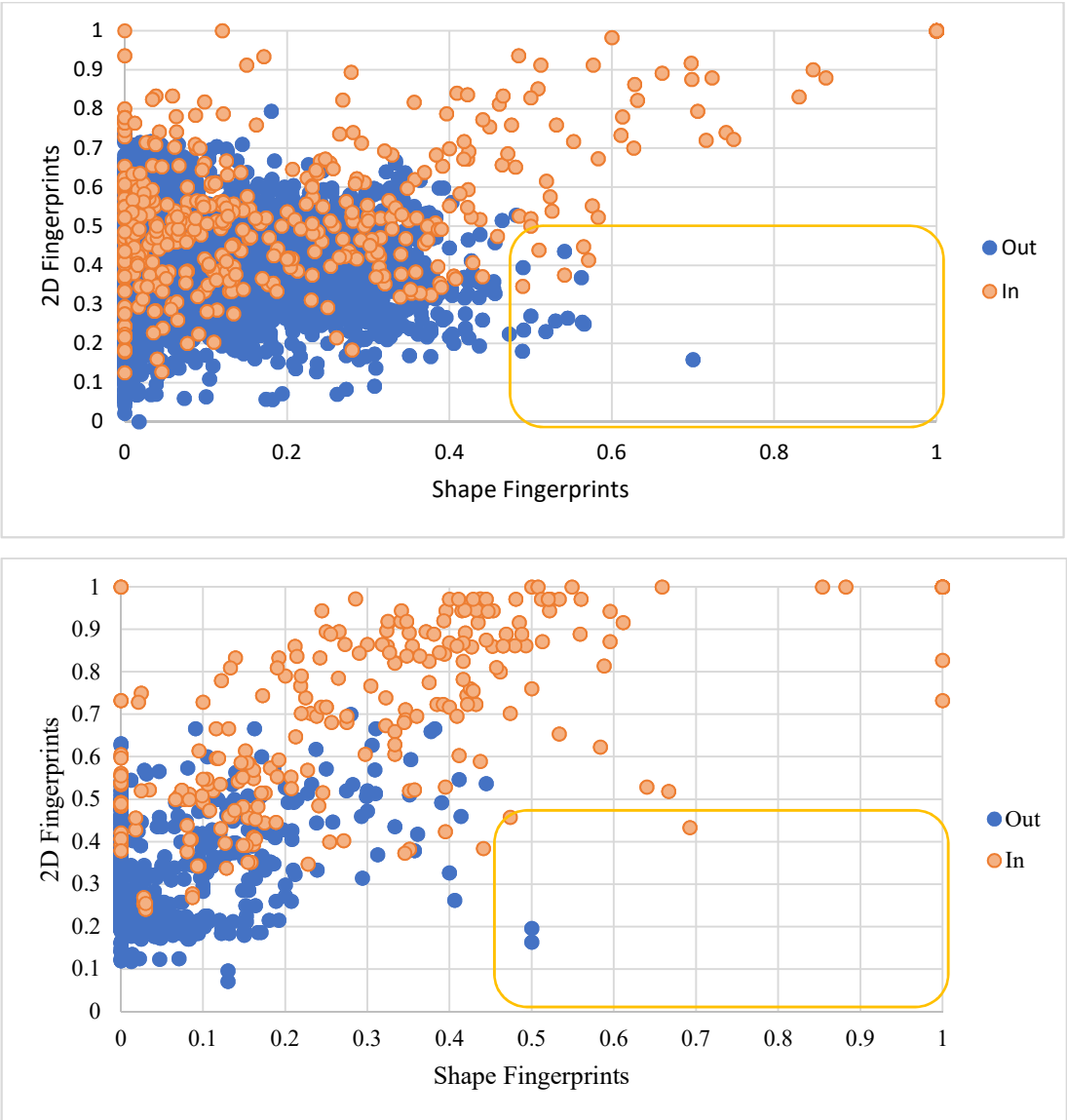


Figure 2-16. Plot showing ST scores obtained by both methods: Shape Fingerprints and MACCS166 fingerprints (2D fingerprints) for each comparison in Test Set 1 (top) and Test Set 2 (bottom). Points in red correspond to compound pairs that share biological activity those in blue do not.

To investigate the complementarity between the two fingerprint types, the Tanimotos between pairs of molecules have been computed with both methods. These are plotted against one another in **Figure 2-16**. Many pairs of molecules with shared biological activity (colored red) have high similarity according to both methods, which is unsurprising. There are a small number of examples of molecules with low shape similarity but high 2D fingerprint similarity that share biological activity but most interestingly, there are also a small number with high shape similarity and relatively

low 2D fingerprint similarity. The orange box on each of the plots in **Figure 2-16** highlights these examples. These are connections that represent scaffold hopping.

One example of a pair of structures for each of the test sets is shown in **Figure 2-17** and **Figure 2-18**. In **Figure 2-17**, an example of a connection between a neuraminidase molecule that contains an aromatic core and one with a monosaccharide core is a very clear example of scaffold hopping between inhibitors that are likely to have different physical properties while the indole and ortho-substituted phenol pair in **Figure 2-18** show that these ring-opening scaffold hops can also be detected by shape fingerprints.

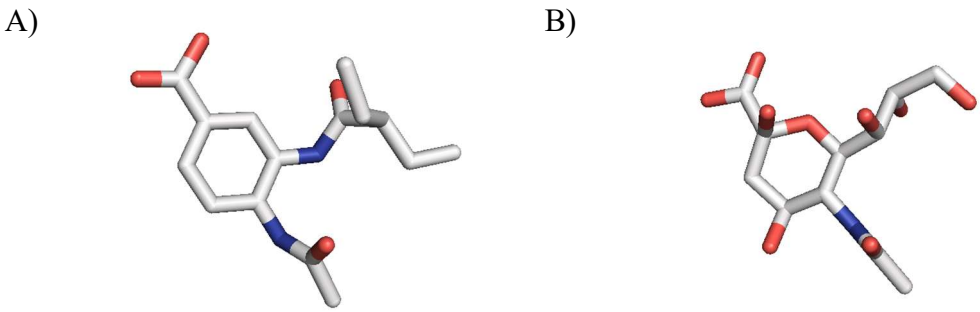


Figure 2-17. The structures of molecules binding to Neuraminidase with pdb codes: 1b9s (A) and1nsc (B).

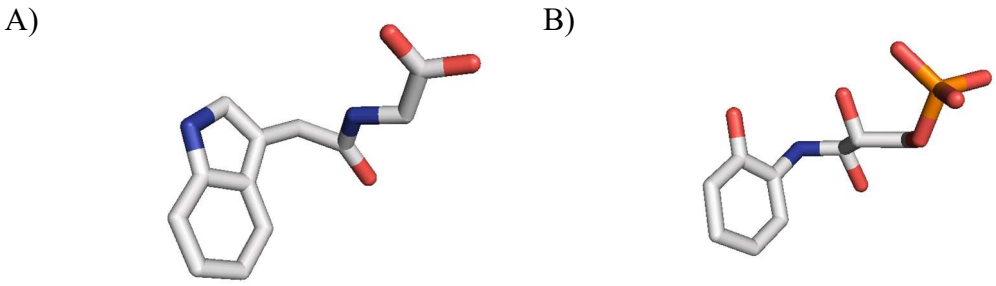


Figure 2-18. The structures of molecules binding to Tryptophan Synthase with pdb codes: 1k7e (A) and1tjp (B).

The combination of shape and 2D fingerprints was also investigated. Logistic regression (using the combination of Test Set 1 and Test Set 2) linked values of FT (for shape fingerprints) and Similarity Tanimoto (for 2D fingerprints) with the likelihood of shared activity. When two molecules are compared, the highest probability (either shape or 2D) was selected in each case. In this way, the calculated AUC values improved for Test Set 1 to 0.74 and Test Set 2 to 0.94. The two methods

provide useful complementarity and combine the ability to make useful connections between molecules with shared chemical structures and those with shared shape and thus permit both clustering and scaffold hopping.

2.6. Conclusions

The chapter details the implementation and validation of the shape fingerprints method, a promising method for describing and comparing molecular shape. The method is able to distinguish subsets of compounds that share biological activity with good levels of accuracy, considering that only the shape of molecules is considered; no other features are represented in these calculations. The obtained AUC values were 0.64 and 0.85 for Test Set 1 and Test Set 2, respectively. This suggests that shape is a strong influence on biological activity, as envisaged by the lock-and-key concept. Shape fingerprints are a useful method to apply this concept and are able to group compounds that are likely to share biological activity. The AUC analysis (as well as examination of logistic regression plots) has permitted the identification of the best performing Shape Database: SD10. The optimum settings involve DT set to 0.65 and BOV to 0.60. The Shape Database 10 performs well when crystal structures are used but also in case of conformations of ligands generated form SMILES with AUC values of 0.61 and 0.61 (AV and MV method respectively) for Test Set 1and 0.77 and 0.78 for Test Set 2, which is comparable with 2D fingerprint methods. However, the ability of shape fingerprints to find molecules similar in shape which could not be found using 2D fingerprints (the different chemistry) shows great potential in scaffold hopping.

The best Shape Database can be accessed via our GitHub repository: https://github.com/LeachResearchGroup/ShapeFingerprints.

Chapter 3

Shape Multipoles

3.1. Introduction

Shape multipoles[71] is a fast computational method that can be used to describe the shape of compounds using only numbers. This comes with many advantages including low storage needs and fast comparisons that require only simple mathematical operations. The shape multipoles[71] method describes the distribution of a molecule's volume using centroids and multipoles (monopole, dipoles, quadrupole, octupole moments), which are computed using a Gaussian description of a molecule.[70] Like in physics, where the electric dipole moment is used to describe the distribution of charge within a system, as shown in **Equation 3-1**, the analogous equation can be written based on Gaussian density (**Equation 3-2**).

Equation 3-1. the electric dipole moment, defined as the first order of multipole expansion, where ρ^{elec} is electrostatic charge density and r is the Cartesian coordinates of a point.

$$p = \int r\rho^{elec}(r)dr$$

Equation 3-2. The first order term in the shape multipole expansion, where V is defined as the Gaussian volume, r_α, r_β, r_γ, etc. are the Cartesian coordinates of a point and ρ^g is the Gaussian density of a molecule.

$$S_\alpha^{(1)} = \frac{1}{V} \int r_\alpha \rho_M^g (r) dr$$

The Gaussian volume from **Equation 3-2** is the zeroth order term describing the shape (monopole) and can be defined as in **Equation 3-3**, showing its independence from a change in origin for the coordinate system.

Equation 3-3. The zeroth moment, the Gaussian volume.

$$V = \int \rho_M^g (r) dr$$

However, the higher order moments should not be defined before specifying the origin of the molecule, its centroid. The first order moment (**Equation 3-2**) can be therefore adjusted (**Equation 3-4**) to indicate its dependency on the centre of the molecule. The choice of $S^{(1)}$ as an origin with coordinates R=(X,Y,Z), leads to vanishing of the first order moment. Now, having defined the centroid of the molecule, there can be defined higher order moments.

Equation 3-4. The first order moment.

$$S^{(1)\prime} = \frac{1}{V} \int r' \rho_M^g (r) dr = S^{(1)} - R$$

Higher order terms can be defined as in **Equation 3-5**, which can be simply transformed into second (**Equation 3-6**) and third (**Equation 3-7**) order of the shape multipole expansion, which are known as the shape quadrupole and shape octupole, respectively.

Equation 3-5. Shape Nth order multipole.

$$S_{\alpha_1 \alpha_2 \dots \alpha_n}^{(n)} = \frac{1}{V} \int r_{\alpha_1} r_{\alpha_2} \dots r_{\alpha_n} \rho_M^g (r) dr$$

Equation 3-6. Shape Quadrupole, the second order moment.

$$S_{\alpha\beta}^{(2)} = \frac{1}{V}\int r_\alpha r_\beta \rho_M^g\,(r)dr$$

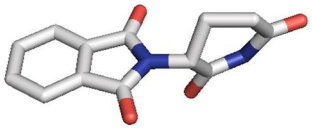
Equation 3-7. Shape Octupole, the third order moment.

$$S_{\alpha\beta\gamma}^{(3)} = \frac{1}{V}\int r_\alpha r_\beta r_\gamma \rho_M^g\,(r)dr$$

As can be seen in the **Figure 3-1** and also based on the **Equation 3-6**, the quadrupole's components (Q_x, Q_y, Q_z) describe how the matter is distributed along the three orthogonal axes such that linear molecules (such as 1,3,5-hexatriyne) have one large component, flat molecules (like benzene) have two large components and spherical molecules (like Buckminsterfullerene) have three large components. Their octupole moments are really small or equal to zero. This suggests that octupoles are more appropriate to identify asymmetric spatial distribution of shape, being almost neglected in symmetric molecules. However, the shape quadrupoles look the same for two enantiomers. As shown in **Figure 3-1**, some of the octupole components for (*R*)-(+)-thalidomide (E) and (*S*)-(−)-thalidomide (F) have the same value but different sign, which indicate the differences in shape of those two enantiomers and indicates that shape octupoles describe some of the unsymmetrical distribution of matter in a molecule.

A)	B)	C)
		
<Quadrupoles> 4.74, 4.68, 4.66 <Octupoles> 0.01, 0.00, -0.00, -0.00, 0.00, -0.00, -0.00, -0.01, 0.00, -0.00	<Quadrupoles> 1.59, 1.59, 0.62 <Octupoles> 0.00, 0.00, -0.00, -0.00, -0.00, 0.00, 0.00, -0.00, -0.00, -0.00	<Quadrupoles> 5.39, 0.62, 0.62 <Octupoles> 0.00, 0.00, 0.00, 0.00, -0.00, 0.00, 0.00, 0.00, -0.00, -0.00

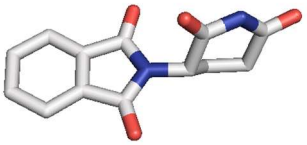
D)



<Quadrupoles>
8.78, 1.63, 1.15

<Octupoles>
1.81, 0.30, 0.29, -1.53, -0.73, -1.39, -0.08, 1.17, 0.01, -0.24

E)



<Quadrupoles>
8.78, 1.63, 1.15

<Octupoles>
1.82, 0.30, -0.29, -1.52, 0.71, -1.39, 0.08, 1.17, 0.01, 0.24

Figure 3-1. Example of calculated shape quadrupoles for Buckminsterfullerene (A) with three almost equal quadrupole components, benzene (B) with two components slightly greater than the third and hexa-1,3,5-triyne (C) with only one outstandingly higher quadrupole component and shape octupoles for (R)-(+)-thalidomide (D) and (S)-(−)-thalidomide (E).

3.2.Components of shape multipoles

Shape multipoles were computed using the Shape toolkit provided by Openeye[78] for a few chosen molecules from Test Set 1 (as described in chapter 2). The reason for this was to examine the components of shape quadrupoles and their ability to distinguish similar molecular shape.

The shape multipoles method performs surprisingly well considering the amount of components it includes. Hence, the components of shape quadrupoles and octupoles for molecules with similar shape, shown in **Figure 3-2** were compared and the values were stored in **Table 3-1**. This suggests that the components of shape quadrupoles are capable of characterizing the broad shape features of molecules. The corresponding components do not deviate much from each other for very similar shapes of molecules. However, shape octupoles are not so straightforward to interpret but likely carry more specific and accurate information about the shape of compounds.

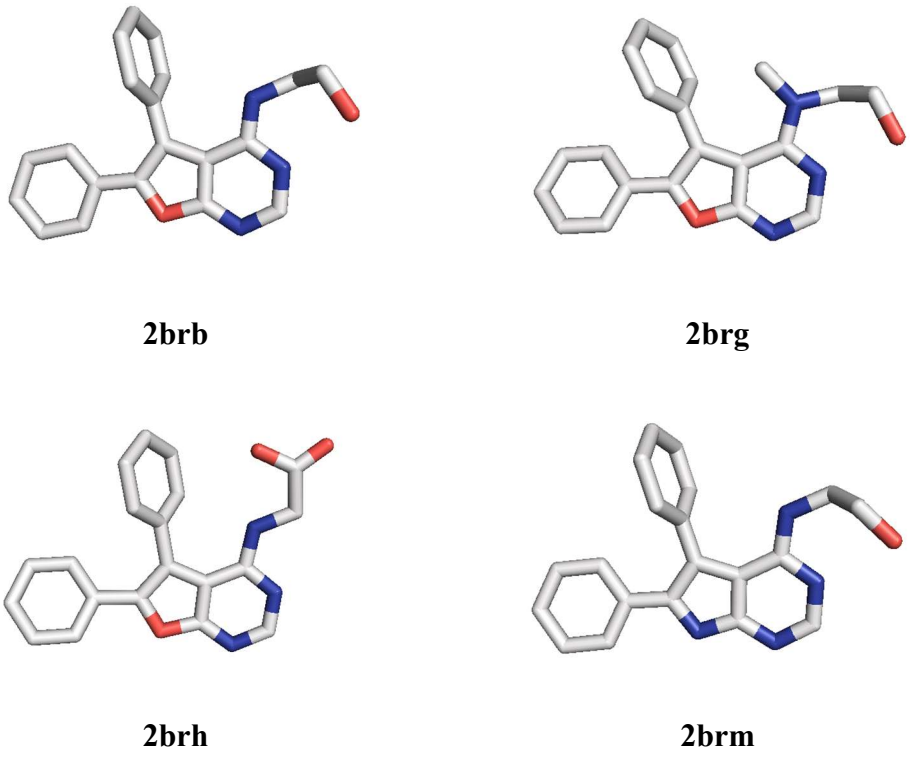


Figure 3-2. The chemical structures of selected Checkpoint kinase (Chk1) ligands.

Table 3-1. Shape quadrupoles for some ligands that bind to Checkpoint kinase (Chk1).

Ligand	Q1	Q2	Q3
2brb	9.25	4.40	0.91
2brg	9.73	4.17	0.89
2brh	9.90	4.75	0.78
2brm	9.26	4.25	0.92

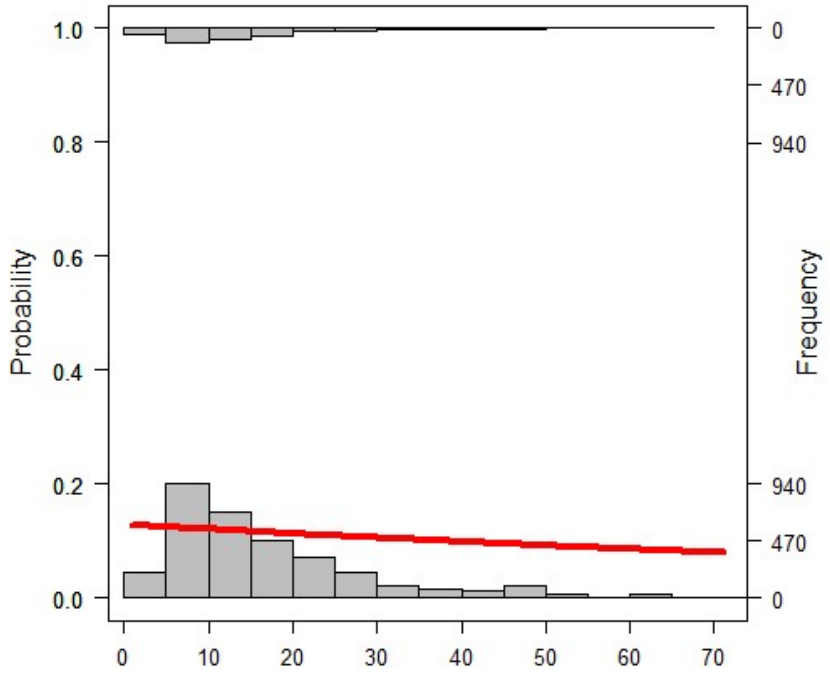
3.3. Test Sets

To evaluate the performance of shape multipoles, we used three test sets: 1) a set described by Taylor et al.[79], which was devised to test pharmacophore models and consists of 10 sets of ligand-protein complexes 2) a group from the Astex diversity set[80], which includes 45 molecules binding to 4 selected proteins and 3) a set of 4 groups of protein-ligand complexes from a set described by Head et al.[87]

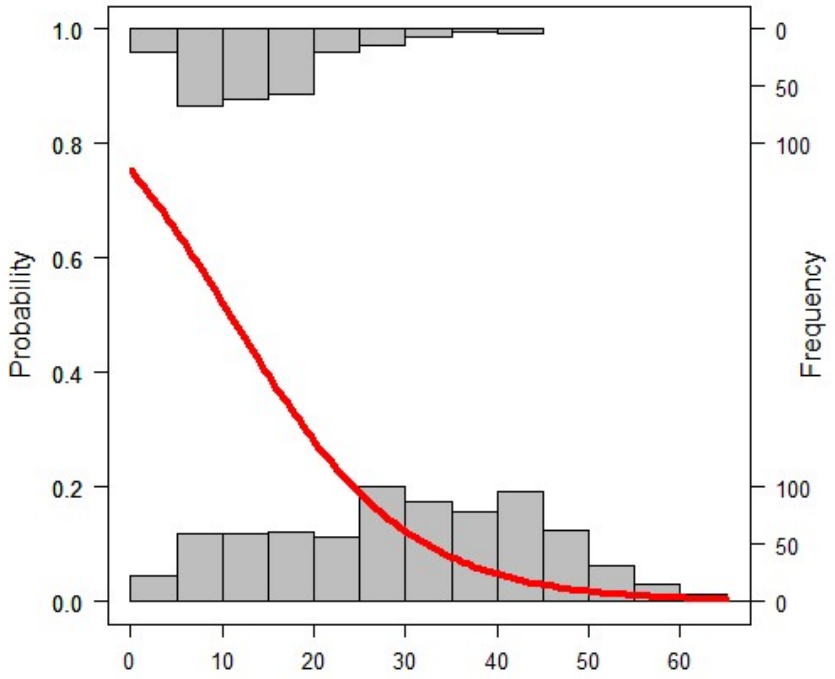
Shape octupoles were generated for every compound in three test sets. Molecules were compared with each other based on the calculated multipoles. The Euclidean distance between shape octupoles were used to determine the similarity between the shapes of each structure. Euclidean distance is equal to 0 for the most similar compounds but has no upper bound for dissimilar molecules.

The performance of the method was evaluated using logistic regression and AUC values from ROC plots, which were generated using R.[82] Logistic regression was used to link the Euclidean distance with the proportion of structures that bind to the same protein, which corresponds to the likelihood of the two molecules sharing biological activity. Ideally, the Euclidean distance values would be low for all molecules that bind to the same protein and high for compounds which do not share activity.

A)



B)



C)

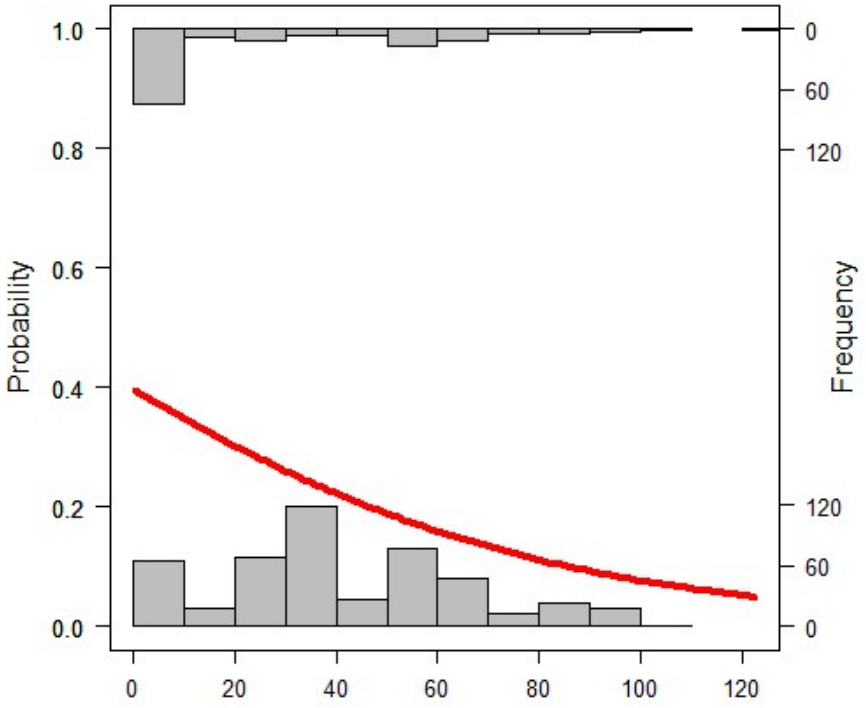
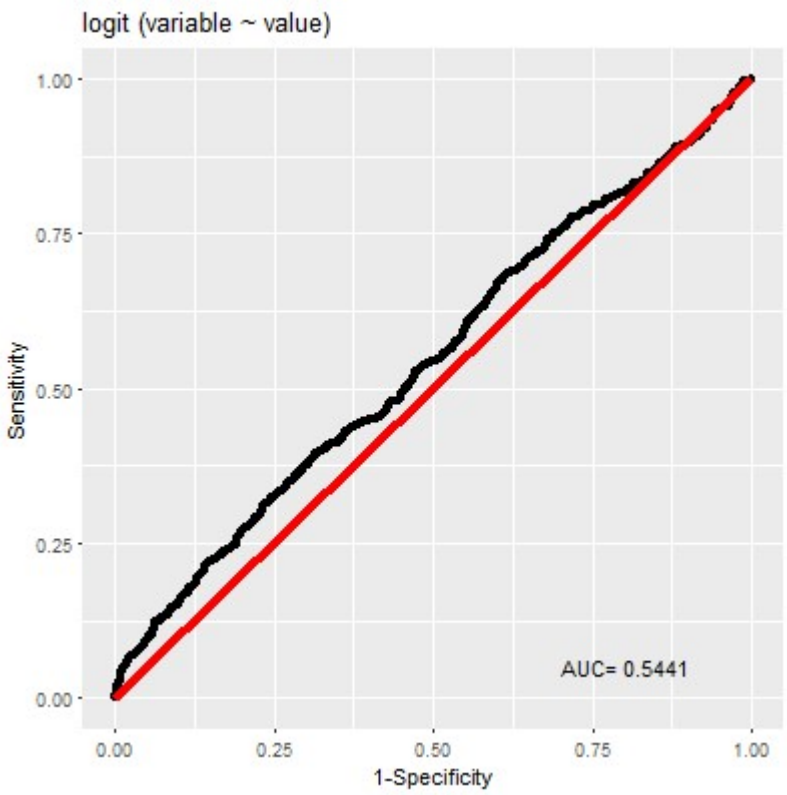


Figure 3-3. Logistic regression plots for 3 Test Sets when using ΔOctupoles: Test Set 1 (A), Test Set 2 (B), Test Set 3 (C).

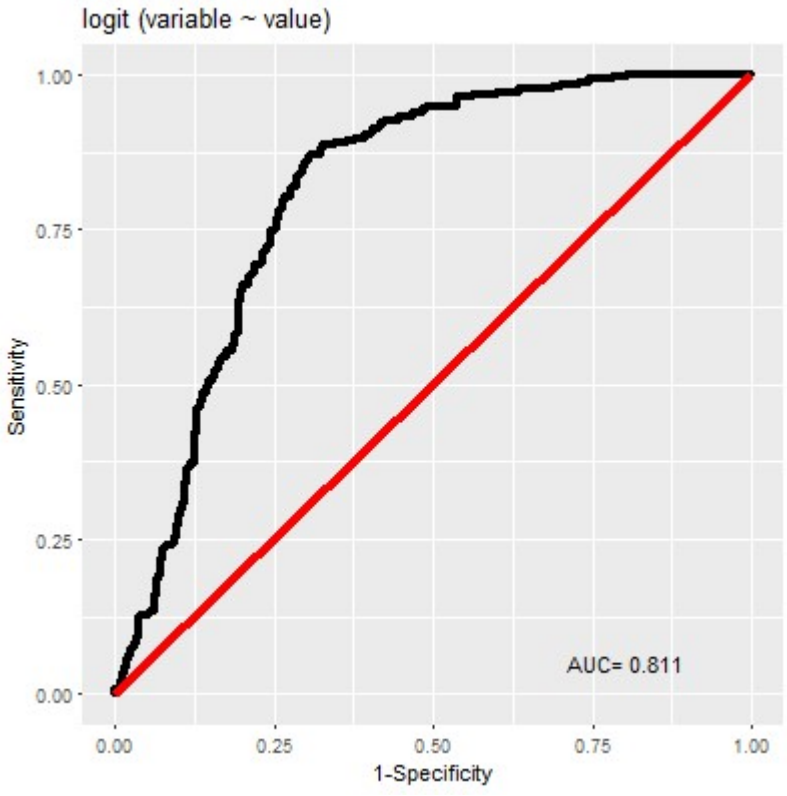
The plots in **Figure 3-3** show the distribution of Euclidean distance values for molecules that share activity (upper histogram) and those that do not (lower histogram). The logistic regression plot for shape octupoles suggests that they do not

provide a good distinction. Only in case of Test Set 2, the probability of two molecules that bind to same protein having low ΔOctupoles reaches almost 80%. However, that might be the result of the wide range of the Euclidean distance values.

A)



B)



C)

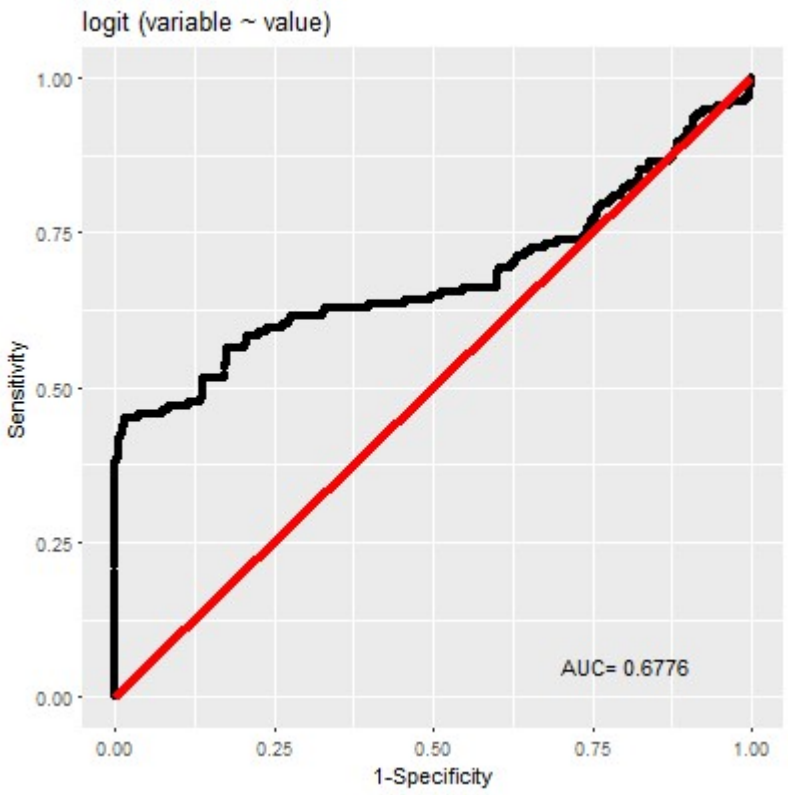


Figure 3-4. ROC curves for 3 Test Sets when using ΔOctupoles: Test Set 1 (A), Test Set 2 (B) and Test Set 3 (C).

The ROC curve is a fundamental tool for diagnostic test evaluation. In a ROC curve the true positive rate (Sensitivity) is plotted as a function of the false positive rate (100-Specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve (AUC) is a measure of how well a method is able to distinguish molecules that share biological activity from those that do not. A perfect method would achieve an AUC value of 1, a completely random method 0.5. Test Set 1 results in an AUC value of 0.54, which is not much better from a random distribution. Higher results were obtained for Test Set 2 – 0.81 and Test Set 3 – 0.68. This is slightly lower than the results obtained using the shape fingerprint method: 0.64 and 0.85 for Test Set 1 and 2, respectively. Test Set 3 was not used with the shape fingerprint method as it produced bit strings with too low (or even zero) bit density, therefore cannot be compared here.

3.4. Enantiomers

In order to check whether the shape multipole method could be applicable to find differences in shape between enantiomers, three sets were used that were extracted from the CHEMBL database[88] based on their measured IC_{50} values: the ligands of acetylcholinesterase (AChE), human ether-a-go-go-related gene potassium channel 1 (hERG) and dipeptidyl peptidase IV (DPP-IV). These comprised 448, 190 and 136 pairs of enantiomers, respectively.

Conformations of all molecules were generated by OMEGA[83] from canonical SMILES with default settings. The structures that failed the generation process due to unspecified stereochemistry had been rejected. All the conformations were optimized using Szybki[78] before calculating shape multipoles.

3.4.1. Human ether-a-go-go-related gene potassium channel 1 (hERG)

The set of compounds, inhibitors of hERG, was downloaded from the CHEMBL database[88] and were selected by the requirement to have a measured IC_{50}. The enantiomer pairs were found by text manipulation. Initially, the set consisted of 448 pairs of enantiomers out of which, duplicates and those without any IC_{50} values have been discarded from the further analysis leaving 105 distinct pairs. For those pairs the shape multipoles were calculated, both shape quadrupoles and octupoles. As mentioned previously, the shape quadrupoles for any enantiomer pairs are either the same or very similar, the further analysis was focused on comparing the shape octupoles of each pair. The plot in the **Figure 3-5** shows the relationship between the difference in octupoles and the change of pIC_{50} for each pair of compounds.

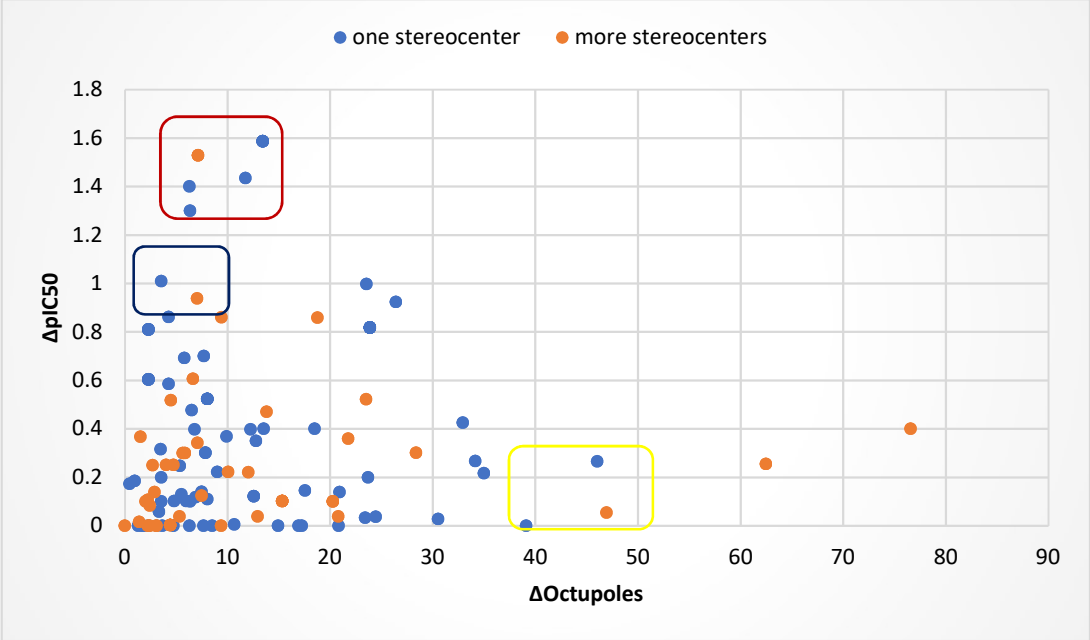
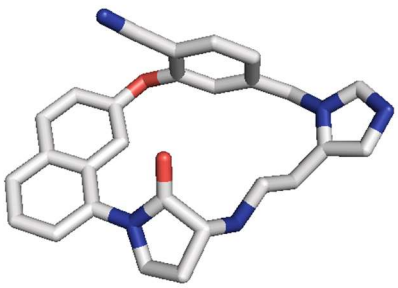


Figure 3-5. The graph showing the ΔpIC50 as a function of ΔOctupoles for enantiomer pairs from hERG dataset.

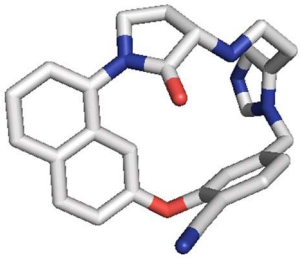
In the **Figure 3-5**, it is noticeable that the division between pairs with only one stereocentre and more than one is not clear. There are some points that particularly stand out and will be discussed below. Those are marked in red, yellow and blue circles in the plot in the **Figure 3-5**.

In the red circle in **Figure 3-5**, there are enantiomer pairs with high ΔpIC50 and a medium difference in octupoles. As shown in the **Figure 3-6**, the difference in the octupoles is not as high as would be expected based on the difference in pIC_{50} these pairs have quite similar shapes.

1A)



1B)



2A) 2B)

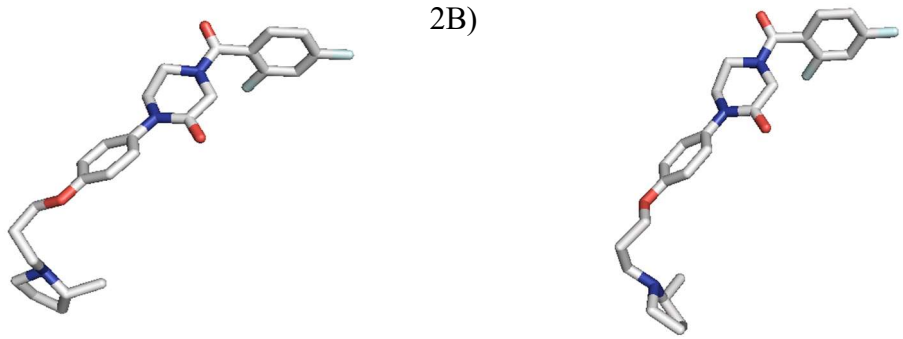


Figure 3-6. The structures of enantiomer pairs: CHEMBL498260 (1A) and CHEMBL55826 (1B); CHEMBL239299 (2A) and CHEMBL239724 (2B).

In the blue circle in **Figure 3-5**, there are enantiomer pairs with a medium ΔpIC50 and quite low difference in octupoles, which, as in the examples above, could be explained by similarities in overall shape of these particular enantiomer pairs. This can be seen in the **Figure 3-7**.

1A) 1B)

2A) 2B)

3A) 3B)

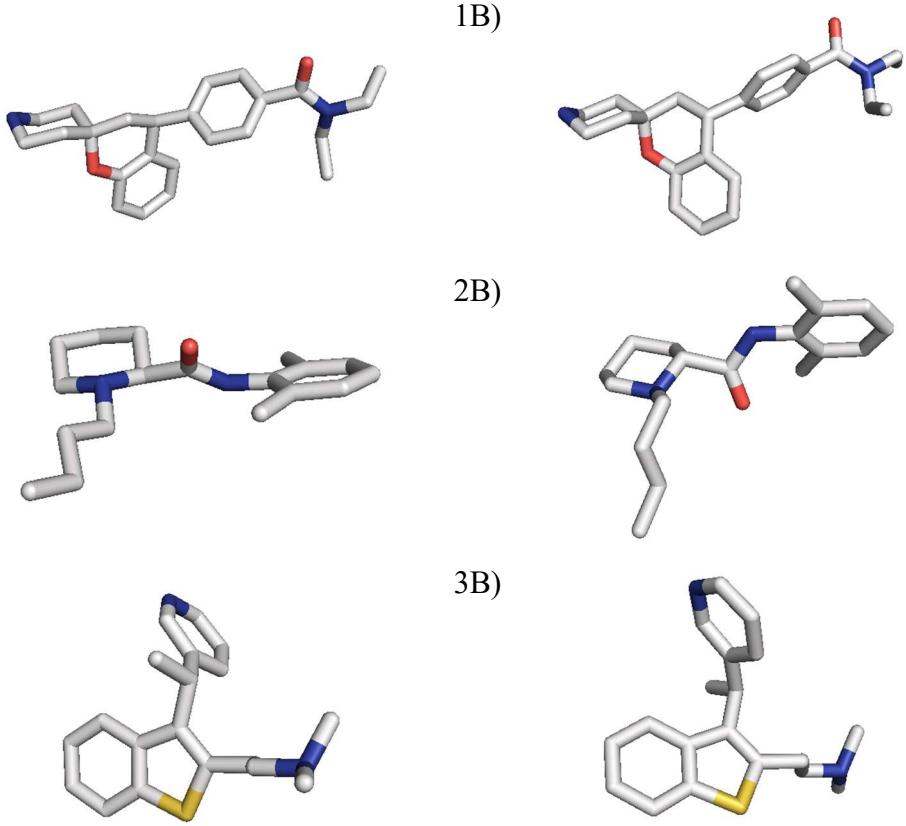


Figure 3-7. The structures of enantiomer pairs: CHEMBL550471 (1A) and CHEMBL556648 (1B); CHEMBL1200749 (2A) and CHEMBL2447962 (2B); CHEMBL1091777 (3A) and CHEMBL1091778 (3B).

In the yellow circle in **Figure 3-5**, are enantiomer pairs with low ΔpIC50 and high difference in octupoles. As shown in the examples in the **Figure 3-8**, the differences in shapes of these pairs are not that great as is indicated by the calculated octupole difference.

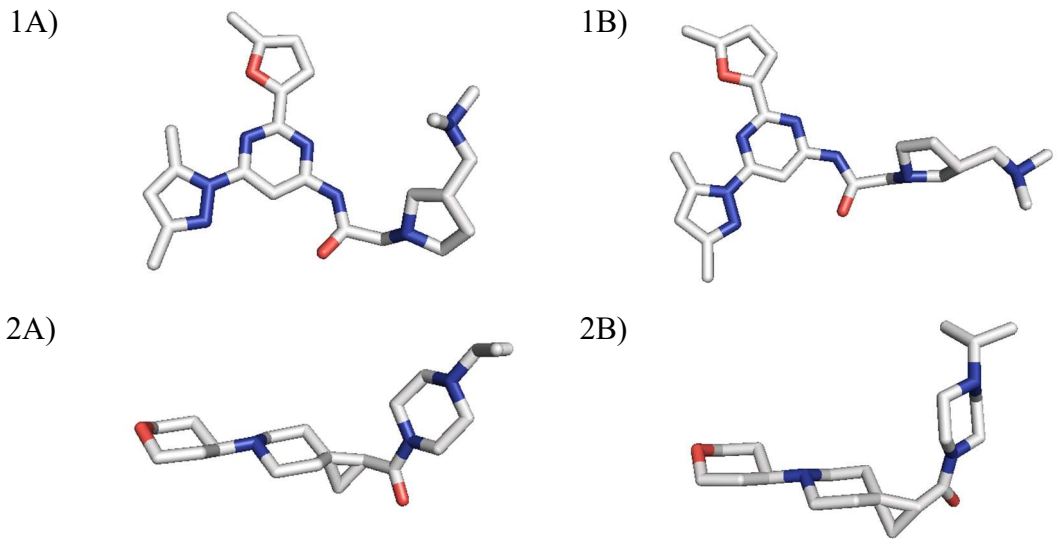
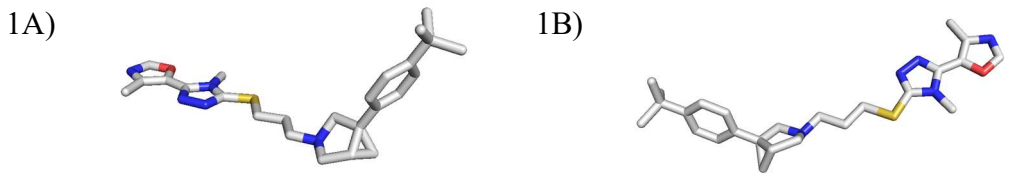


Figure 3-8. The structures of enantiomer pairs: CHEMBL272637 (1A) and CHEMBL429761 (1B); CHEMBL3124968 (2A) and CHEMBL3127672 (2B).

The most extreme values of ΔOctupoles are obtained for enantiomer pairs with more than one stereocentre. As can be seen in the **Figure 3-5**, there are two pairs with ΔOctupole equal to 76.58 and 62.46: CHEMBL1079823 and CHEMBL1079824 pair and CHEMBL2010844 and CHEMBL2010845 pair, respectively. This is mostly caused by the flexibility of the molecules - enantiomers have different shapes after optimization process, which results in high difference in octupoles.



2A) 2B)

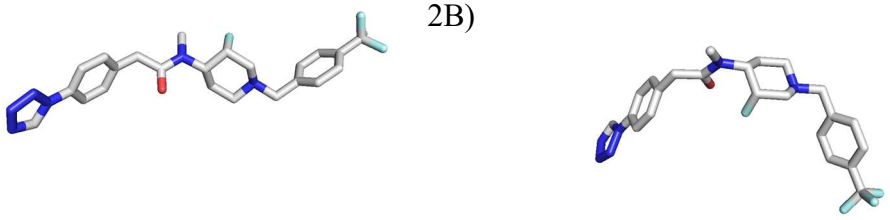


Figure 3-9. The structures of enantiomer pairs: CHEMBL1079823 (1A) and CHEMBL1079824 (1B); CHEMBL2010844 (2A) and CHEMBL2010845 (2B).

3.4.2. Acetylcholinesterase (AChE)

The set of AChE ligands from the CHEMBL database[88] consisted of 190 pairs, which were found by simple text manipulation. The pairs without specified IC_{50} values and those that were considered as duplicates have been removed, leaving 27 distinct pairs.

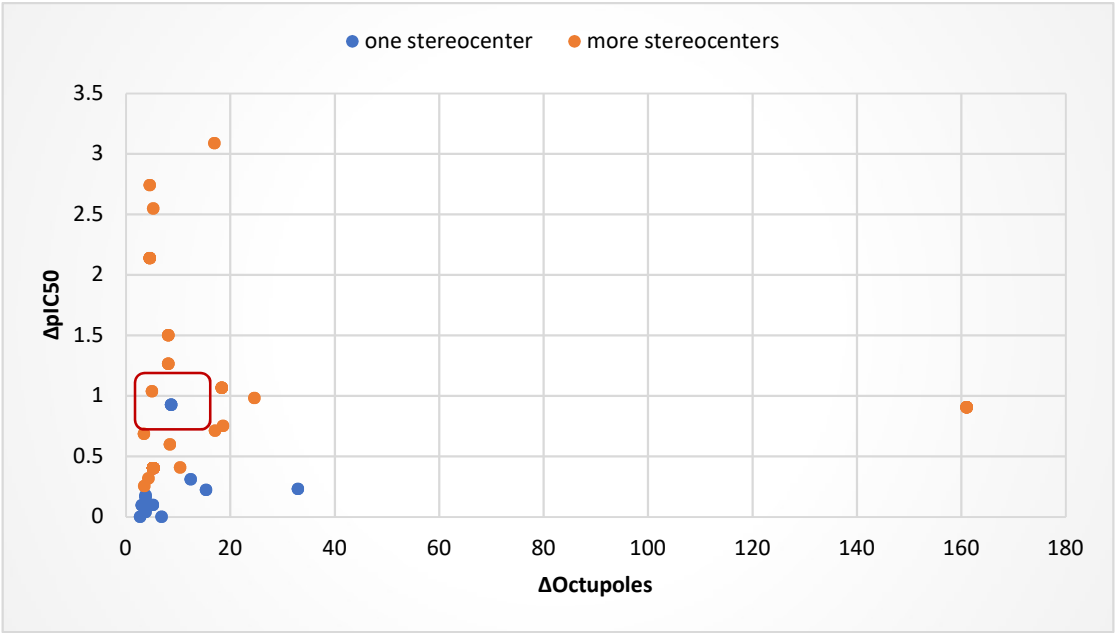


Figure 3-10. The graph showing the ΔpIC50 as a function of ΔOctupoles for enantiomer pairs from AChE dataset.

There is a clear distinction between pairs with one stereocentre and those with more than one stereocentre in the AChE set, visible in **Figure 3-10**. However, as the ΔpIC50 tends to have a higher value for pairs with more stereocentres, the difference in Octupoles does not grow linearly with it. There is one outstanding point on the plot,

marked in the red circle, with ΔpIC50 = 0.93 and ΔOctupoles = 8.68. The structures of the pair can be seen in **Figure 3-11**. It is worth noting that this particular pair has quite a high difference in Quadrupoles (1.09), which usually gives values much closer to 0 for two enantiomers.

A) B)

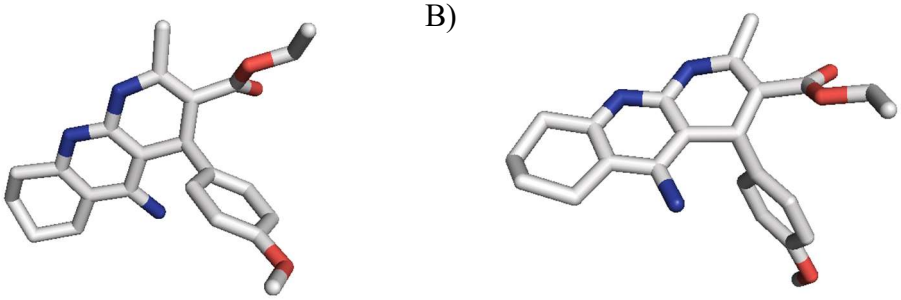
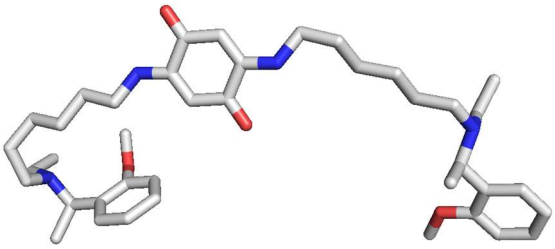


Figure 3-11. The structures of enantiomer pair: CHEMBL470715 (A) and CHEMBL490359 (B).

The most extreme value of ΔOctupoles is obtained for enantiomer pair with more than one stereocentre, as it can be seen in the **Figure 3-5**.The value of ΔOctupole is equal to 161.03 for CHEMBL540178 and CHEMBL556581 pair (**Figure 3-12**). Similarly as in other cases, the generated and optimized conformations differ in shape for each enantiomer due to flexibility of compounds and therefore comparing them resulted in high ΔOctupole value.

1A)



1B)

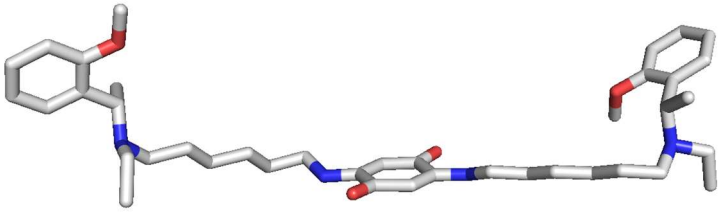


Figure 3-12. The structures of enantiomer pair: CHEMBL540178 (1A) and CHEMBL556581 (1B).

3.4.3. Dipeptidyl peptidase IV (DPP-IV)

The set of DPP-IV ligands initially consisted of 136 pairs, which were found by text manipulation. After removing duplicate pairs without specified bioactivity data, which was essential for further analysis, there were 34 distinct pairs left. Shape octupoles were calculated for them and the relation between the difference in octupoles and the difference in pIC_{50} was studied.

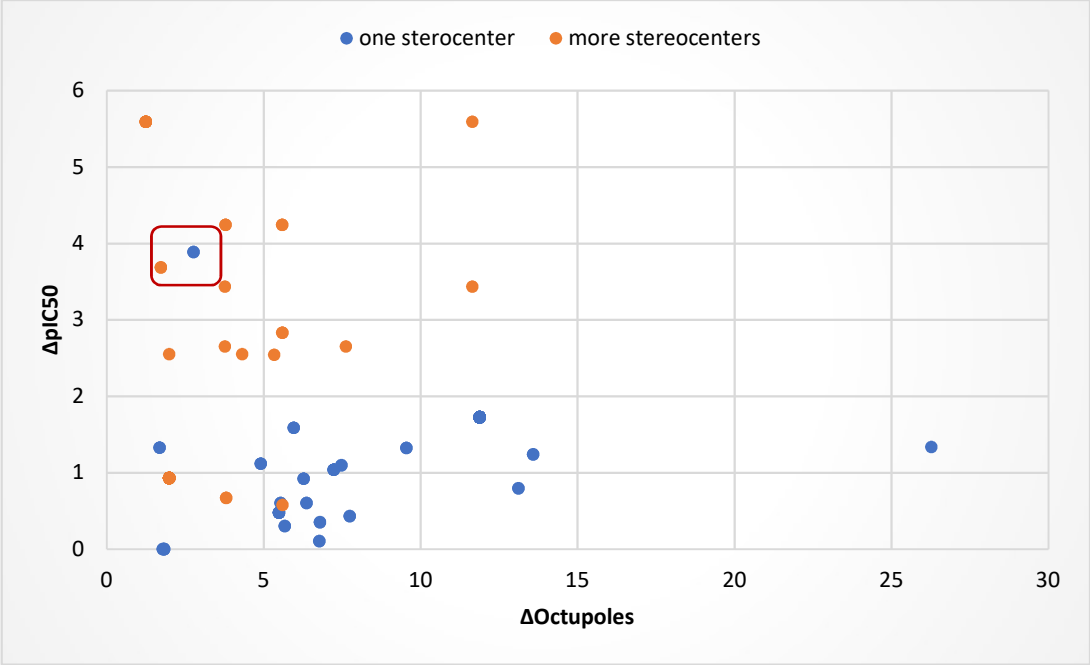


Figure 3-13. The graph showing the ΔpIC50 as a function of ΔOctupoles for enantiomer pairs from the DPP-IV dataset.

As in the case of the AChE set, the plot in **Figure 3-13** has a noticeable distinction between pairs with different numbers of stereocentres. With one exception, generally for pairs with only one stereocentre the difference in Octupoles grow almost linearly with the ΔpIC50.

A) B)

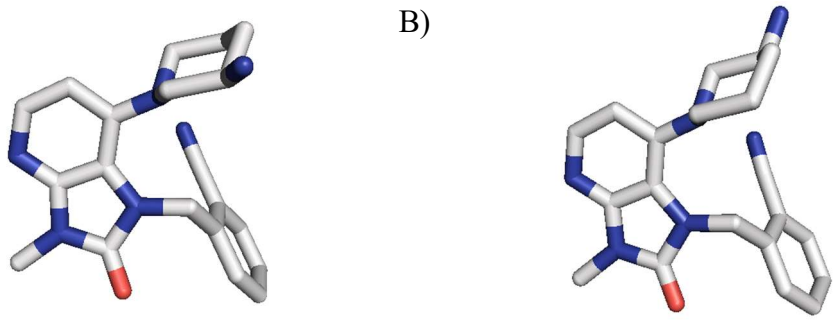


Figure 3-14. The structures of enantiomer pair: CHEMBL3329627 (A) and CHEMBL3329694 (B).

The exception is the pair of enantiomers marked in red box on the plot in **Figure 3-13**, with the structures visible in **Figure 3-14**.

The highest value of ΔOctupoles (26.27) has a pair of enantiomers: CHEMBL428936 and CHEMBL399348. Such difference could be a result of flexibility of cyclic rings (**Figure 3-15**).

1A) 1B)

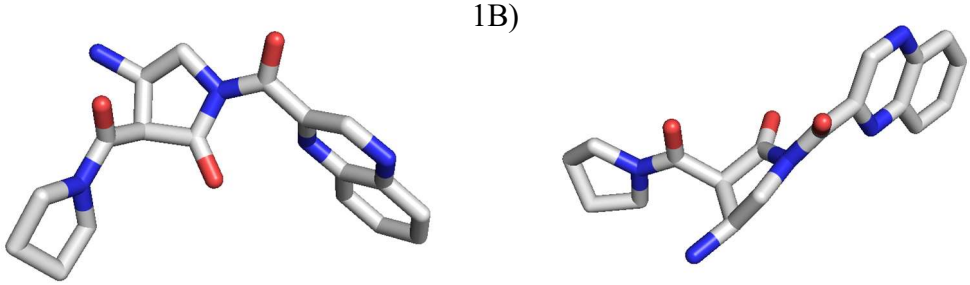


Figure 3-15. The structures of enantiomer pair: CHEMBL428936 (1A) and CHEMBL399348 (1B).

3.5. Conclusions

The shape multipole method is a fast computational method to describe the shape of molecules by using only numbers and therefore it requires low storage needs and comparison is performed by simple mathematical operations. To describe the shape, it uses only 13 values (3 quadrupole components and 10 octupole components). While the quadrupole components describe the distribution of matter in a system along the axes x, y, z, and therefore do not contain too specific information, the octupole

components describe the deviations of matter from axes in a system more accurately and thus can explain the differences in shape and activity between enantiomers.

The shape multipoles method performs surprisingly well in grouping the compounds based on shared biological activity, considering the amount of components it includes. The obtained AUC values of 0.54, 0.81 and 0.68 for Test Set 1, Test Set 2 and Test Set 3, respectively are slightly lower than the results obtained using the shape fingerprint method: 0.64 and 0.85 for Test Set 1 and 2, respectively (Test Set 3 was not used).

The investigation of using shape multipoles in order to find differences in shape between enantiomers showed potential, however requires better comparison metrics in order to be more effective.

Chapter 4

Application of shape fingerprints

4.1. Introduction

In order to be useful, the method needs to show a good performance in addressing some scientific problems. Without applying it, and having proved good performance, to real problems that affect the chemistry world, it is hard to recommend the approach as a solution. This would diminish the likelihood of the shape fingerprint method being used. Consequently, in this chapter a few of the possible applications of the shape fingerprint method will be explored. These include solubility predictions, virtual screening or simply grouping compounds with shared biological activity.

4.2. DUD-E diverse set

The quantitative assessment of performance of a lot of computational methods remains challenging.[89] The Directory of Useful Decoys (DUD) was designed to measure how known ligands rank versus a set of decoy molecules.[90] Here, the set from the DUD-E database was used in an alternative way. The collection of decoys was not included and only actives were taken into consideration. The idea behind this was to investigate the ability of the shape fingerprints method to group compounds binding to different targets, similarly as in chapter 2, where the validation of the method was based on the

ability to group compounds with similar biological activity. However, the set presented here is much larger than used previously. This should examine how well the approach works on various size sets with more shape diversities.

4.2.1. Results

The DUD-E diverse set consists of 8 targets: serine/threonine-protein kinase AKT (AKT1), beta-lactamase (AMPC), cytochrome P450 3A4 (CP3A4), C-X-C chemokine receptor type 4 (CXCR4), the glucocorticoid receptor (GCR), human immunodeficiency virus type 1 protease (HIVPR), human immunodeficiency virus type 1 reverse transcriptase (HIVTR) and kinesin-like protein 1 (KIF11). It consists of 290, 48, 166, 39, 258, 527, 330 and 116 actives, respectively (which is 1774 ligands in total) in AKT1, AMPC, CP3A4, CXCR4, GCR, HIVPR, HIVRT and KIF11, respectively.

The structures of molecules were taken from the DUD-E webpage,[90] provided as structure-data files (SDF). The shape fingerprints were generated for each structure using Shape Database SD10 with DT = 0.65. A bit On value equal to 0.60 was applied, exactly as suggested in the previous chapter to maximize the performance of the method. The ability to group ligands was analysed based on plots (ROC curve and logistic regression) produced in R.[82]

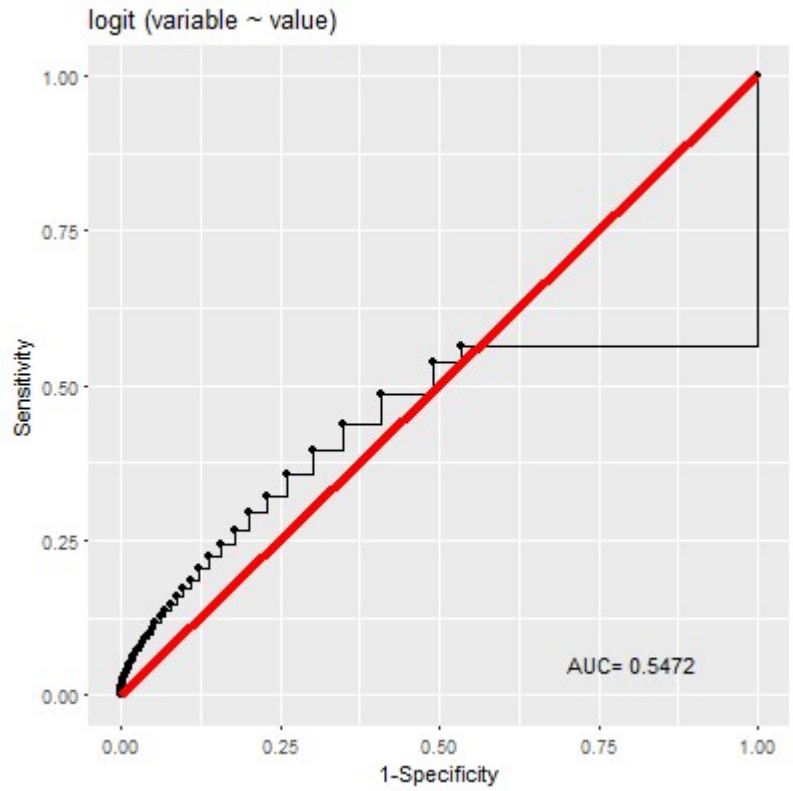


Figure 4-1. The ROC curve for the DUD-E diversity set.

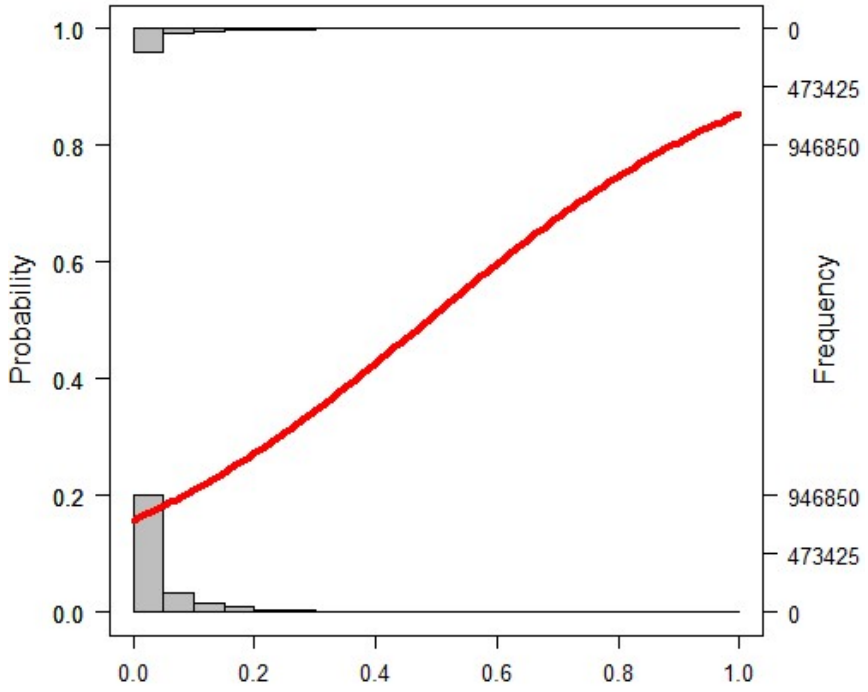


Figure 4-2. Logistic regression plot for the DUD-E diversity set.

As shown in the **Figure 4-1**, the AUC value calculated for the DUD-E diversity set is equal to 0.55. This is a little bit lower value than the ones obtained in previous study in chapter 2 (0.64 and 0.85 for Test Set 1 and Test Set 2, respectively). However, the size of the set presented here is much greater and therefore a value in this range could have been expected. The large jump in the ROC curve is possibly caused by a high percentage of compounds with dissimilar shape yet shared biological activity. The probability curve from the logistic regression plot, in the **Figure 4-2**, reaches the value of 0.8 and it reveals that for FT above 0.5 there is greater than 50% chance of shared biological activity. This behaves slightly better than in the case of Test Set 1 used in evaluation of shape fingerprint method and worse than Test Set 2 described in chapter 2.

4.3. Virtual Screening

Virtual screening is a computational technique used in drug discovery to search libraries of small molecules in order to identify those structures which are most likely to bind to a drug target, typically a protein, receptor or enzyme.[90,91] The virtual screening techniques are rated based on their ability to retrieve a small group of actives from a large collection of structures with similar physicochemical properties but dissimilar 2D topology – decoys.[90]

4.3.1. Results

As virtual screening is a common technique in the drug design process,[29] it was of great interest to check the performance of the shape fingerprint method in it. With the purpose to test whether the shape fingerprint method is able to distinguish the molecules that are active from those that are not, three sets from DUD-E were chosen.[90] These sets include: C-X-C chemokine receptor type 4 (CXCR4), Beta-lactamase (AMPC) and Catechol O-methyltransferase (COMT). The sets contain 40, 48 and 41 actives, respectively and a series of decoys. The shape fingerprints were generated for these sets using Shape Database SD10 with DT = 0.65 and BOV = 0.60. The molecules were compared with each other, resulting in Fingerprint Tanimoto

values. The values were grouped into those arising from comparison of active compounds and those from the comparison of inactives. Based on that, ROC curves were produced for the three sets using R studio.[92]

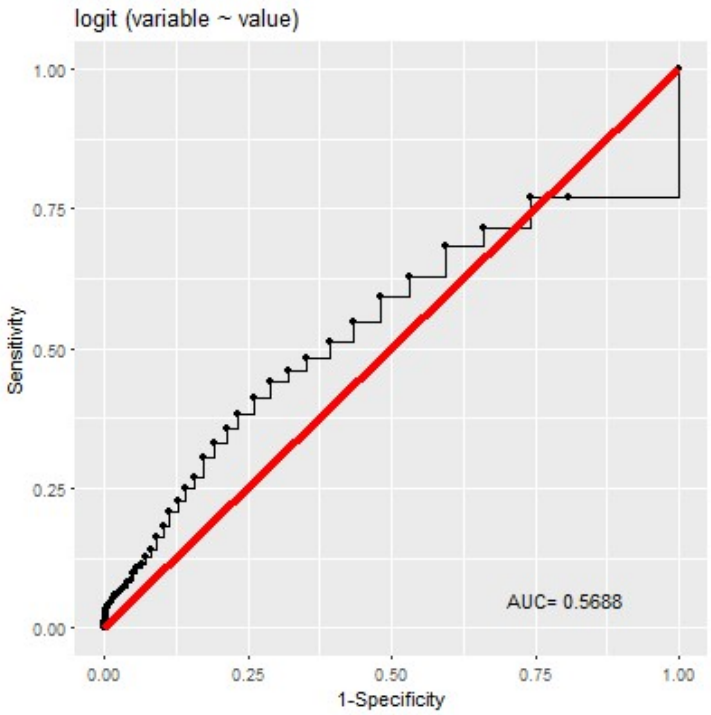


Figure 4-3. The ROC curve for the CXCR4 set.

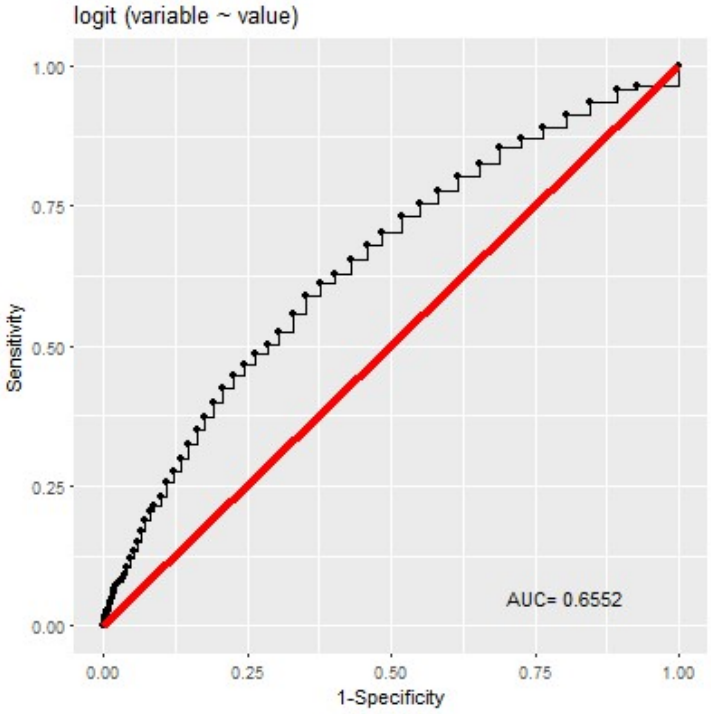


Figure 4-4. The ROC curve for the AMPC set.

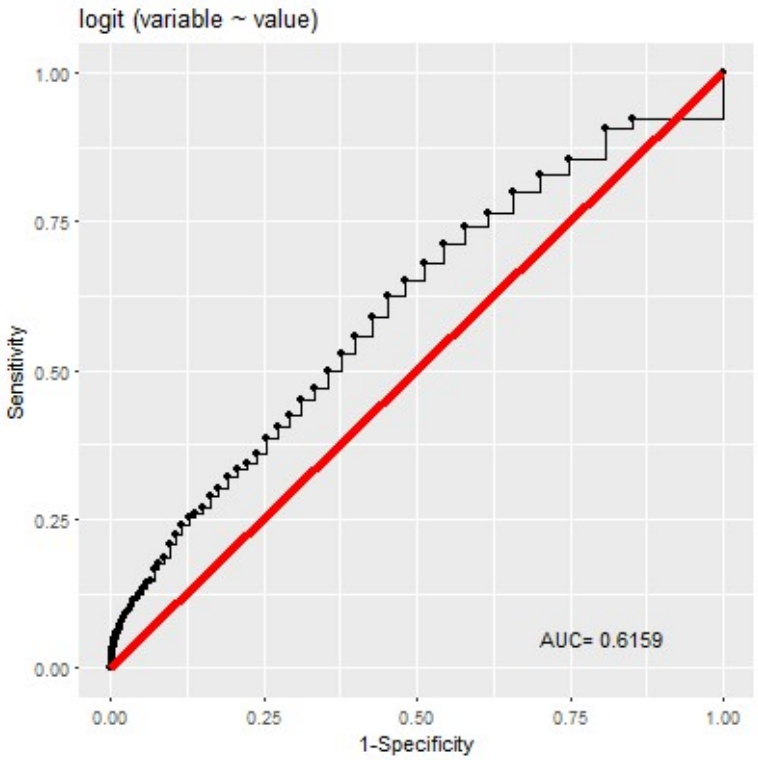


Figure 4-5. The ROC curve for the COMT set.

The shape fingerprint method performed relatively well, considering the size of the sets and the high ratio of decoys per ligand. The AUC values obtained from the ROC curve are 0.57, 0.66 and 0.62 for CXCR4, AMPC and COMT sets, respectively. This gives opportunities that shape fingerprints could be successfully used as a ligand-based virtual screening technique.

4.4. Aqueous Solubility

Solubility prediction has a great importance in the pharmaceutical industry because of its implications in the formulation of drugs and in drug absorption.[93] The solubility of compounds depends on their physical and chemical properties. The interaction of solutes with water and its crystallinity play important roles in determining the solubility of a compound.[94]

Many methods have been developed to estimate aqueous solubility using various physical properties of compounds. The most often used, and also the simplest model, is the general solubility equation (GSE), as in **Equation 4-1**, which was proposed by

S. Yalkowsky.[95,96,97] The equation shows that the solubility of a compound can be calculated from the melting point (MP) and octanol-water partition coefficient (logP) of a compound. The logP values can be either obtained experimentally or calculated using scientific software e.g. MOE.[98]

Equation 4-1. General Solubility Equation, where $S_{W\ is}$ aqueous solubility, MP is the melting point and logP is the octanol-water partition coefficient of the compound.

$$log\ S_W = 0.5 - 0.01(MP - 25) - logP$$

This model, however, requires experimental data and therefore cannot be applied to molecules without such data or new compounds for which those values have not been measured, yet. Therefore, computational methods able to accurately predict solubility are in high demand. Based on the GSE, instead of using melting point which is dependent on the shape of molecules (to the extent that they influence packing in the solid state), it was of great interest to apply shape fingerprints.

4.4.1. Results

All the data for solubility studies were collected from various databases and literature. The data includes information from large solubility datasets: PhysProp,[99] Reaxys Databases[100] and Yalkowsky's Handbook of Aqueous Solubility,[101] as well as many literature published solubility datasets.[102–116] For molecules without any SMILES, these were generated using ChemCell,[117] ChemSpider,[118] ChemDraw or ChemIDPlus.

The dataset contained 102821 measurements of solubility and/or melting points. After taking only unique SMILES strings the number reduced to 94955. Measurements of solubility or melting points that had "less than" or "more than" prefixes were excluded from the dataset and the highest value in ranges of melting point was taken and the lowest value for solubility. The reason for that was to allow prediction based on the most stable polymorph.

4.4.1.1. Training and Test Set to predict solubility – Shape Database 10

The dataset was divided into 1) a set containing logS values 2) compounds having measured MP values 3) compounds with measured values of both logS and MP. The latter set with measured values of both logS and melting point was used as the most suitable for the solubility prediction and its comparison to the GSE model. The set of compounds (945 molecules) was divided into two sets: training (90%) and test set (10%). The shape fingerprints were generated for both sets (training and test) using SD10 and DT= 0.65 and BOV = 0.60. This Shape Database was suggested as producing the best results in chapter 2. The independent values were selected to be calculated logP (computed in MOE)[98] and all the bits from the generated shape fingerprint string. The applied method was 'Enter', which means that all independent variables are entered into the equation in one step. The model built using the training set was stored (**Equation 4-2**, **Table 4-1**) and used on the 95 molecules of the test set to predict their logS values. The model built with clogP and experimental MP values was used as a benchmark (adjusted R^2 = 0.612). The obtained results were compared to experimental values of logS as well as those predicted in MOE[98] in the **Table 4-2**.

Equation 4-2. The linear regression model used for prediction of solubility. The coefficients for each bit can be found in the table below.

$$log\ S_W = -2.140 - 0.427logP + 0.450Bit1 - 0.038Bit10 + \cdots$$

Table 4-1. The table of constant and coefficients used for solubility prediction. Some bits (not listed here) are constants or have missing correlations.

MODEL	UNSTANDARDIZED COEFFICIENTS	
	B	Std. Error
(CONSTANT)	-2.140	0.259
LOGP	-0.427	0.030
BIT1	0.450	0.565

BIT10	-0.038	0.244
BIT12	15.216	13.595
BIT14	-0.433	6.037
BIT17	55.307	22.637
BIT18	-0.716	1.652
BIT22	-16.848	12.075
BIT24	14.071	27.932
BIT26	-1.613	6.214
BIT30	0.066	0.199
BIT32	0.085	0.195
BIT36	0.352	0.221
BIT40	0.498	0.248
BIT45	-59.117	23.955
BIT50	4.157	5.203
BIT51	108.107	37.984
BIT52	-3.234	3.443
BIT54	-1.072	2.497
BIT56	0.038	0.621
BIT58	-0.538	0.213
BIT60	-0.149	0.253
BIT62	0.032	0.266
BIT64	0.062	0.604
BIT66	-4.860	38.180
BIT70	-6.849	38.975
BIT76	-2.325	14.958
BIT81	-25.653	12.229
BIT82	-0.035	1.238
BIT83	11.325	35.828
BIT85	0.315	0.206
BIT87	-38.310	24.825
BIT89	-4.293	6.423
BIT90	-4.127	29.493
BIT91	-54.286	16.998

BIT95	-23.864	17.144
BIT96	-12.968	37.503
BIT98	7.391	7.500
BIT100	-32.690	14.413
BIT102	-0.504	0.234
BIT104	-0.207	1.420
BIT106	3.908	3.329
BIT108	-57.546	30.162
BIT110	-10.662	8.545
BIT112	0.277	0.679
BIT114	0.350	0.979
BIT116	-3.137	5.102
BIT118	-6.980	24.101
BIT121	43.058	49.661
BIT122	-11.305	5.329
BIT135	-10.339	5.360
BIT139	-3.934	8.544
BIT141	-5.110	12.954
BIT142	-0.104	0.706
BIT143	-13.364	11.337
BIT146	0.310	0.294
BIT148	0.512	0.252
BIT152	0.077	0.308
BIT154	-0.700	0.940
BIT155	-13.962	8.102
BIT156	-12.222	16.736
BIT164	24.663	38.976
BIT170	-0.751	0.976
BIT171	-24.871	60.005
BIT172	27.463	23.111
BIT180	-0.280	0.530
BIT182	-1.040	1.329
BIT184	-0.526	0.264

BIT186	1.277	1.042
BIT188	-1.946	0.951
BIT196	20.175	20.918
BIT198	63.068	31.065
BIT200	1.055	6.689
BIT210	53.209	35.716
BIT211	25.563	8.271
BIT212	-19.575	15.010
BIT216	1.224	0.470
BIT219	-29.239	10.192
BIT222	-1.472	5.557
BIT223	-96.309	58.984
BIT224	35.954	36.368
BIT226	-0.403	49.922
BIT228	34.538	21.153
BIT230	-32.121	32.697
BIT231	-35.265	12.060
BIT232	-34.718	20.689
BIT233	54.616	31.301
BIT234	-3.993	4.812
BIT236	-1.981	7.034
BIT237	6.938	34.962
BIT238	14.164	32.404
BIT241	-73.142	41.727
BIT242	27.264	19.186
BIT244	0.031	1.725
BIT248	2.843	2.742
BIT251	-78.610	65.710
BIT254	6.675	8.863
BIT256	11.111	6.737
BIT258	-0.073	0.210
BIT260	-0.312	0.242
BIT266	77.324	35.587

BIT271	-36.416	16.404
BIT273	-33.929	13.445
BIT275	6.173	8.154
BIT276	-1.632	1.490
BIT277	18.698	9.128
BIT278	-1.752	6.751
BIT279	-0.011	21.719
BIT282	66.805	21.526
BIT283	0.247	2.241
BIT286	16.796	7.158
BIT289	6.238	39.858
BIT290	12.659	4.512
BIT292	-38.638	31.352
BIT293	-20.478	11.007
BIT298	160.230	68.863
BIT301	6.828	6.031
BIT303	0.238	0.383
BIT305	-37.211	17.372
BIT307	-0.139	0.317
BIT309	0.945	0.538
BIT311	0.313	1.225
BIT313	0.920	1.073
BIT315	-2.538	1.375
BIT318	-2.425	5.208
BIT322	1.944	6.623
BIT324	-2.777	1.368
BIT326	-102.865	49.191
BIT328	-0.621	1.275
BIT330	-0.609	8.385
BIT331	-56.697	35.842
BIT332	-13.353	20.318
BIT335	-6.050	10.167
BIT337	28.584	18.549

BIT339	1.571	1.687
BIT343	-1.534	2.011
BIT345	-3.408	3.132
BIT347	8.391	23.402
BIT349	-5.736	18.625
BIT353	-0.897	1.478
BIT357	-70.835	41.312
BIT359	-75.363	29.522
BIT361	-40.632	50.047
BIT373	0.102	0.298
BIT375	21.408	12.062
BIT377	0.340	0.458
BIT382	-17.031	20.233
BIT384	-25.139	24.973
BIT390	-22.209	11.659
BIT394	-12.502	13.397
BIT397	30.690	29.705
BIT398	77.154	46.105
BIT400	-2.227	3.107
BIT402	-2.002	3.064
BIT406	-19.401	35.368
BIT408	-13.621	42.432
BIT410	6.244	26.800
BIT413	6.263	15.838
BIT414	-60.398	51.227
BIT416	5.022	4.151
BIT418	0.639	0.940
BIT420	-0.491	0.929
BIT422	-3.144	3.016
BIT424	21.040	19.180
BIT426	-20.310	39.636
BIT427	144.522	72.619
BIT428	-8.963	9.018

BIT431	28.504	18.983
BIT432	0.598	0.507
BIT434	0.013	0.412
BIT436	49.215	50.656
BIT438	-3.109	1.282
BIT440	68.913	33.500
BIT446	56.041	39.203
BIT447	-121.916	54.209
BIT449	0.810	1.267
BIT450	-5.652	3.206
BIT453	0.100	0.277
BIT455	1.198	2.162
BIT457	13.163	4.768
BIT458	-18.746	17.552
BIT461	-13.232	24.046
BIT462	-29.683	36.545
BIT463	22.749	15.935
BIT467	-4.545	22.893
BIT470	-2.315	1.604
BIT472	64.730	45.585
BIT473	23.767	13.416
BIT475	-1.894	17.679
BIT476	-29.894	21.066
BIT479	8.883	26.827
BIT482	-14.772	9.922
BIT483	-42.324	22.849
BIT484	-23.455	53.482
BIT485	-0.004	0.800
BIT493	3.189	8.405
BIT495	19.164	19.644
BIT497	-78.415	43.803
BIT499	22.246	80.955
BIT501	0.267	1.736

BIT505	1.218	4.416
BIT506	97.580	42.207
BIT507	40.542	68.341
BIT508	119.132	44.827
BIT509	-21.783	19.109
BIT510	-5.268	4.630
BIT512	-0.823	0.994
BIT513	13.737	51.130
BIT516	-10.053	10.494
BIT517	8.360	7.415
BIT519	68.383	22.741
BIT521	-0.494	0.228
BIT525	-2.416	6.168
BIT526	-8.253	30.331
BIT529	3.824	5.661
BIT532	-0.360	1.259
BIT533	0.399	1.078
BIT536	-51.697	23.788
BIT537	12.505	15.997
BIT538	23.172	26.463
BIT539	-13.221	10.205
BIT540	59.977	68.899
BIT541	-0.112	0.269
BIT542	-22.167	13.504
BIT543	-17.622	6.871
BIT544	-0.335	2.787
BIT545	-18.085	22.805
BIT547	-111.956	41.659
BIT548	48.690	17.332
BIT549	-13.144	22.787
BIT550	4.094	5.623
BIT551	-7.517	21.034
BIT552	-7.434	33.623

BIT554	10.608	37.121
BIT555	-110.685	41.798
BIT560	-0.484	1.206
BIT562	1.646	1.469
BIT564	0.259	0.253
BIT565	-72.961	54.150
BIT568	17.639	15.559
BIT569	-73.238	51.805
BIT572	-75.776	34.397
BIT575	-0.192	2.351
BIT578	-41.980	23.765
BIT580	10.109	9.320
BIT581	70.458	35.961
BIT584	-24.472	25.270
BIT586	55.233	26.261
BIT587	1.123	18.077
BIT588	0.078	1.313
BIT590	1.728	4.979
BIT591	-0.492	10.946
BIT593	2.023	4.575
BIT595	19.092	16.410
BIT599	-28.539	23.155
BIT600	-11.460	19.116
BIT602	-1.057	1.062
BIT604	-30.529	11.018
BIT606	-38.094	25.873
BIT610	40.970	33.068
BIT611	-14.504	9.554
BIT614	102.874	66.747
BIT617	49.888	16.797
BIT618	4.550	1.781
BIT620	1.484	38.937
BIT623	17.730	16.974

BIT624	21.777	61.291
BIT625	-73.338	20.734
BIT628	18.205	15.800
BIT630	-21.544	43.972
BIT631	-10.750	8.977
BIT632	-0.641	2.969
BIT633	-2.659	13.107
BIT635	12.379	8.990
BIT636	3.412	4.588
BIT638	-28.762	58.877
BIT640	-23.630	11.817
BIT641	2.367	1.777
BIT643	0.622	0.287
BIT645	5.170	4.384
BIT649	-38.197	14.449
BIT650	0.504	5.276
BIT652	-45.255	118.630
BIT653	17.630	30.054
BIT654	-42.579	21.406
BIT656	-0.943	1.834
BIT658	-17.746	9.461
BIT660	0.181	0.314
BIT664	-16.253	32.597
BIT665	-3.833	9.831
BIT667	-2.089	1.748
BIT670	-2.689	1.831
BIT672	-22.434	22.593
BIT674	-0.269	0.239
BIT676	-1.109	1.865
BIT679	-0.174	0.317
BIT681	-10.901	12.234
BIT683	13.840	5.207
BIT684	11.938	4.066

BIT689	-0.253	0.265
BIT690	11.889	23.402
BIT692	-0.121	0.323
BIT695	-0.509	0.384
BIT697	-20.465	14.120
BIT698	-0.049	2.788
BIT699	-8.031	9.606
BIT703	3.605	7.575
BIT704	0.215	1.239
BIT705	-0.421	0.274
BIT706	47.447	41.216
BIT707	45.884	22.485
BIT708	21.285	12.930
BIT709	-61.282	39.672
BIT710	-90.066	54.392
BIT711	15.655	9.592
BIT712	-22.063	17.700
BIT713	166.089	63.720
BIT715	0.561	2.377
BIT717	-29.816	30.397
BIT718	-17.824	28.303
BIT721	-0.367	0.421
BIT723	-4.332	8.099
BIT725	3.119	2.570
BIT727	-0.043	1.282
BIT728	-56.798	26.172
BIT729	53.625	38.742
BIT732	64.354	41.108
BIT736	-0.362	0.636
BIT738	9.713	4.044
BIT739	2.752	2.670
BIT740	-15.217	4.766
BIT741	11.315	21.126

BIT742	-11.613	4.484
BIT744	0.182	2.896
BIT746	34.190	21.169
BIT747	-29.231	17.688
BIT748	-12.813	4.161
BIT749	8.442	9.794
BIT750	12.109	34.250
BIT752	30.048	20.411
BIT753	-0.051	1.474
BIT754	-1.686	4.463
BIT755	9.107	17.176
BIT756	15.010	7.439
BIT759	2.263	4.121
BIT760	-85.591	104.228
BIT762	4.179	1.920
BIT763	-2.358	35.735
BIT764	11.555	6.438
BIT765	19.546	20.575
BIT766	-0.061	0.838
BIT768	18.462	12.206
BIT769	-42.468	20.734
BIT770	-1.480	7.549
BIT772	2.590	4.766
BIT773	12.172	42.781
BIT774	-0.039	0.284
BIT776	26.702	10.972
BIT777	0.148	1.563
BIT778	26.127	10.833
BIT779	21.155	13.862
BIT780	38.078	68.580
BIT781	48.083	27.568
BIT782	18.952	17.923
BIT784	-72.081	49.301

BIT785	-5.006	1.235
BIT786	-10.032	5.492
BIT789	65.163	32.843
BIT790	7.956	4.077
BIT791	11.459	24.915
BIT793	-1.524	20.951
BIT798	3.120	3.577
BIT802	2.564	28.023
BIT803	18.434	21.145
BIT806	-46.640	22.007
BIT810	-8.735	6.134
BIT811	-32.373	32.742
BIT812	-1.314	1.357
BIT813	23.739	15.226
BIT814	-0.186	2.223
BIT816	-15.660	10.751
BIT817	0.117	0.252
BIT818	0.027	2.423
BIT819	-1.894	7.676
BIT820	30.050	22.759
BIT821	4.651	10.902
BIT822	1.634	2.454
BIT823	68.774	29.005
BIT824	-0.128	0.542
BIT825	-30.926	13.436
BIT826	-28.573	17.892
BIT827	-0.060	0.248
BIT828	-0.264	1.496
BIT829	18.120	15.497
BIT830	22.235	10.766
BIT831	-0.227	9.525
BIT833	0.981	0.557
BIT834	11.568	11.027

BIT835	0.329	0.218
BIT837	0.302	6.550
BIT840	0.420	0.212
BIT842	0.985	0.798
BIT843	27.617	11.144
BIT844	-0.519	0.244
BIT846	-2.397	1.777
BIT848	-0.845	1.119
BIT849	-17.248	11.556
BIT850	-5.299	3.072
BIT854	-0.564	3.851
BIT855	-164.664	103.873
BIT859	2.252	7.385
BIT861	-0.044	0.248
BIT863	-0.080	0.404
BIT865	-1.076	0.951
BIT868	-7.943	66.815
BIT870	-0.118	0.401
BIT871	-4.227	5.965
BIT872	11.542	14.107
BIT874	-8.671	5.019
BIT875	5.588	6.832
BIT877	1.592	8.504
BIT878	0.937	2.855
BIT879	-0.234	1.115
BIT880	-4.468	20.371
BIT881	1.638	2.947
BIT882	-8.589	8.570
BIT883	-4.321	6.161
BIT884	-2.518	13.847
BIT886	-1.870	1.864
BIT887	-0.318	0.395
BIT888	-37.957	21.074

BIT889	1.281	1.070
BIT890	22.628	18.008
BIT891	-0.269	0.717
BIT892	20.373	40.041
BIT893	10.456	5.834
BIT894	-19.582	13.416
BIT896	-5.911	14.642
BIT897	5.462	3.920
BIT898	25.219	13.487
BIT899	33.964	11.376
BIT900	3.157	2.273
BIT903	16.273	7.003
BIT904	0.431	1.359
BIT905	-1.784	3.262
BIT906	7.462	8.141
BIT907	-31.213	13.500
BIT909	-5.004	8.424
BIT910	-3.758	10.764
BIT911	-0.363	0.328
BIT912	7.622	14.453
BIT913	0.854	0.286
BIT914	-0.017	2.055
BIT915	1.383	1.004
BIT916	0.420	6.173
BIT918	-17.003	6.700
BIT919	0.648	1.682
BIT920	0.338	0.539
BIT921	4.698	4.857
BIT922	-2.266	2.073
BIT924	-0.608	0.292
BIT925	-0.530	0.349
BIT926	0.923	1.154
BIT927	-12.029	12.135

BIT928	2.708	5.139
BIT929	-0.053	0.717
BIT930	-0.721	2.168
BIT931	0.391	0.454
BIT932	0.235	0.489
BIT933	-0.280	1.653
BIT934	1.635	0.805
BIT935	2.391	3.171
BIT936	9.353	10.611
BIT937	7.660	4.463

Table 4-2. The predicted values of logS by the built model using shape fingerprints. The table also includes values of experimental logS and predicted by MOE.[98]

Molecule	Experimental logS	Predicted logS value	Standard Error	Predicted logS value by MOE
1	-6.23	-5.85	0.13	-5.7
2	-6.21	-5.09	0.16	-4.83
3	-4.39	-3.9	0.32	-5.25
4	-6.19	-6.29	0.46	-6.31
5	-6.16	-6.81	0.17	-8.11
6	-6.15	-6.64	0.48	-4.54
7	-3.68	-6.82	0.46	-3.59
8	-6.08	-4.17	0.21	-6.24
9	-6.04	-3.9	0.32	-3.76
10	-5.14	-5.76	0.18	-6.2
11	-6.01	-6.11	0.15	-6.01
12	-2.97	-4.31	0.26	-4.45
13	-5.47	-3.73	0.27	-5.64
14	-5.84	-5.13	0.19	-6.06
15	-5.97	-7.24	0.56	-6.69
16	-5.96	-7.48	0.57	-4.32

17	-5.93	-4.11	0.65	-5.8
18	-5.79	-4.51	0.15	-4.26
19	-5.85	-2.94	0.22	-3.79
20	-5.91	-3.14	0.5	-5.28
21	-5.38	-4.07	0.34	-5.38
22	-5.87	-4.02	0.52	-6.87
23	-5.88	-3.53	0.37	-4.25
24	-5.84	-3.51	0.41	-5.84
25	-5.83	-5.68	1.04	-6.57
26	-5.82	-4.59	0.45	-7.51
27	-5.8	-4.73	0.2	-4
28	-5.8	-4.79	0.22	-5.35
29	-5.1	-4.81	0.15	-5.11
30	-5.74	-1.67	0.36	-3.75
31	-5.72	-6.52	0.24	-7.86
32	-5.71	-7.01	0.7	-6.25
33	-5.69	-1.64	0.32	-5.95
34	-5.67	-5.24	0.47	-6.3
35	-4.48	-4.48	0.21	-5.43
36	-5.65	-5.34	0.18	-4.83
37	-5.63	-5.24	0.48	-5.5
38	-5.62	-5.71	0.19	-5.16
39	-5.55	-3.85	0.3	-2.77
40	-5.52	-3.86	0.27	-5.31
41	-5.51	-3.64	0.29	-5.96
42	-5.5	-5.2	0.27	-7.11
43	-5.44	-4.79	0.46	-5.89
44	-5.44	-4.37	0.27	-5.16
45	-5.43	-6.68	0.28	-7.3
46	-5.42	-4.27	0.16	-5.97
47	-5.4	-5.28	0.21	-5.58
48	-5.4	-4.3	0.22	-4.64
49	-5.32	-3.02	0.45	-5.05

50	-5.3	-5.11	0.51	-4.01
51	-3.26	-4.26	0.27	-4.59
52	-5.28	-3.5	0.25	-5.41
53	-5.28	-2.53	0.11	-6.39
54	-5.03	-4.18	0.17	-4.41
55	-5.27	-4.15	0.22	-4
56	-5.27	-4.98	0.37	-6.79
57	-5.22	-3.9	0.27	-4.31
58	-5.15	-5.11	0.59	-5.08
59	-5.2	-1.98	0.25	-3.12
60	-5.19	-4.51	0.28	-4.3
61	-5.19	-4.11	0.31	-6.75
62	-5.18	-6.5	0.34	-9.6
63	-3.61	-2.63	0.26	-4.18
64	-1.74	-4.02	0.41	-4.46
65	-5.12	-3.12	0.22	-4.52
66	-5.12	-1.63	0.81	-5.45
67	-5.03	-2.54	0.36	-4.95
68	-5.11	-3.92	0.19	-4.24
69	-5.08	-3.44	0.19	-2.54
70	-5.07	-3.8	0.22	-5.36
71	-4.98	-6.58	0.24	-5.7
72	-5.05	-4.23	0.23	-3.66
73	-2.93	-3.98	0.27	-4.22
74	-5.02	-3.46	0.22	-4.27
75	-4.53	-2.71	0.32	-5.49
76	-5.03	-2.83	0.27	-5.09
77	-5.02	-3.78	0.46	-4.87
78	-5.01	-3.85	0.29	-4.58
79	-5.01	-4.39	0.11	-6.28
80	-4.95	-5.72	0.2	-5.33
81	-4.93	-4.11	0.46	-4.97
82	-4.91	-3.56	0.19	-3.66

83	-4.9	-5.43	0.52	-4.02
84	-4.9	-2.46	0.71	-3.8
85	-4.9	-3.25	0.33	-4.02
86	-2.08	-3.48	0.34	-4.87
87	-4.86	-7.01	0.93	-5.07
88	-4.86	-5.97	0.23	-5
89	-4.85	-4.17	0.42	-4.68
90	-4.85	-4.52	0.18	-4.99
91	-3.25	-3.29	0.28	-3.51
92	-4.83	-5.58	0.19	-4.97
93	-4.37	-4.38	0.46	-4.8
94	-4.8	-4.05	0.84	-3.94
95	-4.79	-5.91	0.7	-4.29

The predicted values do not vary much from those obtained from well-established logS prediction software. Many compounds have either a similar or slightly worse predicted logS values. However, for a few compounds the values are much closer to experimental logS values than the ones predicted by MOE. This was summarized in the form of a plot (**Figure 4-6**).

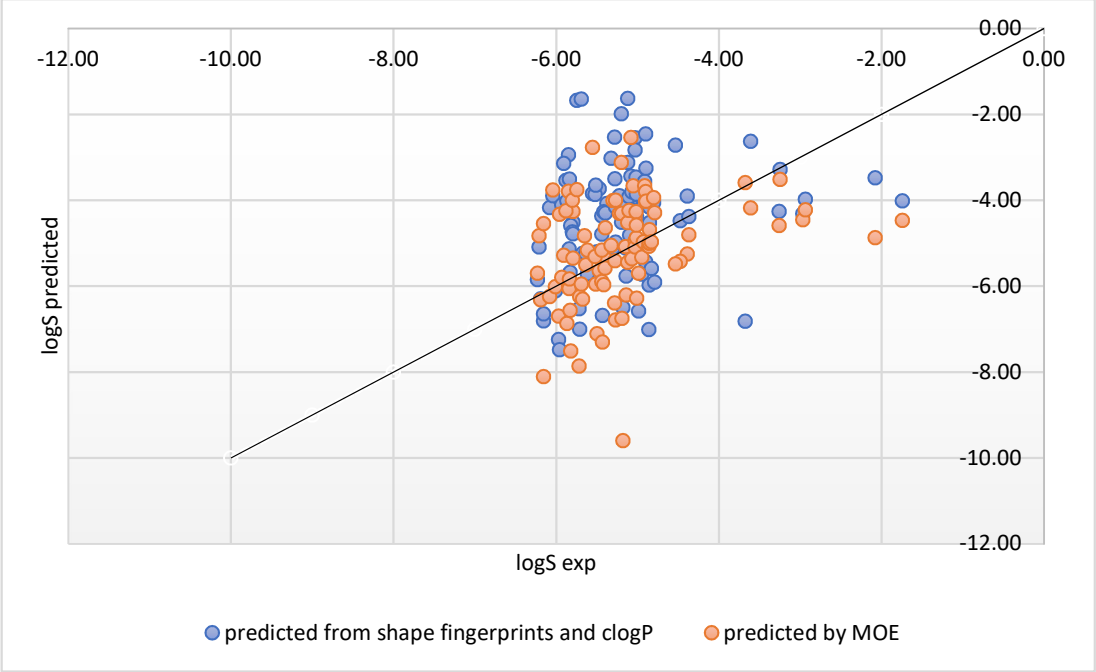
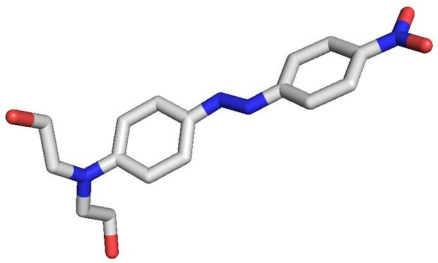


Figure 4-6. The plot of predicted logS values vs. experimental values for shape fingerprints and MOE. The 1:1 line is included.

Some predicted solubility values differ from the experimental ones more than the others. The logS values are not accurately predicted for the structures of molecules shown in the **Figure 4-7**.

A)



B)

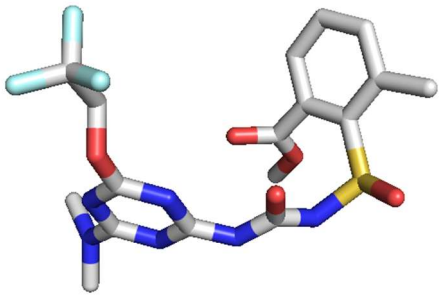


Figure 4-7. The structures of compounds that predicted solubility differs much from experimental: A) predicted logS = -6.82, while experimental logS = -3.68; B) predicted logS = -1.64 and experimental logS = -5.69.

The second set containing all molecules with measured logS had 4194 compounds in the training set and 464 compounds in the test set. The model was built in a similar way as described above. The results are shown in the **Figure 4-8** in comparison to experimental logS values.

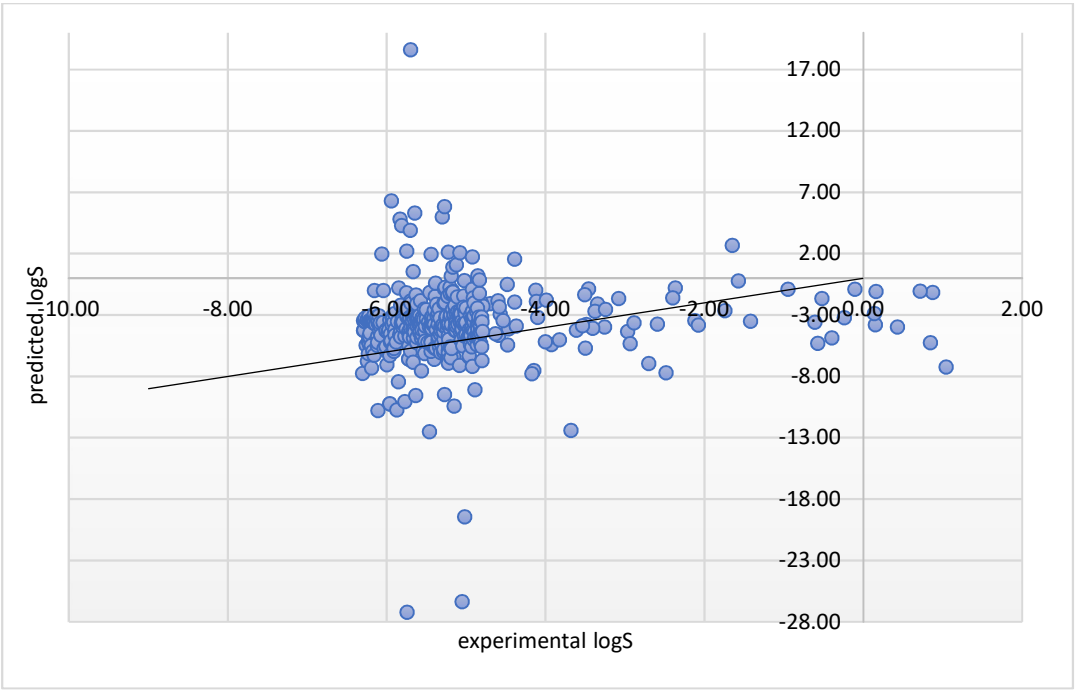


Figure 4-8. The plot of predicted logS values for 464 compounds vs. experimental values for shape fingerprints. The 1:1 line is included.

In second set, some of the logS values were poorly predicted, they vary much from the experimental values. The examples of the structures of the molecules with such

poor prediction are show in the **Figure 4-9**. Most probable reason for some of the extreme results is that the interaction of molecule with water plays more important role than its shape. For others, when the actual solubility is worse than the predicted one, probably form strong intermolecular bonds in the crystal lattice. The consequence of that is the difficulty of predicting solubility using shape fingerprints method.

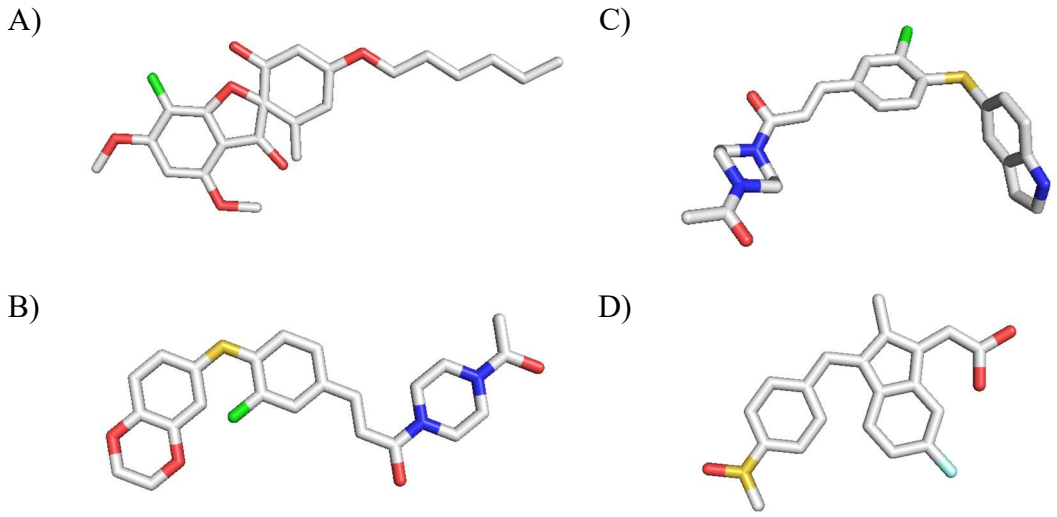


Figure 4-9. The structures of compounds that predicted solubility differs much from experimental: A) predicted logS = 18.61, while experimental logS = -5.70; B) predicted logS = -26.38 and experimental logS = -5.05; C) predicted logS = -27.22 and experimental logS = -5.74; D) predicted logS = -7.26 and experimental logS = 1.04.

The proposed model predicts the aqueous solubility without the use of any experimental data. The only necessary data points are the calculated logP and shape fingerprints. As it produces similar results to MOE, it could be used simultaneously with it to predict quite accurately the values of logS.

4.4.1.2. The set of 100 compounds – all Shape Databases

The shape fingerprints were calculated for each compound from the set of randomly chosen 100 molecules (all 100 molecules had both MP and logS values) using all of the Shape Databases and different settings: 1) DT = 0.50, BOV = 0.50, 2) DT = 0.50, BOV = 0.70, 3) DT = 0.70, BOV = 0.50 and 4) DT = 0.70, BOV = 0.70. The prediction

of solubility of the set of compounds was performed using SPSS,[119]. The independent values were selected to be calculated logP (computed in MOE)[98] and all the bits from the generated shape fingerprint string. The applied method was 'Enter', which means that all independent variables are entered into the equation in one step. The results are shown in **Table 4-3** as adjusted R^2 for each Shape Database with each setting listed above. The model using clogP and MP to predict logS was applied as benchmark, where the obtained R^2 was equal to 0.694.

Table 4-3. The adjusted R^2 values obtained from SPSS.*[119]*

SHAPE DATABASE	DT = 0.50; BOV = 0.50	DT = 0.50; BOV = 0.70	DT = 0.70; BOV = 0.50	DT = 0.70; BOV = 0.70
SD01	0.739	0.676	0.909	0.733
SD02	0.729	0.642	0.915	0.754
SD03	0.71	0.666	0.796	0.719
SD04	0.708	0.641	0.755	0.706
SD05	0.77	0.641	0.79	0.691
SD06	0.766	0.656	0.628	0.747
SD07	0.712	0.643	1	0.754
SD08	0.708	0.647	0.647	0.74
SD09	0.668	0.661	0.737	0.705
SD10	0.746	0.641	0.642	0.78

The highest values of R^2 were obtained for SD07 with DT = 0.70 and BOV = 0.50, and only a little lower for SD01 and SD02 with the same settings. This suggests that using a high DT and slightly lower BOV gives the best results in building models for solubility prediction. However, the high scores (especially R^2 =1) were probably the effect of overfitting, as the number of terms included in model in case was too high compared to the number of compounds used in prediction (100). It is also worth noting that there is no Shape Database that performs the best across all of the applied settings.

4.5. Nuclear Receptors

Nuclear receptors (NRs) are a protein superfamily that bind and respond to certain steroid hormones, e.g. estrogen and progesterone, and a range of other signalling molecules, such as retinoic acid and thyroid hormone.[120,121] The superfamily is classified as transcription factors due to their ability to directly bind DNA and control the expression of genomic DNA. The NRs play an important role in many physiological functions such as cell proliferation, development, metabolism, and reproduction.[120,121] Many NRs also regulate a number of proteins involved in xenobiotic metabolism, which protects the organism against potentially toxic compounds (cytochrome P450 family).[120,122]

Although, there are 48 known NRs encoded in the human genome,[123] for some of them neither physiological function nor natural ligands are known. These are called orphan receptors.[123] This includes e.g. estrogen-related receptor and human nuclear factor 4.

The molecular structure of NRs is very similar. Almost all of the nuclear receptors have two structural domains: a DNA-binding domain (DBD) and C-terminal ligand-binding domain (LBD).[123,124] Members of this superfamily also contain an N-terminal transactivation domain.[123] The DBD domain, which is the most conserved segment of NRs,[124] contains two zinc ions coordinated by four cysteine residues.

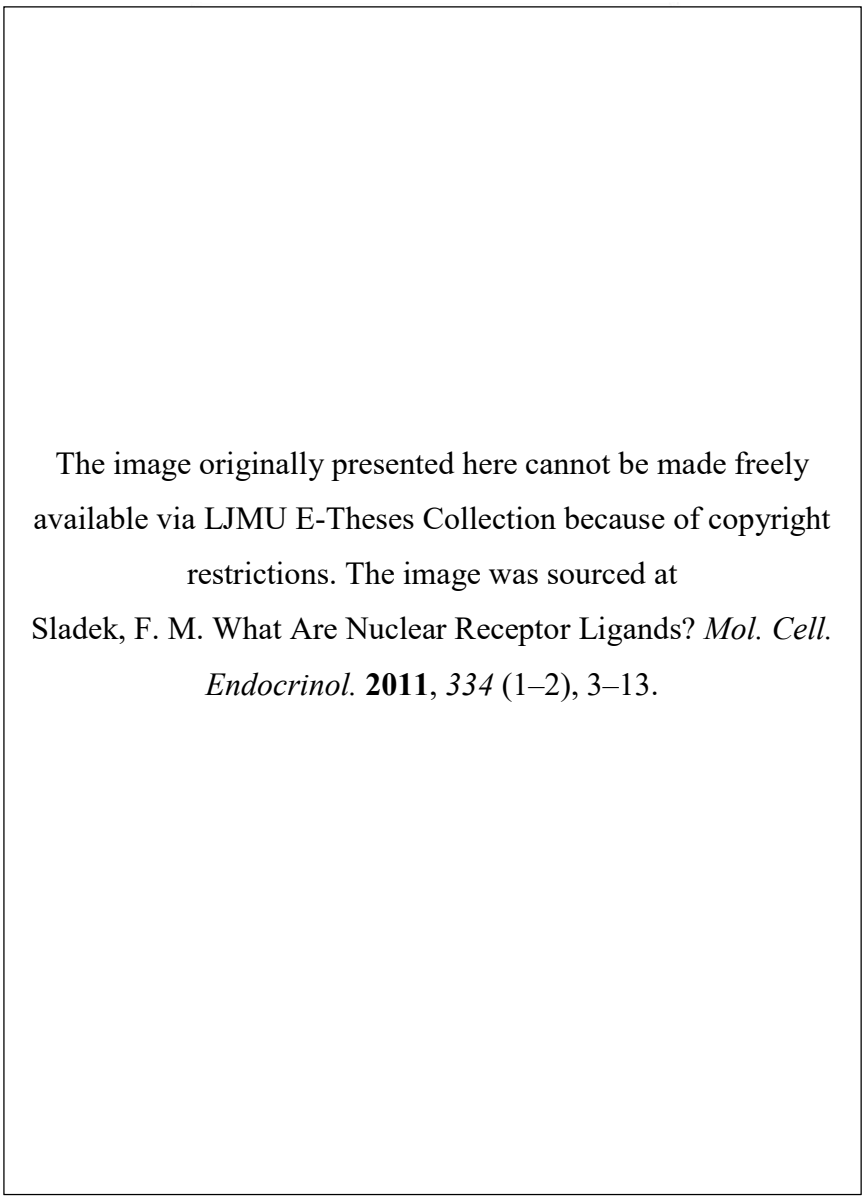


Figure 4-10. The structures of some of the NRs ligands.[125]

As already mentioned, among NR ligands (**Figure 4-10**) there are retinoic acids, steroids and thyroid hormones.[126] They can be characterized as lipophilic, large (200-1600 Da) molecules that need to cross the plasma membrane in order to bind to the hydrophobic pocket of its receptor. [121,125]

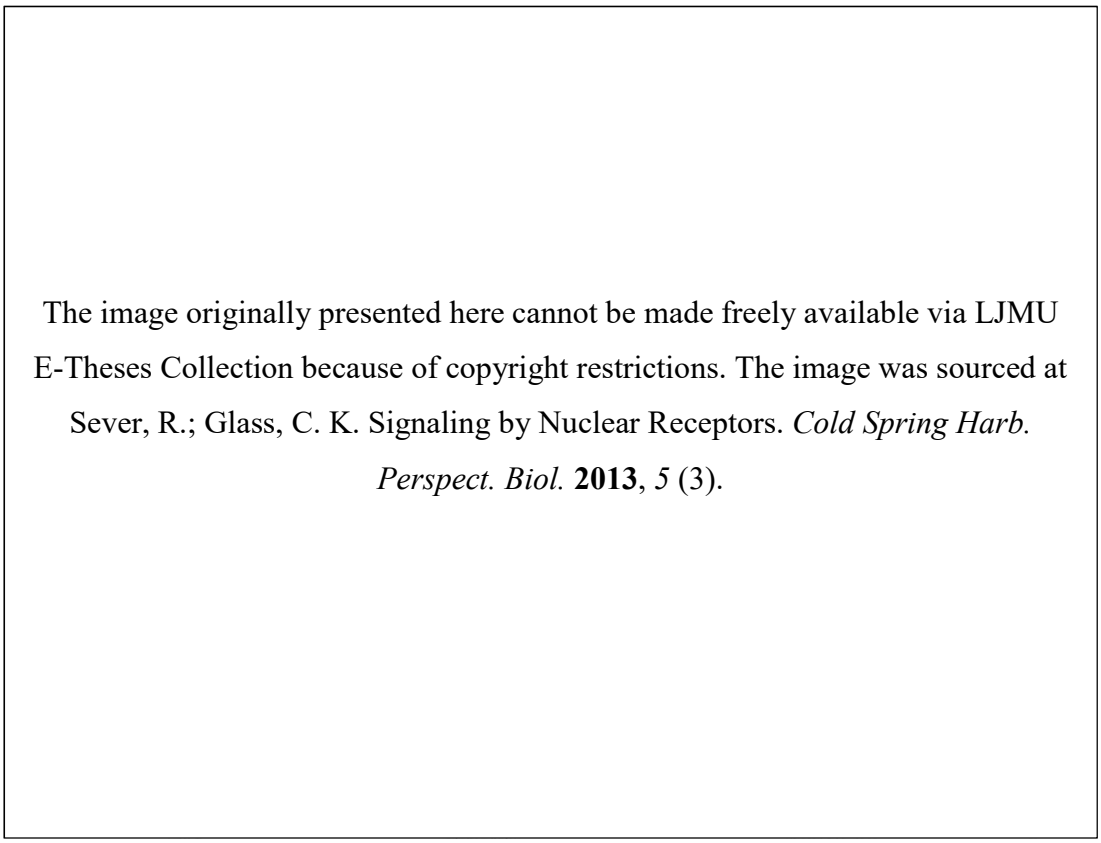


Figure 4-11. Nuclear receptor signaling.[121] Abbreviations: HSP – Heat Shock Protein; ER – Estrogen receptor; RXR – Retinoid X Receptor; NR – Nuclear Receptor; ERE – Estrogen Response Element; TRE – Thyroid Hormone Response Element; LXRE – Liver X Response Element.

When the natural small lipophilic ligands bind to their nuclear receptor, it activates the signalling pathway, as shown in the **Figure 4-11**, based on one of four modes of actions: 1) the ligand frees the receptor from the chaperone, which allows the created complex to enter into the nucleus, where the complex forms interactions with coactivators and the target genes are activated (e.g. the estrogen receptor, the progesterone receptor); 2) ligand by binding to the receptor causes dissociation of the corepressors that interact with it and their replacement with coactivators (e.g. thyroid hormone receptor, retinoid acid receptor); 3) similarly to type 1 but with different organization of the hormone response elements (HREs); 4) bind as monomers to a single half site HREs.[121]

4.5.1. Methods

The database of various NRs ligands was obtained from C. Mellor and F. Steinmetz.[127] The list includes identified NR agonists expanded with data from the ChEMBL database of bioactive molecules.[79] The set comprises of the 22 NRs: the aryl hydrocarbon receptor (AHR), two estrogen receptors (ER): ER-alpha, ER-beta, the glucocorticoid receptor (GR), the progesterone receptor (PR), vitamin D receptor (VDR), the thyroid hormone receptor (TR), three retinoic acid receptors (RAR): RAR-alpha, RAR-beta, RAR-gamma, the pregnane X receptor (PXR), three types of peroxisome proliferator-activated receptors (PPAR): PPAR-alpha, PPAR-gamma, PPAR-delta, two isoforms of the liver X receptor (LXR): LXR-alpha, LXR-beta, farnesoid X receptor (FXR), two thyroid hormone receptors (THR): THR-alpha, THR-beta and three retinoid X receptors (RXR): RXR-alpha, RXR-beta, RXR-gamma. The set was updated into a MySQL database,[128] which stored all ligands with added information of all the receptors that it binds to. This enabled easy access to those ligands that interact with only specific NRs and not with others.

The conformations were generated for every compound using Openeye's OMEGA.[83] The number of maximum conformations was set to 5. The shape fingerprints were calculated with SD10 and settings DT = 0.65 and BOV = 0.60. Two summary values were used when comparing the conformations of ligands: 1) the highest value of FT amongst the array arising from comparisons of all conformations of one molecule with all conformations of the other was selected or 2) the average of those values was selected.

Two approaches were applied. In the first, the shapes of all the ligands for each receptor were compared. This will show the NRs in which the shape of molecules plays a crucial role. In the second, a selected number of ligands from each receptor was compared to a set of structures including ones that bind to each of the NRs and decoys generated by DUD-E.[90]

4.5.2. Comparing the shape of ligands binding to each receptor

The comparison of shapes of ligands binding to the same receptor was performed to examine how important shape is for some of the targets and to find those receptors for which the shape is not an important attribute of its ligands.

Among all of the NRs, the strongest shape similarity between ligands is visible in the case of AHR (**Figure 4-12**), ER-alpha (**Figure 4-13**), ER-beta (**Figure 4-14**), PR (**Figure 4-18**), THR-alpha (**Figure 4-26**) THR-beta (**Figure 4-27**) and GR (**Figure 4-29**). For those targets, the similarity between molecules is quite high for the whole set, especially when using the MV method. The AV method for those NRs shows slightly lower values of ST. In the case of TR (**Figure 4-28**), PXR (**Figure 4-19**), RAR-alpha (**Figure 4-20**), RAR-beta (**Figure 4-21**) RAR-gamma (**Figure 4-22**), RXR-alpha (**Figure 4-23**), RXR-beta (**Figure 4-24**), RXR-gamma (**Figure 4-25**) and VDR (**Figure 4-30**) it can be observed that there are a few groups of ligands which are similar in shape, but this similarity is not shared across the whole set of molecules. The lack of shape similarity can be noticed for FXR (**Figure 4-15**), LXR-alpha (**Figure 4-16**) and LXR-beta (**Figure 4-17**) independently of the comparison method used (AV and MV).

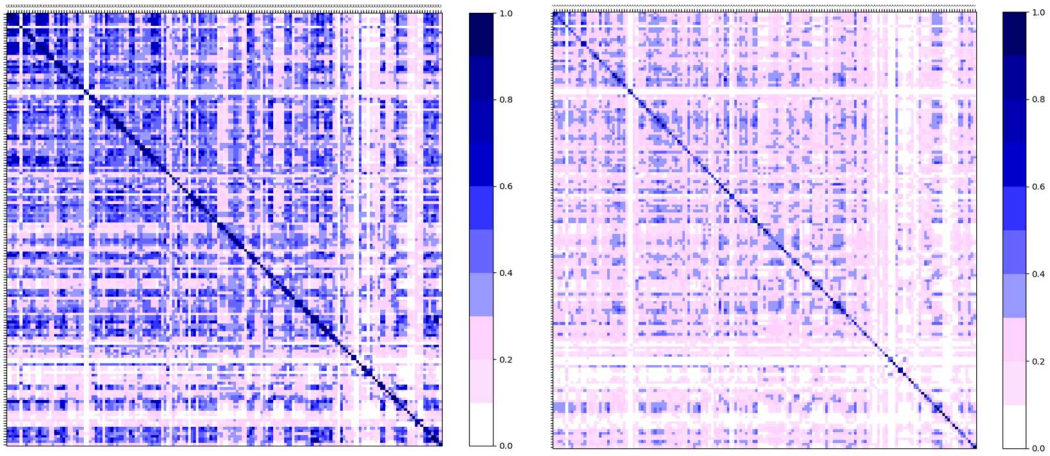


Figure 4-12. Heatmap of STs for AHR ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method.

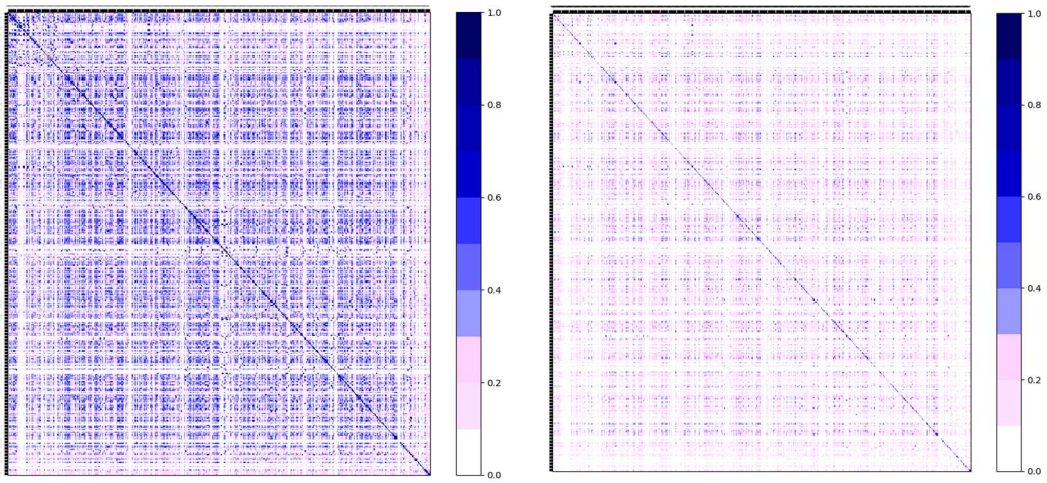


Figure 4-13. Heatmap of STs for ER-alpha ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

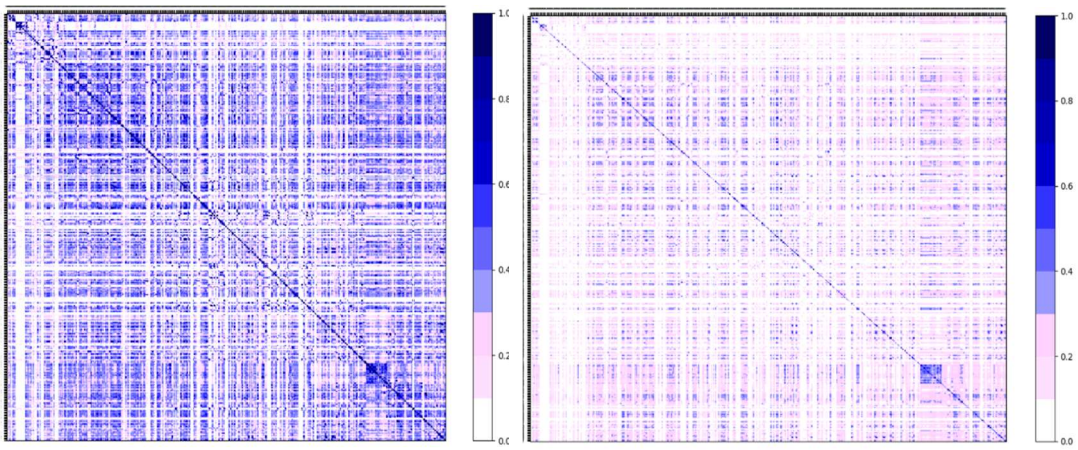


Figure 4-14. Heatmap of STs for ER-beta ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

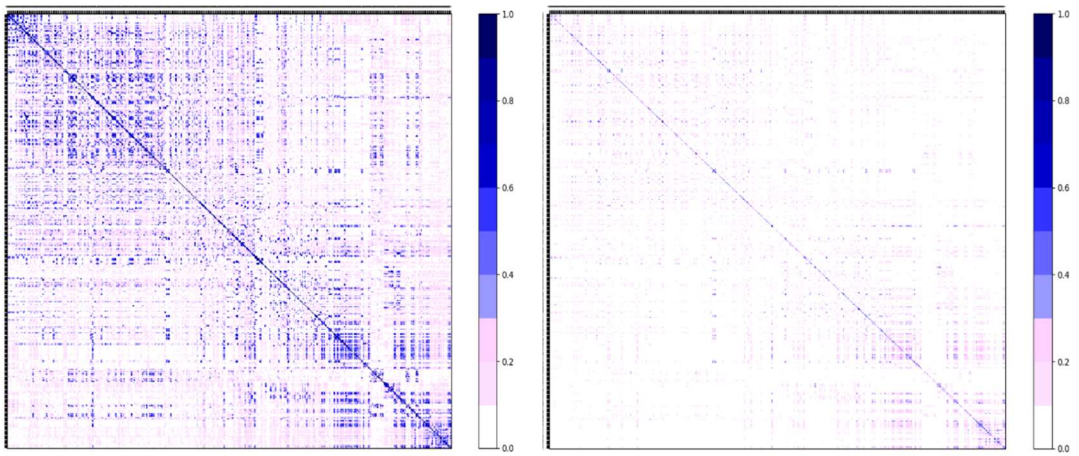


Figure 4-15. Heatmap of STs for FXR ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

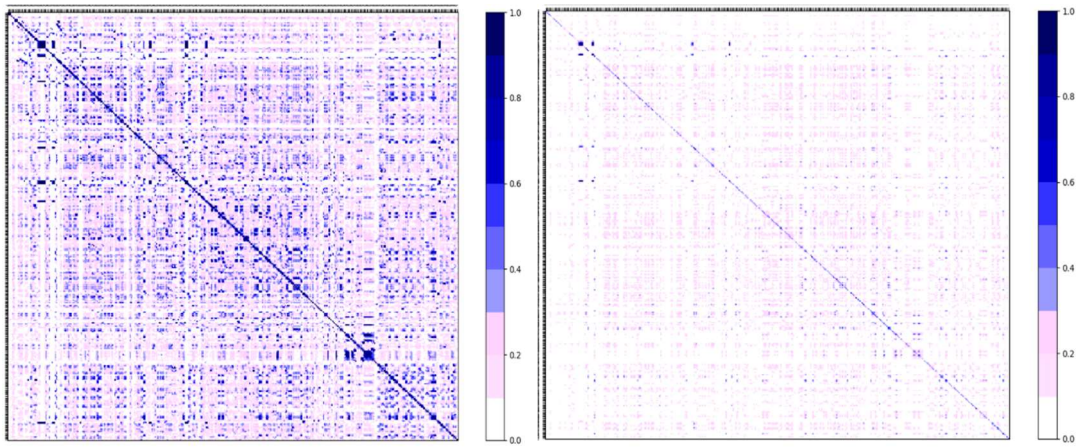


Figure 4-16. Heatmap of STs for LXR-alpha ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

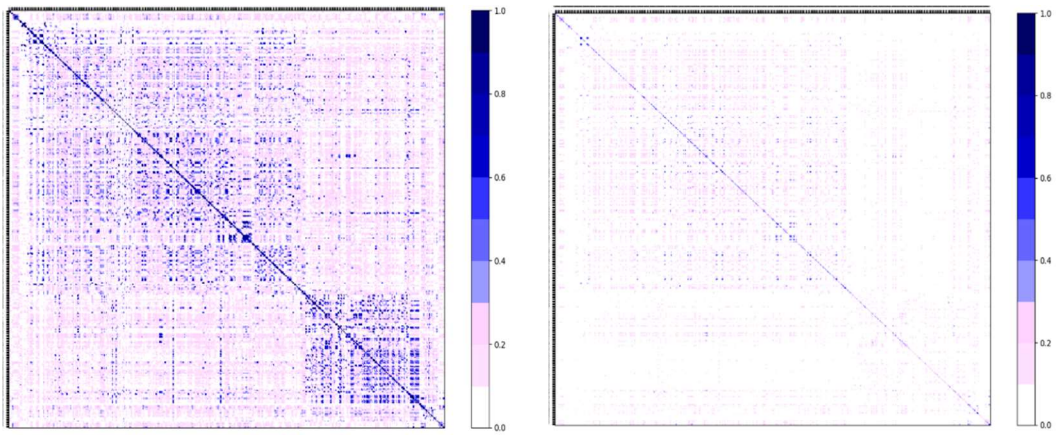


Figure 4-17. Heatmap of STs for LXR-beta ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

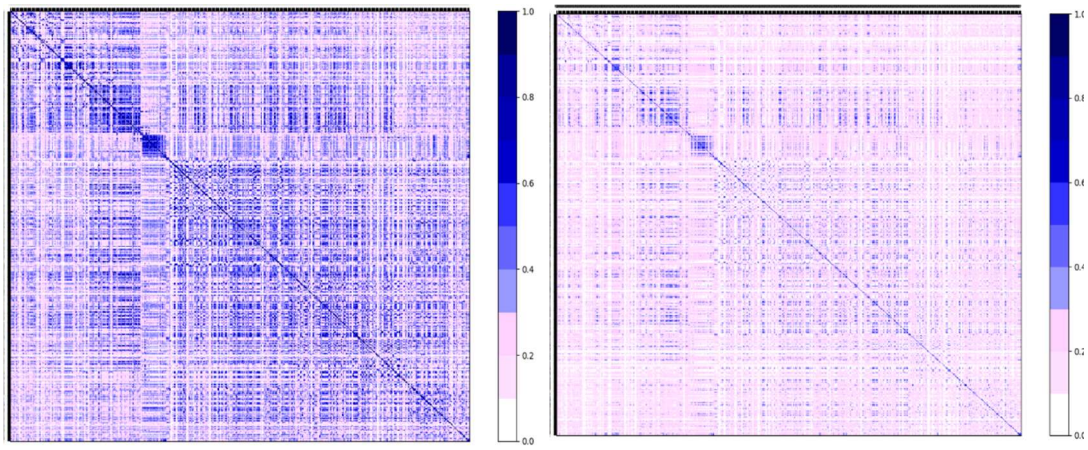


Figure 4-18. Heatmap of STs for PR ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

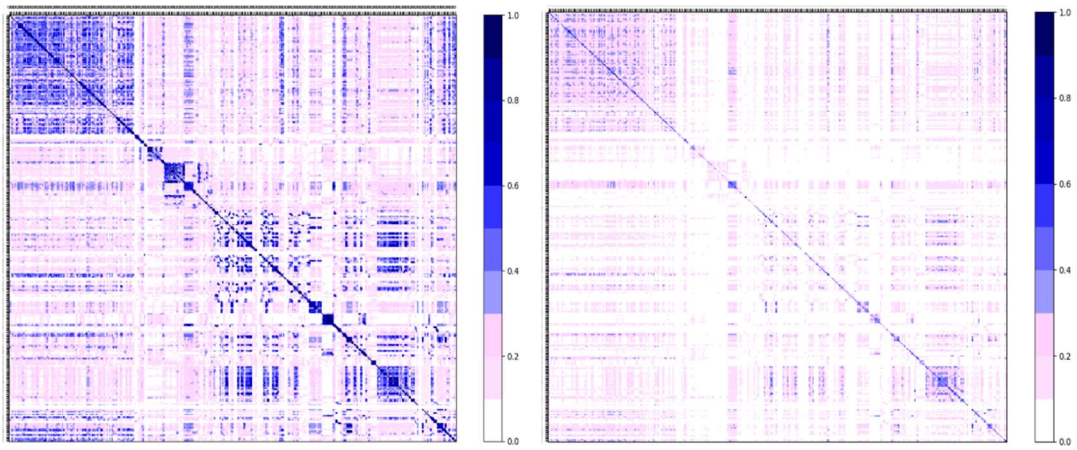


Figure 4-19. Heatmap of STs for PXR ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

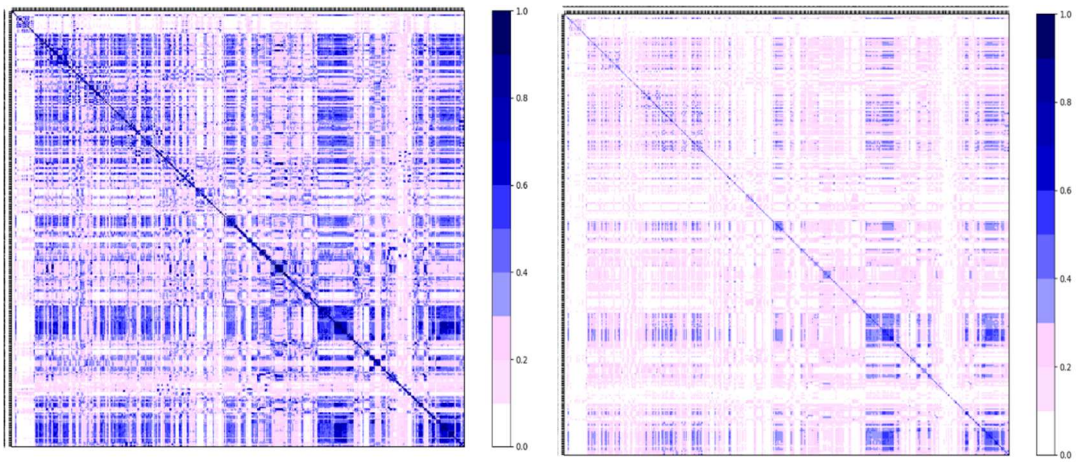


Figure 4-20. Heatmap of STs for RAR-alpha ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

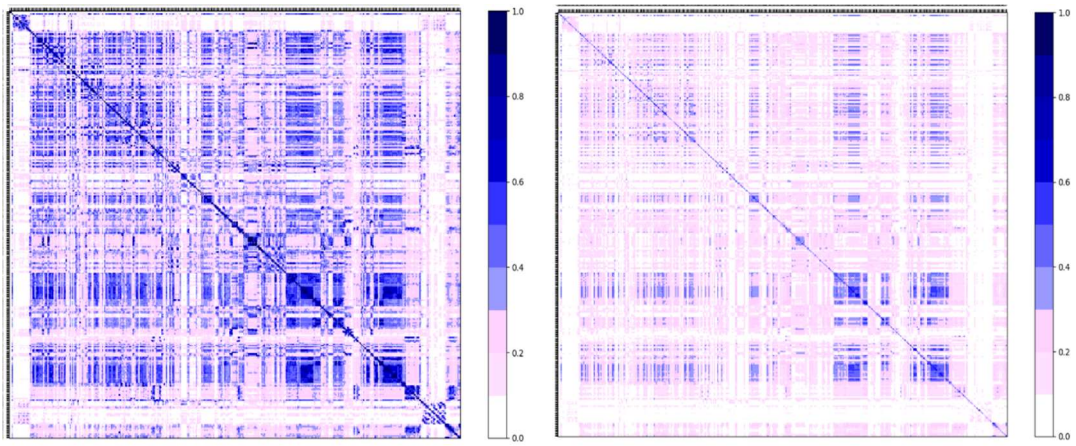


Figure 4-21. Heatmap of STs for RAR-beta ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

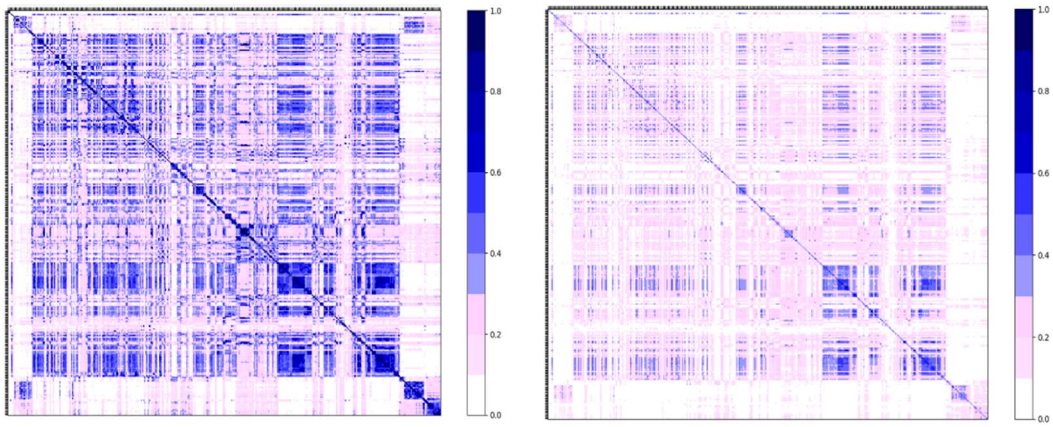


Figure 4-22. Heatmap of STs for RAR-gamma ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

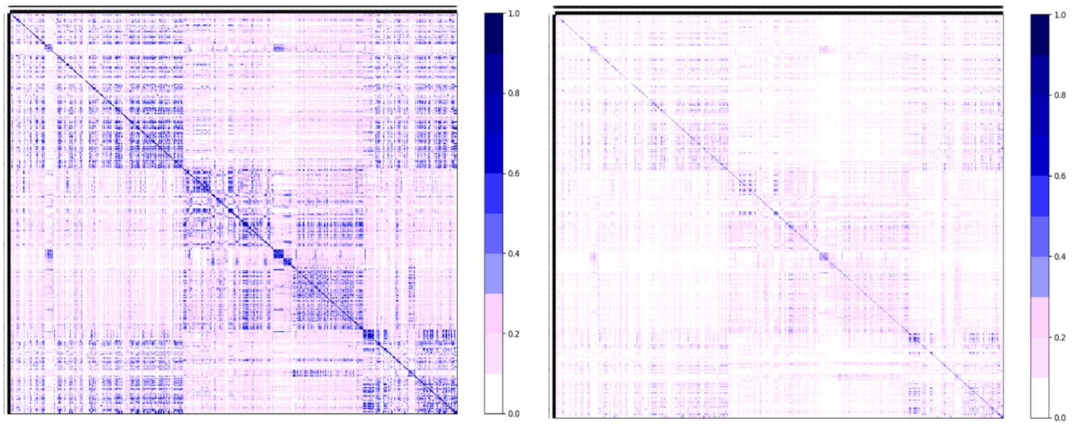


Figure 4-23. Heatmap of STs for RXR-alpha ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

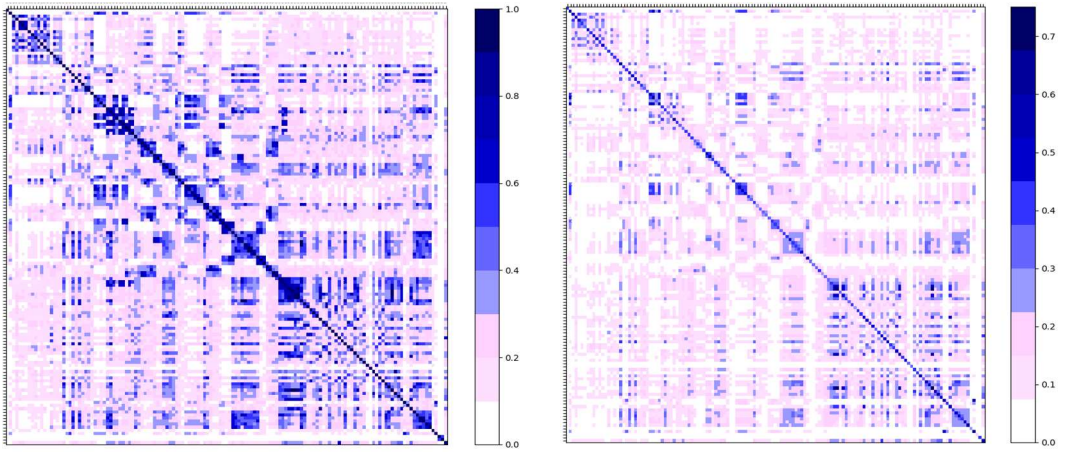


Figure 4-24. Heatmap of STs for RXR-beta ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

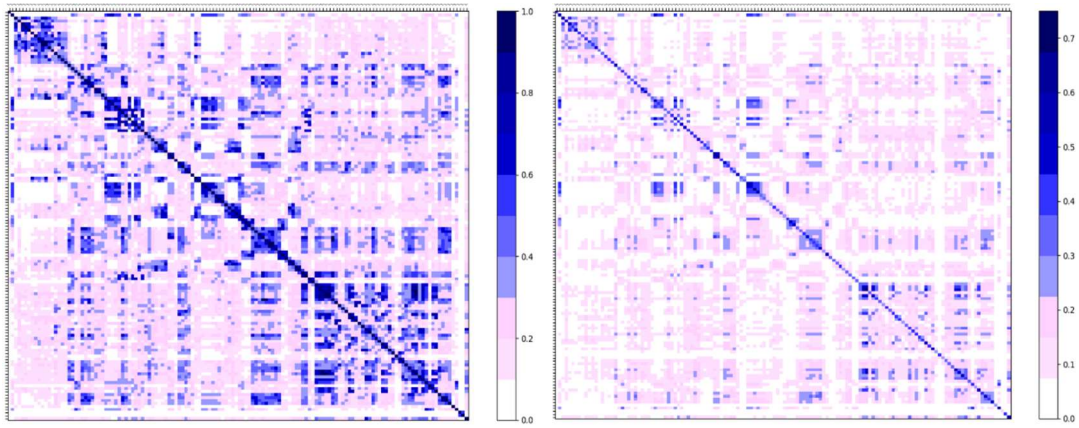


Figure 4-25. Heatmap of STs for RXR-gamma ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

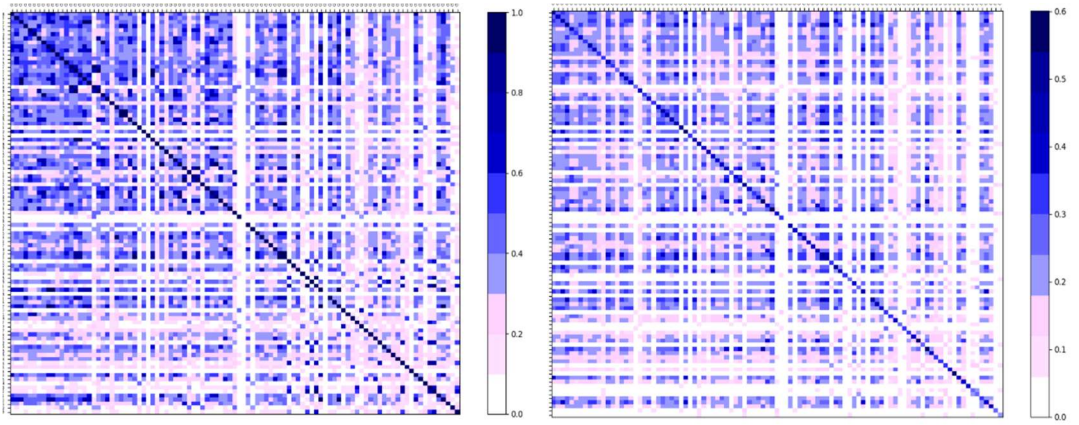


Figure 4-26. Heatmap of STs for THR-alpha ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

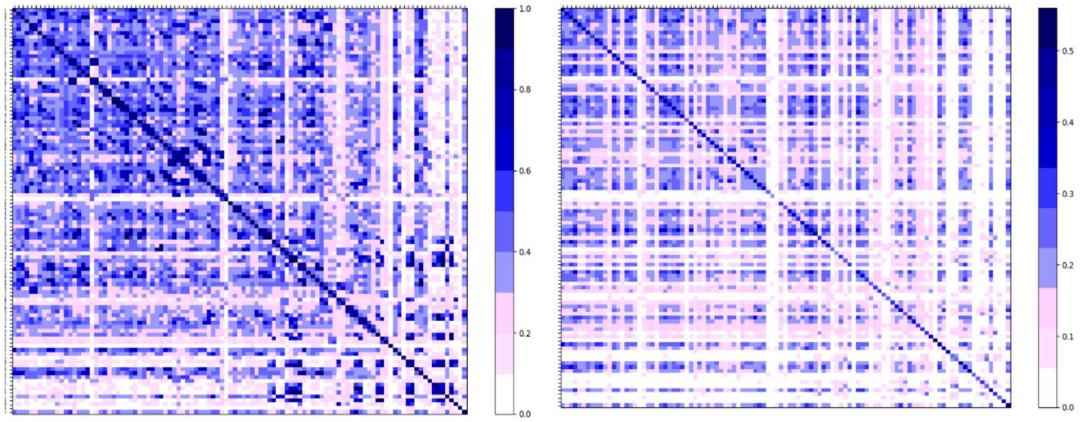


Figure 4-27. Heatmap of STs for THRb ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

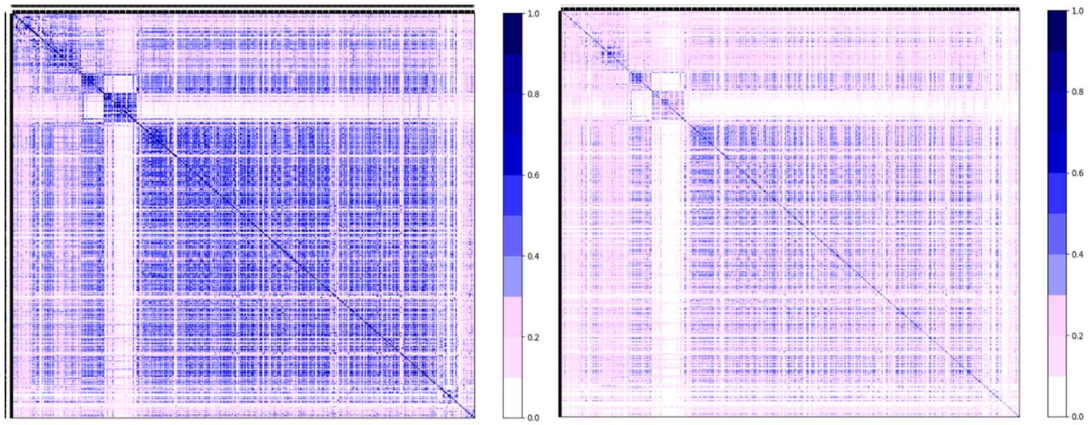


Figure 4-28. Heatmap of STs for TR ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

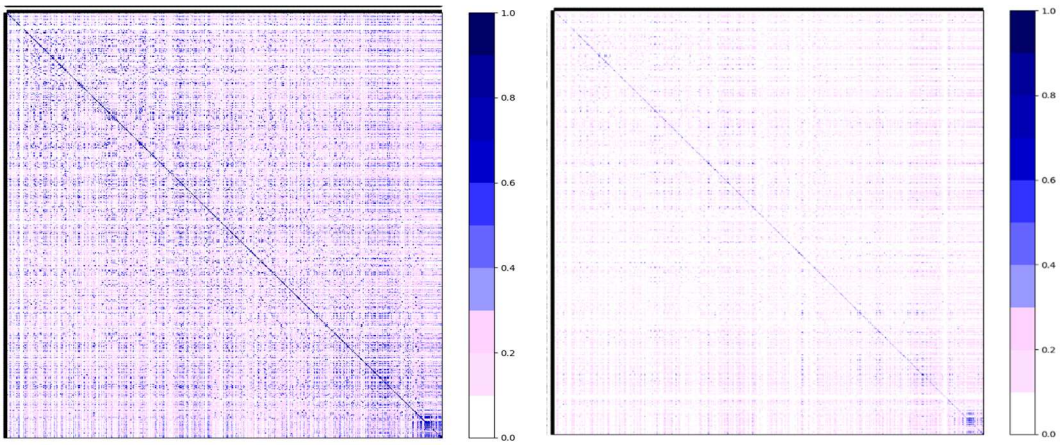


Figure 4-29. Heatmap of STs for GR ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

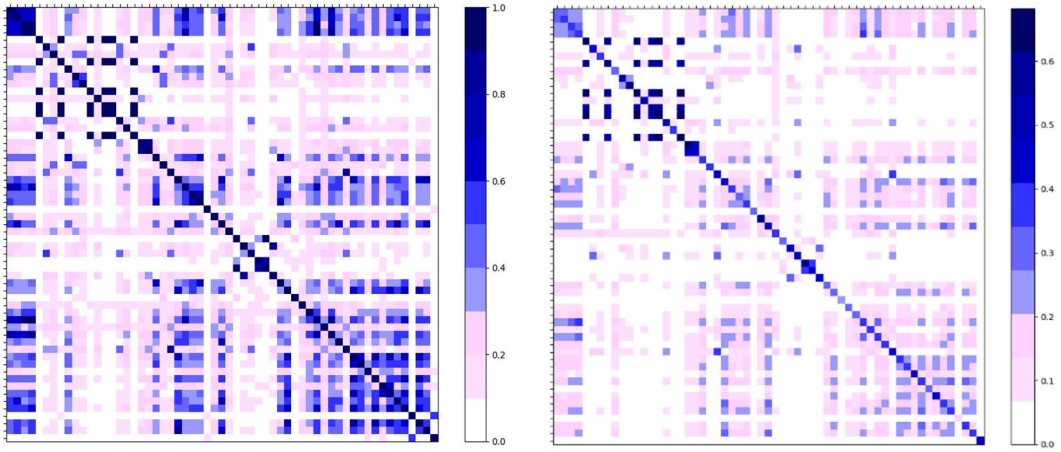


Figure 4-30. Heatmap of STs for VDR ligands when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by the shape fingerprints method.

4.5.3. Virtual Screening of NRs

In order to check the ability to distinguish the ligands binding to one receptor from those binding to the others, the set of 50 compounds from each set was taken and analysed together with a set of molecules binding to all the other receptors. Additionally, a set of decoys was added for better comparison.

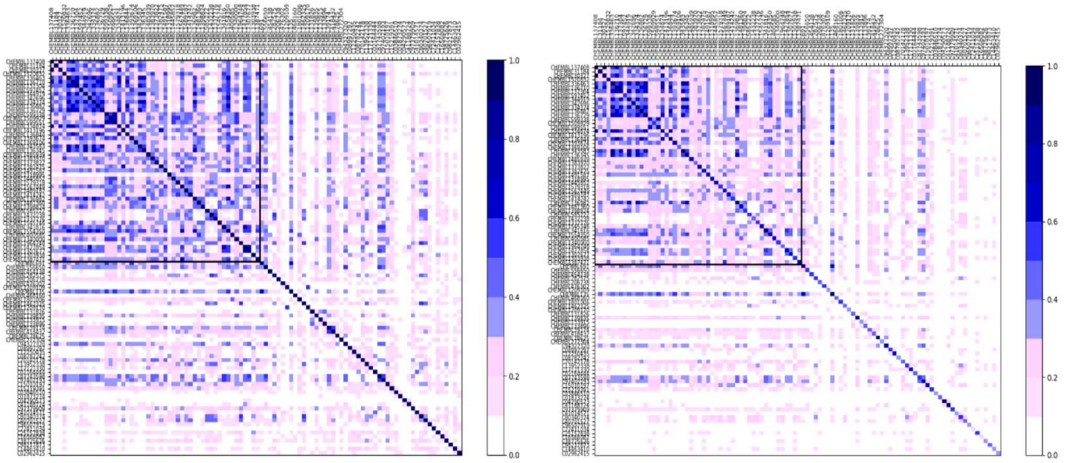


Figure 4-31. The heatmap of comparison of AHR ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of AHR ligands.

The shapes of AHR ligands (**Figure 4-31**) are very similar to each other and differ clearly from the ligands of the rest of the receptors. The similarity between compounds is maintained even when using the AV method to compare the conformations of the molecules, which shows that the generated conformations by OMEGA are quite similar to each other.

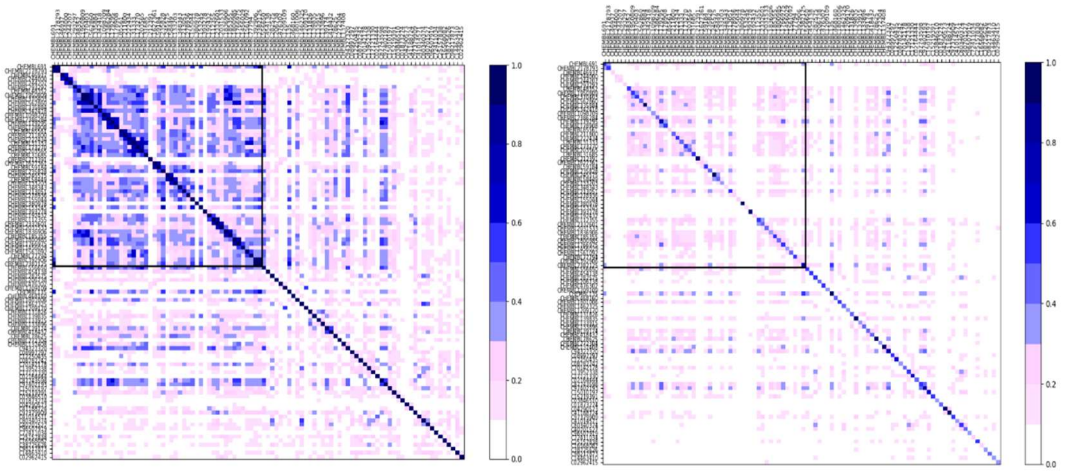


Figure 4-32. The heatmaps of comparison of ER-alpha ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of ER-alpha ligands.

In the case of ER-alpha (**Figure 4-32**), the strong similarities between shapes of its ligands can be observed only when using the MV method, while the AV method does not show any differentiation between the ER-alpha ligands and the ligands of other receptors. It is worth noting that ligands that bind to the ER-beta also have a high Shape Tanimoto with most of the ER-alpha ligands, which is expected as both isoforms have similar binding pockets and therefore their ligands can share similar shape.

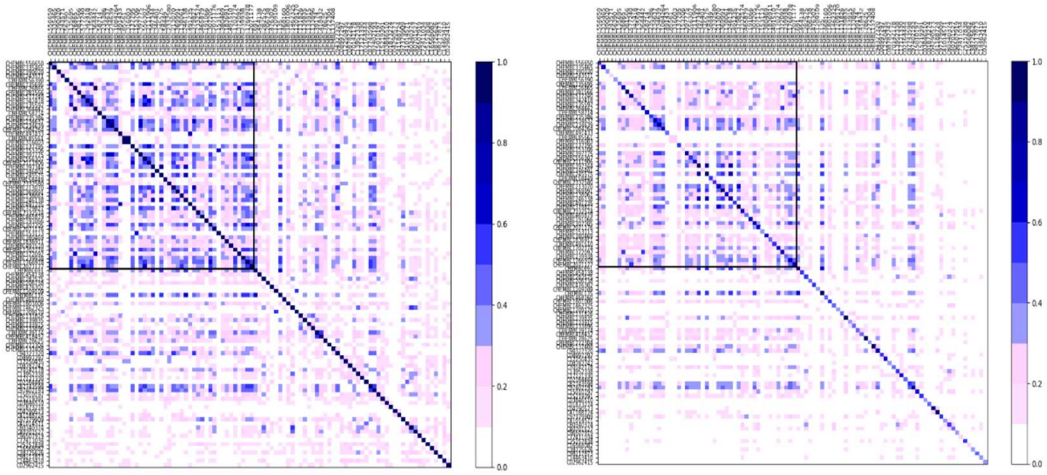


Figure 4-33. The heatmaps of comparison of ER-beta ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of ER-beta ligands.

The ER-beta ligands (**Figure 4-33**) do not share STs that are as high as with ER-alpha, when using the MV method. However, they show much higher similarity when taking the average value of Shape Tanimotos. This suggests that conformations of ER-beta ligands have similar shape and that there might be some more conformational variation tolerated in the ER-beta than in the ER-alpha.

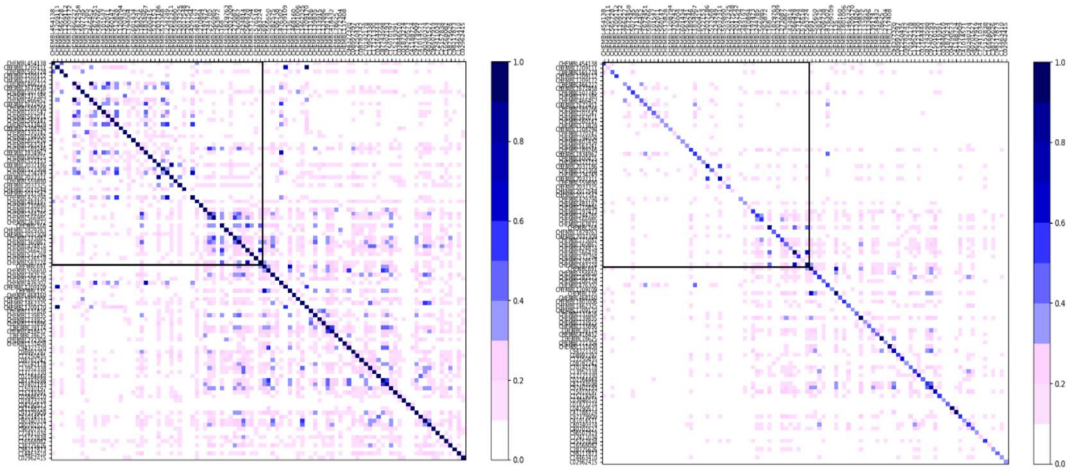


Figure 4-34. The heatmaps of comparison of FXR ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of FXR ligands.

The FXR ligands (**Figure 4-34**) clearly do not show much similarity in shape using either of the applied methods and cannot be distinguished from ligands of other receptors. This could be expected, as the previous study from this chapter showed really low similarities amongst the shape of FXR ligands. A similar situation can be observed in the case of other receptors: GR (**Figure 4-35**), LXR-alpha (**Figure 4-36**), LXR-beta (**Figure 4-37**), PPAR-alpha (**Figure 4-38**), PPAR-delta (**Figure 4-39**) and PPAR-gamma (**Figure 4-40**). The ligands do not have many commonalities in shape and are difficult to distinguish from ligands of other receptors.

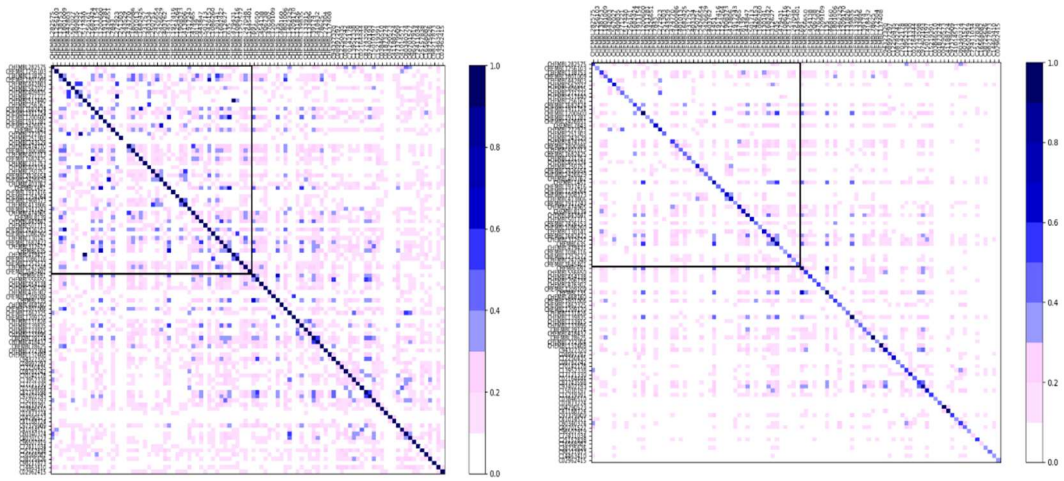


Figure 4-35. The heatmaps of comparison of GR ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The

darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of GR ligands.

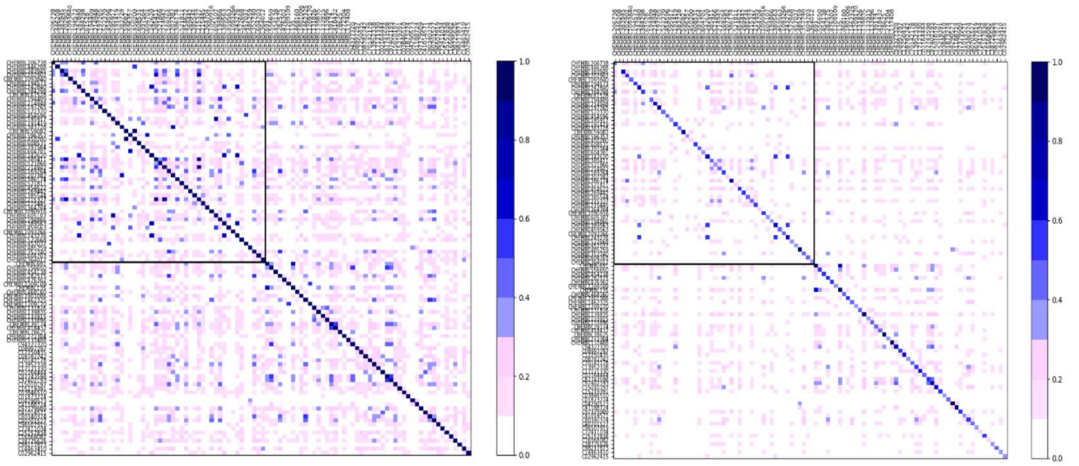


Figure 4-36. The heatmaps of comparison of LXR-alpha ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of LXR-alpha ligands.

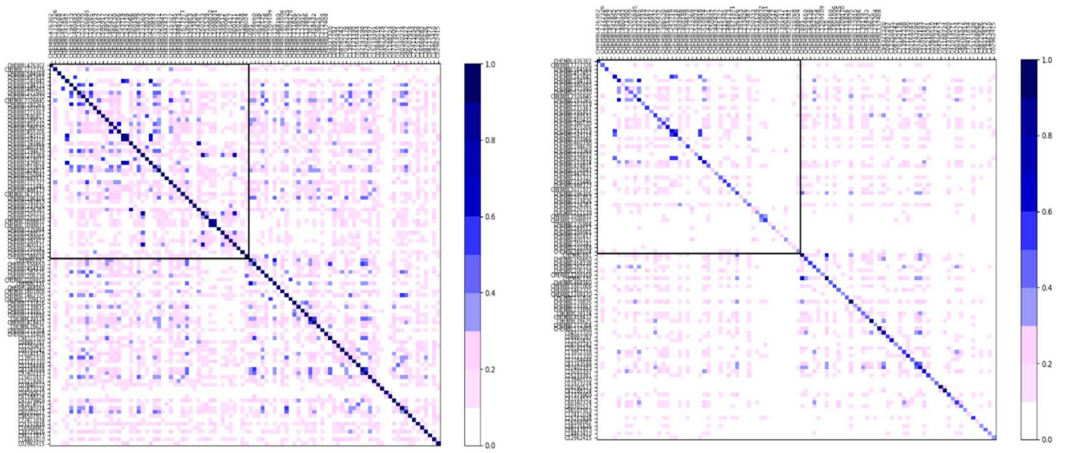


Figure 4-37. The heatmaps of comparison of LXR-beta ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of LXR-beta ligands.

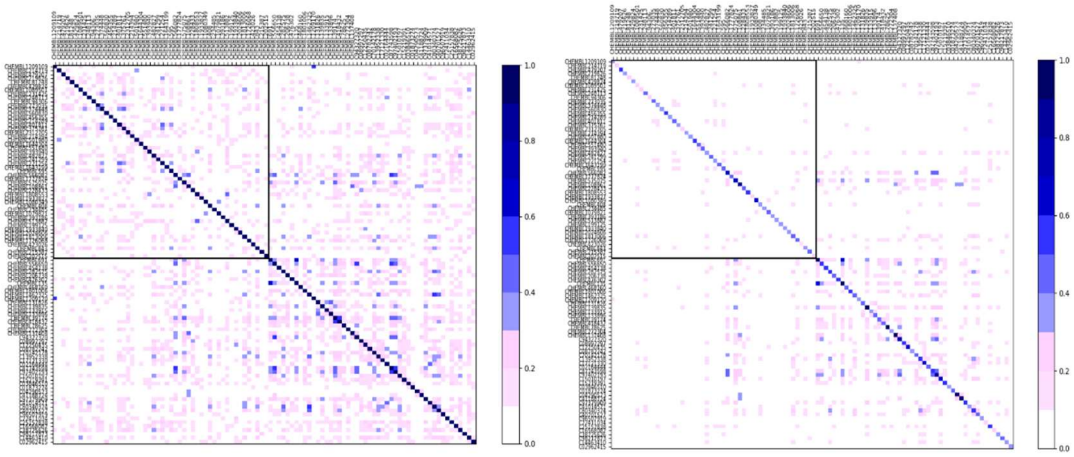


Figure 4-38. The heatmaps of comparison of PPAR-alpha ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of PPAR-alpha ligands.

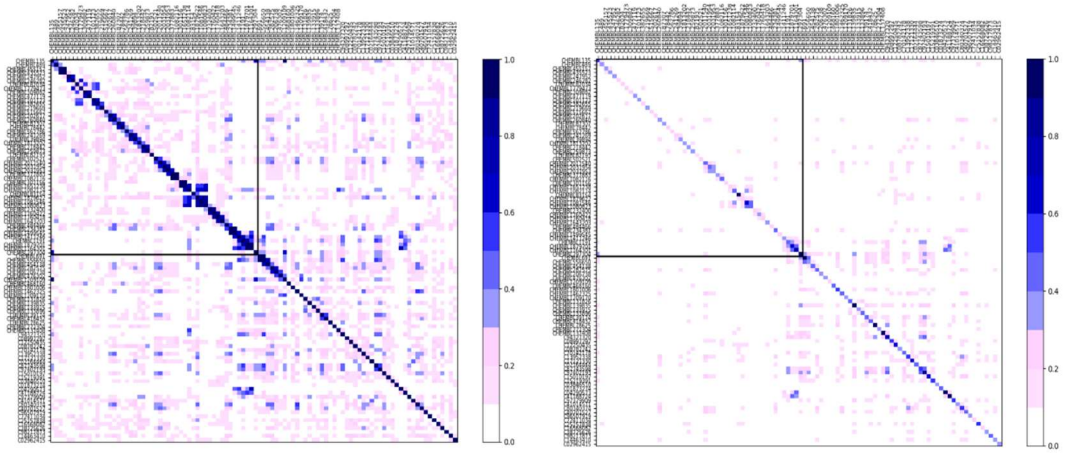


Figure 4-39. The heatmaps of comparison of PPAR-delta ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of PPAR-delta ligands.

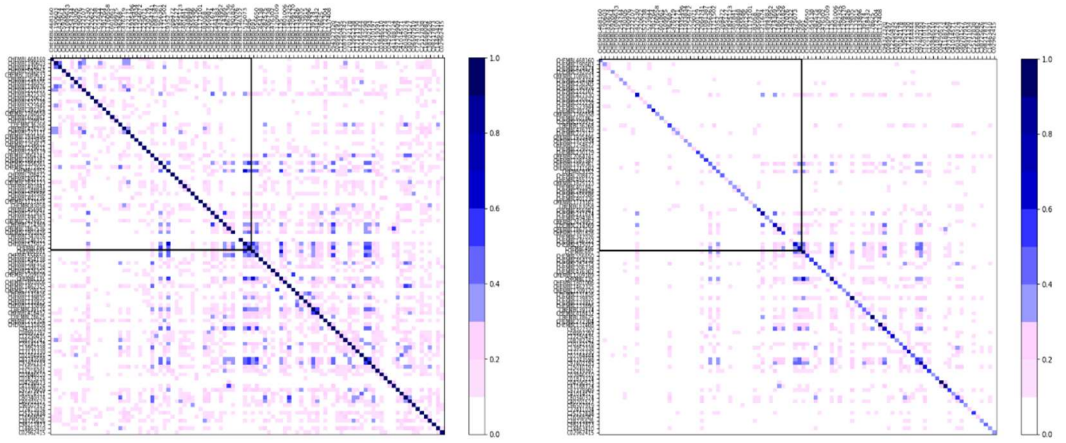


Figure 4-40. The heatmaps of comparison of PPAR-gamma ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of PPAR-gamma ligands.

In the case of PR (**Figure 4-41**), PXR (**Figure 4-42**) RAR-alpha (**Figure 4-43**), RAR-beta (**Figure 4-44**), RAR-gamma (**Figure 4-45**), RXR-alpha (**Figure 4-46**), RXR-beta (**Figure 4-47**) and RXR-gamma (**Figure 4-48**), there is some barely noticeable similarity between ligands however it is small and it is not easy to see the difference from other receptor ligands and decoys.

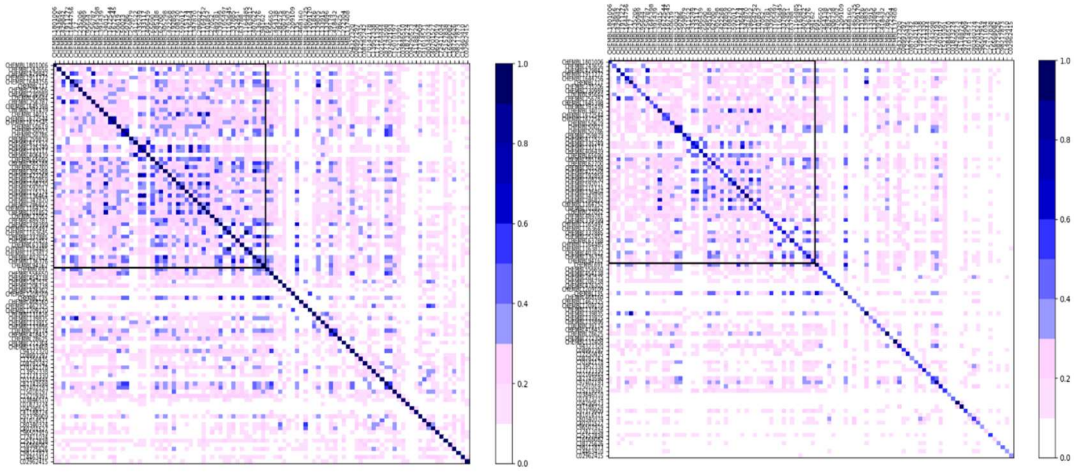


Figure 4-41. The heatmaps of comparison of PR ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of PR ligands.

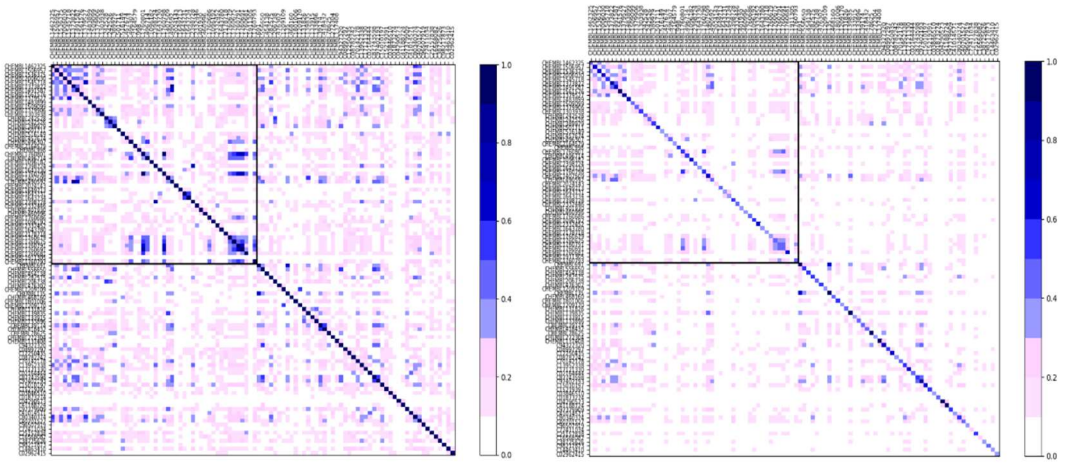


Figure 4-42. The heatmaps of comparison of PXR ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of PXR ligands.

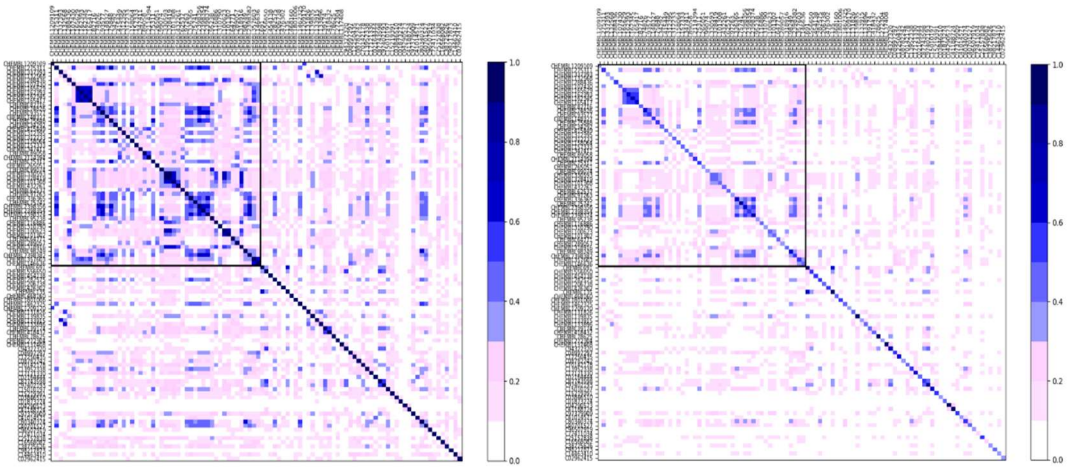


Figure 4-43. The heatmaps of comparison of RAR-alpha ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of RAR-alpha ligands.

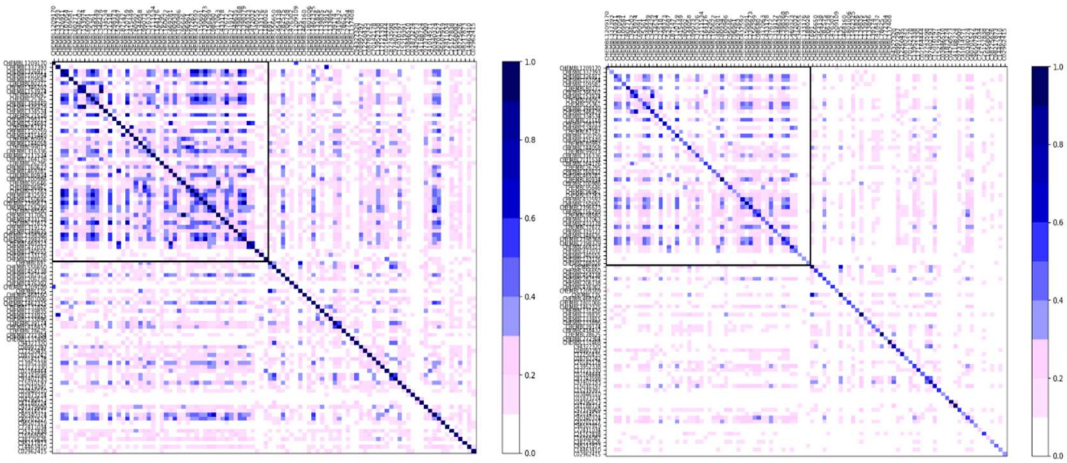


Figure 4-44. The heatmaps of comparison of RAR-beta ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of RAR-beta ligands.

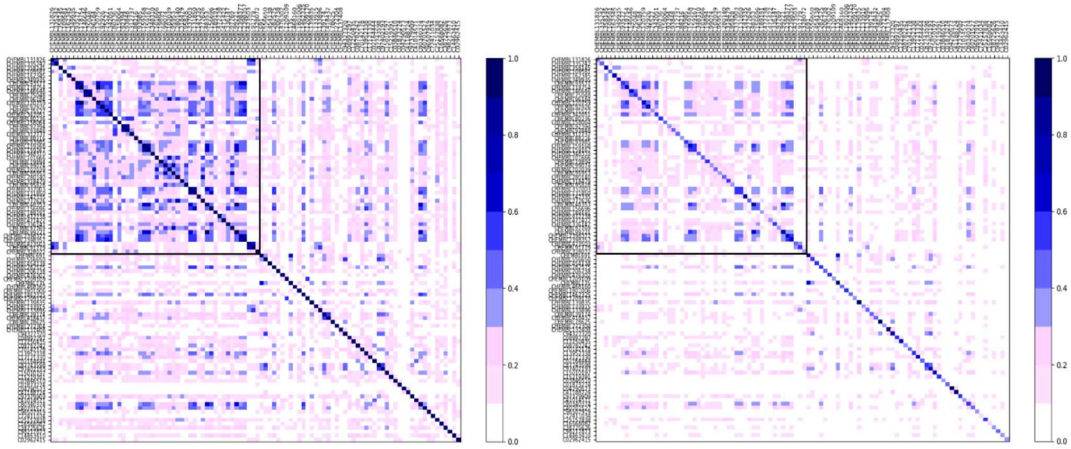


Figure 4-45. The heatmaps of comparison of RAR-gamma ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of RAR-gamma ligands.

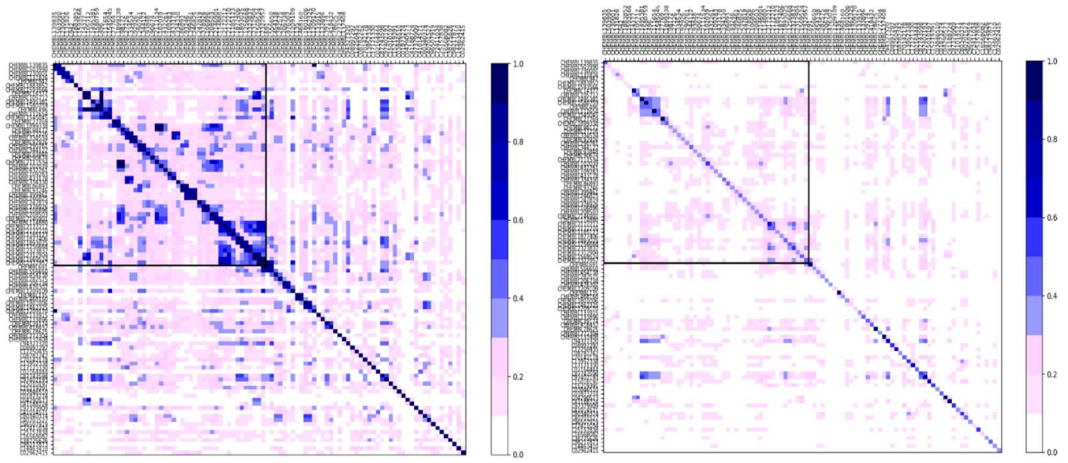


Figure 4-46. The heatmaps of comparison of RXR-alpha ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of RXR-alpha ligands.

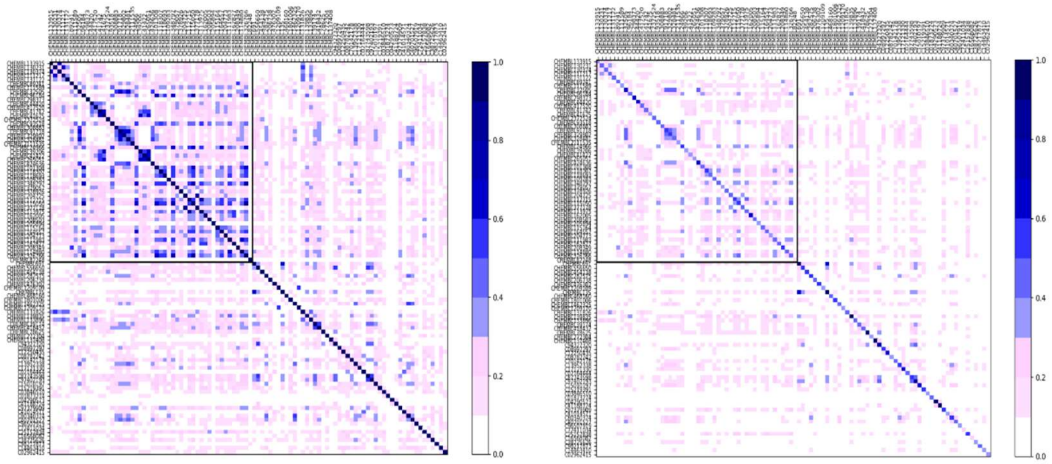


Figure 4-47. The heatmaps of comparison of RXR-beta ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of RXR-beta ligands.

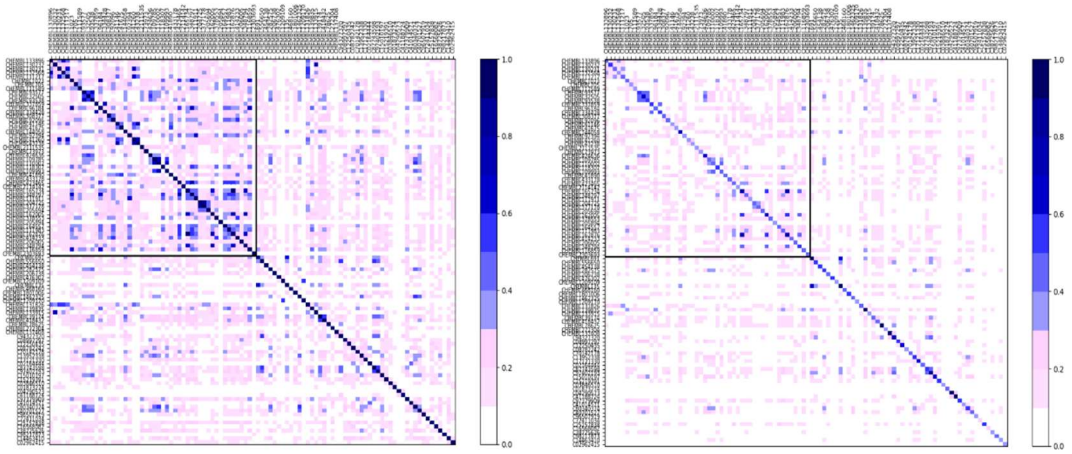


Figure 4-48. The heatmaps of comparison of RXR-gamma ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of RXR-gamma ligands.

The THR-alpha (**Figure 4-49**), THR-beta (**Figure 4-50**) and VDR (**Figure 4-52**) ligands clearly show high shape similarity and are easily distinguished by the shape fingerprints method from other ligands and decoys.

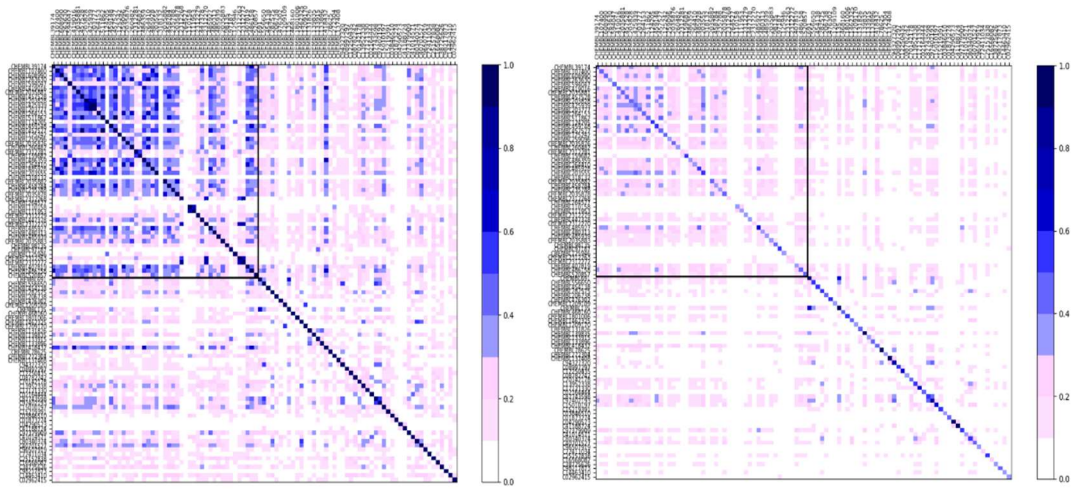


Figure 4-49. The heatmaps of comparison of THR-alpha ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of THR-alpha ligands.

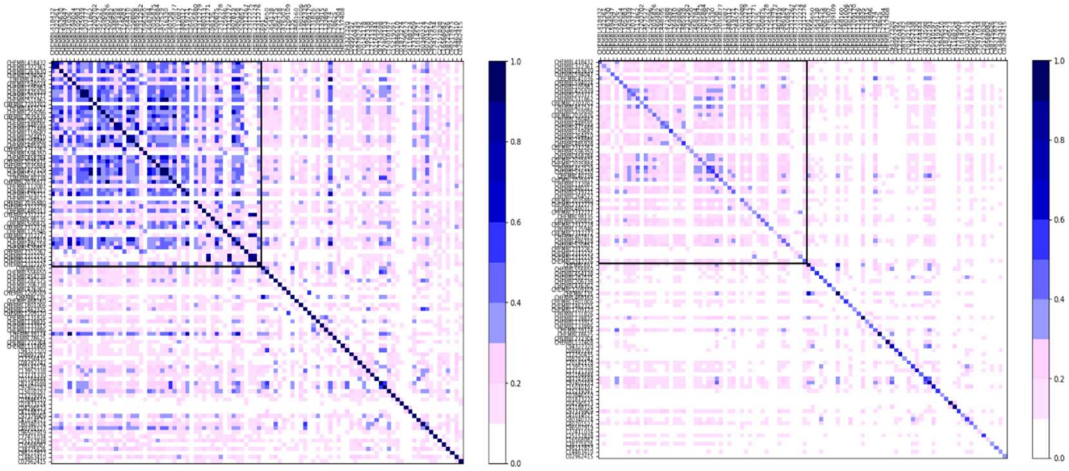


Figure 4-50. The heatmaps of comparison of THR-beta ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of THR-beta ligands.

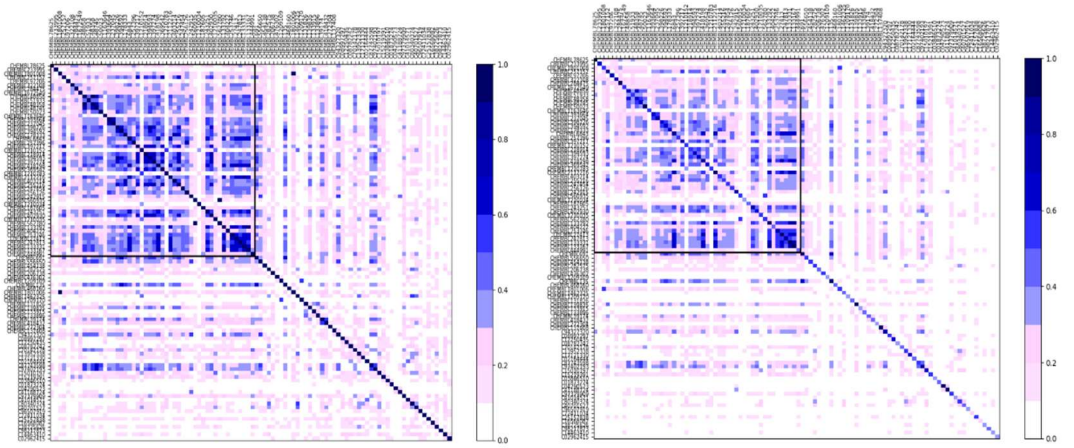


Figure 4-51. The heatmaps of comparison of TR ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of TR ligands.

The VDR ligands are one of the easiest to distinguish from ligands of other receptors, but only when using the MV method of comparison. The calculated STs are high for VDR ligands and really small, close to 0-0.2 ST when compared with shapes of other ligands. In the case of the AV method this is not that easily noticeable. The values are much smaller, however so are the ST values for other ligands – almost all the

comparison between VDR ligands and other receptor ligands have values 0, which indicates complete dissimilarity.

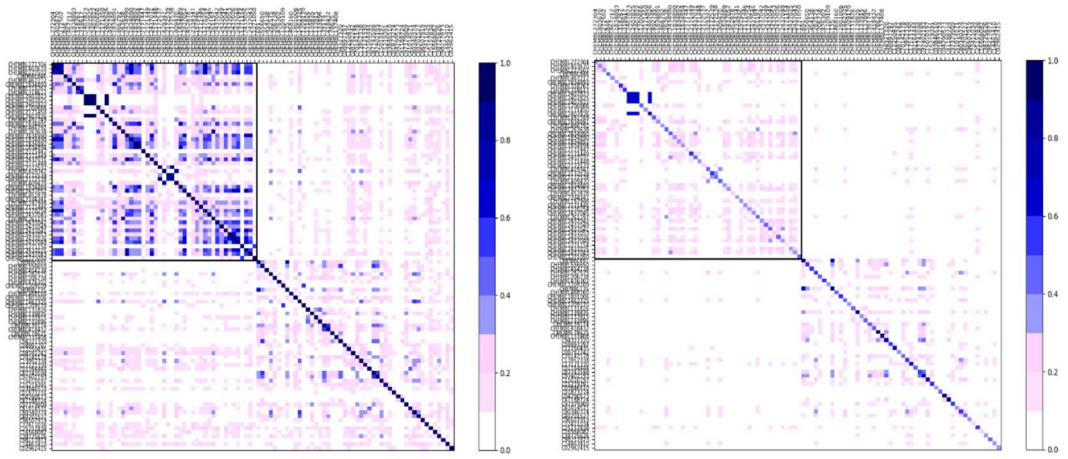


Figure 4-52. The heatmaps of comparison of VDR ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right). The darker the colour, the higher the shape similarity detected by shape fingerprints method. The box (top left) encloses the set that binds of VDR ligands.

These sets of data have been analysed with ROC curves. As can be seen in the **Table 4-4**, the highest AUC values are obtained for AHR, PR, TR. This indicates that the discrimination between ligands binding to these receptors and the set of other ligands and decoys, is the highest in these sets. The values are higher than the ones calculated for Test Sets 1 and 2 in chapter 2. Those were equal to 0.61 and 0.61 for Test Set 1 when using AV and MV approaches, respectively and 0.77 and 0.78 for Test Set 2 when using AV and MV methods, respectively. The AUC values results are complemented by logistic regression plots for each of the NRs. Some of the plots, like **Figure 4-56**, **Figure 4-57**, **Figure 4-58**, **Figure 4-59**, **Figure 4-60**, **Figure 4-61** and **Figure 4-62**, clearly shows the poor discrimination between true positives (ligands) and true negatives (decoys).

Table 4-4. The comparison of AUC values for all NRs using both methods: AV and MV.

NR	AV	MV
AHR	0.83	0.81
ER-ALPHA	0.63	0.72
ER-BETA	0.63	0.64
FXR	0.56	0.54
GR	0.52	0.52
LXR-ALPHA	0.52	0.50
LXR-BETA	0.49	0.51
PPAR-ALPHA	0.55	0.61
PPAR-DELTA	0.59	0.53
PPAR-GAMMA	0.60	0.57
PR	0.76	0.74
PXR	0.56	0.55
RAR-ALPHA	0.65	0.67
RAR-BETA	0.63	0.64
RAR-GAMMA	0.67	0.69
RXR-ALPHA	0.68	0.70
RXR-BETA	0.65	0.67
RXR-GAMMA	0.63	0.65
THR-ALPHA	0.66	0.69
THR-BETA	0.68	0.71
TR	0.78	0.77
VDR	0.63	0.70

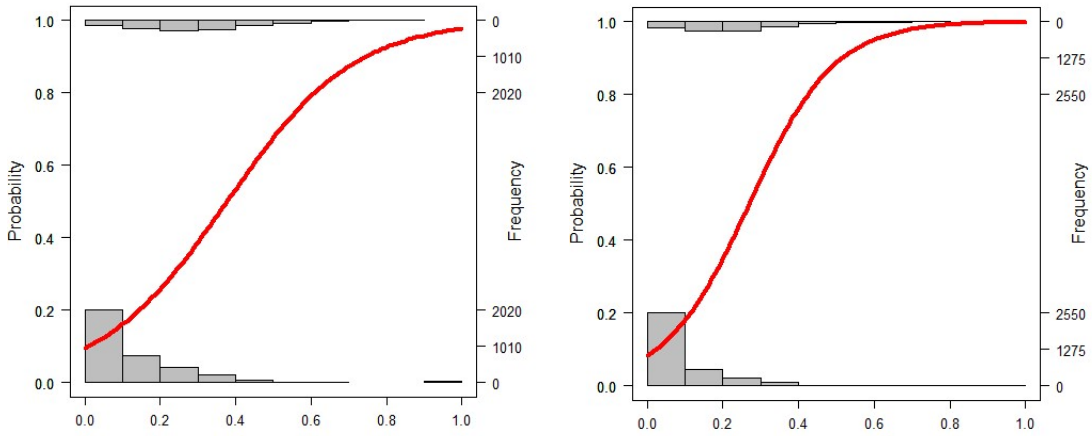


Figure 4-53. The logistic regression plots of comparison of AHR ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

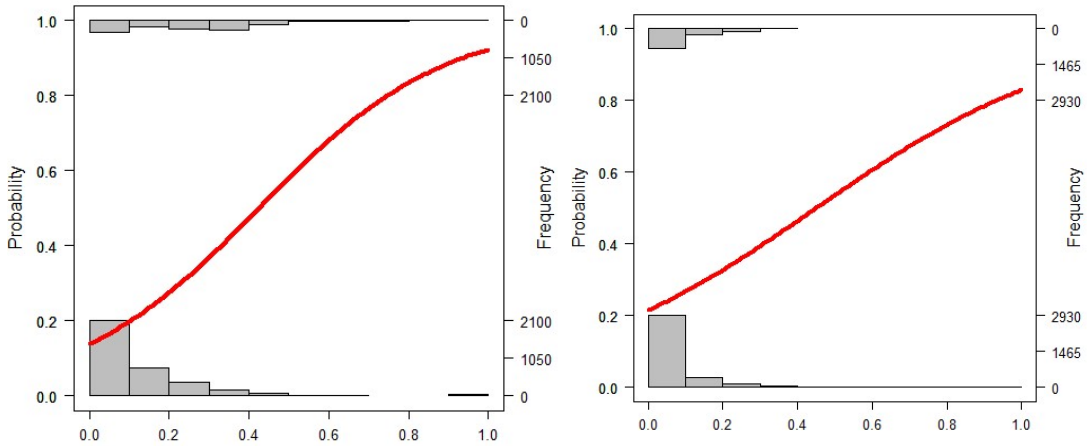


Figure 4-54. The logistic regression plots of comparison of ER-alpha ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

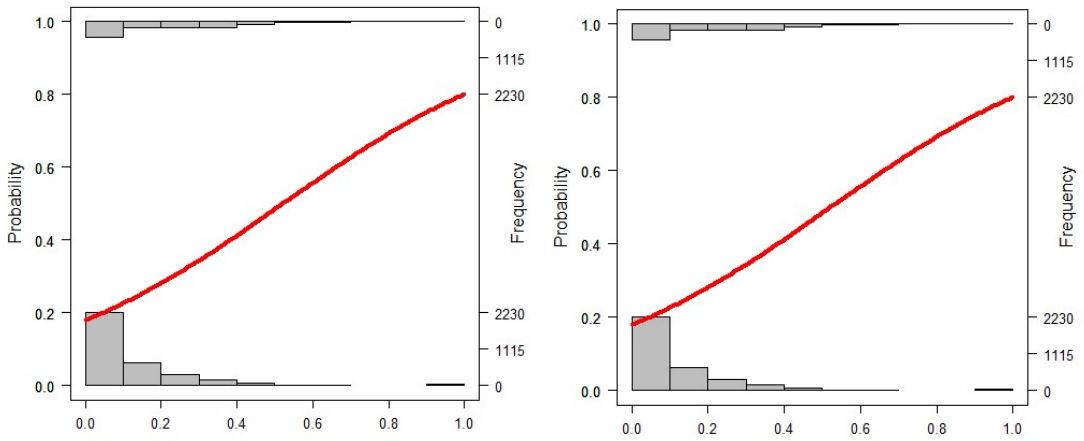


Figure 4-55. The logistic regression plots of comparison of ER-beta ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

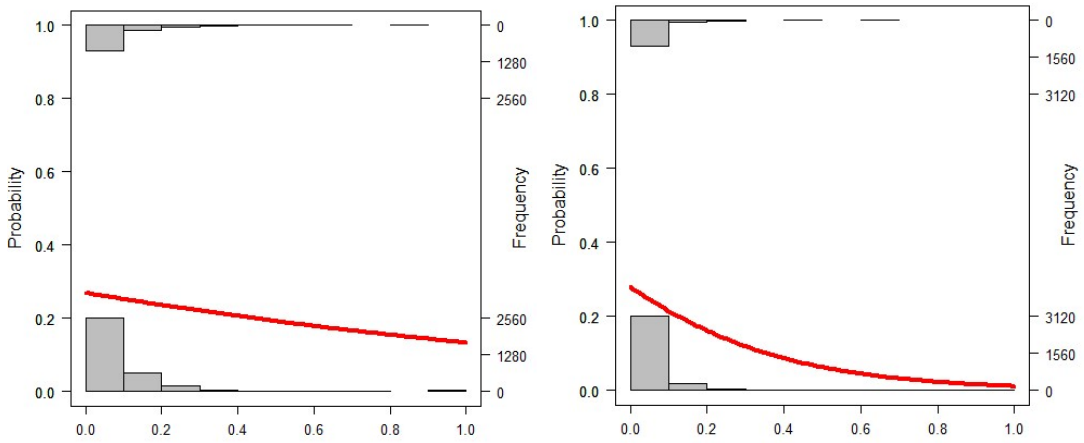


Figure 4-56. The logistic regression plots of comparison of FXR ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

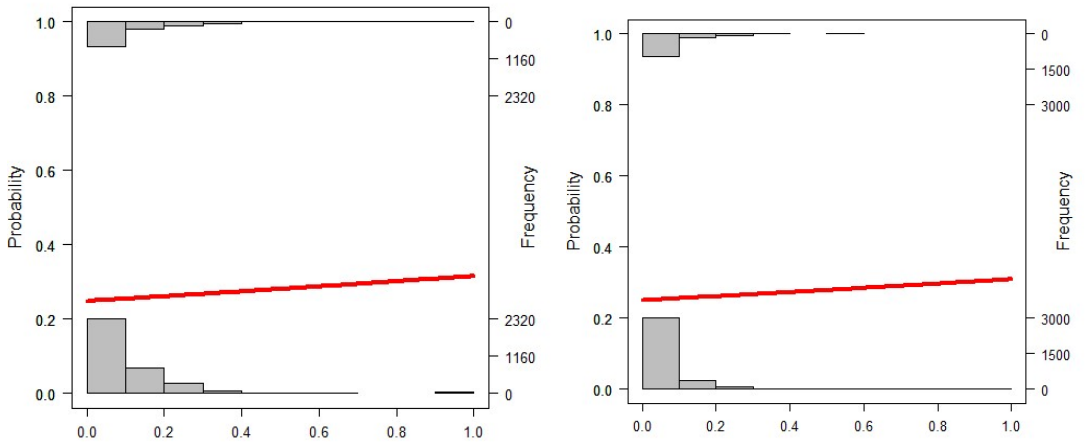


Figure 4-57. The logistic regression plots of comparison of GR ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

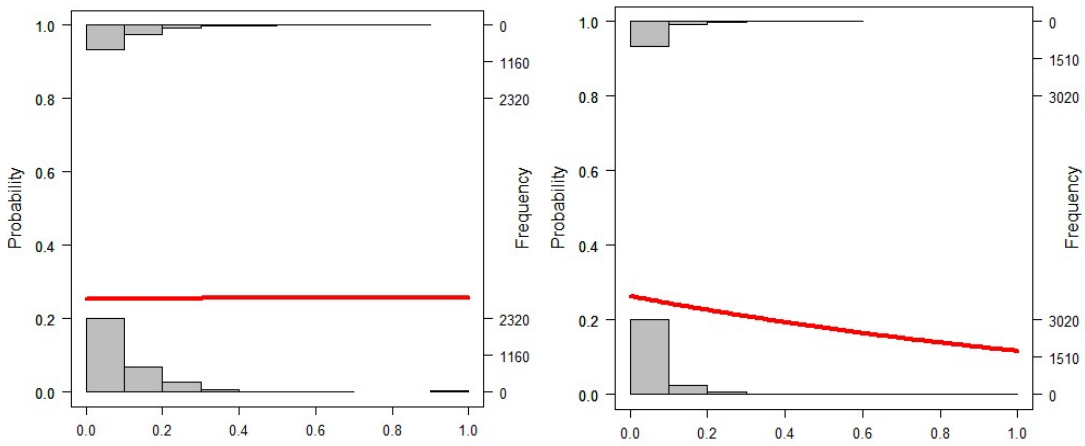


Figure 4-58. The logistic regression plots of comparison of LXR-alpha ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

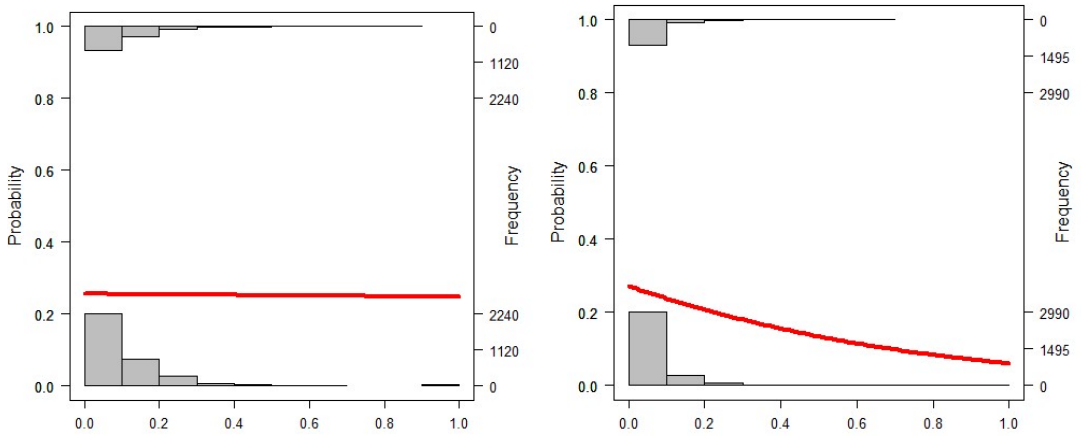


Figure 4-59. The logistic regression plots of comparison of LXR-beta ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

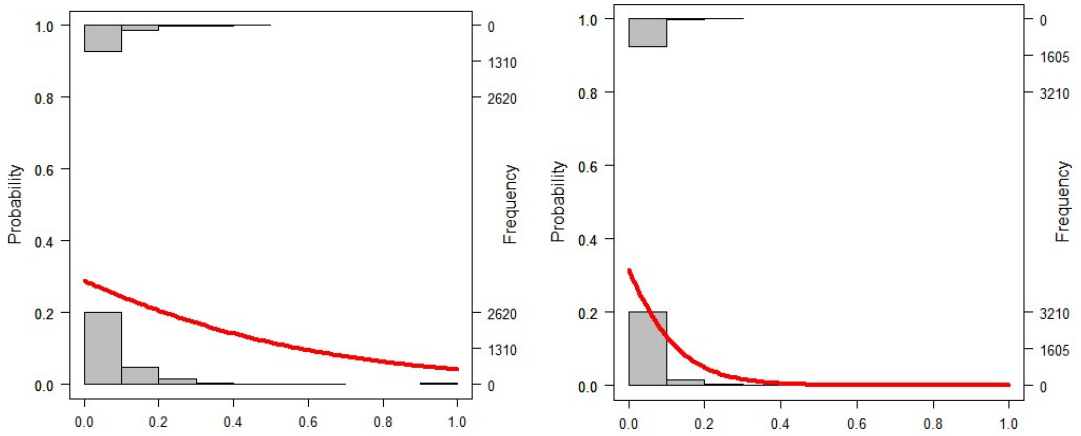


Figure 4-60. The logistic regression plots of comparison of PPAR-alpha ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

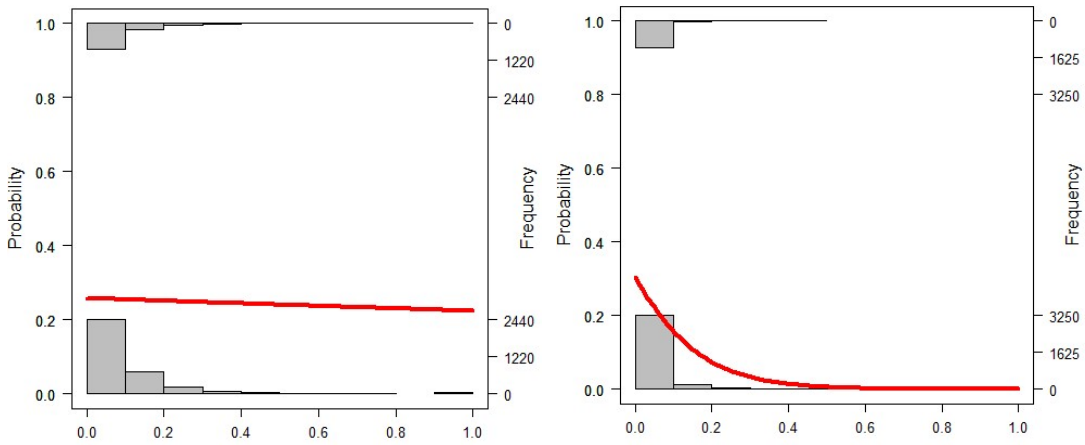


Figure 4-61. The logistic regression plots of comparison of PPAR-delta ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

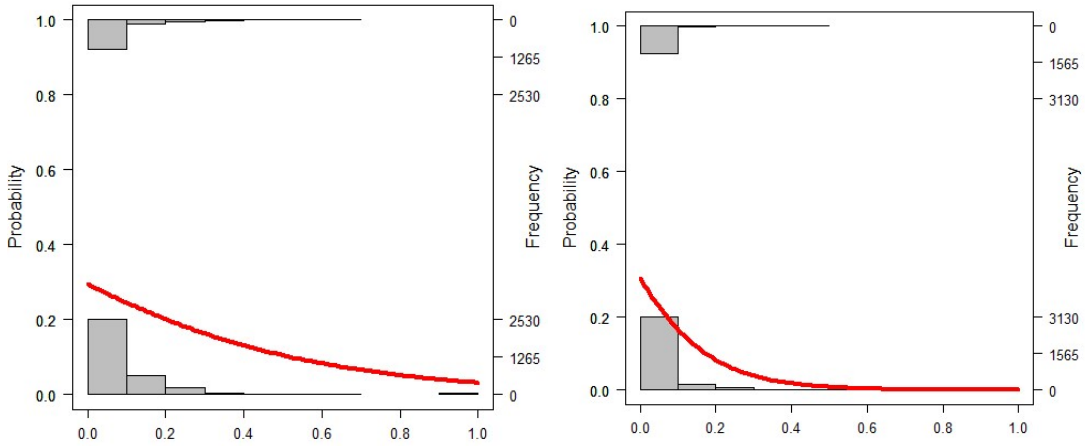


Figure 4-62. The logistic regression plots of comparison of PPAR-gamma ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

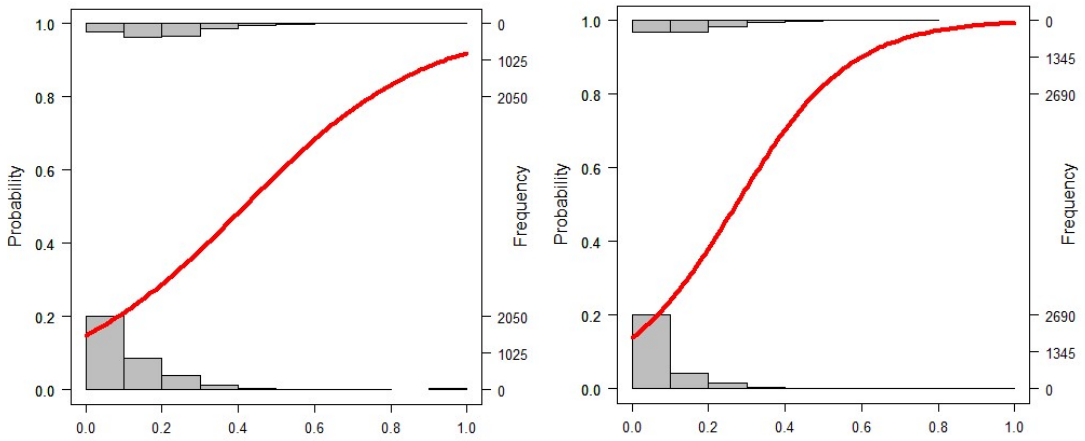


Figure 4-63. The logistic regression plots of comparison of PR ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

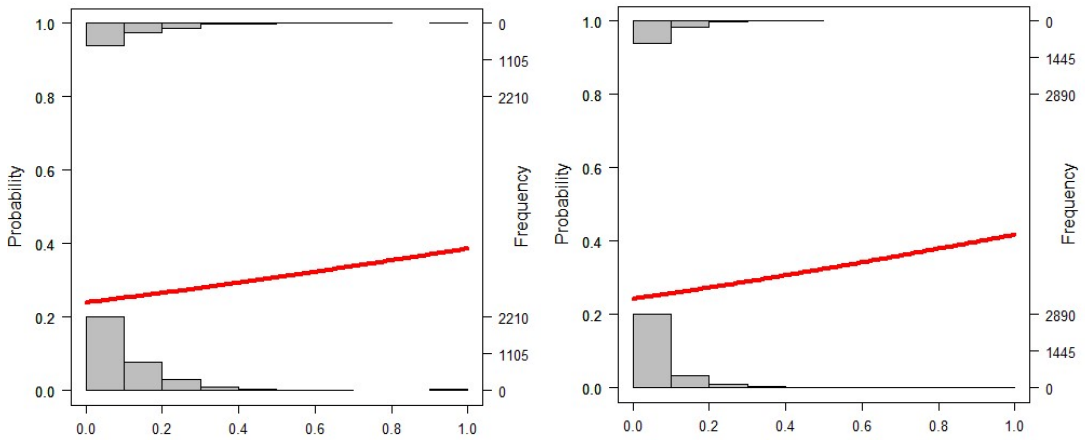


Figure 4-64. The logistic regression plots of comparison of PXR ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

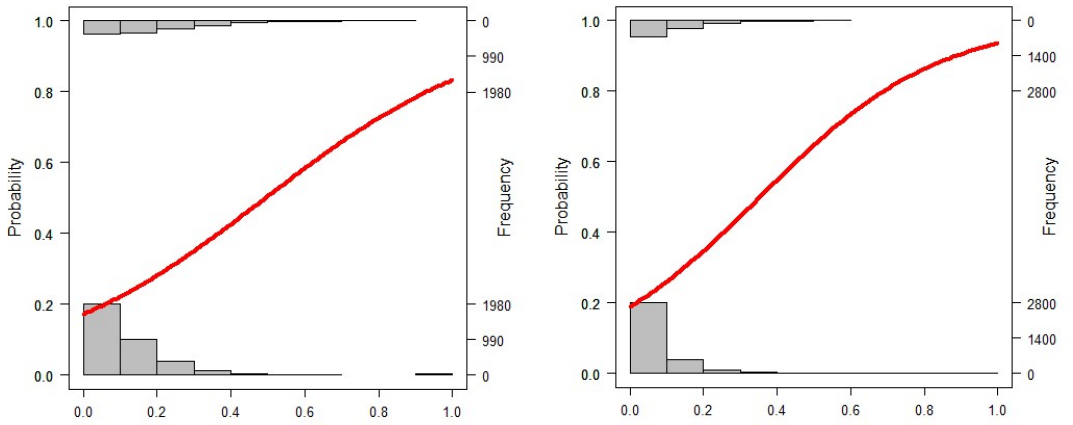


Figure 4-65. The logistic regression plots of comparison of RAR-alpha ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

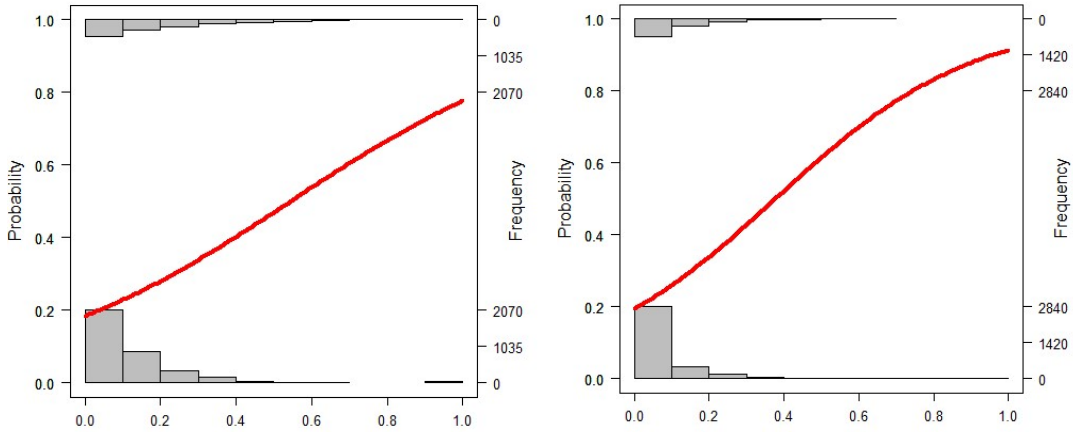


Figure 4-66. The logistic regression plots of comparison of RAR-beta ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

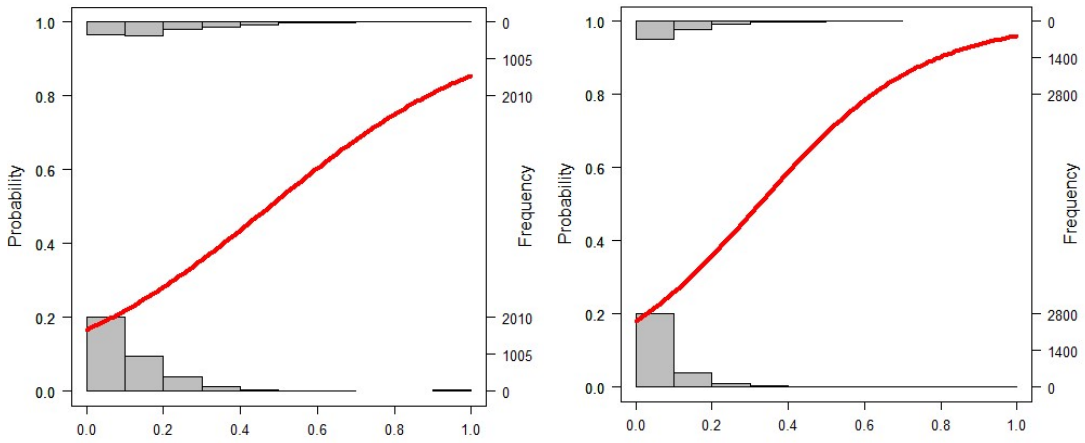


Figure 4-67. The logistic regression plots of comparison of RAR-gamma ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

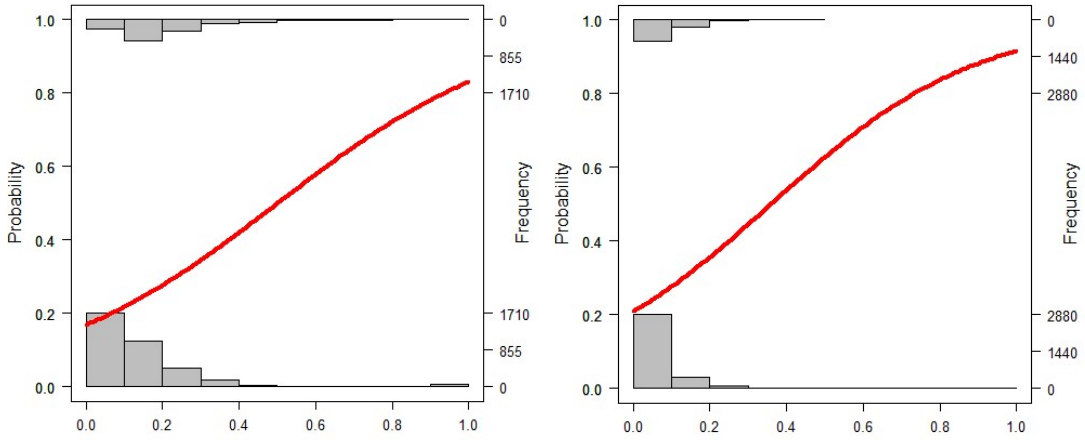


Figure 4-68. The logistic regression plots of comparison of RXR-alpha ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

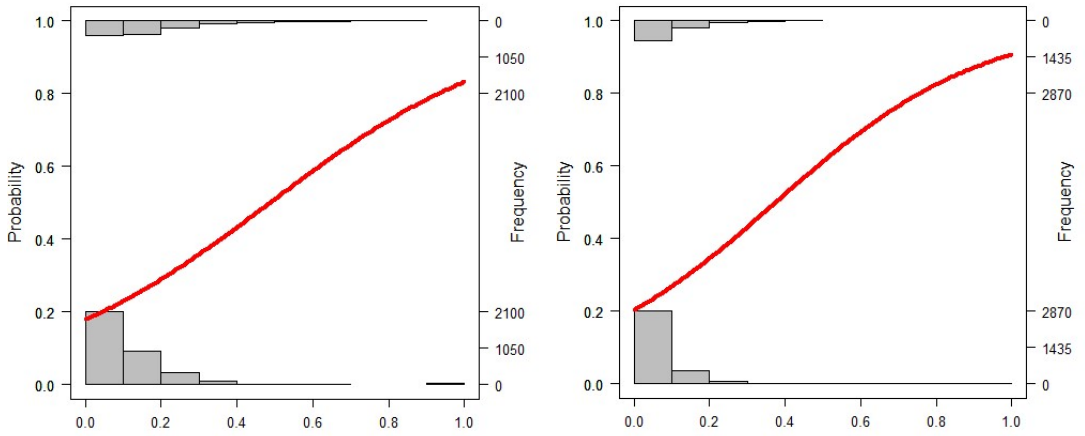


Figure 4-69. The logistic regression plots of comparison of RXR-beta ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

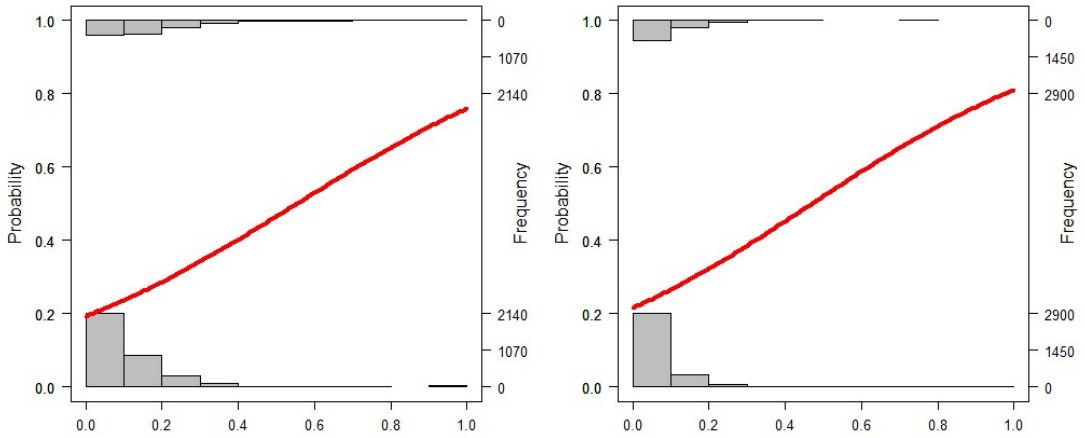


Figure 4-70. The logistic regression plots of comparison of RXR-gamma ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

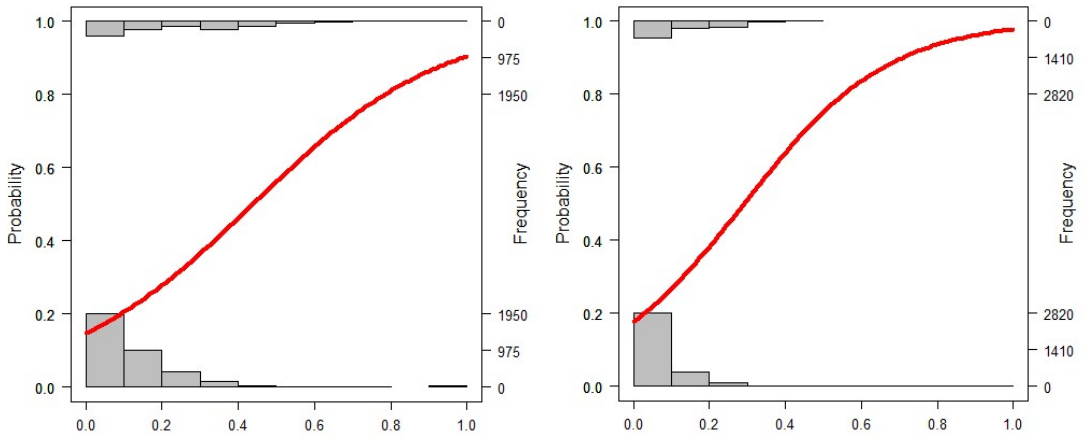


Figure 4-71. The logistic regression plots of comparison of THR-alpha ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

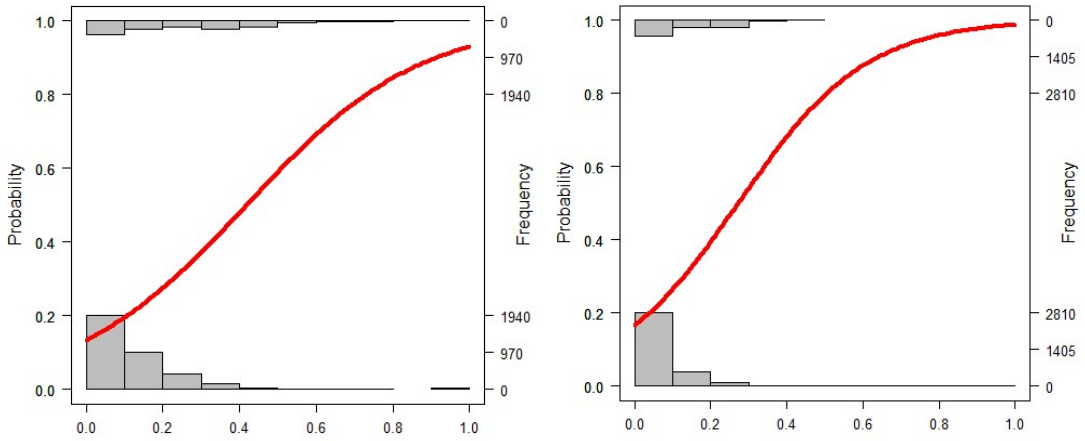


Figure 4-72. The logistic regression plots of comparison of THR-beta ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

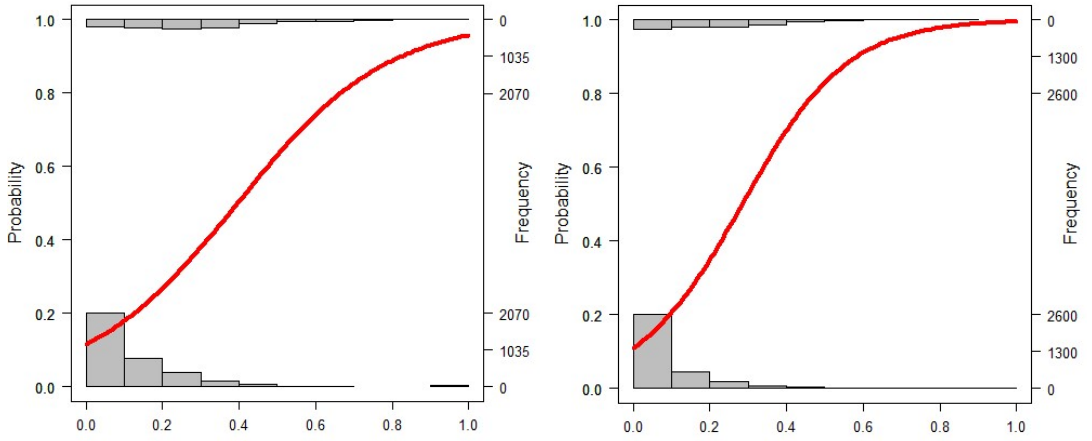


Figure 4-73. The logistic regression plots of comparison of TR ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

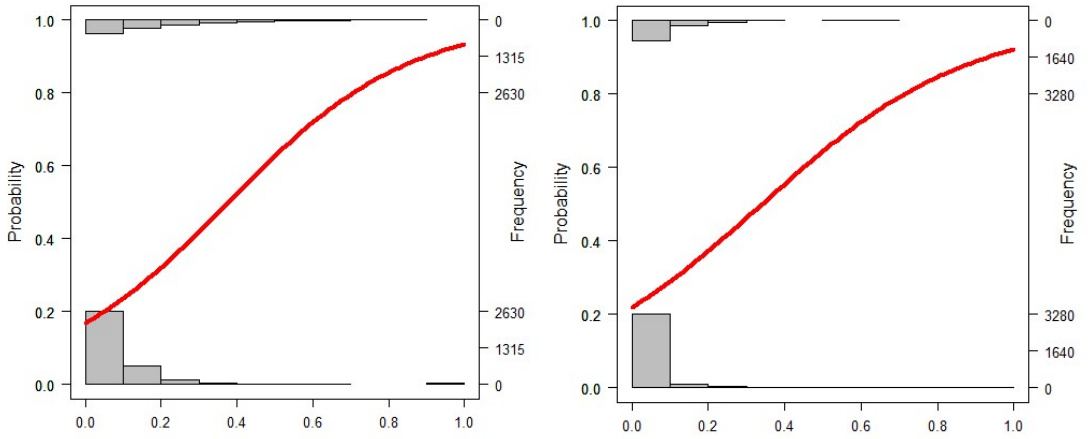


Figure 4-74. The logistic regression plots of comparison of VDR ligands with ligands of other receptors with decoys when using MV (on the left) and AV method (on the right).

4.6. Conclusions

As presented in this chapter, the shape fingerprints method can be successfully applied to solve many problems in the chemistry world. This includes predictions of solubility similarly to well-established approaches, virtual screening and shape similarity searches.

The results presented in this chapter showed that the shape fingerprint method could be successfully applied not only to small sets, like Test Set 1 and Test Set 2 from chapter 2, but also to much larger sets such as DUD-E diverse set (consisting of 8 targets with 1774 ligands in total). Grouping compounds from DUD-E diverse set based on the similar biological activity performed a little worse than in case of smaller sets. However, the obtained AUC value of 0.55 could have been expected as the size of the presented set is much greater than used previously.

The AUC values obtained from virtual screening on three sets taken from DUD-E (AMPC, COMT and CXCR4), consisting of ligands and a series of decoys, reveal the potential of shape fingerprints application as a ligand-based virtual screening technique. The results: 0.57, 0.66 and 0.62 for CXCR4, AMPC and COMT sets respectively are decent considering the size of the sets and the high ratio of decoys per ligand in each set.

Another potential application of shape fingerprints shown in this chapter is the solubility prediction. The predicted logS values are comparable (and in some case the predicted values are much closer to experimental ones) to those obtained using well-established prediction software. As it produces similar results to MOE, it could be used simultaneously with it to predict quite accurately the values of logS. The approach could be improved by including also the chemistry of compounds (using Tanimoto Combo score instead of Shape Tanimoto in generating Shape Database and shape fingerprints). This might reduce number of poorly predicted solubility values for molecules that solubility measurements could be affected by too strong interaction of molecules with solute or too strong intermolecular interactions in crystal lattice and therefore are difficult to predict with shape-only techniques.

Applying shape fingerprint method to 22 sets of NRs ligands showed in which NRs the shape plays crucial role. Among all of the NRs, the strongest shape similarity between ligands is visible in the case of AHR, ER-alpha, ER-beta PR, THR-alpha, THR-beta. Therefore, these sets performed well also in virtual screening of NRs – they had high AUC values and were easily distinguished from ligands of other NRs and decoys. In the case of TR, PXR, RAR-alpha, RAR-beta, RAR-gamma, RXR-alpha, RXR-beta, RXR-gamma and VDR it can be observed that there are a few groups of ligands which are similar in shape, but this similarity is not shared across the whole

set of molecules. The lack of shape similarity can be noticed for FXR, LXR-alpha and LXR-beta.

Chapter 5

Matched Molecular Pairs

5.1. Introduction

Matched Molecular Pairs Analysis (MMPA) is a widely used approach to screen large databases in order to find pairs of molecules with a common structural part but differing by a small, well-defined change in structure and with a known change in properties.[52,51,49,50] This approach assumes that the change of properties is more easily predicted than the absolute values of those properties for each molecule alone. A matched molecular pair (MMP) involves two compounds that have a common core and have a different fragment R as show in the **Figure 5-1**.

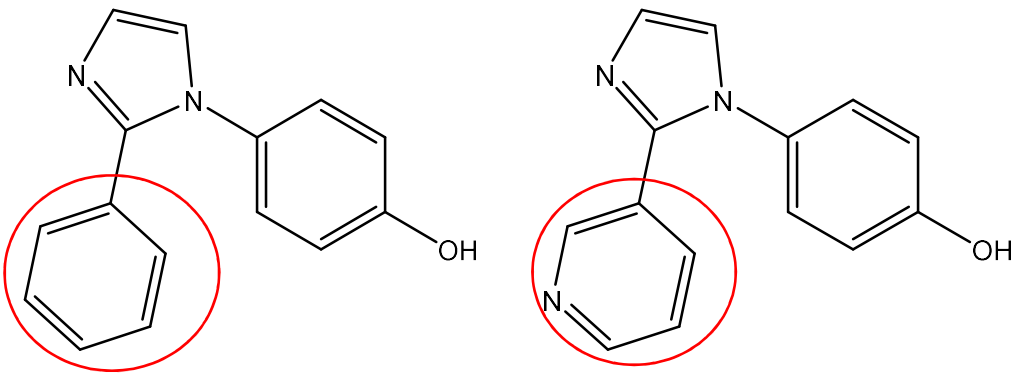


Figure 5-1. An example of pair of MMPs with the changing fragment part marked in red circle.

There are two common approaches used to find MMPs: Fragment and Index (F+I)[53] and Maximum Common Substructure (MCSS).[54] The comparison of how both methods find the MMPs can be found in the **Figure 5-2**.

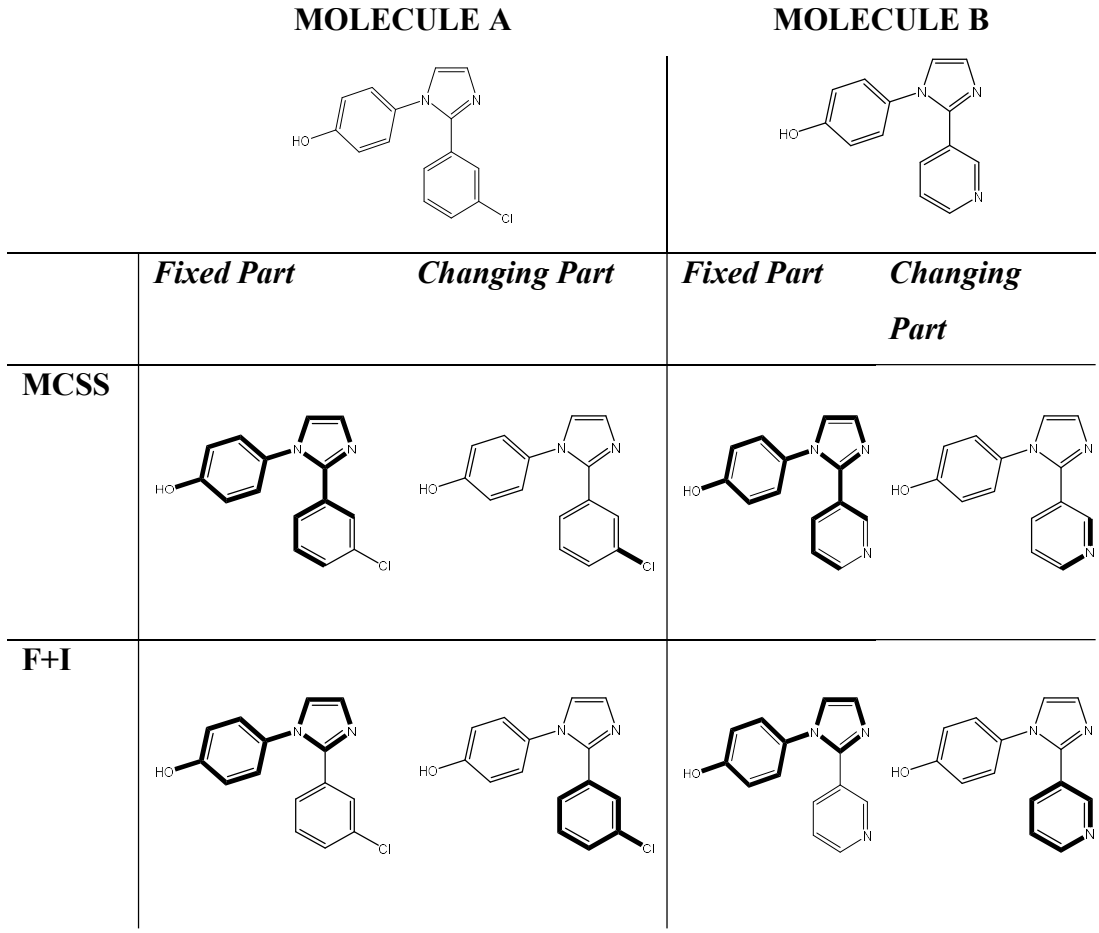


Figure 5-2. The MMP, Molecule A and B, identified by two methods: Fragment and Index (F+I) and Maximum Common Substructure (MCSS) with shown fixed and changing parts.

In the MCSS approach, as used in the WizePairZ algorithm described by Warner et al.,[54] two molecules are compared and the maximum common substructure shared by them is identified – it is called the fixed part or the core. The remaining part is called the changing part as can be seen in the example in **Figure 5-2**, where the structural change is the change from C-Cl to N. In order to compare the molecules using this approach, the molecules are converted into graphs, which enables identifying common substructures between compounds. The structural change identified by the MCSS approach is encoded as SMIRKS, which is a reaction transform languagge.[55]

The disadvantages of the MCSS approach are that the substructure comparison is slow and that most algorithms for finding the MCSS require all of the atoms to be contiguous and therefore prevent pairs in which linkers change from being found.

A second approach was introduced by Hussain and Rea.[53] The algorithm works by generating fragments of the molecule based on predefined rules and then indexing those fragments, as shown in the example in **Figure 5-2**, where the structural change is from Clc1ccccc1 to c1ccnc1. The generated fragments are stored as key – value pairs.[53,56]

In the first step, each molecule is fragmented, by breaking a selected single bond (or bonds). Hussain and Rea[53] used a SMARTS pattern to define the bonds that could be broken and these are limited to acyclic single bonds. The resulting fragments are stored as SMILES strings, which can be manipulated as text. This is a great advantage of the approach: after initial fragmentation, all subsequent steps toward the identification of matched pairs involve only rapid text processing. Another advantage comes from the fact that once a molecule has been fragmented, it can be added to the database, which is then available for comparison to any new molecule or to find new MMPs.

Among the limitations of the fragment and index approach to finding matched pairs are that small changes to rings cannot be identified as pairs, highly substituted core changes are limited by the number of fragmentations considered, and the diversity of the structural changes can be limited by the restrictions that are imposed (usually heavy atom counts or ratio). One specific set of changes that are not readily identified is modifications to macrocyclic rings.

5.2. Methods

In order to find matched molecular pairs, the software developed by MedChemica – MCPairs[129] was used. It has implemented two approaches of finding MMPs: the F+I method described by Hussain and Rea[53] and the MCSS algorithm described by Warner et al.[54] The structural change linking pairs of molecules is encoded by SMIRKS, and these are modified to include differing levels of chemical context. The

chemical context is encoded using SMILES for the atoms in the fixed part that are connected to the changing part. There are four context levels, as shown in **Figure 5-3**.

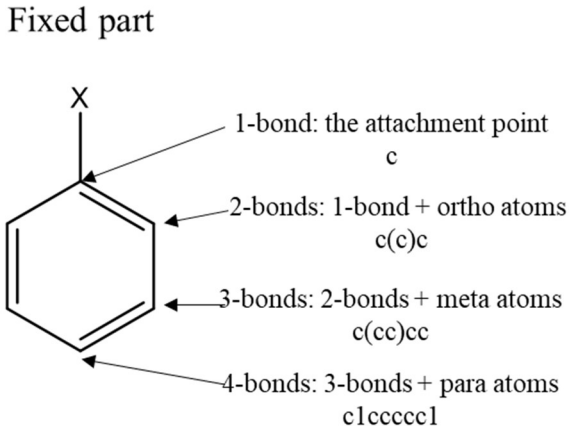


Figure 5-3. Definition of the chemical context for a matched molecular pair.

Both methods need the heavy atom fractions to be set to limit the possible matches. The FI ratio is defined as the number of heavy atoms in the fixed part divided by the count in the changing part. In the case of the MCSS method, the ratio is the count of overlapping heavy atoms to the count of heavy atoms in the smaller molecule. These will be described as f_{F+I} and f_{MCSS} for F+I method and MCSS method, respectively. The outcome of changing these settings for both methods and the optimum settings will be described in section 5.3 and was part of the investigation from the paper included in the appendix.[50]

5.3. Results

Three sets of data extracted from the ChEMBL database[88] were used to perform MMPA: inhibitors of the Epidermal Growth Factor Receptor (EGFR), ligands of the dopamine D1 receptor, and voltage-gated calcium channel subunit alpha Cav3.2. These three were selected from the ChEMBL database[88] based on having measured IC_{50}, K_i and EC_{50} data for EGFR, D1, and Cav 3.2, respectively. These comprised 1010, 903, and 792 individual compounds, respectively.

The very first run of both methods for EGFR with default settings was performed to check the difference in computational time needed to find pairs. It took 6 minutes and

16 seconds for the F+I method and 1 day 18 hours, 36 minutes and 27 seconds for the MCSS method. These results show how fast the F+I method is compared to MCSS, where longer times come from calculating the overlap between the molecules. This might be especially problematic when screening large datasets.

Both methods (MCSS and F+I) have been applied to the sets described above. The exact number of pairs found by each method (and for each set) can be seen in **Table 5-1** and also in **Figure 5-4**. It can be noticed that the higher the f_{F+I}, the greater the number of found pairs, which was expected. It is the opposite in the case of the MCSS method, as shown in **Table 5-2** and in the **Figure 5-5**, where a lower f_{MCSS} leads to a higher number of matched pairs. It comes from the differences in defining ratios f_{F+I} and f_{MCSS}, as explained in section 5.2.

Table 5-1. Number of matched pairs found by the F+I method for all sets of compounds: EGFR, D1 and Cav 3.2.

F_{FI}	EGFR	D1	CAV 3.2
0.1	358	1856	330
0.2	978	4738	874
0.3	2140	11982	2090
0.4	4078	16324	4472
0.5	7964	24312	8174
0.6	15450	35644	16758
0.7	31710	56096	35304
0.8	157500	120678	175328
0.9	381077	236594	298666

Table 5-2. Number of matched pairs found by the MCSS method for all sets of compounds: EGFR, D1 and Cav 3.2.

F_{MCSS}	EGFR	D1	CAV 3.2
0.1	150538	123568	136218
0.2	129752	122867	134039
0.3	78795	86154	47743
0.4	25426	50019	9074
0.5	7564	31528	2584
0.6	3727	24480	1370
0.7	1994	17895	924
0.8	1106	9678	718
0.9	650	4550	503

A high ratio f_{F+I} means that pairs of molecules might be matched by very small fragments, like O, CH_2 or CH_3, which leads to a high number of found pairs using this method. Similarly, low f_{MCSS} can lead to finding pairs with a really small number of overlapped atoms, which obviously results in a higher number of found pairs. It is also possible that setting f_{MCSS} too high might lead to cases where almost all of the heavy atoms of one molecule are within the second molecule and this way the structural change could be only hydrogen or one or two heavy atoms. That could result in too low a number of found pairs being found by the MCSS method. Therefore, it is crucial to set the optimum f_{F+I} and f_{MCSS}.

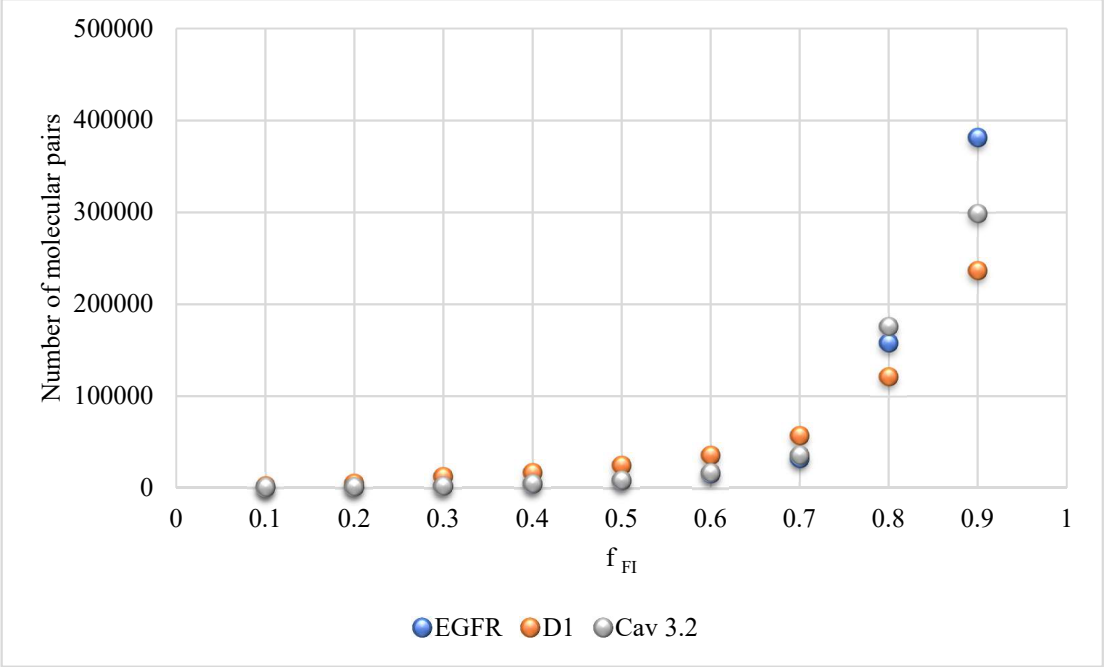


Figure 5-4. Numbers of molecular pairs found by the F+I method for different settings of f_{F+I}.

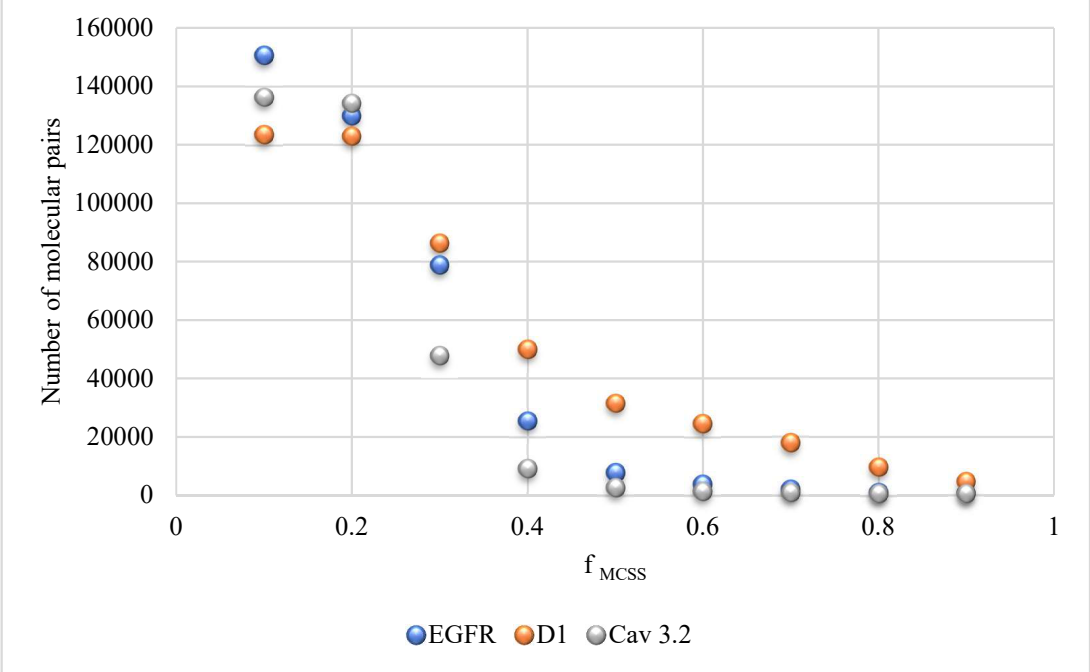


Figure 5-5. Numbers of molecular pairs found by the MCSS method for different settings of f_{MCSS}.

It is also important to analyse the number of pairs found by both approaches and those that were matched using only one of the methods. This analysis can be performed in two ways. One is to simply count the number of pairs found by both methods. A

second is to reduce that number to include only those pairs that are matched in the exact same manner, which means they have the exact same transformation encoded as SMIRKS. The results can be seen in tables: **Table 5-3**, **Table 5-4** and **Table 5-5** for all three sets of compounds. Those are the percentages of pairs found by both approaches. Using high values of f_{F+I} and low values of f_{MCSS} gives the biggest number of pairs found in common. However, it can be seen in **Table 5-3** that using 0.4 and 0.7 f_{f+I} and f_{MCSS} respectively gives over 40% pairs in common (of total number of pairs found by both methods separately). Similarly, in **Table 5-4** and **Table 5-5**, it can be seen that using $f_{F+I} = 0.4$ and $f_{MCSS} = 0.7$ allows the most pairs in common to be found: 33.24% and 55.06% for D1 and Cav 3.2 set, respectively.

When considering the number of the exact same transformations used by both methods, the overlap is smaller than when considering pairs of molecules, which means that a majority of them were paired for different reasons: the F+I method can find changes of large groups (such as substituted phenyl rings) whereas the MCSS method localizes the structural change to the smallest part of the structure that is different between the two molecules in the pair.

Table 5-3. The percentage of matched pairs of molecules found in the EGFR set by both methods using the pair count. In brackets: the percentage of common pairs found by using the exact same transformations.

		FI								
		0.9	**0.8**	**0.7**	**0.6**	**0.5**	**0.4**	**0.3**	**0.2**	**0.1**
MCSS	**0.9**	0.79 (0.10)	1.98 (0.24)	5.84 (1.18)	13.49 (2.37)	21.87 (4.42)	29.88 (7.89)	35.09 (12.87)	43.54 (21.07)	37.64 (23.21)
	0.8	1.31 (0.15)	3.31 (0.37)	9.71 (1.79)	20.82 (3.53)	30.41 (6.45)	38.26 (10.98)	39.97 (15.93)	38.93 (20.01)	23.39 (15.98)
	0.7	2.04 (0.23)	4.96 (0.55)	14.12 (2.61)	26.72 (5.04)	36.69 (8.67)	40.48 (13.09)	32.17 (14.01)	23.26 (14.03)	14.07 (9.95)
	0.6	2.97 (0.31)	6.92 (0.74)	17.89 (3.37)	27.73 (6.18)	32.90 (9.76)	28.36 (11.49)	19.03 (9.87)	13.06 (8.86)	7.69 (5.73)
	0.5	4.64 (0.47)	9.84 (1.11)	21.62 (4.57)	25.25 (7.44)	20.98 (8.20)	15.23 (7.70)	9.79 (5.97)	6.51 (4.88)	3.80 (2.95)
	0.4	7.77 (0.80)	11.98 (1.75)	15.88 (5.22)	11.36 (5.35)	6.96 (3.82)	4.86 (3.04)	3.04 (2.10)	1.96 (1.58)	1.14 (0.91)
	0.3	11.07 (1.57)	12.03 (2.91)	7.92 (3.39)	3.99 (2.32)	2.37 (1.47)	1.61 (1.08)	0.99 (0.72)	0.64 (0.52)	0.37 (0.30)
	0.2	12.14 (2.11)	10.40 (2.98)	5.13 (2.32)	2.57 (1.51)	1.47 (0.93)	0.99 (0.67)	0.61 (0.44)	0.39 (0.32)	0.23 (0.18)
	0.1	12.76 (2.35)	9.23 (2.77)	4.44 (2.05)	2.22 (1.32)	1.26 (0.80)	0.85 (0.58)	0.52 (0.38)	0.34 (0.28)	0.20 (0.16)

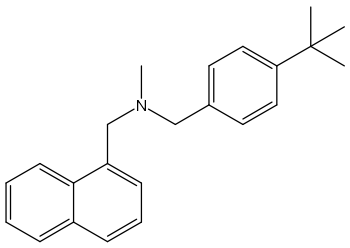
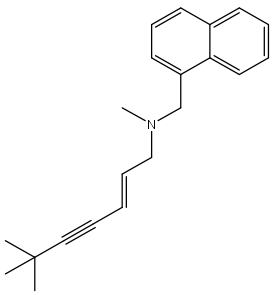
Table 5-4. The percentage of matched pairs of molecules found in the D1 set by both methods using the pair count. In brackets: the percentage of common pairs found by using the exact same transformations.

		FI								
		0.9	**0.8**	**0.7**	**0.6**	**0.5**	**0.4**	**0.3**	**0.2**	**0.1**
MCSS	**0.9**	5.44 (0.71)	7.91 (1.37)	15.66 (2.83)	20.20 (4.27)	23.24 (5.95)	27.36 (8.23)	29.27 (10.33)	26.48 (15.54)	24.26 (16.66)
	0.8	9.01 (1.23)	12.12 (2.33)	22.40 (4.62)	27.53 (6.70)	30.92 (8.93)	35.02 (11.68)	32.69 (13.82)	21.50 (14.13)	11.64 (9.25)
	0.7	11.75 (1.91)	15.30 (3.51)	24.65 (6.58)	28.64 (9.09)	31.11 (11.52)	33.24 (14.07)	29.16 (14.64)	12.11 (9.00)	6.49 (5.40)
	0.6	12.73 (2.17)	16.00 (3.90)	23.00 (6.97)	25.86 (9.34)	27.28 (11.43)	27.25 (12.39)	22.08 (12.00)	9.12 (6.97)	4.93 (4.05)
	0.5	14.49 (2.61)	17.65 (4.60)	23.92 (7.93)	26.13 (10.33)	25.61 (11.23)	21.77 (10.57)	17.60 (10.05)	7.32 (5.62)	3.98 (3.20)
	0.4	13.39 (2.61)	15.12 (4.37)	17.62 (6.91)	17.83 (8.38)	16.59 (8.43)	14.02 (7.62)	11.28 (7.05)	4.64 (3.72)	2.51 (2.06)
	0.3	11.57 (2.65)	11.96 (4.12)	11.60 (5.44)	10.68 (5.90)	9.83 (5.68)	8.26 (4.93)	6.62 (4.46)	2.70 (2.24)	1.46 (1.21)
	0.2	10.2 (2.69)	9.86 (3.86)	8.36 (4.33)	7.59 (4.53)	6.95 (4.26)	5.83 (3.63)	4.66 (3.24)	1.89 (1.60)	1.02 (0.86)
	0.1	10.24 (2.71)	9.83 (3.84)	8.32 (4.31)	7.55 (4.51)	6.92 (4.24)	5.80 (3.61)	4.64 (3.23)	1.88 (1.59)	1.02 (0.85)

Table 5-5. The percentage of matched pairs of molecules found in the Cav 3.2 set by both methods using the pair count. In brackets: the percentage of common pairs found by using the exact same transformations.

		FI								
		0.9	**0.8**	**0.7**	**0.6**	**0.5**	**0.4**	**0.3**	**0.2**	**0.1**
MCSS	**0.9**	0.46 (0.10)	0.73 (0.16)	4.02 (0.80)	9.68 (1.67)	22.59 (3.32)	35.26 (5.79)	48.24 (10.57)	53.70 (19.61)	50.89 (28.81)
	0.8	0.66 (0.12)	1.04 (0.20)	5.74 (0.97)	13.78 (2.00)	32.16 (3.94)	49.15 (6.74)	47.06 (11.75)	40.97 (17.59)	35.73 (22.90)
	0.7	0.84 (0.15)	1.34 (0.25)	7.36 (1.23)	17.58 (2.53)	38.51 (4.91)	55.06 (7.69)	48.76 (11.61)	33.40 (15.57)	27.80 (19.14)
	0.6	1.23 (0.22)	1.95 (0.38)	10.14 (1.81)	22.56 (3.61)	46.19 (6.74)	54.16 (9.09)	36.36 (10.12)	23.34 (12.48)	18.78 (14.12)
	0.5	2.08 (0.37)	3.26 (0.63)	15.20 (2.94)	29.20 (5.44)	41.05 (9.34)	33.07 (7.53)	20.72 (7.49)	12.78 (8.10)	10.05 (8.24)
	0.4	4.96 (0.87)	7.08 (1.41)	21.8 (5.88)	23.61 (8.68)	15.26 (8.74)	10.77 (3.92)	6.29 (3.14)	3.74 (2.81)	2.86 (2.55)
	0.3	11.64 (2.31)	12.85 (3.15)	12.62 (8.47)	6.19 (5.83)	3.30 (2.70)	2.14 (1.02)	1.22 (0.70)	0.71 (0.58)	0.54 (0.50)
	0.2	18.45 (5.06)	16.35 (5.38)	5.86 (9.83)	2.71 (2.49)	1.22 (1.06)	0.77 (0.38)	0.44 (0.26)	0.25 (0.21)	0.19 (0.18)
	0.1	18.75 (5.16)	16.24 (5.34)	5.82 (9.70)	2.67 (2.46)	1.20 (1.04)	0.76 (0.38)	0.43 (0.25)	0.25 (0.20)	0.19 (0.18)

There are a lot of pairs found only by one method and not the other. Some examples are shown in **Figure 5-6**, **Figure 5-7** and **Figure 5-8** for EGFR, D1 and Cav 3.2 sets. There are pairs found only by one of the method when using $f_{MCSS} = 0.7$ for MCSS approach and $f_{F+I} = 0.4$ for F+I approach.

	Molecule A	**Molecule B**
F+I only		
	$N_{fixed} = 18$; $N_{changing} = 6$	$N_{fixed} = 18$; $N_{changing} = 4$

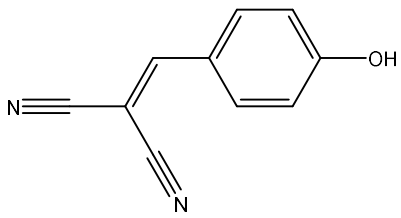
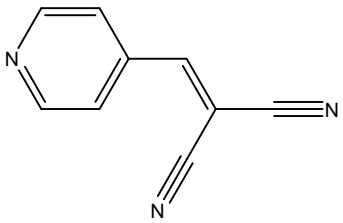
	Molecule C	Molecule D
MCSS only		
	N_{fixed} = 11; $N_{changing}$ = 2	N_{fixed} = 11; $N_{changing}$ = 1

Figure 5-6. Pairs found only by the MCSS (bottom) and F+I (top) methods when using the EGFR set.

In **Figure 5-6**, the pair found by only the F+I approach has a change from aromatic ring c1ccccc1 to acyclic chain C=CC#C (double cut is considered here). The number of heavy atoms in the changing part $N_{changing}$ is 6 for the molecule A and 4 for molecule B. The number of heavy atoms in the whole molecule $N_{molecule}$ is 24 and 22 for molecule A and B, respectively. Thus, the ratio $N_{changing}/N_{molecule}$ is 0.25 for the first molecule and 0.18 for the second. These are below the specified f_{F+I} = 0.4, so the pair is allowed. The fixed and changing part identified by MCSS in this case would be different. The N_{fixed} would be 14 and the ratio $N_{fixed}/N_{molecule}$ would be 0.58 and 0.63 for molecule A and B, respectively. This is below the f_{MCSS} cutoff and therefore the pair would not be allowed.

The pair of compounds found only by the MCSS method have a change in the heteroaromatic ring, as shown in **Figure 5-6**. Such structural change in the aromatic ring would be not possible to find by using the F+I method as it requires breaking the bond to the aromatic ring which gives a changing part that is a large fraction of the molecule. For the first molecule the N_{fixed}, the number of heavy atoms in the fixed part, is 11 and $N_{molecule}$ is 13 and for the second molecule $N_{molecule}$ is 12. Therefore, the ratio, $N_{fixed}/N_{molecule}$ is 0.85 for the first molecule and 0.92 for the second. This is greater than the cutoff f_{MCSS} used in this example (0.7), which is allowed. The pair would not be allowed in the F+I method as the ratio $N_{fixed}/N_{molecule}$ would be 0.54 and 0.5 for molecule C and D, respectively, which is greater than the f_{F+I} cutoff (0.4).

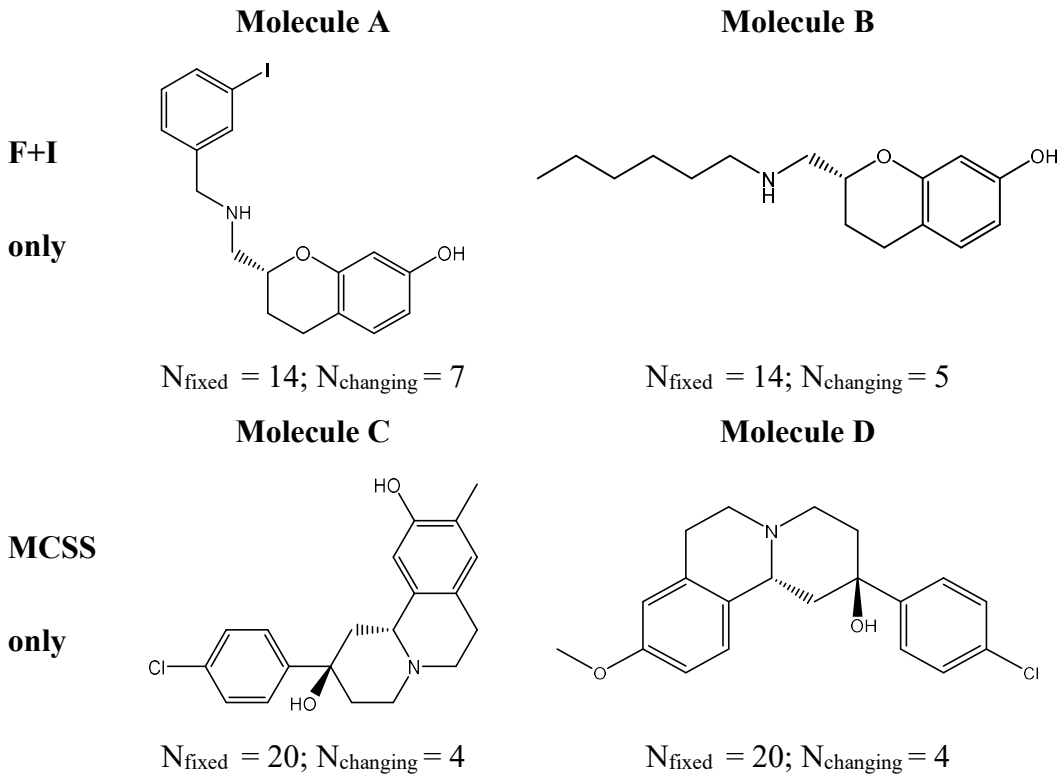


Figure 5-7. Pairs found only by the MCSS (bottom) and F+I (top) methods when using the D1 set.

In **Figure 5-7** the pair of compounds found only by the F+I method has the structural change from c1ccccc1I to CCCCC and the pair found by only the MCSS approach has the structural change includes part of aromatic ring and because of that was not possible to be found by the F+I method. In case of pair found only by the MCSS method, the ratio $N_{fixed}/N_{molecule}$ is 0.83 for both molecules, which is greater than the defined cutoff value (0.7) so the pair is allowed. Using the F+I method, the fragments would be identified differently and the ratio of $N_{changing}/N_{molecule}$ would be 17/24 = 0.71 for both molecules, which is greater than cutoff 0.4 and therefore would not be considered as a pair.

In the case of the pair found only by the F+I method, the ratio $N_{changing}/N_{molecule}$ is 0.33 and 0.26 for molecules A and B respectively, which is below the used f_{F+I} (0.4) and therefore the pair is allowed. When the MCSS approach is applied, the ratio $N_{fixed}/N_{molecule}$ is 0.67 and 0.74. As one of the molecule has a ratio lower than the cutoff value (0.7) therefore the pair is not allowed.

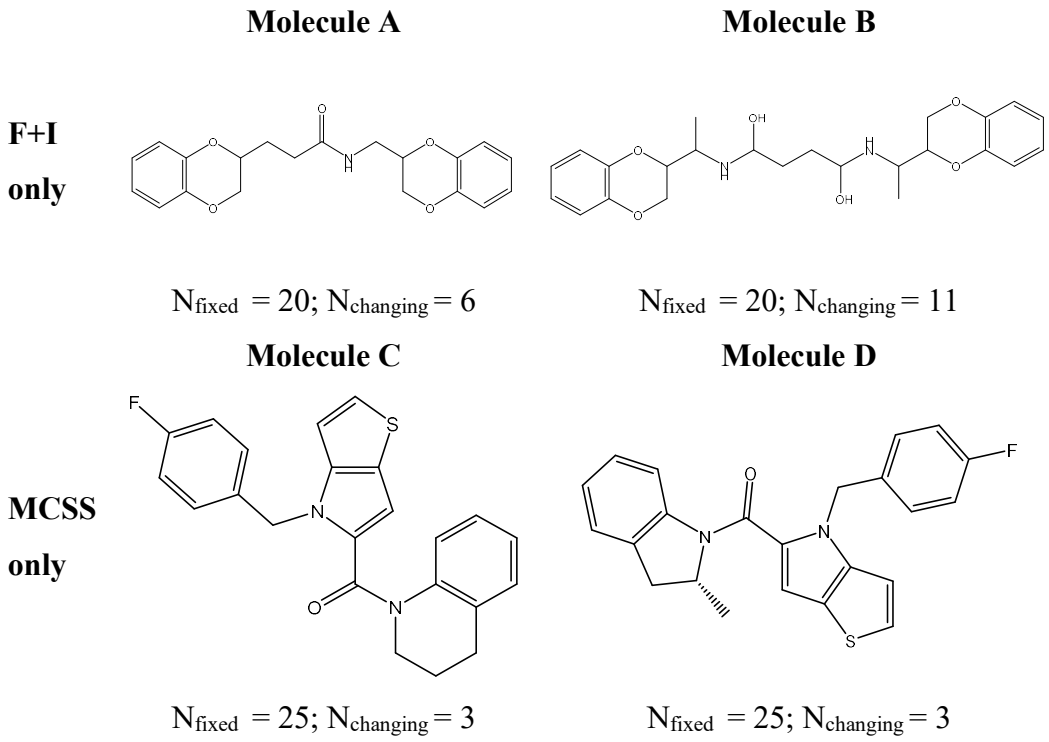


Figure 5-8. Pairs found only by the MCSS (bottom) and F+I (top) methods when using the Cav3.2 set.

In **Figure 5-8** the pair found by only the MCSS method has a change from CCC to CC(C). The pair has the $N_{fixed}/N_{molecule}$ equal to 0.89 and 0.89 for molecule A and B, respectively. As the ratios for the both molecules of the pair are above the specified f_{MCSS} of 0.7, this pair is allowed. The F+I approach would not identify the pair as the obtained ratio would be greater than the f_{F+I} cutoff (in this approach the ratio $N_{changing}/N_{molecule}$ is 12/28 = 0.43 for both molecules) and therefore is not allowed.

In the case of the pair found by the F+I approach, the molecule A has 0.23 and molecule B has 0.35 values for the ratio $N_{changing}/N_{molecule}$. This is below the f_{F+I} = 0.4, thus it is allowed. When using the MCSS approach, this pair would not be allowed as the ratio $N_{fixed}/N_{molecule}$ is 0.38 and 0.32 for molecule A and B, respectively. This is lower than the specified cutoff (0.7) and thus is not allowed.

All the examples show that using both methods for MMPA simultaneously would be more advantageous, as none of the presented examples of found pairs is chemically unreasonable. This would significantly increase the chances of finding pairs worth considering.

5.4. Conclusions

The analyses described in this chapter lead to that conclusion that in order to most thoroughly explore the effect of structural transformations on chemical properties, it might be reasonable to use both approaches, F+I and MCSS, simultaneously. This way it is possible to find MMPs that would be found by one method and not the other, like finding changes of linkers (which is the limitation of MCSS approach but can be found by using F+I method) or small changes in rings or highly substituted core changes that are among the limitations of F+I method but could be found easily using MCSS method.

The choice of optimum settings for both methods is an important decision. A high ratio f_{F+I} allows to find more pairs, but too high f_{F+I} could lead to pairs that might be matched by very small fragments like O, CH_2 or CH_3. Similarly, setting too low f_{MCSS} can lead to finding pairs with a really small number of overlapped atoms. On the other hand, too high f_{MCSS} could result in matched pairs with almost all of the heavy atoms of one molecule are within the second molecule and this way the structural change could be only hydrogen or one or two heavy atoms.

It could also be suggested that the optimum settings should see f_{F+I} and the f_{MCSS} set to 0.4 and 0.7, respectively. Such choice allows to good coverage of chemical space and find high number of matched pairs. As was shown in this chapter, choosing the aforementioned settings gives over ca. 33%, 40% and 55% pairs in common for EGFR, D1 and Cav3.2 set, respectively. Even though the majority of them were paired by each approach for different reasons: the F+I method can find changes of large groups (such as substituted phenyl rings) whereas the MCSS method localizes the structural change to the smallest part of the structure that is different between the two molecules in the pair.

Further analysis was conducted by the colleagues in the research group[50] and lead to slightly different conclusions - the optimum settings are f_{F+I}=0.4 and f_{MCSS}=0.9. The paper with full analysis and results was included in the appendix.

References

(1) Landau, B.; Smith, L. B.; Jones, S. S. The Importance of Shape in Early Lexical Learning. *Cogn. Dev.* **1988**, *3* (3), 299–321.

(2) Zhang, J. X. J.; Hoshino, K. Chapter 1 - Introduction to Molecular Sensors. In *Molecular Sensors and Nanodevices*; Zhang, J. X. J., Hoshino, K., Eds.; William Andrew Publishing: Oxford, **2014**; pp 1–42.

(3) Sowdhamini, R.; Srinivasan, N.; Guruprasad, K.; Rufino, S.; Dhanaraj, V.; Wood, S. P.; Emsley, J.; White, H. E.; Blundell, T. Protein Three-Dimensional Structure and Molecular Recognition: A Story of Soft Locks and Keys. *Pharm. Acta Helv.* **1995**, *69* (4), 185–192.

(4) Tesniere, A.; Apetoh, L.; Ghiringhelli, F.; Joza, N.; Panaretakis, T.; Kepp, O.; Schlemmer, F.; Zitvogel, L.; Kroemer, G. Immunogenic Cancer Cell Death: A Key-Lock Paradigm. *Curr. Opin. Immunol.* **2008**, *20* (5), 504–511.

(5) Fischer, E. Einfluß der Konfiguration auf die Wirkung der Enzyme. I. In *Untersuchungen Über Kohlenhydrate und Fermente (1884–1908)*; Springer, Berlin, Heidelberg, **1909**; pp 836–844.

(6) Pauling, L. Molecular Architecture and Biological Reactions. *Chem. Eng. News Arch.* **1946**, *24* (10), 1375–1377.

(7) Lauria, A.; Tutone, M.; Almerico, A. M. Virtual Lock-and-Key Approach: The in Silico Revival of Fischer Model by Means of Molecular Descriptors. *Eur. J. Med. Chem.* **2011**, *46* (9), 4274–4280.

(8) Putta, S.; Beroza, P. Shapes of Things: Computer Modeling of Molecular Shape in Drug Discovery. *Curr. Top. Med. Chem.* **2007**, *7* (15), 1514–1524.

(9) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; et al. Molecular Shape and Medicinal Chemistry: A Perspective. *J. Med. Chem.* **2010**, *53* (10), 3862–3886.

(10) Istvan, E. S.; Deisenhofer, J. Structural Mechanism for Statin Inhibition of HMG-CoA Reductase. *Science* **2001**, *292* (5519), 1160–1164.

(11) Patani, G. A.; LaVoie, E. J. Bioisosterism: A Rational Approach in Drug Design. *Chem. Rev.* **1996**, *96* (8), 3147–3176.

(12) Proschak, E.; Zettl, H.; Tanrikulu, Y.; Weisel, M.; Kriegl, J. M.; Rau, O.; Schubert-Zsilavecz, M.; Schneider, G. From Molecular Shape to Potent Bioactive Agents I: Bioisosteric Replacement of Molecular Fragments. *ChemMedChem* **2009**, *4* (1), 41–44.

(13) Lipinski, C. A. Chapter 27. Bioisosterism in Drug Design. In *Annual Reports in Medicinal Chemistry*; Bailey, D. M., Ed.; Academic Press, **1986**; Vol. 21, pp 283–291.

(14) Wermuth, C. G.; Ciapetti, P.; Giethlen, B.; Bazzini, P. 2.16 - Bioisosterism. In *Comprehensive Medicinal Chemistry II*; Taylor, J. B., Triggle, D. J., Eds.; Elsevier: Oxford, **2007**; pp 649–711.

(15) Friedman, H. L. Influence of Isosteric Replacements upon Biological Activity. In *First Symposium on Chemical-biological Correlation, May 26-27, 1950*; National Academies, **1951**; pp 295–362.

(16) Langdon, S. R.; Ertl, P.; Brown, N. Bioisosteric Replacement and Scaffold Hopping in Lead Generation and Optimization. *Mol. Inform.* **2010**, *29* (5), 366–385.

(17) Sheng, C.; Che, X.; Wang, W.; Wang, S.; Cao, Y.; Miao, Z.; Yao, J.; Zhang, W. Design and Synthesis of Novel Triazole Antifungal Derivatives by Structure-Based Bioisosterism. *Eur. J. Med. Chem.* **2011**, *46* (11), 5276–5282.

(18) Rudmann, D. G. On-Target and Off-Target-Based Toxicologic Effects. *Toxicol. Pathol.* **2013**, *41* (2), 310–314.

(19) Meng, X.-Y.; Zhang, H.-X.; Mezei, M.; Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr. Comput. Aided Drug Des.* **2011**, *7* (2), 146–157.

(20) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161* (2), 269–288.

(21) Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. Lead Discovery Using Molecular Docking. *Curr. Opin. Chem. Biol.* **2002**, *6* (4), 439–446.

(22) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein−Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49* (20), 5851–5855.

(23) Berry, M.; Fielding, B.; Gamieldien, J. Chapter 27 - Practical Considerations in Virtual Screening and Molecular Docking. In *Emerging Trends in*

Computational Biology, Bioinformatics, and Systems Biology; Tran, Q. N., Arabnia, H., Eds.; Emerging Trends in Computer Science and Applied Computing; Morgan Kaufmann: Boston, **2015**; pp 487–502.

(24) Dar, A. M.; Mir, S. Molecular Docking: Approaches, Types, Applications and Basic Challenges. *J. Anal. Bioanal. Tech.* **2017**, *8* (2), 1–7.

(25) Chen, Y.-C. Beware of Docking! *Trends Pharmacol. Sci.* **2015**, *36* (2), 78–95.

(26) Leelananda, S. P.; Lindert, S. Computational Methods in Drug Discovery. *Beilstein J. Org. Chem.* **2016**, *12* (1), 2694–2718.

(27) Sinha, S.; Vohora, D. Chapter 2 - Drug Discovery and Development: An Overview. In *Pharmaceutical Medicine and Translational Clinical Research*; Academic Press: Boston, **2018;** pp 19–32.

(28) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983–996.

(29) Kristensen, T. G.; Nielsen, J.; Pedersen, C. N. S. Methods For Similarity-Based Virtual Screening. *Comput. Struct. Biotechnol. J.* **2013**, *5* (6), e201302009.

(30) Hughes, J.; Rees, S.; Kalindjian, S.; Philpott, K. Principles of Early Drug Discovery. *Br. J. Pharmacol.* **2011**, *162* (6), 1239–1249.

(31) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66* (1), 334–395.

(32) Acharya, C.; Coop, A.; Polli, J. E.; MacKerell, A. D. Recent Advances in Ligand-Based Drug Design: Relevance and Utility of the Conformationally Sampled Pharmacophore Approach. *Curr. Comput. Aided Drug Des.* **2011**, *7* (1), 10–22.

(33) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal Assignment Methods for Ligand-Based Virtual Screening. *J. Cheminformatics* **2009**, *1*, 14.

(34) Drwal, M. N.; Griffith, R. Combination of Ligand- and Structure-Based Methods in Virtual Screening. *Drug Discov. Today Technol.* **2013**, *10* (3), e395–e401.

(35) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57* (12), 4977–5010.

(36) Danishuddin; Khan, A. U. Descriptors and Their Selection Methods in QSAR Analysis: Paradigm for Drug Design. *Drug Discov. Today* **2016**, *21* (8), 1291–1302.

(37) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39* (16), 3049–3059.

(38) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36.

(39) Franco, P.; Porta, N.; Holliday, J. D.; Willett, P. The Use of 2D Fingerprint Methods to Support the Assessment of Structural Similarity in Orphan Drug Legislation. *J. Cheminformatics* **2014**, *6*, 5.

(40) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58–63.

(41) Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and Comparison of 2D Fingerprints: Insights into Database Screening Performance Using Eight Fingerprint Methods. *J. Mol. Graph. Model.* **2010**, *29* (2), 157–170.

(42) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 1177–1185.

(43) Bajusz, D.; Rácz, A.; Héberger, K. *Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching*; **2017**.

(44) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280.

(45) Vidal, D.; Thormann, M.; Pons, M. LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarities. *J. Chem. Inf. Model.* **2005**, *45* (2), 386–393.

(46) Grant, J. A.; Haigh, J. A.; Pickup, B. T.; Nicholls, A.; Sayle, R. A. Lingos, Finite State Machines, and Fast Similarity Searching. *J. Chem. Inf. Model.* **2006**, *46* (5), 1912–1918.

(47) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.

(48) Kunimoto, R.; Vogt, M.; Bajorath, J. Maximum Common Substructure-Based Tversky Index: An Asymmetric Hybrid Similarity Measure. *J. Comput. Aided Mol. Des.* **2016**, *30* (7), 523–531.

(49) Leach, A. G.; Lukac, I.; Zarnecka, J. M.; Dossetter, A. G.; Griffen, E. J. 3.10 - Matched Molecular Pair Analysis. In *Comprehensive Medicinal Chemistry III*; Chackalamannil, S., Rotella, D., Ward, S. E., Eds.; Elsevier: Oxford, **2017**; pp 221–252.

(50) Lukac, I.; Zarnecka, J.; Griffen, E. J.; Dossetter, A. G.; St-Gallay, S. A.; Enoch, S. J.; Madden, J. C.; Leach, A. G. Turbocharging Matched Molecular Pair Analysis: Optimizing the Identification and Analysis of Pairs. *J. Chem. Inf. Model.* **2017**, *57* (10), 2424–2436.

(51) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA, **2005**; pp 271–285.

(52) Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *J. Med. Chem.* **2006**, *49* (23), 6672–6682.

(53) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50* (3), 339–348.

(54) Warner, D. J.; Griffen, E. J.; St-Gallay, S. A. WizePairZ: A Novel Algorithm to Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry. *J. Chem. Inf. Model.* **2010**, *50* (8), 1350–1357.

(55) Deng, W.; Schneider, G.; So, W. V. Mapping Chemical Structures to Markush Structures Using SMIRKS. *Mol. Inform.* **2011**, *30* (8), 665–671.

(56) Wassermann, A. M.; Dimova, D.; Iyer, P.; Bajorath, J. Advances in Computational Medicinal Chemistry: Matched Molecular Pair Analysis. *Drug Dev. Res.* **2012**, *73* (8), 518–527.

(57) Böhm, H.-J.; Flohr, A.; Stahl, M. Scaffold Hopping. *Drug Discov. Today Technol.* **2004**, *1* (3), 217–224.

(58) Hu, Y.; Stumpfe, D.; Bajorath, J. Recent Advances in Scaffold Hopping. *J. Med. Chem.* **2017**, *60* (4), 1238–1246.

(59) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem. Int. Ed.* **1999**, *38* (19), 2894–2896.

(60) Zhao, H. Scaffold Selection and Scaffold Hopping in Lead Generation: A Medicinal Chemistry Perspective. *Drug Discov. Today* **2007**, *12* (3), 149–155.

(61) Sun, H.; Tawa, G.; Wallqvist, A. Classification of Scaffold-Hopping Approaches. *Drug Discov. Today* **2012**, *17* (7), 310–324.

(62) Martin, Y. C.; Muchmore, S. Beyond QSAR: Lead Hopping to Different Structures. *QSAR Comb. Sci.* **2009**, *28* (8), 797–801.

(63) Yang, S.-Y. Pharmacophore Modeling and Applications in Drug Discovery: Challenges and Recent Advances. *Drug Discov. Today* **2010**, *15* (11–12), 444–450.

(64) Fei, J.; Zhou, L.; Liu, T.; Tang, X.-Y. Pharmacophore Modeling, Virtual Screening, and Molecular Docking Studies for Discovery of Novel Akt2 Inhibitors. *Int. J. Med. Sci.* **2013**, *10* (3), 265–275.

(65) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110* (18), 5959–5967.

(66) Zhao, X.; Chen, M.; Huang, B.; Ji, H.; Yuan, M. Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA) Studies on A1A-Adrenergic Receptor Antagonists Based on Pharmacophore Molecular Alignment. *Int. J. Mol. Sci.* **2011**, *12* (10), 7022–7037.

(67) Kubinyi, H. Comparative Molecular Field Analysis (CoMFA). In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH Verlag GmbH, **2003**; pp 1555–1574.

(68) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37* (24), 4130–4146.

(69) Haigh, J. A.; Pickup, B. T.; Grant, J. A.; Nicholls, A. Small Molecule Shape-Fingerprints. *J. Chem. Inf. Model.* **2005**, *45* (3), 673–684.

(70) Grant, J. A.; Pickup, B. T. A Gaussian Description of Molecular Shape. *J. Phys. Chem.* **1995**, *99* (11), 3503–3510.

(71) Grant, J. A.; Pickup, B. T. Gaussian Shape Methods. In *Computer Simulation of Biomolecular Systems*; Computer Simulations of Biomolecular Systems; Springer, Dordrecht, **1997**; pp 150–176.

(72) *ROCS 3.2.1.4: OpenEye Scientific Software, Santa Fe, NM. Http://Www.Eyesopen.Com.*

(73) Copi, C. J.; Huterer, D.; Starkman, G. D. Multipole Vectors--a New Representation of the CMB Sky and Evidence for Statistical Anisotropy or Non-Gaussianity at 2<=l<=8. *Phys. Rev. D* **2004**, *70* (4).

(74) Willett, P. Similarity Searching Using 2D Structural Fingerprints. *Methods Mol. Biol. Clifton NJ* **2011**, *672*, 133–158.

(75) Hu, G.; Kuang, G.; Xiao, W.; Li, W.; Liu, G.; Tang, Y. Performance Evaluation of 2D Fingerprint and 3D Shape Similarity Methods in Virtual Screening. *J. Chem. Inf. Model.* **2012**, *52* (5), 1103–1113.

(76) Hu, Y.; Stumpfe, D.; Bajorath, J. Recent Advances in Scaffold Hopping. *J. Med. Chem.* **2017**, *60* (4), 1238–1246.

(77) Haigh, J. A.; Pickup, B. T.; Grant, J. A.; Nicholls, A. Small Molecule Shape-Fingerprints. *J. Chem. Inf. Model.* **2005**, *45* (3), 673–684.

(78) *OpenEye Toolkits 2017.Oct.1 OpenEye Scientific Software, Santa Fe, NM. Http://Www.Eyesopen.Com*.

(79) Taylor, R.; Cole, J. C.; Cosgrove, D. A.; Gardiner, E. J.; Gillet, V. J.; Korb, O. Development and Validation of an Improved Algorithm for Overlaying Flexible Molecules. *J. Comput. Aided Mol. Des.* **2012**, *26* (4), 451–472.

(80) Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. Protein−Ligand Docking against Non-Native Protein Conformers. *J. Chem. Inf. Model.* **2008**, *48* (11), 2214–2225.

(81) Hajian-Tilaki, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Casp. J. Intern. Med.* **2013**, *4* (2), 627–635.

(82) *R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Http://Www.R-Project.Org/.*

(83) *OMEGA 2.5.1.4: OpenEye Scientific Software, Santa Fe, NM. Http://Www.Eyesopen.Com.*

(84) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. The Development of Versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31* (2), 187–204.

(85) *MDDR, MDL Information System.*

(86) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand Depot: A Data Warehouse for Ligands Bound to Macromolecules. *Bioinformatics* **2004**, *20* (13), 2153–2155.

(87) Head, R. D.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M.; Marshall, G. R. VALIDATE: A New Method for the Receptor-Based Prediction of Binding Affinities of Novel Ligands. *J. Am. Chem. Soc.* **1996**, *118* (16), 3959–3969.

(88) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42* (D1), D1083–D1090.

(89) Jain, A. N.; Nicholls, A. Recommendations for Evaluation of Computational Methods. *J. Comput. Aided Mol. Des.* **2008**, *22* (3–4), 133–139.

(90) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582–6594.

(91) McGregor, M. J.; Luo, Z.; Jiang, X. Virtual Screening in Drug Discovery. In *Drug Discovery Research*; Huang, Z., Ed.; John Wiley & Sons, Inc., **2007**; pp 63–88.

(92) *RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL Http://Www.Rstudio.Com/.*

(93) Alantary, D.; Yalkowsky, S. Calculating the Solubilities of Drugs and Drug-Like Compounds in Octanol. *J. Pharm. Sci.* **2016**, *105* (9), 2770–2773.

(94) Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubility by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41* (2), 354–357.

(95) Yalkowsky, S. H. Estimation of the Aqueous Solubility of Organic Compounds. In *Physical Property Prediction in Organic Chemistry*; Springer, Berlin, Heidelberg, **1988**; pp 469–480.

(96) Jain, N.; Yang, G.; Machatha, S. G.; Yalkowsky, S. H. Estimation of the Aqueous Solubility of Weak Electrolytes. *Int. J. Pharm.* **2006**, *319* (1), 169–171.

(97) Yang, G.; Ran, Y.; Yalkowsky, S. H. Prediction of the Aqueous Solubility: Comparison of the General Solubility Equation and the Method Using an Amended Solvation Energy Relationship. *J. Pharm. Sci.* **2002**, *91* (2), 517–533.

(98) *Molecular Operating Environment (MOE), 2013.08; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2018.*

(99) Bloch, D. Computer Software Review. Review of PHYSPROP Database (Version 1.0). *J. Chem. Inf. Comput. Sci.* **1995**, *35* (2), 328–329.

(100) *Reaxys*; [Frankfurt, Germany] ; [New York, NY] : Elsevier.

(101) Yalkowsky, S. H.; He, Y.; Jain, P. *Handbook of Aqueous Solubility Data, Second Edition*; CRC Press, **2016**.

(102) Ran, Y.; Jain, A.; Yalkowsky, S. H. Solubilization and Preformulation Studies on PG-300995 (An Anti-HIV Drug). *J. Pharm. Sci.* **2005**, *94* (2), 297–303.

(103) Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. Why Are Some Properties More Difficult To Predict than Others? A Study of QSPR Models of Solubility, Melting Point, and Log P. *J. Chem. Inf. Model.* **2008**, *48* (1), 220–232.

(104) McDonagh, J. L.; Nath, N.; De Ferrari, L.; van Mourik, T.; Mitchell, J. B. O. Uniting Cheminformatics and Chemical Theory To Predict the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *J. Chem. Inf. Model.* **2014**, *54* (3), 844–856.

(105) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. Representation/Prediction of Solubilities of Pure Compounds in Water Using Artificial Neural Network−Group Contribution Method. *J. Chem. Eng. Data* **2011**, *56* (4), 720–726.

(106) Hansen, N. T.; Kouskoumvekaki, I.; Jørgensen, F. S.; Brunak, S.; Jónsdóttir, S. Ó. Prediction of PH-Dependent Aqueous Solubility of Druglike Molecules. *J. Chem. Inf. Model.* **2006**, *46* (6), 2601–2609.

(107) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility from Structure. *Adv. Drug Deliv. Rev.* **2002**, *54* (3), 355–366.

(108) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility from Monte Carlo Simulations. *Bioorg. Med. Chem. Lett.* **2000**, *10* (11), 1155–1158.

(109) Muratov, E. N.; Kuz'min, V. E.; Artemenko, A. G.; Kovdienko, N. A.; Gorb, L.; Hill, F.; Leszczynski, J. New QSPR Equations for Prediction of Aqueous Solubility for Military Compounds. *Chemosphere* **2010**, *79* (8), 887–890.

(110) Jain, A.; Yalkowsky, S. H. Estimation of Melting Points of Organic Compounds-II. *J. Pharm. Sci.* **2006**, *95* (12), 2562–2618.

(111) Newby, D.; Freitas, A. A.; Ghafourian, T. Decision Trees to Characterise the Roles of Permeability and Solubility on the Prediction of Oral Absorption. *Eur. J. Med. Chem.* **2015**, *90*, 751–765.

(112) Peterson, D. L.; Yalkowsky, S. H. Comparison of Two Methods for Predicting Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (6), 1531–1534.

(113) Jain, P.; Sepassi, K.; Yalkowsky, S. H. Comparison of Aqueous Solubility Estimation from AQUAFAC and the GSE. *Int. J. Pharm.* **2008**, *360* (1), 122–147.

(114) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 266–275.

(115) Kramer, C.; Heinisch, T.; Fligge, T.; Beck, B.; Clark, T. A Consistent Dataset of Kinetic Solubilities for Early-Phase Drug Discovery. *ChemMedChem* **2009**, *4* (9), 1529–1536.

(116) Johnson, S. R.; Chen, X.-Q.; Murphy, D.; Gudmundsson, O. A Computational Model for the Prediction of Aqueous Solubility That Includes Crystal Packing, Intrinsic Solubility, and Ionization Effects. *Mol. Pharm.* **2007**, *4* (4), 513–523.

(117) *ChemCell, Collaborative Drug Discovery, Inc., 2010, Https://Github.Com/Cdd/Chemcell*.

(118) Pence, H. E.; Williams, A. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87* (11), 1123–1124.

(119) *IBM Corp. Released 2015. IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp.*

(120) Kramer, I. M. Chapter 8 - Nuclear Receptors. In *Signal Transduction (Third Edition)*; Academic Press: Boston, **2016**; pp 477–527.

(121) Sever, R.; Glass, C. K. Signaling by Nuclear Receptors. *Cold Spring Harb. Perspect. Biol.* **2013**, *5* (3).

(122) Woods, C. G.; Heuvel, J. P. V.; Rusyn, I. Genomic Profiling in Nuclear Receptor-Mediated Toxicity. *Toxicol. Pathol.* **2007**, *35* (4), 474–494.

(123) Huang, P.; Chandra, V.; Rastinejad, F. Structural Overview of the Nuclear Receptor Superfamily: Insights into Physiology and Therapeutics. *Annu. Rev. Physiol.* **2010**, *72*, 247–272.

(124) Rastinejad, F.; Huang, P.; Chandra, V.; Khorasanizadeh, S. Understanding Nuclear Receptor Form and Function Using Structural Biology. *J. Mol. Endocrinol.* **2013**, *51* (3), T1–T21.

(125) Sladek, F. M. What Are Nuclear Receptor Ligands? *Mol. Cell. Endocrinol.* **2011**, *334* (1–2), 3–13.

(126) Heinzel, N. N. and T. Nuclear Receptors: Overview and Classification http://www.eurekaselect.com/91278/article (accessed Feb 2, 2018).

(127) Mellor, C.; Cronin, M.; Steinmetz, F. Identification of in Silico Structural Alerts for Liver Steatosis Induced by Nuclear Receptor Agonists. *Toxicol. Lett.* **2014**, *229*, S162.

(128) Widenius, M.; Axmark, D.; Arno, K. *MySQL Reference Manual*; O'Reilly Media, Inc., 2002.

(129) *MCPairs, Version 2.0; MedChemica Ltd.: Macclesfield, U.K., 2016*.