

Evaluation of Similarity Measures for Ligand-Based Virtual Screening

A study submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

at



The
University
Of
Sheffield.

The University of Sheffield

by

Lucyantie Mazalan

Supervisor: Prof. Peter Willett, PhD

Co-Supervisors: John Holliday, PhD and Laura Sbaffi, PhD

Information School

November 2017

Acknowledgments

This thesis becomes a reality with the kind support and help of many individuals. I would like to sincerely thank all of them.

First and foremost, I am highly indebted to Prof Peter Willett, Dr John Holliday and Dr Laura Scaffi for their guidance and constant support in completing the journey of this endeavour. Peter, in particular, had been very understanding and patience to me during the challenges that I encountered in this journey. I am truly blessed and honoured to be supervised by all of them.

Many thanks and appreciations to the former and present members of the Chemoinformatics Group: Prof Val Gillet, Christina Founti, Dr Antonio de la Vega de Leon, Dr James Wallace, Dr Matthew Seddon, Gian Marco Ghiandoni, James Webster, Jessica Stacey, Dr Nor Samsiah Mohd Sani, Christos Kannas and Dr Edmund Duesbury who have willingly helped me with their abilities and encouragement. The experience working with all of you at Michael Lynch Laboratory will never be forgotten. Special thanks to Dr Andrew Bell from Sheffield Methods Institute for his knowledge and guidance towards the successful research collaboration. Appreciation to Dr Will Furnass and Dr Mike Croucher from Research Software Engineering Sheffield for their knowledge and technical support on HPC.

My heartiest gratitude goes to my dear husband Mohd Aswad and my mom Zainab for loving me unconditionally and reminds me to be positive and keep going. This journey will not be a reality without them. To all my family and friends who have helped and support me in every possible manner, may all of them be rewarded more, if not equally.

This thesis is dedicated to my son Adam Yusoff, the long awaited gift that have made me stronger and more fulfilled than I could have ever imagined.

Abstract

Nearest neighbour searching is a fundamental concept for many ligand-based virtual screening applications. The system searches for the nearest molecule by quantifying their similarity using various molecular representations and similarity coefficients. These similarity measures are the key components of the system where the variability and the characteristic of the components affect the effectiveness of the search.

The first aim of this thesis was to investigate the effects of 2D fingerprint dimensionality on the effectiveness of chemoinformatics applications and the contributing factors were analysed. Two nearest neighbour search applications, similarity searching and molecular clustering were conducted. Various types of coefficients were used to measure the similarity and distances of the chemical dataset. It was observed that the effectiveness of the similarity search and clustering applications varied depending on the coefficient used to measure the degree of similarity or distances. The sparseness of the representations also affects the similarity measures. The second aim of the study was to quantify the relative importance of the components influencing 2D fingerprint similarity searching and this research was carried out using cross-classified modeling. Effectiveness values produced by different types of 2D fingerprints and similarity coefficients were used to model the more important component. The bioactivity of the molecule was the most important factor identified, followed by the reference structure. Evaluation between the fingerprint representation and the similarity coefficient revealed that the fingerprint had a greater role in determining the effectiveness of the similarity searching than the similarity coefficient. This research contributes to the knowledge of similarity measures in the chemoinformatics domain on the impact of high dimensional space and the similarity search components. This contribution provides a practical implication on the effectiveness of the similarity search application in particular and ligand-based virtual screening applications.

Table of Contents

Acknowledgments.....	i
Abstract	iii
Table of Contents	v
List of Figures.....	ix
List of Tables.....	xiii
Chapter 1 Introduction.....	1
1.1 Background.....	1
1.2 Basis of Chemoinformatics	2
1.3 Aims of Research	4
1.4 Organisation of Thesis	7
Chapter 2 Similarity Searching in Chemoinformatics	9
2.1 Virtual Screening	9
2.2 Molecular Similarity	11
2.2.1 Representation and Descriptors	13
2.2.2 Weighting Scheme	23
2.2.3 Similarity Coefficient	26
2.3 Similarity Searching Application.....	33
2.3.1 Clustering.....	33
2.3.2 Molecular Diversity	34
2.4 Evaluation Measurement.....	35
2.5 Conclusion.....	37
Chapter 3 Nearest Neighbour Searching in High Dimensionality	39
3.1 Introduction.....	39
3.2 Issues with High Dimensionality Data	39
3.3 Effectiveness of Nearest Neighbour Search in High Dimensionality Data	43
3.3.1 Distance Measure Approach.....	44
3.3.2 Approximate Nearest Neighbour Approach	45

3.4 Dimensionality Reduction Approach.....	46
3.4.1 Feature Selection Method.....	46
3.4.2 Projective Method.....	48
3.4.3 Binary Fingerprint Dimensional Reduction Approach.....	52
3.5 Conclusion.....	54
Chapter 4 Methodology.....	57
4.1 Introduction.....	57
4.2 Dataset.....	57
4.2.1 MDDR.....	58
4.2.2 WOMBAT.....	58
4.2.3 ChEMBL.....	59
4.3 Molecular Representation.....	64
4.4 Similarity and Distance Measures.....	68
4.4.1 Similarity Coefficients.....	68
4.4.2 Distance Coefficients.....	69
4.5 Experimental Procedure.....	75
4.5.1 Procedure of Similarity Searching.....	75
4.5.2 Procedure of Clustering.....	75
4.6 Evaluation Method.....	77
4.6.1 Enrichment Factor.....	77
4.6.2 F-Measure.....	77
4.6.3 QPI-Measure.....	78
4.7 Statistical Method.....	79
4.7.1 Spearman's Rank Correlation.....	79
4.7.2 Kendall's W Test.....	79
4.7.3 Sign Test.....	80
4.7.4 The Wilcoxon Signed-rank Test.....	81
4.8 Conclusion.....	83
Chapter 5 Investigation into the Effect of Dimensionality on the Effectiveness of Similarity Searching.....	85
5.1 Introduction.....	85
5.2 Experimental Design.....	85

Table of Contents

5.3 Results and Discussion.....	86
5.3.1 Analysis of Spearman’s Rank Correlation.....	86
5.3.2 Analysis of Kendall’s W Test.....	87
5.3.3 Effect of Dimensionality on the Effectiveness of Similarity Searching.....	96
5.4 Conclusion.....	109
Chapter 6 Investigation into the Effect of Dimensionality on the Effectiveness of Clustering.....	111
6.1 Introduction.....	111
6.2 Experimental Design.....	112
6.2.1 Clustering Method.....	113
6.2.2 Cluster Analysis.....	116
6.3 Results and Discussion.....	118
6.3.1 Effects of Low Dimensionality on the Effectiveness of Clustering.....	124
6.3.2 Effects of High Dimensionality on the Effectiveness of Clustering.....	126
6.3.3 Effects of Clustering Partition on F Measure and QPI Measure.....	140
6.4 Conclusion.....	140
Chapter 7 Investigation into the Relative Importance of the Similarity Search Components using a Cross-Classified Multilevel Model.....	143
7.1 Introduction.....	143
7.2 Cross-Classified Multilevel Modeling.....	144
7.3 Model Implementation.....	146
7.3.1 MCMC Estimation.....	147
7.3.2 MCMC Diagnostics.....	148
7.4 Experimental Design.....	150
7.5 Initial Model.....	150
7.5.1 Relative Importance of Similarity Measures.....	152
7.5.2 Estimation of the Individual Activity Class Effect.....	154
7.5.3 Estimation of the Individual Fingerprint Effect.....	157
7.5.4 Estimation of the Individual Similarity Coefficient Effect.....	159
7.6 Extended Model I.....	160
7.6.1 Relative Importance of Similarity Measures.....	161
7.6.2 Estimation of the Individual Fingerprint Effect.....	165

Table of Contents

7.6.3 Estimation of the Individual Similarity Coefficient Effect.....	167
7.7 Extended Model II	168
7.7.1 Relative Importance between Fingerprint and Similarity Coefficient	169
7.8 Conclusion.....	171
Chapter 8 Summary and Future Work.....	173
8.1 Introduction.....	173
8.2 Overall Summary of Work and Findings	173
8.3 Implication of Results	176
8.4 Contribution to Knowledge	177
8.5 Strengths and Limitations.....	178
8.6 Suggestion for Future Research.....	178
References.....	181
Appendix A Additional Results of Chapter 5.....	203
Appendix B Additional Results of Chapter 6.....	211
Appendix C Additional Results of Chapter 7	229

List of Figures

Figure 2-1 Example of 2D Binary Fingerprints.....	18
Figure 2-2 Example of Fragment Dictionary in Fragment Based Dictionary Fingerprints.....	19
Figure 2-3 Example of Hashed Fingerprints.....	20
Figure 3-1 Effect of the Curse of Dimensionality Phenomenon.....	40
Figure 3-2 Query point and its nearest neighbour (from Beyer et al. 1999).....	41
Figure 3-3 Another query point and its nearest neighbour (from Beyer et al. 1999).....	42
Figure 3-4 Example of the projection from high dimensional to low dimensional variables using principal component analysis.....	50
Figure 3-5 Binary Fingerprint Folding Steps.....	53
Figure 5-1 Performance of the 31 similarity coefficients, as obtained from MDDR dataset, ordered from best (high mean rank values) to worst (low mean rank values).....	91
Figure 5-2 Performance of the 31 similarity coefficients, as obtained from WOMBAT dataset, ordered from best (high mean rank values) to worst (low mean rank values).....	93
Figure 5-3 Performance of the 31 similarity coefficients, as obtained from ChEMBL dataset, ordered from best (high mean rank values) to worst (low mean rank values).....	95
Figure 5-4 A subset of average enrichment values using top 1% of the ranked dataset in searches for the eleven MDDR activity classes using various Morgan Radius 2 fingerprint dimensions.....	97

Figure 5-5 Identification of identical scaffold using Murcko scaffold between the existing active molecules retrieved in a lower dimension and new active molecule retrieved in a higher dimension.....	101
Figure 5-6 A comparison of the Simple Matching similarity values for two molecules (inactive and active) to illustrate the effect of global similarity measure.....	105
Figure 5-7 Line plot measuring the average bits set, average enrichment curves and bit collision rate based on the average of 10 random molecules for MDDR dataset using various Morgan R2 fingerprint dimensions.....	108
Figure 6-1 General workflow of high dimensional chemical data clustering implementation using ShARC	115
Figure 6-2 ShARC performance for various high dimensional chemical data clustering based on Ward's algorithm using MDDR dataset of 10,254 molecules measured by Euclidean distance coefficient.....	116
Figure 6-3 Hierarchical cluster dendrogram with the red horizontal dotted line indicating the level of partition to define the number of clusters.....	117
Figure 6-4 Effects of dimensionality on Ward's clustering measured by (a) <i>F</i> -measure and (b) <i>QPI</i> -measure for MDDR dataset using various distance coefficients	122
Figure 6-5 Distribution histograms of pairwise distances for molecules in MDDR represented by various fingerprint dimensions and measured by Euclidean distance coefficient.....	128
Figure 6-6 Distribution histograms of pairwise distances for molecules in MDDR represented by various fingerprint dimensions and measured by Hamming distance coefficient.....	135
Figure 7-1 Diagram illustrating the influence variables of the enrichment factor in similarity search application.....	145

List of Figures

Figure 7-2 Comparison of two visual diagnostics for monitoring chain trajectories of one model which runs for different iterations; (a) 10,000 iterations and (b) 500,000 iterations	149
Figure 7-3 Caterpillar plot of the activity class-level residuals with 95% Bayesian credible intervals for ChEMBL dataset	155
Figure 7-4 Caterpillar plot of the fingerprint-level residuals with 95% Bayesian credible intervals for ChEMBL dataset.....	157
Figure 7-5 Caterpillar plots of the similarity coefficient-level residuals with 95% Bayesian credible intervals for ChEMBL dataset	159
Figure 7-6 Bar chart comparing the relative importance between the fingerprint and similarity coefficient effects for 15 activity classes of ChEMBL dataset.....	162
Figure 7-7 Heat map summarising the ranking of the variable effects for level 3 (fingerprint) for 15 activity classes of ChEMBL dataset.....	165
Figure 7-8 Heat map summarising the ranking of the variable effects for level 2 (similarity coefficient) for 15 activity classes of ChEMBL dataset.....	167
Figure A-1 A subset of average enrichment values using top 1% of the ranked dataset in searches for the fourteen WOMBAT activity classes using various Morgan Radius 2 fingerprint dimensions (Refer to Table A-1 for detail values) ...	204
Figure A-2 A subset of average enrichment values using top 1% of the ranked dataset in searches for the fifteen ChEMBL activity classes using various Morgan Radius 2 fingerprint dimensions (Refer to Table A-2 for detail values).....	206
Figure A-3 Line plot measuring the average bits set, average enrichment curves and bit collision rate based on the average of 10 random molecules for WOMBAT dataset using various Morgan R2 fingerprint dimensions (Refer to Table A-3 for detail values)	209
Figure A-4 Line plot measuring the average bits set, average enrichment curves and bit collision rate based on the average of 10 random molecules for ChEMBL	

dataset using various Morgan R2 fingerprint dimensions (Refer to Table A-4 for detail values).....	209
Figure B-1 Effects of dimensionality on Group Average clustering measured by (a) <i>F</i> -measure and (b) <i>QPI</i> -measure for MDDR dataset using various distance coefficients (Refer to Table B-1 for detail values).....	214
Figure B-2 Effects of dimensionality on Ward's clustering measured by (a) <i>F</i> -measure and (b) <i>QPI</i> -measure for WOMBAT dataset using various distance coefficients (Refer to Table B-2 for detail values).....	220
Figure B-3 Effects of dimensionality on Group Average clustering measured by (a) <i>F</i> -measure and (b) <i>QPI</i> -measure for WOMBAT dataset using various distance coefficients (Refer to Table B-3 for detail values).....	226
Figure C-1 Caterpillar plot of the fingerprint-level residuals with 95% Bayesian credible intervals for 15 activity classes of ChEMBL dataset.....	232
Figure C-2 Caterpillar plots of the similarity coefficient-level residuals with 95% Bayesian credible intervals for 15 activity classes of ChEMBL dataset	234

List of Tables

Table 2-1 Common Binary Similarity Coefficient (Holliday et al., 2003; Leach & Gillet, 2007)	32
Table 4-1 MDDR dataset with 11 activity classes	61
Table 4-2 WOMBAT dataset with 14 activity classes	62
Table 4-3 ChEMBL dataset with 15 activity classes	63
Table 4-4 Fingerprints used in this study (Riniker & Landrum, 2013; Landrum, 2016)	66
Table 4-5 The list of the binary coefficients	70
Table 4-6 The list of the distance coefficients (Jones et al., 2001)	74
Table 5-1 Spearman's rank correlations result	87
Table 5-2 Kendall's <i>W</i> results for the top 1% based on the average actives retrieved for MDDR dataset. Mean rank indicates the value of mean EF1% obtained from the EF1% values averaged over the 11 activity classes in the dataset. For each fingerprint dimension, the highest ranked similarity measure is marked by blue box and the lowest ranked by orange box for ease of reference.	90
Table 5-3 Kendall's <i>W</i> results for the top 1% based on the average actives retrieved for WOMBAT dataset. Mean rank indicates the value of mean EF1% obtained from the EF1% values averaged over the 14 activity classes in the dataset. For each fingerprint dimension, the highest ranked similarity measure is marked by blue box and the lowest ranked by orange box for ease of reference.	92
Table 5-4 Kendall's <i>W</i> results for the top 1% based on the average actives retrieved for ChEMBL dataset. Mean rank indicates the value of mean EF1%	

List of Tables

obtained from the EF1% values averaged over the 15 activity classes in the dataset. For each fingerprint dimension, the highest ranked similarity measure is marked by blue box and the lowest ranked by orange box for ease of reference..... 94

Table 5-5 Average enrichment values using top 1% of the ranked dataset in searches for the eleven MDDR activity classes using various Morgan R2 fingerprint dimensions. For each fingerprint dimension, the highest average enrichment value is marked by green colour and the lowest value by red colour for ease of reference..... 98

Table 5-6 Identification of identical scaffold based on the active molecules retrieved using a single reference from the Renin activity class of the MDDR dataset 100

Table 5-7 Average bits set and bit collision rate based on the average of 10 molecules for MDDR dataset using various Morgan R2 fingerprint dimensions... 107

Table 6-1 Effectiveness value of Ward's clustering measured by (a) *F*-measure and (b) *QPI*-measure for the MDDR dataset using various distance coefficients and fingerprint dimensions. The range of the standard deviation, σ , for the mean *F* is between 0.022 and 0.446 119

Table 6-2 Summary statistics of bits set and bit collision rate for (a) 10,254 molecules in MDDR dataset and (b) 13,813 molecules in WOMBAT dataset using various Morgan R2 fingerprint dimensions..... 125

Table 6-3 Summary statistics for distribution of pairwise distance measured by Euclidean [D4] distance coefficient for MDDR dataset using various fingerprint dimensions..... 127

Table 6-4 Summary statistics for distribution of pairwise distance measured by Hamming [D5] distance coefficient for MDDR dataset using various fingerprint dimensions..... 134

Table 7-1 Variables used in this study 150

List of Tables

Table 7-2 Variance estimation of similarity search components (4 level cross-classified model) for ChEMBL dataset.....	153
Table 7-3 Variance estimation of similarity search components (3 level cross-classified model) for 15 activity classes	163
Table A-1 Average enrichment values using top 1% of the ranked dataset in searches for the fourteen WOMBAT activity classes using various Morgan R2 fingerprint dimensions. For each fingerprint dimension, the highest average enrichment value is marked by green colour and the lowest value by red colour for ease of reference	203
Table A-2 Average enrichment values using top 1% of the ranked dataset in searches for the fifteen ChEMBL activity classes using various Morgan R2 fingerprint dimensions. For each fingerprint dimension, the highest average enrichment value is marked by green colour and the lowest value by red colour for ease of reference	205
Table A-3 Average bits set and bit collision rate based on the average of 10 molecules for WOMBAT dataset using various Morgan R2 fingerprint dimensions	207
Table A-4 Average bits set and bit collision rate based on the average of 10 molecules for ChEMBL dataset using various Morgan R2 fingerprint dimensions.....	208
Table B-1 Effectiveness value of Group Average clustering measured by (a) <i>F</i> -measure and (b) <i>QPI</i> -measure for the MDDR dataset using various distance coefficients and fingerprint dimensions. The range of the standard deviation, σ , for the mean <i>F</i> is between 0.000 and 0.625	211
Table B-2 Effectiveness value of Ward's clustering measured by (a) <i>F</i> -measure and (b) <i>QPI</i> -measure for the WOMBAT dataset using various distance coefficients and fingerprint dimensions. The range of the standard deviation, σ , for the mean <i>F</i> is between 0.055 and 0.336	217

List of Tables

Table B-3 Effectiveness value of Group Average clustering measured by (a) *F*-measure and (b) *QPI*-measure for the WOMBAT dataset using various distance coefficients and fingerprint dimensions. The range of the standard deviation, σ , for the mean *F* is between 0.000 and 0.466 223

Table C-1 Variance estimation of similarity search components (3 level cross-classified model) for 150 reference structures 229

Chapter 1 Introduction

1.1 Background

The discovery of new medications for many diseases such as depression and gastrointestinal disorders has increased the health, quality of life and life expectancy of patients. All of this was made possible through *drug discovery* processes conducted by various pharmaceutical companies for many decades.

Drug discovery is a process that aims to identify new drug candidates for a disease in pharmaceutical industry. The modern drug discovery pipeline consists of seven steps: (1) target identification, (2) target validation, (3) hit and lead identification, (4) lead optimisation, (5) pre-clinical testing, (6) clinical testing and (7) new drug application (NDA) and food and drug administration (FDA) approval (Rao & Srinivas, 2011).

The first step in this process is the target identification, which identifies and understands the role of a potential therapeutic drug target (i.e., a protein involved in a particular disease). Next step is to validate the target in order to make sure that the properties of the target produces the desired therapeutic effect. This is followed by the hit and lead identification, and lead optimisation, which involve the target and lead compound interactions. Hit and lead identification is a process of evaluating the initial screening hits assessed by technology-based approaches like high-throughput screening. The hits are often undergoing limited optimisation to identify promising lead compounds. For example, the limited optimisation may improve the binding affinities for biological target of initial screening hits (Craeto, 2016).

The lead optimisation involves more extensive techniques such as docking to improve the characteristics (i.e., ADMET - structure-based absorption, distribution, metabolism, excretion and toxicity) and the efficacy (i.e., bioactivity or bioavailability) of the drug. In this process, the quantitative structure-activity relationship (QSAR) methods are used to study the features of molecule that influence the ADMET characteristics. The docking and scoring computations will then be applied on the three-dimensional structures resulted from the

QSAR study to produce drug-like lead compounds (Moroy et al., 2012). The result of this process is the identification of final compounds that will be selected for clinical trials.

Finally, the NDA provides all information for the FDA, which approves that the new drug is safe and effective to be used. The drug discovery process can take about twelve to fifteen years and costs the pharmaceutical company about US\$2,870 million (2013 dollars currency) per compound brought to the market (DiMasi et al., 2016).

The need for screening larger compound libraries to increase the number of marketable drugs has encouraged the emergence of *high throughput screening (HTS)*. Through HTS process, hundreds of thousands of compounds can be screened per drug target per year. The technology was developed in the 1980s and the HTS capacity evolved greatly in the 1990s. The evolution includes focusing on small compound libraries and expands into improving several fundamental technologies such as high density microplates, high performance microliter dispensers, imaging and laboratory automation (Carnero, 2006).

The increase of HTS capacity has allowed thousands of compounds to be tested at the same time. This has led to the use of *combinatorial chemistry (CC)* technologies to produce more new compounds in a shorter time. Using this technology, a large array of compounds from sets of different types of building blocks is repeatedly produced in a systematic way (Terrett et al., 1995). Although there are millions of new compounds created, the drug discovery process could not be enhanced due to the lack of chemical diversity and drug-like compounds in the compound libraries. Therefore, various computational approaches are needed to process chemical structure in order to create a highly diverse and drug-like chemical compound library. One of these approaches, and the focus of this thesis, is chemoinformatics.

1.2 Basis of Chemoinformatics

Chemoinformatics is known as the application of informatics methods to solve chemical problems (Gasteiger, 2006). As defined by Brown (1998, p. 375),

chemoinformatics is *“the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization”*. In simple terms, chemoinformatics can be understood as a computational approach and scientific discipline that interface between chemistry, computer science and information science to process chemical data structure (Vogt & Bajorath, 2012).

The main focus of chemoinformatics is the manipulation of two-dimensional (2D) or three-dimensional (3D) chemical structures for searching, modeling and statistics (Willett, 2011a). The implementation of chemoinformatics approaches is not limited to research in chemistry and pharmaceutical domains. It has been adapted to other domains such as food sciences, agrochemicals and perfumes.

For example, the approaches have been used to: (1) process and characterise the structure-property relationship of chemicals relevant to food chemistry (Martinez-Mayorga & Medina-Franco, 2009; (Martinez-Mayorga, Peppard, Ramírez-Hernández, Terrazas-Álvarez, & Medina-Franco, 2014), (2) predict the toxicity of aquatic pesticides (Casalegno et al., 2006) and (3) predict sensory characteristics of chemical structures (Keller et al., 2017).

These studies contribute to the development of, among others, better food or supplements for health productivity, effective fertilizers for agricultural productivity and chemical agents for perfumed products. A latest review on chemoinformatics applications of QSAR in food and agricultural sciences was recently published by Kar et al., (2017).

The rise of computational technology has improved the ways in which chemoinformatics analysis is conducted and can be optimised (Chen, 2006). The growth of big data analysis has encouraged chemoinformatics studies to venture into more sophisticated methods such as deep learning for analysing chemical information (Gawehn et al., 2016; Goh et al., 2017).

1.3 Aims of Research

Molecular similarity is an important concept in chemoinformatics based on the “*Similar Property Principle*” (Johnson & Maggiora, 1990). According to this principle, molecules that have similar structures are likely to have similar properties. This principle underlies many chemoinformatics applications involving searching for the nearest neighbour molecule of a specified query molecule such as similarity searching and clustering (Willett et al., 1998).

The search for nearest neighbour molecules involves two important components: (1) the molecular representations or descriptors and (2) the similarity or distance coefficients. The process involves a comparison between the representations of two molecules using one of many existing coefficients. These coefficients measure the degree of similarity of the two molecules, in which the standard coefficient for chemoinformatics applications has been the Tanimoto coefficient (Willett, 2014). Chapter 2 introduces different similarity searching techniques and reviews different molecular representations and coefficients that are used in chemoinformatics applications.

One of the main obstacles of the nearest neighbour search is the “*curse of dimensionality*”, a term coined by Richard Bellman (Bellman, 1961). The phenomenon occurs when the performance of nearest neighbour search decreases as the dimensionality of the data representation increases (Agrawal et al., 1998; Weber et al., 1998). Beyer et al. (1999) reported that, as the dimensionality of the data increases, the ratio of the distance of a query point to its nearest neighbour and to its furthest neighbour tends to unity when measured by arbitrary distance measure. France et al. (2012) suggested that the effects of the nearest neighbour searching vary considerably, depending on the nature of the similarity coefficient that is used. Chapter 3 reviews issues of nearest neighbour search concerning high dimensionality data. It also introduces methods for dimensional reduction, including methods applicable to chemoinformatics datasets.

In chemoinformatics applications, a single molecule structure can be represented by multi-dimensional representations or descriptors (Todeschini & Consonni, 2000). These dimensions can be much higher than the object representations in most applications of pattern recognition and data mining. Despite the use of high dimensionality representations, nearest neighbour searches in the chemoinformatics domain have been found to be effective. Sastry et al. (2010) suggested that the use of larger bits representation is more effective than 1024 bits when searching for nearest neighbour using 2D binary fingerprints.

Therefore, a substantial study on the effect of dimensionality on the effectiveness of the nearest neighbour search application involving chemical datasets is essential to understand the reason why the behaviour seems to contradict the effect observed by the curse of dimensionality. To the researcher's knowledge, there has been no study conducted as such, and any possible behaviour to the changes of the dimensionality remains unclear.

Hence, the first aim of this study is to investigate the effect of dimensionality on the effectiveness of nearest neighbour search in chemoinformatics applications. Chapter 4 describes the methodology of the investigations. The investigations were conducted on two different applications and discussed in two different chapters: (1) similarity search in Chapter 5 and (2) molecular clustering in Chapter 6. These applications can be considered as involving large numbers of nearest neighbour searches.

The specific research objectives for the first aim are as follows:

- To provide a detailed, step by step evaluation of the effects of changing dimensions of 2D fingerprints on the effectiveness of the applications.
- To analyse the effects of using various types of similarity (or distance) coefficients on the effectiveness of the application when changing the dimensionality of the 2D fingerprints.

- To identify other potential factors contributing to the effects of changing the dimensionality of the 2D fingerprints on the effectiveness of the applications.

Next, as mentioned earlier, the search for nearest neighbour molecules involves two important components, i.e., the molecular representations and the similarity coefficients. Many studies have evaluated the effects of using different types of molecular representations or different types of similarity coefficients by varying only a single component. Todeschini et al. (2012) compared various types of similarity coefficients used for comparing the similarity of 2D fingerprints, while Hert et al. (2004) and Riniker and Landrum (2013) evaluated different 2D fingerprints used as molecular representations for similarity measures. Sastry et al. (2010) on the other hand, compared various combinations of parameter settings which include both 2D fingerprints and similarity coefficients. The research set out to determine the most generally useful parameter settings for the effectiveness of the similarity searching.

In other domains, researchers have investigated the relative importance of different components which contributed the performances of various applications (Garner & Raudenbush, 1991; Leckie, 2009; Bell et al., 2016). A novel method called *cross-classified multilevel modeling* has made it possible to investigate the relative importance of different sources of influences on a response (Goldstein, 1987; 2011). However, in the chemoinformatics domain, the relative importance between the similarity search components remains inconclusive. Despite their importance, this issue has not yet been investigated.

The reasons above have motivated the second aim of this study, which is to use cross-classified multilevel modeling to model the relative importance of similarity measure components. Different from previous comparison studies, this study considers both 2D fingerprints and similarity coefficients, and uses a novel statistical method in order to model their relative importance in determining the effectiveness of similarity-based virtual screening. The findings are reported in Chapter 7.

The specific research objectives for the second aim are as follows:

- To demonstrate the use of cross-classified multilevel modeling for the analysis of relative importance of various similarity search components.
- To identify the more important component between the 2D fingerprints and similarity coefficients in determining the effectiveness of the similarity measures.

The conclusions that can be drawn from the work conducted in this thesis are summarised in Chapter 8, along with suggestions for future research.

1.4 Organisation of Thesis

The dissertation is organised as follows:

Chapter 2 begins by discussing the concept of virtual screening applications in chemoinformatics. This involves the key components of molecular similarity application that are molecular representation and descriptor, weighting scheme and similarity coefficient. It also introduces the basic concept of two other chemoinformatics applications, that are clustering and molecular diversity.

Chapter 3 is concerned with nearest neighbour searching in high dimensionality. It discusses issues concerning high dimensionality data and methods for dimensional reduction.

Chapter 4 presents the methodology of the experiments conducted in this thesis. This includes the introduction of the chemical datasets (i.e., MDDR, WOMBAT and ChEMBL), molecular representations, similarity and distance measures, application procedures, evaluation methods and statistical methods.

Chapter 5 is the first experimental chapter on the investigation of the effect of high dimensionality on the effectiveness of the similarity search application. The results are analysed and discussed within this chapter.

Chapter 6 expands the investigation in the previous chapter and looks at the effect of high dimensionality on the effectiveness of the clustering application. The results are analysed and compared between different clustering methods implemented.

Chapter 7 introduces cross-classified multilevel modeling and uses this method to identify the relative importance of similarity search components in determining the effectiveness of a similarity search.

Finally, Chapter 8 provides the reader with the conclusions of this thesis, its limitations and an overview of possible future research directions.

Chapter 2 Similarity Searching in Chemoinformatics

2.1 Virtual Screening

Virtual screening is an *in silico* technique in chemoinformatics which aims to identify and prioritize candidate compounds for *in vitro* experiments. It uses computational methods to search large sets of chemical compounds in order to find compounds that are most likely to be bioactive. HTS, on the other hand, screens large numbers or sets of chemical compounds in the laboratory experiment, which involves a controlled environment and equipment. The increasing size of compound databases has led to the implementation of virtual screening using high-performance computing, which can involve advanced computer processors and parallel programming. This approach is more cost effective to drug discovery than the traditional HTS (Heikamp et al., 2013).

The types and amounts of data that are available determine the virtual screening method. First, similarity-searching methods are used when only a single active molecule is available. Second, pharmacophore methods are used when there are several active molecules with associated structures available. Third, machine-learning methods are used when significant numbers of both active and inactive molecules are available. Finally, docking methods are used when the 3D structure of the biological target is available. Categorised into two groups, similarity searching, pharmacophore mapping and machine learning are examples of ligand-based virtual screening (Ripphausen et al., 2011), while docking is a structure-based virtual screening method (Lyne, 2002).

Similarity searching identifies compounds in a database that are structurally similar to the target structure. The approach implements a quantitative comparison between the *target structure* with each structure in the database to produce a ranking of database compounds in decreasing order of similarity to the target, which is usually a known active structure. The top of the list are the nearest neighbours to the target structure, which exhibit the most structural resemblance. Willett (2014) summarised the main components of similarity

measures in similarity searching (Willett, 2014). Recent research studies have considered the technique of combining different approaches, i.e., data fusion to improve the effectiveness of similarity searching. Data fusion can be used to combine different similarity measures, e.g. combining different fingerprints, or different virtual screening methods. The approach captures different chemical information resulting to the highest-ranked hits from the combinations. Hence, this optimal search and combination may increase the performance (Cereto-Massagué et al., 2015a).

Pharmacophore methods aim to identify the key common features from a set of active molecules that bind to an identical target molecule. The common features, which represent the essential interactions between the ligand and a specific molecular target, were extracted from 3D structures of known active molecules. Thus, one can make an assumption that the other molecules which contain the similar pattern may also exhibit the same biological activities. The main advantage of this method is to provide better understanding on target and ligand interactions as well as improving the screening hit rates during *in vitro* experiments (Langer et al., 2004).

Machine learning also aims to analyse the structural characteristics of molecules but for the purpose of classifying the active or inactive compounds. This method works by developing and training a model using machine learning methods. It requires input of a training set, which consists of a set of molecules that had previously been tested and shown to be either active or inactive. These training set molecules are then analysed to develop a decision rule that is used to classify new molecules (the test set). Geppert et al. (2010) surveyed data mining approaches which are applicable to machine learning in compound classification. Their analysis focused on the novel algorithms and methods of data mining that are support vector machines, Bayesian classifiers, decision trees and inductive logic programming.

Docking programs identify 3D structures that are complementary to, and are predicted to bind to, the 3D protein active site. Docking is performed by the search algorithm and the scoring function. The docking algorithm is used to

determine an optimal position and conformation of the ligand in the active site. Following this, the scoring function evaluates the conformation of the positioned ligand in the active site and its interactions. Several studies have reviewed in-depth methods and applications of scoring and docking (Kitchen et al., 2004; Ghosh et al., 2006). Cheng et al. (2012) suggested a few practical aspects to improve docking programs while (Wójcikowski, Ballester, & Siedlecki, 2017) proposed a new machine-learning scoring function that improves the performance of virtual screening and the prediction of binding affinity.

Ranking the truly active molecules as high as possible and inactive ones as low as possible has become one of the issues in virtual screening. This is because virtual screening evaluates large amounts of chemical data, in which the number of actives retrieved is important. A study by Scior et al. (2012) mentioned several drawbacks of various aspects in virtual screening methods which related to this issue. Among the possible solutions, as suggested, are careful preparations of database, correct parameter settings and good choice of algorithm for implementation.

As described above, the similarity searching approach is used to rank the active molecules in a chemical database. Having introduced what is meant by this approach, the chapter will now move on to describe the similarity searching approach in detail and discuss its main components in the next section.

2.2 Molecular Similarity

The past decades have seen the rapid development of molecular similarity in chemical structures database research. Molecular similarity is a concept that aims to identify molecules which have the same bioactivity as a bioactive target structure.

Molecular similarity is a concept based on the similar property principle that was first presented by Johnson and Maggiora (1990). The principle states molecules that are structurally similar are likely to have similar properties. This also indicates that the nearest neighbours of a bioactive target structure are also

likely to possess that same bioactivity. One of the exceptions to this concept is called the *activity cliffs* (Stumpfe & Bajorath, 2012). In general, an activity cliff is a pair of structurally similar compounds having a large difference in potency. It happens when a small change in molecular structure causes large changes to its activity. However, despite this exception, the impact of activity cliffs provides researchers with fundamental information to understand the underlying structure-activity relationship (SAR) of the datasets (Cruz-Monteagudo et al., 2014).

The significant contribution of the similar property principle to the lead generation and optimisation efforts can be the reason why the principle remains applicable to the development of molecular similarity applications. The most important application of molecular similarity is probably similarity searching as introduced in Section 2.1. It was developed as a way of overcoming the limitations of substructure searching, i.e., finding all molecules in a database that contain a user-defined query substructure (Leach & Gillet, 2007).

The main component of the similarity searching approach is the measure used to quantify the similarity between the target structure and each database structure. A measure comprises these components: molecular descriptors, weighting scheme and similarity coefficient. Molecular descriptors are used to represent characteristics of molecules that are being compared in a computer readable format. The weighting schemes, on the other hand, prioritise the contributions of different parts of the representation. The similarity coefficient is used to quantify the degree of structural resemblance between pairs of molecules (Willett, 2014).

The search starts with calculating the degree of similarity between the target structure and each of the molecules in the database. Following this, the database is ranked in order of decreasing similarity. As the principle stated, the top ranked molecules, which are the nearest neighbour molecules, are considered as the most similar to the target structure's bioactivity. Results of this search, which are the top ranked molecules, are therefore selected for the subsequent experimental testing (Willett, 2009).

Stumpfe and Bajorath (2011) discussed important principles of similarity searching and reviewed major categories of searching methods, i.e., molecular representation and descriptors of similarity searching (e.g., 2D and 3D). The review highlighted several reasons for the development and application of similarity searching, e.g., similarity searching can be applied when little or nothing is known about compound structure-activity relationship. This view was mentioned earlier by Sheridan and Kearsley (2002), who pointed out the similar reason for the establishment of similarity methods in the pharmaceutical setting. It has also been suggested by Stumpfe and Bajorath (2011) that the chemoinformatics community needs to establish calculation standards and evaluation criteria that enable a meaningful comparison for different similarity search methods.

The next sections focus on the detail of (a) different types of representation and descriptors (b) implementation of weighting schemes (c) various groups of similarity coefficients as the key components of the similarity measures that lie at the heart of the similarity searching approach.

2.2.1 Representation and Descriptors

A molecule's structure is an important data for chemoinformatics applications, e.g., similarity searching. To enable the computer to process such applications, a molecule's structure is represented by a machine-readable format, which can be identified by a *unique compound identifier*. One of the common identifiers is referred to as a CAS Registry Number, which is a numeric identifier designated by the Chemical Abstract Service (CAS) (Chemical Abstracts Service, 2015). Warr (2011) pointed out several limitations for these compound identifications: (i) complexity of the identifier for chemical structure processing and (ii) meaningless identifier to the chemists.

The limitations of compound identification motivate the widespread implementation of encoding molecular structures into more meaningful and unique molecular representation. A few examples of encoded molecular structures are *line notations* (a linear string of alphanumeric symbols) and

connection tables (a table form of molecular graph). Simplified Molecular Input Line Entry System (SMILES) is one of the well-known line notations because of its easy implementation while connection tables are often used by common file formats, e.g., Structure-Data File (SDF), for describing molecule structure information (Weininger, 1988).

Molecular descriptors, on the other hand, are numerical values that characterize properties of molecules. As stated by Brown (2009), "*molecular descriptor are descriptions of molecules that aim to capture the salient aspects of molecules for application with statistical methods*".

Molecular descriptors can be classified into 1D (whole molecule), 2D and 3D. Todeschini and Consonni (2000) have briefly introduced various types of descriptors. For implementation, a wide range of software has been developed for generating and calculating molecular descriptors for the use of molecular similarity applications (Steinbeck et al., 2003; Yap, 2011; Cao et al., 2013; Vasilyev et al., 2014).

2.2.1.1 1D Descriptors

1D descriptors define a molecule by a single value. Pipeline Pilot can be used to calculate (or model) a molecule's structure or its chemical properties using certain mathematical (or modeling) functions to produce 1D descriptors, i.e., *structural features* or *physicochemical properties*. There are various examples of 1D descriptors, i.e., simple integer counts (e.g., number of atoms, bonds and ring assemblies) and chemical properties that could be in either integer or real values (e.g., $\log P$ and molecular weight).

$\log P$ (octanol-water partition coefficient), for example, is a chemical property that quantifies molecular hydrophobicity. It determines the activity and transport of drugs, e.g., drug absorption, bioavailability and hydrophobic drug-receptor interactions.

Although 1D is the most simple and computationally fast descriptor (Leach & Gillet, 2007), it does suffer from a number of flaws. A single such descriptor on

its own is an insufficient molecular discriminant (Willett, 2014). Hence, a molecule will normally be represented by a vector, each element of which represents a single 1D descriptor. The values are calculated and normalised using certain mathematical functions or models to ensure that all of the attributes in the molecular representation are measured on the same scale (Chu et al., 2009).

However because of the advantage and importance, many researches are still using 1D descriptors as part of their QSAR studies (Nicolotti & Carotti, 2006) as well as the components in rule-based approaches (Bajorath, 2001). For example, four physicochemical parameters, i.e., molecular weight and sum of nitrogen, oxygen, and hydrogen-bond acceptors were used by Lipinski et al. (2012) in the experiment of solubility and permeability prediction in drug discovery.

2.2.1.2 2D Descriptors

A molecular graph representation provides a useful way of organizing molecular structure for 2D molecular database analysis (Bemis & Murcko, 1996). It consists of sets of nodes and edges, which represents a molecule's framework. The nodes of the graph correspond to the molecule atoms, while the edges correspond to the chemical bonds of the atoms. This information, therefore, becomes the basis of many 2D descriptors. Examples of 2D descriptors are topological indices and structural fragments as described in this section.

Topological indices or *connectivity indices* are single-valued 2D descriptors that are calculated based on the molecular graph of a chemical structure. Topological indices aim to characterize molecules based on size, degree of branching, flexibility and overall shape as a whole. A typical way to calculate a topological index is by multiplying the values or some function of adjacent vertices such as square root, and then summed across all edges (Dearden, 2017). In 1947, Wiener reported the first example of topological indices, i.e., the Wiener Index (Wiener, 1947). The *Wiener Index* is defined as the sum over all topological

distances in the molecule. It counts the number of bonds between each pair of atoms and sums the distances between each pair. It can be calculated using the following Eq. (1),

$$W = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N D_{ij} \quad (1)$$

where N is the number of atoms in the molecules, subscripts i and j are the atoms and D is the shortest path distance between i and j .

Another example is the *molecular connectivity index*, which is one of the well-known topological indices that was first reported by (Randić, 1975). The molecular connectivity index is defined as the sum of bond contributions calculated from the vertex degrees (number of graph edges) of each atom in the hydrogen suppressed (non-hydrogen atoms) molecular graph.

As suggested by Kier and Hall, (2001) and Estrada (2002), the molecular connectivity index is a good measurement for the molecular surface area (i.e. a measure of molecular size) and is rich in molecular structure information. The molecular area is useful in measuring the extension of intermolecular interactions. The molecular connectivity index is also valuable in quantifying the relationship between structure and physical properties.

By drawing on the concept of molecular connectivity index, a simple example of connectivity index calculation is described by (Livingstone, 2000). First, each atom in a molecule is assigned a degree of connectivity, which indicates the number of adjacent non-hydrogen atoms (hydrogen-suppressed). Second, the *bond connectivity*, C_k , for each bond in the structure is calculated by taking the reciprocal of the square root of the product of the connectivities of the atoms. The calculation is given by the following Eq. (2),

$$C_k = \frac{1}{\sqrt{(\delta_i \delta_j)}} \quad (2)$$

where δ_i and δ_j refer to the degree of connectivity to each atom i and j . Finally, the molecular *connectivity index*, χ , for a molecule is calculated by summation of the bond connectivities over all of its bonds given by Eq. (3),

$$\chi = \sum_{k=1}^N C_k \quad (3)$$

Extended chi indices were developed to overcome one of the issues with the molecular connectivity index, i.e. direct representation of molecular structure, which require more than single index of molecular connectivity indices to encode structure information (Hall & Kier, 2001). They aim to provide greater sensitivity to structure variation by adopting an algorithm similar to the molecular connectivity index algorithm. Extended chi indices involve a set of chi indices that encode a wide range of structure features for a molecular characterization.

However recently, (Randić, 2014) suggested that single topological indices may be suitable for molecular similarity studies. The research outlined a general approach for constructing '*generalized connectivity indices*' that was used as a single molecular descriptor for molecular characterisation. The new topological descriptor is also appropriate for screening huge combinatorial libraries due to its conceptual and computational simplicity.

The second example of a 2D descriptor is based on structured fragments. For *structural fragment* descriptors, a molecule is characterised by its fragment substructures. The occurrence of these fragment substructures is derived from a connection table and encoded into a 2D vector of elements called a fingerprint. Each 2D fingerprint element describes the presence or absence of molecular

features, thus two molecules are considered similar if their fingerprints share common values for many of the constituent elements (Willett, 2014).

2D fingerprints became the most common descriptors used for molecular similarity due to their simplicity and efficiency. Many researchers have reviewed and studied various aspects of 2D fingerprints in molecular similarity, which includes 2D fingerprint comparisons and their application in similarity searching (Duan et al., 2010; Willett, 2014; Cereto-Massagué et al., 2015).

There are many types of 2D fingerprints; the most common fingerprints are binary (Hert et al., 2004). Binary fingerprints are represented by a bit string, which encodes the present features by '1' and '0' for the absent ones (Figure 2-1). Binary fingerprints are especially useful, as there are highly efficient computer science algorithms that work with binary strings.

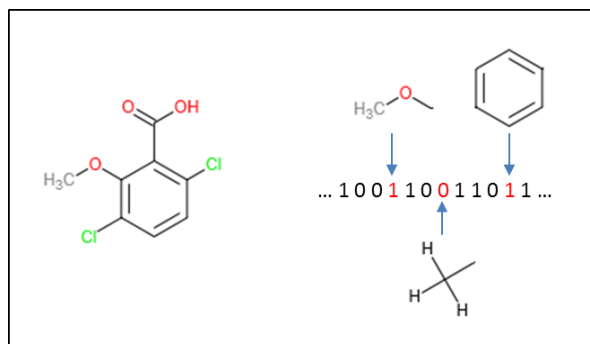


Figure 2-1 Example of 2D Binary Fingerprints

2D binary fingerprints can be classified into fragment based dictionary fingerprints or hashed fingerprints. *Fragment based dictionary* fingerprints are based on pre-defined fragments. Each bit position in the fingerprint corresponds to a specific substructure fragment. The fragment dictionary contains different predefined molecular fragments (Figure 2-2).

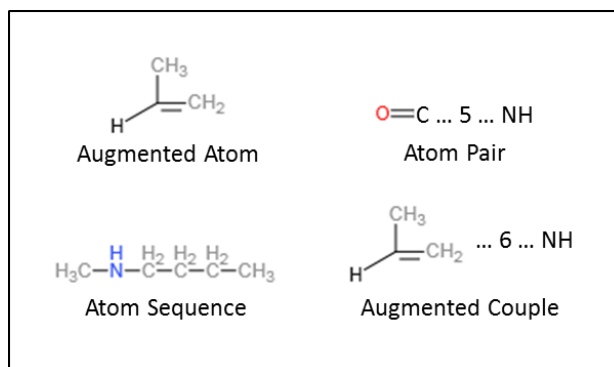


Figure 2-2 Example of Fragment Dictionary in Fragment Based Dictionary Fingerprints

Common examples of fragment based dictionary fingerprints are MDL MACCS keys (Keys, 2002) and BCI keys (Barnard & Downs, 1997). For example, MDL 166-key structural key (known as MACCS keys) defines 166 fragments that are considered important in medicinal chemistry.

A number of authors have attempted to implement fragment based dictionary fingerprints in their experiment. Durant et al. (2002) have demonstrated that reoptimised MDL fingerprints have shown an improvement in the performance when applied to the standard 166 and 960-bit keysets in molecular similarity application.

In contrast, hashed fingerprints do not need a fragment dictionary. Each fragment is processed using several hash functions that each set one or more bits in the fingerprint (Figure 2-3). Based on a specified length of bond connection, each fragment in a molecule is analysed for its linear path. These paths are hashed to produce the bits in a fingerprint. Fragments and bits in the bit string are mapped by many-to-many.

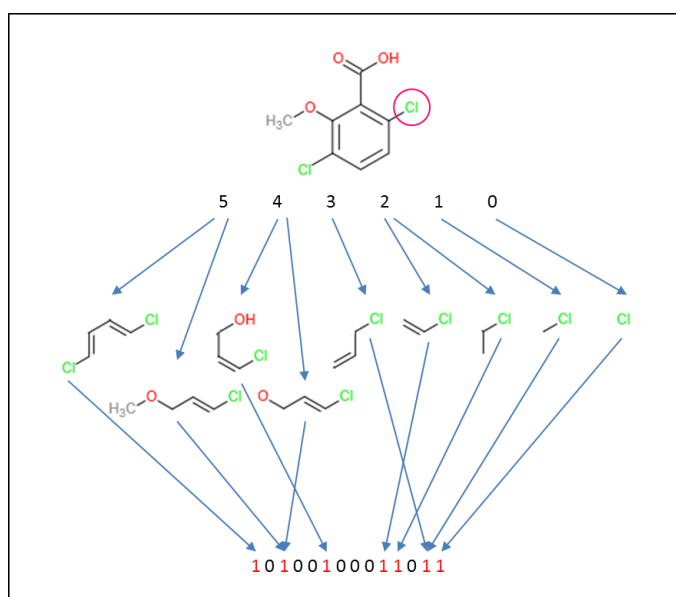


Figure 2-3 Example of Hashed Fingerprints

A common example of hashed fingerprints is the Daylight fingerprint (James et al., 1995). In the Daylight algorithm, the fingerprint is derived from hashing all possible linear paths for a given length of bond connection. The fingerprint is then hashed into a fixed length of bit string. Fingerprints may be folded to decrease the length and increase the bit density. Typical sizes for Daylight fingerprints are 512 or 1024 bits in length depending on the hashing algorithm.

2.2.1.3 3D Descriptors

In 3D similarity searching systems, the geometric patterns of functional groups in molecules is one of the contemporary methods used to derive 3D descriptors (Bajorath, 2001). These patterns are chosen based on their importance to specific molecule activities. Many studies have implemented the 3D descriptors to find the correlation between similarities of individual compounds and their biological activities (Kubinyi, 1997; Nicolotti & Carotti, 2006; Almeida et al., 2014). The common examples for 3D descriptors are 3D pharmacophore, 3D fingerprint and electrostatic interaction fields.

A *pharmacophores* is the spatial arrangement of atoms or groups in a small molecule that are responsible for its biological activity (Martin, 1992). The key

importance of pharmacophore representations is the type of features (e.g., hydrophobic) and distance between the features (e.g., distance matrix) (Bender & Glen, 2004). A pharmacophore query is searched against 3D conformations of database compounds. It preenumerates multiple conformations for each compound in the database to identify compounds that have similar chemical features to the query. This process requires prior knowledge (hypothesis) of the features, which determine the activity. The hypothesis of the features can be derived from the pharmacophore elucidation methods, which involve the preparation of data set, generation of possible pharmacophores and pharmacophore validation.

3D fingerprints captures pharmacophore arrangements derived from systematic conformational analysis of test molecules. In 3D pharmacophore fingerprints, each bit position is assigned to an individual pharmacophore pattern of predefined feature points and inter-feature distance ranges. The bit is set to '1' if the conformational ensemble of a molecule satisfies the features and distance ranges of a given pattern and vice versa (Cereto-Massagué et al., 2015b).

Electrostatic interaction fields, which are derived from 3D grid representations, are another example of descriptors in 3D similarity studies. In this approach, interaction field energies from each grid point of query and test compound are calculated. Based on the result, both interaction fields are then aligned to best match interaction energies. Despite being time consuming, this type of descriptor provides a global measurement of molecular similarity and continues to interest many studies (Cheeseright et al., 2006).

The 3D descriptors, which are based on molecular shapes, are also widely implemented in molecular shape similarity applications (Finn & Morris, 2013). One of the common approaches is to use a mathematical function, e.g., Gaussian function, to calculate the volume of a molecule as a descriptor (Grant et al., 1996).

The 3D descriptors provide different degree of molecular information as compared to the 2D descriptors that are based on molecular graphs. For example, the intermolecular forces that are important for ligand-receptor

binding are more dependent on the 3D structural properties rather than the presence of 2D fragments (Brown & Martin, 1997). However, 3D descriptors suffer from several important drawbacks, e.g. high in computational cost because of its intensive calculations. This also includes finding correct common features and ability to align molecules in a 3D similarity searching.

2.2.1.4 Effect of Descriptor Correlations

The selection of the descriptors has become one of the important steps in chemoinformatics applications. This is because the use of highly correlated descriptors can affect the data representation and analysis. Several reviews have also suggested to avoid the use of highly correlated descriptors (Xu & Hagler, 2002; Maldonado et al., 2006; Leach & Gillet, 2007; Clarke et al., 2008).

Correlation methods offer an effective way to measure the degree of the linear correlation between two variables (descriptors). The sign and the value of the correlation coefficient describe the direction and the degree of the correlation. Pearson correlation is one of the common measures used to calculate the correlation (Field, 2013). The calculation for the Pearson correlation is defined in Eq. (4):

$$r = \frac{cov_{xy}}{\sigma_x \sigma_y} \quad (4)$$

where r is the correlation coefficient and cov_{xy} is the covariance of the two variables divided by the product of their standard deviations. The covariance is calculated by multiplying the deviations of one variable by the corresponding deviations of a second variable. The averaged sum of combined deviations is then divided by the number of observation (Field, 2013). A coefficient of +1 indicates a perfect positive correlation, while the coefficient of -1 indicates a perfect negative correlation. A coefficient of 0 indicates no linear correlation between the measured variables.

The correlation matrix is used to represent the pairwise correlation when multiple variables are being measured (Leach & Gillet, 2007). For each entry in the matrix, the calculation for the correlation coefficient is performed using another variation of Eq. (4) as defined in Eq. (5):

$$r = \frac{\sum_{k=1}^N [(x_{i,k} - \bar{x}_i)(x_{j,k} - \bar{x}_j)]}{\sqrt{\sum_{k=1}^N (x_{i,k} - \bar{x}_i)^2 \sum_{k=1}^N (x_{j,k} - \bar{x}_j)^2}} \quad (5)$$

where r is the correlation coefficient between variables x_i and x_j .

(Kümmel et al., 2011) used a correlation matrix to eliminate the highly correlated variables in the multivariable data analysis. This method calculates a pairwise correlation matrix for all of the variables. Next, it determines a pair of variables with the highest correlation coefficient. For these two variables, this method calculates the sum of all correlation coefficients to all other variables. The variable with the highest sum of correlation coefficients is then eliminated. This method was used to reduce the number of variables. Thus, it was repeated until the desired number of variables is reached.

2.2.2 Weighting Scheme

The weighting scheme is another main component in molecular similarity searching, which is important for prioritisation of features in molecular similarity (Maggiora et al., 2014). The *weighting scheme* aims to emphasise the differences between the various components of a molecular representation. It assigns different degrees of importance to the various components of molecular representations. If applied to molecular features, a certain feature in a molecule is considered more important than other features if it has higher weight assigned to it.

There have been a few types of weighting scheme discussed in the molecular similarity domain. First, a weighting scheme based on the number of times that a fragment occurs in an individual molecule. Second, a weighting scheme based

on the number of times that a fragment occurs in the entire database. Third, a weighting scheme based on the total number of fragments within a molecule (Willett et al., 1986). Extensive experiments have been carried out by Arif et al. (2009) focusing on weighting of fragments on the basis of their frequencies of occurrence in molecules. The work continues with an introduction of *inverse frequency weighting*, which discussed specifically the use of weights that assign greatest importance to the substructural fragments that occur least frequently in the compound database (Arif et al., 2010).

The next subsections describe how weighting schemes, are being implemented in binary and non-binary fingerprints for molecular similarity purposes. These sections require an understanding of the different types of fingerprints.

2.2.2.1 Binary Fingerprints

In 2D *binary fingerprints*, the weighting scheme is applied to encode merely the presence and absence (*incidences*) of topological substructures in a molecule. Although binary fingerprints are an extremely simple type of structural representation, they contain sufficient information for effective similarity searching to be successfully carried out. Ewing et al. (2006) have demonstrated the development of a set of new 2D fingerprints for virtual screening, which involved weighting in order to assess the range of frequencies encoded for drug-like molecules. In another study, binary fingerprints have also been used for similarity coefficient analysis (Todeschini et al., 2012).

However, binary fingerprints may not be able to describe the relative degree of importance of substructure fragment occurrence in a molecule. This disadvantage limits the identification of which fragments are making higher contribution to the overall degree of similarity and which are not. The weighted fingerprint (count fingerprint) overcomes this limitation. It is introduced and described in the next section.

2.2.2.2 Weighted Fingerprints

The *weighted fingerprint* is another type of 2D fingerprint, which encodes the substructural fragments in a molecule based on their *occurrence* rather than the incidence. It aims to differentiate the level of contribution from each substructure fragment in a molecule. The weighted fingerprint, which is commonly referred to as the *count fingerprint*, yields an integer or real vector rather than a binary fingerprint.

In the weighted fingerprint, a high-weighted fragment that is common to both target structure and database compounds determines the importance of that fragment, among others in both molecules. Thus, this fragment provides greater contribution to the overall degree of similarity than the low-weighted fragments.

Arif et al. (2009) investigated the effect of weighted fingerprints using individual molecule fingerprints. They have concluded that the weighted fingerprints are more effective than the non-weighted, conventional binary fingerprints in molecular similarity searching. The result suggests the standardization of raw occurrence frequencies to maximise the effectiveness. They also found that small variations in weighting scheme could potentially affect the magnitude of the Tanimoto coefficient due to its defined mathematical formulation.

Arif et al. (2010) have further investigated the *inverse frequency weighting*, which considers the occurrence of fragments within the entire database by assigning the greatest weights to those substructural fragments that occur least frequently in the screened database. The experiment found that if two molecules have in common a fragment that occurs only rarely in the database as a whole, then they should be regarded as being more similar than if they have in common a fragment that occurs very frequently.

2.2.2.3 Standardisation Method

Standardisation is a mathematical function that is implemented in molecular similarity searching as well as in many other domains of data mining (Su et al., 2009). In molecular similarity, *standardisation* aims to ensure that all of the attributes comprising a molecular representation are measured on the same scale. This is to avoid any variable domination in similarity calculation, which involves descriptors measured on different scales.

Standardisation calculates on real-valued or integer-valued data of molecular representation such as different types of physicochemical attributes. Examples of these attributes include the logP, molecular weight and number of rotatable bonds. One of the most common standardisation methods in molecular similarity is *Z standardisation* (Milligan et al., 1988). It computes the mean and standard deviation for molecular representation attributes into zero and unity, respectively. To get a z-score, subtract the mean from each data value and divide by the standard deviation. The new set of data is then comparable for the similarity calculation.

Previous research investigated the effectiveness of standardization in chemical clustering and similarity searching, and concluded that the choice of standardisation method is not a critical component of procedures for molecular clustering and searching. This is because there is no consistent performance benefit that is likely to be obtained from the use of any particular standardization method (Chu et al., 2009).

2.2.3 Similarity Coefficient

The effectiveness of measurement in molecular similarity is highly dependent on the third component described in this section, the similarity coefficient. The similarity coefficient provides the quantitative measure of the degree of structural relatedness between two comparable molecules. The usefulness of similarity coefficients has been addressed in various applications such as similarity, clustering and molecular diversity (Todeschini et al., 2012; Haranczyk et al., 2008; Matter, 1997).

Studies focusing on the comparative studies between similarity coefficients have been conducted in various methodologies. Early work by Willett et al. (1986) compared the effectiveness of six similarity coefficients for intermolecular structural similarity. Haranczyk et al. (2008) also reported the relative performance of association and correlation coefficients in their clustering and compound selection studies. Al Khalifa et al. (2009) continued the work by investigating the relative performance of similarity coefficients on non-binary data using (dis)similarity-based techniques. Todeschini et al. (2012) recently analysed and compared a large number of similarity coefficients for binary fingerprint similarity searching.

The similarity coefficient may be divided into three main categories, which are based on the practical uses: (i) *association coefficient*, if the molecular query needs to measure the compound's degree of association; (ii) *correlation coefficient*, if the molecular query requires a degree of proportionality and independence; (iii) *distance coefficient*, if the molecular query seeks for distance between the target compound and itself in the descriptor space (Ellis et al., 1993; Willett et al., 1998; Holliday et al., 2002). Coefficients for each category are described below.

2.2.3.1 Association Coefficient

The *Association coefficient* aims to measure similarity according to the number of common features between the two representations. It reflects the association or resemblance of two molecules that are being compared.

There are many types of association coefficient, with the *Tanimoto coefficient* being the most effective due to its simplicity and accuracy in binary similarity searching (Willett et al., 1998). The Tanimoto coefficient, also known as the *Jaccard coefficient*, can be used with both binary and weighted variables (Al Khalifa et al., 2009). The binary variant of the Tanimoto coefficient is defined by Eq. (6):

$$T_c(A, B) = \frac{c}{(a + b - c)} \quad (6)$$

where a is the number of set bits in fingerprint A (target compound), b is the number of set bits in fingerprint B (compared compound) and c is the number of set bits common to both fingerprints. For binary similarity measurement, the output value ranges between 0 to +1, where the highest similarity is indicated by the value +1. In a non-binary case (i.e., using non-binary descriptors), the Tanimoto coefficient is defined as Eq. (7):

$$T_c(A, B) = \frac{\sum a_i b_i}{\sum a_i^2 + \sum b_i^2 - \sum a_i b_i} \quad (7)$$

where the summation of all elements in the fingerprint is divided by the magnitude of fingerprint A added to the magnitude of fingerprint B , minus the summation of all elements. For non-binary similarity measurements, the output value ranges between $-1/3$ to +1. More examples of common association coefficients used for binary variables in chemoinformatics are listed in Table 2-1.

Willett (2006) has demonstrated the effectiveness of various similarity coefficients when applied to binary similarity searching. The research concludes that Tanimoto is effective for 2D fingerprint similarity searching. However, research by Todeschini et al. (2012) suggest that other coefficients are potentially effective for the similarity searching of binary fingerprints.

The latter outcome is similar to that experimented with non-binary descriptors. Likewise, Holliday et al. (2012) also found out that another coefficient, the Cosine coefficient, is more robust than the Tanimoto coefficient when applied to weighted fingerprint similarity searching. It is reported that the Cosine coefficient's screening abilities are much less affected by the precise nature of the weights applied to the fingerprints for both target structure and database structures, which has become the limitation of the Tanimoto coefficient.

2.2.3.2 Distance Coefficient

The *Distance coefficient* is also referred to as the *dissimilarity coefficient*. It aims to measure the difference between the two representations. There are many types of distance coefficients, which are based on simple geometric interpretation. The *Euclidean distance coefficient* is one of the examples used in many applications including molecular similarity and multivariate statistics (Champely et al., 2002). The binary variant of the Euclidean distance coefficient is defined as in Eq. (8):

$$E_C(A, B) = \sqrt{a + b - 2c} \quad (8)$$

where a is the number of set bits in fingerprint A (target compound), b is the number of set bits in fingerprint B (compared compound) and c is the number of set bits common to both fingerprints. The output value ranges between 0 to N , where N is the total bit length. The minimum value of 0 indicating that two compounds are identical, and the maximum value of N indicating the most dissimilarity. In a non-binary case (i.e., using non-binary descriptors), the Euclidean distance coefficient is defined by Eq. (9):

$$E_C(A, B) = \left[\sum_{i=1}^n |x_i - y_i|^2 \right]^{1/2} \quad (9)$$

where a_i is the value for each fragment of fingerprint A (target compound) and b_i is the value for each fragment of fingerprint B (compared compound). For non-binary similarity measurements, the output value ranges between 0 to ∞ , where the minimum of 0 indicates that two compounds are identical. More examples of common distance coefficients used for binary variables in chemoinformatics are listed in Table 2-1.

Distance coefficients are used to measure the distance between structures in a molecular space. Since it is difficult to visualise the geometry of a space of M dimensions when M is more than 3, the validity of geometric distances between

objects in a hyperspace of M dimensions are said to be preserved if the coefficient that is used has the property of a *metric*. If a distance coefficient fulfils a few properties it can be described as a metric (Willett et al., 1998). The properties are: (i) distance values must be zero or positive, and the distance from an object to itself must be zero; (ii) distance values must be symmetric; (iii) distance values must obey the triangular inequality and (iv) distance between non-identical objects must be greater than zero.

Interestingly, some distance coefficients are complementary to an association coefficient. Based on the coefficient value, subtraction from unity can be performed to convert between association coefficients to distance coefficients. An example of a coefficient complementary to the Tanimoto coefficient is the *Soergel distance coefficient*. In the case of bit vectors, the Soergel distance coefficient is one minus the Tanimoto coefficient (Cheng et al., 1996).

2.2.3.3 Correlation Coefficient

The *Correlation coefficient* aims to identify the correlation between the sets of values characterising each of a pair of molecules. It calculates the degree of correlation in terms of the proportionality and independence between the sets of values used to describe the pair of compounds.

There are many types of correlation coefficient; the *Pearson correlation coefficient* is probably the least biased for dissimilarity analysis (Maldonado et al., 2006). The binary variant of the Pearson correlation coefficient is defined by Eq. (10):

$$P_c(A, B) = \frac{nc - ab}{\sqrt{ nab(n - b)(n - a) }} \quad (10)$$

where a is the number of set bits in fingerprint A (target compound), b is the number of set bits in fingerprint B (compared compound) and n is the total bit length. The values for correlation coefficient range between -1 to +1. Results of the coefficient calculation determine (i) -1, anti-correlated; (ii) 0, no correlation;

or (iii) +1, perfectly correlated, between the database compound and the target structure. Like the other coefficients, the value of attributes may also rescale into the range of 0 to 1. More examples of common correlation coefficients used for binary variables in chemoinformatics are listed in Table 2-1.

.

Table 2-1 Common Binary Similarity Coefficient (Holliday et al., 2003; Leach & Gillet, 2007)

Type	Name	Equation	Value range
Association Coefficient	Jaccard/ Tanimoto	$\frac{c}{a + b - c}$	0 to +1
	Cosine	$\frac{c}{\sqrt{ab}}$	0 to +1
	Dice	$\frac{2c}{a + b}$	0 to +1
	Russell/ Rao	$\frac{c}{n}$	0 to 1
	Forbes	$\frac{cn}{ab}$	0 to ∞
	Simpson	$\frac{c}{\min(a, b)}$	0 to 1
Distance Coefficient	Euclidean	$\sqrt{a + b - 2c}$	n to 0
	Soergel	$\frac{a + b - 2c}{a + b - c}$	1 to 0
	Hamming/ Manhattan/ City-Block	$a + b - 2c$	n to 0
Correlation Coefficient	Pearson	$\frac{nc - ab}{\sqrt{ nab(n - b)(n - a) }}$	-1 to 1
	Yule	$\frac{nc - ab}{cd + (a - c)(b - c)}$	-1 to 1
	Dennis	$\frac{nc - ab}{\sqrt{ nab }}$	0 to ∞

* The definitions apply to the combination of bit-string of length n where a is the number of set bits in A (target string), b is the number of set bits in B (compared string), c is the number of set bits common to both strings and d is the number of set bits in neither string.

2.3 Similarity Searching Application

Two other important applications that are developed from the molecular similarity approach and widely implemented in the chemoinformatics domain, are (i) clustering and (ii) molecular diversity. Both applications are described in the next section.

2.3.1 Clustering

Clustering provides a simple and effective overview of the range of structural types in a molecular database. It helps to save cost and rationalise the basis for molecular biological testing (Willett, 2011). A representative molecule of a cluster is selected for the biological testing. If the representative proves to be bioactive, then the other molecules in the same cluster will be tested. But if the representative is not bioactive, then the other molecules in the same cluster will be disregarded from the biological testing.

In chemoinformatics, clustering is used as a tool for molecular database analysis. It aims to identify clusters of molecules that exhibit strong intra-cluster similarities as well as strong inter-cluster dissimilarities (Willett, 2014). The review by Downs and Barnard offers a comprehensive introduction to clustering methods in the chemoinformatics context (Downs & Barnard, 2002). Many comparative studies have been conducted on the performance of different clustering methods when applied to chemoinformatics datasets, with the first undertaken by Willett (1987). Clustering is also widely implemented as a multivariate statistical analysis tool in other domains (Di Giuseppe et al., 2014).

For each compound in the dataset, the clustering process for compound selection includes: (i) generation of descriptors, (ii) calculation of similarity or distance, (iii) compound clustering using a cluster algorithm and (iv) selection of one compound from each cluster as a representative of the subset (Leach & Gillet, 2007). There are various methods available for molecular clustering, which groups compounds by means of distances in the descriptor or fingerprint

space. The methods can be classified into (i) hierarchical clustering or (ii) non-hierarchical clustering methods.

In the *hierarchical clustering* methods, each molecule (or cluster of molecules) merges with other similar molecules resulting in a cluster of two molecules or clusters of molecules. There are two types of this clustering, which are *agglomerative (bottom-up)* and *divisive (top-down)*. *Ward's* method is one of the best-known hierarchical agglomerative clustering methods (Bajorath, 2001). Although it is widely implemented in chemical database clustering, *Ward's* method consumes more computational resources as compared to the non-hierarchical clustering methods described below.

The *non-hierarchical clustering* method is another approach, *K-means* method is one of the examples for a non-hierarchical clustering method. In the *K-means* clustering algorithm, the number of clusters is denoted by the value of '*k*'. First, the '*k*' points are selected at random. The remaining molecules are assigned to the nearest '*k*' point. This will give the initial sets of '*k*' clusters. Then, the method calculates the *centroid* for each cluster. Each molecule is reassigned to the nearest centroid. The centroids are then recalculated for relocation and the procedure repeated until a cluster condition is satisfied (Leach & Gillet, 2007). The advantage of this method is the ability to process large databases with low computational demand.

Recent reviews from MacCuish and MacCuish (2014) suggested a few potential research areas for molecular clustering, which include bi-clustering for feature selection and polypharmacology as well as determining SAR clusters. The bi-clustering algorithm is commonly used in gene expression and bioinformatics applications. It uses a dataset to generate sets of: (i) samples and (ii) features. Bi-clustering provides better data representation and allows the molecular similarity based on subset of attributes.

2.3.2 Molecular Diversity

Molecular diversity is a technique used to maximize the diversity of the molecules for biological testing. This technique selects the diverse compounds

by calculating the (dis)similarities between pairs of molecules in the dataset. A diverse subset of molecules in a dataset is selected by considering their inter-molecular structural similarities (Willett, 2005).

The *cluster-based selection* method is a typical approach for selecting adverse subset together with a few others, which are: (i) partition-based selection, (ii) dissimilarity-based selection and (iii) optimisation-based selection (Maldonado et al., 2006).

The *partition-based selection* method matches and assigns each molecule into a partition that was created based on a defined set of molecular properties, in which a compound representative is selected from each partition. This method can be used to find the difference between databases, but is limited to low dimensional datasets.

The *dissimilarity-based selection* method chooses the most dissimilar molecule from the earlier molecule selected. This approach results in a subset that contains most diverse molecules. The *optimisation-based selection* method, on the other hand, predefines the diversity measurement based on optimisation procedure. The key importance of the optimisation procedure relies on a diversity function, in which the MaxMin maximum-dissimilarity algorithm was identified by Snarey et al. (1997), as the most effective algorithm based on its operation and ability to process very large datasets.

2.4 Evaluation Measurement

An important criterion of any similarity searching application is the ability to retrieve a significantly higher number of active compounds than if selected at random. The measurement of this criterion can be evaluated using various methods that are available, e.g., Enrichment Factor (EF), Receiver Operator Characteristic (ROC), Robust Initial Enhancement (RIE) and the Boltzmann-Enhanced Discrimination of ROC (BEDROC).

The *enrichment factor (EF)* is one of the common evaluation methods used in virtual screening application because of its simple calculation and

straightforward interpretation (Kirchmair et al., 2009). It measures the active compounds retrieved compared to active compounds from random selection. The calculation of the EF is defined in Eq. (11):

$$EF = \frac{AR}{R} \quad (11)$$

where AR is the number of active compounds retrieved, and R is the number of actives expected based on random selection, for a given cut off value. The typical cut off values for this method are 1% and 5% (Geppert et al., 2010).

The *receiver operator characteristic (ROC)* is a widely used method for evaluation in machine learning applications (Witten & Frank, 2000). It generates a detection rate between hit rate and false rate by plotting the percentage of the total number of true positives as the vertical axis (i.e., active compounds retrieved) against the percentage of total number of false positives as the horizontal axis (i.e., inactive compounds retrieved) (Witten & Frank, 2000). The calculation for the percentage of true positives is defined by Eq. (12):

$$\text{Percentage of true positive rate} = \frac{TP}{(TP + FN)} \times 100\% \quad (12)$$

where TP is the number of true positives and FN is the number of false negatives (i.e., active compounds that are not retrieved). The calculation for the percentage of false positives is defined by Eq. (13):

$$\text{Percentage of false positive rate} = \frac{FP}{(FP + TN)} \times 100\% \quad (13)$$

where FP is the number of false positives and TN is the number of true negatives (i.e., inactive compounds that are not retrieved).

The *robust initial enhancement (RIE)* is another evaluation method that was developed to discriminate 'early recognition' in the correct order, i.e., rank

actives early in an ordered list (Sheridan et al., 2001). This method uses a continuously decreasing exponential weight as a function of rank that places heavier weight on early ranked actives. The calculation of the RIE is defined in Eq. (14):

$$RIE = \frac{\frac{1}{n} \sum_{i=1}^n e^{-\alpha x_i}}{\frac{1}{N} \left[\frac{1 - e^{-\alpha}}{\frac{\alpha}{e^{\frac{\alpha}{N}} - 1}} \right]} \quad (14)$$

where $x_i = \frac{r_i}{N}$ is the relative rank of the i th active and α is a tuning parameter (Zhao et al., 2009).

However, this method is dependent on the exponential weight and ratio of actives to inactives (Riniker & Landrum, 2013). Thus, the *Boltzmann-enhanced discrimination of ROC (BEDROC)* method of evaluation is derived to avoid the dependency on the ratio of actives to inactives by forcing the RIE to be bounded by 0 and 1 (Truchon & Bayly, 2007). The calculation of the BEDROC is defined by Eq. (15):

$$BEDROC = RIE \times \frac{\frac{1}{N} \sinh(\alpha/2)}{\cosh(\alpha/2) - \cosh\left(\alpha/2 - \alpha \frac{n}{N}\right)} + \frac{1}{1 - e^{\alpha\left(\frac{N-n}{N}\right)}} \quad (15)$$

The focus of this research evaluation is to identify the number of actives retrieved from the similarity searching application rather than identifying the ranking order of the actives retrieved. Thus, the enrichment factor was chosen to evaluate the effectiveness of the proposed research method in this thesis's subsequent chapters.

2.5 Conclusion

This chapter has introduced the key components, methods, applications and evaluation measurements for molecular similarity in virtual screening. It has

shown that molecular representation and descriptor, weighting scheme and similarity coefficients are the main components of any similarity searching system. The literature showed that the effectiveness of a similarity search relies on the components, which many reported as the similarity coefficients. This can be seen from previous comparative studies mentioned in Section 2.2.3. Taken together, these key components are implemented as a basis to any similarity search applications.

Chapter 3 Nearest Neighbour Searching in High Dimensionality

3.1 Introduction

The main task of an information retrieval application is to use a dataset to search for relevant information. The objects in the dataset are usually represented by a large number of variables, i.e., high dimensionality in the variable space. Nearest neighbour searching is one of the applications that involves searching for data in high dimensional datasets (Clarke et al., 2008; Willett et al., 1998). However, the effects of performance in high dimensional datasets have become an issue for many years.

This chapter intends to describe the concepts and issues of nearest neighbour searching in high dimensionality datasets. These include the possible methods and solutions that can be applied in chemoinformatics applications. The overall structure of this chapter takes the form of three sections. It starts with the introduction to issues in high dimensionality datasets, followed by the review of previous research on the effectiveness of nearest neighbour search in high dimensionality. The final section introduces and discusses several approaches for nearest neighbour search in high dimensionality. This chapter also provides important insights for the methodology of this research investigation.

3.2 Issues with High Dimensionality Data

Dimensionality refers to the number of variables used to characterise the objects in a dataset (Leach et al., 2007). High dimensionality involves the use of a large number of variables to represent a dataset. Chemoinformatics datasets are also known for their representation using high dimensionality descriptors (Todeschini & Consonni, 2000). These descriptors describe the characteristics of a molecular compound in many aspects as discussed in Chapter 2.

Despite the ability to describe data in various ways, there are several issues inherent in high dimensionality analysis. One is the “*curse of dimensionality*”, introduced in the 1960s (Bellman, 1961). The curse of dimensionality is a

phenomenon that arose during the analysis of data in high dimensional space. In this phenomenon, the degree of compactness of a dataset becomes sparser as the dimensionality of the dataset increases.

The phenomenon is often interpreted to cause the decrease in the performance of high dimensionality applications. Figure 3-1 illustrates an example of a variation of performance level for an application using n dimensional features. The performance increases up to the dimension of m . The performance starts decreasing with each continuous increment of dimension to n . Here, the optimal performance of the application is produced when the dimension of features is equal to m .

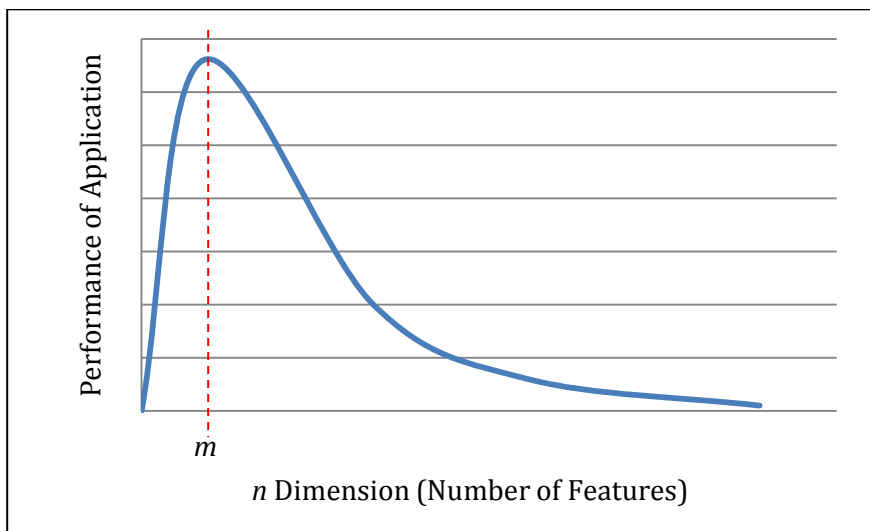


Figure 3-1 Effect of the Curse of Dimensionality Phenomenon

Clarke et al. (2008) discuss several properties of high dimensional data space in the context of gene data. Among the properties are: (i) the performance of several statistical learning techniques degrades as the dimensionality increases and (ii) the scalability of distance measures in Euclidean space is generally poor when the dimensionality is increased.

The effect of dimensionality on the nearest neighbour search was investigated by Beyer et al. (1999). The study proved that as the dimensionality increases,

the difference of the distances between the nearest and the furthest neighbours to the query object becomes insignificant, while the variance of the distance distributions converges to zero. The experimental results showed that the nearest neighbour search becomes meaningless with as few as 10 to 20 dimensions when tested on a synthetic dataset of one million data points.

The importance of determining the nearest neighbour is illustrated in the following figures. In Figure 3-2, the nearest neighbour point to the query point can be identified more clearly compared to the scenario in Figure 3-3. Although the nearest neighbour point in Figure 3-3 is well-identified based on the location of the circle, the difference between the distance of the nearest neighbour and the distances of the remaining points in the dataset to the query point is so small. Hence, this scenario affects the confidence level when determining the nearest neighbour of a query point.

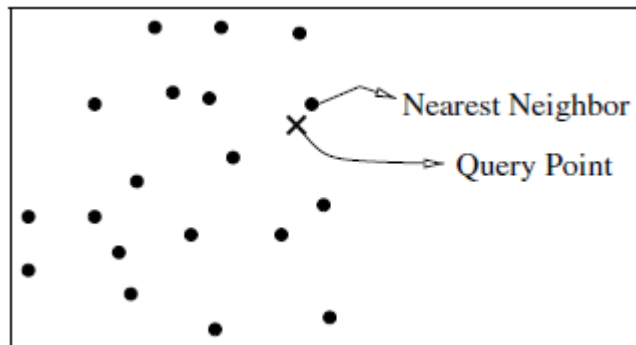


Figure 3-2 Query point and its nearest neighbour (from Beyer et al. 1999)

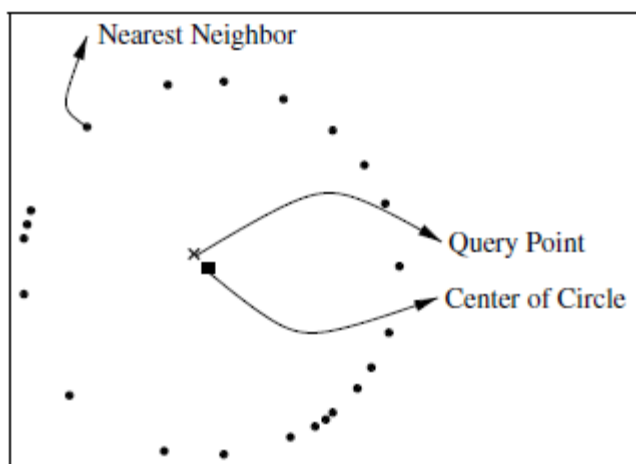


Figure 3-3 Another query point and its nearest neighbour (from Beyer et al. 1999)

The sparse sampling in high dimensions also creates the “*empty space phenomenon*”, that is, the density of data in a compartment of space decreases during a partition dimension (Rupp et al., 2009). The partition dimension divides each dimension into two compartments. In this process, the number of compartments increases exponentially as the dimensionality increases. It is important that each compartment contain at least one data point. Thus, a calculation of the maximum covered dimension can be used to estimate the *maximum number of dimensions* in a dataset. This is to ensure that each compartment has a minimum of one data point. The calculation can be defined by Eq. (16):

$$d_{max} = \lceil \log_2(n) \rceil \quad (16)$$

where d is the dimensionality of the compound descriptor and n is the size of dataset. Rupp et al. (2009) uses an example of a common molecule dataset, which contains $10^8 = 100,000,000$ molecules. The above equation is used for the calculation. The maximum number of dimensions is calculated to be 26 dimensions.

The above calculation is a general estimation that does not consider the distribution of the dataset. However, the estimation of a maximum number of

dimensions for the dataset that has an independent and uniform distribution of data can be measured differently. It is defined as the probability that at least one compartment is shared by two or more molecules. The probability can be calculated by Eq. (17):

$$P(d) = 1 - \binom{m}{n} \frac{n!}{m^n} \quad (17)$$

where d is the dimensionality of the compound descriptor, n is the size of dataset and $m = 2^d$.

Regardless of various problems in high dimensionality, the increased size of data and improvements in methods and software have generated many interesting high dimensionality studies in a number of domains (Mikolajczyk et al., 2005; Palmer et al., 2013; (Audain, Sanchez, Vizcaíno, & Perez-Riverol, 2014). In particular, a study by Godden and Bajorath (2006) supports the success of virtual screening methods in extremely high dimensionality chemical representations. The study investigated molecular similarity using a simple distance approach. The experiment selects a *centre* of a group of compounds with similar activity in high dimensional space. Euclidean distances were calculated between each compound in the dataset to the centre. This produces a distance-based ranking, indicating the molecular similarity ranking. A set of 123 descriptors was used in this experiment containing 1D, 2D and 3D descriptors. These descriptors were generated from the compounds in the Molecular Drug Data Report (MDDR) dataset. The result showed that this method successfully ranked compounds according to the biological activity in high dimensional space.

3.3 Effectiveness of Nearest Neighbour Search in High Dimensionality Data

A nearest neighbour search in high dimensional data aims to find the closest match to the query object in multivariable datasets. The curse of dimensionality affects nearest neighbour search in many applications. When dimensionality

increases, the nearest neighbour search tends to be meaningless when, among others, the data space is sparse, i.e., scattered (Weber et al., 1998; Hinneburg et al., 2000). As a result, the difference between the distances of nearest and farthest points to the query object in high dimensional space approximates to zero (Beyer et al., 1999). This section reviews previous studies related to the effectiveness of nearest neighbour search in high dimensional datasets. It identifies existing approaches and effectiveness criteria, which are implemented in the search.

3.3.1 Distance Measure Approach

Aggarwal et al. (2001) analysed the general behaviour and effects of using various distance metrics on the nearest neighbour searching in high dimensional data mining datasets. The investigation was conducted using different L_k distance metrics: fraction ($k < 1$), Manhattan ($k = 1$) and Euclidean ($k = 2$) on a uniformly distributed dataset. The effectiveness criterion measured for this experiment is the ratio of distance between the nearest and farthest neighbours. The higher ratio indicates higher effectiveness of the nearest neighbour search. The results of the above study showed that the fraction distance metric provides the highest effectiveness. This was followed by the Manhattan and Euclidean distance metrics.

In a more recent study, France et al. (2012) further investigated the effectiveness of nearest neighbour recovery on clustering of high dimensional document datasets. The study was conducted using the Euclidean and Manhattan distance functions. Additional metrics such as cosine and correlation distance metrics were also used as similarity measures. The effectiveness criterion measured for this experiment is the number of nearest neighbours found.

Similar to Aggarwal et al. (2001), the results showed that the Manhattan distance metric resulted in the highest effectiveness of nearest neighbour search. A comparison was also made between the correlation and cosine metrics. It was found that the correlation metric produced better results than

the cosine metric on the nearest neighbour search. The above study also recommended data standardisation to enhance the effectiveness of neighbourhood classification.

3.3.2 Approximate Nearest Neighbour Approach

Another approach to the nearest neighbour search in high dimensionality datasets is based on *approximate nearest neighbour*. This approach may return near optimal nearest neighbour but is more efficient than linear search in high dimensionality (Muja & Lowe, 2009).

Indyk and Motwani (1998) introduced an approximate nearest neighbour search based on a hashing technique called the *locality-sensitive hashing (LSH)* method. This is followed by an improved LSH method on the execution time by Gionis et al. (1999). This approach uses a hash function in order to identify the nearest object to the query objects. The objects in a dataset are hashed into hash values and mapped into hash tables. The closest object to the query is identified based on the probability of their collision in the table entry, i.e., *bucket*. The experiment conducted by Gionis et al. (1999) on an image dataset showed that the method performed well even with more than 50 dimensions.

A series of investigations have been conducted by Muja and Lowe (2009; 2014) on several tree-based algorithms for approximate nearest neighbour search in high dimensionality. These include *multiple randomized kd-tree* and *hierarchical k-means tree* algorithms, which are different based on the way that the search region is constructed. Multiple randomized kd-tree splits data on the dimension randomly from the first D dimensions, which contains data with the greatest variance. Hierarchical k-means tree splits the objects recursively using k-means clustering. The nearest neighbour searches are then performed within the regions that have been constructed.

The experiments conducted on real-world image datasets by Muja and Lowe (2009; 2014) were evaluated based on: (1) the precision of the search, i.e., the percentage of *exact* nearest neighbours returned by the approximate method and (2) the performance, i.e., the search time over linear search time. The

performances and precisions (i.e., 81% and 85%) of the nearest neighbour searches have been found to increase for as high as 4,096 dimensions.

The above studies highlight existing approaches, i.e., distance measurements and approximation nearest neighbour approaches. They have been used to investigate the effectiveness of different nearest neighbour searches in high dimensionality datasets of different domains. They also indicate a few effectiveness criteria used to measure the effectiveness of the nearest neighbour search. The following section introduces an approach, which involves the dimensional reduction of high dimensionality datasets.

3.4 Dimensionality Reduction Approach

The issues of high dimensional data decrease the performance of any data analysis, e.g., the nearest neighbour search. One of the solutions reviewed by Clarke et al. (2008) is to reduce the original set of variables into a new set of uncorrelated variables using dimensional reduction methods. The purpose of these methods is to reduce the high dimensional variables into a lower number of dimensional variables. These contain the most meaningful information to describe the pattern of the datasets and for better data interpretation (Howe et al., 2007).

Dimensional reduction methods have been widely implemented in many areas, including image and text analysis (Bingham et al., 2001). Fodor (2002) reviewed the state-of-the-art for dimensional reduction in statistics, signal processing and machine learning areas. There are two main categories of dimensional reduction methods: (i) feature selection methods and (ii) projective methods.

3.4.1 Feature Selection Method

The *feature selection method* is an approach that reduces the feature dimensionality. These new, reduced features preserve the meanings of the features. It selects the most relevant features or subset of features from original high dimensional features. Advantages of this approach include (i) facilitating data visualization and understanding, and (ii) defying the curse of

dimensionality to improve prediction performance. Guyon and Elisseeff (2003) discussed several methods for feature selection. These include the variable ranking and variable subset selection.

3.4.1.1 Variable Ranking Method

The *variable ranking method* implements a ranking criterion. It measures the goodness of linear fit of individual features and then results in a ranking of features. An example of the variable ranking criterion is the coefficient of determination, R^2 , which indicates the fraction of variance explained by individual features. One of the advantages of this method is that it is computationally efficient as it only requires computation and the sorting of ranking scores.

3.4.1.2 Variable Subset Selection Method

The *variable subset selection method* includes a “wrapper” methodology. This uses the prediction performance of a given learning machine to assess the relative usefulness of subsets of variables. This methodology may include the following steps:

Step 1 : Select a subset of features;

Step 2 : Evaluate the performance for the selected subset using an objective function;

Step 3 : Repeat Steps 1 & 2 until predefined termination condition is met;

Step 4 : Return the subset that yields the best performance.

One of the limitations of this method is that it is intensive in computation. Several strategies have been implemented to overcome this limitation. One example is a *backward elimination approach*. In 2013, Vogt and Bajorath (2013) implemented this strategy for the variable subset selection in fingerprint similarity searching. It begins by selecting all features and then evaluates performance of the application. The process is repeated after each feature is

being individually removed. This implementation produces a subset of reduced fingerprint representation, which is able to increase the performance of the similarity searching.

3.4.2 Projective Method

The *projective method*, on the other hand, reduces the dimensionality by combining features of all variables. There are two types of combinations: a linear or non-linear combination. *Linear combination methods* use the least-square regression line in their computation. The linear fit minimises the sum of squares of the measured data. *Non-linear combination methods*, on the other hand, use the properties of data. It reproduces the distances of high dimension variables in the low dimension variables (Maaten et al., 2009). Linear combination methods are more attractive compared to non-linear combination methods. This is because they are simple in computation and analytically tractable.

3.4.2.1 Linear Dimensional Reduction Approach

A common method of *linear dimensional reduction* is the *Principal Component Analysis (PCA)*. It is the most widely used linear dimensional reduction method and is considered the most effective in its group because of its ability to reduce *mean-square error*, i.e., the difference of squared error loss (Fodor, 2002).

PCA aims to seek a projection that preserves as much of the data information as possible. It measures the multidimensional data and reduces it to lower dimensions. The aim is to remove the correlations between descriptors (Bayada et al., 1999). This method also reveals the correlations and relationships between data, thus providing easier interpretations (Akella & DeCaprio, 2010).

PCA uses a covariance matrix of the multivariable descriptors to compute the orthogonal projections (principal components) with the least squared error. If dimensional reduction is needed, the original data is projected into the perpendicular lines (the reduced dimensions). This results in a set of data with

the highest variance (Wold et al., 1987). The computation of principal components involves a few steps (Smith, 2002; Andrew, 2015; Nimrod, 2014):

Step 1 : Computation of the *covariance matrix*

Step 2 : Computation of the *eigenvectors* and selection of reduced number of dimensions

The eigenvectors are the uncorrelated *linear combinations* and are referred to as the *principal components*. These are derived from the original variables in decreasing order of importance. The eigenvalues are the *variances* of each eigenvector to each variable. For n variables, as many as n eigenvectors can be computed from the $n \times n$ matrix of variables. The calculation of eigenvectors is defined by Eq. (18):

$$PC_i = \sum_{j=1}^n c_{i,j} x_j \quad (18)$$

where PC_i is the i th eigenvector, $c_{i,j}$ is the covariance matrix and x_j is the eigenvalue for n variables.

The first principal component, PC_1 , maximises the variance in the data. It is represented by the largest eigenvalue. The second principal component, PC_2 is orthogonal to the first. It contains as much of the remaining variance as possible (i.e., second largest eigenvalue). This is followed by the rest of the principal components, which are ordered in decreasing eigenvalues (Leach & Gillet, 2007).

The eigenvalues are useful in determining the selection of the reduced number of principal components, k . It is based on the percentage of variance retained from the data. This is typically represented by the value of above 90%. It also indicates how well the reduced dimensions, k , approximate the original dataset.

Step 3 : Projection of original data into the reduced dimensions

The reduced k dimensions form a matrix with the k eigenvectors in the rows and the variables in the columns. The eigenvectors are arranged in a descending order of corresponding eigenvalues. The original data, which is also represented by a matrix, has the original variables in the row and the data points in the column.

The multiplication of both matrices produces the projection of the original data into the reduced dimensions. It represents the final data in a matrix that has the eigenvectors in the rows and the data points in the columns. The projection is defined by Eq. (19):

$$\widehat{D}_{i,j} = PC_{i,k} \times D_{k,j} \quad (19)$$

where $\widehat{D}_{i,j}$ is the final data, which is projected by the principal components, $PC_{i,k}$ is the reduced dimensions and $D_{k,j}$ is the original data.

The calculation of principal components requires the variable's standardization to have a mean of zero and standard deviation of one. This is because the result of variance depends on the scale of the variable. Thus, it is important to have an equal contribution between the variables. Figure 3-4 illustrates an example of the projection from high dimensions (3 dimensions) to low dimensions (2 dimensions) using PCA (Matthias, 2014).

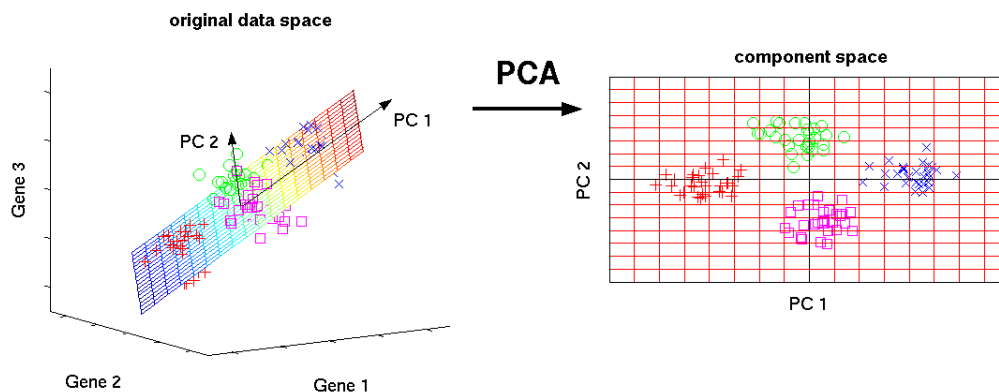


Figure 3-4 Example of the projection from high dimensional to low dimensional variables using principal component analysis

PCA has been used in cheminformatics applications for molecular descriptor reduction. Bayada et al. (1999) implemented the PCA as a method to remove the descriptor's correlations in clustering analysis. Ten principal components that represent 87% of the variance from a diverse database, i.e., the Available Chemicals Directory (ACD) database, were identified from 86 descriptors. The combinations of the ten principal components were used as a new set of descriptors for each compound. The compounds were then clustered using several clustering methods. The result using Ward's algorithm and ten principal components was more effective in separating biological activities than random selection.

The Bajorath group have implemented PCA for the reduction and combination of both molecular descriptors and binary fingerprints (Xue et al., 1999a; Xue et al., 1999b; Xue & Bajorath, 2000). However, there are more effective methods of molecular fingerprint reduction. These have been described elsewhere (Baldi et al., 2007; Swamidass & Baldi, 2007; Geppert et al., 2010). For the purpose of this research, these methods will be introduced and discussed in Section 3.5.3.

Linear discriminant analysis (LDA) is another example of a linear combination method. This method aims to seek projections of low dimensionality. This low dimensionality preserves as much of the class discriminatory information that best separates the data. The result achieves maximum data discrimination by maximizing the ratio between class distances to the within-class distances (Balakrishnama & Ganapathiraju, 1998).

In comparison to PCA, LDA results in the direction that maximizes the difference between two classes, which is more applicable for data classification. PCA on the other hand, results in the direction that maximizes the variance in the data and generates new variables that represent maximum variance in the dataset.

3.4.2.2 Non-Linear Dimensional Reduction Approach

An example of a common method in drug discovery for *non-linear reduction* is *Multidimensional Scaling (MDS)* (Xu et al., 2002). This method aims to model the dissimilarity and similarity relationships between two sets of variables by

rescaling the distance. It reproduces approximate distances between original high dimension and new generated low dimension, by (i) generating projection of low dimension coordinates, and then (ii) modifying distance between the original and projected coordinates for optimization (Leach et al., 2007). The key component for the reproduction of the distance is the optimization procedure, using a stress function, e.g., Kruskal (1964). The stress function is a sum-of-squares error function. It measures the degree of correspondence between the original and the projected coordinates. The output of a stress function must not exceed a threshold value to ensure the optimisation.

Another example of a non-linear method is *Locally Linear Embedding (LLE)*. It transforms high dimensional data to a low dimension, while retaining the surrounding neighbourhood. One of the advantages of this method is that it preserves the neighbourhood mapping. It provides the underlying structure identification, i.e., the small scale resembles a Euclidean space of data in a specific dimension (Roweis and Saul, 2000).

3.4.3 Binary Fingerprint Dimensional Reduction Approach

The molecular fingerprint has been the most effective molecular representation for many chemoinformatics applications as noted in Section 2.2.1. Molecular fingerprints are typically represented by a very long binary bit length, i.e., 512 or 1024 bits. These indicate the fingerprint's dimensions. Several methods have been introduced to reduce the dimensions of a molecular fingerprint. Geppert et al. (2010) described several methods, which include folding, hashing (James et al., 1995) as well as reduction based on a statistical fingerprint model (Baldi et al., 2007).

One of the most common methods of binary fingerprint reduction is *folding*. This method takes the original number of fingerprint bits and folds it to a reduced number, using the *modulo operator*. Let F be the original number of fingerprint bits, \hat{F} is the reduced number of fingerprint bits and N is the length of bits in the reduced fingerprint. A bit in \hat{F} with index, j , is set to 1 if there is at least one bit of F_i set to 1, where $F_i \bmod N$ is equal to index j .

Figure 3-5 illustrates the binary fingerprint folding steps. The original fingerprint bits F_i , which has a size of 16-bits, is reduced to fingerprint bits \hat{F}_j , where $N=4$ is the length of \hat{F}_j . The bit position of \hat{F}_1 and \hat{F}_3 are set to 1 because there are bits in F_i that are set to 1, when $F_i \bmod N$ is equal to index \hat{F}_1 and \hat{F}_3 .

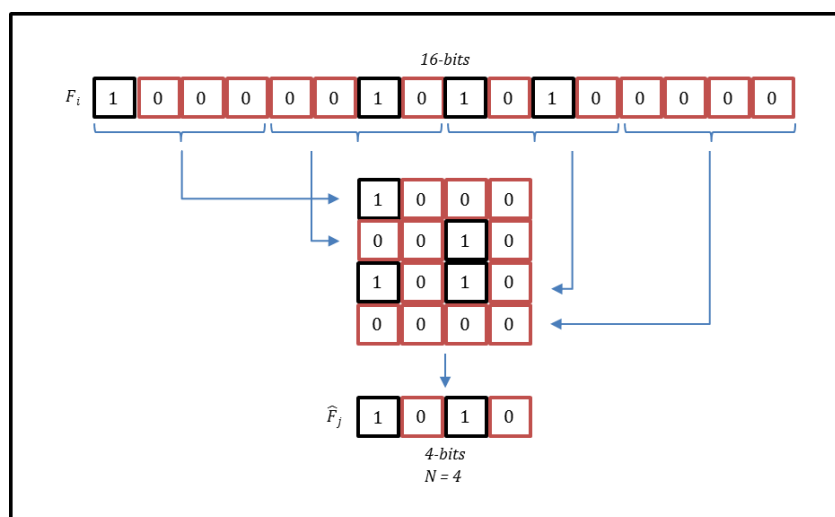


Figure 3-5 Binary Fingerprint Folding Steps

The reduced binary fingerprints can provide a rapid search in the chemical database. However, one limitation of this method is that it ignores the weighted information of the bits. This, however, can be solved by bit rearrangement, using a *hashing algorithm* or *random permutation*. Nevertheless, this method is the most effective for an application that treats all bits equally, e.g., the specific ordering of the bits is not important (Swamidass & Baldi, 2007).

Bit dependency is one of the reasons for bit fingerprint reduction. This is because, the dependant bits can affect the similarity measurement. The bit dependencies are the universal presences of a bit given the presence of another. Chen and Golovlev (2013) analysed the bit dependencies of 881 bits structural keys from PubChem dataset. The study showed a method to identify and eliminate the dependant bits.

First, the frequency of occurrence of each bit was tabulated from the matrix of bit values. The number of bits that were set for each compound was also noted. Next, to identify the dependencies, each of the bit positions (A) was selected and checked against all other bit positions (B). Positions (B) in which bits were set when set in the selected bit position (A) were noted. Thus, each bit position is not dependent upon itself. The two way dependencies were identified by examining all pairs of bit positions. The pairs which bits were always identically set are the two way dependencies bits. The dependent bits within the structural keys were stripped. The number for set independent bits for each compound was then recorded.

Similarity searching using the Tanimoto similarity measure was then experimented on both the complete 881 keys and the subset of 160 non-dependant bits. The results showed that the similarity search using the set of non-dependant bits affect the similarity scores. It returns a large numbers of nearly identical compounds. However, this does not mean that the non-dependant set is better because the similarity searches resulted in different compounds as compared to the similarity searches using the complete keys. Further analysis on the non-dependant bits based on bit occurrence frequencies showed that a non-dependant bit can also be the most common bit and often encodes features similar to the dependant bits.

3.5 Conclusion

This chapter focused on the concept of nearest neighbour search in high dimensional datasets. It was seen that high dimensional datasets cause difficulties in data interpretation and visualization. This is because, as the dimension of data increases, the density of data decreases. As a result, this phenomenon degrades the performance of nearest neighbour search applications.

The third section of this chapter reviewed a number of studies conducted to identify the effects of the nearest neighbour search as dimensionality increased.

One of the solutions for high dimensionality datasets is to reduce highly dimensional descriptors into a lower number of dimensions.

The appropriate dimensional reduction methods are discussed in the final section of this chapter. This study will evaluate the effect of changing the dimensions of molecular representations on the effectiveness of nearest neighbour searching. Thus, the methods introduced in this chapter provide ideas on how to reduce the molecular representations and descriptors. They can also be used for the binary and non-binary data representation, which are the common molecular representations in chemoinformatics applications.

Chapter 4 Methodology

4.1 Introduction

This chapter will outline the experimental design used for the three investigations reported in this thesis. The three investigations are: 1. the effects of dimensionality on the effectiveness of similarity searching (reported in Chapter 5); 2. the effects of dimensionality on the effectiveness of clustering (reported in Chapter 6); 3. the relative importance of the fingerprint and the similarity coefficient components on the effectiveness of similarity searching using cross-classified multilevel model analysis (reported in Chapter 7).

This chapter provides the details of methodology which are common to all three chapters mentioned above in terms of the databases, molecular representations, and similarity (and distance) coefficients. All evaluation methods will be introduced in this chapter together with the statistical methods.

4.2 Dataset

Three chemical datasets have been used within the investigations, i.e., the MDL Drug Data Report (MDDR) (*MDL Drug Data Report*, 2005), the WORld of Molecular BioAcTivity (WOMBAT) (“World of Molecular Bioactivity,” 2011) and the ChEMBL dataset (Gaulton et al., 2012). These datasets are commonly used within the chemoinformatics research group at the University of Sheffield.

Each dataset is described separately in the subsections below. Each description also includes a table that contains information about: (i) the activity class with its abbreviation, (ii) the number of active molecules in each activity class, (iii) the number of distinct scaffolds present in the class and (iv) the value of mean pairwise similarity (MPS). The distinct scaffolds describe the core structure that is the central component of a molecule. This is a substantial substructure that contains the important molecular material to ensure that the functional groups are in a desired geometric arrangement and therefore produce similar biological properties. This study used the definition of scaffold by Bemis and Murcko (1996). The MPS value describes the diversity of each activity class in a

dataset. It is measured based on the inter-molecular similarities using the standard UNITY 2D fingerprints and the Tanimoto coefficient. The mean intra-set similarity is then calculated and noted. A higher MPS value means higher inter-molecular similarity and vice versa.

4.2.1 MDDR

The MDDR dataset is a commercial dataset produced by BIOVIA and Thomson Reuters (“BIOVIA Datasets | Sourcing Datasets: BIOVIA Available Chemicals Directory (ACD),” n.d.). The dataset contains molecules compiled from resources such as patent literature, journals, meetings and congresses. The activity data is qualitative, i.e., a molecule is active if it is known to exhibit a specific activity and assumed to be inactive if no activity has been reported.

The MDDR dataset utilised in this study was the version from 1995, which contained 102,540 molecules and 11 activity classes. It was used in the previous studies by Todeschini et al. (2012) and Holliday et al., (2015). As shown in Table 4-1, the Renin activity class is known to be the most homogeneous (highest MPS value, i.e., 0.57), while the Cyclooxygenase activity class is the most heterogeneous in this dataset (lowest MPS value, i.e., 0.27).

The first investigation on the similarity search application in Chapter 5 uses a total of 102,540 molecules and 11 activity classes. The second investigation on the clustering application in Chapter 6 uses 10% of the molecules in the dataset that are randomly selected. This yields a dataset containing a total of 10,254 molecules. This is because the large number of pairwise distance calculations in the clustering applications demands a lot of computation. As a result, the subset of the dataset contained between 36 and 125 active molecules, depending on the activity class.

4.2.2 WOMBAT

The WOMBAT dataset is a leading small molecule chemogenomics dataset released by Sunset Molecular (“World of Molecular Bioactivity,” 2011). The dataset contains molecules extracted from important drug-discovery journals

such as the *Journal of Medicinal Chemistry* and *Bioorganic & Medicinal Chemistry*. The activity data is quantitative, e.g., a molecule is assumed to be active if an associated IC₅₀ (the half maximal inhibitory concentration) value is equal or more than a defined threshold value (or inactive if the activity value is less than the threshold value).

The WOMBAT dataset used in this study has been described and compiled by Gardiner et al. (2009). A molecule is marked to be active or inactive for a specific activity class based on the drug potency. A threshold of pIC₅₀ at 5.0 is defined. For each activity class, molecules with pIC₅₀ \geq 5.0 are marked as active for that class, and molecules with pIC₅₀ $<$ 5.0 are removed from that class. The resulting database contained a total of 138,127 molecules reduced from the original version which has 186,117 molecules by removing duplicated molecules.

There are 14 activity classes used throughout the study (Chapters 5 and 6), of which eleven classes are similar to the MDDR and three others are the additional activity classes. Like the MDDR dataset, the Renin activity class is also known to be the most homogeneous with the highest MPS value, i.e., 0.59, while the Cyclooxygenase activity class is the most heterogeneous with the lowest MPS value, i.e., 0.32 (Table 4-2).

The first investigation on the similarity search application in Chapter 5 uses a total of 138,127 molecules. For similar reason as the MDDR dataset, the second investigation on the clustering application in Chapter 6 uses 10% of the molecules in the dataset that are selected at random, yielding a dataset containing a total of 13,813 molecules. Hence, the subset of the dataset contained between 14 and 113 active molecules, depending on the activity class.

4.2.3 ChEMBL

The ChEMBL dataset is one of the largest publicly available Open Data datasets created by the European Bioinformatics Institute (EMBL-EBI). It consists of a large number of drug-like bioactive compounds compiled from the main published literature on a regular basis. The ChEMBL dataset used in this study is

ChEMBL 18, which was released on 2 April 2014 and available for download at <https://www.ebi.ac.uk/chembl/>. It contains a total of 1,352,681 molecules. For this experiment, the molecules are quantitatively selected based on three properties: (i) homo sapiens target organism; (ii) compounds with $pIC_{50} \geq 5.0$ and (iii) compounds with a confidence score equal to nine (Williams, 2014). The confidence score for the ChEMBL dataset is a score value that reflects the target type assigned to a particular assay and the assurance that the target assigned is the correct target for that assay.

The first and third investigations in Chapters 5 and 7 used only 10% from the total number of molecules in this dataset that are randomly selected for two reasons: (1) for a comparable number of compounds used for the MDDR and WOMBAT datasets and (2) to avoid intensive computation as the searches involve repetition of very highly dimensional fingerprints. The resulting database contained a total of 134,362 molecules. Similar activity classes to MDDR and WOMBAT were used including one additional activity class resulted in a total of 15 activity classes. Among the 15 activity classes, Type-1 Angiotensin II activity class is known to be the most homogeneous with the highest MPS value of 0.52, while Cyclooxygenase-1 activity class is the most heterogeneous with lowest MPS value of 0.28 (Table 4-3).

Table 4-1 MDDR dataset with 11 activity classes

No.	Activity Class (with abbreviations)	Number of Active Molecules	Number of Scaffolds	MPS
1	5HT3 Antagonists (5HT3)	752	417	0.35
2	5HT1A Agonists (5HT1A)	827	450	0.34
3	5HT Reuptake Inhibitors (5HT)	359	181	0.35
4	D2 Antagonists (D2)	395	258	0.35
5	Renin Inhibitors (Renin)	1130	554	0.57
6	Angiotensin II AT1 Antagonists (AT1)	943	464	0.40
7	Thrombin Inhibitors (Thrombin)	803	425	0.42
8	Substance P Antagonists (SubP)	1246	586	0.40
9	HIV Protease Inhibitors (HIVP)	750	461	0.45
10	Cyclooxygenase Inhibitors (COX)	636	282	0.27
11	Protein Kinase C Inhibitors (PKC)	453	171	0.32

Table 4-2 WOMBAT dataset with 14 activity classes

No.	Activity Class (with abbreviations)	Number of Active Molecules	Number of Scaffolds	MPS
1	5HT3 Antagonists (5HT3)	220	117	0.38
2	5HT1A Agonists (5HT1A)	592	224	0.40
3	Acetylcholinesterase Inhibitors (AChE)	503	220	0.37
4	D2 Antagonists (D2)	910	324	0.37
5	Renin Inhibitors (Renin)	474	253	0.59
6	Angiotensin II AT1 Antagonists (AT1)	724	253	0.44
7	Thrombin Inhibitors (Thrombin)	421	196	0.42
8	Substance P Antagonists (SubP)	558	186	0.43
9	HIV Protease Inhibitors (HIVP)	1128	473	0.44
10	Cyclooxygenase Inhibitors (COX)	965	220	0.32
11	Protein Kinase C Inhibitors (PKC)	142	31	0.57
12	Phosphodiesterase Inhibitors (PDE)	596	270	0.36
13	Matrixmetalloprotease Inhibitors (MMP1)	694	280	0.44
14	Factor Xa Inhibitors (FXA)	842	328	0.39

Table 4-3 ChEMBL dataset with 15 activity classes

No.	Activity Class (with abbreviations)	Number of Active Molecules	Number of Scaffolds	MPS
1	Serotonin 3a (5-HT3a) Receptor (5HT3)	213	90	0.35
2	Serotonin 1a (5-HT1a) Receptor (5HT1A)	1483	641	0.37
3	Serotonin Transporter (5HT)	2447	687	0.34
4	Acetylcholinesterase (AChE)	739	400	0.36
5	Dopamine D2 Receptor (D2)	1858	815	0.35
6	Renin (Renin)	982	291	0.45
7	Type-1 Angiotensin II Receptor (AT1)	106	60	0.52
8	Thrombin (Thrombin)	838	472	0.35
9	Neurokinin 1 Receptor (SubP)	847	316	0.43
10	Human Immunodeficiency Virus Type 1 Protease (HIVP)	2157	904	0.43
11	Cyclooxygenase-1 (COX)	139	63	0.28
12	Protein Kinase C Alpha (PKC)	211	76	0.42
13	Phosphodiesterase 4a (PDE)	254	100	0.31
14	Matrix Metalloproteinase-1 (MMP1)	395	157	0.40
15	Coagulation Factor X (FXA)	1502	603	0.39

4.3 Molecular Representation

The MorganR2 fingerprints (i.e., RDKit equivalent of ECFP_4-like) have been used as a molecular representation in all investigations. The fingerprints were generated using the RDKit standard Morgan fingerprints from the KNIME software (Landrum, 2016), which applies the Morgan algorithm that uses the connectivity information similar to those used for the well-known ECFP family of fingerprints. The only difference is about the atom typing definition to the ECFP fingerprints, i.e., isotope information is added and the valance-hydrogen count parameter is removed. A radius of two has been chosen when generating the Morgan fingerprints, which is similar to the ECFP_4 fingerprint found in Pipeline Pilot (Rogers & Hahn, 2010). The fingerprints were folded based on the size of the convention power of two, which is aligned to the word sizes on hardware and computer libraries.

To investigate the effect of changing the dimensionality of molecular representation in Chapters 5 and 6, a set of different fingerprint bit sizes was used. The set was prepared to avoid bit collisions, i.e., two different chemical features setting the same bit. Bit collisions can happen when folding the fingerprints to a particular size, which possibly results in a loss of information. In this study, meaningful information is important in assessing the effect of dimensionality to similarity searching. Although inevitable, the bit collisions can be reduced by increasing the number of fingerprint bit size to a larger number of bit spaces (Sastry et al., 2010).

The thirteen different folded dimensions that were generated are: 32 (2^5) bits, 64 (2^6) bits, 128 (2^7) bits, 256 (2^8) bits, 512 (2^9) bits, 1,024 (2^{10}) bits, 2,048 (2^{11}) bits, 4,096 (2^{12}) bits, 8,192 (2^{13}) bits, 16,384 (2^{14}) bits, 32,768 (2^{15}) bits, 65,536 (2^{16}) bits, 131,072 (2^{17}) bits. Throughout this thesis, the power of two convention will be used to represent the fingerprint dimensions or sizes, e.g., 2^{10} .

The third investigation in Chapter 7 also used MorganR2 fingerprints and nine other types of fingerprints in order to observe the relative importance of the

similarity search components. In total, ten different types of fingerprints have been used in the third investigation as listed in Table 4-4 Fingerprints used in this study (Riniker & Landrum, 2013; Landrum, 2016)⁴. All fingerprints were generated for a size of 1,024 (2^{10}) bits using the RDKit from the KNIME software (Landrum, 2016).

Table 4-4 Fingerprints used in this study (Riniker & Landrum, 2013; Landrum, 2016)

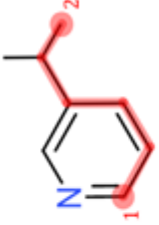
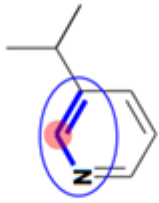
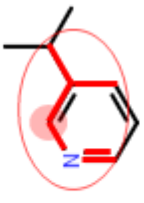
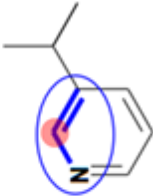
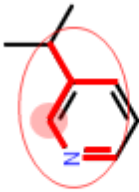
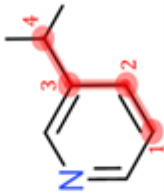
No.	Name	Abbreviation	Fingerprint	Type	Description	Example
1	AtomPair	AtomPair	Similarity	Topological	<p>Encodes (i) the atom types (i.e. the element, the number of heavy-atom neighbours and the number of π-electrons) and (ii) the distance (i.e. number of bonds) between two atoms. All possible pairs of atoms in the molecule are encoded and the distances are the number of bonds in the shortest path between each pair.</p> <p>Example:</p> <ul style="list-style-type: none"> Atom type 1: C with 2 neighbours and 1 π-electron Atom type 2: C with 1 neighbour and 0 π-electron Number of bonds: Atom type 1 is 5 bonds from atom type 2 	 <p>C, 2, 1 - 5 - C, 1, 0</p>
2	Avalon	Avalon	Substructure	Topological	Similar to Daylight fingerprints. Captures all possible pathways up to a fixed length through a molecule.	
3	FeatMorganR1	FMorganR1	Similarity	Circular	<p>Encodes the feature invariants of circular atom environments up to 1 bond radius from the central atom. Generated based on the Morgan algorithm (FCFP2-like). The feature invariants are:</p> <ul style="list-style-type: none"> the donor the acceptor the aromatic the halogen the basic the acidic 	
4	FeatMorganR2	FMorganR2	Similarity	Circular	<p>Encodes the feature invariants of circular atom environments up to 2 bond radius from the central atom. Generated based on the Morgan algorithm (FCFP4-like). The feature invariants are:</p> <ul style="list-style-type: none"> the donor the acceptor the aromatic the halogen the basic the acidic 	

Table 4-4 (continued)

No.	Name	Abbreviation	Fingerprint	Type	Description	Example
5	Layered	Layered	Substructure	Topological	An experimental substructure-matching fingerprint. A topological fingerprint that captures the number of bonds, atoms, rings and aromaticity in a molecular graph.	
6	MorganR1	MorganR1	Similarity	Circular	Encodes the connectivity invariants of circular atom environments up to 1 bond radius from the central atom. Generated based on the Morgan algorithm (ECFP2-like). The connectivity invariants are: <ul style="list-style-type: none"> • the element • the number of heavy-atom neighbours • the number of hydrogens • the isotopes • the ring information 	
7	MorganR2	MorganR2	Similarity	Circular	Encodes the connectivity invariants of circular atom environments up to 2 bond radius from the central atom. Generated based on the Morgan algorithm (ECFP4-like). The connectivity invariants are: <ul style="list-style-type: none"> • the element • the number of heavy-atom neighbours • the number of hydrogens • the isotopes • the ring information 	
8	Pattern	Pattern	Substructure	Topological	Topological fingerprint optimized for substructure screening.	
9	RDKit	RDKit	Substructure	Topological	Encodes (i) the atom types (i.e. the atomic number and the aromaticity state) and (ii) the bond types (i.e. the atom types and the bond types) of all branched and linear molecular subgraphs. Relative to Daylight fingerprint.	
10	Torsion	Torsion	Similarity	Topological	Encodes the atom types (i.e. the element, the number of heavy-atom neighbours and the number of π -electrons) of four atoms that formed a torsion. All fragments of this fingerprint contain four atoms. Example: <ul style="list-style-type: none"> • Atom type 1: C with 2 neighbours and 1 π-electron • Atom type 2: C with 2 neighbours and 1 π-electron • Atom type 3: C with 3 neighbours and 1 π-electron • Atom type 4: C with 3 neighbours and 0 π-electron 	

4.4 Similarity and Distance Measures

4.4.1 Similarity Coefficients

The similarity measures in Chapter 5 were initially calculated using all 51 similarity coefficients as previously compared by Todeschini et al. (2012). The coefficients are those suitable for the type of binary representation used in this experiment. Several of the coefficients are the most common measurements used for binary data types, e.g., the Jaccard-Tanimoto coefficient. As noted in Chapter 2, the Jaccard-Tanimoto coefficient has been the most effective measurement in binary similarity searching.

The formulation of the similarity coefficients used in this experiment may consist of the components of a , b , c , d and p . The definition of the components is based on Todeschini et al. (2012). Each component indicates:

- a = the number of common presence features between molecules x and y
- b = the number of features which molecule x has and molecule y lacks
- c = the number of features which molecule y has and molecule x lacks
- d = the number of common absence features between molecules x and y
- p = the total number of features (dimensions) that is equal to the summation of a , b , c and d

Table 4-5 provides the following information: ID, symbol, name, formula, two coefficient definitions and the metricity. The first definition was based on the symmetric and asymmetric definition of the Tversky index (Tversky, 1977). It indicates that an index (i.e., coefficient) is symmetric if $S_{xy} = S_{yx}$ and asymmetric if $S_{xy} \neq S_{yx}$. As such, the coefficients were denoted based on the formulation, i.e., symmetric if both component b and c are weighted equally, and asymmetric if not. This is because b and c represent unique features of molecules that are being compared, e.g., b is the number of unique features of molecule x and c is the number of unique features of molecule y . Thus, the

condition of $S_{xy} = S_{yx}$ will be satisfied if the coefficient considers both unique features of the compared molecules. The second definition was based on Todeschini et al. (2012). It defines a coefficient based on the formulation as: (i) symmetric if components a and d are equally considered, (ii) asymmetric if only a is considered and (iii) intermediate if both a and d are considered, but d is underweighted with respect to a . The metric properties have already been discussed in Chapter 2. The coefficient IDs in Table 4-5 will be used to refer to the similarity coefficients throughout the study in Chapters 5 and 7.

Based on the statistical test conducted in Chapter 5, 20 similarity coefficients were found to be monotonic with other coefficients. Therefore, these coefficients have been excluded from being further investigated. As a result, only 31 non-monotonic similarity coefficients from 51 similarity coefficients were used in the investigations in Chapters 5 and 7. The retained coefficients were marked with an asterisk in the ID column in Table 4-5.

4.4.2 Distance Coefficients

The clustering algorithm in Chapter 6 used the distances of the molecules as a basis for grouping molecules in which two molecules that are closer will be clustered together. Therefore, ten distance coefficients have been implemented in this experiment to measure the pairwise distance between the molecules in the clustering procedure. The distance coefficients are available in the distance computations package library from SciPy (Jones et al., 2001). The distance coefficients are listed in Table 4-6, which describes the molecules x and y as represented by an n -binary vector, i.e., dimension. The binary vector element x_i contains the presence or absence of the i -th binary in x (and similarly for molecule y).

Table 4-5 The list of the binary coefficients

No.	ID	Symbol	Name	Formula	Index Definition A (Tversky)	Index Definition B (Todeschini)	Metric
1	*B1	SM	Sokal-Michener, Simple Matching	$S_{SM} = \frac{a+d}{p}$	A	S	M
2	B2	RT	Rogers-Tanimoto	$S_{RT} = \frac{a+d}{p+b+c}$	S	S	M
3	*B3	JT	Jaccard-Tanimoto	$S_{JT} = \frac{a}{a+b+c}$	S	A	M
4	B4	GLE	Gleason-Dice-Sorensen	$S_{GLE} = \frac{2a}{2a+b+c}$	S	A	N
5	*B5	RR	Russel-Rao	$S_{RR} = \frac{a}{p}$	A	A	M
6	*B6	FOR	Forbes	$S_{FOR} = \frac{pa}{(a+b)(a+c)}$	S	A	M
7	*B7	SIM	Simpson	$S_{SIM} = \frac{a}{\min\{(a+b), (a+c)\}}$	A	A	N
8	*B8	BB	Braun-Blanquet	$S_{BB} = \frac{a}{\max\{(a+b), (a+c)\}}$	A	A	M
9	*B9	DK	Driver-Kroeber, Ochiai, Cosine	$S_{DK} = \frac{a}{\sqrt{(a+b)(a+c)}}$	S	A	N
10	*B10	BUB	Baroni-Urbani-Buser	$S_{BUB} = \frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$	S	I	M
11	*B11	KUL	Kulczynski	$S_{KUL} = \frac{1}{2} \left[\frac{a}{a+b} + \frac{a}{a+c} \right]$	S	A	N
12	B12	SS1	Sokal-Sneath	$S_{SS1} = \frac{a+2b+2c}{2a+2d}$	S	A	M
13	B13	SS2	Sokal-Sneath	$S_{SS2} = \frac{2a+2d}{p+a+d}$	A	S	N
14	B14	JA	Jaccard	$S_{JA} = \frac{3a}{3a+b+c}$	S	A	N
15	*B15	FAI	Faith	$S_{FAI} = \frac{a+0.5d}{p}$	A	I	M

The asterisk mark in the ID column indicates the non-monotonic similarity coefficients that were retained and used in the investigations in Chapter 5 and 7.

Table 4-5 (continued)

No.	ID	Symbol	Name	Formula	Index Definition A (Tversky)	Index Definition B (Todeschini)	Metric
16	*B16	MOU	Mountford	$S_{MOU} = \frac{2a}{ab + ac + 2bc}$	S	A	M
17	*B17	MIC	Michael	$S_{MIC} = \frac{4(ad - bc)}{(a + d)^2 + (b + c)^2}$	S	Q	N
18	*B18	RG	Rogot-Goldberg	$S_{RG} = \frac{a}{2a + b + c} + \frac{d}{2d + b + c}$	S	S	M
19	*B19	HD	Hawkins-Dotson	$S_{HD} = \frac{1}{2} \left[\frac{a}{a + b + c} + \frac{b + c + d}{b + c + d} \right]$	S	S	M
20	*B20	YU1	Yule	$S_{YU1} = \frac{ad - bc}{ad + bc}$	S	Q	N
21	B21	YU2	Yule	$S_{YU2} = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$	S	Q	M
22	*B22	FOS	Fossum	$S_{FOS} = \frac{p(a - 0.5)^2}{(a + b)(a + c)}$	S	A	M
23	*B23	DEN	Dennis	$S_{DEN} = \frac{ad - bc}{\sqrt{p(a + b)(a + c)}}$	S	Q	M
24	B24	CO1	Cole	$S_{CO1} = \frac{ad - bc}{(a + c)(c + d)}$	S	Q	N
25	*B25	CO2	Cole	$S_{CO2} = \frac{ad - bc}{(a + b)(b + d)}$	S	Q	N
26	*B26	DIS	Dispersion	$S_{DIS} = \frac{ad - bc}{p^2}$	S	Q	N
27	B27	GK	Goodman-Kruskal	$S_{GK} = \frac{2 \min(a, d) - b - c}{2 \min(a, d) + b + c}$	S	S	N
28	*B28	SS3	Sokal-Sneath	$S_{SS3} = \frac{1}{4} \left[\frac{a}{a + b} + \frac{a}{a + c} + \frac{d}{b + d} + \frac{d}{c + d} \right]$	S	S	M
29	*B29	SS4	Sokal-Sneath	$S_{SS4} = \frac{a}{\sqrt{(a + b)(a + c)} \sqrt{(b + d)(c + d)}}$	S	S	M
30	*B30	PHI	Pearson-Heron	$S_{PHI} = \frac{ad - bc}{\sqrt{(a + b)(a + c)(c + d)(b + d)}}$	S	Q	M

The asterisk mark in the ID column indicates the non-monotonic similarity coefficients that were retained and used in the investigations in Chapter 5 and 7.

Table 4-5 (continued)

No.	ID	Symbol	Name	Formula	Index Definition A (Tversky)	Index Definition B (Todeschini)	Metric
31	B31	DI1	Dice, Wallace, Post-Snijders	$S_{DI1} = \frac{a}{(a+b)}$	A	A	N
32	B32	DI2	Dice, Wallace, Post-Snijders	$S_{DI2} = \frac{a}{(a+c)}$	A	A	N
33	*B33	SOR	Sorgenfrei	$S_{SOR} = \frac{a^2}{(a+b)(a+c)}$	S	A	N
34	*B34	COH	Cohen	$S_{COH} = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$	S	Q	N
35	B35	PE1	Peirce	$S_{PE1} = \frac{ad-bc}{(a+b)(c+d)}$	S	Q	N
36	*B36	PE2	Peirce	$S_{PE2} = \frac{ad-bc}{(a+c)(b+d)}$	S	Q	N
37	*B37	MP	Maxwell-Pilliner	$S_{MP} = \frac{2(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)}$	S	Q	M
38	*B38	HL	Harris-Lahey	$S_{HL} = \frac{a(2d+b+c)}{2(a+b+c)} + \frac{d(2a+b+c)}{2(b+c+d)}$	S	S	N
39	B39	CT1	Consonni-Todeschini	$S_{CT1} = \frac{\ln(1+a+d)}{\ln(1+p)}$	A	S	M
40	B40	CT2	Consonni-Todeschini	$S_{CT2} = \frac{\ln(1+p) - \ln(1+b+c)}{\ln(1+p)}$	S	S	N
41	B41	CT3	Consonni-Todeschini	$S_{CT3} = \frac{\ln(1+a)}{\ln(1+p)}$	A	A	N
42	*B42	CT4	Consonni-Todeschini	$S_{CT4} = \frac{\ln(1+a)}{\ln(1+a+b+c)}$	S	A	N
43	*B43	CT5	Consonni-Todeschini	$S_{CT5} = \frac{\ln[1+ad] - \ln[1+bc]}{\ln(1+p^2/4)}$	S	S	M
44	B44	AC	Austin_Colwell	$S_{AC} = \frac{2}{\pi} \arcsin \left(\sqrt{\frac{a+d}{p}} \right)$	A	S	M

The asterisk mark in the ID column indicates the non-monotonic similarity coefficients that were retained and used in the investigations in Chapter 5 and 7.

Table 4-5 (continued)

No.	ID	Symbol	Name	Formula	Index Definition A (Tversky)	Index Definition B (Todeschini)	Metric
45	B45	HAM	Hamann, Holley-Guilford, Hubert	$S_{HAM} = \frac{a+d-b-c}{p}$	S	S	M
46	*B46	MCC	McConaughy	$S_{MCC} = \frac{a^2 - bc}{(a+b)(a+c)}$	S	A	N
47	B47	GL	Gower-Legendre	$S_{GL} = \frac{a + 0.5(b+c) + d}{a+d}$	S	S	N
48	B48	BU2	Baroni-Urbani-Buser	$S_{BU2} = \frac{\sqrt{ad} + a - b - c}{\sqrt{ad} + a + b + c}$	S	I	M
49	B49	JOH	Johnson	$S_{JOH} = \frac{a}{a+b} + \frac{a+c}{a+c}$	S	A	N
50	B50	SCO	Scott	$S_{SCO} = \frac{4ad - (b+c)^2}{(2a+b+c)(2d+b+c)}$	S	S	M
51	*B51	MAA	van der Maarel	$S_{MAA} = \frac{2a-b-c}{2a+b+c}$	S	A	N

The asterisk mark in the ID column indicates the non-monotonic similarity coefficients that were retained and used in the investigations in Chapter 5 and 7.

Table 4-6 The list of the distance coefficients (Jones et al., 2001)

No.	ID	Symbol	Name	Formula
1	D1	BC	Bray-Curtis	$D_{BC} = \frac{\sum_{i=1}^n x_i - y_i }{\sum_{i=1}^n x_i + y_i }$
2	D2	CB	City-Block	$D_{CB} = \sum_{i=1}^n x_i - y_i $
3	D3	COS	Cosine	$D_{COS} = 1 - \frac{\sum_{i=1}^n x_i y_i}{[\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2]^{1/2}}$
4	D4	EUC	Euclidean	$D_{EUC} = \left[\sum_{i=1}^n x_i - y_i ^2 \right]^{1/2}$
5	D5	HAM	Hamming	$D_{HAM} = \frac{\sum_{i=1}^n x_i - y_i }{n}$
6	D6	JAC	Jaccard	$D_{JAC} = \frac{\sum_{i=1}^n x_i - y_i }{\sum_{i=1}^n x_i y_i + \sum_{i=1}^n x_i - y_i }$
7	D7	KUL	Kulsinski	$D_{KUL} = \frac{\sum_{i=1}^n x_i - y_i - \sum_{i=1}^n x_i y_i + n}{\sum_{i=1}^n x_i - y_i + n}$
8	D8	RT	Rogers-Tanimoto	$D_{RT} = \frac{2 \sum_{i=1}^n x_i - y_i }{\sum_{i=1}^n x_i y_i + (n - (\sum_{i=1}^n x_i y_i + \sum_{i=1}^n x_i - y_i)) + 2 \sum_{i=1}^n x_i - y_i }$
9	D9	RR	Russell-Rao	$D_{RR} = \frac{n - \sum_{i=1}^n x_i y_i}{n}$
10	D10	SS	Sokal-Sneath	$D_{SS} = \frac{2 \sum_{i=1}^n x_i - y_i }{\sum_{i=1}^n x_i y_i + 2 \sum_{i=1}^n x_i - y_i }$

The definitions describe the molecules x and y as represented by an n -binary vector, i.e., dimension. The binary vector element x_i contains the presence or absence of the i -th binary in x (and similarly for molecule y).

4.5 Experimental Procedure

4.5.1 Procedure of Similarity Searching

The experiment carried out in Chapter 5 replicates the virtual screening based similarity searching application, which calculates the similarity values between a reference structure and each structure in a dataset. Ten random reference structures from each activity class were used for the similarity searching.

Also, each similarity search was conducted for different fingerprint dimensions as described in Section 4.3. The similarity values were calculated based on different similarity coefficients as described in section 4.4.1. The similarity values computed were used to rank the molecules in decreasing order. A threshold was applied to retrieve a fixed number of top-ranked molecules, i.e., top 1%. Numbers of active molecules within the retrieved list were used to measure the effectiveness of the search based on the enrichment factor. The enrichment factors were then averaged over the ten searches and the value denoted by the symbol $\overline{EF}_{1\%}$. For the first investigation, the total number of similarity searches using all three datasets, thirteen fingerprint dimensions and fifty-one similarity coefficients was 265,200.

A similar similarity search procedure was applied in the third experiment as reported in Chapter 7. The difference was that the searches were conducted using ten types of fingerprints which were represented by one size of dimension (i.e., 2^{10} or 1,024 bits), measured by only 31 similarity coefficients and using only the ChEMBL dataset (which has 15 activity classes). This investigation yielded a total number of 46,500 similarity searches.

4.5.2 Procedure of Clustering

The agglomerative hierarchical non-overlapping clustering method has been chosen as the method for clustering the molecules in Chapter 6. Based on this method, each molecule (or cluster of molecules) merges bottom-up with other similar molecules. The merges were determined by different types of methods, resulting in a cluster of two molecules or clusters of several molecules. The

procedure is non-overlapping, which means that a molecule can occur only in one cluster. Two types of algorithms were implemented in the experiment, which are the Ward's algorithm and the Group Average algorithm.

Ward's algorithm has been the most widely used clustering algorithm in chemoinformatics applications (Brown & Martin, 1996; Bayada et al., 1999). It has also been found to perform better than other non-hierarchical cluster algorithms in terms of its predictive ability (Downs et al., 1994). Based on Ward's algorithm, the clusters are grouped so as to minimise the total variance for each cluster (Ward, 1963). At each process, a pair of clusters is chosen whose merger leads to the minimum change in total variance. The variance of a cluster is measured as the sum of the squared deviations from the mean of the cluster. For a cluster, c , of N_c objects where each object j is represented by a vector $r_{c,j}$, the mean (or centroid) of the cluster, \bar{r}_c and the intracluster variance, v_c are determined by Eq. (20) and Eq. (21):

$$\bar{r}_c = \frac{1}{N_c} \sum_{j=1}^{N_c} r_{c,j} \quad (20)$$

$$v_c = \sum_{j=1}^{N_c} (|r_{c,j} - \bar{r}_c|)^2 \quad (21)$$

The total variance is measured as the sum of the intracluster variances for each cluster. For each iteration, a pair of clusters is chosen whose merger leads to the minimum change in total variance.

Ward's algorithm tends to produce spherical clusters which may not accurately reflect the true shape of the clusters present in the dataset (Willett, 1987). For this reason, further experiment has been conducted using the Group Average algorithm. In this algorithm, the intercluster distance is measured as the average of the distances between all pairs of compounds in the two clusters. As a result, each cluster member has a smaller average distance to the remaining members of that cluster than to all members of any other cluster. The results

from both Ward's and Group Average algorithms were considered in order to identify a comparable and conclusive finding about the experiment.

4.6 Evaluation Method

4.6.1 Enrichment Factor

In the investigations described in Chapters 5 and 7, the *enrichment factor* (*EF*) was chosen to measure the effectiveness of the similarity search application. This method is commonly used when the number of actives retrieved is more important than the active ranking order. It measures the active compounds retrieved compared to active compounds from a random selection. The calculation of the *EF* is defined by Eq. (22):

$$EF = \frac{AR}{R} \quad (22)$$

where *AR* is the number of active compounds retrieved, and *R* is the number of actives expected based on random selection, for a given cut off value. The typical cut off value used in these experiments is 1%. The search effectiveness for each representation was measured by the mean enrichment factor when averaged over the ten searches for each activity class.

4.6.2 F-Measure

The *F*-measure was first devised to evaluate methods for document clustering in information retrieval (van Rijsbergen, 1979). It evaluates the extent to which a method clustered together molecules that belonged to the same activity class.

Assume that a cluster contains *n* molecules, that *a* of these are active and that there is a total of *A* molecules with the chosen activity. The precision, *P*, and the recall, *R*, for that cluster are then calculated by Eq. (23):

$$P = \frac{a}{n} \quad \text{and} \quad R = \frac{a}{A} \quad (23)$$

F is the harmonic mean of P and R that is calculated by Eq. (24):

$$F = \frac{2PR}{P + R} \quad (24)$$

This calculation is carried out for each cluster. The F -measure is the maximum value obtained across all clusters. This value describes the single cluster that provides the best combination of precision and recall for the current bioactivity assuming both P and R are of equal importance.

4.6.3 QPI-Measure

QPI -measure is a method for evaluating the clustering effectiveness that was developed from the QCI (*Quality Clustering Index*) (Varin et al., 2008). It is used to evaluate the performance of a clustering algorithm by measuring the separation between active and inactive molecules resulting from the use of a clustering method.

In this approach, an *active cluster* is defined as a non-singleton cluster where the percentage of active molecules in the cluster is greater than the percentage of active molecules in the database as a whole. Let p be the number of active molecules in the active clusters, q the number of inactive molecules in the active clusters, r the number of active molecules in the inactive clusters (i.e., clusters that are not active clusters) and s the number of singletons that are active molecules. The quality partition index, QPI , is then calculated by Eq. (25):

$$QPI = \frac{p}{p + q + r + s} \quad (25)$$

This calculation will result in a high value when the active molecules are clustered tightly together and separated from the inactive molecules.

The QPI -measure describes the entire set of clusters, while the F -measure describes the single best cluster. These approaches have been used to evaluate

the performance of molecular clustering by several previous studies in chemoinformatics domain (Chu et al., 2012; Gan et al., 2014). For each algorithm (i.e., Ward's and Group Average), the clusters were generated for all 260 combinations of fingerprint dimensions measured by ten distance coefficients for two datasets to obtain each of the six partitions of 500, 600, 700, 800, 900 and 1000 clusters. The F and QPI values were also computed for each cluster partition. Both evaluation methods were implemented in the second investigation in this thesis as reported in Chapter 6.

4.7 Statistical Method

4.7.1 Spearman's Rank Correlation

The Spearman's rank correlation test was used to identify monotonicity, i.e., when two different similarity coefficients produce the same similarity rankings, which is another important characteristic of a similarity coefficient. Similarity search results for each similarity coefficient measuring similar fingerprint size and reference molecule were chosen. The results were tested using the Spearman's rank correlation. The monotonic coefficients were identified and grouped together, i.e., coefficients with correlation value = 1. This statistical test was implemented in the first investigation as reported in Chapter 5.

4.7.2 Kendall's W Test

The Kendall's W test was used to test the significance of the performance of each similarity coefficient. The test was done using IBM SPSS version 22 (IBM Corp. *IBM SPSS Statistics for Windows*, 2013) and by measuring the $\overline{EF}_{1\%}$ from all activity classes using all fingerprint dimensions. For each dataset, the mean $\overline{EF}_{1\%}$ obtained from all fingerprint dimensions were averaged and the similarity coefficients were ranked based on their average mean $\overline{EF}_{1\%}$ value. The similarity coefficient with the largest average mean $\overline{EF}_{1\%}$ value would be ordered as the highest in the row (i.e., first in the rank position) and vice versa. This statistical test was also implemented in the first investigation as reported in Chapter 5.

4.7.3 Sign Test

The Sign test was used to validate the significance of the contribution between the compound representations and the similarity coefficients in determining the performance of similarity searching using the cross classified multilevel modeling in Chapter 7. It was implemented to measure the contribution in order to make a conclusion about which factor is more important.

The Sign test is based on the *direction* of the differences between the two components to test the following null hypothesis, H_0 using Eq. (26):

$$P[X_i > Y_i] = P[X_i < Y_i] = \frac{1}{2} \quad (26)$$

where P is the number of pairs which have X_i or Y_i scores greater or less than the other for two different components that are to be compared, X and Y . In this test, the *sign* of the difference between each pair of X_i and Y_i scores is noted as positive (+) or negative (-). H_0 is true if half of the differences are negatives and half are positives. H_0 is rejected if too few differences of one sign occur.

In the case of “tie” occurrences, all tied pairs are dropped from the analysis and the sample size (i.e., number of pairs), N is reduced correspondingly. In other words, N is the number of pairs whose differences show a sign (+ or -). This is because it is not possible to discriminate between the values of a tied pair.

Two different methods can be used to determine the probability associated with the occurrence of data, which depends on the sample size. For a small sample size of $N \leq 35$, the probability can be determined by reference to the binomial distribution with $p = q = \frac{1}{2}$. The significance of the probability values can be looked up by referring to the binomial distribution table (Siegel & Castellan Jr, 1988).

For a large sample size of $N > 35$, the probability can be determined by normal approximation to the binomial distribution and measured using the z-score in Eq. (27):

$$z = \frac{2x \pm 1 - N}{\sqrt{N}} \quad (27)$$

where N is the number of pairs and x is the number of fewer signs, for which +1 is used when $x < \frac{N}{2}$ and -1 when $x > \frac{N}{2}$. The significance of the obtained z value can be looked up by referring to the normal distribution table (Siegel & Castellan Jr, 1988).

The sign test may be either one-tailed or two-tailed. In a one-tailed test, the alternative hypothesis states which sign (+ or -) will occur more frequently. The two-tailed test predicts the frequencies with which the two signs occur that will be significantly different.

In the study in Chapter 7, the sign test was conducted to evaluate the differences of variances of the two components. Each variance acts as a judge of the similarity search effectiveness, where the significance of the differences is measured by the number of (i) fingerprint level > similarity coefficient level, (ii) fingerprint level = similarity coefficient level and (iii) fingerprint level < similarity coefficient level. The two-tailed test was considered for the sign test in which the probability values obtained from the lookup tables are doubled. The test was done using IBM SPSS version 22. Detailed explanation about the implementation of the test is explained separately in the corresponding sections in Chapter 7.

4.7.4 The Wilcoxon Signed-rank Test

In addition to the Sign test, the Wilcoxon signed-rank test was also implemented to validate the significance of the contribution between the compound representations and the similarity coefficients in determining the performance of similarity searching. The Wilcoxon signed-rank test is a more powerful test that can be used to compare two sets of components which not only utilises the *direction* of the preferences of a component, but also includes the relative *magnitude* of the direction in the comparison. Hence, it gives more weight to a pair, which shows larger difference than a smaller one.

In order to carry out this test, first the sign's differences d_i of each pair X_i and Y_i need to be determined, where $d_i = X_i - Y_i$. All resulting d_i values will be ranked without regard to sign with 1 being the smallest $|d_i|$. Next, the sign of the difference is affixed to each rank to indicate which rank is positive or negative from d_i .

The null hypothesis H_0 is true when there exist equal values of summation of positive d_i as well as negative d_i . Here, N is again the number of non-zero d_i , which is used in defining these two statistics:

$$T^+ = \text{the sum of the positive } d_i\text{'s ranks}$$

$$T^- = \text{the sum of the negative } d_i\text{'s ranks}$$

Since the sum of all of the ranks is $\frac{N(N+1)}{2}$, then $T^- = \frac{N(N+1)}{2} - T^+$. The H_0 is rejected when the T^+ or T^- is too small, i.e., when either summation of the ranks is different from the other.

The "tie" case may occur when the two scores of any pair are equal, i.e., $X_i - Y_i = d_i = 0$. The same practice with the sign test will be followed, which excludes the tied pairs from the analysis and reduces the number of pairs, N correspondingly. Another tie case can occur when two or more differences, d 's are of the same magnitude. For this case, the same rank, which is the average of the ranks of the same d 's, will be assigned.

For a small sample size of $N \leq 15$, the probability value is determined based on the sum of the positive d_i 's ranks, T^+ which can be looked up by referring to the probabilities table for critical values of T^+ for the Wilcoxon signed-ranks test (Siegel & Castellan Jr, 1988). The one-tailed test is appropriate if the direction of the differences has been predicted in advance.

For a large sample size of $N > 15$, the probability of the sum of the positive ranks, T^+ can be determined by normal approximation and measured using the z -score (Eq. (28)):

$$z = \frac{T^+ - \frac{N(N + 1)}{4}}{\sqrt{\frac{N(N + 1)(2N + 1)}{24}}} \quad (28)$$

where N is the number of pairs and the significance of the obtained z value can be looked up by referring to the normal distribution table (Siegel & Castellan Jr, 1988).

If the probability value is less than or equal to the significance level, α , then the H_0 can be rejected in favour of the alternative hypothesis by concluding that there is a significant difference between components X and Y and that either X or Y has shown better performance than the other.

Similar to the sign test, the two-tailed test was considered for the Wilcoxon signed-ranked test in the third investigation in Chapter 7. IBM SPSS version 22 was used to compute the statistical test, making it a very useful statistical software for carrying out such analysis. Detailed explanation about the implementation of the test is explained separately in the corresponding sections in Chapter 7.

4.8 Conclusion

This chapter presented the methodologies involved in the investigations reported in this thesis. It introduced the datasets that have been tested, the experimental design involved, the evaluation and the statistical methods that have been implemented. The other experimental details, which vary depending on the investigations conducted, will be introduced separately in each experimental chapter.

Chapter 5 Investigation into the Effect of Dimensionality on the Effectiveness of Similarity Searching

5.1 Introduction

The effects of the curse of dimensionality have been discussed in Chapter 3. The previous study has reported that the effectiveness of a nearest neighbour search application decreases as the dimensionality increases (Donoho, 2000). This study will investigate the effect of changing the dimensionality of molecular representations on the effectiveness of virtual screening based similarity search applications.

This study seeks to test the hypothesis that as the dimensionality increases, the effectiveness of the nearest neighbour searches decreases. In contrast, studies carried out in the chemoinformatics domain have shown that similarity searching is found to be effective using high dimensional molecular representation (Willett, 2011b). Thus, the aim of this study is to identify the characteristics of chemical datasets that contribute to the effectiveness of the application in high dimensionality. It also explains the observed performance using various molecular dimensions and similarity coefficients, which simulate a practical virtual screening process.

5.2 Experimental Design

In this investigation, the experiments simulate virtual screening experiments, which calculate the similarity between a reference structure and each structure in a dataset. The experiments were carried out for all activity classes from three datasets, i.e., MDDR, WOMBAT and ChEMBL. These datasets have been introduced in Chapter 4, along with the similarity searching procedures.

Each compound in the datasets was represented using the binary fingerprint, i.e., ECFP₄-like (MorganR2) fingerprint. To investigate effect of changing the dimensionality of molecular representations on the effectiveness of similarity

search applications, thirteen different fingerprint sizes have been used in this study. These fingerprints have been introduced in Chapter 4.

To observe the performance using various similarity coefficients, 51 similarity coefficients were implemented to measure the similarity of the compounds. These coefficients have been used in the previous study by Todeschini et al. (2012) and introduced in Chapter 4.

5.3 Results and Discussion

5.3.1 Analysis of Spearman's Rank Correlation

The Spearman's rank correlation test has been carried out for all similarity coefficients used in this experiment as described in Chapter 4. Table 5-1 shows twenty nine coefficients that were grouped into nine monotonic groups. All coefficients in the same group were monotonic to each other. Twenty-two other coefficients are the singletons, i.e., non-monotonic coefficients. As can be seen from Table 4-5, several coefficients were derived by a very similar equations. For example, the B3 (JT) and B4 (GLE/DICE) are monotonic based on their formulation which differs in the weightings of the component a .

Only one coefficient from each group, i.e., the best known coefficient, and the singletons were retained for the results and discussion. The total number of retained coefficients is 31, which are shown in bold in the Table 5-1. Several correlated groups are in agreement with the previous study by Todeschini et al. (2012), e.g., B3 (JT), B4 (GLE/DICE), B12 (SS1) and B14 (JA).

Table 5-1 Spearman's rank correlations result

Monotonic Group	Correlated Similarity Coefficients	<i>p</i> value
1	B1 , B2, B13, B39, B40, B44, B45, B47	1
2	B3 , B4, B12, B14, B27	1
3	B5 , B31, B41	1
4	B6 , B24, B32	1
5	B10 , B48	1
6	B11 , B49	1
7	B18 , B50	1
8	B20 , B21	1
9	B26 , B35	1
Singletons	B7, B8, B9, B15, B16, B17, B19, B22, B23, B25, B28, B29, B30, B33, B34, B36, B37, B38, B42, B43, B46, B51	-

5.3.2 Analysis of Kendall's *W* Test

The Kendall's *W* tests have been carried out for the mean $\overline{EF}_{1\%}$ values of all similarity coefficients as explained in Chapter 4. For each fingerprint dimension, the similarity coefficient with the largest average mean $\overline{EF}_{1\%}$ value would be ordered first in the rank position. For example, in Table 5-2, the B18 coefficient has the largest value of average mean $\overline{EF}_{1\%}$, i.e., 24.59 (refer to the second last column). Hence, it is ordered as the highest in the row (i.e., rank position 1). The B7 coefficient is ordered as the lowest in the row (i.e., rank position 31) because it has the smallest value of average mean $\overline{EF}_{1\%}$, i.e., 3.80 (refer to the second last column). In addition, the table also presents the mean $\overline{EF}_{1\%}$ and rank position of the similarity coefficients obtained for each dimension. The other values, i.e., the *W*, χ^2 and significant values were also recorded.

For the MDDR average mean values, with $k = 31$, $N = 11$ and the searches with $\overline{EF}_{1\%}$, the test yields the values of *W* between the range of 0.433 to 0.613 and χ^2 between 142.73 to 202.38. The values are highly significant with value $p \leq$

0.001. Thus, the results from Table 5-2 suggest the following rankings (see Figure 5-1):

B18 > B38 > B34 > B3 > B19 > B37 > B8 > B29 > B42 > B30 > B51 > B22 > B33 > B9 > B10 > B23 > B28 > B26 > B11 > B17 > B46 > B25 > B16 > B43 > B15 > B20 > B36 > B5 > B6 > B1 > B7

It is interesting to see that the B3 (JT) coefficient demonstrated a good performance in the similarity search using seven fingerprint dimensions, i.e., 2^8 , 2^9 , 2^{10} , 2^{12} , 2^{13} , 2^{14} , 2^{15} bits. Of all seven dimensions, 2^{14} bits dimension equals to the highest W value of 0.613 while the χ^2 value yielded is 202.38. This has also been the highest W value calculated for all thirteen dimensions investigated in the MDDR dataset. However, the B3 coefficient was ranked the fourth in the final rank position because the final rank position is based on the average mean values. The B1 coefficient was the worst for nine out of all thirteen dimensions (from 2^9 until 2^{17}) with the highest W and χ^2 values obtained from the same dimension, i.e., 2^{14} .

Table 5-3 and Figure 5-2 shows results for WOMBAT dataset suggested the following rankings in both tabular and graphical form. With $k = 31$, $N = 14$ and the searches with $\overline{EF}_{1\%}$, the test yields values for W between the range of 0.506 to 0.726 and χ^2 between 212.50 to 304.82 (all values have $p \leq 0.001$):

B38 > B18 > B34 > B3 > B37 > B19 > B42 > B8 > B29 > B22 > B30 > B9 = B33 > B51 > B10 > B23 > B26 > B28 > B11 > B17 > B46 > B25 > B15 > B16 > B43 > B5 > B20 > B36 > B1 > B7 > B6

For the WOMBAT dataset, the B42 coefficient demonstrated a good performance in the similarity search using seven fingerprint dimensions starting from 2^{11} until 2^{17} dimensions. Of all seven dimensions, 2^{11} equals to the highest W value of 0.659 while the χ^2 value yielded is 276.96. Similar to the MDDR ranking, the B1 coefficient was also the worst for the same nine dimensions, i.e., 2^9 until 2^{17} .

And finally, for ChEMBL with $k = 31$, $N = 15$ and the searches with $\overline{EF}_{1\%}$, the test yields values for W between the range of 0.392 to 0.676 and χ^2 between 176.50 to 304.03 (all values have $p \leq 0.001$). Results from Table 5-4 suggest the following rankings (see Figure 5-3):

B38 > B42 > B3 > B18 > B34 > B37 > B19 > B26 > B22 > B29 > B9 > B33 > B8 >
B30 = B51 = B17 > B10 > B25 > B23 > B28 > B11 > B46 > B16 > B43 > B20 >
B15 > B5 > B36 > B6 > B1 > B7

Both B38 and B42 coefficients demonstrated good performances in the similarity search using the eight high dimensions from 2^{10} until 2^{17} bits for the ChEMBL dataset. The test for bit dimension of 2^{12} yields the highest W value of 0.676 and the χ^2 value is 304.03 which was demonstrated by the B38 coefficient. Similar to the MDDR and WOMBAT rankings, the B1 coefficient was also the worst for the same nine bits dimensions, i.e., 2^9 until 2^{17} .

Overall, the average Kendall's W rankings using all thirteen dimensions as mentioned above seem comparable. For searches with $\overline{EF}_{1\%}$ across all fingerprint dimensions, B38 performs extremely well in all datasets, except for MDDR where B18 is shown to be the best performer. B7 is the worst similarity coefficient suggested to be used for MDDR and ChEMBL while B6 is the worst suggested for WOMBAT. When referring to the previous study, the best performance and the worst performance using the 2^{10} bit dimension for MDDR is in line with the Todeschini *et al.*'s finding (i.e., B3 as the best performance and B1 as the worst performance).

Table 5-2 Kendall's W results for the top 1% based on the average actives retrieved for MDDR dataset. Mean rank indicates the value of mean $EF_{1\%}$ obtained from the $EF_{1\%}$ values averaged over the 11 activity classes in the dataset. For each fingerprint dimension, the highest ranked similarity measure is marked by blue box and the lowest ranked by orange box for ease of reference.

Similarity Measures	Kendall's W Mean Rank for MDDR																																	Average mean rank	Average position										
	2^5			2^6			2^7			2^8			2^9			2^{10}			2^{11}			2^{12}			2^{13}			2^{14}			2^{15}					2^{16}			2^{17}						
	mean	rank	position	mean	rank	position	mean	rank	position	mean	rank	position	mean	rank	position	mean	rank	position	mean	rank	position	mean	rank	position	mean	rank	position	mean	rank	position	mean	rank	position	mean	rank	position	mean	rank	position	mean	rank	position	mean	rank	position
B18	25.41	2	24.77	2	24.14	2	23.32	2	23.50	2	23.59	3	26.09	3	25.55	3	25.05	2	25.05	2	24.86	5	24.86	5	24.86	5	24.64	5	24.95	1.5	25.09	1.5	25.09	1.5	24.52	1	24.42	2							
B38	24.64	5	23.05	6	23.23	3	22.05	6	22.50	5	24.09	2	25.59	2	25.77	1	25.59	4	24.77	4	26.05	4	25.50	3	25.05	3.5	24.95	1.5	24.50	1.5	25.09	1.5	25.09	1.5	24.19	3	23.59	4							
B34	26.18	1	23.27	4	22.18	5	21.68	7	23.09	4	23.59	3.5	23.09	5	24.77	4	26.05	4	25.50	3	25.05	3.5	25.05	3.5	25.05	3.5	24.95	1.5	24.50	1.5	25.09	1.5	25.09	1.5	23.59	4	23.59	4							
B3	16.95	14.5	17.77	15.6	19.86	10	23.50	1	25.18	1	25.68	1	26.27	1	26.59	1.5	25.68	1	26.27	1	26.59	1.5	25.68	1	25.27	1	24.50	6	24.36	6	24.36	6	24.36	6	23.59	4	23.59	4							
B19	24.77	4	26.00	1	21.59	6	18.73	15.5	20.23	13	20.77	13	21.45	9	23.09	7	23.68	6	23.95	6	24.41	5	25.23	6	24.45	6	24.68	3.5	24.55	5	24.55	5	23.09	4	23.09	4	22.94	6	22.94	6					
B37	24.18	6	21.14	8	19.55	11	19.73	12.5	21.68	6	21.86	7.5	22.55	6	23.95	6	23.95	6	24.41	5	25.23	6	24.45	6	24.68	3.5	24.77	4	24.77	4	22.94	6	22.94	6	22.94	6	22.94	6	22.94	6					
B8	20.09	9	18.64	14	17.18	15	20.05	11	21.59	8	22.00	6	25.41	2	23.55	6	23.27	7	24.18	7	23.77	7	24.00	7	24.18	7	24.00	7	24.18	7	22.15	7	22.15	7	22.15	7	22.15	7	22.15	7					
B29	23.95	7	21.95	7	22.23	4	22.45	4	20.95	10	20.91	12	19.91	13	19.64	11	19.27	12	18.95	14	19.59	13	19.73	13	19.55	14	20.70	14	20.70	14	20.70	14	20.70	14	20.70	14	20.70	14	20.70	14					
B42	15.73	16	13.68	18	13.27	22	18.23	17	21.09	9	21.86	7.5	22.50	7	22.14	8	22.18	8	22.09	8	23.14	8	23.23	8	22.68	8	20.14	9	20.14	9	20.14	9	20.14	9	20.14	9	20.14	9	20.14	9	20.14	9			
B30	24.82	3	20.50	10	18.68	12	18.73	15.5	18.41	14	19.36	14	19.00	14	18.55	14	17.77	15	19.27	11	19.82	12	19.86	11	19.59	12.5	19.57	10	19.57	10	19.57	10	19.57	10	19.57	10	19.57	10	19.57	10	19.57	10			
B51	16.95	14.5	17.77	15.6	20.05	9	23.09	3	23.18	3	22.82	5	20.86	12	20.18	10	18.00	13	17.05	16	17.00	15	16.95	15	17.00	15	19.30	11	19.30	11	19.30	11	19.30	11	19.30	11	19.30	11	19.30	11	19.30	11			
B22	14.41	20	11.68	24	14.59	20	20.73	8	21.64	7	21.64	9	21.50	8	21.91	9	21.09	9	20.82	9	19.00	14	19.18	14	19.59	12.5	19.06	12	19.06	12	19.06	12	19.06	12	19.06	12	19.06	12	19.06	12	19.06	12			
B33	15.14	17.5	13.64	19.2	16.50	17	20.59	9.5	20.91	11.5	21.05	10.5	20.91	10.5	20.91	10.5	20.91	10.5	20.91	10.5	20.91	10.5	20.91	10.5	20.91	10.5	18.91	13	18.91	13	18.91	13	18.91	13	18.91	13	18.91	13	18.91	13	18.91	13			
B9	15.14	17.5	13.64	19.2	16.41	18	20.59	9.5	18.36	15	15.73	18	17.05	16	16.64	16	16.59	16	17.23	15	16.73	16	16.23	16	16.36	16	18.82	15	18.82	15	18.82	15	18.82	15	18.82	15	18.82	15	18.82	15	18.82	15			
B10	22.77	8	24.41	3	24.41	3	22.18	5	22.18	5	22.18	5	18.36	15	15.73	18	17.05	16	16.64	16	16.59	16	17.23	15	16.73	16	16.23	16	16.36	16	16.36	16	16.36	16	16.36	16	16.36	16	16.36	16					
B23	14.77	19	13.27	21	15.00	19	14.95	20	16.00	17	17.64	15	17.45	15	18.18	15	17.91	14	20.05	10	21.05	9	19.91	9	19.91	9	19.82	9	19.82	9	19.82	9	19.82	9	19.82	9	19.82	9	19.82	9	19.82	9			
B28	18.68	10	19.68	11	17.68	14	16.55	18	15.32	19	15.91	17	14.64	18	14.73	18	14.86	18	14.41	18	13.95	18	13.95	18	14.14	18	15.73	17	15.73	17	15.73	17	15.73	17	15.73	17	15.73	17	15.73	17	15.73	17			
B26	18.18	11	19.55	12	21.18	8	19.73	12.5	15.95	18	14.82	19	14.36	19	13.32	19	12.68	20	12.64	19.5	13.27	19.5	13.86	19.5	13.86	19.5	15.65	18	15.65	18	15.65	18	15.65	18	15.65	18	15.65	18	15.65	18	15.65	18			
B11	13.36	21	8.95	26.5	10.68	27	15.09	19	16.05	16	16.91	16	15.64	17	15.04	17	15.00	17	14.77	17	14.32	17	14.23	17	14.23	17	14.18	19	14.18	19	14.18	19	14.18	19	14.18	19	14.18	19	14.18	19	14.18	19			
B17	12.82	23	16.36	17	18.50	13	11.45	23	10.14	25	10.23	24	11.45	22	12.05	21	12.68	20	12.36	21	13.00	21	13.59	21	13.59	21	12.94	20	12.94	20	12.94	20	12.94	20	12.94	20	12.94	20	12.94	20	12.94	20			
B46	13.14	22	9.14	25	10.91	26	14.77	21	14.77	20	13.95	20	12.32	21	11.91	22	12.14	22	11.82	22	11.55	22	11.77	22	11.64	22	12.29	21	12.29	21	12.29	21	12.29	21	12.29	21	12.29	21	12.29	21	12.29	21			
B25	9.36	27	8.64	28	8.68	28	9.55	26	12.18	22	12.32	21	12.82	20	12.55	20	12.68	20	12.64	19.5	13.27	19.5	13.86	19.5	13.86	19.5	11.72	22	11.72	22	11.72	22	11.72	22	11.72	22	11.72	22	11.72	22	11.72	22			
B16	17.09	13	20.77	9	16.68	16	11.82	22	10.00	26	9.05	26	8.73	27	8.77	26	8.55	26	8.77	26	8.91	26	8.91	26	8.59	26	11.28	23	11.28	23	11.28	23	11.28	23	11.28	23	11.28	23	11.28	23	11.28	23			
B43	6.64	28	12.68	22	12.91	23	11.32	24	11.95	23	11.32	23	11.23	23	11.23	23	10.91	24	10.64	23	10.59	23	11.05	23	11.32	23	11.06	24	11.06	24	11.06	24	11.06	24	11.06	24	11.06	24	11.06	24	11.06	24			
B15	10.41	25	8.95	26.5	21.36	7	19.41	14	13.91	21	9.82	25	9.23	26	8.45	27	8.00	27	8.27	27	8.09	27	8.09	27	8.18	27	10.94	25	10.94	25	10.94	25	10.94	25	10.94	25	10.94	25	10.94	25	10.94	25			
B20	2.41	30.5	12.23	23	12.73	24	11.14	25	11.59	24	10.86	23	11.09	24	11.14	24	11.14	24	10.32	24	10.41	24	10.50	24	10.64	24	10.48	26	10.48	26	10.48	26	10.48	26	10.48	26	10.48	26	10.48	26	10.48	26			
B36	9.95	26	18.91	13	12.00	25	8.50	27	6.95	27	7.00	28	6.82	28	6.64	28	6.64	28	6.68	28	6.32	28	6.45	28	6.36	28	8.40	27	8.40	27	8.40	27	8.40	27	8.40	27	8.40	27	8.40	27	8.40	27			
B5	6.00	29	1.73	31	1.55	31	2.45	30	6.64	28	8.77	27	9.73	25	9.95	25	10.00	25	10.05	25	10.00	25	10.27	25	10.27	25	7.49	28	7.49	28	7.49	28	7.49	28	7.49	28	7.49	28	7.49	28	7.49	28			
B6	12.18	24	8.14	29	6.45	29	5.55	29	5.00	29	5.68	29	5.77	29	5.59	29	6.00	29	5.73	29	6.00	29	5.82	29	5.82	29	6.44	29	6.44	29	6.44	29	6.44	29	6.44	29	6.44	29	6.44	29	6.44	29			
B1	17.45	12	23.18	5	13.82	21	5.82	28	3.00	31	1.95	31	1.64	31	1.55	31	1.55	31	1.55	31	1.55	31	1.45	31	1.45	31	5.84	30	5.84	30	5.84	30	5.84	30	5.84	30	5.84	30	5.84	30	5.84	30			
B7	2.41	30.5	1.91	30	2.00	30	2.27	31	3.32	30	3.77	30	4.14	30	5.45	30	4.68	30	4.58	30	4.82	30	4.95	30	4.82	30	3.80	31	3.80	31	3.80	31	3.80	31	3.80	31	3.80	31	3.80	31	3.80	31			
W	0.547		0.520		0.433		0.493		0.507		0.543		0.560		0.573		0.608		0.613		0.601		0.581		0.583		0.584		0.584		0.584		0.584		0.584		0.584		0.584		0.584				
χ^2	180.67		171.64		142.73		162.62		167.25																																				

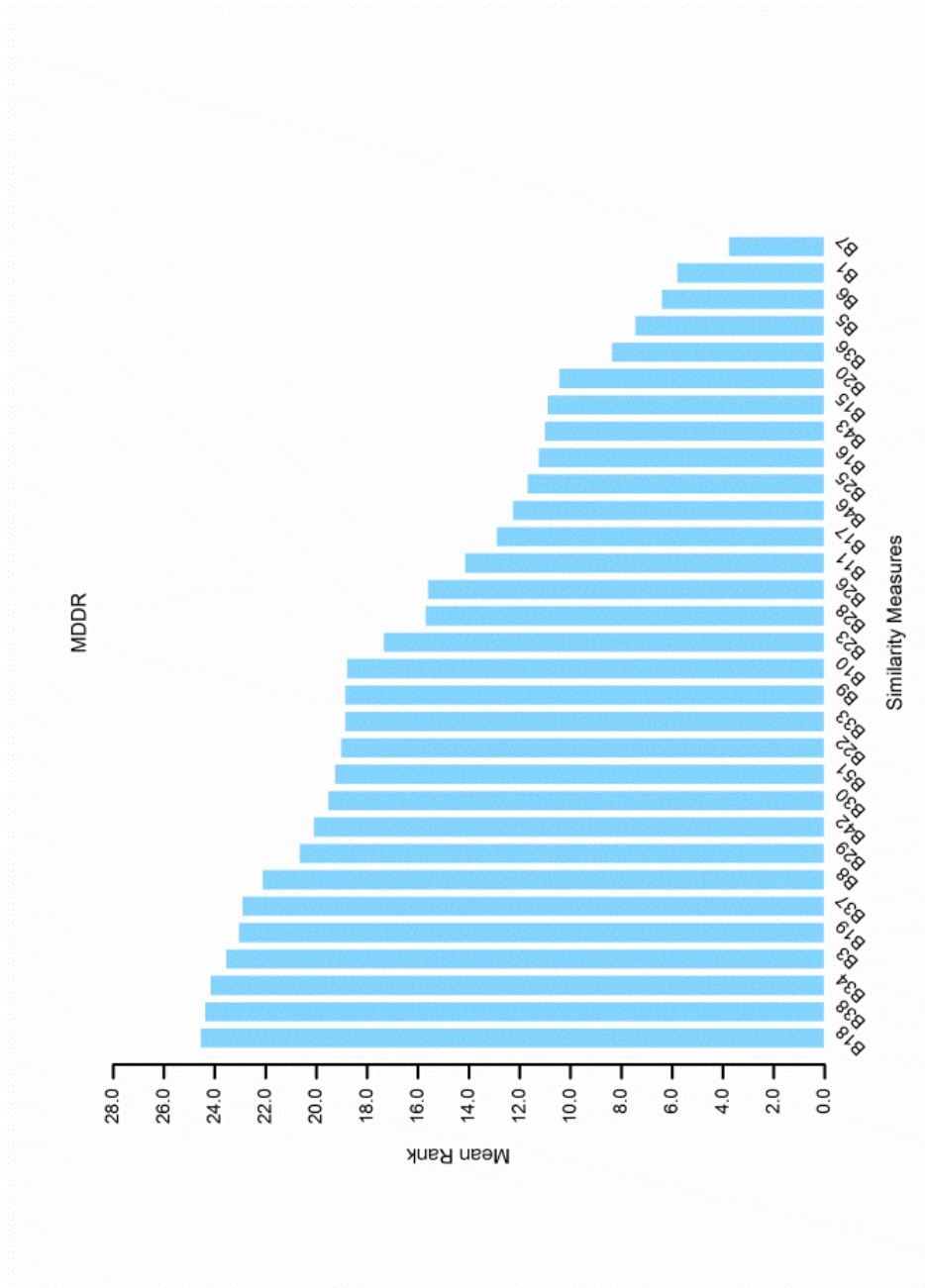


Figure 5-1 Performance of the 31 similarity coefficients, as obtained from MDDR dataset, ordered from best (high mean rank values) to worst (low mean rank values)

Table 5-3 Kendall's W results for the top 1% based on the average actives retrieved for WOMBAT dataset. Mean rank indicates the value of mean $\overline{EF}_{1\%}$ obtained from the $\overline{EF}_{1\%}$ values averaged over the 14 activity classes in the dataset. For each fingerprint dimension, the highest ranked similarity measure is marked by blue box and the lowest ranked by orange box for ease of reference.

Similarity Measures	Kendall's W Mean Rank for WOMBAT														Average													
	2 ⁵		2 ⁶		2 ⁷		2 ⁸		2 ⁹		2 ¹⁰		2 ¹¹		2 ¹²		2 ¹³		2 ¹⁴		2 ¹⁵		2 ¹⁶		2 ¹⁷		mean rank	rank position
	mean rank	rank position	mean rank	rank position	mean rank	rank position	mean rank	rank position	mean rank	rank position	mean rank	rank position	mean rank	rank position	mean rank	rank position	mean rank	rank position	mean rank	rank position	mean rank	rank position	mean rank	rank position	mean rank	rank position	mean rank	rank position
B38	23.96	4.5	25.00	5	27.32	1	23.54	6	24.21	3.5	25.18	3	24.18	4	25.21	2	25.04	2	25.04	2	24.32	3	24.07	3	24.11	3	24.76	1
B18	25.46	1	27.29	1	25.14	3	25.18	1	25.29	2	24.75	4.5	24.14	5	23.21	5	23.71	5	23.71	5	23.29	4	23.21	7	22.68	6	24.44	2
B34	24.36	3	26.46	4	24.00	5	23.64	5	24.21	3.5	24.36	6	23.89	5	23.36	4	23.82	4	23.82	4	23.14	5	23.43	5.5	22.68	6	23.90	3
B3	15.04	14.5	16.64	16.5	20.32	11.5	24.46	3	26.46	1	25.75	1	24.89	2	24.50	3	24.43	3	24.43	3	24.68	2	24.43	2	24.14	2	23.20	4
B37	23.96	4.5	22.04	9	21.21	9	21.36	8	23.04	8	23.71	7	23.32	7	23.11	6	23.36	6	23.36	6	22.86	6	23.43	5.5	22.68	6	22.86	5
B19	25.00	2	27.04	2	22.18	8	19.64	13	21.29	12	21.46	10	22.11	8.5	21.50	9	21.82	9	21.82	9	22.82	7	23.46	4	23.14	4	22.63	6
B42	14.71	16	12.82	18	12.54	21	18.82	16	23.39	6	24.75	4.5	26.36	1	25.82	1	25.86	1	25.86	1	26.46	1	26.61	1	26.25	1	22.36	7
B8	19.86	12	19.39	13	15.46	18	19.43	14	23.07	7	25.50	2	23.75	6	21.79	8	22.39	7.5	22.39	7.5	22.36	8	22.29	8	22.21	9	21.78	8
B29	23.46	6	24.79	6	24.68	4	24.57	2	22.46	9	20.21	13	23.75	6	20.86	10	19.79	12	19.79	12	20.32	11	19.82	11	19.93	11	21.46	9
B22	14.14	20	10.75	23	14.18	20	20.25	12	22.00	10	22.86	8	22.11	8.5	21.93	7	22.39	7.5	22.39	7.5	22.18	9	22.11	9	22.25	8	19.92	10
B30	22.21	8	21.46	10	20.68	10	19.21	15	19.04	14	18.61	14	17.93	15	18.68	14	19.39	13	19.39	13	18.86	14	19.39	13	19.93	11	19.58	11
B9	14.36	18.5	12.79	19.5	16.57	15	21.32	9.5	21.29	12	21.18	11.5	20.32	11	20.71	11.5	19.82	10.5	19.82	10.5	20.32	11	19.82	11	19.75	13.5	19.13	12.5
B33	14.36	18.5	12.79	19.5	16.50	16	21.32	9.5	21.29	12	21.18	11.5	20.39	10	20.71	11.5	19.82	10.5	19.82	10.5	20.32	11	19.82	11	19.75	13.5	19.13	12.5
B51	15.04	14.5	16.64	16.5	20.32	11.5	23.86	4	23.68	5	21.50	9	19.68	12	17.93	15	17.00	15	17.00	15	16.86	15	17.11	15	17.07	15	18.90	14
B10	22.64	7	23.32	7	25.36	2	22.21	7	18.18	15	15.07	18	14.89	19	13.86	22	14.14	22	14.14	22	13.96	22	13.96	22	13.96	22	17.33	15
B23	13.96	21	11.86	21	15.68	17	15.11	19	16.71	16	17.36	15	13.71	21	13.86	22	14.14	22	14.14	22	13.96	22	13.96	22	13.96	22	17.33	15
B26	18.82	13	23.00	8	23.96	6	20.29	11	16.18	18	15.89	17	15.29	18	14.71	19	14.71	19	14.71	19	14.64	19.5	14.64	19.5	15.21	19.5	17.17	17
B28	21.39	10	20.11	11	18.75	14	15.96	18	15.68	19	14.11	19.5	15.50	17	15.57	17	16.18	17	16.18	17	16.11	17	15.82	17	16.11	17	16.66	18
B11	13.68	22	7.36	28	8.54	26.5	14.96	20	16.29	17	16.54	16	15.36	17	15.96	16	16.79	16	16.79	16	16.36	16	16.36	16	16.25	16	16.64	19
B17	8.07	28	17.68	14	19.54	13	11.29	22	8.96	25	9.43	24.5	11.21	22	14.32	21	14.36	21	14.36	21	14.14	21	14.00	21	14.64	21	13.12	20
B46	14.39	17	7.46	27	8.54	26.5	14.64	21	15.07	20	14.11	19.5	12.21	23	11.93	23	11.61	23	11.61	23	11.32	23	11.50	23	11.39	23	12.15	21
B25	10.36	25	7.71	26	5.86	28	8.36	26	10.79	22	11.61	21	13.96	20	14.71	20	14.71	20	14.71	20	14.64	19.5	14.64	19.5	15.21	19.5	11.99	22
B15	11.43	23	8.93	25	23.71	7	18.75	17	11.57	21	10.25	22	8.93	26	8.32	26	8.18	26	8.18	26	8.18	26	8.21	26	8.25	26	11.05	23
B16	19.96	11	19.68	12	14.96	19	9.57	25	7.57	26	8.14	26	6.36	27	7.04	27	7.14	27	7.14	27	7.21	27	7.25	27	7.32	27	9.94	24
B43	9.29	26	11.07	22	11.25	23	10.00	23	9.43	23	9.86	23	9.54	24	9.54	24	9.61	24	9.61	24	9.57	24	9.89	24	9.61	24	9.83	25
B5	6.96	29	1.29	31	1.36	31	1.57	31	4.82	27	8.00	27	10.39	23	14.14	19	15.50	18	14.82	18	15.43	18	15.50	18	15.39	18	9.63	26
B20	4.18	30.5	10.21	24	10.86	24	9.64	24	4.18	24	9.43	24.5	9.29	25	9.39	25	9.32	25	9.32	25	9.14	25	9.21	25	9.07	25	9.07	27
B36	10.64	24	16.93	15	8.86	25	5.96	27	4.50	28	4.46	29	5.14	29	5.25	29	4.93	29	4.93	29	4.93	29	4.86	29	4.93	29	6.58	28
B1	22.00	9	27.00	3	12.07	22	5.25	28	2.93	31	2.21	31	2.54	31	1.46	31	1.21	31	1.21	31	1.43	31	1.43	31	1.43	31	6.34	29
B7	4.18	30.5	2.00	30	1.93	30	2.25	30	4.39	29	5.57	28	6.36	28	5.93	28	6.36	28	6.36	28	6.57	28	6.43	28	6.32	28	4.96	30
B6	8.11	27	4.50	29	3.68	29	3.57	29	3.07	30	2.96	30	4.46	30	4.39	30	4.21	30	4.21	30	4.50	30	4.50	30	4.39	30	4.27	31
w	0.506		0.726		0.654		0.641		0.690		0.667		0.578		0.572		0.582		0.582		0.578		0.581		0.568		0.568	
χ^2	212.50		304.82		274.51		269.41		289.80		280.28		242.87		240.12		244.49		242.77		242.77		243.85		238.39		238.39	
p	2.21E-29		2.97E-47		2.64E-41		2.61E-40		2.69E-44		1.97E-42		3.59E-35		1.21E-34		1.75E-35		1.75E-35		3.75E-35		2.32E-35		2.60E-34		2.60E-34	

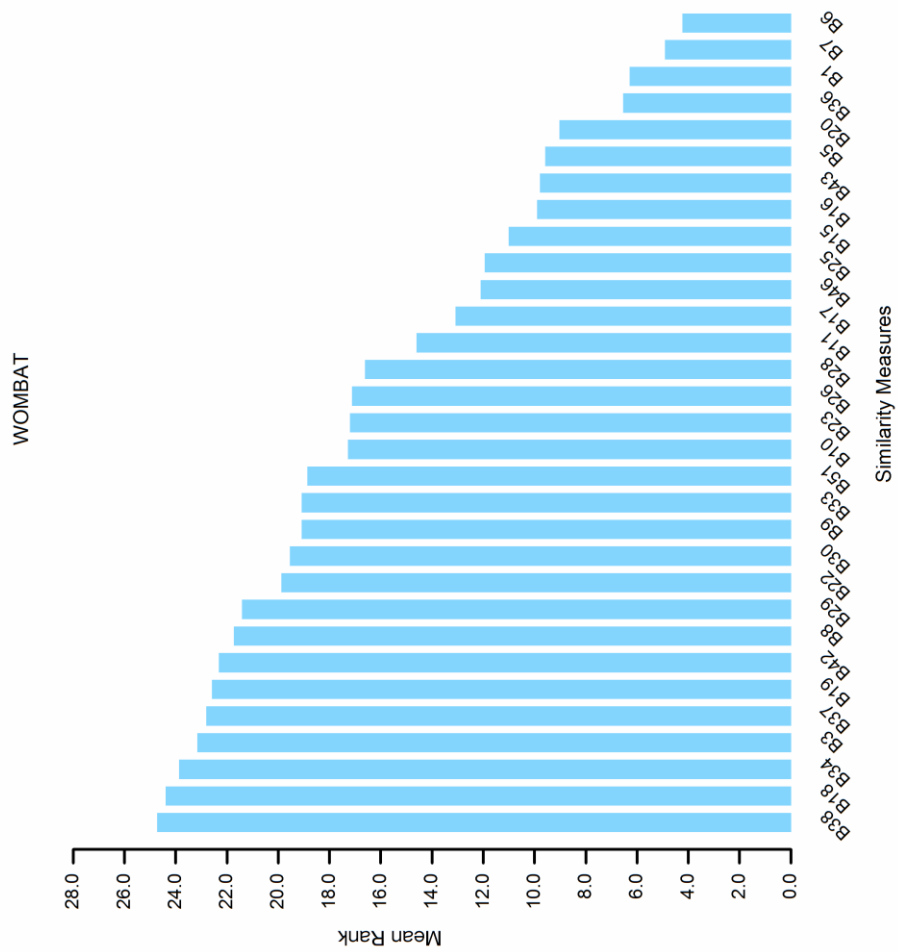


Figure 5-2 Performance of the 31 similarity coefficients, as obtained from WOMBAT dataset, ordered from best (high mean rank values) to worst (low mean rank values)

Table 5-4 Kendall's W results for the top 1% based on the average actives retrieved for ChEMBL dataset. Mean rank indicates the value of mean $EF_{1\%}$ obtained from the $EF_{1\%}$ values averaged over the 15 activity classes in the dataset. For each fingerprint dimension, the highest ranked similarity measure is marked by blue box and the lowest ranked by orange box for ease of reference.

Similarity Measures	Kendall's W Mean Rank for ChEMBL															Average												
	2^5		2^6		2^7		2^8		2^9		2^{10}		2^{11}		2^{12}		2^{13}		2^{14}		2^{15}		2^{16}		2^{17}			
	mean rank	position	mean rank	position	mean rank	position	mean rank	position	mean rank	position	mean rank	position	mean rank	position	mean rank	position	mean rank	position	mean rank	position	mean rank	position	mean rank	position	mean rank	position	mean rank	position
B38	2187	9	2067	7	2597	1	2560	1	2590	2	2693	1	2643	1	2487	3	2577	1	2483	2	2513	2	2617	1	2377	4	2486	1
B42	1763	15	1747	15	2017	11	2273	6	2447	3	2647	3	2530	3	2580	1	2567	2	2623	1	2617	1	2617	1	2617	1	2386	2
B3	1767	13.5	2037	8.5	2417	4	2507	2	2603	3	2483	3	2507	2	2400	5.5	2410	5.5	2377	6	2380	3.5	2273	7	2365	3	2365	3
B18	2353	6	2163	5	2003	12	2123	10	2367	4	2240	8	2417	4	2500	2	2480	3	2403	3.5	2380	3.5	2390	2.5	2331	4	2331	4
B34	2503	4	2177	4	1880	13	2063	11	2233	7	2270	6	2427	6	2440	4	2463	4	2403	3.5	2377	5	2390	2.5	2307	5	2307	5
B37	2443	5	1920	10	1837	14	1900	13	2113	12	2133	9	2350	6	2450	5	2410	5.5	2397	5	2360	6	2370	5	2237	6	2237	6
B19	2510	3	2287	2	1673	17	1760	14	1760	14	1843	15	2137	8	2237	7	2307	7	2330	7	2293	7	2333	6	2140	7	2140	7
B26	1593	20	1917	11	2507	2	2420	3	2177	10	2250	7	2063	12	2153	9	2070	10	2093	11	2097	11.5	2100	10	2119	8	2119	8
B22	1740	17	1580	20	2070	8	2290	5	2313	6	2300	5	2290	7	2170	8	2127	8.5	2130	9	2157	8	2160	8	2112	9	2112	9
B29	2517	2	2180	3	2143	7	2193	7	2153	11	1953	12	1953	13	1847	15	1823	15	1880	15	1853	16	1907	15	2027	10	2027	10
B9	1780	11.5	1737	17	2067	9.5	2243	7.5	2213	8.5	2083	10	1913	14.5	1943	13.5	1847	13.5	1927	13.5	1893	13.5	1923	13.5	1972	11	1972	11
B33	1780	11.5	1737	17	2067	9.5	2243	7.5	2213	8.5	2060	11	2073	10.5	1913	14.5	1847	13.5	1927	13.5	1893	13.5	1923	13.5	1971	12	1971	12
B8	1907	10	1483	21	1353	21	1640	17	1900	13	2387	4	2083	9	1970	12	2180	8	2173	8	2150	9	2073	12	1956	13	1956	13
B30	2347	7	1863	12	1703	16	1740	15	1823	14	1700	16	1747	15	1757	16	1740	16	1783	16	1857	15	1877	17	1825	15	1825	15
B51	1767	13.5	2037	8.5	2410	5	2320	4	2330	5	1863	13	1680	18	1643	18	1607	18	1520	18	1517	18	1480	18	1825	15	1825	15
B17	1090	24	1737	17	2180	6	1523	21	1420	22	1583	18	1717	17	2027	11	2053	11.5	2100	12	2093	11	2100	10	1825	15	1825	15
B10	2570	1	2357	1	2433	3	1940	12	1587	19	1580	19	1450	19	1450	19	1427	19	1430	19	1423	19	1420	19	1741	17	1741	17
B25	720	28	1160	26	1290	22	1280	23	1677	16	1850	14	1837	14	1837	14	1837	14	2067	10	2053	11.5	2093	11	2100	10	1718	18
B23	1297	21	983	28	1073	24	1540	20	1663	17	1603	17	1727	16	1687	17	1703	17	1723	17	1797	17	1887	16	1570	19	1570	19
B28	2310	8	1797	13	1537	18	1577	19	1457	20	1380	21	1380	21	1380	21	1293	21	1370	21	1323	21	1323	20	1486	20	1486	20
B11	1630	19	1337	22.5	1407	19.5	1670	16	1653	18	1573	20	1503	20	1437	20	1390	20	1337	20	1327	20	1317	21	1456	21	1456	21
B46	1650	18	1337	22.5	1407	19.5	1607	18	1433	21	1170	22	1077	22.5	993	23	993	23	980	24	987	24	990	24	1201	22	1201	22
B16	1760	16	2143	6	1283	23	947	24.5	777	26	787	27	783	26	733	26	737	26	787	26	787	26	787	26	1006	23	1006	23
B43	873	26	1290	24	1043	25	947	24.5	893	24	967	23	1017	24	950	24	987	24	957	25	993	23	1003	23	994	24	994	24
B20	160	30.5	1217	25	1010	26	933	26	867	25	953	24	990	25	910	25	960	24	967	25	983	25	987	25	915	25	915	25
B15	963	25	1030	27	1823	15	1350	22	927	23	853	26	740	27	673	27	673	27	620	27	620	27	623	27	883	26	883	26
B5	357	29	273	30	177	31	333	30	720	27	937	26	1077	22.5	1190	22	1243	22	1173	22	1190	22	1250	22	854	27	854	27
B36	1153	23	1773	14	810	27	547	27	463	28	520	28	417	29	437	29	480	28	493	28.5	537	28	527	28	668	28	668	28
B6	793	27	433	29	447	29	413	29	273	30	387	30	310	30	313	30	363	30	443	30	450	30	453	30	422	29	422	29
B1	1157	22	1597	19	737	28	453	28	207	31	177	31	147	31	130	31	160	31	147	31	147	31	147	31	412	30	412	30
B7	160	30.5	207	31	200	30	263	31	347	29	430	29	483	28	460	28	470	29	493	28.5	503	29	497	29	386	31	386	31
w	0.618		0.392		0.550		0.586		0.652		0.618		0.676		0.629		0.645		0.645		0.623		0.614		0.614		0.614	
X ²	277.92		176.50		247.56		263.48		293.60		278.08		304.03		282.96		290.42		290.42		283.60		280.41		276.43		276.43	
p	5.70E-42		1.11E-22		4.47E-36		3.72E-39		4.81E-45		5.31E-42		4.25E-47		5.90E-43		2.05E-44		2.02E-44		4.41E-43		1.86E-42		1.12E-41		1.12E-41	

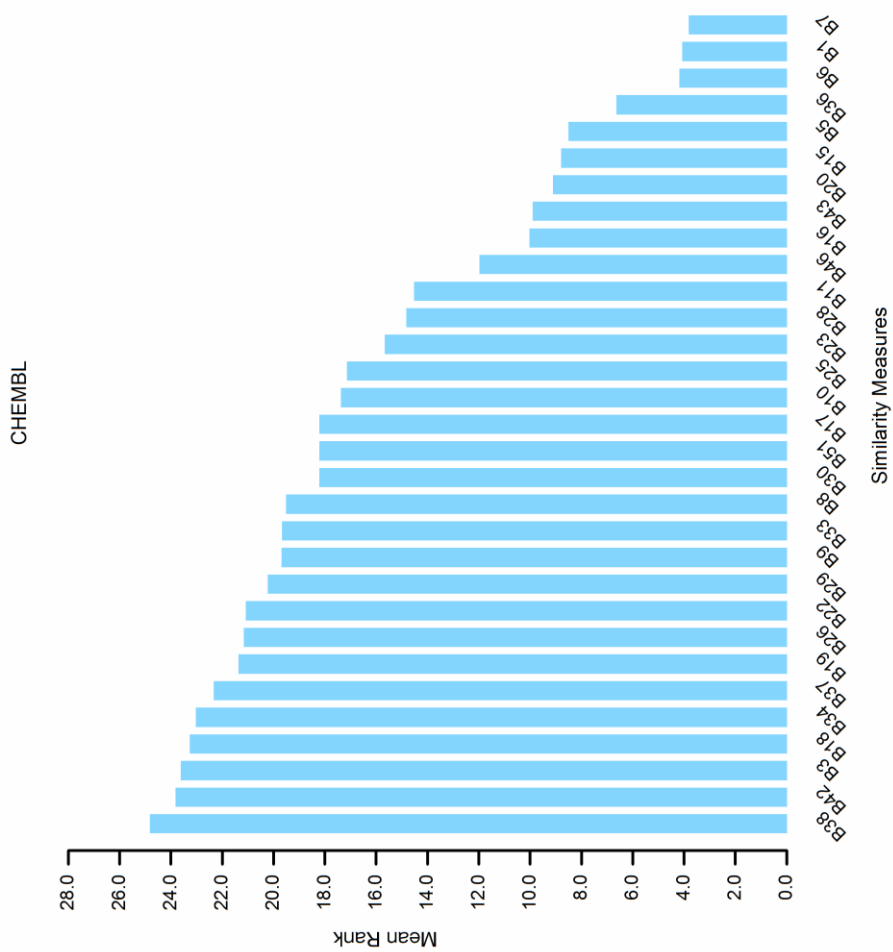


Figure 5-3 Performance of the 31 similarity coefficients, as obtained from ChEMBL dataset, ordered from best (high mean rank values) to worst (low mean rank values)

5.3.3 Effect of Dimensionality on the Effectiveness of Similarity Searching

Figure 5-4 A subset of average enrichment values using top 1% of the ranked dataset in searches for the eleven MDDR activity classes using various Morgan Radius 2 fingerprint dimensions illustrates the effectiveness of similarity searching over the changes of the dimensionality for the MDDR dataset. It presents a subset of effectiveness to show the main trends resulted from the experiments. Detailed values for all results are available in Table 5-5. The enrichment values were averaged over 10 searches for 11 activity classes. There was a significant trend that the effectiveness of similarity searching increases as the dimensionality increases. The effectiveness remains consistent for fingerprint dimensions from 2^{12} until 2^{17} bits. This behaviour was shown by twenty-nine similarity coefficients. It is also interesting to see that there was a slight drop in the effectiveness using two similarity coefficients, i.e., B1 (SM) and B15 (FAI). A similar trend for the similarity search results using the WOMBAT and ChEMBL datasets can be found in Appendix A (results in Table A-1 are illustrated by Figure A-1 for WOMBAT dataset and Table A-2 by Figure A-2 for ChEMBL dataset).

In general, the observed behaviour showed that changing the dimensionality of the Morgan R2 fingerprint did not suffer from the curse of dimensionality. However, the results indicate that the effectiveness maybe affected by the similarity coefficients.

Further analysis was carried out to investigate the reasons that contribute to the trends. This chapter will first discuss the increase effects followed by the decrease effects and the consistent effects that were obtained as the dimension increases. This was made either by: (i) investigating the characteristic of the molecule in the dataset that contribute to such effects, (ii) investigating the formulation of the similarity coefficients or (iii) analysing the bit collision in the datasets.

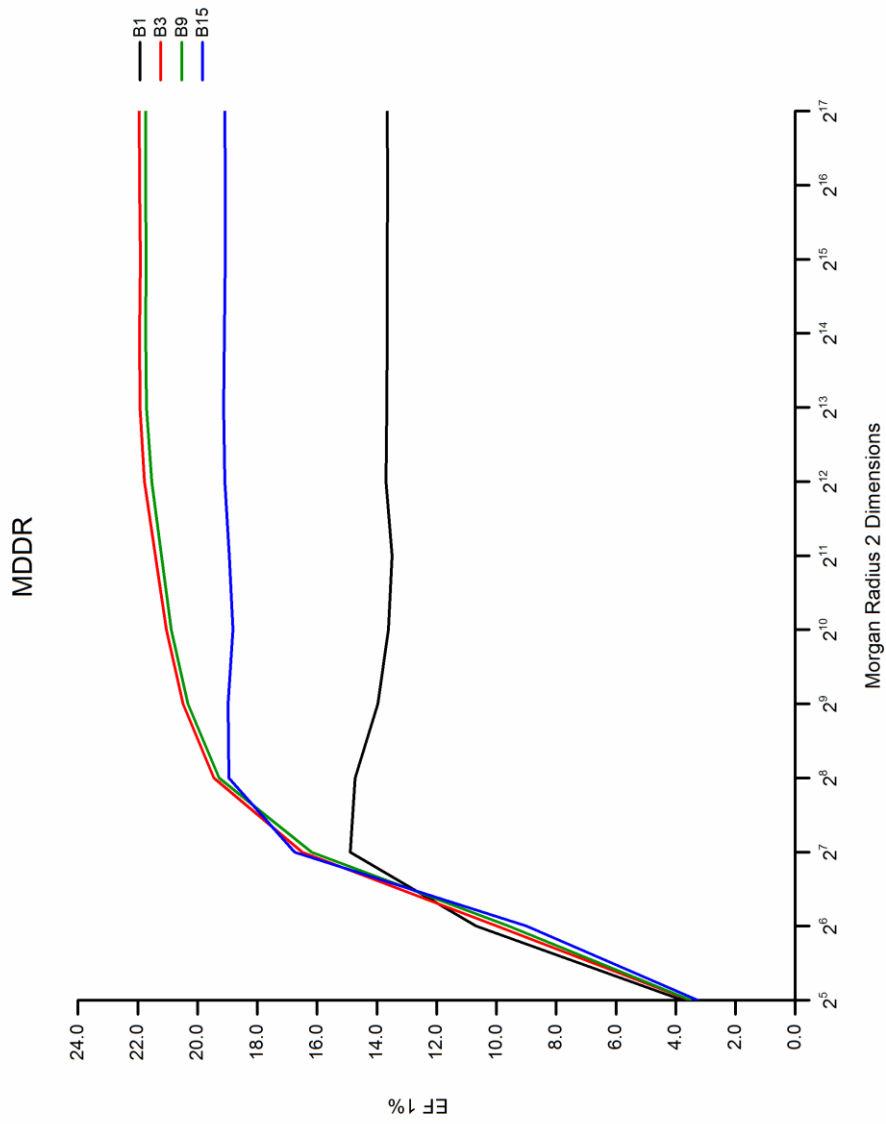


Figure 5-4 A subset of average enrichment values using top 1% of the ranked dataset in searches for the eleven MDDR activity classes using various Morgan Radius 2 fingerprint dimensions

Table 5-5 Average enrichment values using top 1% of the ranked dataset in searches for the eleven MDDR activity classes using various Morgan R2 fingerprint dimensions. For each fingerprint dimension, the highest average enrichment value is marked by green colour and the lowest value by red colour for ease of reference

No	Similarity Coefficients	Morgan R2 Dimensions MDDR - EF 1%														
		2 ⁵	2 ⁶	2 ⁷	2 ⁸	2 ⁹	2 ¹⁰	2 ¹¹	2 ¹²	2 ¹³	2 ¹⁴	2 ¹⁵	2 ¹⁶	2 ¹⁷		
1	B1	3.77	10.67	14.89	14.73	13.97	13.62	13.49	13.70	13.67	13.66	13.65	13.64	13.65		
2	B3	3.52	9.98	16.49	19.45	20.49	19.45	20.49	21.78	21.93	21.93	21.91	21.95	21.95		
3	B5	1.83	3.38	8.00	13.41	17.21	13.41	17.21	19.05	19.97	20.38	20.69	20.69	20.70		
4	B6	2.21	4.76	9.21	12.27	13.80	12.27	13.80	14.81	15.45	16.05	16.16	16.18	16.21		
5	B7	1.10	3.25	7.96	12.56	15.11	12.56	15.11	16.45	17.05	17.70	17.86	17.83	17.89		
6	B8	3.84	10.07	15.97	18.88	20.18	18.88	20.18	20.95	21.56	21.95	22.03	22.10	22.11		
7	B9	3.49	9.59	16.18	19.27	20.32	19.27	20.32	20.89	21.20	21.54	21.71	21.74	21.73		
8	B10	3.89	10.71	16.58	18.93	19.59	18.93	19.59	20.58	20.26	20.57	20.74	20.75	20.68		
9	B11	3.43	9.04	15.58	18.91	19.98	18.91	19.98	20.58	20.89	21.28	21.46	21.44	21.41		
10	B15	3.26	8.99	16.74	18.95	18.97	18.95	18.97	18.82	18.94	19.09	19.12	19.10	19.08		
11	B16	3.45	10.20	15.60	17.02	17.47	17.02	17.47	17.84	18.22	18.63	18.77	18.74	18.75		
12	B17	2.16	7.72	16.30	18.21	18.81	18.21	18.81	19.57	20.27	20.72	21.03	21.07	21.09		
13	B18	3.91	10.36	16.22	19.05	20.23	19.05	20.23	20.88	21.30	21.72	21.90	21.93	21.95		
14	B19	3.89	10.53	16.02	18.65	19.81	18.65	19.81	20.63	21.18	21.66	21.85	21.92	21.95		
15	B20	1.10	8.76	14.60	16.99	17.92	16.99	17.92	18.49	18.92	19.39	19.50	19.45	19.50		
16	B22	3.48	9.48	16.11	19.29	20.38	19.29	20.38	20.92	21.26	21.60	21.78	21.80	21.77		
17	B23	2.78	7.19	14.09	17.92	19.56	17.92	19.56	20.49	21.02	21.48	21.67	21.74	21.73		
18	B25	2.26	8.39	14.85	17.92	19.29	17.92	19.29	19.92	20.37	20.75	21.06	21.10	21.13		
19	B26	3.06	9.15	16.38	19.35	20.09	19.35	20.09	20.39	20.63	20.79	21.06	21.10	21.13		
20	B28	3.60	9.86	15.75	18.56	19.70	18.56	19.70	20.41	20.79	21.22	21.44	21.43	21.41		
21	B29	3.90	10.03	16.18	19.10	20.19	19.10	20.19	20.81	21.19	21.53	21.71	21.74	21.73		
22	B30	3.89	9.87	15.79	18.74	19.91	18.74	19.91	20.70	21.11	21.51	21.68	21.74	21.73		
23	B33	3.49	9.59	16.18	19.27	20.32	19.27	20.32	20.89	21.20	21.54	21.71	21.74	21.73		
24	B34	3.93	10.22	16.10	18.98	20.19	18.98	20.19	20.87	21.30	21.71	21.89	21.93	21.95		
25	B36	2.77	9.98	14.28	15.22	15.30	15.22	15.30	15.53	15.82	16.23	16.25	16.24	16.26		
26	B37	3.90	9.89	15.85	18.85	20.11	18.85	20.11	20.83	21.29	21.70	21.88	21.92	21.95		
27	B38	3.86	9.82	16.51	19.46	20.58	19.46	20.58	21.17	21.51	21.83	21.94	21.93	21.96		
28	B42	3.50	9.59	15.94	19.27	20.57	19.27	20.57	21.21	21.65	22.01	22.16	22.18	22.19		
29	B43	2.49	8.78	14.60	16.99	17.92	16.99	17.92	18.50	18.92	19.39	19.49	19.45	19.51		
30	B46	3.43	9.04	15.59	18.89	19.93	18.89	19.93	20.46	20.71	21.02	21.18	21.15	21.12		
31	B51	3.52	9.98	16.49	19.44	20.42	19.44	20.42	20.91	21.20	21.51	21.63	21.62	21.62		
Average		3.18	9.00	15.07	17.89	18.98	17.89	18.98	19.60	20.00	20.39	20.54	20.56	20.57		

5.3.3.1 Observation of Retrieved Compounds

In the literature, the increase of search performance in high dimensionality has been reported to be associated with the intrinsic (“fractal”) dimensionality of the data, not the dimensionality of the address space (Korn et al., 2001). In relation to the chemical data, fractals in the form of iterated substructures or fragments exist in a compound. Compounds that have similar bioactivity are also likely to have similar fractals (substructures) exhibited in the compounds (Johnson & Maggiora, 1990). Based on these relations, it would be useful to conduct an investigation on the molecular intrinsic structure to explain the characteristic of the chemical data that contribute to the increasing trend.

The ECFP₄-like (Morgan R2) fingerprints did not give any direct information about the structure of the molecule. This can be because of a few reasons. First, it is not possible to directly decode the integer identifiers (and the bits) of the ECFPs to a particular feature that it represents. Second, the relationship between the bit fingerprint and the molecule structure may not always be one-to-one during the generation of the ECFP fingerprints. Hence, it is difficult to identify the structures by analysing the bit fingerprint based on the bit position (Rogers & Hahn, 2010).

There are however, other ways to identify the similar fractals in a compound. That is using the SMILES representation or the molecular scaffold. In chemoinformatics, the Murcko scaffold has been used to define the frameworks of a molecule (Bemis & Murcko, 1996). It can also be used to find the common features present in molecules. Thus, for this reason we will investigate the Murcko scaffold of the molecules to identify the characteristics of the chemical data.

A few examples of molecules have been chosen to be analysed. These molecules are the active molecules retrieved from the $EF_{1\%}$ resulting from the similarity search using a single reference. The similarity of these molecules was measured using the B3 (JT) coefficient. The B3 coefficient has been chosen as an example of similarity measure based on three reasons: (i) it shows a resemblance of

increasing effectiveness, (ii) it ranks the highest in the MDDR and WOMBAT datasets for the commonly used 1024 bits fingerprint and (iii) it is the most effective similarity measure in the literature.

The identification of similar features that exist in each increasing dimension was conducted. For the first dimension (i.e., 2^5 bits), the active molecules retrieved and the distinct scaffolds of the active molecules retrieved were recorded. Next, we identified the new active molecules retrieved for the next dimension (i.e., 2^6 bits). These are the new actives which were retrieved using the 2^6 bits but not retrieved when searched using the 2^5 bits. The distinct scaffolds of the new active molecules retrieved were compared with the distinct scaffolds of the previous active molecules retrieved. The number of similar scaffolds was recorded, i.e., identical scaffolds that exist in the active molecules retrieved in the previous and current dimensions. The process was continued for the next following dimension.

Table 5-6 Identification of identical scaffold based on the active molecules retrieved using a single reference from the Renin activity class of the MDDR dataset

No.	Morgan R2 Dimensions	Number of Actives Retrieved	Number of Scaffolds	Number of New Actives Retrieved	Number of New Actives Retrieved with Identical Scaffold
1	2^5	160	120	-	-
2	2^6	565	318	431	32
3	2^7	769	392	236	38
4	2^8	798	402	88	14
5	2^9	798	404	46	17
6	2^{10}	777	389	25	9
7	2^{11}	790	393	21	12
8	2^{12}	795	394	15	7
9	2^{13}	796	395	7	2
10	2^{14}	797	396	3	1
11	2^{15}	797	396	1	1
12	2^{16}	799	397	2	1
13	2^{17}	799	397	0	0

The result, as shown in Table 5-6, indicates that identical scaffolds to the previous active molecules retrieved exist in the new active molecules retrieved for each increasing dimension. For example, one of the two new active molecules retrieved in the higher dimension (2^{16} bits) has the identical scaffold

to the existing active molecules retrieved in the lower dimension (i.e., 2^{15} bits). Figure 5-5 illustrates an example of identical scaffolds that have been found. It provides the illustrations of the original molecule and its Murcko scaffold. The first two rows are the existing active molecules retrieved using the 2^{15} dimension. The third row is the new active molecule retrieved using the 2^{16} dimension.

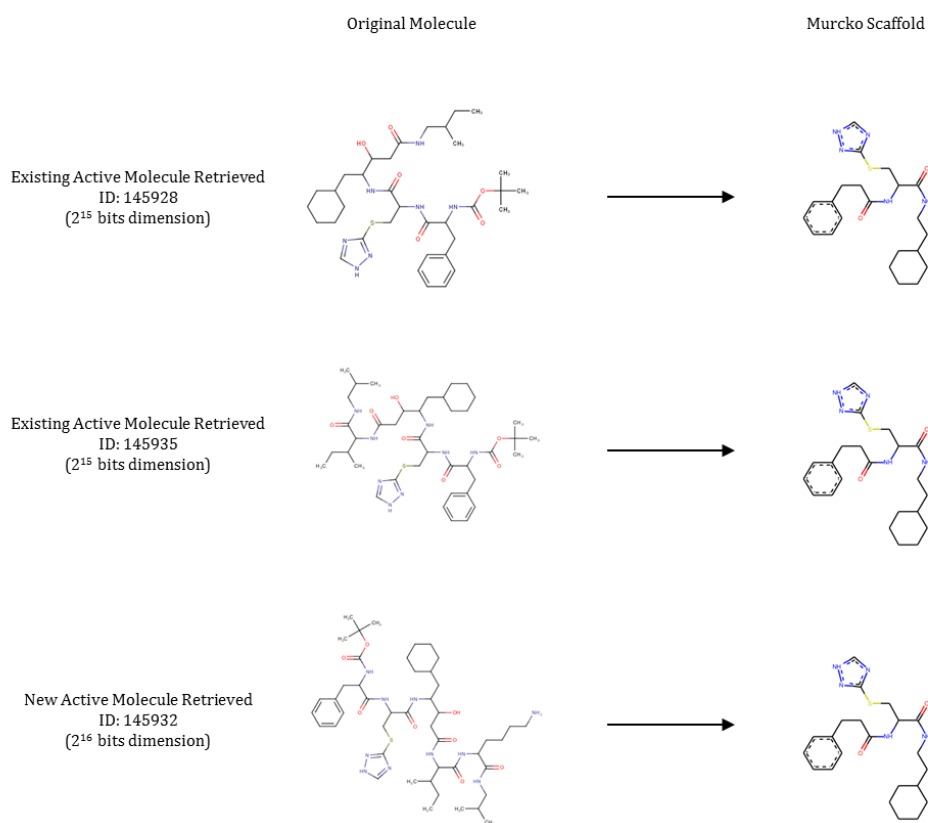


Figure 5-5 Identification of identical scaffold using Murcko scaffold between the existing active molecules retrieved in a lower dimension and new active molecule retrieved in a higher dimension

As observed, the new retrieved molecule has an identical scaffold to the other two existing retrieved molecules. These scaffolds can be used to represent the intrinsic feature (substructure) of the molecules. There is, however, a single exception in the last dimension, i.e., 2^{17} . This is because the active molecules retrieved were the identical active molecules retrieved in the previous dimension, i.e., 2^{16} . Thus, there is no new active molecule retrieved to be analysed.

It is also worth to mention that there was a loss of the active molecules retrieved when searched using a higher dimension. These are the actives which were retrieved using the lower dimension but not retrieved when searched using the higher dimension. For example, there were 798 active molecules retrieved using the 2^9 dimension and 777 active molecules retrieved using the 2^{10} dimension. This indicates that several active molecules were not retrieved even when the dimension has been increased. These findings may show a possible behaviour of clumping effect in the database due to the analogous of molecular scaffolds. However, the interpretation cannot be extrapolated to all dimensions as the similar behaviour was not observed in a higher dimension.

Taken together, these results suggest that there is an association between the increases of search performance with the intrinsic dimensionality of the data. The nearest neighbour search in high dimensions can still be effective for a chemical dataset if the molecules have similar intrinsic features (structures).

However, these findings do not show the occurrence of the curse of dimensionality. It is possible, therefore, that this outcome is contrary to the curse of dimensionality as no evidence of decrease in the performance of high dimensionality was detected.

5.3.3.2 Effect of Similarity Coefficient

The next discussion on the decrease trends involves the understanding of the global and local similarity. Hence, it is worth explaining about the global and local similarity before discussing about the results. In general, *global similarity* measures the similarity of two objects using the complete vectors (i.e., the object representations). In contrast, *local similarity* measures the similarity of two objects by looking for the best internal matching region between the two vectors. In the former case, the similarity indicates the total percentage of match while the latter indicates the percentage matches of the internal region.

The review by Maggiora et al. (2014) interpret and provide examples of global and local similarities in molecular similarity. The computation of global similarity is generally derived from structural information associated with the

entire compounds. On the other hand, the local similarity focuses only on selected fragments or functionalities of the molecules. In relation to this experiment, the global similarity measures the similarity of two molecules associated with the entire fingerprint whilst the local similarity focuses only on selected bits in the fingerprint.

When focusing on the decreasing trends, it can be seen from Table 5-5 that the B1 (SM) coefficient resulted in a decreased effect starting at 2^{10} bits for the MDDR dataset. A similar observation can be found using the WOMBAT dataset in Table A-1 (Appendix A). The decreasing effect for the ChEMBL dataset using the similar coefficient starts from 2^9 bits as shown in Table A-2 (Appendix A). This coefficient has also resulted in the lowest $\overline{EF}_{1\%}$ value for the last eight fingerprint dimensions in the MDDR and WOMBAT datasets, i.e., 2^{10} until 2^{17} as compared to the other coefficients. For the ChEMBL dataset, the B1 coefficient has also resulted in the lowest $\overline{EF}_{1\%}$ value for the last nine fingerprint dimensions, i.e., 2^9 until 2^{17} fingerprint dimensions. This is in the agreement with the previous study, which ranked B1 coefficient among the lowest rank of similarity coefficient to be used (Todeschini et al., 2012).

The B1 coefficient is measured according to the following formulation in Eq. (29):

$$S_{SM} = \frac{a + d}{p} \quad (29)$$

where S_{SM} is the similarity value, a is the number of common bits set, d is the number of common bits unset and p is the total bits size (dimension). This coefficient has the components a and d in its numerator and denominator, which means it compares the number of matching bits (both set and unset) with the entire possible bits dimension. This also means that it evaluates the similarity between two molecules based on their similarity relative to the possible whole dimensions (i.e., global similarity). This is different to evaluate the similarity relative to the internal matching features (i.e., local similarity) which is effectively measured by the other coefficients, e.g., B3 (Jaccard-

Tanimoto) and B9 (Cosine). As shown in Table 5-5, Table A-1 and Table A-2, these coefficients have resulted in increasing effectiveness in similarity searching using all three datasets, correspondingly.

In this study, the results using the B1 coefficient did show a minor resemblance to the curse of dimensionality. There is however, a possible explanation for this effect. As the dimensionality increases, the distribution of the data becomes increasingly sparse with the increasing number of zero attributes, i.e., $d \rightarrow p$. As a result, a global similarity between two molecules can be increased and approaches to unity because of the existence of zero attributes.

In relation to the virtual screening experiment, it is possible for an inactive molecule to be measured more similar to the reference molecule if it has a larger number of common zero bits, i.e., bits unset (due to the sparsity) although it was structurally different. As a result, the inactive molecules will be ranked to the top of the ranking while the active molecules were not. This could probably be the reason why there were less active molecules retrieved as the dimensionality increases hence the decreases of the effectiveness of similarity searching.

To illustrate this effect we show in Figure 5-6 three molecules which were measured by the B1 coefficient in this experiment. The similarity value (S_{SM}), number of common bits set (a), number of common bits unset (d), number of total bits (p) and the similarity ranking between the molecules are also shown. The inactive molecule has a larger similarity value as compared to the active molecule. As a result, the inactive molecule is ranked higher than the active molecule. One possible reason is because it has more common unset bits (d) which can increase the similarity value when measured using the B1 coefficient.

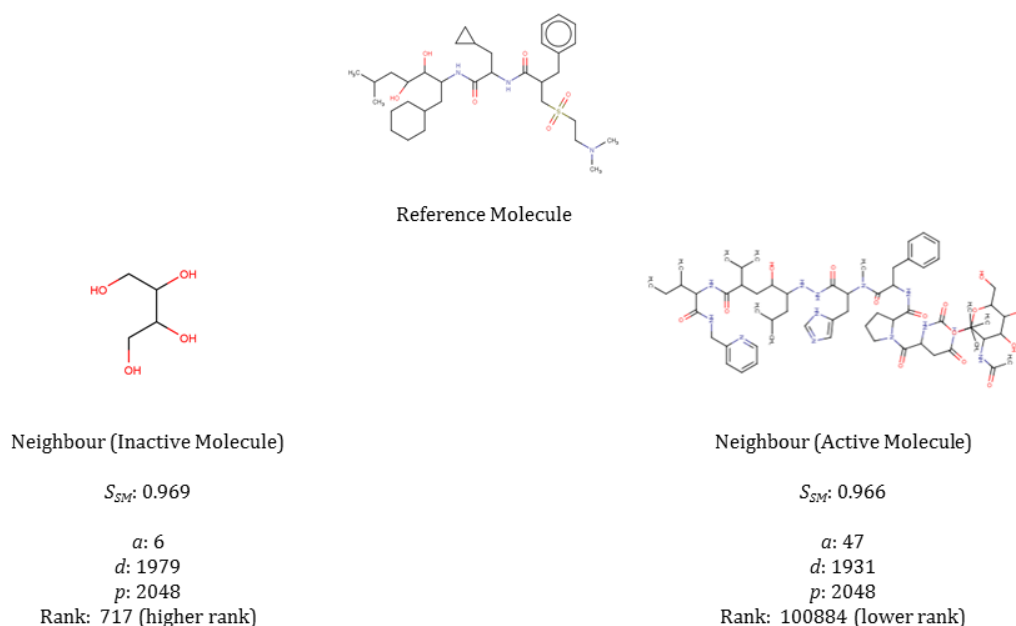


Figure 5-6 A comparison of the Simple Matching similarity values for two molecules (inactive and active) to illustrate the effect of global similarity measure

There is also another similarity coefficient which has shown a similar result to the B1 coefficient, i.e., B15 (FAI). The possible reason for this behaviour is because of the formulation of the coefficient. The B15 coefficient is measured according to the following formulation in Eq. (30):

$$S_{FAI} = \frac{a + 0.5d}{p} \quad (30)$$

The formulation of this coefficient only differs in terms of the weighting of the component d as compared with the formulation of the B1 coefficient, i.e., equal to half of the number of common unset bits. However, as the dimensionality increases, the inactive molecules which have more zero bits will possibly still be ranked higher as compared to the active molecules. This is because the coefficient is still measuring the similarity associated with the whole dimension. Hence, this produced similar trends of reduced effectiveness that can be observed in Figure 5-4 for the MDDR dataset, and for WOMBAT and ChEMBL in Appendix A (Figure A-1 and Figure A-2, correspondingly). The other two

coefficients, i.e., B35 (PE1) and B45 (HAM) which have similar formulation to the B1 coefficient have been excluded. This is because they were monotonic to the B1 coefficient as listed in Table 5-1.

5.3.3.3 Effect of Fingerprint's Bit Collision

Finally, we further investigate the constant effects starting with the 2^{11} bits size. This is done by measuring the average bit collisions of all references used in this experiment, across all dimensions. In general, the number of bits set will increase with the size of the addressable space until there are no collisions. The bit collision is calculated as follows in Eq. (31):

$$\text{Bit collision rate}_i = \bar{x}_i - \bar{x}_{i-1} \quad (31)$$

$$i = \{2^5, 2^6, 2^7, 2^8, 2^9, 2^{10}, 2^{11}, 2^{12}, 2^{13}, 2^{14}, 2^{15}, 2^{16}, 2^{17}\}$$

$$x = \text{set bits}$$

Table 5-7 shows the bits set, average bits set and average bit collisions calculated from the MDDR dataset. A higher value of bit collision rate indicates a higher bit collision in the particular fingerprint dimension and vice versa. As can be seen, more collisions were particularly apparent for fingerprint sizes of 2^5 until 2^{10} bits. There were almost zero bit collisions for fingerprint sizes of 2^{11} until 2^{16} bits, and zero bit collisions for 2^{17} bits fingerprint. These results suggest that 2^{17} bits is large enough to ensure that, in most cases, there will be no collision occurring and even 2^{12} bits have very few collisions. This result is almost similar to the other two datasets used in this experiment (Table A-3 and Table A-4 in Appendix A).

Table 5-7 Average bits set and bit collision rate based on the average of 10 molecules for MDDR dataset using various Morgan R2 fingerprint dimensions

Activity Class [Average 10 Reference]	Morgan R2 Fingerprint Dimension (MDDR)																
	2^5	2^6	2^7	2^8	2^9	2^{10}	2^{11}	2^{12}	2^{13}	2^{14}	2^{15}	2^{16}	2^{17}				
5HT	25.00	32.00	36.00	40.00	44.00	45.00	45.00	45.00	45.00	45.00	45.00	45.00	45.00	45.00	45.00	45.00	45.00
5HT1A	25.70	33.30	39.40	44.80	46.50	47.60	47.90	47.90	48.10	48.20	48.20	48.20	48.20	48.20	48.20	48.20	48.20
5HT3	25.70	34.50	39.50	43.50	45.40	46.00	46.60	46.60	46.60	46.60	46.60	46.60	46.60	46.60	46.60	46.60	46.60
AT1	26.70	36.90	46.10	52.00	54.70	57.10	57.90	58.90	59.00	59.50	59.60	59.60	59.60	59.60	59.60	59.60	59.60
COX	24.50	32.10	37.70	41.20	42.70	43.40	44.00	44.40	44.50	44.50	44.60	44.70	44.70	44.70	44.70	44.70	44.70
D2	24.40	35.00	42.70	47.80	49.50	50.10	50.60	50.80	50.90	50.90	50.90	50.90	50.90	50.90	50.90	50.90	50.90
HIVP	28.30	40.50	51.20	57.10	59.90	61.90	62.60	63.80	63.90	63.90	63.90	63.90	63.90	63.90	63.90	63.90	63.90
PKC	24.30	33.50	39.20	43.80	45.50	46.30	46.80	46.90	47.10	47.10	47.10	47.10	47.10	47.10	47.10	47.10	47.10
Renin	28.60	43.00	54.20	61.90	67.00	69.50	70.80	72.10	72.20	72.30	72.30	72.30	72.30	72.30	72.30	72.30	72.30
SubP	27.10	38.80	48.50	53.80	56.40	57.80	58.30	58.40	58.40	58.50	58.50	58.50	58.50	58.50	58.50	58.50	58.50
Thrombin	28.20	41.10	51.90	58.60	62.60	64.90	65.80	66.40	66.70	66.90	66.90	66.90	66.90	66.90	66.90	66.90	66.90
Average	26.23	36.43	44.22	49.50	52.20	53.60	54.21	54.65	54.76	54.85	54.87	54.88	54.88	54.88	54.88	54.88	54.88
Bit Collision Rate		10.20	7.79	5.28	2.70	1.40	0.61	0.45	0.11	0.09	0.02	0.01	0.01	0.01	0.01	0.01	0.00

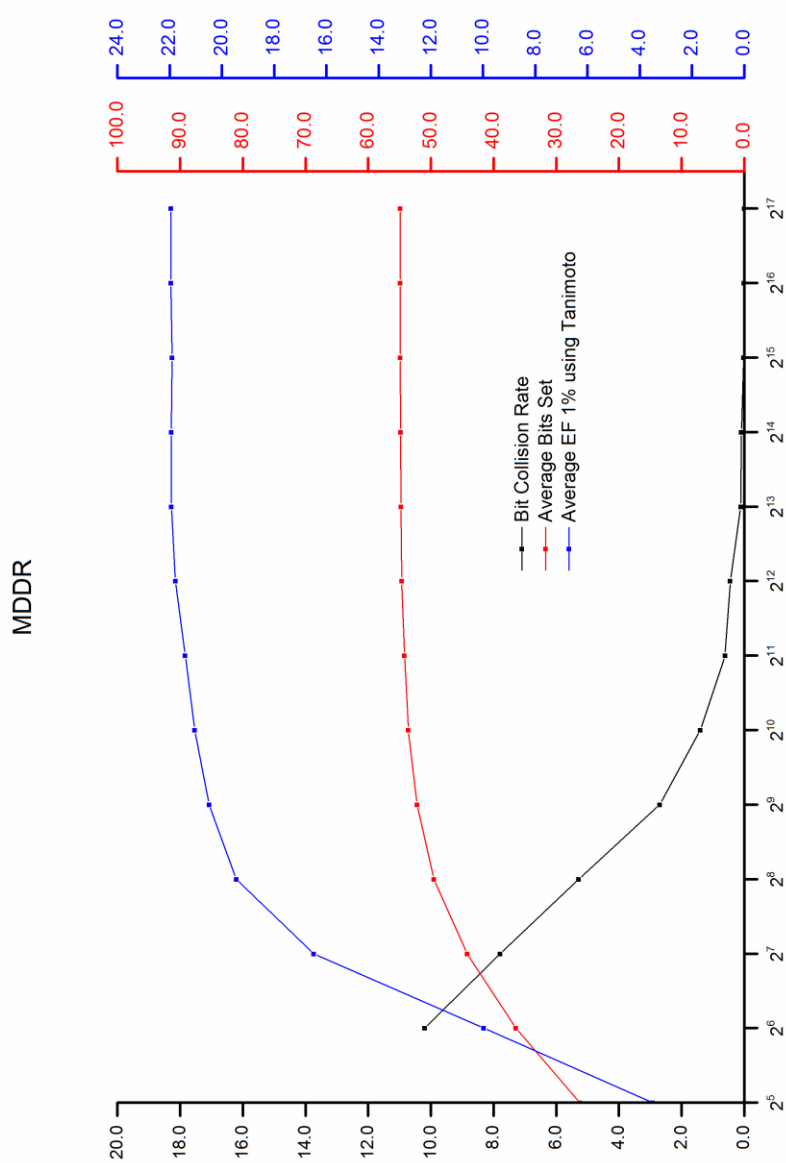


Figure 5-7 Line plot measuring the average bits set, average enrichment curves and bit collision rate based on the average of 10 random molecules for MDDR dataset using various Morgan R2 fingerprint dimensions

It is further shown that the effect of the bit collisions and bits set to the similarity search values. Figure 5-7 shows the effect of the addressable bit space (fingerprint dimensions) on the $\overline{EF}_{1\%}$ across all 11 activity classes in the MDDR dataset using the B3 (JT) similarity coefficient. As can be seen, there were constant effects to the $\overline{EF}_{1\%}$ from 2^{11} bits fingerprint until the final dimensions. A possible reason for this is because of the similar number and the position of bits set starting from the 2^{11} bit fingerprint. Hence, the similarity value measured will also be the same. A similar trend can also be observed in Figure A-3 and Figure A-4 for the WOMBAT and ChEMBL datasets.

5.4 Conclusion

This chapter investigates the effect of changing the dimensionality of molecular representations on the effectiveness of virtual screening based similarity search applications. Overall, the results suggest that the effectiveness of the chemical search was not affected by the curse of the dimensionality phenomenon. The effect of changing the dimension related to two possible reasons: (i) the molecular representation and (ii) the formulation of the similarity coefficient.

First, the use of Morgan R2 fingerprint as the molecular representation does not decrease the effectiveness of the similarity search application. As defined in Chapter 4, the Morgan R2 representation encodes the connectivity invariants of circular atom environments for a molecule up to two bond radius from its central atom. The fingerprints were then folded into certain bit dimensions. At a certain number of bits, increasing the fingerprint dimensions only increases the bit spaces to describe the information of a molecule. The information captured however, is limited by the function of the Morgan algorithm, which is two bond radii in the case of the study. This was supported by the analysis of the bit collisions in Section 5.3.3. The analysis showed the possible number of bits required to capture the information of a sample of molecules used in this study and its relation with the effectiveness of the similarity search application. Other molecular representations or descriptors may have different effects on the performance of the similarity search application. The physicochemical descriptors for example, capture different properties of a molecule. The use of

high dimensionality of physicochemical descriptors may have a different effect to the performance of the similarity search application.

Second, the effectiveness of the similarity search application increased as the dimensionality increases when measured by the similarity coefficients tested in this experiment. The only exception is when the similarity is measured by the global similarity coefficient, which measures the similarity of the molecules associated with the entire fingerprint, i.e., whole dimensions. As discussed in Section 5.3.3, as the dimensionality increases, the distribution of the data becomes increasingly sparse with the increasing number of zero attributes. Hence, the number of zero attributes will affect the global similarity measure of the molecules in high dimensionality fingerprint representation.

The above conclusion was made based on the experimental work for the similarity search application. The following chapter will describe the effect of dimensionality on the effectiveness of other virtual screening applications. The study will allow the investigation and conclusion to be made on other common types of virtual screening applications, i.e., molecular clustering.

Chapter 6 Investigation into the Effect of Dimensionality on the Effectiveness of Clustering

6.1 Introduction

Clustering the molecular structures in a chemical database provides a way of identifying and viewing the groups that are present in a chemical dataset. Clustering helps to save costs and rationalise the basis for molecular biological testing. A representative molecule of a cluster is selected for the biological testing. If the representative proves to be bioactive, then the other molecules in the same cluster will be tested. But if the representative is not bioactive, then the other molecules in the same cluster will be disregarded from the biological testing (Willett, 1987; Downs & Willett, 1994; Downs & Barnard, 2002; MacCuish & MacCuish, 2014).

The clustering procedure involves grouping molecules based on their distance, i.e., closest molecules (as most similar) will be grouped together. The pairwise distance approximations between the molecules can be measured using various distance coefficients. One of the most commonly used coefficients is the Euclidean distance, which measures the straight line distance between two molecules. The other common coefficient is the City Block (or Manhattan) distance that measures the distance in x and the distance in y in the xy coordinates. This is similar when moving in a city where one has to move around the buildings instead of moving straight through the buildings to reach the destination.

Different clustering methods require different types of distance (or similarity) coefficients to measure the distances (or similarity) between molecules. Therefore, in the chemoinformatics domain, many studies have been conducted using different types of coefficients depending on the clustering algorithms, and also on different types of clustering method (Downs et al., 1994; Brown & Martin, 1996; Bayada et al., 1999; Chu et al., 2012; Gan et al., 2014).

The effects of high dimensional data and distance coefficient on document clustering have been studied by France et al. (2012). These authors found that

increased dimensionality aids the clustering performance dependent upon the particular dataset being examined. The study also reported that different effects on the clustering performances were obtained using different distance coefficients.

In the chemoinformatics context, many virtual screening applications have been successfully conducted even though the molecules are represented by very high dimensional representations (Willett, 2011). The cluster application, in particular, is a method that can be used with high dimensionality descriptors such as the binary fingerprint. However, the effect of the application performance using high dimensional data has not yet been investigated. Furthermore, as far as the research in chemoinformatics is concerned, there is no work carried out on the effect of high dimensionality in the effectiveness of the molecular clustering application.

This chapter will investigate the effect of changing the dimensionality of molecular representations on the effectiveness of the molecular clustering applications. The purpose is to test the hypothesis that as the dimensionality increases, the effectiveness of the application decreases. The aim of this study is to identify the characteristics of chemical datasets that contribute to the effectiveness of the molecular clustering application in high dimensionality. It also aims to explain the observed performances using various molecular dimensions and distance coefficients, which simulate a practical clustering procedure.

6.2 Experimental Design

The experiments were carried out to replicate the clustering application, which calculates the distance between all possible pairs of molecules in the dataset. These distance proximities, which were measured by various distance coefficients were used to build an agglomerative hierarchical non-overlapping clustering. In a virtual screening application, a representative molecule of the cluster will then be selected as a sample for the biological testing. These experiments were carried out for subsets of data from two datasets, i.e., MDDR

and WOMBAT. These datasets have been introduced in Chapter 4 including the molecular clustering procedures.

Similar molecular representation in Chapter 5 has been used in this study. Each compound in the datasets was represented using the binary fingerprint, i.e., ECFP_4-like (MorganR2) fingerprint, and folded into thirteen different fingerprint sizes as introduced in Chapter 4.

Ten distance coefficients were used to measure the pairwise distances of the compounds, which allow observations on various clustering performance using different distance coefficients. These coefficients have been introduced in Chapter 4 and listed in Table 4-6.

6.2.1 Clustering Method

Chapter 4 has introduced the two clustering methods used in this study, i.e., Ward's and Group Average algorithms. The following steps summarise the clustering procedures applied to this experiment:

Summary of clustering procedure.

- Step 1: Each molecule, \mathbf{x} , is assigned a class label, l_k , identifying its activity class. The set of all labels for a database \mathfrak{a} is $= \{l_1, \dots, l_k\}$, where k is the number of activity classes. For example, the MDDR dataset used in this experiment has 11 activity classes. Hence, the set of all labels for the database \mathfrak{a} is $= \{l_1, \dots, l_{11}\}$. A similar procedure was performed for the WOMBAT dataset which has 14 activity classes, yielding a set of labels $\mathfrak{a} = \{l_1, \dots, l_{14}\}$.
- Step 2: Each molecule, \mathbf{x} , is converted into a specific type and length of fingerprint representation, i.e., Morgan R2. The fingerprint consists of a binary vector of n dimensions: $\mathbf{x} = (x_1, \dots, x_n)$.
- Step 3: The pairwise distance matrix of all possible pairs of molecules in the database is measured using the ten distance coefficients listed in Table 4.6. This procedure was repeated for each fingerprint dimension.
- Step 4: The closest molecules were clustered based on the chosen clustering method. The clustering is repeated until there is only a single cluster. This procedure was repeated for each fingerprint dimension.
- Step 5: The generated cluster for each fingerprint dimension was analysed and evaluated using two evaluation methods that were introduced in Chapter 4.
-

In terms of computational resources, Ward's agglomerative hierarchical algorithm consumes more computational resources compared to the non-hierarchical clustering methods. For N molecules in a dataset, the stored-matrix algorithm for the procedure requires storage (or memory) space proportional to N^2 , which is written as " $O(N^2)$ ", and the time to perform the clustering is proportional to N^3 ($O(N^3)$). This becomes a severe restriction if the algorithm is to be implemented on large data sets.

Due to the computer intensive calculations, the Ward's procedures in the current experiment were implemented on the Sheffield Advanced Research Computer (ShARC) cluster of the University of Sheffield. The high performance computing was developed and managed by the Research Software Engineering Group, Faculty of Engineering of the University of Sheffield. Figure 6-1 shows the general workflow of cluster implementation using ShARC. Each job contains a batch script of single or task array jobs that requests the high performance computing's scheduler for CPU and execution time resources, job notification configuration and user environment creation, which install specific modules and libraries for the implementation (Figure 6-1). The application was coded using the Python language and the hierarchical clustering package from SciPy has been used to generate the Ward's clustering (Jones et al., 2001).

The performance of the ShARC implementation has been recorded. Figure 6-2 shows the example of performance based on CPU memory and time usage when used to cluster the dataset in this experiment that contains 10,254 molecules for different fingerprint dimensions using the Euclidean distance coefficient.

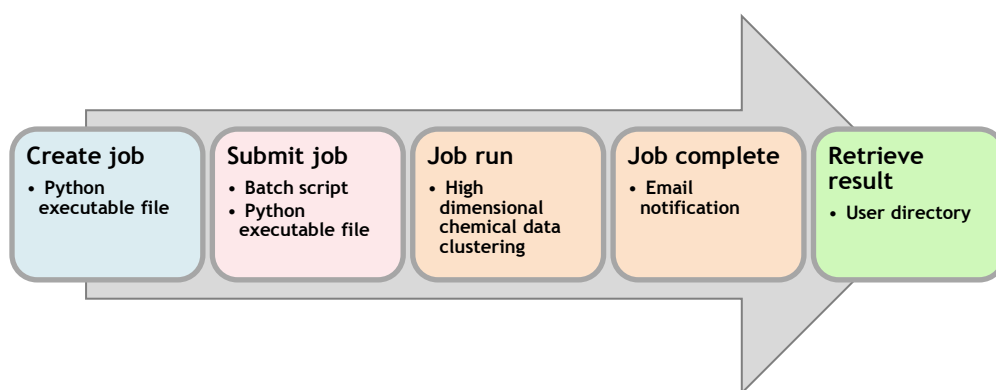


Figure 6-1 General workflow of high dimensional chemical data clustering implementation using ShARC

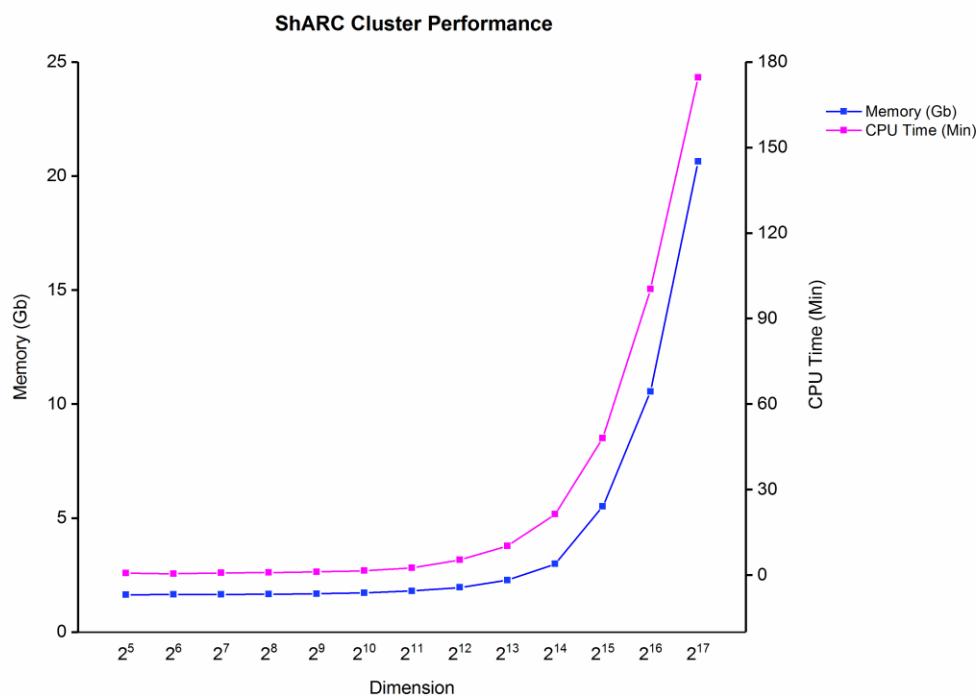


Figure 6-2 ShARC performance for various high dimensional chemical data clustering based on Ward's algorithm using MDDR dataset of 10,254 molecules measured by Euclidean distance coefficient

It is not surprising to see that the highest increase in the usage of computational resources is observed for the fingerprint dimensions above 2^{15} bits. This is because the sizes of the dimensions are very high (65,536 and 131,072 bits). This requires more memory and time for the computer to convert the initial molecule representation into the fingerprint descriptors, measure the pairwise distance and clustering. However, in this implementation, the overall memory and time have taken much less than expected, suggesting that this is becoming less of a restriction for a large dataset.

6.2.2 Cluster Analysis

A common way to visualise the cluster for analysis is by drawing a *dendrogram*, which displays the distance level at which there was a combination of objects and clusters (Leach & Gillet, 2007). Figure 6-3 shows an example of a cluster dendrogram in which the y-axis indicates the distance level and x-axis indicates

the clustered molecules. The dendrogram is being read bottom up to see at which distance molecules have been combined. For example, in Figure 6-3, molecules *b*, *c* and *e* are combined at a distance level of 1.5 while *a* and *d* at distance level of 2.0. Molecules *f* and *g* are the examples of two singletons (until a distance level of 3.0 when *f* merges with *a-e*).

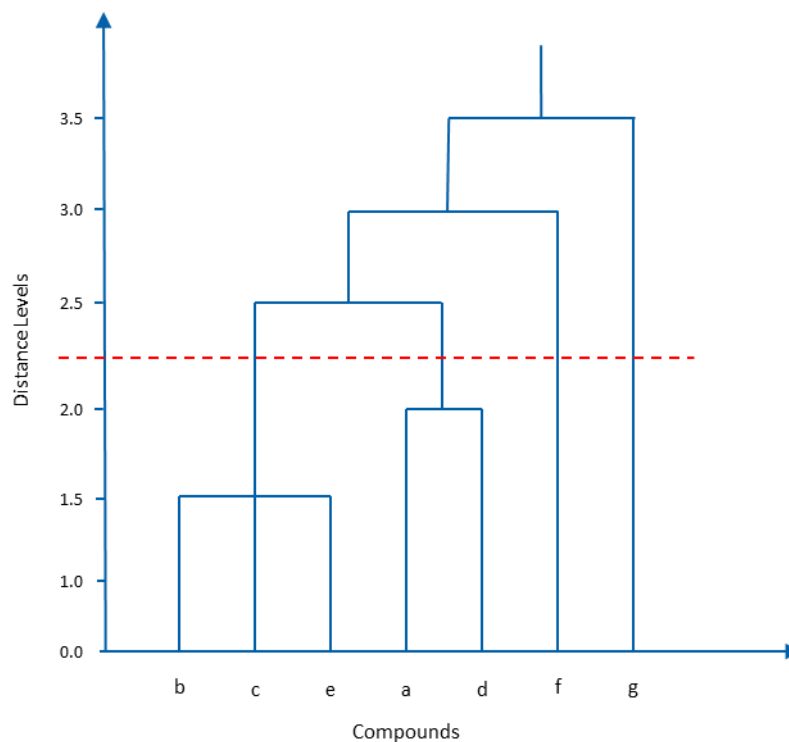


Figure 6-3 Hierarchical cluster dendrogram with the red horizontal dotted line indicating the level of partition to define the number of clusters

Cluster analysis can be performed on the cluster partitions which contain the number of clusters. Any desired number of clusters can be obtained by ‘cutting’ the dendrogram at the proper distance level. For example, the red dotted line in Figure 6-3 indicates such a horizontal line, resulting in four clusters. In the SciPy package library, the number of clusters can be determined simply by setting a threshold value in a function that indicates the number of clusters required (Jones et al., 2001).

In this experiment, the procedures described in Section 6.2.1 yielded 520 classifications from Ward’s and Group Average clustering methods using two

datasets, one type of fingerprint representation which has thirteen fingerprint dimensions and measured by ten distance coefficients. A partition value was applied to the cluster hierarchies to obtain cluster partitions that contain a set of 500, 600, 700, 800, 900 and 1000 clusters following the previous research by Chu, et al. (2012). The analysis and cluster evaluation were conducted based on these cluster partitions.

Two methods have been used to evaluate the effectiveness of the clustering application in this experiment: (i) *F*-measure and (ii) *QPI*-measure (*Quality Partition Index*). These methods have been introduced in Chapter 4.

6.3 Results and Discussion

The *F*-measure and the *QPI*-measure were used to evaluate the effectiveness of the molecular clustering in this experiment. The mean *F* and *QPI* values were averaged over the eleven activity classes in the MDDR dataset and the values resulted from Ward's clustering are shown in columns (a) *F*-Measure and (b) *QPI*-Measure in Table 6-1. The range of standard deviation for the mean *F* is also reported above the table. The results were presented for all distance coefficients and fingerprint dimensions where the best-performing fingerprint dimension for each partition in each column of the table is italicised, bold-faced and marked in red. In addition, Figure 6-4 represents the results in Table 6-1, visualising the effects of the clustering performances over different fingerprint dimensions.

As mentioned in section 6.2.1, further experiments have been conducted using the Group Average algorithm, the results of which are given in Table B-1 and Figure B-1 in Appendix B. Using similar clustering algorithms and evaluation methods, the results averaged over the fourteen activity classes in the WOMBAT dataset are listed and visualised in the tables and figures in Appendix B (Table B-2 and Figure B-2 for Ward's clustering, Table B-3 and Figure B-3 for Group Average clustering).

Table 6-1 Effectiveness value of Ward’s clustering measured by (a) *F*-measure and (b) *QPI*-measure for the MDDR dataset using various distance coefficients and fingerprint dimensions. The range of the standard deviation, σ , for the mean *F* is between 0.022 and 0.446

Distance Coefficients	Fingerprint Dimensions	Partition											
		(a) <i>F</i> -Measure					(b) <i>QPI</i> -Measure						
		500	600	700	800	900	1000	500	600	700	800	900	1000
[D1] Bray-Curtis	2 ⁵	0.645	0.749	0.756	0.783	0.822	0.897	0.133	0.138	0.143	0.148	0.151	0.156
	2 ⁶	0.766	0.843	0.860	0.997	0.997	0.993	0.194	0.206	0.215	0.221	0.228	0.238
	2 ⁷	1.039	1.046	1.063	1.123	1.123	1.141	0.247	0.262	0.269	0.287	0.299	0.311
	2 ⁸	1.006	0.988	1.009	1.074	1.168	1.107	0.283	0.306	0.316	0.326	0.335	0.338
	2 ⁹	1.029	1.045	1.046	1.058	1.085	1.091	0.290	0.307	0.319	0.334	0.337	0.345
	2 ¹⁰	0.996	1.021	1.106	1.106	1.127	1.207	0.286	0.299	0.309	0.332	0.339	0.344
	2 ¹¹	1.023	1.023	1.043	1.063	1.074	1.075	0.299	0.325	0.338	0.346	0.356	0.368
	2 ¹²	0.983	1.053	1.060	1.066	1.056	1.091	0.283	0.301	0.311	0.327	0.340	0.346
	2 ¹³	1.044	1.057	1.085	1.091	1.099	1.098	0.290	0.314	0.325	0.330	0.340	0.349
	2 ¹⁴	1.075	1.136	1.114	1.148	1.150	1.135	0.304	0.308	0.333	0.345	0.343	0.353
	2 ¹⁵	1.054	1.066	1.100	1.102	1.104	1.090	0.310	0.314	0.327	0.336	0.348	0.355
	2 ¹⁶	1.018	1.046	1.101	1.104	1.104	1.090	0.300	0.320	0.333	0.335	0.347	0.350
	2 ¹⁷	1.053	1.099	1.099	1.102	1.103	1.089	0.300	0.322	0.332	0.332	0.344	0.351
	[D2] City-Block	2 ⁵	0.764	0.831	0.831	0.937	0.942	0.969	0.141	0.146	0.152	0.157	0.163
2 ⁶		1.016	0.960	1.032	1.019	1.060	1.076	0.199	0.209	0.219	0.228	0.233	0.241
2 ⁷		1.089	1.089	1.065	1.069	1.093	1.140	0.271	0.286	0.290	0.298	0.308	0.315
2 ⁸		0.888	0.936	0.954	0.961	0.983	0.983	0.281	0.308	0.311	0.318	0.321	0.340
2 ⁹		0.997	1.019	1.019	1.056	1.060	1.070	0.275	0.296	0.298	0.312	0.325	0.344
2 ¹⁰		0.947	0.965	1.001	1.004	1.072	1.144	0.299	0.303	0.318	0.325	0.328	0.338
2 ¹¹		0.971	1.091	1.124	1.161	1.180	1.153	0.283	0.292	0.307	0.319	0.334	0.334
2 ¹²		0.876	0.951	1.032	1.067	1.078	1.085	0.287	0.302	0.315	0.330	0.345	0.347
2 ¹³		0.896	0.996	1.014	1.039	1.046	1.097	0.299	0.305	0.323	0.332	0.352	0.353
2 ¹⁴		0.878	0.901	0.956	1.003	1.032	1.073	0.275	0.289	0.318	0.336	0.347	0.347
2 ¹⁵		0.870	0.909	0.963	1.032	1.061	1.101	0.274	0.289	0.306	0.331	0.349	0.356
2 ¹⁶		0.870	0.905	0.946	1.017	1.017	1.073	0.294	0.304	0.314	0.329	0.340	0.358
2 ¹⁷		0.898	0.898	0.957	1.003	1.003	1.073	0.291	0.293	0.318	0.329	0.348	0.356
[D3] Cosine		2 ⁵	0.653	0.834	0.909	0.983	0.991	1.008	0.136	0.142	0.149	0.152	0.155
	2 ⁶	0.810	0.854	0.909	0.915	0.911	0.937	0.195	0.205	0.213	0.220	0.226	0.236
	2 ⁷	1.191	1.135	1.153	1.172	1.210	1.219	0.261	0.262	0.274	0.284	0.291	0.301
	2 ⁸	1.014	1.011	1.047	1.086	1.070	1.160	0.281	0.290	0.305	0.316	0.319	0.332
	2 ⁹	1.011	1.064	1.052	1.048	1.071	1.122	0.298	0.307	0.319	0.330	0.340	0.341
	2 ¹⁰	1.016	1.013	1.071	1.112	1.112	1.104	0.285	0.305	0.322	0.341	0.352	0.355
	2 ¹¹	1.049	1.056	1.054	1.054	1.054	1.059	0.282	0.310	0.331	0.342	0.353	0.363
	2 ¹²	1.047	1.055	1.101	1.068	1.074	1.112	0.291	0.304	0.319	0.337	0.345	0.366
	2 ¹³	1.045	1.069	1.099	1.127	1.133	1.119	0.285	0.307	0.328	0.330	0.345	0.354
	2 ¹⁴	1.039	1.056	1.073	1.112	1.134	1.119	0.296	0.306	0.318	0.325	0.339	0.351
	2 ¹⁵	1.062	1.062	1.134	1.152	1.168	1.119	0.313	0.327	0.340	0.342	0.348	0.360
	2 ¹⁶	1.059	1.059	1.124	1.153	1.168	1.119	0.307	0.324	0.338	0.342	0.349	0.363
	2 ¹⁷	1.057	1.057	1.122	1.152	1.168	1.119	0.300	0.311	0.339	0.349	0.355	0.366
	[D4] Euclidean	2 ⁵	0.691	0.705	0.740	0.740	0.761	0.800	0.145	0.149	0.156	0.162	0.169
2 ⁶		0.935	1.001	1.019	1.029	1.082	1.105	0.199	0.217	0.224	0.234	0.240	0.244
2 ⁷		0.938	1.017	1.042	1.068	1.088	1.073	0.248	0.262	0.281	0.289	0.295	0.306
2 ⁸		0.893	1.022	1.037	1.037	1.113	1.112	0.266	0.290	0.312	0.329	0.335	0.341
2 ⁹		1.050	1.056	1.098	1.085	1.068	1.090	0.297	0.312	0.315	0.330	0.337	0.345
2 ¹⁰		0.960	1.095	1.103	1.144	1.106	1.130	0.281	0.316	0.334	0.338	0.344	0.351
2 ¹¹		0.909	1.008	1.024	1.034	1.011	0.990	0.291	0.319	0.327	0.335	0.342	0.352
2 ¹²		0.931	1.032	1.079	1.079	1.036	1.042	0.290	0.317	0.336	0.335	0.337	0.348
2 ¹³		0.895	1.041	1.047	1.047	1.042	1.046	0.284	0.314	0.326	0.344	0.345	0.343
2 ¹⁴		0.891	1.010	1.027	1.042	1.047	0.993	0.296	0.325	0.330	0.339	0.343	0.343
2 ¹⁵		0.870	0.951	1.028	1.040	1.046	1.046	0.278	0.301	0.315	0.331	0.333	0.338
2 ¹⁶		0.870	0.981	1.011	1.015	1.015	0.979	0.277	0.302	0.311	0.337	0.337	0.340
2 ¹⁷		0.891	0.969	1.031	1.031	1.026	0.990	0.284	0.307	0.319	0.341	0.348	0.348

The best-performing fingerprint dimension in each column of the table is italicised, bold-faced and marked in red for ease of reference.

Table 6-1 (continued)

Distance Coefficients	Fingerprint Dimensions	Partition												
		(a) <i>F</i> -Measure						(b) <i>QPI</i> -Measure						
		500	600	700	800	900	1000	500	600	700	800	900	1000	
[D5] Hamming	2 ⁵	0.764	0.831	0.831	0.937	0.942	0.969	0.141	0.146	0.152	0.157	0.163	0.165	
	2 ⁶	1.016	0.960	1.032	1.019	1.060	1.076	0.199	0.209	0.219	0.228	0.233	0.241	
	2 ⁷	1.089	1.089	1.065	1.069	1.093	1.140	0.271	0.286	0.290	0.298	0.309	0.315	
	2 ⁸	0.888	0.936	0.954	0.961	0.983	0.983	0.281	0.308	0.311	0.317	0.321	0.340	
	2 ⁹	0.997	1.021	1.019	1.056	1.060	1.070	0.275	0.296	0.298	0.312	0.325	0.344	
	2 ¹⁰	0.975	0.981	1.001	1.004	1.072	1.144	0.299	0.304	0.317	0.331	0.333	0.338	
	2 ¹¹	0.971	1.091	1.124	1.161	1.180	1.136	0.281	0.292	0.307	0.315	0.334	0.331	
	2 ¹²	0.857	0.951	1.032	1.032	1.078	1.085	0.283	0.302	0.315	0.324	0.344	0.348	
	2 ¹³	0.852	0.903	0.996	1.014	1.051	1.097	0.295	0.298	0.313	0.329	0.342	0.354	
	2 ¹⁴	0.837	0.875	0.945	0.945	0.945	1.021	0.270	0.280	0.312	0.312	0.312	0.345	
	2 ¹⁵	0.833	0.909	0.909	0.909	0.909	0.909	0.267	0.289	0.289	0.289	0.289	0.289	
	2 ¹⁶	0.841	0.841	0.841	0.841	0.841	0.841	0.287	0.287	0.287	0.287	0.287	0.287	
	2 ¹⁷	0.818	0.818	0.818	0.818	0.818	0.818	0.238	0.238	0.238	0.238	0.238	0.238	
	[D6] Jaccard	2 ⁵	0.717	0.734	0.774	0.774	0.760	0.772	0.135	0.138	0.143	0.148	0.153	0.158
		2 ⁶	0.779	0.779	0.860	0.889	1.013	1.044	0.190	0.200	0.214	0.217	0.232	0.236
		2 ⁷	1.062	1.068	1.092	1.107	1.127	1.127	0.275	0.291	0.304	0.305	0.310	0.310
		2 ⁸	0.976	1.011	1.037	1.010	1.041	1.041	0.262	0.283	0.295	0.304	0.321	0.328
2 ⁹		1.071	1.065	1.060	1.099	1.101	1.110	0.285	0.307	0.317	0.333	0.346	0.346	
2 ¹⁰		1.034	1.127	1.115	1.131	1.150	1.150	0.277	0.289	0.311	0.329	0.340	0.345	
2 ¹¹		1.010	1.067	1.071	1.075	1.076	1.076	0.305	0.313	0.329	0.346	0.360	0.363	
2 ¹²		1.013	1.088	1.099	1.075	1.122	1.132	0.296	0.323	0.336	0.338	0.348	0.353	
2 ¹³		1.023	1.051	1.095	1.138	1.112	1.112	0.283	0.306	0.319	0.325	0.342	0.350	
2 ¹⁴		0.997	1.061	1.064	1.077	1.079	1.079	0.281	0.315	0.332	0.333	0.345	0.344	
2 ¹⁵		0.995	1.058	1.068	1.083	1.079	1.079	0.283	0.304	0.334	0.354	0.357	0.352	
2 ¹⁶		1.012	1.054	1.091	1.106	1.106	1.106	0.285	0.308	0.333	0.351	0.353	0.353	
2 ¹⁷		1.008	1.050	1.076	1.101	1.101	1.101	0.288	0.318	0.331	0.343	0.347	0.355	
[D7] Kulsinski		2 ⁵	0.640	0.763	0.800	0.817	0.847	0.856	0.139	0.147	0.150	0.153	0.159	0.163
		2 ⁶	0.925	0.979	0.999	1.080	1.080	1.106	0.200	0.207	0.221	0.223	0.230	0.239
		2 ⁷	1.000	1.000	1.002	1.038	1.067	1.080	0.263	0.272	0.275	0.287	0.295	0.299
		2 ⁸	0.932	0.970	0.972	1.039	1.104	1.110	0.272	0.290	0.303	0.311	0.319	0.332
	2 ⁹	0.920	0.969	0.972	1.021	1.038	1.047	0.294	0.306	0.314	0.333	0.334	0.337	
	2 ¹⁰	0.943	1.025	1.073	1.073	1.104	1.091	0.274	0.301	0.316	0.309	0.322	0.337	
	2 ¹¹	0.939	1.091	1.102	1.147	1.147	1.129	0.295	0.305	0.331	0.351	0.364	0.364	
	2 ¹²	0.908	1.063	1.066	1.079	1.079	1.079	0.298	0.308	0.322	0.332	0.338	0.341	
	2 ¹³	0.917	1.053	1.086	1.086	1.109	1.109	0.272	0.286	0.303	0.331	0.347	0.347	
	2 ¹⁴	0.966	0.995	1.119	1.119	1.119	1.073	0.271	0.301	0.311	0.311	0.311	0.350	
	2 ¹⁵	0.897	1.040	1.040	1.040	1.040	1.040	0.276	0.311	0.311	0.311	0.311	0.311	
	2 ¹⁶	0.896	0.896	0.896	0.896	1.096	1.096	0.264	0.264	0.264	0.264	0.334	0.334	
	2 ¹⁷	0.947	0.947	0.947	0.947	0.947	0.947	0.275	0.275	0.275	0.275	0.275	0.275	
	[D8] Rogers-Tanimoto	2 ⁵	0.660	0.681	0.741	0.799	0.818	0.835	0.144	0.148	0.150	0.156	0.162	0.168
		2 ⁶	0.835	0.884	0.891	0.977	1.013	1.020	0.205	0.213	0.219	0.221	0.231	0.237
		2 ⁷	0.960	0.992	1.023	1.023	1.045	1.093	0.268	0.284	0.294	0.308	0.311	0.325
		2 ⁸	0.849	0.871	0.887	0.941	0.964	0.964	0.286	0.296	0.311	0.318	0.330	0.344
2 ⁹		0.931	1.011	1.022	1.049	1.063	1.131	0.293	0.303	0.325	0.330	0.337	0.348	
2 ¹⁰		0.980	1.048	1.017	1.043	1.043	1.127	0.279	0.295	0.324	0.336	0.344	0.358	
2 ¹¹		0.973	1.023	1.069	1.069	1.127	1.150	0.277	0.292	0.300	0.311	0.328	0.327	
2 ¹²		0.919	0.955	1.051	1.099	1.130	1.085	0.283	0.306	0.317	0.332	0.352	0.357	
2 ¹³		0.852	0.983	1.014	1.039	1.051	1.097	0.281	0.297	0.311	0.341	0.347	0.352	
2 ¹⁴		0.878	0.905	0.927	0.956	1.032	1.073	0.273	0.279	0.303	0.319	0.338	0.348	
2 ¹⁵		0.838	0.870	0.963	0.963	0.963	1.049	0.254	0.275	0.302	0.302	0.302	0.347	
2 ¹⁶		0.839	0.905	0.905	0.905	0.905	0.905	0.262	0.305	0.305	0.305	0.305	0.305	
2 ¹⁷		0.875	0.875	0.875	0.875	0.875	0.875	0.281	0.281	0.281	0.281	0.281	0.281	

The best-performing fingerprint dimension in each column of the table is italicised, bold-faced and marked in red for ease of reference.

Table 6-1 (continued)

Distance Coefficients	Fingerprint Dimensions	Partition												
		(a) <i>F</i> -Measure						(b) <i>QPI</i> -Measure						
		500	600	700	800	900	1000	500	600	700	800	900	1000	
[D9] Russell-Rao	2 ⁵	0.619	0.708	0.704	0.741	0.810	0.888	0.132	0.136	0.141	0.143	0.146	0.149	
	2 ⁶	0.900	0.935	0.971	0.983	1.032	1.121	0.195	0.200	0.205	0.212	0.219	0.223	
	2 ⁷	1.026	1.045	1.051	1.059	1.087	1.087	0.274	0.287	0.290	0.301	0.303	0.313	
	2 ⁸	0.934	0.984	0.981	0.986	0.986	1.052	0.292	0.305	0.315	0.328	0.333	0.340	
	2 ⁹	0.980	0.979	1.022	1.020	1.067	1.063	0.290	0.302	0.320	0.329	0.331	0.349	
	2 ¹⁰	0.953	0.968	1.004	1.063	1.112	1.080	0.283	0.300	0.311	0.324	0.336	0.343	
	2 ¹¹	0.939	1.065	1.050	1.097	1.097	1.102	0.281	0.304	0.312	0.326	0.330	0.335	
	2 ¹²	0.962	1.029	1.087	1.094	1.094	1.094	0.281	0.291	0.318	0.325	0.344	0.352	
	2 ¹³	0.920	1.052	1.099	1.095	1.095	1.095	0.293	0.294	0.298	0.324	0.336	0.336	
	2 ¹⁴	0.990	1.023	1.060	1.060	1.060	1.083	0.286	0.301	0.324	0.324	0.324	0.356	
	2 ¹⁵	0.885	1.012	1.012	1.012	1.012	1.012	0.258	0.290	0.290	0.290	0.290	0.290	
	2 ¹⁶	1.016	1.016	1.016	1.016	1.044	1.044	0.281	0.281	0.281	0.281	0.322	0.322	
	2 ¹⁷	0.945	0.945	0.945	0.945	0.945	0.945	0.293	0.293	0.293	0.293	0.293	0.293	
	[D10] Sokal-Sneath	2 ⁵	0.704	0.745	0.761	0.832	0.899	0.899	0.139	0.145	0.149	0.152	0.156	0.160
		2 ⁶	1.118	1.118	1.153	1.159	1.145	1.126	0.215	0.219	0.226	0.228	0.240	0.244
		2 ⁷	0.985	1.030	1.031	1.034	1.128	1.151	0.248	0.256	0.272	0.280	0.288	0.304
		2 ⁸	1.020	1.038	1.034	1.064	1.106	1.107	0.272	0.289	0.299	0.320	0.330	0.337
2 ⁹		1.001	1.049	1.062	1.066	1.064	1.069	0.269	0.284	0.296	0.316	0.326	0.328	
2 ¹⁰		1.035	1.110	1.106	1.106	1.106	1.064	0.291	0.319	0.335	0.342	0.337	0.360	
2 ¹¹		1.011	1.059	1.086	1.115	1.142	1.115	0.289	0.300	0.316	0.341	0.328	0.339	
2 ¹²		1.048	1.082	1.114	1.129	1.156	1.163	0.288	0.304	0.322	0.332	0.339	0.355	
2 ¹³		1.028	1.082	1.101	1.101	1.098	1.114	0.290	0.310	0.341	0.337	0.348	0.358	
2 ¹⁴		0.933	1.030	1.034	1.056	1.103	1.109	0.283	0.311	0.327	0.334	0.338	0.354	
2 ¹⁵		0.978	1.012	1.062	1.078	1.064	1.114	0.279	0.299	0.309	0.321	0.322	0.339	
2 ¹⁶		0.978	1.043	1.045	1.047	1.064	1.114	0.266	0.289	0.307	0.320	0.333	0.343	
2 ¹⁷	0.975	1.045	1.048	1.069	1.125	1.135	0.276	0.299	0.314	0.314	0.330	0.343		

The best-performing fingerprint dimension in each column of the table is italicised, bold-faced and marked in red for ease of reference.

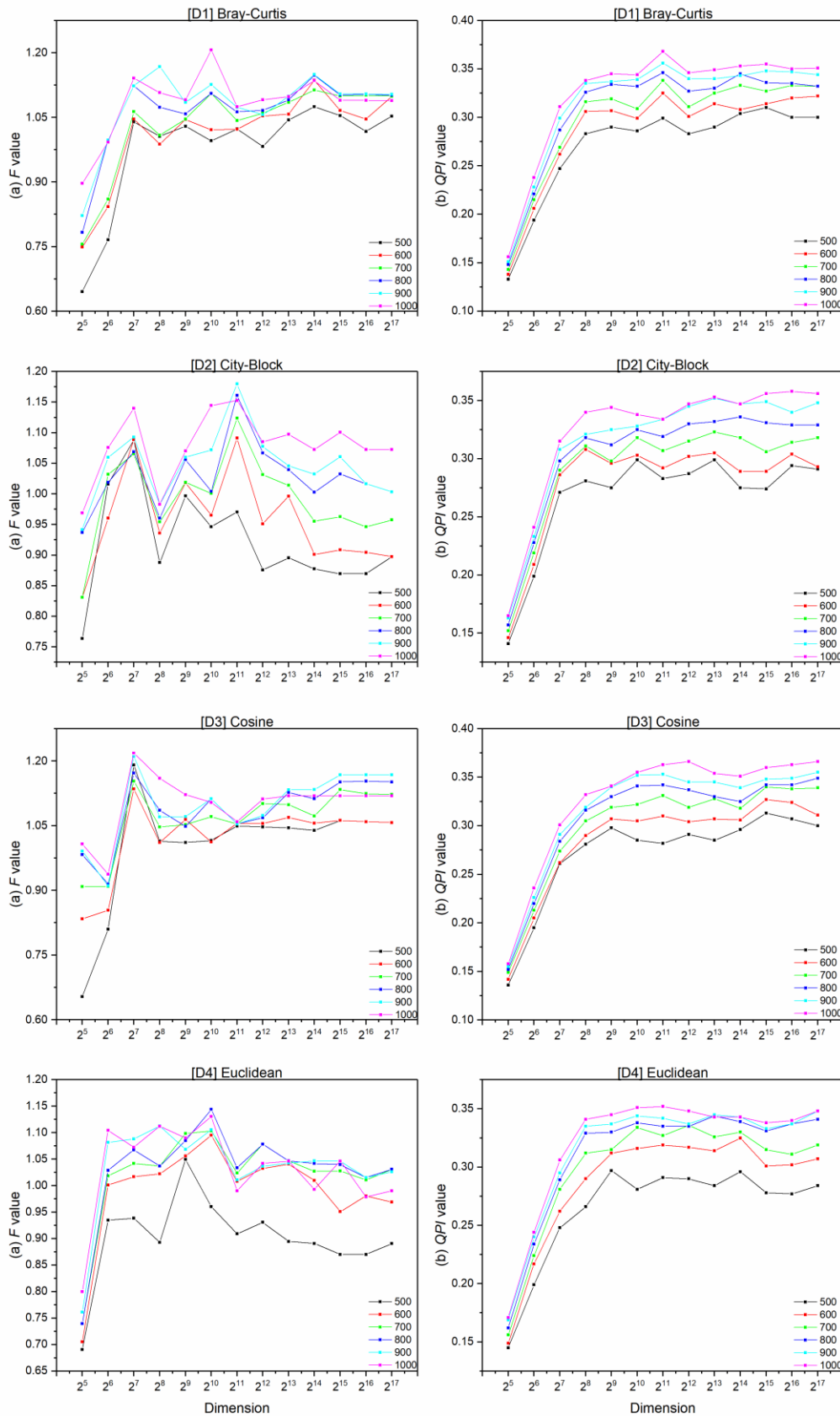


Figure 6-4 Effects of dimensionality on Ward's clustering measured by (a) F -measure and (b) QPI -measure for MDDR dataset using various distance coefficients

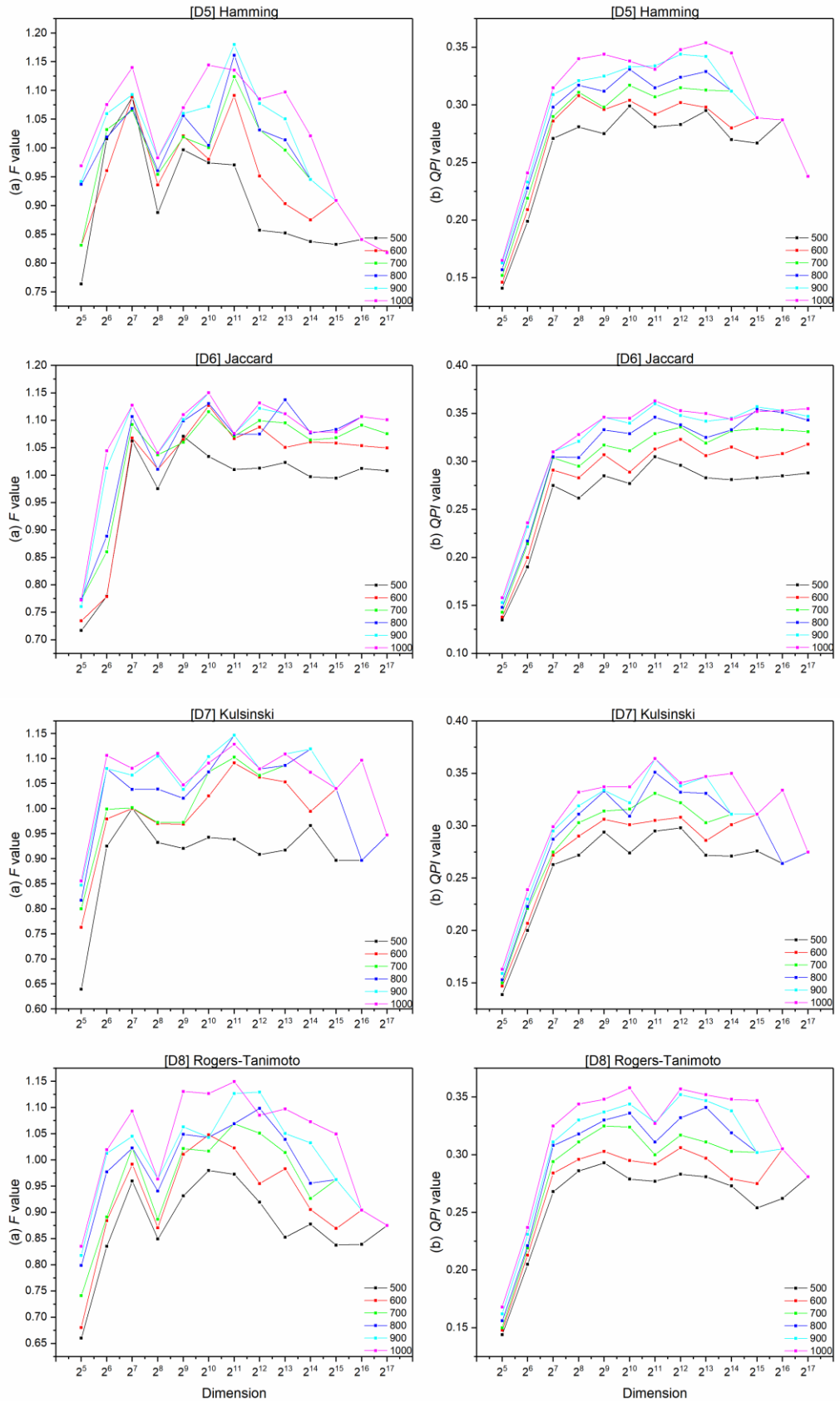


Figure 6-4 (continued)

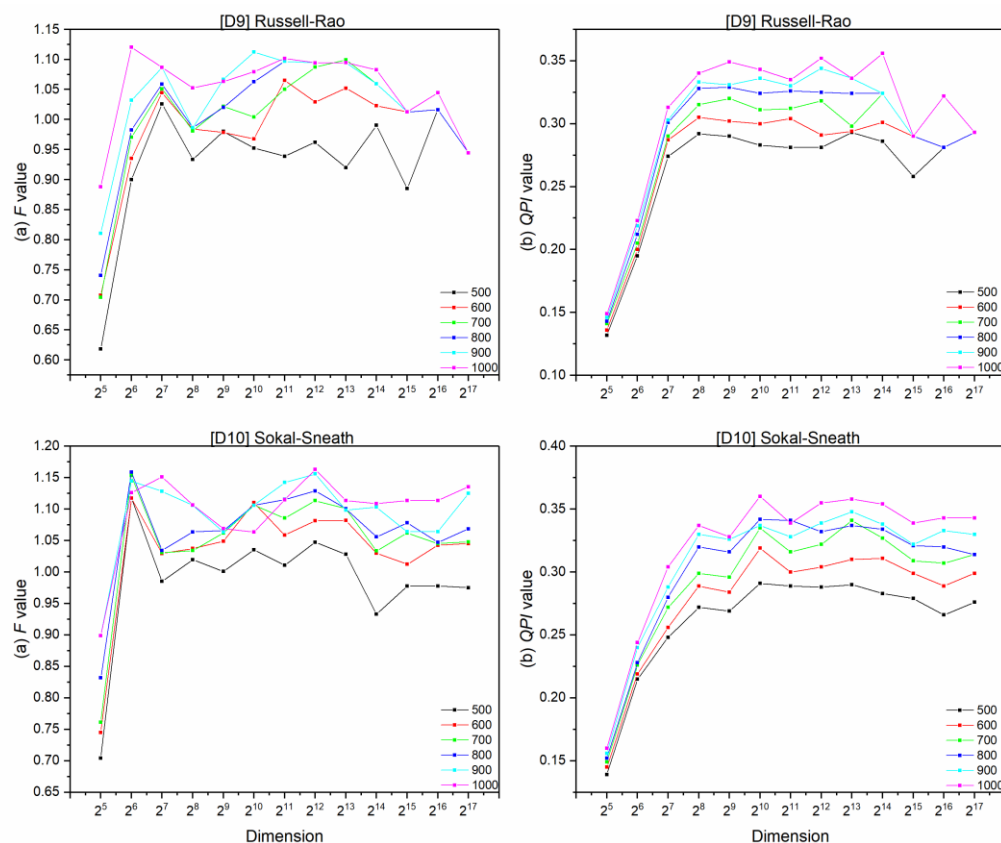


Figure 6-4 (continued)

6.3.1 Effects of Low Dimensionality on the Effectiveness of Clustering

The inspection of Figure 6-4 shows a common general behaviour across all distance coefficients and hierarchical partitions. Lowest clustering performance was obtained from the lowest fingerprint dimension considering both evaluation criteria.

The possible reason for this behaviour is the fewer bit vector spaces of the lowest dimension, which only has 32 (i.e., 2^5) bits space, and hence involves very large numbers of collisions when bits are being set. This is considered small to represent the information of 10,254 molecules belonging to the MDDR dataset and 13,813 molecules in the WOMBAT dataset used in this experiment. Hence, there is a possibility that most of the bits will be utilised to represent the features in the molecules or most of the molecules will have the same bit sets in the fingerprint.

The bits dimension of Morgan R2 fingerprints used in this experiment were analysed. Table 6-2 lists the summary statistics obtained from analysis of bits set for the molecules in the MDDR and WOMBAT datasets. In addition, it provides the bit collision rate for each dimension that was obtained by subtracting the average bits set of a lower dimension from the average bits set of a higher dimension.

Table 6-2 Summary statistics of bits set and bit collision rate for (a) 10,254 molecules in MDDR dataset and (b) 13,813 molecules in WOMBAT dataset using various Morgan R2 fingerprint dimensions

Dataset	Bits Set	Morgan R2 Fingerprint Dimension												
		2 ⁵	2 ⁶	2 ⁷	2 ⁸	2 ⁹	2 ¹⁰	2 ¹¹	2 ¹²	2 ¹³	2 ¹⁴	2 ¹⁵	2 ¹⁶	2 ¹⁷
(a) MDDR	Min	8	9	9	9	10	10	10	10	10	10	10	10	10
	Max	32	62	94	127	148	157	162	165	165	166	166	166	166
	Average	25.62	35.60	43.31	48.18	50.61	52.00	52.68	53.37	53.50	53.58	53.61	53.62	53.63
	Bit Collision Rate		9.98	7.71	4.86	2.44	1.39	0.67	0.70	0.13	0.07	0.03	0.01	0.01
(b) WOMBAT	Min	6	8	8	8	8	8	8	8	8	8	8	8	8
	Max	32	63	114	163	192	206	220	222	222	223	223	223	223
	Average	25.11	34.62	41.86	46.33	48.61	49.89	50.53	51.17	51.28	51.35	51.39	51.40	51.40
	Bit Collision Rate		9.52	7.23	4.47	2.29	1.28	0.64	0.64	0.10	0.07	0.04	0.01	0.01

It can be seen that the lowest fingerprint dimension (i.e., 2⁵) of both datasets has molecules with a maximum number of bits set of 32 bits. Similar behaviour can be seen from the fingerprint dimension of 2⁶, which has a maximum number of bits set of 62 bits for MDDR and 63 bits for WOMBAT. In addition, the average number of bits set increases and the bit collision rate decreases to zero as the dimensionality increases.

This indicates that the use of low fingerprint dimensions can result in a maximum utilisation of bits fingerprint, therefore increasing the chances of higher bit collisions. As a result, this will affect the pairwise distance calculation between the molecules since the distances between a molecule and its nearest and furthest molecules can be difficult to distinguish. Hence, the performance of the clustering using lower dimensions will also be affected, explaining the behaviour observed in Figure 6-4 for the MDDR dataset and similarly from the WOMBAT dataset in Appendix B.

6.3.2 Effects of High Dimensionality on the Effectiveness of Clustering

The results from the *QPI* measure are discussed because they provide general interpretations of the separation between the actives and inactives in the MDDR dataset. As shown in Figure 6-4, two distinct trends on the effects of dimensionality on the effectiveness of clustering can be observed.

First, the effectiveness of clustering increased as the fingerprint dimension increases until it reached a maximum *QPI* value and remains thereafter. This behaviour can be observed by using six distance coefficients, which are Bray-Curtis [D1], City-Block [D2], Cosine [D3], Euclidean [D4], Jaccard [D6] and Sokal-Sneath [D10].

Second, the cluster performance increased as the fingerprint dimension increases followed by a decrease after it reached a maximum *QPI* value, which can be seen by using the other four distance coefficients that are Hamming [D5], Kulsinski [D7], Rogers-Tanimoto [D8] and Russell-Rao [D9].

The trends observed varied depending on the coefficients used to measure the pairwise distance of the molecules in the dataset. Two distance coefficients were chosen as the examples in this discussion, i.e., the Euclidean [D4] and Hamming [D5] distance coefficients, which represent the distinct behaviours.

As listed in Table 4-6, the Euclidean [D4] and Hamming [D5] distance coefficients are defined by Eq. (32) and Eq. (33):

$$D_{EUC} = \left[\sum_{i=1}^n |x_i - y_i|^2 \right]^{1/2} \quad (32)$$

$$D_{HAM} = \frac{\sum_{i=1}^n |x_i - y_i|}{n} \quad (33)$$

In relation to the fingerprint dimensionality, Hamming [D5] is different from Euclidean [D4] because it measures the differences between two molecules

from the overall dimensions. Based on the Hamming [D5] formulation, the distance between two molecules will be transformed into a much shorter distance in very high dimensional space compared to the distance measured in a lower dimensional space. These assumptions are investigated separately in the following Sections 6.3.2.1 and 6.3.2.2.

6.3.2.1 Analysis of Distance Measures by Euclidean Distance Coefficient

The pairwise distances of the molecules in the MDDR dataset measured by the Euclidean [D4] distance coefficient for each fingerprint dimension were analysed using the histogram distribution. Table 6-3 lists the statistical information about the distribution, which includes the mean, standard deviation, minimum and maximum distance values. The difference between the maximum and minimum distances for an extreme case is also included. In addition, Figure 6-5 represents the histogram distribution plot for the distance values against the frequency of the observations for each dimension.

Table 6-3 Summary statistics for distribution of pairwise distance measured by Euclidean [D4] distance coefficient for MDDR dataset using various fingerprint dimensions

Distance Coefficient	Fingerprint Dimensions	Mean Distance	Standard Deviation	Minimum Distance	Maximum Distance	(Maximum - Minimum) Distance
[D4] Euclidean	2^5	3.025	0.542	0.000	5.385	5.385
	2^6	5.236	0.408	0.000	7.348	7.348
	2^7	6.986	0.508	0.000	9.592	9.592
	2^8	8.089	0.716	0.000	11.747	11.747
	2^9	8.678	0.854	0.000	14.000	14.000
	2^{10}	9.004	0.944	0.000	15.133	15.133
	2^{11}	9.166	0.995	0.000	15.843	15.843
	2^{12}	9.283	1.018	0.000	16.248	16.248
	2^{13}	9.318	1.027	0.000	16.371	16.371
	2^{14}	9.338	1.033	0.000	16.492	16.492
	2^{15}	9.348	1.036	0.000	16.523	16.523
	2^{16}	9.352	1.037	0.000	16.523	16.523
	2^{17}	9.354	1.037	0.000	16.583	16.583

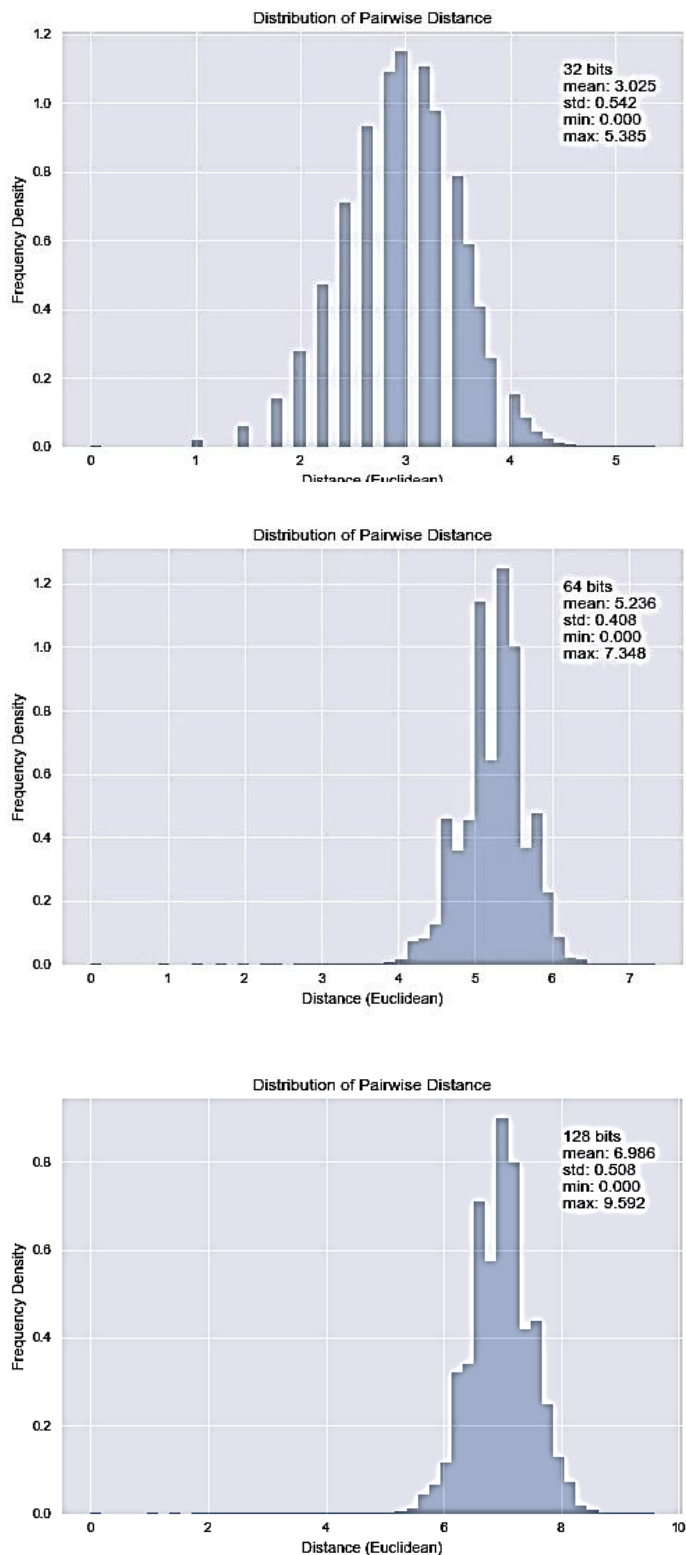


Figure 6-5 Distribution histograms of pairwise distances for molecules in MDDR represented by various fingerprint dimensions and measured by Euclidean distance coefficient

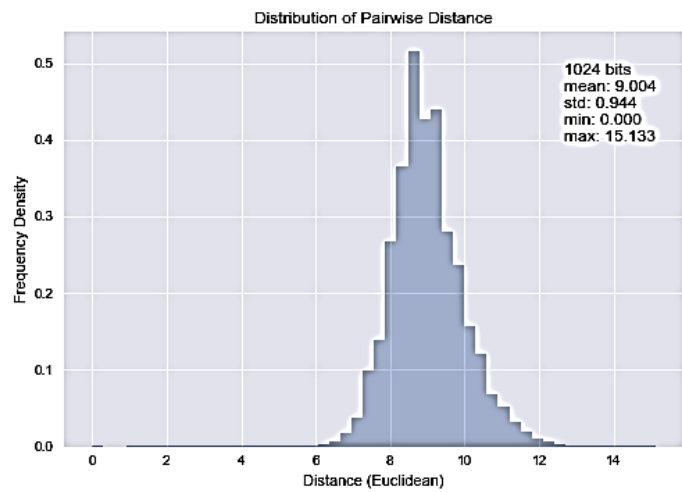
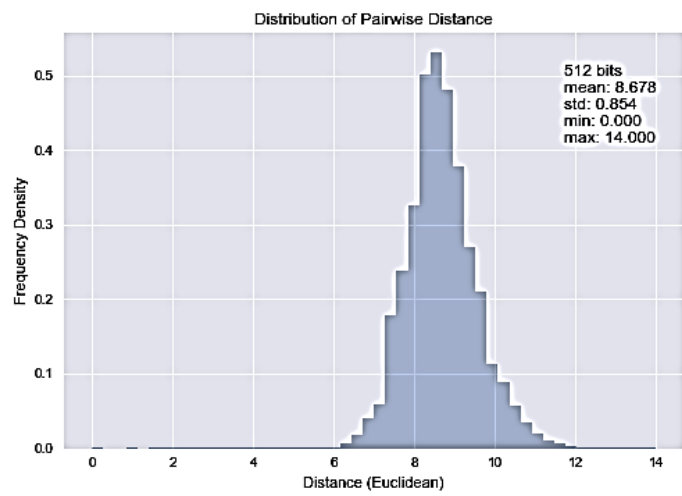
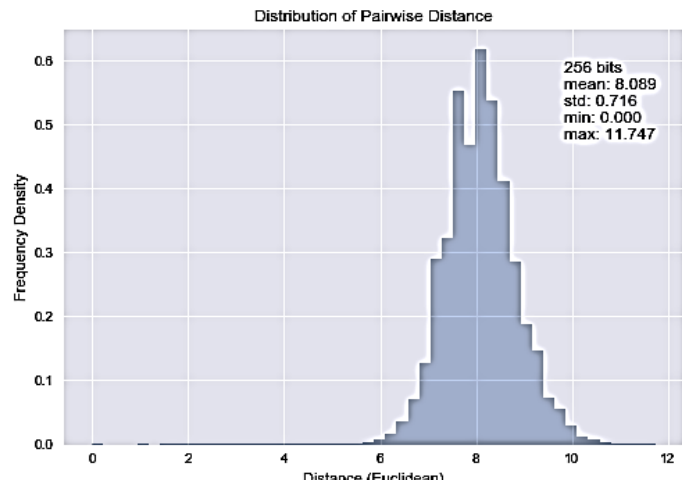


Figure 6-5 (continued)

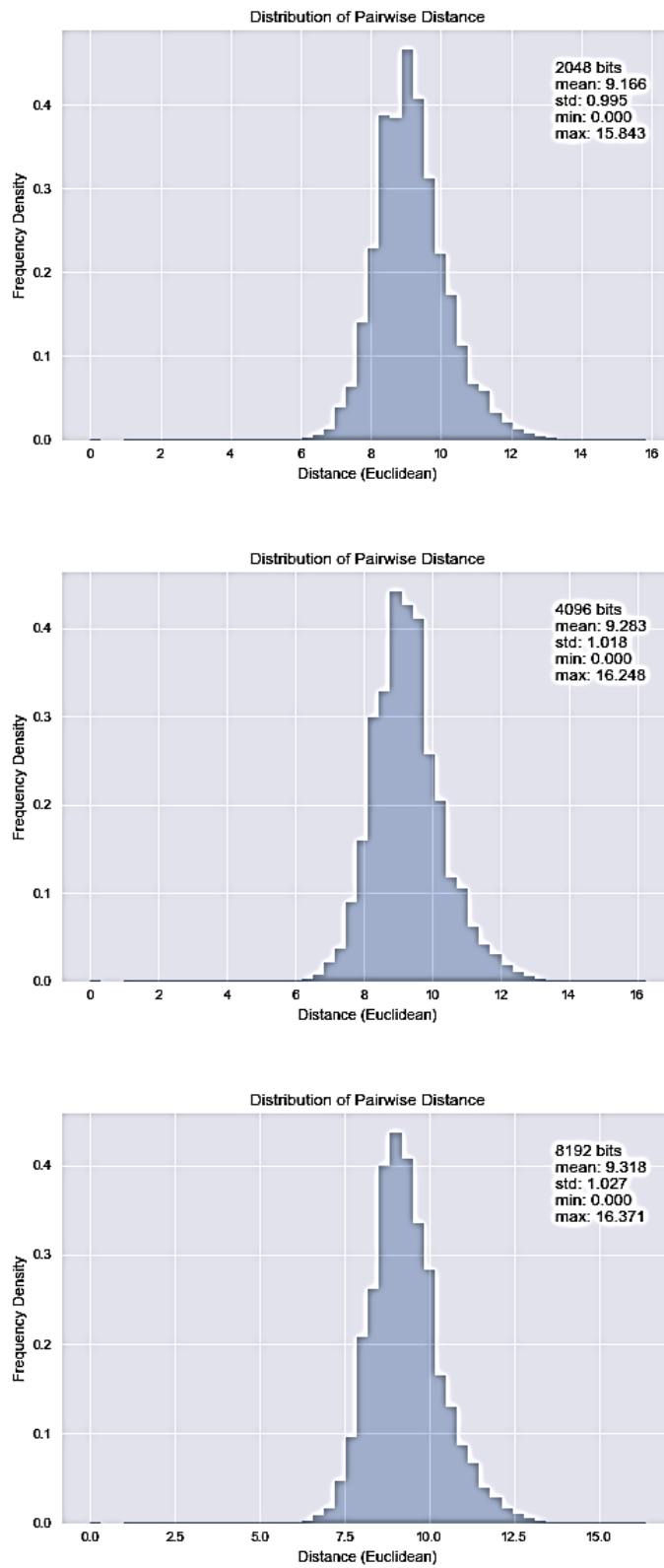


Figure 6-5 (continued)

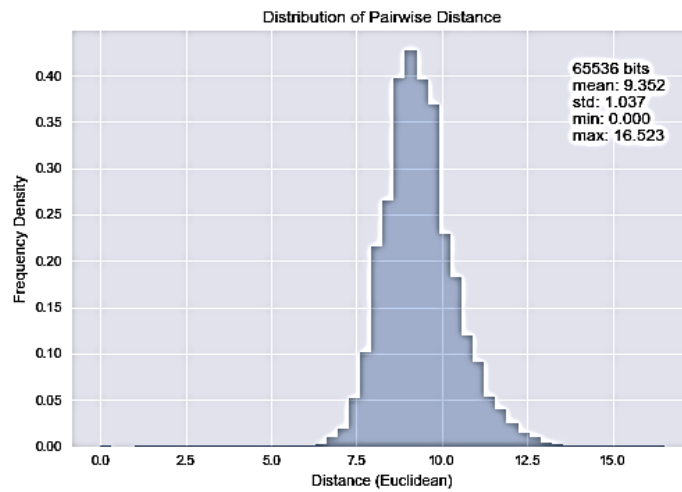
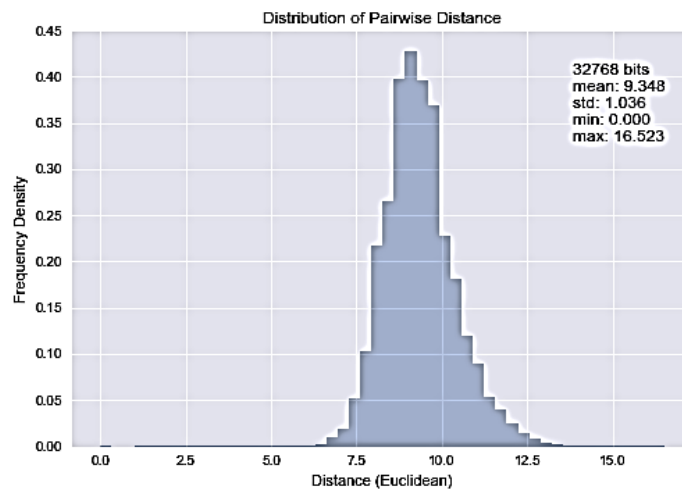
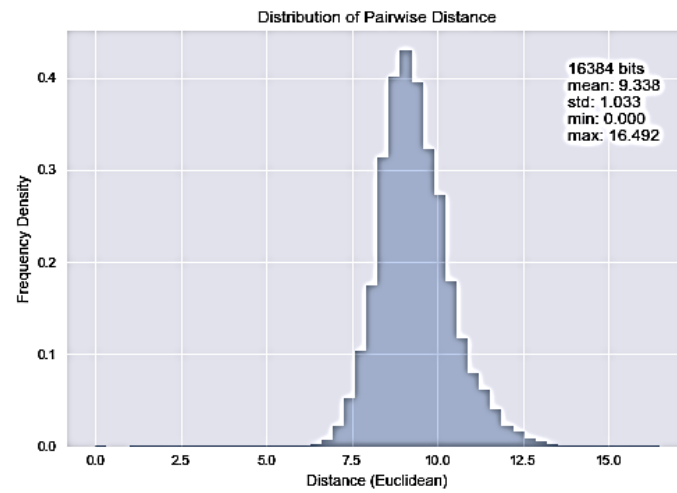


Figure 6-5 (continued)

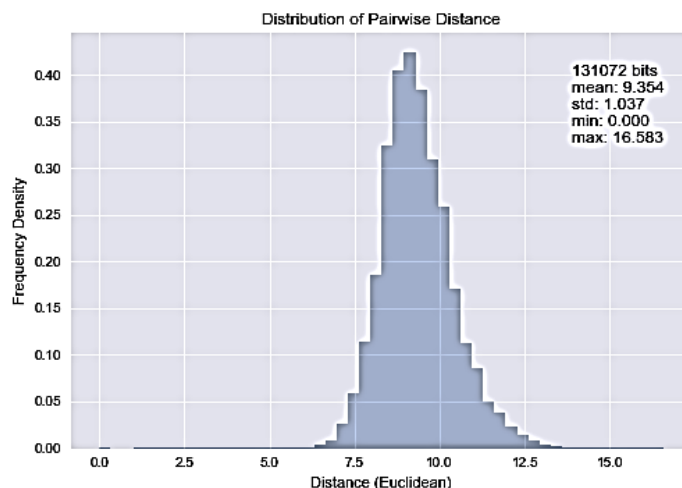


Figure 6-5 (continued)

Inspection of Figure 6-5 shows an overall symmetric and well spread pattern of distributions when measured by the Euclidean [D4] distance coefficient for all dimensions. The centre and variance values of the distributions can be obtained from the mean and standard deviation values provided in each plot.

As listed in Table 6-3, the mean values for the distances increased from 3.025 to 9.354 as the dimensionality increases. However, the increase of mean values is very small from the 2^{10} until 2^{17} bits dimension. This indicates that, the increase of bits dimension increases the average distance values until a certain dimension and remains constant thereafter.

The standard deviation values have also increased from 0.542 to 1.037. This similar behaviour indicates that there is more variance of distances in the higher dimensionality space than in the lower dimensions. In this condition, a better separation between the molecules is likely to be seen in the dataset. As an effect, the clustering process will likely be effective because it can distinguish between the nearest and the furthest molecule in the higher dimensional space.

Considering the effect of clustering, it is assumed that these criteria enable better discrimination between the molecules when Ward's algorithm is used. This is because the algorithm considers distance values in determining the

minimal variance when performing the merger, which can be more effectively quantified when better discrimination is available. Therefore, these criteria have resulted in the first trend observed in Figure 6-4. The effectiveness of clustering increased as the fingerprint dimension increases until it reached a maximum *QPI* value and remains constant thereafter.

In general, the results from the *F* measure show a similar pattern of effectiveness to the *QPI* measure, with the exception of being more variable due to the effects of different homogeneity classes when averaging the *F* values. The highest *F* value corresponds to the average of optimal cluster for each dimension in the MDDR dataset.

Finally, similar trends can be observed by using the Bray-Curtis [D1], City-Block [D2], Cosine [D3], Jaccard [D6] and Sokal-Sneath [D10] distance coefficients suggesting similar behaviour on the distance distributions.

The findings were also consistent for the results using the Group Average algorithm, suggesting the consistency of the findings using another algorithm that considers the distances for the merger. Similarly, results using the WOMBAT dataset also suggest the consistency of the findings on different datasets. The corresponding tables and figures can be found in Appendix B.

6.3.2.2 Analysis of Distances Measured by Hamming Distance Coefficient

Pairwise distance distribution measured by Hamming [D5] distance was analysed to quantify the second trend observed in Figure 6-4, i.e., the effectiveness of the clustering increased as the fingerprint dimension increases followed by a decrease after it reached a maximum *QPI* value. Table 6-4 lists the statistical information about the distribution and Figure 6-6 plots the histogram distribution.

Table 6-4 Summary statistics for distribution of pairwise distance measured by Hamming [D5] distance coefficient for MDDR dataset using various fingerprint dimensions

Distance Coefficient	Fingerprint Dimensions	Mean Distance	Standard Deviation	Minimum Distance	Maximum Distance	(Maximum - Minimum) Distance
[D5] Hamming	2 ⁵	0.295	0.100	0.000	0.906	0.906
	2 ⁶	0.431	0.066	0.000	0.844	0.844
	2 ⁷	0.383	0.055	0.000	0.719	0.719
	2 ⁸	0.258	0.046	0.000	0.539	0.539
	2 ⁹	0.148	0.030	0.000	0.383	0.383
	2 ¹⁰	0.080	0.017	0.000	0.224	0.224
	2 ¹¹	0.042	0.009	0.000	0.123	0.123
	2 ¹²	0.021	0.005	0.000	0.064	0.064
	2 ¹³	0.011	0.002	0.000	0.033	0.033
	2 ¹⁴	0.005	0.001	0.000	0.017	0.017
	2 ¹⁵	0.003	0.001	0.000	0.008	0.008
	2 ¹⁶	0.001	0.000	0.000	0.004	0.004
	2 ¹⁷	0.011	0.000	0.000	0.002	0.002

Figure 6-6 shows a different behaviour compared to the previous discussion in Section 6.3.2.1. As the dimensionality increases towards the highest dimension, the distribution of distances between the molecules changing from symmetric to relatively uniform. As listed in Table 6-4, with the exception of the 2⁵ bits dimension, the mean values for the distances decrease as the dimensionality increases from 0.431 to 0.011. This indicates that, the increase of bits dimension decreases the average distance values when measured by the Hamming [D5] distance coefficient.

Another important behaviour is the change of the variances of the distributions, which decrease from the standard deviation value of 0.100 to 0.000. This indicates three behaviours: (1) in general, the low standard deviation value means that almost most of the distances are very close to the average distance, (2) there were less variances of distances in the higher dimensionality spaces than in the lower dimensions and (3) there were zero variances in the two highest dimensional spaces, i.e., 2¹⁶ and 2¹⁷ bits.

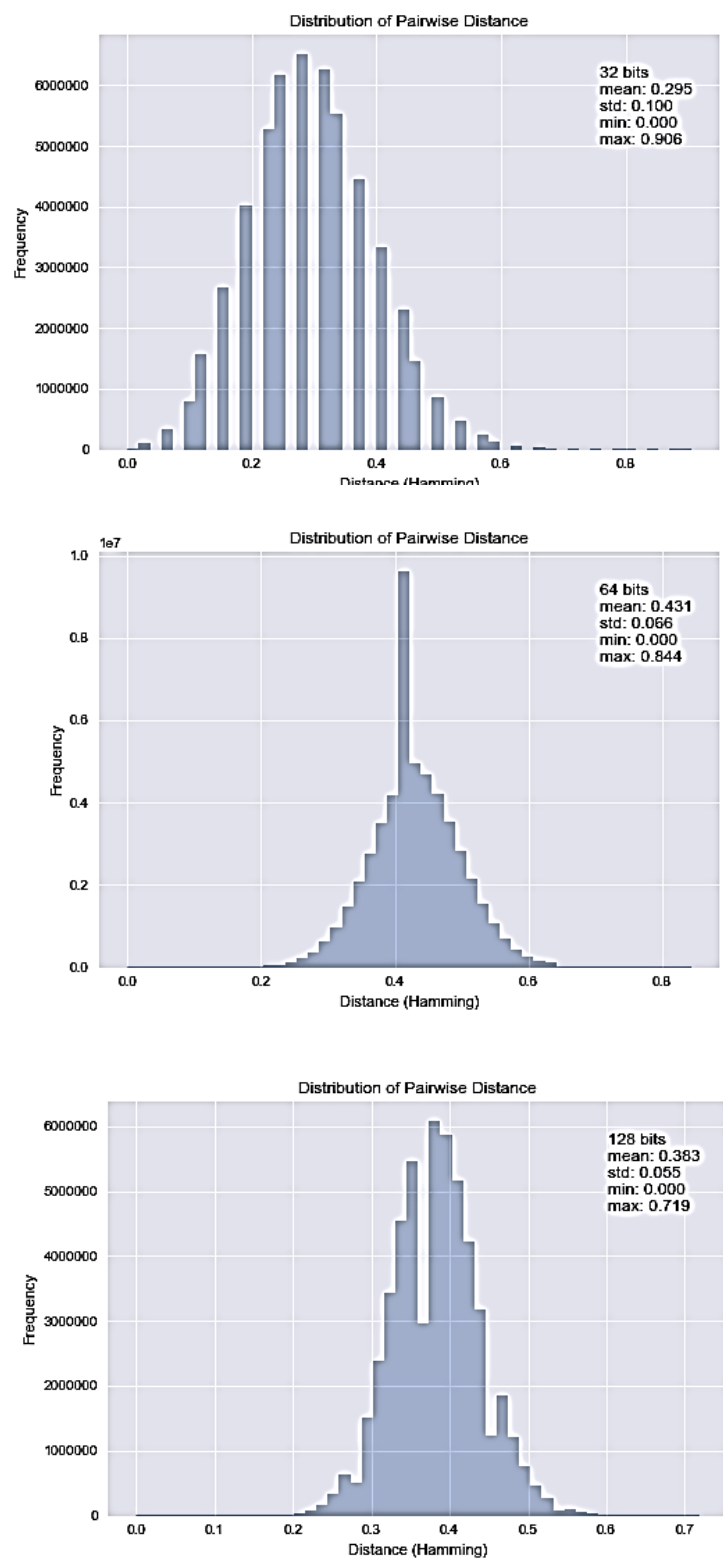


Figure 6-6 Distribution histograms of pairwise distances for molecules in MDDR represented by various fingerprint dimensions and measured by Hamming distance coefficient

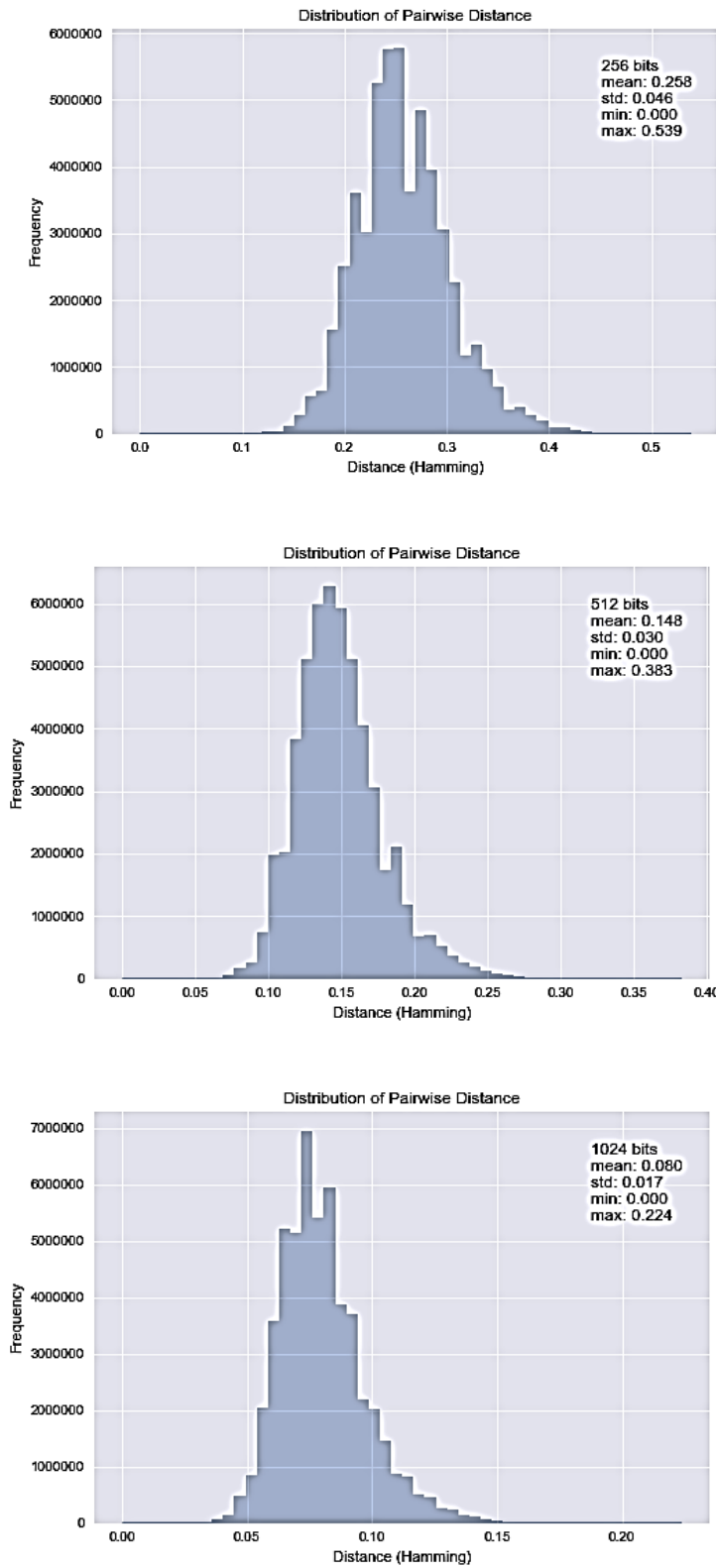


Figure 6-6 (continued)

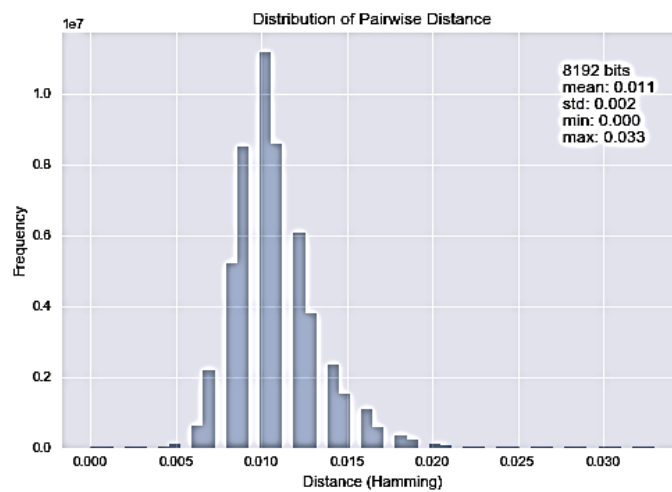
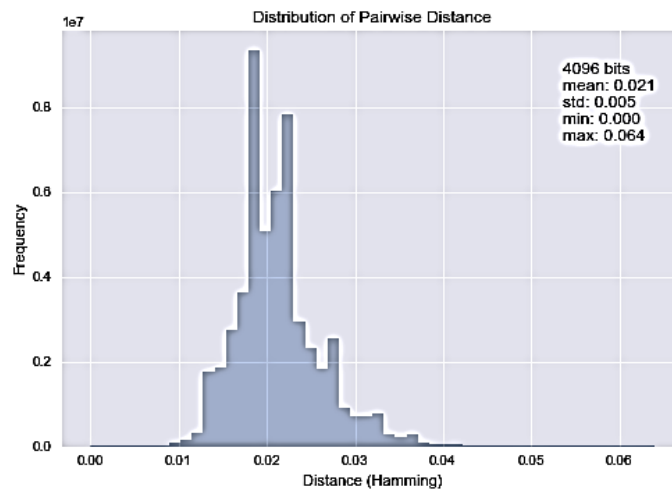
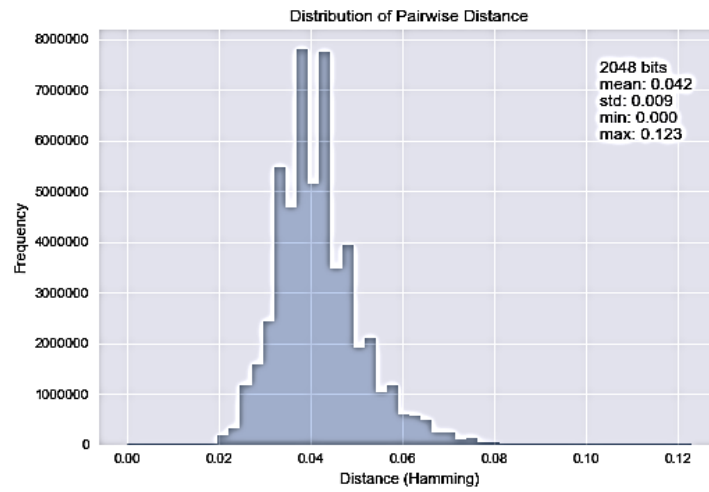


Figure 6-6 (continued)

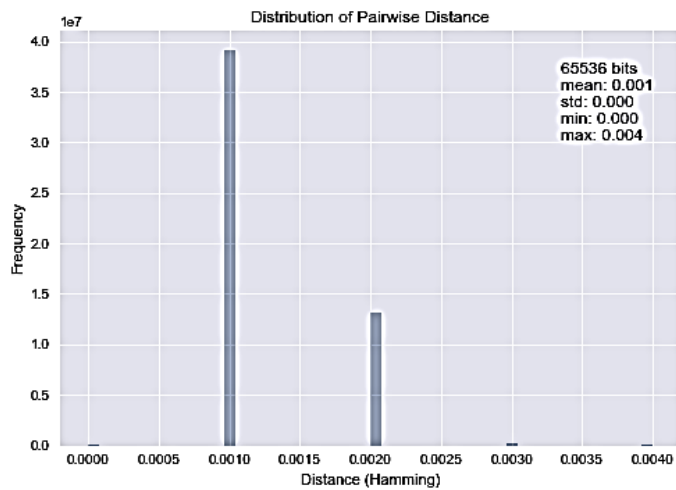
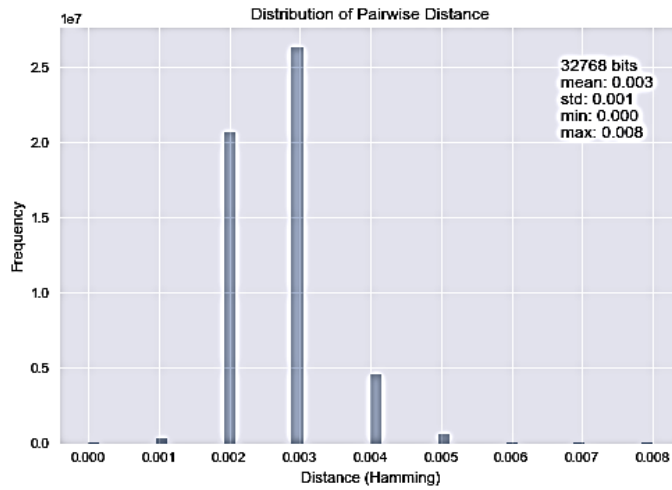
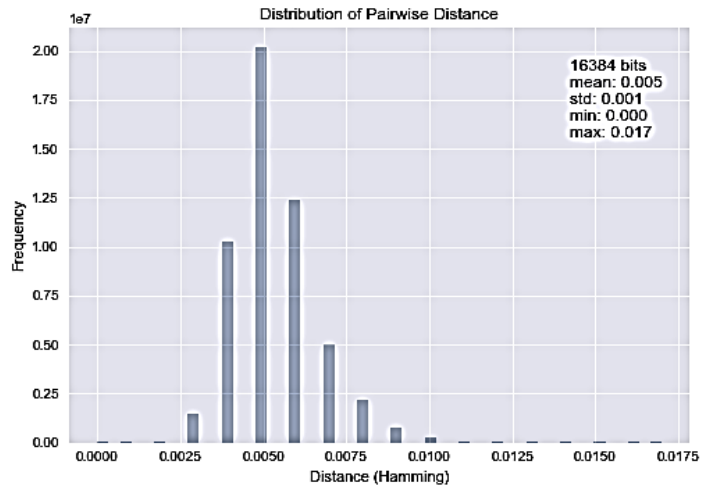


Figure 6-6 (continued)

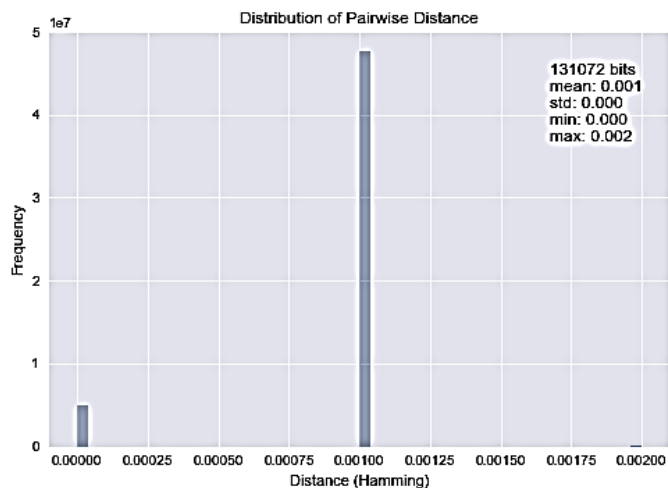


Figure 6-6 (continued)

In this situation, no substantial separation between the molecules can be found in the higher dimensional space. It is expected that the relative difference of the distances of the closest and furthest neighbours is zero. As an effect, the clustering process will likely be not effective because it is almost impossible to distinguish between the nearest or the furthest molecule (or even the active or inactive molecules) because they are all approximately at the same distance level.

These criteria affect the Ward's clustering because non-discrimination between the molecules resulted in the difficulty to quantify the minimal variance for the merger. Therefore, this could be the basis for the behaviour observed in the second trend evaluated by both *QPI* and *F* methods in Figure 6-4.

Finally, similar trends were observed by using the Kulsinski [D7], Rogers-Tanimoto [D8] and Russell-Rao [D9] distance coefficients suggesting similar behaviour to the distance distributions. In addition, similar findings using the Group Average algorithm and the WOMBAT dataset can be found in Appendix B.

6.3.3 Effects of Clustering Partition on F Measure and QPI Measure

A general observation on Figure 6-4 indicates that the effectiveness values of both F and QPI measures increased as the number of cluster partitions increases from 500 to 1000 partitions across almost all dimensions and distance coefficients. This can be seen by the coloured line plots, which are mostly plotted in a sequence of the lowest effectiveness value being from 500 partitions and increasing up to 1000 partitions, i.e., black (500), red (600), green (700), blue (800), turquoise (900) and magenta (1000). Similar behaviour can be observed from using the Group Average algorithm and the WOMBAT dataset in Appendix B.

It can be seen that in most cases, the larger number of cluster partitions have resulted in the higher values of QPI and F measures. These results demonstrate the effectiveness of small clusters in separating the actives and inactives, and identifying the best cluster with a balance of precision and recall. Therefore, this finding suggests the use of a larger number of cluster partition to obtain optimum effectiveness for molecular clustering.

6.4 Conclusion

The molecular clustering application implements the distance between molecules as a basis for grouping the molecules. Many studies have been conducted on clustering involving the search for efficient cluster algorithms and effect of distance coefficient. These applications typically involved high dimensional descriptors as the molecular representation. However, to the researcher's knowledge, there are no previous studies conducted on the effect of a high dimensionality dataset on the performance of the molecular clustering in the chemoinformatics context.

This chapter investigated the effect of changing the dimensionality of molecular representations on the effectiveness of molecule clustering applications. It aimed to observe the performance of the clustering application using various descriptor dimensions and distance coefficients used in this application.

The findings suggest two main conclusions. First, the effectiveness of molecular clustering increases with the increase of the fingerprint dimension until it reached a certain maximum value and remains at similar levels thereafter. This finding suggests that the molecular cluster performance is not affected by the changes of the fingerprint dimension. This finding is in line with the result obtained in the previous experiment in Chapter 5, which investigated the effectiveness of the similarity search application in high dimensionality.

Second, the findings are varying depending on the distance coefficient that is used to measure the distance of molecules during the clustering procedure. The effectiveness of molecular clustering decreases when the distance of the molecules is measured by the distance coefficients, which measure the distance of the molecules over the molecular fingerprint dimensions. This also suggests that, as the dimensionality increases, the ratio of distances between a molecule to its nearest and furthest neighbours becomes unity when measured by these types of distance coefficients. Hence, is it difficult to cluster molecules represented by very high dimensions as the distances between the molecules become incomparable.

This chapter also suggests two additional conclusions. First, the need to avoid the use of very small fingerprint dimensions, e.g., 2^5 or 2^6 bits dimension, which can result in more bit collisions, hence affecting the effectiveness of the molecular clustering. Second, smaller clusters are more effective than larger clusters in separating the actives and inactives in a dataset.

Chapter 7 Investigation into the Relative Importance of the Similarity Search Components using a Cross-Classified Multilevel Model

7.1 Introduction

Previous studies have evaluated the effects of different types of compound representations and similarity coefficients on similarity measures (Hert et al., 2004; Todeschini et al., 2012; Riniker and Landrum, 2013). The performance of a similarity measure is affected by the choice of both compound representation and similarity coefficient.

The molecular fingerprints are the most effective compound representations that describe compound features in several different ways. The performance of a similarity measure depends on the ability of the molecular fingerprints to describe the molecules (Riniker & Landrum, 2013). The similarity coefficients, on the other hand, are the mathematical measures that are derived from different formulations. The ability to quantify the degree of similarity for the similarity coefficients has been evaluated in previous research (Todeschini et al., 2012).

However, the measure of contribution to the overall effectiveness in similarity measure between the similarity components has not been investigated. Thus, this chapter aims to analyse the measure of contribution between the compound representations (i.e., molecular fingerprints) and the similarity coefficients to the enrichment factor. The investigation seeks to identify which component in the similarity measure matters more than the other. The results from the similarity search application will be investigated via a cross-classified multilevel approach to measure the contribution between the similarity components.

7.2 Cross-Classified Multilevel Modeling

Multilevel modeling is a statistical tool that is designed to model data based on its influence factor (Goldstein, 1987). In this approach, the influence factors are treated as different levels. The initial structure of the model involves a pure hierarchical data structure where data is nested within the higher levels. For example, a student is nested within the school. Hence, the student's achievement can be influenced by the school that the student attended (Goldstein, 2011).

However, in many cases, data can involve other potential influence factors which are not purely nested in the form of a hierarchical data structure. There can also be more than one type of influence factor in each level. For example, a student who attended more than one type of school in different neighbourhoods. In this case, the student's achievement can be influenced by the schools they attended and the neighbourhood they lived in. Incorporating neighbourhood as a further level is not straightforward since schools and neighbourhoods are not strictly nested within one another.

Cross-classified multilevel modeling can be used to model and analyse such complex non-hierarchical data structures (Goldstein, 1987). It has been applied to investigate various potential influence factors in areas such as education (Garner & Raudenbush, 1991; Leckie, 2009) and sport (Bell et. al., 2016). This model decomposes the total variance of the *response variable* into separate components in order to estimate the variance contributed by each influence factor, i.e., *influence variable*. It measures the proportion of the observed response variation that lies at a given level of the model and represents the percentage variance explained by the levels. Hence, it allows making conclusions about the relative importance of different sources of influence (different levels) on the response (Goldstein, 2011). As well as assessing the overall influence of a given level, the model can estimate and rank the magnitude of individual random effects (i.e., different types of influence variables).

In the chemoinformatics context, each enrichment value of a similarity search can be produced in part by a combination of compound representation, similarity coefficient and weighting scheme. In this case, the effectiveness of the similarity searches will generally be influenced by the molecular representation, the similarity coefficient and the weighting scheme. Common similarity search applications will involve many types of molecular representations, similarity coefficients and weighting schemes. There are also other potential influence factors that have an effect on the enrichment value, such as the bioactivity of the molecule and the specific reference structure used.

Since the effectiveness of a similarity search can result from many components, it is important to investigate which component is strongly affecting the effectiveness. As mentioned in Section 7.1, the measure of contribution to the overall effectiveness in similarity measure between the similarity components has not been investigated in previous studies.

The cross-classified multilevel model can be used to identify the importance of these similarity measure components that contribute to the effectiveness of similarity-based virtual screening. The total variation in similarity search effectiveness (i.e., response variable) can be modelled as the sum of contributions from various influence variables that are the molecular representation, similarity coefficient, bioactivity and weighing scheme (Figure 7-1).

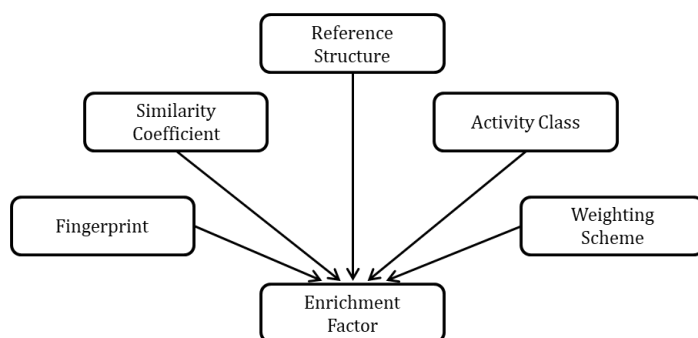


Figure 7-1 Diagram illustrating the influence variables of the enrichment factor in similarity search application

However, the focus of this chapter is to evaluate the relative importance between the compound representation (i.e., binary fingerprint) and the similarity coefficient components. The weighting scheme will not be implemented in the similarity search as the compound representation is not weighted fingerprints, i.e. integer or real values fingerprints that denote the relative importance of the fragments. Hence, its role as an influence variable will not be investigated in this study.

7.3 Model Implementation

The cross-classified models were run in MLwiN version 2.36 (Rasbash et. al., 2012) using the `runmlwin` command in Stata (Leckie & Charlton, 2013). The MLwiN is a software package that allows users to set up, fit and manipulate multilevel models. The parameter variances are estimated based on a *Bayesian* algorithm using *Markov chain Monte Carlo (MCMC)* estimation (Browne, 2015).

In the Bayesian algorithm, the probability of finding a certain value for the unknown parameter given the data (i.e., *posterior probability*), is proportional to the probability prior to the experiment (i.e., *prior probability*) multiplied by the *likelihood function*. In relation to the cross-classified model, each parameter of the model is equivalent to the unknown parameter in the Bayesian algorithm. For example, a cross-classified model such as the one defined below has three parameters to be estimated:

$$y_i = \beta_0 + u_{j2} + u_{j1} + e_i \quad (34)$$

where y_i is the response variable, β_0 is the fixed parameter and u_{j2} , u_{j1} and e_i are considered as the unknown parameters in the Bayesian algorithm. The algorithm measures the posterior probability distribution for each parameter and combines the posterior probability distributions into the *joint posterior distribution*.

7.3.1 MCMC Estimation

The MCMC method is a simulation-based procedure that aims to generate sample points (i.e., draws) in the space defined by the joint posterior distribution of all the parameters. The generation of the sample point is based on the *proposal distribution* as defined by Eq. (35):

$$\text{Draw } \theta_t \sim N(\theta_{t-1}, \sigma) \quad (35)$$

where $\text{Draw } \theta_t$ is a sample point of parameter, θ , for iteration, t , and σ is an arbitrary deviation. The sample points are generated using a user defined *starting value*, for example, $\theta_t = 1$. This then makes a large number of iterated random estimates; each iteration produces a new estimation value that is dependent on the estimation value from its previous iteration, θ_{t-1} .

These random estimates form a summary of the underlying distributions. It is then possible to calculate the posterior means and the standard deviations of the complete posterior distributions. MCMC is implemented because of its ability to handle more complex statistical models and structures.

In this method, the initial sample points may not be from the desired posterior distribution. It depends on the starting values in which the chain of iterations may take some time to converge. The period before a chain has converged is known as the *burn-in*. This part of the chain will be discarded. The remaining chain is known as the *monitoring chain*. The summary statistics of the monitoring chain provide the means and standard deviations for the model parameters.

A longer monitoring period can assure that the method has fully explored the parameter space and the chain has converged to yield a reliable estimate, that is, the chain is not trending in a particular direction. In MLwiN, the default value for burn-in length is 500 iterations and monitoring chain length is 5000 iterations. However, the length of iterations can be increased for better convergence.

7.3.2 MCMC Diagnostics

A wide range of MCMC diagnostics can be used to check the convergence of MCMC models. This is important to give an indication of whether the chain has been run for long enough to provide robust values for the mean and standard deviation of the estimated parameters. This experiment used two MCMC diagnostics that are commonly used in cross-classified model analysis (Rasbash et. al., 2012).

First, a visual inspection of the monitoring chain trajectories window in MLwiN for each parameter estimated was performed. Through visualisation inspection, an equilibrium pattern or stationary distribution in the trajectories indicates that the chains have sufficiently converged.

A second common diagnostic is the quantification of the *effective sample size (ESS)*. During the MCMC iterations, it is common for the value of the draw to be correlated with the value of the preceding draw i.e., *autocorrelation*. This is because each subsequent sample is drawn by using the current sample as mentioned in Section 7.3.1.

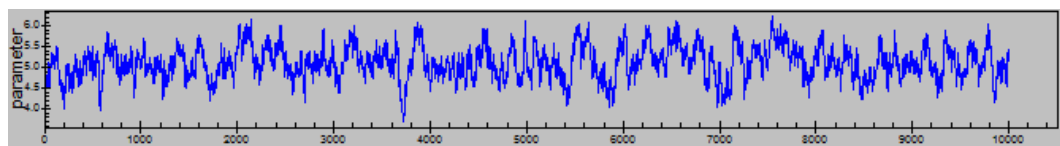
The ESS measures the number of iterations in a way that accounts for the autocorrelation of the chain. It is automatically calculated in MLwiN using the implementation by Browne (2015). It defines the ESS as the number of iterations, n , divided by a measure of the correlation of the chain called the autocorrelation time, ρ_k :

$$ESS(n) = \frac{n}{1 + 2 \sum_{k=1}^{\infty} \rho_k} \quad (36)$$

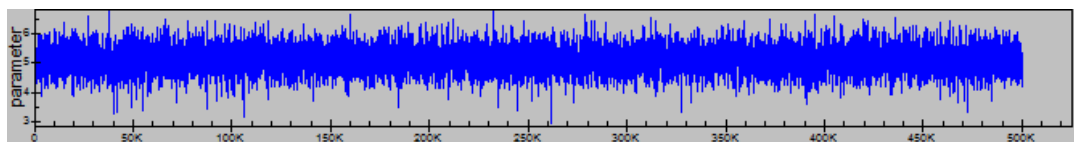
A higher ESS number indicates high independence (or less autocorrelation) in the chain and thus provides more information about the posterior distribution. It is common practice to terminate the simulation once the ESS is greater than a pre-specified threshold. This experiment uses a rule of thumb for sample size of

at least 400 iterations for all parameters. It is considered enough for the model to make a reasonable estimation of the posterior mean.

The example given in Figure 7-2 shows a comparison of the visual diagnostics for one model which runs for two different numbers of iterations; (a) 10,000 and (b) 500,000. For both trajectories, the X axis represents the number of iterations and the Y axis represents the draws from the parameter estimate.



(a)



(b)

Figure 7-2 Comparison of two visual diagnostics for monitoring chain trajectories of one model which runs for different iterations; (a) 10,000 iterations and (b) 500,000 iterations

The monitoring chain trajectories in Figure 7-2 (a and b) are the examples of trajectories which have resulted from the current experiment. The trajectories showed different behaviour as explained below:

(a) Inconsistent-looking graph which has the estimated posterior mean = 5.113 and ESS = 156 iterations. This is considered low (i.e., not enough) for the model to make a reasonable estimation of the posterior mean as the effective sample size is less than 400 iterations. Whilst there is no trending, the chain is not long enough to promote confidence in the results.

(b) Consistent-looking graph which has the estimated posterior mean = 5.126 and ESS = 7,062 iterations. Here, the chain is much longer and the ESS is also

higher. This is considered enough for the model to make a reasonable estimation of the posterior mean as it produces an effective sample size much higher than 400.

7.4 Experimental Design

The existing similarity search experiment uses several types of fingerprints and similarity coefficients that are combined with each other. There are ten different types of fingerprints, which describe a compound's different features as listed in Table 4-4. The features were hashed into the bits in the binary fingerprints. All fingerprints were generated for a size of 1024 bits using the RDKit from the KNIME software (Landrum, 2016). The thirty-one similarity coefficients used in this experiment were the same as the previous experiment described in Chapter 5.

Ten random reference compounds from each of 15 activity classes in the ChEMBL dataset were used for the similarity search, resulting in a total of 46,500 similarity searches (i.e., 10 reference compounds, 15 activity classes, 10 types of fingerprints and 31 types of similarity coefficients). The effectiveness of these similarity searches was measured based on the top 1% enrichment factor ($EF_{1\%}$). The variables used in this study are summarised in Table 7-1.

Table 7-1 Variables used in this study

Variables	Descriptions
Fixed part variable	
Enrichment factor	The dependent variable: The overall effectiveness of each similarity measure
Constant	The variable associated with the intercept coefficient
Random part variables	
Activity class	The chemical dataset grouped by similar biological properties (e.g. 5HT)
Fingerprint	The representation of chemical compound (e.g. ECFP_4)
Similarity coefficient	The measurement that quantifies the degree of similarity (e.g. Tanimoto)
Reference structure	The chemical compound used as reference structure in similarity search

7.5 Initial Model

An initial cross-classified model with four levels was implemented for all similarity search results. The model will decompose the total variance of the

enrichment values (i.e., the response variable) into separate activity classes, fingerprints, similarity coefficients and similarity searches (i.e., the reference structures) variance components (i.e., the influence variables). A basic, null model can be expressed as:

$$\begin{aligned}
 ef_i &= \beta_0 + u_{classid(j3)}^{(4)} + u_{fpid(j2)}^{(3)} + u_{coefid(j1)}^{(2)} + e_i \\
 u_{classid(j3)}^{(4)} &\sim N(0, \sigma_{u^{(4)}}^2) \\
 u_{fpid(j2)}^{(3)} &\sim N(0, \sigma_{u^{(3)}}^2) \\
 u_{coefid(j1)}^{(2)} &\sim N(0, \sigma_{u^{(2)}}^2) \\
 e_i &\sim N(0, \sigma_e^2)
 \end{aligned} \tag{37}$$

where ef_i is the observed enrichment value for a given similarity search i ($i = 1, \dots, 46,500$), β_0 is the mean $EF_{1\%}$ across all activity classes, fingerprints and similarity coefficients, $u_{classid(j3)}^{(4)}$ ($classid(j3) = 1, \dots, 15$) is the effect of similarity search i 's activity class, $u_{fpid(j2)}^{(3)}$ ($fpid(j2) = 1, \dots, 10$) is the effect of similarity search i 's fingerprint, $u_{coefid(j1)}^{(2)}$ ($coefid(j1) = 1, \dots, 31$) is the effect of similarity search i 's similarity coefficient, and e_i is the level 1 residual error term, incorporating other factors (and random variation) that affect the enrichment value. The activity class, fingerprint, similarity coefficient and residual error are assumed independent and normally distributed with zero means and constant variances.

The proportion of the observed response variation can be measured at activity class, fingerprint, similarity coefficient and similarity search levels. As a result, it is possible to establish the relative importance of the activity class, fingerprint, similarity coefficient and level one residual variation as sources of variations to the enrichment values. Furthermore, the magnitude and ranking of individual

activity classes, fingerprints and similarity coefficients can also be examined using their individual random effects.

The model was fitted in the MLwiN software using the MCMC method as described in Section 7.3. A starting value of 1 and the default value for the burn-in length of 500 iterations were used. The model was run for 500,000 iterations of monitoring chain length. These values were found to be sufficient for the chains to have converged (i.e., monitored by the consistent-look of visual diagnostics in the model trajectories window). The ESS value was over 800 for all parameters of the model which indicates the number of independence (or less autocorrelation) samples in the 500,000 iterations. The results are presented and discussed in the following sections.

7.5.1 Relative Importance of Similarity Measures

The results from fitting the initial model in Eq. (37) for all 46,500 similarity searches using the ChEMBL dataset are listed in Table 7-2. It reports the variances and standard errors estimated for each level (i.e., component) in the model. Hence, the comparison of the relative importance between the activity class, fingerprint and similarity coefficient can be observed based on the estimated variance in each level.

As shown in Table 7-2, the mean $EF_{1\%}$ across all levels is estimated to be 12.725, with a standard error of 1.857. The effect of L4 variance (i.e., between-activity class variance) is estimated as 54.170 (S.E. = 23.635). The effect of L3 variance (i.e., between-fingerprint variance) is estimated as 4.689 (S.E. = 3.042) while the effect of L2 variance (i.e., between-similarity coefficient variance) is estimated as 4.222 (S.E. = 1.772). The residual error, i.e., reference compound as affect to the enrichment value, is estimated as 92.473 with a standard error of 0.607.

Table 7-2 Variance estimation of similarity search components (4 level cross-classified model) for ChEMBL dataset

Model No.	Dataset	Intercept (Mean EF)		Effect L4 (Activity Class)		Effect L3 (Fingerprint)		Effect L2 (Similarity Coefficient)		Effect L1 (Residual Error)	
		Variance	S.E.	Variance	S.E.	Variance	S.E.	Variance	S.E.	Variance	S.E.
1	ChEMBL	12.725	1.857	54.170	23.635	4.689	3.042	4.222	1.772	92.473	0.607

Overall, the residual errors are larger compared to the variances estimated for the activity class, fingerprint and similarity coefficient levels. The variance estimated for activity class level is also larger compared to the fingerprint and similarity coefficient variances. This indicates larger disparities between the activity classes and the reference structures as compared to the fingerprint and similarity coefficient components. However, the difference between the fingerprint and similarity coefficient variances is relatively small. This shows that the fingerprint and the similarity coefficient are almost equally important to the enrichment value when considered across the entire dataset of searches.

The larger variation for the residual error (i.e., 92.473) can be due to the iterations of the model and the different structure of the reference compounds. All similarity values that were fitted in this cross-classified model were measured from 15 different activity classes which have different properties. As mentioned in Section 7.4, ten reference compounds were chosen randomly from each activity class to be measured in the similarity search experiment. These reference compounds were structurally different depending on which activity class they belong to. Hence, the residual error is large as it is affected by the nature of the activity class. This is supported by the estimated variance for the activity class level, which is the second largest after the residual error (i.e., 54.170). The variance between-activity class and individual ranking will be discussed in the next section.

7.5.2 Estimation of the Individual Activity Class Effect

Figure 7-3 presents the caterpillar plot for the activity class variable effect (i.e., level 4) estimated by the model. The plots in the diagram indicate the ranking of different types of activity classes used in this experiment. They were ordered by the value of *residuals* (i.e., predicted activity class effect). The horizontal scale indicates the rank order with vertical scale surrounded by 95% Bayesian confidence interval (CI) limits.

The residual value represents the difference when compared to an average activity class, i.e., the grid line y -axis equal to zero. Higher residual values indicate a better rank position. The activity class with the highest residual value will be ranked highest and can be considered the best according to the model. The activity classes with the CIs that do not overlap the grid line y -axis equal to zero (i.e., the average line) are considered 'better than the average'.

As shown in Figure 7-3, the highest ranked activity class is AT1. This is followed by the SubP, MMP1, HIVP, PKC, Thrombin, 5HT3, AChE, PDE4, COX, 5HT1A, D2, Renin, FXA and 5HT activity classes. According to the model, four out of 15 activity classes are considered better than the average (i.e., activity classes with CIs that do not overlap the average line). These are the top four activity classes, i.e., the AT1, SubP, MMP1, HIVP activity classes.

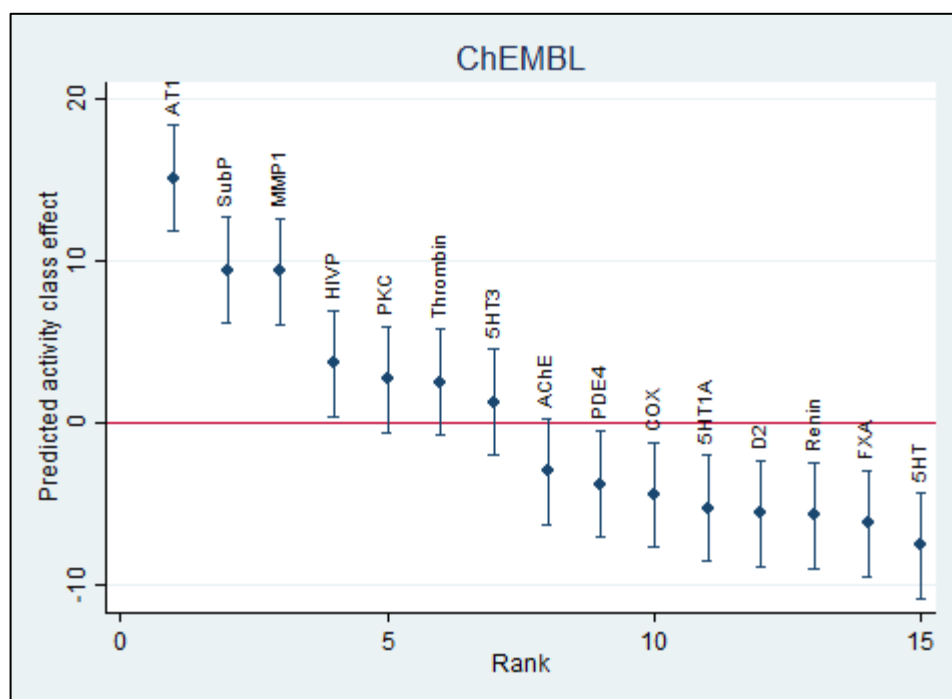


Figure 7-3 Caterpillar plot of the activity class-level residuals with 95% Bayesian credible intervals for ChEMBL dataset

It is also interesting to observe that all but one homogeneous class (i.e., Renin) were ranked highest. This also indicates that the $EF_{1\%}$ results produced by the homogeneous classes (except for Renin) are higher than the heterogeneous classes, in which the $EF_{1\%}$ has been used as the response variable for the cross-classified model implementation. The level of homogeneity for each activity class in the ChEMBL dataset can be referred to the mean pairwise similarity value (MPS) in Table 4-3 in Chapter 4.

There is a possible reason for this occurrence. In similarity search applications, the homogeneous class is expected to produce higher $EF_{1\%}$ results than the heterogeneous class. This is because the compounds that belong to the homogeneous class are structurally more similar than the compounds that belong to the heterogeneous class. It will be easier to differentiate between the actives from inactives for the homogeneous class compared to the heterogeneous class. Hence, the performance of the similarity searches for homogeneous activity class is higher.

However, in the case of Renin, further observation of the $EF_{1\%}$ values resulted from this activity class showed that they are relatively low compared to the other $EF_{1\%}$ values resulting from the other homogeneous classes. This is probably due to the reference structures that have been randomly selected for the Renin activity class. The use of these reference structures may affect the effectiveness of the similarity search results and also the ranking of Renin in the cross-classified model.

The fingerprint and the similarity coefficient components have been found to be almost equally important in this model. However, it would still be interesting to observe the individual effect of the various types of fingerprints and similarity coefficients, across the entire dataset of searches. Therefore, the following sections present the results of an individual effect for each component.

7.5.3 Estimation of the Individual Fingerprint Effect

Figure 7-4 presents the caterpillar plot for the fingerprint variable effect (i.e., level 3) estimated by the model. As shown in Figure 7-4, the highest ranked fingerprint is MorganR2. This is followed by the FeatMorganR2, MorganR1, Torsion, Atom Pair, FeatMorganR1, Avalon, Layered, RDKit and Pattern fingerprints. According to the model, three out of ten fingerprints are considered better than the average (i.e., fingerprints with CIs that do not overlap the average line). These are the top three fingerprints, i.e., MorganR2, FeatMorganR2 and MorganR1.

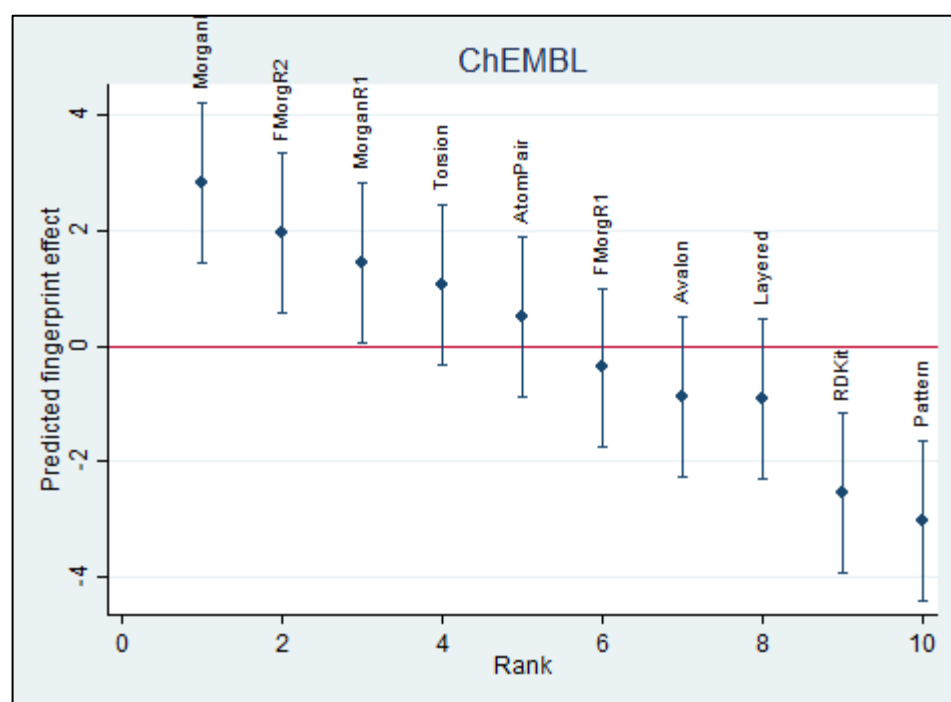


Figure 7-4 Caterpillar plot of the fingerprint-level residuals with 95% Bayesian credible intervals for ChEMBL dataset

From the rankings, it can be observed that all six fingerprints that were ranked on the top are the similarity types of fingerprints (refer Table 4-4). The remaining four fingerprints that were ranked lower are the substructure types of fingerprints. Three of the four circular fingerprints, Morgan R2,

FeatMorganR2 and MorganR1, were indeed found to be significantly better than the average in this model. It is also observed that one of the topological types of fingerprints, Torsion, has been ranked among the top circular fingerprints. This implies that Torsion fingerprint has a certain degree of discrimination ability, which is similar to circular fingerprints.

The finding of the top ranked fingerprints supports previous research that has examined the comparison of 2D fingerprints used for similarity-based virtual screening with multiple reference structures (Hert et al., 2004). The study conducted on the MDDR dataset found that the circular types of fingerprint are generally more effective, with the best results obtained from the ECFP_4 fingerprints (i.e., the Morgan R2 in this investigation).

A more recent research, which implemented the similar types of 2D fingerprints used in this study, has also been reviewed. Riniker and Landrum (2013) developed an open-source platform for virtual screening to evaluate the performance of 12 commonly used fingerprints. Six of the 12 types of 1024 bits fingerprint used in the previous study have been used in this experiment, i.e., Atom Pair, Torsion, RDKit, Avalon, ECFP_4, FCFP_4. For RDKit, the maximum path length that was used in the previous study (i.e., path length of 5) is different than the path length used in this experiment. This is because this experiment uses the default maximum path length which is 4.

Riniker and Landrum (2013) have found that the circular fingerprints are generally ranked higher by the enrichment factor as the evaluation method; which are consistent with the finding in this research. Another interesting finding is that the Torsion fingerprint has been found to be exceptionally ranked among the top fingerprints by all of the evaluation methods used. This matches the finding of this study, in which the Torsion has been ranked among the top circular fingerprints as shown in Figure 7-4.

7.5.4 Estimation of the Individual Similarity Coefficient Effect

Figure 7-5 presents the caterpillar plot for the similarity coefficient variable effect (i.e., level 2) estimated by the model. The highest ranked fingerprint is the B37 (Maxwell-Pilliner) similarity coefficient as shown in Figure 7-5. This is followed by the B38, B34, B26, B18, B30, B29, B19, B42, B3, B28, B22, B9, B33 and B10 similarity coefficients. These fifteen similarity coefficients are estimated to be significantly better than the average by the model. The remaining similarity coefficients in the ranking were B51, B11, B8, B17, B46, B23, B15, B25, B16, B43, B20, B1, B36, B5, B7 and B6.

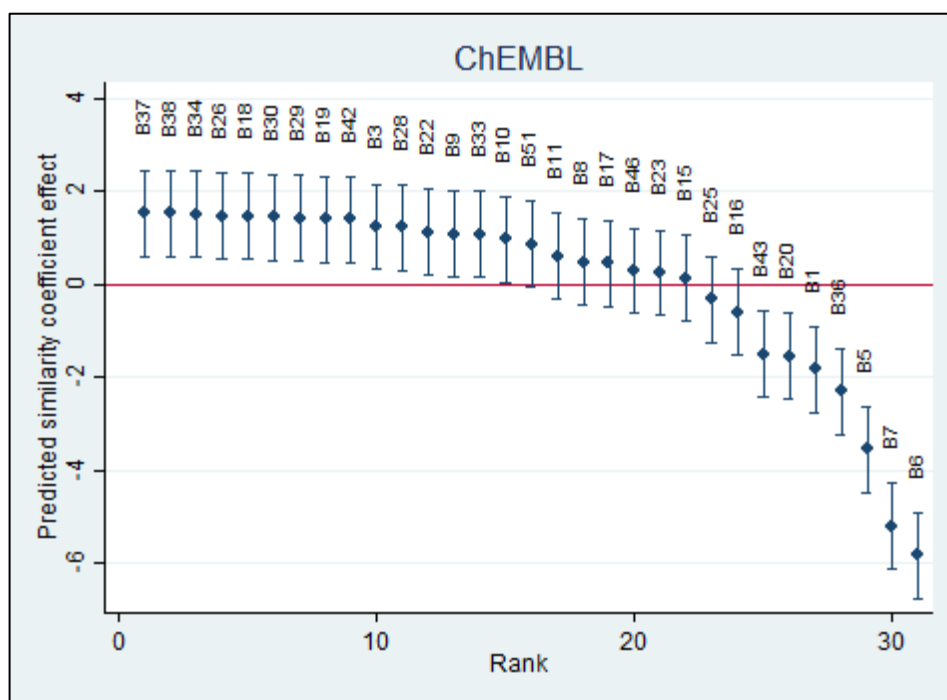


Figure 7-5 Caterpillar plots of the similarity coefficient-level residuals with 95% Bayesian credible intervals for ChEMBL dataset

Further observation of Figure 7-5 showed that many of the higher ranked similarity coefficients are plotted almost equally on the same horizontal line. This is another way of illustrating the variances resulting from using the similarity coefficients. It indicates that there are almost equally similar

variances and small differences of the residual values among the similarity coefficients.

A previous study by Todeschini et al. (2012) has observed that the similarity coefficients B38 (Harris-Lahey) and B42 (CT4) yield very good results on their retrieval abilities in similarity based virtual screening using the WOMBAT dataset. Both similarity coefficients were also superior to the well-established B3 (Jaccard-Tanimoto). The present findings seem to agree with Todeschini et al. (2012), who showed that the B38 similarity coefficient is ranked among the top (i.e., second rank with variance = 1.528). The first ranked is B37 with a variance of 1.529. The B42 similarity coefficient was at rank nine. All of the similarity coefficients were still ranked higher than B3 (i.e., 10th in rank).

7.6 Extended Model I

Results from the previous section have shown that the variation of the cross-classified model is highly influenced by the activity class component. Therefore, further analysis was conducted to investigate the importance of the components independently of the activity classes. Fifteen three-level models were developed and implemented in this analysis, one for each activity class in the ChEMBL dataset. Each model uses only the total number of 3,100 $EF_{1\%}$ values resulting from the similarity searches for a particular activity class. The model can be expressed by:

$$\begin{aligned} ef_i &= \beta_0 + u_{fpid(j_2)}^{(3)} + u_{coefid(j_1)}^{(2)} + e_i \\ u_{fpid(j_2)}^{(3)} &\sim N(0, \sigma_{u^{(3)}}^2) \\ u_{coefid(j_1)}^{(2)} &\sim N(0, \sigma_{u^{(2)}}^2) \\ e_i &\sim N(0, \sigma_e^2) \end{aligned} \tag{38}$$

where the response variable, ef_i is the observed enrichment value for a given similarity search i ($i = 1, \dots, 3,100$), β_0 is the mean $EF_{1\%}$ across all fingerprints and similarity coefficients, $u_{fpid(j2)}^{(3)}$ ($fpid(j2) = 1, \dots, 10$) is the effect of similarity search i 's fingerprint, $u_{coefid(j1)}^{(2)}$ ($coefid(j1) = 1, \dots, 31$) is the effect of similarity search i 's similarity coefficient, and e_i is the level 1 residual error term, incorporating other factors (and random variation) that affect the enrichment value. The fingerprint, similarity coefficient and residual error are assumed independent and normally distributed with zero means and constant variances.

The models produced the proportion of the observed response variation and individual random effects at fingerprint, similarity coefficient and similarity search levels. Hence, the relative importance and individual random effects can only be examined on these levels. The model was fitted in the MLwiN software with the similar settings as described by the previous model. The results are discussed in the following sections.

7.6.1 Relative Importance of Similarity Measures

The results from fitting the model in Eq. (38) for all activity classes of the ChEMBL dataset are listed in Table 7-3. It reports the variances and standard errors estimated for each parameter (i.e., component) of all fifteen cross-classified models. The relative importance between the fingerprint and the similarity coefficient levels can be compared in Figure 7-6.

The values in Table 7-3 indicate that the estimated variances and standard errors of mean $EF_{1\%}$ vary depending on the nature of the activity classes. The highest mean $EF_{1\%}$ across all fingerprints and all similarity coefficients is from the AT1 activity class (i.e., the most homogeneous with MPS = 0.52). The estimated variance for this activity class is 27.806 with a standard error of 2.184. The lowest mean $EF_{1\%}$ (i.e., variance 5.125 of and S.E. of 0.398) is from the 5HT activity class which is one of the heterogeneous classes in the ChEMBL dataset (i.e., MPS = 0.34). The mean $EF_{1\%}$ variances resemble the ranking of

activity class obtained in Figure 7-3 if the variances are sorted in descending order.

Further observation of Figure 7-6 shows that in the majority of cases the fingerprint effect is higher than the similarity coefficient effect. This can be seen for the 5HT1A, 5HT3, AChE, D2, HIVP, MMP1, PDE4, PKC, Renin, SubP and Thrombin activity classes. The differences of these variance levels were also very large. The remaining four activity classes have higher similarity coefficient effects than the fingerprint effects, i.e., 5HT, AT1, COX and FXA. However, in contrast with the other eleven activity classes, the differences of these variance levels are relatively small

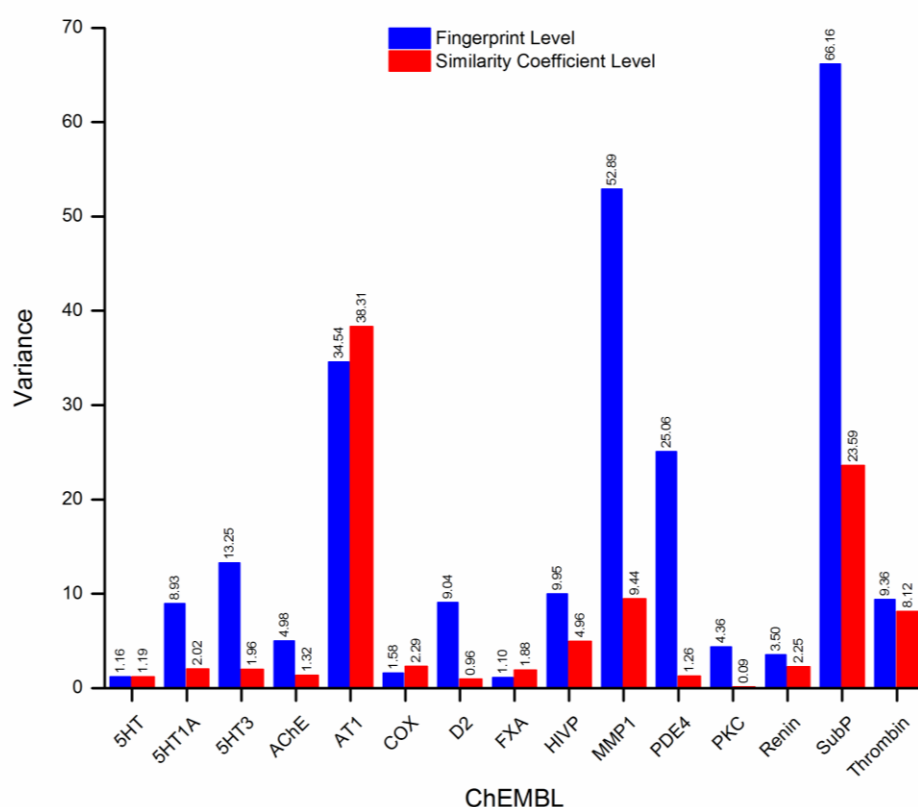


Figure 7-6 Bar chart comparing the relative importance between the fingerprint and similarity coefficient effects for 15 activity classes of ChEMBL dataset

Table 7-3 Variance estimation of similarity search components (3 level cross-classified model) for 15 activity classes

Model No.	Activity Class	MPS	Intercept (Mean EF) Variance	S.E.	Effect L3 (Fingerprint) Variance	S.E.	Effect L2 (Similarity Coefficient) Variance	S.E.	Effect L1 (Residual Error) Variance	S.E.
1	5HT	0.34	5.125	0.398	1.162	0.765	1.194	0.359	10.253	0.262
2	5HT1A	0.37	7.417	0.984	8.925	5.665	2.018	0.638	28.080	0.718
3	5HT3	0.35	13.983	1.206	13.247	8.896	1.962	0.985	147.707	3.781
4	ACHe	0.36	9.662	0.746	4.984	3.252	1.323	0.470	36.753	0.940
5	AT1	0.52	27.806	2.184	34.537	22.541	38.306	11.333	257.478	6.582
6	COX	0.28	8.244	0.491	1.584	1.075	2.286	0.698	23.194	0.593
7	D2	0.35	7.072	0.973	9.041	5.732	0.956	0.341	27.008	0.690
8	FXA	0.39	6.497	0.424	1.098	0.770	1.882	0.584	22.409	0.573
9	HIVP	0.43	16.373	1.080	9.951	6.351	4.958	1.481	38.332	0.980
10	MMP1	0.40	22.068	2.365	52.889	33.178	9.438	2.865	89.451	2.287
11	PDE4	0.31	8.929	1.583	25.062	15.625	1.258	0.417	24.397	0.624
12	PKC	0.42	15.359	0.712	4.360	3.185	0.091	0.173	189.193	4.815
13	Renin	0.45	6.962	0.658	3.496	2.295	2.250	0.704	28.582	0.731
14	SubP	0.43	22.115	2.734	66.164	42.232	23.585	7.274	263.219	6.729
15	Thrombin	0.35	15.174	1.100	9.361	6.009	8.118	2.377	45.596	1.166

Third column is a mean pairwise similarity value (MPS) that indicates the diversity of each activity class. Higher MPS value means higher inter-compound similarity (i.e., more homogeneous, threshold value of 0.40) and vice versa. The grey box indicates larger variance when compared between the variance estimated for L3 and L2 while the italic and bold faced indicate largest variance when compared between the variance estimated for L3, L2 and the residual error within the same activity class.

A sign test and a Wilcoxon signed-rank test were conducted to evaluate the differences of variances of the two components (i.e., fingerprint and similarity coefficient). For the Wilcoxon signed-rank test, the SPSS application will automatically measure the significance of the data using the large-sample test although there are only 15 pairs of observations ($N = 15$). This is acceptable as the large-sample test for the Wilcoxon signed-rank test appears to produce a good approximation even for relatively small samples (Siegel & Castellan Jr, 1988). In the present context, each variance acts as a judge of the effectiveness of the various activity classes, where the significant of the differences is measured by the number of (i) fingerprint level > similarity coefficient level, (ii) fingerprint level = similarity coefficient level and (iii) fingerprint level < similarity coefficient level.

The sign test resulted in the significance of the probability value of $\rho = 0.118$ that is higher than the significant level of $\alpha = .05$. This indicates that there is no significant difference in variances between the two components considering all 15 cross-classified models using the sign test. However, the result of the significance of the probability value using the Wilcoxon signed-rank test is $\rho = 0.008$. This means that when measured using the Wilcoxon signed-rank test, the variances between the two components are significantly different at $\alpha = .01$ level considering all 15 cross-classified models. The possible reason for this is because the Wilcoxon sign-rank test considers the direction and the relative magnitude in its measurement which makes it more powerful than the sign test.

The results in Table 7-3 also show that the estimated variances for the residual errors were still large compared to the fingerprint and the similarity coefficient levels. All but one residual error value are higher than the L3 and L2 variances (where fourteen residual values were emphasised in italic and bold faced in column 10). This indicates that even after the separation of activity classes, in most of the cases the variation influenced by the reference structure far outweighs the influence of the fingerprint and the similarity coefficient. This is supported by the results from one of the statistical tests, i.e., the sign test, that has shown no significant difference between the fingerprint and the similarity

coefficient components. Therefore, another investigation, which developed a different cross-classified model for each reference structure, has been conducted and will be described in Section 7.7.

7.6.2 Estimation of the Individual Fingerprint Effect

Figure 7-7 presents the heat map of fingerprint level reflecting the ranking of the fingerprints across all activity classes according to the model. The rows indicate the types of fingerprints while the columns represent the activity classes. Each cell point in the heat map represents the rank position, i.e., low rank positions tend towards darker green tones while high rank positions tend to hotter orange and red tones.

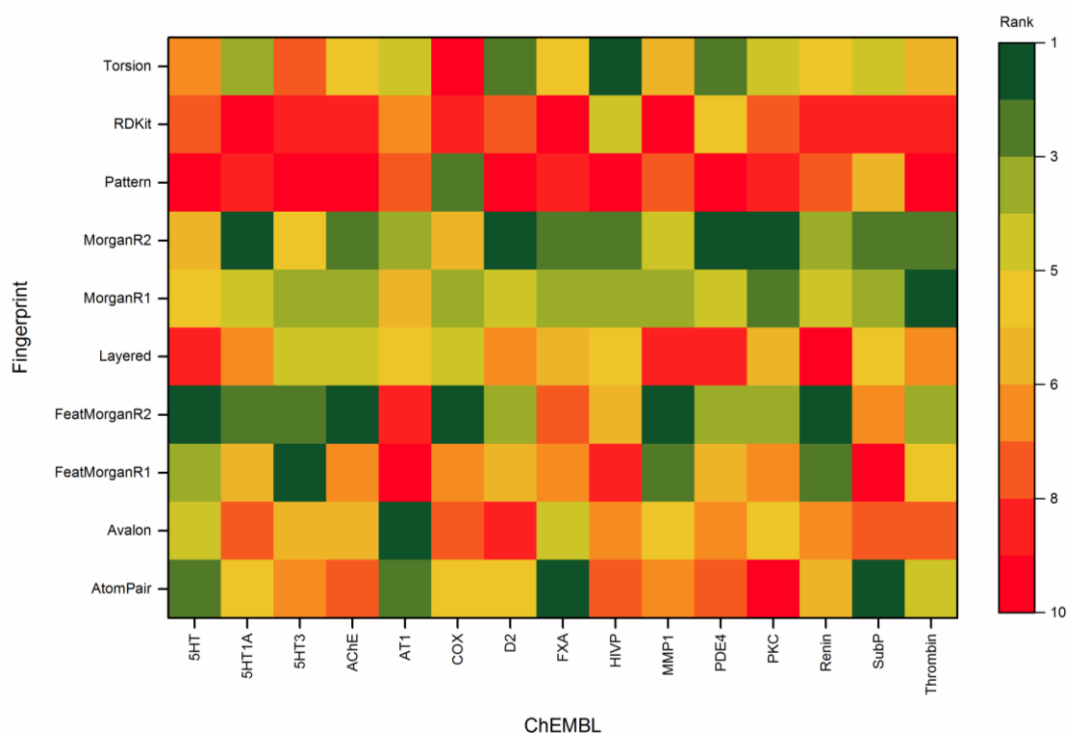


Figure 7-7 Heat map summarising the ranking of the variable effects for level 3 (fingerprint) for 15 activity classes of ChEMBL dataset

Overall, it can be seen that different fingerprints are best for different activity classes. Interestingly, five fingerprints were observed to reveal a consistent

ranking across all activity classes. The MorganR1, MorganR2 and FeatMorganR2 fingerprints were found to be mostly ranked among the top six fingerprints while the RDKit and Pattern fingerprints were found to be mostly ranked among the lowest. The top ranked fingerprints for most activity classes are the similarity fingerprints (i.e., circular type) while the low ranked fingerprints are the substructure fingerprints (i.e., topological type).

The details of these ranks are presented in Figure C-1 in Appendix C. The figure illustrates the caterpillar plots of the level 3 variable effects (i.e., the fingerprints) for each activity class. The plots in the diagrams were ordered by the value of residuals (i.e., predicted fingerprint effect). The horizontal scale indicates the rank order with vertical scale surrounded by 95% Bayesian confidence interval (CI) limits. The average fingerprint was determined by the same method explained in the Section 7.5.2 (second paragraph).

The variance of the fingerprint level for each of the activity class relates to the value of the fingerprint residual in the caterpillar plots. An activity class which has a high value of variance in the fingerprint level will also have high residual values between the fingerprints. By referring to Table 7-3, the SubP activity class has the highest value of variance in the fingerprint level that is 66.164 while FXA has the lowest value of variance, i.e., 1.098.

A previous study by Hert et al. (2004) has found the FeatMorganR2 (FCFP_4) fingerprints being better for heterogeneous classes while MorganR1 (EFCFP_2) being better for homogeneous classes in the MDDR dataset. In this study, FeatMorganR2 has been found to be very effective in both heterogeneous and homogeneous classes, e.g., 5HT (Model 1), AChE (Model 4), COX (Model 6), MMP1 (Model 10) and Renin (Model 13) in Figure C-1. The MorganR1 fingerprints have also been found to be among the most effective for both types of activity classes.

7.6.3 Estimation of the Individual Similarity Coefficient Effect

Figure 7-8 presents the heat map of similarity coefficient level reflecting the ranking of the coefficients across all activity classes. The rows indicate the types of similarity coefficients while the columns represent the activity classes. Similar to Figure 7-7, each cell point in the heat map represents the rank position, i.e., low rank positions tend towards darker green tones while high rank positions tend to hotter orange and red tones.

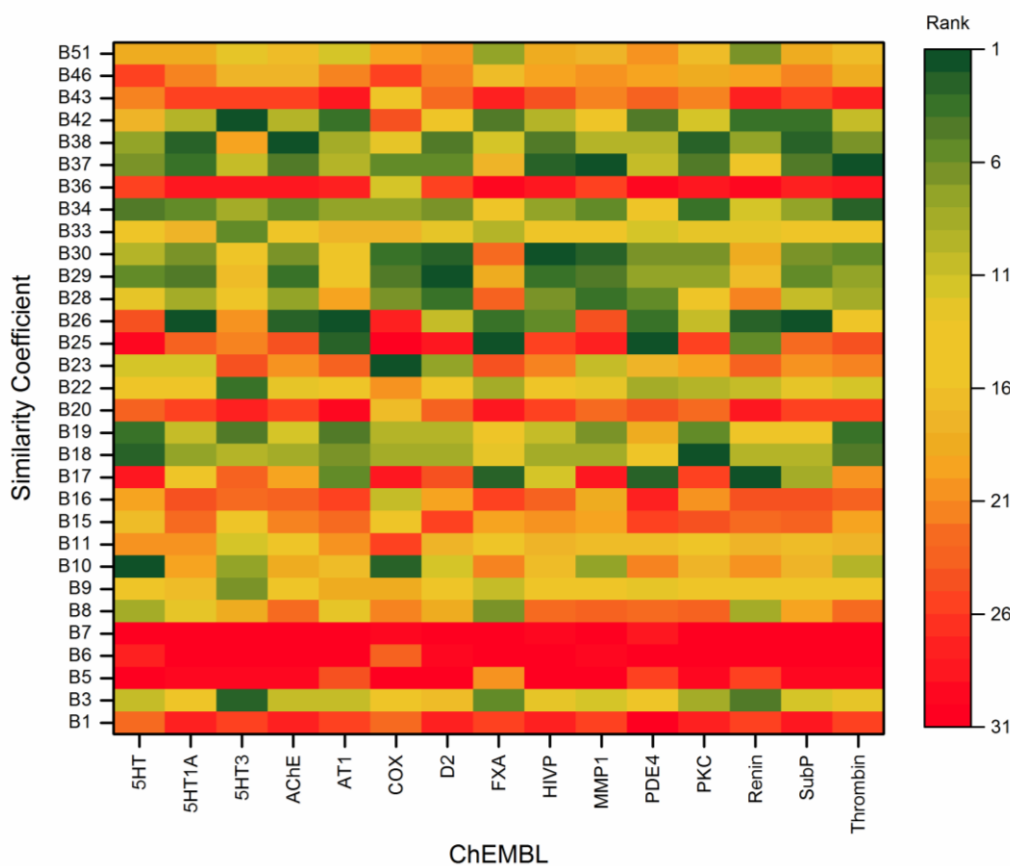


Figure 7-8 Heat map summarising the ranking of the variable effects for level 2 (similarity coefficient) for 15 activity classes of ChEMBL dataset

It can be observed that the higher or lower ranked similarity coefficients are easily identified across all activity classes. The B3 (Jaccard-Tanimoto), B18 (Rogot–Goldberg), B19 (Hawkins–Dotson), B34 (Cohen), B37 (Maxwell-Pilliner) and B38 (Harris-Lahey) similarity coefficients were visually observed to be

consistently ranked higher. B1 (Sokal–Michener, Simple Matching), B5 (Russel–Rao), B6 (Forbes), B7 (Simpson), B20 (Yule) and B36 (Peirce) were found to be consistently ranked lowest across all activity classes.

The details of these ranks are presented in Figure C-2 (Appendix C). The figure illustrates the caterpillar plots of the level 2 variable effects (i.e., the similarity coefficients) for each activity class. The highest value of variance in the similarity coefficient level can be observed from homogeneous classes, e.g., ATI (Model 5) and SubP (Model 14) in Figure C-2. For both classes, B26 has shown to be the highest ranked coefficient and B6 as the lowest ranked coefficient. For the most heterogeneous class COX (Model 6), the B23 has shown to be the highest ranked coefficient and the B5 coefficient at the lowest rank.

Model 12 which represents the PKC activity class has shown an interesting observation. The variances among the similarity coefficients were low and almost equally the same. These were indicated by the value of the residuals, which were nearly zero and showed an equal horizontal pattern in the caterpillar plot. This is the only case that has shown this behaviour across all activity classes.

In comparison with the previous study, B26 has also been found to work well and performed better than B3 in homogeneous classes of the MDDR and WOMBAT datasets (Todeschini et al., 2012). The B38 and B42 similarity coefficient were also found to rank higher in both homogeneous and heterogeneous classes and mostly ranked higher than the B3 coefficient.

7.7 Extended Model II

The previous results in Table 7-3 have shown that the estimated residual errors were still large compared to the fingerprint and similarity coefficient variances. This indicates that the similarity search experiments were still influenced by the variation of the reference compounds. The sign test also showed that the fingerprint is not significantly different from the similarity coefficient component.

Thus, this section will make a conclusion about the variances between the fingerprint and the similarity coefficient levels independently of the reference structures. Another 150 three-level models were developed, one for each reference structure. Each model uses only the total number of 310 $EF_{1\%}$ values resulting from the similarity searches based on a single reference compound.

Using the same model expression in Eq. (38), the response variable, ef_i is the observed enrichment value for a given similarity search i in which i equal to 1 until 310. The other parameters in the model remain the same. The models also produced the proportion of the observed response variation and individual random effects at fingerprint, similarity coefficient and similarity search levels. A better conclusion can be drawn about which component is more important between the fingerprint and the similarity coefficient based on these models. Next section discusses the results of the models.

7.7.1 Relative Importance between Fingerprint and Similarity Coefficient

The results from Table C-1 (Appendix C) report the variances estimated for each level of all 150 cross-classified models. A general inspection of the table showed that many variances estimated for the fingerprint were higher than the variances estimated for the similarity coefficient and residual errors. This can be seen by the models that have the L3 variances emphasised in italic and bold face.

Comparison between the fingerprint and similarity coefficient variances also showed that the fingerprint variance was superior to the similarity coefficient variance, which can be observed by 136 models, which have the L3 variances marked by grey boxes. Only 14 models have L2 variances higher than the L3 variances.

The same statistical tests were repeated to evaluate these performances. Both the sign test and the Wilcoxon signed-rank test indicate that there were statistically significant differences in variances between the two components

considering all 150 cross-classified models ($z = -10.024$, $\rho < .01$ for the sign test and $z = -9.880$, $\rho < .01$ for the Wilcoxon signed-ranks test). Hence, these tests showed that the fingerprint component is significantly more important than the similarity coefficient component in this study.

In addition to the changes of the variances above, it has also been observed that the residual errors have lessened compared to the L3 and L2 variances in most models. Only 18 out of 150 models have the residual errors larger than the other two levels of variances. This indicates that there were no higher variances between the reference structures as seen in the previous model in Section 7.5.2 because the current models were modelled based on each reference structure. Of all cases, only one homogeneous activity class still has many models with higher residual errors than the L3 and L2 variances, i.e., PKC (7 models). The other homogeneous classes were SubP (1 model), AT1 (2 models), HIVP (1 model) and Renin (3 models). The heterogeneous classes were 5HT (1 model), FXA (2 models) and Thrombin (1 model).

The results from using 150 models showed that the use of different reference structures can result in substantial difference in the more important component of a similarity search. A robust conclusion was made considering all of the reference structures used in this experiment. Hence, it highlights another important finding that the role of the number of reference structures is an important factor in the comparative study of similarity measures.

Arif et al. (2013) conducted a study that ranks different similarity measures based on the effectiveness of the similarity searches resulting from the use of different number of reference structures. The study found that rankings produced by the results of using all reference structures could be substantially different from the results of using a small number of reference structures.

The findings in the current experiment seem to support the findings by Arif et al. (2013). The models have shown that different reference structures can result in different identification of the relative importance between similarity

measures, and therefore, a conclusion can only be made using a considerably large number of reference structures.

7.8 Conclusion

This chapter has carried out a detailed investigation into the relative importance between the fingerprint representation and the similarity coefficient components in similarity-based virtual screening. The experiment involved the use of cross-classified multilevel modeling to estimate the variances produced by various factors contributing to the similarity search. These variances were analysed to identify the importance of the components.

The main findings in this study indicate that the fingerprint component is more important than the similarity coefficient in determining the effectiveness of similarity based-virtual screening. Based on the implemented dataset, the results suggest MorganR2 (ECFP_4) as the best fingerprint and B37 as the best similarity coefficient.

Compared with the previous studies by Hert et al. (2004), Riniker and Landrum (2013) and Todeschini et al. (2012), this study carried out a different investigation that combines both similarity search components, i.e., the molecular representation and the similarity coefficient. Many of the results from this study seem to match those observed in earlier studies.

Another important finding in this study also highlights the role of different reference structures in determining the relative importance of similarity measures. The use of large number of reference structures has allowed a robust conclusion to be made on the main findings, which seem to agree with the previous study by Arif et al. (2013). Therefore, the number of reference structures in determining the effectiveness of similarity search can be a basis for future studies of similarity search in virtual screening.

In addition to these practical findings, it was also observed that the influences of the biological activity and the reference structure were also very important. These influences have been shown by the high variances estimated by the

models in Sections 7.5 and 7.6. However, the generalisability of these results is subject to limitation such as the non-normality of the residuals which can be investigated in the future.

Hence, apart from the novelty of the cross-classified multilevel modelling and its implementation in chemoinformatics research, this chapter also highlights the importance of the similarity search component to help improving similarity-based virtual screening.

Chapter 8 Summary and Future Work

8.1 Introduction

This thesis has conducted three investigations: (1) The effects of dimensionality on the effectiveness of similarity search applications (reported in Chapter 5); (2) The effects of dimensionality on the effectiveness of clustering applications (reported in Chapter 6); (3) The relative importance of the fingerprint and the similarity coefficient components on the effectiveness of similarity searching using cross-classified multilevel model analysis (reported in Chapter 7). This chapter summarises the overall key findings.

8.2 Overall Summary of Work and Findings

The search for nearest neighbour molecules in chemoinformatics applications involves two important components: (1) the molecular representations or descriptors and (2) the similarity or distance coefficients (Willett et al., 1998). The molecules are usually represented by a very high dimensionality representation (Todeschini & Consonni, 2000). For example, a common number of bits used for a 2D binary circular fingerprint is 1024 bits. However, the fingerprint dimension could be higher depending on the space required to represent the structure of the molecule (Sastry et al., 2010). The similarity or distance coefficients quantify the similarity or the distances of the molecules based on various formulations which consider different attributes of the fingerprint representation (Todeschini et al., 2012). The search process starts with converting the molecules into various types of representations and then measuring the similarity (or distance) between the molecules using different types of coefficients. Based on the underlying similar property principle, the nearest molecule which has the closest distance (or is most similar) to the query molecule is considered as the molecule with the most similar properties to the query molecule (Johnson & Maggiora, 1990). The nearest neighbour search has become the foundation of many chemoinformatics applications such as similarity searching and clustering.

In other domains, increasing the dimensionality of data representations has been found to decrease the effectiveness of nearest neighbour searches, a phenomenon known as the curse of dimensionality (Bellman, 1961). It happens when the ratio of the distance of a query point to its nearest neighbour and to its furthest neighbour tends to unity measured by a distance coefficient (Agrawal et al., 1998; Weber et al., 1998; Beyer et al., 1999). Hence, the effectiveness of the nearest neighbour search decreases and the results become meaningless, i.e., difficult to distinguish between the nearest (most similar) or the furthest (most dissimilar) neighbour since the distances are almost the same (Clarke et al., 2008).

However, the effectiveness of nearest neighbour searches in the chemoinformatics domain does not seem to be affected by the use of high dimensionality representations. This behaviour has led this researcher to investigate the effect of nearest neighbour search in high dimensionality chemical datasets. Despite the proven effectiveness of the nearest neighbour search in chemoinformatics applications, a detailed study was needed to investigate the effects of nearest neighbour search when increasing the dimensionality of chemical datasets. This includes evaluating the effects of using different similarity or distance coefficients to the effectiveness of the searches. Experimental Chapters 5 and 6, were hence investigating the first aim of this study, i.e., the effects of dimensionality on the effectiveness of similarity searching and clustering applications.

The first experiment in Chapter 5 conducted a similarity search using three chemical datasets. Each molecule in the datasets was represented by thirteen different dimensions of ECFP₄-like binary fingerprints. The similarity between the reference molecules and the rest of the molecules in the datasets was measured using thirty-one non-monotonic similarity coefficients. The effectiveness of the application was evaluated based on the $\overline{EF}_{1\%}$ ranked molecules.

It was observed that an increase in fingerprint dimensions increases the effectiveness of the similarity search up to a certain fingerprint dimension

which is maintained thereafter. The evidence from this study suggested that this behaviour depends on the number of bits that is required to represent the information of the molecules. In addition to these findings, the variations in performance are due to the characteristics of the similarity coefficients used as the similarity measure. The use of a similarity coefficient that measures the internal (or local) representation of the molecules has proven not to be affected by the sparsity of high dimensional data. Instead, it can be used to identify the molecules with similar scaffolds or having a similar local structure to the query molecule.

Further investigations were performed in Chapter 6 to study the effects of dimensionality on the effectiveness of another chemoinformatics application, i.e., molecular clustering. Similar to the experiment conducted in Chapter 5, the molecules were represented using thirteen dimensions of an ECFP₄-like binary fingerprints and clustered by two clustering methods. The pairwise distances were measured by ten distance coefficients and the effectiveness was measured based on the ability to separate the actives/ inactives and the identification of the single best cluster.

The experiments revealed that the effectiveness of the clustering application in high dimensionality varies depending on the nature of the distance coefficient. Distance coefficients which measure the proportion of distances between two molecules from the overall dimensions tend to decrease the performance of the application in very high fingerprint dimensions. A detailed investigation of the distribution of the distances of two distance coefficients resulted in the identification of two significant behaviours. The results showed that, for a certain type of distance coefficient, as the dimensionality increases, it is difficult to discriminate the distances between the nearest or the furthest molecules as their distances were almost similar. This strengthens the conclusion made for the investigation reported in Chapter 5 that the variation of the effectiveness depends on the nature of the similarity measures.

With regards to the second aim of this study, as mentioned in Section 1.3, the molecular fingerprint and similarity coefficient are among the key components

of a similarity search application. Many comparative studies have investigated the effect of varying the components of the searches (Hert et al., 2004; Riniker and Landrum, 2013; Todeschini et al., 2012). However, the studies focused on varying a single component while the other components were held constant in the investigations.

Hence, Chapter 7 was designed to determine the relative importance of the components influencing 2D fingerprint similarity searching. A novel statistical approach called cross-classified multilevel modeling was adapted to model the results of similarity searches from all possible combinations of 2D fingerprints and similarity coefficients used in this experimental chapter. In contrast to previous comparative studies, this research considered all variations of components in the investigation.

It was found that the activity class plays the greatest role in determining the effectiveness of the application followed by the reference structure, then the fingerprints and finally the similarity coefficients. Further analysis was carried out to assess the most important factor between the fingerprint and the similarity coefficient and showed that the fingerprint component is significantly more important than the similarity coefficient. This study also supports previous findings by Arif et al. (2013) that more reference structures should be used in comparative studies of similarity measures.

8.3 Implication of Results

The results from the high dimensional effect studies in Chapters 5 and 6 seem to contradict the curse of dimensionality phenomenon. In general, the increase of the dimensionality did not decrease the performance of the similarity searches. An implication of this is the possibility that the effectiveness was influenced by the coefficients used to measure the similarity or the distance of the molecules. Hence, these conclusions support the influence of the similarity coefficient in high dimensional similarity measure as suggested by France et al. (2012).

The findings also suggested that the number of bits in the fingerprint and the types of similarity measure can have a significant impact on the performance of

nearest neighbour search in virtual screening applications. This is especially the case for searches involving high dimensionality with sparse binary representations. Hence, any research involving such representations should consider pre-analysing the binary fingerprint for bit collision (as conducted in this study) and carefully choose the coefficient for the similarity measure before performing the nearest neighbour searching.

In addition, the conclusions made from Chapter 7 implied that the cross-classification multilevel modeling, which has proven to be very useful in social science research, was also effective in this study. Such an approach is able to quantify the importance of components for similarity searching applications. Hence, this method can be used by researchers in the chemoinformatics domain to identify the components that could improve other virtual screening applications.

8.4 Contribution to Knowledge

The findings from this study make several contributions to the current literature in chemoinformatics context and in other domains.

First, with appropriate dimensions of representations and suitable coefficients to measure the neighbourhood of the molecules, the effectiveness of the search can be improved. Researchers may consider higher dimensions than the commonly used 1024 bits fingerprint to represent the chemical dataset as also suggested by Sastry et al. (2010).

Second, the findings of this investigation support those of earlier studies on high dimensionality data analysis that the effect of nearest neighbour search in high dimensionality is influenced by the neighbourhood measures (France et al., 2012).

Third, this is the first study reporting the use of cross-classified multilevel modeling to analyse various factors concerning chemical datasets and virtual-based screening applications. It quantifies the importance of activity classes, 2D fingerprints, similarity coefficients and reference structures on the effectiveness

of similarity searches, which has not been studied previously in the chemoinformatics domain. It also identifies the variances of 2D fingerprint and similarity coefficient effects and suggests the relative importance of these two components. In addition, the findings of this investigation confirm the suggestion made by the previous study that more reference structures should be used in comparative research of similarity measures (Arif et al., 2013).

8.5 Strengths and Limitations

The key strengths of this study are that: (1) this is the first time an extended study of dimensionality effects was conducted in the chemoinformatics domain, and (2) it is also the first ever research in which the cross-classified multilevel modeling was implemented in a chemoinformatics domain.

On the other hand, this work is subject to at least two limitations. First, the processing of high dimensionality data requires high computational resources of processing time and memory. However, this might not be the case if the experiment is conducted using high performance computing. Second, the current implementation of the cross-classified multilevel modeling is limited by the use of one chemical dataset. The implementation involving other datasets might provide more evidence for the conclusion.

8.6 Suggestion for Future Research

It is recommended that further research be undertaken in the following areas:

1. The current study investigated the effects using various dimensions of 2D fingerprint. The results corroborate those of a previous study that highlighted the importance of using the proper number of 2D fingerprint dimensions (Sastry et al., 2010). The dimension of the 2D fingerprints can be considered as another influencing factor in determining the effectiveness of the similarity search application. Therefore, in future work, the dimension of the fingerprints can be added as another level modelled by the cross-classified method. This is to quantify the importance of the dimension and suggest the best dimension that might be used to optimise the similarity

search application. In addition, the work can be extended to investigate the effects of dimensionality on the effectiveness of chemoinformatics applications using 3D fingerprints.

2. There has been interest in optimising the similarity search application by evaluating the best combination of components (Riniker & Landrum, 2013; Sastry et al., 2010). Previously, this has been done by performing all the possible combinations of components and comparing the results using basic statistical methods. Alternatively, the cross-classified multilevel modeling has the ability to provide such an investigation in a different way. That is, by adding more levels to the models of any possible interactions between the similarity search components. For example, in order to identify the best combination of molecular representation and similarity coefficient, an analysis is conducted, a result is achieved by adding a new level to the model that represents the combination of different types of representation and coefficient. Upon the completion of the iterations, the model will produce the rank and variances for all possible combinations of representations and coefficients. Based on this rank, the best pair of performers can be identified and its relative importance can be measured by the level of the variances compared to the other combinations. It might be possible to use the best combinations identified for the purpose of optimising the similarity search application.
3. Many previous studies in molecular clustering have compared and evaluated different clustering methods with the focus on identifying an effective method for grouping chemical data (e.g., Willett, 1987; Chu et al., 2012). However, there are other aspects that need to be considered for optimising the molecular clustering. MacCuish and MacCuish wrote a review that emphasised the importance of the molecular representation and the similarity measure used in the clustering process (MacCuish & MacCuish, 2014). This highlights another perspective that is important in influencing the effectiveness of the clustering application. Therefore, a further study focusing on the identification of relative importance of components that influence molecular clustering is suggested. The cross-classified multilevel

modeling can be implemented to quantify the more important factors between the molecular representations, distance measures and clustering methods. In addition, the number of cluster partitions can be added as another component because it has been shown to have an influence on the current study in Chapter 6.

References

- Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International Conference on Database Theory* (Vol. 1, pp. 420–434). Yorktown Heights, N.Y.: Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/3-540-44503-X_27
- Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data* (pp. 94–105). New York, NY, USA: ACM. <https://doi.org/10.1145/276304.276314>
- Akella, L. B., & DeCaprio, D. (2010). Cheminformatics approaches to analyze diversity in compound screening libraries. *Current Opinion in Chemical Biology*, 14(3), 325–330. <https://doi.org/10.1016/j.cbpa.2010.03.017>
- Al Khalifa, A., Haranczyk, M., & Holliday, J. (2009). Comparison of nonbinary similarity coefficients for similarity searching, clustering and compound selection. *Journal of Chemical Information and Modeling*, 49(5), 1193–1201. <https://doi.org/10.1021/ci8004644>
- Almeida, M. O., Maltarollo, V. G., de Toledo, R. A., Shim, H., Santos, M. C., & Honorio, K. M. (2014). Medicinal electrochemistry: Integration of electrochemistry, medicinal chemistry and computational chemistry. *Current Medicinal Chemistry*, 21(20), 2266–2275.
- Andrew, N. (2015). Principal component analysis problem formulation [Video file]. Retrieved April 13, 2015, from <https://www.coursera.org/learn/machine-learning>
- Arif, S. M., Holliday, J. D., & Willett, P. (2009). Analysis and use of fragment-occurrence data in similarity-based virtual screening. *Journal of Computer-Aided Molecular Design*, 23(9), 655–668. <https://doi.org/10.1007/s10822-009-9285-0>

References

- Arif, S. M., Holliday, J. D., & Willett, P. (2010). Inverse frequency weighting of fragments for similarity-based virtual screening. *Journal of Chemical Information and Modeling*, *50*(8), 1340–1349. <https://doi.org/10.1021/ci1001235>
- Arif, S. M., Holliday, J. D., & Willett, P. (2013). Comparison of chemical similarity measures using different numbers of query structures. *Journal of Information Science*, *39*(1), 7–14. <https://doi.org/10.1177/0165551512470042>
- Audain, E., Sanchez, A., Vizcaíno, J. A., & Perez-Riverol, Y. (2014). A survey of molecular descriptors used in mass spectrometry based proteomics. *Current Topics in Medicinal Chemistry*, *14*(3), 388–397.
- Bajorath, J. (2001). Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *Journal of Chemical Information and Computer Sciences*, *41*(2), 233–245. <https://doi.org/10.1021/ci0001482>
- Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis—a brief tutorial. Retrieved November 22, 2017, from http://www.music.mcgill.ca/~ich/classes/mumt611_07/classifiers/lda_theory.pdf
- Baldi, P., Benz, R. W., Hirschberg, D. S., & Swamidass, S. J. (2007). Lossless compression of chemical fingerprints using integer entropy codes improves storage and retrieval. *Journal of Chemical Information and Modeling*, *47*(6), 2098–2109. <https://doi.org/10.1021/ci700200n>
- Barnard, J. M., & Downs, G. M. (1997). Chemical fragment generation and clustering software. *Journal of Chemical Information and Computer Sciences*, *37*(1), 141–142. <https://doi.org/10.1021/ci960090k>
- Bayada, D. M., Hamersma, H., & van Geerestein, V. J. (1999). Molecular diversity and representativity in chemical databases. *Journal of Chemical Information and Computer Sciences*, *39*(1), 1–10. <https://doi.org/10.1021/ci980109e>

References

- Bell, A., Smith, J., Sabel, C. E., & Jones, K. (2016). Formula for success: Multilevel modelling of Formula One Driver and Constructor performance, 1950–2014. *Journal of Quantitative Analysis in Sports*, 12(2), 99–112. <https://doi.org/10.1515/jqas-2015-0050>
- Bellman, R. E. (1961). *Adaptive control processes: A guided tour*. Princeton University Press.
- Bemis, G. W., & Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry*, 39(15), 2887–2893. <https://doi.org/10.1021/jm9602928>
- Bender, A., & Glen, R. C. (2004). Molecular similarity: A key technique in molecular informatics. *Organic & Biomolecular Chemistry*, 2(22), 3204–3218. <https://doi.org/10.1039/B409813G>
- Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is “nearest neighbor” meaningful? In *International Conference on Database Theory* (pp. 217–235). Springer Berlin Heidelberg.
- Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: Applications to image and text data. *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 245–250. <https://doi.org/10.1145/502512.502546>
- BIOVIA Databases | Sourcing Databases: BIOVIA Available Chemicals Directory (ACD). (n.d.). Retrieved June 15, 2014, from <http://accelrys.com/products/collaborative-science/databases/sourcing-databases/biovia-available-chemicals-directory.html>
- Brown, F. K. (1998). Chemoinformatics: What is it and how does it impact drug discovery. *Annual Reports in Medicinal Chemistry*, 33, 375–384.
- Brown, N. (2009). Chemoinformatics—an introduction for computer scientists. *ACM Computing Surveys*, 41(2), 8:1-8:38. <https://doi.org/10.1145/1459352.1459353>

References

- Brown, R. D., & Martin, Y. C. (1996). Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compound selection. *Journal of Chemical Information and Computer Sciences*, *36*(3), 572–584. <https://doi.org/10.1021/ci9501047>
- Brown, R. D., & Martin, Y. C. (1997). The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *Journal of Chemical Information and Computer Sciences*, *37*(1), 1–9. <https://doi.org/10.1021/ci960373c>
- Browne, W. J. (2015). MCMC Estimation in MLwiN, v2.32. Retrieved January 15, 2017, from <http://www.bris.ac.uk/cmm/media/software/mlwin/downloads/manuals/2-32/mcmc-web.pdf>
- Cao, D. S., Liang, Y. Z., Yan, J., Tan, G. S., Xu, Q. S., & Liu, S. (2013). PyDPI: Freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies. *Journal of Chemical Information and Modeling*, *53*(11), 3086–3096. <https://doi.org/10.1021/ci400127q>
- Carnero, A. (2006). High throughput screening in drug discovery. *Clinical and Translational Oncology*, *8*(7), 482–490. <https://doi.org/10.1007/s12094-006-0048-2>
- Casalegno, M., Sello, G., & Benfenati, E. (2006). Top-priority fragment QSAR approach in predicting pesticide aquatic toxicity. *Chemical Research in Toxicology*, *19*(11), 1533–1539. <https://doi.org/10.1021/tx0601814>
- Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015a). Molecular fingerprint similarity search in virtual screening. *Methods*, *71*, 58–63. <https://doi.org/10.1016/j.ymeth.2014.08.005>
- Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015b). Molecular fingerprint similarity search in virtual screening. *Methods*, *71*, 58–63. <https://doi.org/10.1016/j.ymeth.2014.08.005>

References

- Champely, S., & Chessel, D. (2002). Measuring biological diversity using euclidean metrics. *Environmental and Ecological Statistics*, 9(2), 167–177. <https://doi.org/10.1023/A:1015170104476>
- Cheeseright, T., Mackey, M., Rose, S., & Vinter, A. (2006). Molecular field extrema as descriptors of biological activity: Definition and validation. *Journal of Chemical Information and Modeling*, 46(2), 665–676. <https://doi.org/10.1021/ci050357s>
- Chemical Abstracts Service. (2015). CAS registry and CAS registry number. Retrieved April 22, 2015, from <https://www.cas.org/content/chemical-substances/faqs#q3>
- Chen, N. G., & Golovlev, V. (2013). Structural key bit occurrence frequencies and dependencies in PubChem and their effect on similarity searches. *Molecular Informatics*, 32(4), 355–361. <https://doi.org/10.1002/minf.201300006>
- Chen, W. L. (2006). Chemoinformatics: Past, present, and future. *Journal of Chemical Information and Modeling*, 46(6), 2230–2255. <https://doi.org/10.1021/ci060016u>
- Cheng, C., Maggiora, G., Lajiness, M., & Johnson, M. (1996). Four association coefficients for relating molecular similarity measures. *Journal of Chemical Information and Computer Sciences*, 36(4), 909–915. <https://doi.org/10.1021/ci9604605>
- Cheng, T., Li, Q., Zhou, Z., Wang, Y., & Bryant, S. H. (2012). Structure-based virtual screening for drug discovery: A problem-centric review. *The AAPS Journal*, 14(1), 133–141. <https://doi.org/10.1208/s12248-012-9322-0>
- Chu, C. W., Holliday, J. D., & Willett, P. (2009). Effect of data standardization on chemical clustering and similarity searching. *Journal of Chemical Information and Modeling*, 49(2), 155–161. <https://doi.org/10.1021/ci800224h>

References

- Chu, C. W., Holliday, J. D., & Willett, P. (2012). Combining multiple classifications of chemical structures using consensus clustering. *Bioorganic & Medicinal Chemistry*, 20(18), 5366–5371. <https://doi.org/10.1016/j.bmc.2012.03.010>
- Clarke, R., Resson, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., & Wang, Y. (2008). The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8(1), 37–49. <https://doi.org/10.1038/nrc2294>
- Crasto, A. M. (2016). Drug discovery, hit to lead. Retrieved February 21, 2018, from <https://www.slideshare.net/anthonycrasto64/drug-discovery-hit-to-lead>
- Cruz-Monteaudo, M., Medina-Franco, J. L., Pérez-Castillo, Y., Nicolotti, O., Cordeiro, M., & Borges, F. (2014). Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discovery Today*. <https://doi.org/10.1016/j.drudis.2014.02.003>
- Dearden, J. C. (2017). The Use of Topological Indices in QSAR and QSPR Modeling. In *Advances in QSAR Modeling* (pp. 57–88). Springer, Cham. https://doi.org/10.1007/978-3-319-56850-8_2
- Di Giuseppe, M. G., Troiano, A., Troise, C., & De Natale, G. (2014). K-means clustering as tool for multivariate geophysical data analysis. An application to shallow fault zone imaging. *Journal of Applied Geophysics*, 101, 108–115. <https://doi.org/10.1016/j.jappgeo.2013.12.004>
- DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, 47(Supplement C), 20–33. <https://doi.org/10.1016/j.jhealeco.2016.01.012>
- Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. In *AMS Math Challenges Lecture* (pp. 1–32).
- Downs, G. M., & Barnard, J. M. (2002). Clustering methods and their uses in computational chemistry. *Reviews in Computational Chemistry*, 18, 1–40.

References

- Downs, G. M., & Willett, P. (1994). Clustering of chemical structure databases for compound selection. *Advanced Computer-Assisted Techniques in Drug Discovery*, 111–130.
- Downs, G. M., Willett, P., & Fisanick, W. (1994). Similarity searching and clustering of chemical-structure databases using molecular property data. *Journal of Chemical Information and Computer Sciences*, 35(5), 1094–1102. Retrieved from <http://pubs.acs.org/doi/pdf/10.1021/ci00021a011>
- Duan, J., Dixon, S. L., Lowrie, J. F., & Sherman, W. (2010). Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *Journal of Molecular Graphics and Modelling*, 29(2), 157–170. <https://doi.org/10.1016/j.jmglm.2010.05.008>
- Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6), 1273–1280. <https://doi.org/10.1021/ci010132r>
- Ellis, D., Furner-Hines, J., & Willett, P. (1993). Measuring the degree of similarity between objects in text retrieval systems. *Perspectives in Information Management*, 3(2), 128–149. Retrieved from <http://works.bepress.com/furner/8>
- Estrada, E. (2002). Physicochemical interpretation of molecular connectivity indices. *Journal of Physical Chemistry A*, 106(39), 9085–9091. <https://doi.org/10.1021/jp026238m>
- Ewing, T., Baber, J. C., & Feher, M. (2006). Novel 2D fingerprints for ligand-based virtual screening. *Journal of Chemical Information and Modeling*, 46(6), 2423–2431. <https://doi.org/10.1021/ci060155b>
- Field, A. (2013). *Discovering statistics using SPSS*. Sage publications.
- Finn, P. W., & Morris, G. M. (2013). Shape-based similarity searching in chemical databases. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(3), 226–241. <https://doi.org/10.1002/wcms.1128>

References

- Fodor, I. K. (2002). A survey of dimension reduction techniques. Retrieved November 6, 2014, from <http://www.llnl.gov/tid/lof/documents/pdf/240921.pdf>
- France, S. L., Carroll, J. D., & Xiong, H. (2012). Distance metrics for high dimensional nearest neighborhood recovery: Compression and normalization. *Information Sciences*, *184*(1), 92–110. <https://doi.org/10.1016/j.ins.2011.07.048>
- Gan, S., Cosgrove, D. A., Gardiner, E. J., & Gillet, V. J. (2014). Investigation of the use of spectral clustering for the analysis of molecular data. *Journal of Chemical Information and Modeling*, *54*(12), 3302–3319. <https://doi.org/10.1021/ci500480b>
- Gardiner, E. J., Gillet, V. J., Haranczyk, M., Hert, J., Holliday, J. D., Malim, N., ... Willett, P. (2009). Turbo similarity searching: Effect of fingerprint and dataset on virtual-screening performance. *Statistical Analysis and Data Mining*, *2*(2), 103–114. <https://doi.org/10.1002/sam.10037>
- Garner, C. L., & Raudenbush, S. W. (1991). Neighborhood effects on educational attainment: A multilevel analysis. *Sociology of Education*, *64*(4), 251–262. <https://doi.org/10.2307/2112706>
- Gasteiger, J. (2006). The central role of chemoinformatics. *Chemometrics and Intelligent Laboratory Systems*, *82*(1), 200–209. <https://doi.org/10.1016/j.chemolab.2005.06.022>
- Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., ... Overington, J. P. (2012). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, *40*(D1), D1100–D1107. <https://doi.org/10.1093/nar/gkr777>
- Gawehn, E., Hiss, J. A., & Schneider, G. (2016). Deep learning in drug discovery. *Molecular Informatics*, *35*(1), 3–14. <https://doi.org/10.1002/minf.201501008>

References

- Geppert, H., Vogt, M., & Bajorath, J. (2010). Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *Journal of Chemical Information and Modeling*, *50*(2), 205–216. <https://doi.org/10.1021/ci900419k>
- Ghosh, S., Nie, A., An, J., & Huang, Z. (2006). Structure-based virtual screening of chemical libraries for drug discovery. *Current Opinion in Chemical Biology*, *10*(3), 194–202. <https://doi.org/10.1016/j.cbpa.2006.04.002>
- Gionis, A., Indyk, P., & Motwani, R. (1999). Similarity search in high dimensions via hashing. *Conference on Very Large Data Bases*, *99*, 518–529.
- Godden, J. W., & Bajorath, J. (2006). A distance function for retrieval of active molecules from complex chemical space representations. *Journal of Chemical Information and Modeling*, *46*(3), 1094–1097. <https://doi.org/10.1021/ci050510i>
- Goh, G. B., Hodas, N. O., & Vishnu, A. (2017). Deep learning for computational chemistry. *Journal of Computational Chemistry*, *38*(16), 1291–1307. <https://doi.org/10.1002/jcc.24764>
- Goldstein, H. (1987). Multilevel covariance component models. *Biometrika*, *74*(2), 430–431.
- Goldstein, H. (2011). *Multilevel statistical models*. John Wiley & Sons.
- Grant, J. A., Gallardo, M. A., & Pickup, B. T. (1996). A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *Journal of Computational Chemistry*, *17*(14), 1653–1666. [https://doi.org/10.1002/\(SICI\)1096-987X\(19961115\)17:14<1653::AID-JCC7>3.0.CO;2-K](https://doi.org/10.1002/(SICI)1096-987X(19961115)17:14<1653::AID-JCC7>3.0.CO;2-K)
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

References

- Hall, L. H., & Kier, L. B. (2001). Issues in representation of molecular structure: The development of molecular connectivity. *Journal of Molecular Graphics and Modelling*, 20(1), 4–18. [https://doi.org/10.1016/S1093-3263\(01\)00097-3](https://doi.org/10.1016/S1093-3263(01)00097-3)
- Haranczyk, M., & Holliday, J. (2008). Comparison of similarity coefficients for clustering and compound selection. *Journal of Chemical Information and Modeling*, 48(3), 498–508. <https://doi.org/10.1021/ci700413a>
- Heikamp, K., & Bajorath, J. (2013). The future of virtual compound screening. *Chemical Biology & Drug Design*, 81(1), 33–40. <https://doi.org/10.1111/cbdd.12054>
- Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., & Schuffenhauer, A. (2004). Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Organic & Biomolecular Chemistry*, 2(22), 3256–3266. <https://doi.org/10.1039/B409865J>
- Hinneburg, A., Aggarwal, C. C., & Keim, D. A. (2000). What is the nearest neighbor in high dimensional spaces? In *26th International Conference on Very Large Databases* (pp. 506–515).
- Holliday, J. D., Hu, C.-Y., & Willett, P. (2002). Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Combinatorial Chemistry & High Throughput Screening*, 5(2), 155–166. <https://doi.org/10.2174/1386207024607338>
- Holliday, J. D., Salim, N., Whittle, M., & Willett, P. (2003). Analysis and display of the size dependence of chemical similarity coefficients. *Journal of Chemical Information and Computer Sciences*, 43(3), 819–828. <https://doi.org/10.1021/ci034001x>
- Holliday, J. D., Sani, N., & Willett, P. (2015). Calculation of substructural analysis weights using a genetic algorithm. *Journal of Chemical Information and Modeling*, 55(2), 214–221. <https://doi.org/10.1021/ci500540s>

References

- Holliday, J. D., Willett, P., & Xiang, H. (2012). Interactions between weighting scheme and similarity coefficient in similarity-based virtual screening. *International Journal of Chemoinformatics and Chemical Engineering*, 2(2), 28–41. <https://doi.org/10.4018/ijcce.2012070103>
- Howe, T. J., Mahieu, G., Marichal, P., Tabruyn, T., & Vugts, P. (2007). Data reduction and representation in drug discovery. *Drug Discovery Today*, 12(1–2), 45–53. <https://doi.org/10.1016/j.drudis.2006.10.014>
- IBM Corp. IBM SPSS Statistics for Windows. (2013). (Version 20.0). Armonk, NY: IBM Corp.
- Indyk, P., & Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing* (pp. 604–613). New York, NY, USA: ACM. <https://doi.org/10.1145/276698.276876>
- James, C. A., Weininger, D., & Delany, J. (1995). *Daylight theory manual*. Daylight Chemical Information Systems.
- Johnson, M. A., & Maggiora, G. M. (1990). *Concepts and applications of molecular similarity*. New York, NY: Wiley.
- Jones, E., Oliphant, E., & Peterson, P. (2001). SciPy: Open Source Scientific Tools for Python. Retrieved May 16, 2017, from <http://www.scipy.org/>
- Kar, S., Roy, K., & Leszczynski, J. (2017). On applications of QSARs in food and agricultural sciences: History and critical review of recent developments. In *Advances in QSAR Modeling* (pp. 203–302). Springer, Cham. https://doi.org/10.1007/978-3-319-56850-8_7
- Keller, A., Gerkin, R. C., Guan, Y., Dhurandhar, A., Turu, G., Szalai, B., ... Meyer, P. (2017). Predicting human olfactory perception from chemical features of odor molecules. *Science*, 355(6327), 820–826. <https://doi.org/10.1126/science.aal2014>

References

- Kier, L. B., & Hall, L. H. (2001). Molecular connectivity: Intermolecular accessibility and encounter simulation. *Journal of Molecular Graphics and Modelling*, 20(1), 76–83. [https://doi.org/10.1016/S1093-3263\(01\)00102-4](https://doi.org/10.1016/S1093-3263(01)00102-4)
- Kirchmair, J., Distinto, S., Markt, P., Schuster, D., Spitzer, G. M., Liedl, K. R., & Wolber, G. (2009). How to optimize shape-based virtual screening: Choosing the right query and including chemical information. *Journal of Chemical Information and Modeling*, 49(3), 678–692. <https://doi.org/10.1021/ci8004226>
- Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery*, 3(11), 935–949. <https://doi.org/10.1038/nrd1549>
- Korn, F., Pagel, B. U., & Faloutsos, C. (2001). On the “dimensionality curse” and the “self-similarity blessing.” *IEEE Transactions on Knowledge and Data Engineering*, 13(1), 96–111. <https://doi.org/10.1109/69.908983>
- Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1–27. <https://doi.org/10.1007/BF02289565>
- Kubinyi, H. (1997). QSAR and 3D QSAR in drug design part 1: Methodology. *Drug Discovery Today*, 2(11), 457–467. [https://doi.org/10.1016/S1359-6446\(97\)01079-9](https://doi.org/10.1016/S1359-6446(97)01079-9)
- Kümmel, A., Selzer, P., Beibel, M., Gubler, H., Parker, C. N., & Gabriel, D. (2011). Comparison of multivariate data analysis strategies for high-content screening. *Journal of Biomolecular Screening*, 16(3), 338–347. <https://doi.org/10.1177/1087057110395390>
- Landrum, G. (2016, April). RDKit Documentation. Release 2016.03.1. Retrieved February 6, 2016, from http://www.rdkit.org/RDKit_Docs.current.pdf
- Langer, T., & Wolber, G. (2004). Pharmacophore definition and 3D searches. *Drug Discovery Today: Technologies*, 1(3), 203–207. <https://doi.org/10.1016/j.ddtec.2004.11.015>

References

- Leach, A. R., & Gillet, V. J. (2007). *An introduction to chemoinformatics*. Dordrecht: Springer.
- Leckie, G. (2009). The complexity of school and neighbourhood effects and movements of pupils on school differences in models of educational achievement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(3), 537–554. <https://doi.org/10.1111/j.1467-985X.2008.00577.x>
- Leckie, G., & Charlton, C. (2013). Runmlwin: A program to run the MLwiN multilevel modelling software from within Stata. *Journal of Statistical Software*, 52(11), 1–40.
- Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (2012). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 64, Supplement, 4–17. <https://doi.org/10.1016/j.addr.2012.09.019>
- Livingstone, D. J. (2000). The characterization of chemical structures using molecular properties. A survey. *Journal of Chemical Information and Computer Sciences*, 40(2), 195–209.
- Lyne, P. D. (2002). Structure-based virtual screening: An overview. *Drug Discovery Today*, 7(20), 1047–1055. [https://doi.org/10.1016/S1359-6446\(02\)02483-2](https://doi.org/10.1016/S1359-6446(02)02483-2)
- MACCS keys. (2002). MDL Information Systems Inc.
- MacCuish, J. D., & MacCuish, N. E. (2014). Chemoinformatics applications of cluster analysis. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(1), 34–48. <https://doi.org/10.1002/wcms.1152>
- Maggiore, G., Vogt, M., Stumpfe, D., & Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(8), 3186–3204. <https://doi.org/10.1021/jm401411z>

References

- Maldonado, A. G., Doucet, J. P., Petitjean, M., & Fan, B. T. (2006). Molecular similarity and diversity in chemoinformatics: From theory to applications. *Molecular Diversity*, *10*(1), 39–79. <https://doi.org/10.1007/s11030-006-8697-1>
- Martin, Y. C. (1992). 3D database searching in drug design. *Journal of Medicinal Chemistry*, *35*(12), 2145–2154. <https://doi.org/10.1021/jm00090a001>
- Martinez-Mayorga, K., & Medina-Franco, J. L. (2009). Chapter 2 chemoinformatics—applications in food chemistry. *Advances in Food and Nutrition Research*, *58*, 33–56. [https://doi.org/10.1016/S1043-4526\(09\)58002-3](https://doi.org/10.1016/S1043-4526(09)58002-3)
- Martinez-Mayorga, K., Peppard, T. L., Ramírez-Hernández, A. I., Terrazas-Álvarez, D. E., & Medina-Franco, J. L. (2014). Chemoinformatics analysis and structural similarity studies of food-related databases. In *Foodinformatics* (pp. 97–110). Springer International Publishing. https://doi.org/10.1007/978-3-319-10226-9_3
- Matter, H. (1997). Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *Journal of Medicinal Chemistry*, *40*(8), 1219–1229. <https://doi.org/10.1021/jm960352+>
- Matthias, S. (2014). PCA - Principal component analysis. Retrieved November 5, 2014, from http://www.nlpc.org/pca_principal_component_analysis.html
- MDL Drug Data Report*. (2005). MDL Information Systems/Symyx: Santa Clara, CA.
- Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(10), 1615–1630. <https://doi.org/10.1109/TPAMI.2005.188>
- Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, *5*(2), 181–204. <https://doi.org/10.1007/BF01897163>

References

- Moroy, G., Martiny, V. Y., Vayer, P., Villoutreix, B. O., & Miteva, M. A. (2012). Toward in silico structure-based ADMET prediction in drug discovery. *Drug Discovery Today*, *17*(1), 44–55. <https://doi.org/10.1016/j.drudis.2011.10.023>
- Muja, M., & Lowe, D. G. (2009). Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference Computer Vision Theory Application* (Vol. 1, pp. 331–340).
- Muja, M., & Lowe, D. G. (2014). Scalable nearest neighbor algorithms for high dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(11), 2227–2240. <https://doi.org/10.1109/TPAMI.2014.2321376>
- Nicolotti, O., & Carotti, A. (2006). QSAR and QSPR studies of a highly structured physicochemical domain. *Journal of Chemical Information and Modeling*, *46*(1), 264–276. <https://doi.org/10.1021/ci0502931>
- Nimrod, R. (2014). Principal components analysis. Retrieved April 14, 2015, from <http://www.tau.ac.il/~rubi/PCA.pdf>
- Palmer, A. D., Bunch, J., & Styles, I. B. (2013). Randomized approximation methods for the efficient compression and analysis of hyperspectral data. *Analytical Chemistry*, *85*(10), 5078–5086. <https://doi.org/10.1021/ac400184g>
- Randić, M. (1975). Characterization of molecular branching. *Journal of the American Chemical Society*, *97*(23), 6609–6615. <https://doi.org/10.1021/ja00856a001>
- Randić, M. (2014). On of molecular similarity based on a single molecular descriptor. *Chemical Physics Letters*, *599*, 1–6. <https://doi.org/10.1016/j.cplett.2014.03.022>
- Rao, V. S., & Srinivas, K. (2011). Modern drug discovery process: An in silico approach. *Journal of Bioinformatics and Sequence Analysis*, *3*(5), 89–94. Retrieved from <http://www.academicjournals.org/journal/JBSA/article-abstract/6368F575575>

References

- Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2012). A User's Guide to MLwiN, v2.26. Centre for Multilevel Modelling, University of Bristol. Retrieved December 12, 2016, from <http://www.bristol.ac.uk/cmm/software/mlwin/download/manuals.html>
- Riniker, S., & Landrum, G. A. (2013). Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, 5, 26. <https://doi.org/10.1186/1758-2946-5-26>
- Ripphausen, P., Nisius, B., & Bajorath, J. (2011). State-of-the-art in ligand-based virtual screening. *Drug Discovery Today*, 16(9–10), 372–376. <https://doi.org/10.1016/j.drudis.2011.02.011>
- Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5), 742–754. <https://doi.org/10.1021/ci100050t>
- Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500), 2323–2326. <https://doi.org/10.1126/science.290.5500.2323>
- Rupp, M., Schneider, P., & Schneider, G. (2009). Distance phenomena in high-dimensional chemical descriptor spaces: Consequences for similarity-based approaches. *Journal of Computational Chemistry*, 30(14), 2285–2296. <https://doi.org/10.1002/jcc.21218>
- Sastry, M., Lowrie, J. F., Dixon, S. L., & Sherman, W. (2010). Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *Journal of Chemical Information and Modeling*, 50(5), 771–784. <https://doi.org/10.1021/ci100062n>
- Scior, T., Bender, A., Tresadern, G., Medina-Franco, J. L., Martínez-Mayorga, K., Langer, T., ... Agrafiotis, D. K. (2012). Recognizing pitfalls in virtual screening: A critical review. *Journal of Chemical Information and Modeling*, 52(4), 867–881. <https://doi.org/10.1021/ci200528d>

References

- Sheridan, R. P., & Kearsley, S. K. (2002). Why do we need so many chemical similarity search methods? *Drug Discovery Today*, 7(17), 903–911. [https://doi.org/10.1016/S1359-6446\(02\)02411-X](https://doi.org/10.1016/S1359-6446(02)02411-X)
- Sheridan, R. P., Singh, S. B., Fluder, E. M., & Kearsley, S. K. (2001). Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *Journal of Chemical Information and Computer Sciences*, 41(5), 1395–1406. <https://doi.org/10.1021/ci0100144>
- Siegel, S., & Castellan Jr, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.
- Smith, L. I. (2002). *A tutorial on principal components analysis*. Cornell University, USA.
- Snarey, M., Terrett, N. K., Willett, P., & Wilton, D. J. (1997). Comparison of algorithms for dissimilarity-based compound selection. *Journal of Molecular Graphics and Modelling*, 15(6), 372–385. [https://doi.org/10.1016/S1093-3263\(98\)00008-4](https://doi.org/10.1016/S1093-3263(98)00008-4)
- Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An open-source Java library for chemo and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2), 493–500. <https://doi.org/10.1021/ci025584y>
- Stumpfe, D., & Bajorath, J. (2011). Similarity searching. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(2), 260–282. <https://doi.org/10.1002/wcms.23>
- Stumpfe, D., & Bajorath, J. (2012). Exploring activity cliffs in medicinal chemistry: Miniperspective. *Journal of Medicinal Chemistry*, 55(7), 2932–2942. <https://doi.org/10.1021/jm201706b>
- Su, C., Zhan, J., & Sakurai, K. (2009). Importance of data standardization in privacy-preserving k-means clustering. In *International Conference on Database Systems for Advanced Applications* (pp. 276–286). Springer Berlin Heidelberg.

References

- Swamidass, S. J., & Baldi, P. (2007). Mathematical correction for fingerprint similarity measures to improve chemical retrieval. *Journal of Chemical Information and Modeling*, 47(3), 952–964. <https://doi.org/10.1021/ci600526a>
- Terrett, N. K., Gardner, M., Gordon, D. W., Kobylecki, R. J., & Steele, J. (1995). Combinatorial synthesis — the design of compound libraries and their application to drug discovery. *Tetrahedron*, 51(30), 8135–8173. [https://doi.org/10.1016/0040-4020\(95\)00467-M](https://doi.org/10.1016/0040-4020(95)00467-M)
- Todeschini, R., Ballabio, D., & Consonni, V. (2006). Distances and other dissimilarity measures in chemometrics. In *Encyclopedia of Analytical Chemistry*. John Wiley & Sons, Ltd. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/9780470027318.a9438/abstract>
- Todeschini, R., & Consonni, V. (2000). *Handbook of molecular descriptors*. Weinheim: Wiley-VCH.
- Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., & Willett, P. (2012). Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets. *Journal of Chemical Information and Modeling*, 52(11), 2884–2901. <https://doi.org/10.1021/ci300261r>
- Truchon, J.-F., & Bayly, C. I. (2007). Evaluating virtual screening methods: Good and bad metrics for the “early recognition” problem. *Journal of Chemical Information and Modeling*, 47(2), 488–508. <https://doi.org/10.1021/ci600426e>
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327.
- van der Maaten, L. J., Postma, E. O., & van den Herik, H. J. (2009). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(1–41), 66–71.
- van Rijsbergen, C. J. (1979). *Information Retrieval*. London Butterworths.

References

- Varin, T., Saettel, N., Villain, J., Lesnard, A., Dauphin, F., Bureau, R., & Rault, S. (2008). 3D Pharmacophore, hierarchical methods, and 5-HT(4) receptor binding data. *Journal of Enzyme Inhibition and Medicinal Chemistry*, 23(5), 593–603. <https://doi.org/10.1080/14756360802204748>
- Vasilyev, A., & Stevanovic, D. (2014). MathChem: A python package for calculating topological indices. *MATCH Communications Mathematical Computer Chemistry*, 71(3), 657–680.
- Vogt, M., & Bajorath, J. (2012). Chemoinformatics: A view of the field and current trends in method development. *Bioorganic & Medicinal Chemistry*, 20(18), 5317–5323. <https://doi.org/10.1016/j.bmc.2012.03.030>
- Vogt, M., & Bajorath, J. (2013). Similarity searching for potent compounds using feature selection. *Journal of Chemical Information and Modeling*, 53(7), 1613–1619. <https://doi.org/10.1021/ci4003206>
- Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244. <https://doi.org/10.1080/01621459.1963.10500845>
- Warr, W. A. (2011). Representation of chemical structures. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(4), 557–579. <https://doi.org/10.1002/wcms.36>
- Weber, R., Schek, H. J., & Blott, S. (1998). A quantitative analysis and performance study for similarity-search methods in high-dimensional space. *24th International Conference on Very Large Databases*, 98, 194–205.
- Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1), 31–36. <https://doi.org/10.1021/ci00057a005>
- Wiener, H. (1947). Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69(1), 17–20. <https://doi.org/10.1021/ja01193a005>

References

- Willett, P. (1987). *Similarity and clustering in chemical information systems*. New York, NY, USA: John Wiley & Sons, Inc.
- Willett, P. (2005). Chemoinformatics techniques for data mining in files of two-dimensional and three-dimensional chemical molecules. In *The Third Conference on the Foundations of Information Science*. MDPI.
- Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today*, 11(23–24), 1046–1053. <https://doi.org/10.1016/j.drudis.2006.10.005>
- Willett, P. (2009). Similarity methods in chemoinformatics. *Annual Review of Information Science and Technology*, 43(1), 1–117. <https://doi.org/10.1002/aris.2009.1440430108>
- Willett, P. (2011a). Chemoinformatics: A history. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(1), 46–56. <https://doi.org/10.1002/wcms.1>
- Willett, P. (2011b). Similarity searching using 2D structural fingerprints. In *Chemoinformatics and Computational Chemical Biology* (pp. 133–158). Humana Press.
- Willett, P. (2011c). Similarity-based data mining in files of two-dimensional chemical structures using fingerprint measures of molecular resemblance. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3), 241–251. <https://doi.org/10.1002/widm.26>
- Willett, P. (2014). The calculation of molecular structural similarity: Principles and practice. *Molecular Informatics*, 33(6–7), 403–413. <https://doi.org/10.1002/minf.201400024>
- Willett, P., Barnard, J. M., & Downs, G. M. (1998). Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38(6), 983–996. <https://doi.org/10.1021/ci9800211>

References

- Willett, P., & Winterman, V. (1986). A comparison of some measures for the determination of inter-molecular structural similarity measures of inter-molecular structural similarity. *Quantitative Structure-Activity Relationships*, 5(1), 18–25. <https://doi.org/10.1002/qsar.19860050105>
- Williams, A. (2014). *Designing and implementing a usable interface for test databases used by chemoinformatics researchers at the University of Sheffield's information school* (Master's thesis). University of Sheffield, United Kingdom.
- Witten, I. H., & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with Java implementations*. San Francisco, Calif.: Morgan Kaufmann.
- Wójcikowski, M., Ballester, P. J., & Siedlecki, P. (2017). Performance of machine-learning scoring functions in structure-based virtual screening. *Scientific Reports*, 7, srep46710. <https://doi.org/10.1038/srep46710>
- Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1–3), 37–52. [https://doi.org/10.1016/0169-7439\(87\)80084-9](https://doi.org/10.1016/0169-7439(87)80084-9)
- World of Molecular Bioactivity. (2011). Retrieved January 7, 2016, from <http://www.sunsetmolecular.com/>
- Xu, J., & Hagler, A. (2002). Chemoinformatics and drug discovery. *Molecules*, 7(8), 566–600. <https://doi.org/10.3390/70800566>
- Xue, L., & Bajorath, J. (2000). Molecular descriptors for effective classification of biologically active compounds based on principal component analysis identified by a genetic algorithm. *Journal of Chemical Information and Computer Sciences*, 40(3), 801–809. <https://doi.org/10.1021/ci000322m>
- Xue, L., Godden, J., Gao, H., & Bajorath, J. (1999). Identification of a preferred set of molecular descriptors for compound classification based on principal component analysis. *Journal of Chemical Information and Computer Sciences*, 39(4), 699–704. <https://doi.org/10.1021/ci980231d>

References

- Xue, L., Godden, J. W., & Bajorath, J. (1999). Database searching for compounds with similar biological activity using short binary bit string representations of molecules. *Journal of Chemical Information and Computer Sciences*, 39(5), 881–886. <https://doi.org/10.1021/ci990308d>
- Yap, C. W. (2011). PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7), 1466–1474. <https://doi.org/10.1002/jcc.21707>
- Zhao, W., Hevener, K. E., White, S. W., Lee, R. E., & Boyett, J. M. (2009). A statistical framework to evaluate virtual screening. *BMC Bioinformatics*, 10(1), 225. <https://doi.org/10.1186/1471-2105-10-225>

Appendix A Additional Results of Chapter 5

Table A-1 Average enrichment values using top 1% of the ranked dataset in searches for the fourteen WOMBAT activity classes using various Morgan R2 fingerprint dimensions. For each fingerprint dimension, the highest average enrichment value is marked by green colour and the lowest value by red colour for ease of reference

No	Similarity Coefficients	Morgan R2 Dimensions WOMBAT - EF 1%													
		2 ⁵	2 ⁶	2 ⁷	2 ⁸	2 ⁹	2 ¹⁰	2 ¹¹	2 ¹²	2 ¹³	2 ¹⁴	2 ¹⁵	2 ¹⁶	2 ¹⁷	
1	B1	9.03	21.03	25.38	25.05	23.60	22.95	22.64	22.76	22.71	22.59	22.64	22.64	22.64	22.64
2	B3	7.78	18.99	26.79	30.72	32.05	32.66	33.27	33.72	33.77	33.83	33.84	33.84	33.85	33.84
3	B5	3.61	3.72	10.67	19.68	25.24	29.24	30.91	32.13	32.44	32.52	32.62	32.68	32.68	32.64
4	B6	5.93	12.07	17.81	21.07	22.55	23.70	24.40	25.10	25.29	25.30	25.42	25.45	25.46	25.46
5	B7	2.80	4.23	12.42	19.23	22.94	25.13	26.13	26.92	27.11	27.26	27.33	27.36	27.28	27.28
6	B8	8.43	19.60	26.43	30.73	32.26	33.25	34.21	34.70	34.76	34.83	34.80	34.81	34.83	34.83
7	B9	7.45	18.14	26.23	30.35	31.70	32.24	32.85	33.29	33.39	33.39	33.38	33.39	33.36	33.36
8	B10	8.54	20.59	27.45	30.19	31.07	31.39	31.88	32.14	32.22	32.18	32.17	32.16	32.12	32.12
9	B11	7.15	16.89	24.84	29.52	31.06	31.56	32.10	32.55	32.67	32.72	32.71	32.71	32.68	32.68
10	B15	6.62	17.36	27.19	30.03	30.13	29.76	29.80	29.96	30.09	30.04	30.04	30.03	30.03	30.03
11	B16	8.53	19.80	25.84	27.11	27.58	27.95	28.25	28.74	28.98	28.95	29.01	29.04	29.05	29.05
12	B17	6.22	17.43	26.59	28.20	28.38	29.76	31.13	31.98	32.35	32.55	32.60	32.63	32.61	32.61
13	B18	9.41	20.67	27.19	30.45	31.82	32.48	33.14	33.67	33.74	33.82	33.83	33.84	33.83	33.83
14	B19	9.42	20.75	26.96	29.81	31.32	32.18	33.00	33.58	33.70	33.79	33.82	33.84	33.83	33.83
15	B20	2.80	17.17	24.17	26.79	27.97	28.60	29.06	29.64	29.82	29.87	29.87	29.91	29.89	29.89
16	B22	7.38	17.95	26.08	30.33	31.74	32.33	32.94	33.38	33.46	33.47	33.48	33.48	33.45	33.45
17	B23	7.91	16.50	24.41	28.82	30.86	31.78	32.62	33.17	33.32	33.36	33.37	33.39	33.36	33.36
18	B25	4.30	15.10	23.39	27.52	29.33	30.51	31.52	32.05	32.38	32.58	32.64	32.67	32.65	32.65
19	B26	8.88	19.21	27.11	30.39	31.15	31.45	31.91	32.28	32.38	32.58	32.64	32.67	32.65	32.65
20	B28	8.68	19.62	26.31	29.40	30.77	31.33	31.97	32.52	32.65	32.70	32.71	32.71	32.67	32.67
21	B29	9.05	20.12	27.07	30.41	31.63	32.16	32.80	33.24	33.39	33.39	33.38	33.39	33.37	33.37
22	B30	8.91	19.76	26.42	29.75	31.25	31.94	32.69	33.20	33.34	33.37	33.37	33.39	33.37	33.37
23	B33	7.45	18.14	26.22	30.35	31.70	32.24	32.85	33.29	33.39	33.39	33.38	33.39	33.36	33.36
24	B34	9.32	20.43	27.00	30.23	31.71	32.44	33.11	33.67	33.75	33.82	33.82	33.84	33.83	33.83
25	B36	5.42	19.30	23.90	24.66	24.47	24.73	24.90	25.33	25.41	25.35	25.45	25.48	25.49	25.49
26	B37	9.04	19.84	26.49	29.96	31.58	32.37	33.07	33.65	33.74	33.82	33.82	33.84	33.83	33.83
27	B38	9.55	20.03	27.60	30.80	32.09	32.77	33.35	33.76	33.85	33.86	33.85	33.85	33.84	33.84
28	B42	7.63	18.12	25.77	30.24	32.10	32.85	33.61	34.15	34.19	34.24	34.26	34.29	34.27	34.27
29	B43	4.83	17.21	24.18	26.81	27.98	28.61	29.06	29.64	29.82	29.88	29.90	29.92	29.91	29.91
30	B46	7.18	16.89	24.84	29.50	30.99	31.33	31.84	32.20	32.26	32.20	32.20	32.22	32.17	32.17
31	B51	7.78	18.99	26.79	30.69	31.97	32.45	32.96	33.35	33.38	33.36	33.33	33.34	33.34	33.34
Average		7.32	17.60	24.82	28.35	29.71	30.46	31.10	31.60	31.73	31.77	31.80	31.81	31.80	31.80

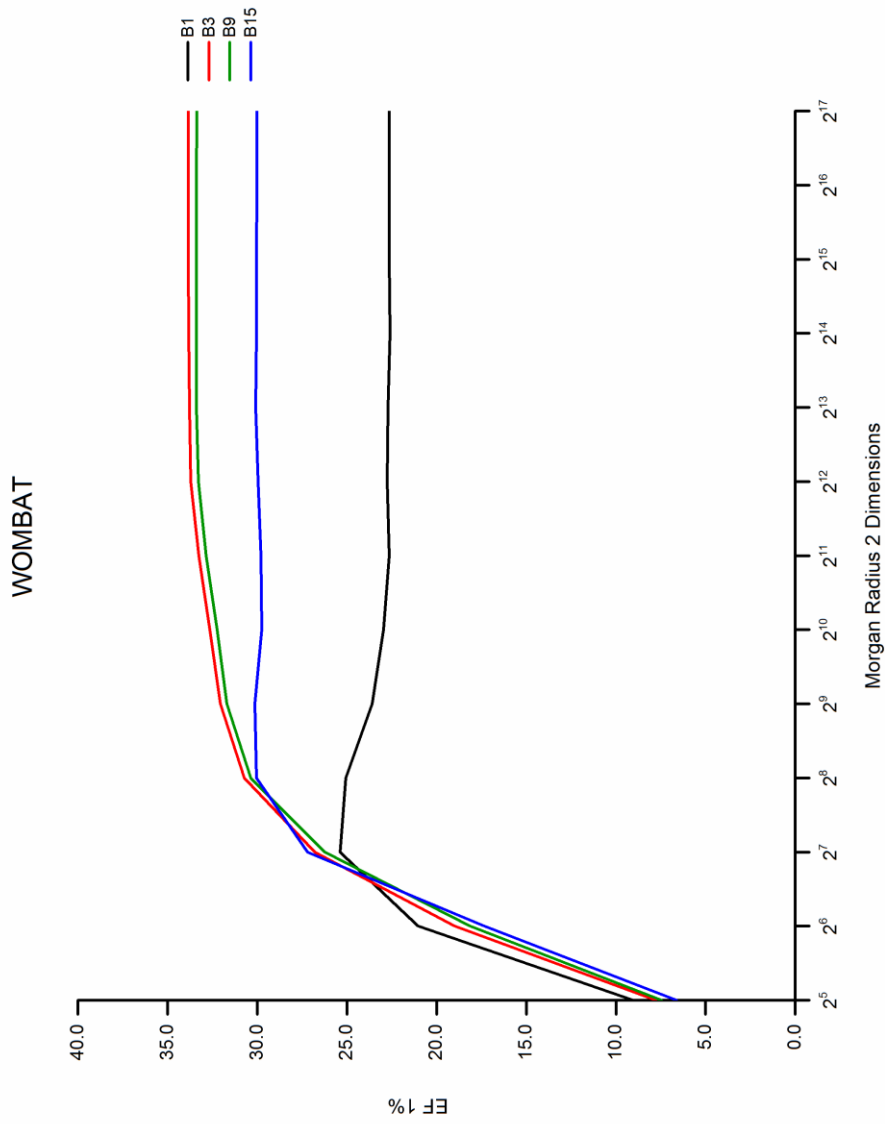


Figure A-1 A subset of average enrichment values using top 1% of the ranked dataset in searches for the fourteen WOMBAT activity classes using various Morgan Radius 2 fingerprint dimensions (Refer to Table A-1 for detail values)

Table A-2 Average enrichment values using top 1% of the ranked dataset in searches for the fifteen ChEMBL activity classes using various Morgan R2 fingerprint dimensions. For each fingerprint dimension, the highest average enrichment value is marked by green colour and the lowest value by red colour for ease of reference

No	Similarity Coefficients	Morgan R2 Dimensions ChEMBL - EF 1%														
		2 ⁵	2 ⁶	2 ⁷	2 ⁸	2 ⁹	2 ¹⁰	2 ¹¹	2 ¹²	2 ¹³	2 ¹⁴	2 ¹⁵	2 ¹⁶	2 ¹⁷		
1	B1	3.45	8.45	10.64	10.53	10.09	9.65	9.38	9.48	9.46	9.38	9.37	9.36	9.34	9.34	
2	B3	3.87	9.00	13.23	15.46	16.79	17.24	17.77	18.40	18.20	18.12	18.21	18.19	18.21	18.21	
3	B5	0.73	3.34	6.56	10.35	13.25	15.00	16.14	16.83	16.86	16.87	16.92	16.91	16.96	16.96	
4	B6	2.03	4.60	7.39	9.36	10.48	11.02	11.35	11.76	11.86	11.89	11.92	11.91	11.92	11.92	
5	B7	0.28	3.14	6.58	9.90	11.57	12.37	12.69	13.05	13.04	13.10	13.13	13.08	13.07	13.07	
6	B8	4.05	8.08	12.05	14.64	16.33	17.37	17.69	18.17	18.25	18.17	18.26	18.30	18.30	18.30	
7	B9	3.84	8.74	13.09	15.34	16.56	16.96	17.40	17.94	17.84	17.74	17.80	17.80	17.82	17.82	
8	B10	4.34	9.26	13.20	14.96	16.12	16.49	16.94	16.94	16.78	16.64	16.64	16.60	16.60	16.58	
9	B11	3.75	8.36	12.65	14.99	16.08	16.35	16.67	17.19	17.09	17.09	17.15	17.14	17.16	17.16	
10	B15	3.15	8.14	12.84	14.36	14.43	14.11	14.12	14.36	14.26	14.12	14.17	14.14	14.13	14.13	
11	B16	3.83	9.08	12.16	12.96	13.48	13.54	13.72	14.07	14.06	14.10	14.14	14.13	14.13	14.13	
12	B17	2.45	7.58	13.23	14.65	15.60	16.55	17.40	18.16	18.07	18.15	18.13	18.15	18.18	18.18	
13	B18	4.36	8.70	12.95	15.15	16.51	17.07	17.68	18.33	18.17	18.13	18.21	18.19	18.22	18.22	
14	B19	4.38	8.82	12.61	14.68	16.00	16.73	17.51	18.20	18.14	18.10	18.19	18.19	18.22	18.22	
15	B20	0.28	7.59	11.27	12.91	13.82	13.99	14.30	14.75	14.71	14.78	14.79	14.77	14.77	14.77	
16	B22	3.83	8.68	13.09	15.38	16.62	17.06	17.48	18.06	17.93	17.86	17.91	17.89	17.90	17.90	
17	B23	3.18	6.80	11.27	14.22	15.82	16.50	17.14	17.79	17.76	17.72	17.78	17.78	17.82	17.82	
18	B25	1.56	7.76	12.15	14.46	15.91	16.83	17.53	18.22	18.10	18.18	18.16	18.18	18.21	18.21	
19	B26	3.51	8.19	13.38	15.53	16.71	17.32	17.70	18.28	18.10	18.18	18.16	18.18	18.21	18.21	
20	B28	4.17	8.38	12.58	14.56	15.72	16.09	16.56	17.09	17.03	17.08	17.15	17.14	17.16	17.16	
21	B29	4.39	8.57	13.08	15.23	16.43	16.86	17.36	17.94	17.83	17.74	17.79	17.79	17.82	17.82	
22	B30	4.35	8.39	12.70	14.76	16.11	16.60	17.22	17.83	17.78	17.73	17.78	17.79	17.82	17.82	
23	B33	3.84	8.74	13.09	15.34	16.56	16.96	17.40	17.94	17.84	17.74	17.80	17.80	17.82	17.82	
24	B34	4.42	8.64	12.84	15.07	16.42	17.01	17.65	18.30	18.16	18.13	18.21	18.19	18.22	18.22	
25	B36	3.03	8.60	10.69	11.36	11.64	11.65	11.63	11.93	12.00	11.97	11.99	11.98	11.98	11.98	
26	B37	4.39	8.37	12.74	14.95	16.34	16.95	17.64	18.30	18.15	18.12	18.21	18.18	18.22	18.22	
27	B38	4.10	8.48	13.44	15.58	17.00	17.55	17.92	18.45	18.23	18.18	18.23	18.22	18.23	18.23	
28	B42	3.86	8.76	13.02	15.43	16.98	17.77	18.21	18.83	18.67	18.65	18.79	18.79	18.80	18.80	
29	B43	2.72	7.62	11.28	12.92	13.83	14.00	14.30	14.76	14.72	14.78	14.79	14.77	14.78	14.78	
30	B46	3.75	8.36	12.65	14.95	15.96	15.97	16.17	16.51	16.41	16.41	16.48	16.46	16.47	16.47	
31	B51	3.87	9.00	13.23	15.40	16.66	16.80	17.21	17.73	17.56	17.44	17.53	17.51	17.50	17.50	
Average		3.35	7.94	11.99	14.04	15.21	15.68	16.11	16.63	16.55	16.53	16.57	16.56	16.58	16.58	

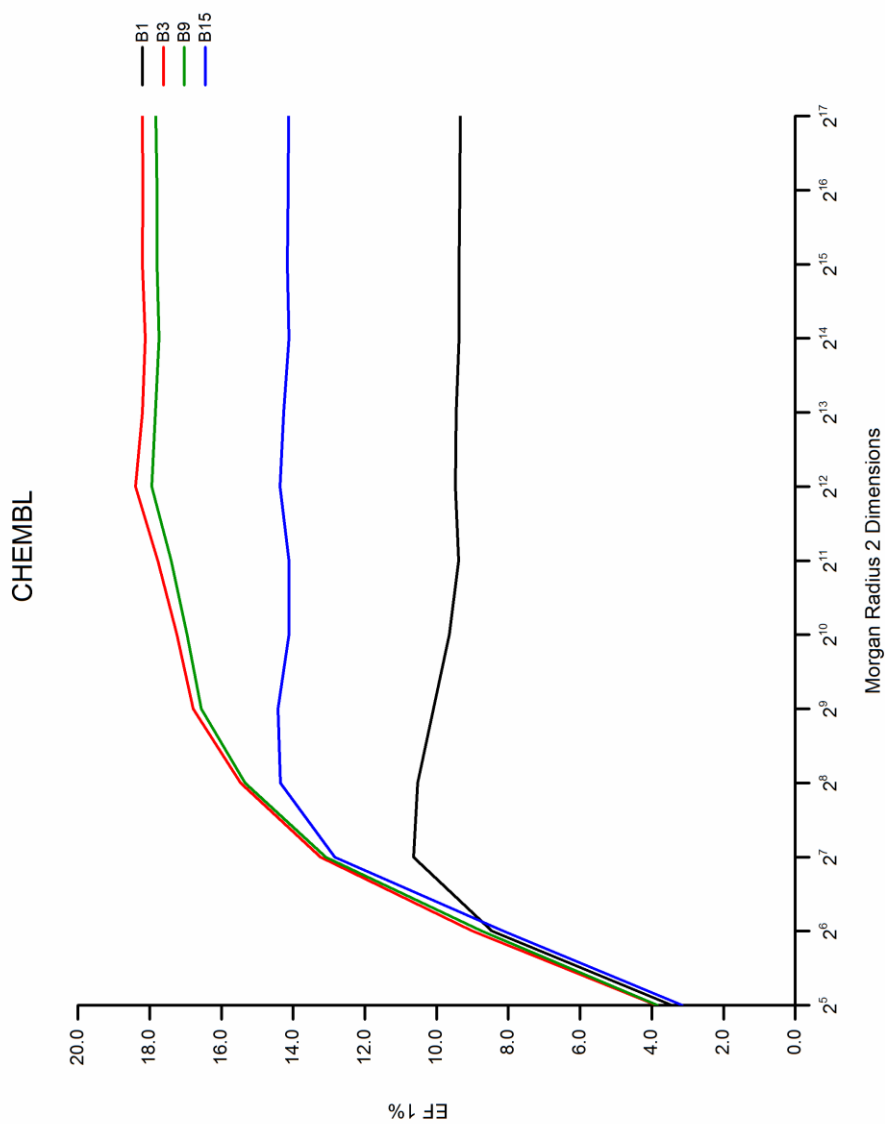


Figure A-2 A subset of average enrichment values using top 1% of the ranked dataset in searches for the fifteen ChEMBL activity classes using various Morgan Radius 2 fingerprint dimensions (Refer to Table A-2 for detail values)

Table A-3 Average bits set and bit collision rate based on the average of 10 molecules for WOMBAT dataset using various Morgan R2 fingerprint dimensions

Activity Class [Average 10 Reference]	Morgan R2 Fingerprint Dimension (WOMBAT)															
	2^5	2^6	2^7	2^8	2^9	2^{10}	2^{11}	2^{12}	2^{13}	2^{14}	2^{15}	2^{16}	2^{17}	2^{18}	2^{19}	2^{20}
5HT1A	26.00	34.70	41.80	47.70	49.70	50.50	51.60	51.60	51.80	51.80	51.90	51.90	51.90	51.90	51.90	51.90
5HT3	24.80	33.50	38.10	42.00	43.10	43.50	43.80	43.80	43.90	43.90	43.90	43.90	43.90	43.90	43.90	43.90
AChE	25.80	35.00	42.00	45.40	47.50	48.70	49.50	49.80	49.90	49.90	49.90	49.90	49.90	49.90	49.90	49.90
ANG	27.40	40.80	52.00	58.90	62.90	65.50	66.90	67.40	67.40	67.50	67.50	67.50	67.50	67.50	67.50	67.50
COX	22.90	29.80	34.60	37.10	39.30	39.70	40.20	40.40	40.40	40.50	40.60	40.60	40.60	40.60	40.60	40.60
D2	26.70	35.80	41.90	48.00	50.70	52.00	52.80	53.00	53.10	53.10	53.20	53.20	53.20	53.20	53.20	53.20
FXA	27.20	38.80	47.10	53.00	57.20	58.60	60.20	61.20	61.30	61.30	61.40	61.40	61.40	61.40	61.40	61.40
HIVP	27.70	41.60	50.70	58.60	61.70	63.00	64.00	64.90	65.00	65.20	65.20	65.20	65.30	65.30	65.30	65.30
MMP1	24.80	35.30	44.00	47.90	50.60	52.00	52.80	54.20	54.30	54.40	54.40	54.40	54.50	54.50	54.50	54.50
PDE4	25.20	35.20	41.30	46.60	48.20	49.40	49.60	50.10	50.10	50.20	50.20	50.30	50.30	50.30	50.30	50.30
PKC	22.10	30.30	34.60	35.80	37.60	39.00	40.20	40.20	40.40	40.40	40.40	40.40	40.40	40.40	40.40	40.40
Remin	30.30	46.10	59.30	68.80	73.40	78.60	80.30	82.00	82.40	82.50	82.50	82.50	82.50	82.50	82.50	82.50
SubP	27.30	40.60	49.60	56.00	57.60	59.10	59.70	59.80	59.80	59.80	60.00	60.00	60.00	60.00	60.00	60.00
Thrombin	28.50	42.50	53.80	58.90	61.60	63.70	65.00	65.70	65.80	65.80	65.80	65.80	65.80	65.80	65.80	65.80
Average	26.19	37.14	45.06	50.34	52.94	54.52	55.47	56.01	56.11	56.16	56.21	56.22	56.23	56.23	56.23	56.23
Bit Collision Rate		10.95	7.91	5.28	2.60	1.59	0.95	0.54	0.11	0.05	0.04	0.01	0.01	0.01	0.01	0.01

Table A-4 Average bits set and bit collision rate based on the average of 10 molecules for ChEMBL dataset using various Morgan R2 fingerprint dimensions

Activity Class [Average 10 Reference]	Morgan R2 Fingerprint Dimension (ChEMBL)															
	2^5	2^6	2^7	2^8	2^9	2^{10}	2^{11}	2^{12}	2^{13}	2^{14}	2^{15}	2^{16}	2^{17}	Bits Set	Collision Rate	
5HT	24.90	34.20	40.10	44.00	45.40	46.10	46.40	47.00	47.00	47.00	47.00	47.00	47.00	47.00	47.00	47.00
5HT1A	25.10	34.50	42.00	47.00	49.80	50.40	50.90	51.00	51.40	51.40	51.40	51.40	51.40	51.40	51.40	51.40
5HT3	23.50	30.90	35.40	38.30	41.20	41.90	42.30	42.60	42.60	42.60	42.60	42.60	42.60	42.60	42.60	42.60
ACHE	25.30	34.60	40.80	44.60	47.00	47.80	48.30	48.40	48.50	48.50	48.50	48.50	48.50	48.50	48.50	48.50
AT1	28.30	42.80	53.00	61.20	66.60	69.20	70.40	71.60	71.60	71.60	71.60	71.60	71.60	71.60	71.60	71.60
COX	21.40	28.50	32.80	34.90	36.20	36.90	37.20	37.80	38.00	38.00	38.00	38.00	38.00	38.00	38.00	38.00
D2	23.40	30.00	35.50	39.10	40.60	41.20	41.80	42.10	42.20	42.30	42.30	42.30	42.30	42.30	42.30	42.30
FXA	27.60	40.90	49.90	56.50	59.20	60.40	61.90	62.00	62.20	62.20	62.30	62.30	62.30	62.30	62.30	62.30
HIVP	22.70	32.60	40.80	45.90	48.20	49.30	49.30	49.90	50.10	50.10	50.10	50.10	50.10	50.10	50.10	50.10
MMP1	24.80	35.40	42.00	46.30	48.30	49.80	50.00	51.10	51.10	51.20	51.20	51.20	51.20	51.20	51.20	51.20
PDE4	24.70	33.00	41.20	47.50	49.00	50.30	50.40	50.40	50.40	50.80	50.80	50.80	50.80	50.80	50.80	50.80
PKC	24.30	34.70	41.50	46.30	48.40	49.60	51.00	51.80	52.00	52.00	52.00	52.00	52.00	52.00	52.00	52.00
Renin	27.90	41.80	52.60	60.20	64.90	67.50	68.50	69.60	70.10	70.20	70.20	70.20	70.20	70.20	70.20	70.20
SubP	27.00	38.00	46.40	52.50	55.00	55.50	56.10	56.30	56.30	56.30	56.30	56.30	56.30	56.30	56.30	56.30
Thrombin	29.40	44.80	57.20	62.70	65.10	66.90	68.10	69.10	69.20	69.50	69.50	69.50	69.50	69.50	69.50	69.50
Average	25.35	35.78	43.41	48.47	50.99	52.19	52.84	53.38	53.51	53.59	53.60	53.61	53.63	53.63	53.63	53.63
Bit Collision Rate		10.43	7.63	5.05	2.53	1.19	0.65	0.54	0.13	0.07	0.01	0.01	0.02	0.01	0.01	0.02

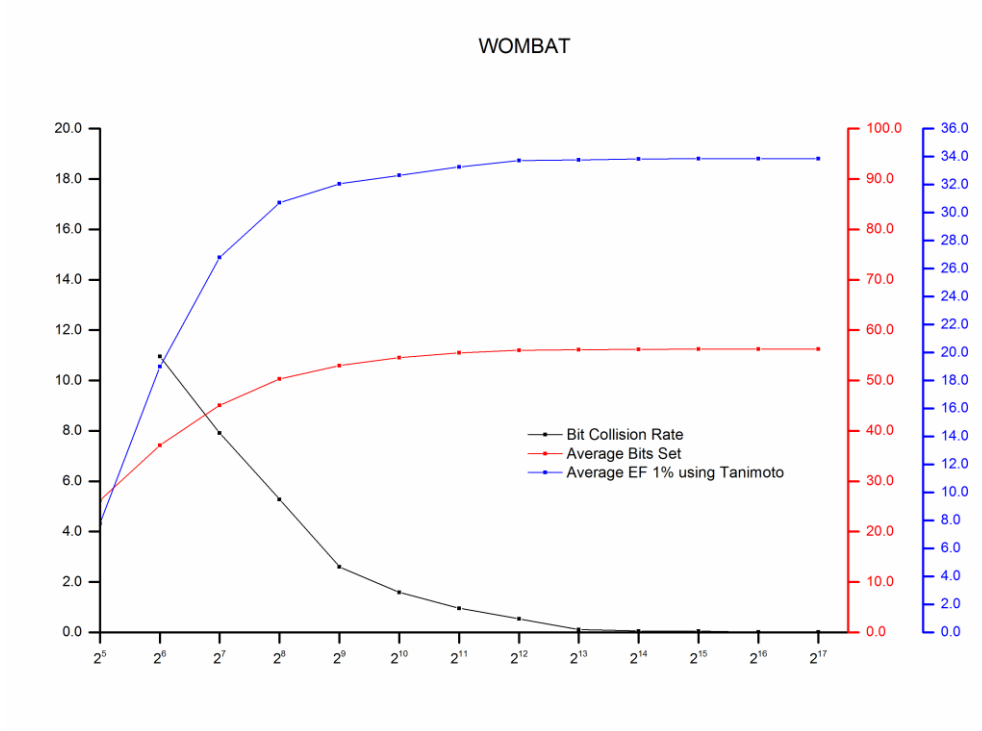


Figure A-3 Line plot measuring the average bits set, average enrichment curves and bit collision rate based on the average of 10 random molecules for WOMBAT dataset using various Morgan R2 fingerprint dimensions (Refer to Table A-3 for detail values)

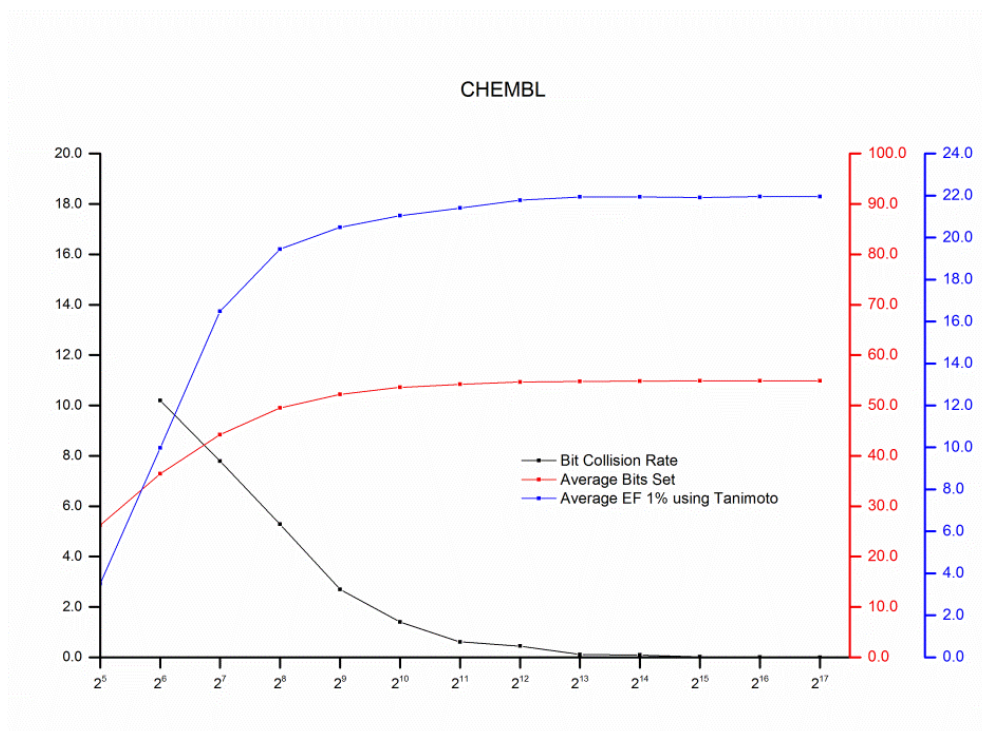


Figure A-4 Line plot measuring the average bits set, average enrichment curves and bit collision rate based on the average of 10 random molecules for ChEMBL dataset using various Morgan R2 fingerprint dimensions (Refer to Table A-4 for detail values)

Appendix B Additional Results of Chapter 6

Table B-1 Effectiveness value of Group Average clustering measured by (a) F -measure and (b) QPI -measure for the MDDR dataset using various distance coefficients and fingerprint dimensions. The range of the standard deviation, σ , for the mean F is between 0.000 and 0.625

Distance Coefficients	Fingerprint Dimensions	Partition											
		(a) F -Measure						(b) QPI -Measure					
		500	600	700	800	900	1000	500	600	700	800	900	1000
[D1] Bray-Curtis	2 ⁵	0.539	0.605	0.665	0.713	0.750	0.869	0.097	0.101	0.105	0.109	0.113	0.116
	2 ⁶	0.802	0.887	0.974	1.060	1.060	1.043	0.109	0.112	0.122	0.131	0.140	0.146
	2 ⁷	0.785	0.877	0.943	0.972	0.981	0.896	0.142	0.161	0.175	0.189	0.206	0.229
	2 ⁸	0.741	0.830	0.876	0.938	0.977	0.971	0.203	0.225	0.248	0.282	0.305	0.314
	2 ⁹	0.818	0.891	0.931	0.967	0.977	1.009	0.231	0.244	0.260	0.270	0.283	0.300
	2 ¹⁰	0.898	0.934	0.934	0.924	0.924	1.049	0.235	0.248	0.256	0.260	0.274	0.290
	2 ¹¹	0.918	0.937	0.946	0.945	1.003	1.039	0.229	0.256	0.261	0.260	0.273	0.289
	2 ¹²	0.891	0.987	0.987	0.888	0.911	0.951	0.229	0.243	0.246	0.257	0.277	0.286
	2 ¹³	0.914	0.954	0.965	0.996	1.007	0.987	0.235	0.240	0.253	0.273	0.279	0.283
	2 ¹⁴	0.842	0.954	0.952	1.053	1.046	1.046	0.227	0.252	0.254	0.266	0.270	0.281
	2 ¹⁵	0.830	0.921	0.946	1.003	1.018	1.103	0.232	0.243	0.259	0.270	0.280	0.286
	2 ¹⁶	0.850	0.918	0.944	0.970	1.015	1.103	0.227	0.254	0.258	0.268	0.280	0.293
	2 ¹⁷	0.915	0.923	0.947	0.973	1.006	1.008	0.226	0.246	0.259	0.268	0.277	0.292
[D2] City-Block	2 ⁵	0.495	0.507	0.577	0.612	0.643	0.758	0.112	0.114	0.118	0.123	0.126	0.129
	2 ⁶	0.851	0.946	0.956	0.983	1.007	1.008	0.172	0.179	0.191	0.198	0.211	0.217
	2 ⁷	1.072	1.126	1.079	1.038	1.044	1.045	0.276	0.273	0.294	0.296	0.305	0.316
	2 ⁸	0.851	0.862	0.908	0.976	0.976	1.021	0.314	0.276	0.286	0.293	0.313	0.330
	2 ⁹	0.959	0.947	1.020	1.091	1.064	1.064	0.324	0.332	0.327	0.343	0.347	0.348
	2 ¹⁰	0.840	0.887	0.943	0.921	0.928	0.976	0.309	0.327	0.339	0.347	0.363	0.367
	2 ¹¹	0.880	0.889	0.931	0.929	1.053	1.086	0.309	0.324	0.325	0.345	0.359	0.361
	2 ¹²	0.950	1.018	0.961	1.001	1.039	1.087	0.299	0.315	0.324	0.313	0.353	0.372
	2 ¹³	0.982	1.012	0.971	0.990	1.042	1.043	0.300	0.325	0.316	0.337	0.349	0.357
	2 ¹⁴	0.963	0.963	0.933	1.001	1.030	1.096	0.308	0.335	0.334	0.338	0.349	0.379
	2 ¹⁵	0.962	0.963	1.002	1.019	1.055	1.121	0.310	0.337	0.348	0.348	0.354	0.339
	2 ¹⁶	0.963	0.982	0.972	0.988	1.022	1.070	0.306	0.327	0.339	0.339	0.344	0.330
	2 ¹⁷	0.954	0.956	0.968	0.994	1.032	1.070	0.309	0.334	0.342	0.347	0.353	0.371
[D3] Cosine	2 ⁵	0.737	0.791	0.843	0.843	0.877	0.878	0.098	0.101	0.104	0.109	0.113	0.118
	2 ⁶	0.721	0.812	0.857	0.860	0.864	0.871	0.109	0.116	0.123	0.133	0.141	0.145
	2 ⁷	0.727	0.825	0.865	0.884	0.854	0.996	0.150	0.164	0.173	0.183	0.189	0.244
	2 ⁸	0.774	0.823	0.863	0.972	0.968	0.978	0.208	0.222	0.238	0.270	0.274	0.293
	2 ⁹	0.855	0.870	0.875	0.887	0.900	0.909	0.225	0.244	0.256	0.265	0.278	0.288
	2 ¹⁰	0.845	0.906	0.974	0.979	1.007	1.109	0.221	0.244	0.259	0.271	0.273	0.281
	2 ¹¹	0.839	0.929	1.012	0.987	0.988	1.056	0.235	0.241	0.260	0.270	0.279	0.287
	2 ¹²	0.863	0.850	0.895	0.855	0.865	0.875	0.226	0.233	0.251	0.268	0.284	0.284
	2 ¹³	0.861	0.961	0.990	1.018	1.032	1.032	0.229	0.237	0.255	0.266	0.276	0.294
	2 ¹⁴	0.905	0.955	0.977	1.049	1.061	1.063	0.230	0.243	0.262	0.283	0.285	0.290
	2 ¹⁵	0.820	0.994	1.016	1.036	1.023	1.031	0.245	0.264	0.272	0.275	0.275	0.309
	2 ¹⁶	0.908	0.989	0.998	1.050	1.068	1.030	0.246	0.240	0.265	0.268	0.275	0.284
	2 ¹⁷	0.881	0.968	0.998	1.050	1.068	1.030	0.231	0.237	0.264	0.268	0.275	0.283
[D4] Euclidean	2 ⁵	0.753	0.818	0.857	0.874	0.874	0.935	0.109	0.115	0.121	0.124	0.128	0.132
	2 ⁶	0.808	0.835	0.848	0.920	0.943	0.955	0.173	0.185	0.194	0.201	0.217	0.225
	2 ⁷	0.908	0.955	0.987	0.996	1.037	1.043	0.254	0.264	0.278	0.287	0.283	0.310
	2 ⁸	0.797	0.888	0.903	0.934	1.035	1.098	0.297	0.304	0.295	0.278	0.292	0.307
	2 ⁹	0.910	0.965	1.011	1.053	1.038	1.064	0.319	0.324	0.329	0.336	0.342	0.335
	2 ¹⁰	0.922	0.976	0.983	0.950	0.968	1.010	0.313	0.323	0.329	0.341	0.369	0.381
	2 ¹¹	0.877	0.910	1.009	1.013	1.025	1.124	0.313	0.328	0.341	0.342	0.339	0.369
	2 ¹²	0.924	0.887	0.908	0.928	1.018	1.052	0.313	0.313	0.327	0.336	0.341	0.375
	2 ¹³	0.882	0.919	0.901	0.913	1.034	1.034	0.306	0.330	0.336	0.347	0.360	0.356
	2 ¹⁴	0.990	1.023	0.977	1.058	1.081	1.065	0.311	0.337	0.344	0.344	0.359	0.377
	2 ¹⁵	1.016	0.993	0.994	1.043	1.050	1.113	0.302	0.333	0.342	0.337	0.341	0.388
	2 ¹⁶	1.036	1.072	1.018	1.043	1.058	1.052	0.295	0.333	0.327	0.346	0.357	0.390
	2 ¹⁷	1.037	1.072	1.018	1.043	1.058	1.046	0.295	0.326	0.321	0.341	0.351	0.378

The best-performing fingerprint dimension in each column of the table is italicised, bold-faced and marked in red for ease of reference.

Table B-1 (continued)

Distance Coefficients	Fingerprint Dimensions	Partition											
		(a) <i>F</i> -Measure						(b) <i>QPI</i> -Measure					
		500	600	700	800	900	1000	500	600	700	800	900	1000
[D5] Hamming	2 ⁵	0.495	0.507	0.577	0.612	0.643	0.758	0.110	0.113	0.118	0.123	0.126	0.128
	2 ⁶	0.849	0.946	0.956	0.983	1.007	1.008	0.172	0.180	0.186	0.198	0.211	0.217
	2 ⁷	1.072	1.126	1.079	1.038	1.038	1.045	0.254	0.273	0.293	0.291	0.304	0.314
	2 ⁸	0.851	0.862	0.908	0.915	0.976	1.021	0.314	0.274	0.279	0.293	0.312	0.329
	2 ⁹	0.959	0.947	1.020	1.082	1.064	1.064	0.324	0.332	0.324	0.345	0.347	0.347
	2 ¹⁰	0.840	0.872	0.943	0.907	0.928	0.976	0.311	0.324	0.333	0.343	0.363	0.367
	2 ¹¹	0.825	0.940	0.910	0.931	0.929	1.053	0.307	0.325	0.328	0.325	0.347	0.359
	2 ¹²	0.950	0.950	0.962	0.962	1.011	1.011	0.299	0.299	0.325	0.325	0.349	0.349
	2 ¹³	0.949	0.949	0.949	0.990	0.990	0.990	0.298	0.298	0.298	0.340	0.340	0.340
	2 ¹⁴	0.764	0.764	0.764	0.764	0.764	0.764	0.228	0.228	0.228	0.228	0.228	0.228
	2 ¹⁵	0.248	0.248	0.906	0.906	0.906	0.906	0.084	0.084	0.336	0.336	0.336	0.336
	2 ¹⁶	0.758	0.758	0.758	0.758	0.758	0.758	0.146	0.146	0.146	0.146	0.146	0.146
	2 ¹⁷	0.161	0.161	0.161	0.161	0.161	0.161	0.000	0.000	0.000	0.000	0.000	0.000
	[D6] Jaccard	2 ⁵	0.588	0.584	0.622	0.698	0.716	0.766	0.097	0.101	0.104	0.109	0.113
2 ⁶		0.806	0.872	0.880	0.901	0.993	1.017	0.109	0.114	0.118	0.129	0.142	0.151
2 ⁷		0.736	0.798	0.930	0.949	0.968	0.930	0.145	0.160	0.183	0.196	0.209	0.222
2 ⁸		0.734	0.922	0.910	0.954	0.944	0.967	0.204	0.232	0.253	0.270	0.283	0.293
2 ⁹		0.874	0.837	0.846	0.932	0.989	1.039	0.244	0.243	0.253	0.264	0.275	0.286
2 ¹⁰		0.909	0.919	0.986	0.966	0.987	1.099	0.240	0.257	0.275	0.278	0.288	0.300
2 ¹¹		0.870	0.825	0.822	0.839	0.849	0.890	0.229	0.248	0.260	0.268	0.273	0.277
2 ¹²		0.933	0.952	0.952	0.907	0.949	0.936	0.229	0.244	0.249	0.258	0.275	0.287
2 ¹³		0.930	0.948	0.972	1.030	1.030	1.020	0.234	0.246	0.252	0.267	0.282	0.289
2 ¹⁴		0.904	0.949	0.971	0.983	1.026	1.071	0.249	0.241	0.266	0.272	0.281	0.290
2 ¹⁵		0.936	0.938	0.959	0.980	1.038	1.103	0.226	0.245	0.259	0.266	0.277	0.284
2 ¹⁶		0.914	0.932	0.932	0.980	1.015	1.103	0.233	0.250	0.256	0.262	0.274	0.285
2 ¹⁷		0.837	0.932	0.935	0.983	1.015	1.038	0.208	0.239	0.257	0.260	0.270	0.281
[D7] Kulsinski		2 ⁵	0.553	0.592	0.699	0.761	0.846	0.846	0.088	0.090	0.092	0.093	0.094
	2 ⁶	0.627	0.747	0.809	0.847	0.865	0.866	0.090	0.092	0.095	0.097	0.099	0.103
	2 ⁷	0.586	0.657	0.709	0.801	0.878	0.879	0.095	0.099	0.103	0.106	0.110	0.117
	2 ⁸	0.700	0.742	0.725	0.759	0.773	0.795	0.101	0.106	0.117	0.129	0.135	0.139
	2 ⁹	0.836	0.818	0.874	0.875	0.875	0.815	0.119	0.124	0.134	0.147	0.147	0.157
	2 ¹⁰	0.688	0.768	0.768	0.850	0.850	0.850	0.120	0.136	0.136	0.149	0.149	0.149
	2 ¹¹	0.456	0.648	0.648	0.648	0.648	0.808	0.096	0.134	0.134	0.134	0.134	0.211
	2 ¹²	0.569	0.569	0.569	0.569	0.569	0.569	0.109	0.109	0.109	0.109	0.109	0.109
	2 ¹³	0.161	0.161	0.786	0.786	0.786	0.786	0.081	0.081	0.170	0.170	0.170	0.170
	2 ¹⁴	0.386	0.386	0.386	0.386	0.386	0.386	0.090	0.090	0.090	0.090	0.090	0.090
	2 ¹⁵	0.161	0.161	0.161	0.161	0.161	0.161	0.000	0.000	0.000	0.000	0.000	0.000
	2 ¹⁶	0.161	0.161	0.161	0.161	0.161	0.161	0.000	0.000	0.000	0.000	0.000	0.000
	2 ¹⁷	0.161	0.161	0.161	0.161	0.161	0.161	0.000	0.000	0.000	0.000	0.000	0.000
	[D8] Rogers-Tanimoto	2 ⁵	0.725	0.773	0.802	0.813	0.951	0.969	0.113	0.115	0.118	0.122	0.128
2 ⁶		0.860	0.902	0.859	0.920	0.947	0.990	0.183	0.190	0.195	0.202	0.220	0.223
2 ⁷		0.902	0.962	0.995	1.003	1.023	1.027	0.253	0.259	0.274	0.283	0.307	0.307
2 ⁸		0.756	0.850	0.917	0.926	1.030	1.090	0.282	0.270	0.278	0.289	0.311	0.319
2 ⁹		0.923	0.975	0.964	1.031	1.039	1.059	0.304	0.329	0.336	0.344	0.349	0.345
2 ¹⁰		0.851	0.958	0.979	0.943	0.959	1.006	0.309	0.320	0.342	0.345	0.373	0.373
2 ¹¹		0.843	0.925	0.950	0.950	0.989	1.019	0.301	0.321	0.322	0.326	0.333	0.353
2 ¹²		0.947	0.960	0.904	0.933	1.019	1.042	0.293	0.311	0.323	0.320	0.309	0.318
2 ¹³		0.934	0.934	0.954	0.954	1.010	1.010	0.296	0.296	0.337	0.337	0.346	0.346
2 ¹⁴		0.996	0.996	0.996	1.058	1.058	1.058	0.302	0.302	0.302	0.352	0.352	0.352
2 ¹⁵		0.776	0.776	0.776	0.776	0.776	0.776	0.245	0.245	0.245	0.245	0.245	0.245
2 ¹⁶		0.317	0.317	1.033	1.033	1.033	1.033	0.083	0.083	0.335	0.335	0.335	0.335
2 ¹⁷		0.629	0.629	0.629	0.629	0.629	0.629	0.162	0.162	0.162	0.162	0.162	0.162

The best-performing fingerprint dimension in each column of the table is italicised, bold-faced and marked in red for ease of reference.

Table B-1 (continued)

Distance Coefficients	Fingerprint Dimensions	Partition											
		(a) <i>F</i> -Measure					(b) <i>QPI</i> -Measure						
		500	600	700	800	900	1000	500	600	700	800	900	1000
[D9] Russell-Rao	2 ⁵	0.546	0.622	0.698	0.699	0.699	0.774	0.084	0.085	0.086	0.087	0.087	0.088
	2 ⁶	0.562	0.633	0.634	0.729	0.805	0.834	0.086	0.087	0.089	0.090	0.092	0.093
	2 ⁷	0.601	0.646	0.700	0.765	0.816	0.816	0.089	0.091	0.093	0.097	0.098	0.101
	2 ⁸	0.600	0.656	0.671	0.694	0.767	0.867	0.095	0.101	0.103	0.109	0.113	0.115
	2 ⁹	0.692	0.678	0.683	0.714	0.717	0.737	0.106	0.115	0.124	0.132	0.140	0.150
	2 ¹⁰	0.734	0.734	0.774	0.774	0.806	0.806	0.117	0.117	0.130	0.130	0.150	0.150
	2 ¹¹	0.580	0.580	0.580	0.580	0.800	0.800	0.122	0.122	0.122	0.122	0.179	0.179
	2 ¹²	0.667	0.667	0.667	0.667	0.667	0.667	0.117	0.117	0.117	0.117	0.117	0.117
	2 ¹³	0.161	0.161	0.816	0.816	0.816	0.816	0.081	0.081	0.168	0.168	0.168	0.168
	2 ¹⁴	0.397	0.397	0.397	0.397	0.397	0.397	0.088	0.088	0.088	0.088	0.088	0.088
	2 ¹⁵	0.161	0.161	0.161	0.161	0.161	0.161	0.000	0.000	0.000	0.000	0.000	0.000
	2 ¹⁶	0.161	0.161	0.161	0.161	0.161	0.161	0.000	0.000	0.000	0.000	0.000	0.000
	2 ¹⁷	0.161	0.161	0.161	0.161	0.161	0.161	0.000	0.000	0.000	0.000	0.000	0.000
[D10] Sokal-Sneath	2 ⁵	0.574	0.654	0.675	0.699	0.738	0.756	0.097	0.100	0.104	0.108	0.112	0.116
	2 ⁶	0.641	0.805	0.902	0.903	0.997	1.019	0.108	0.116	0.125	0.131	0.139	0.143
	2 ⁷	0.729	0.858	0.859	0.911	0.936	0.994	0.141	0.158	0.183	0.203	0.206	0.221
	2 ⁸	0.850	0.814	0.858	0.950	0.982	0.966	0.211	0.243	0.249	0.272	0.291	0.295
	2 ⁹	0.853	0.839	0.921	1.000	1.014	1.036	0.235	0.248	0.255	0.269	0.282	0.290
	2 ¹⁰	0.791	0.817	0.860	0.873	0.916	0.999	0.239	0.247	0.262	0.265	0.268	0.278
	2 ¹¹	0.878	0.814	0.829	0.831	0.868	0.928	0.231	0.240	0.249	0.258	0.267	0.270
	2 ¹²	0.894	0.932	0.985	0.940	0.996	1.007	0.223	0.241	0.244	0.262	0.273	0.285
	2 ¹³	1.028	0.900	0.936	0.952	0.988	1.009	0.232	0.241	0.253	0.260	0.271	0.285
	2 ¹⁴	0.786	0.907	0.963	0.974	0.966	1.020	0.219	0.255	0.266	0.269	0.277	0.284
	2 ¹⁵	0.899	0.941	0.937	0.952	0.957	1.064	0.224	0.254	0.271	0.282	0.281	0.285
	2 ¹⁶	0.945	0.959	0.955	0.970	0.955	1.064	0.231	0.246	0.272	0.281	0.284	0.289
	2 ¹⁷	0.904	0.921	0.924	0.927	0.955	1.064	0.235	0.251	0.266	0.280	0.287	0.291

The best-performing fingerprint dimension in each column of the table is italicised, bold-faced and marked in red for ease of reference.

Appendix B Additional Results of Chapter 6

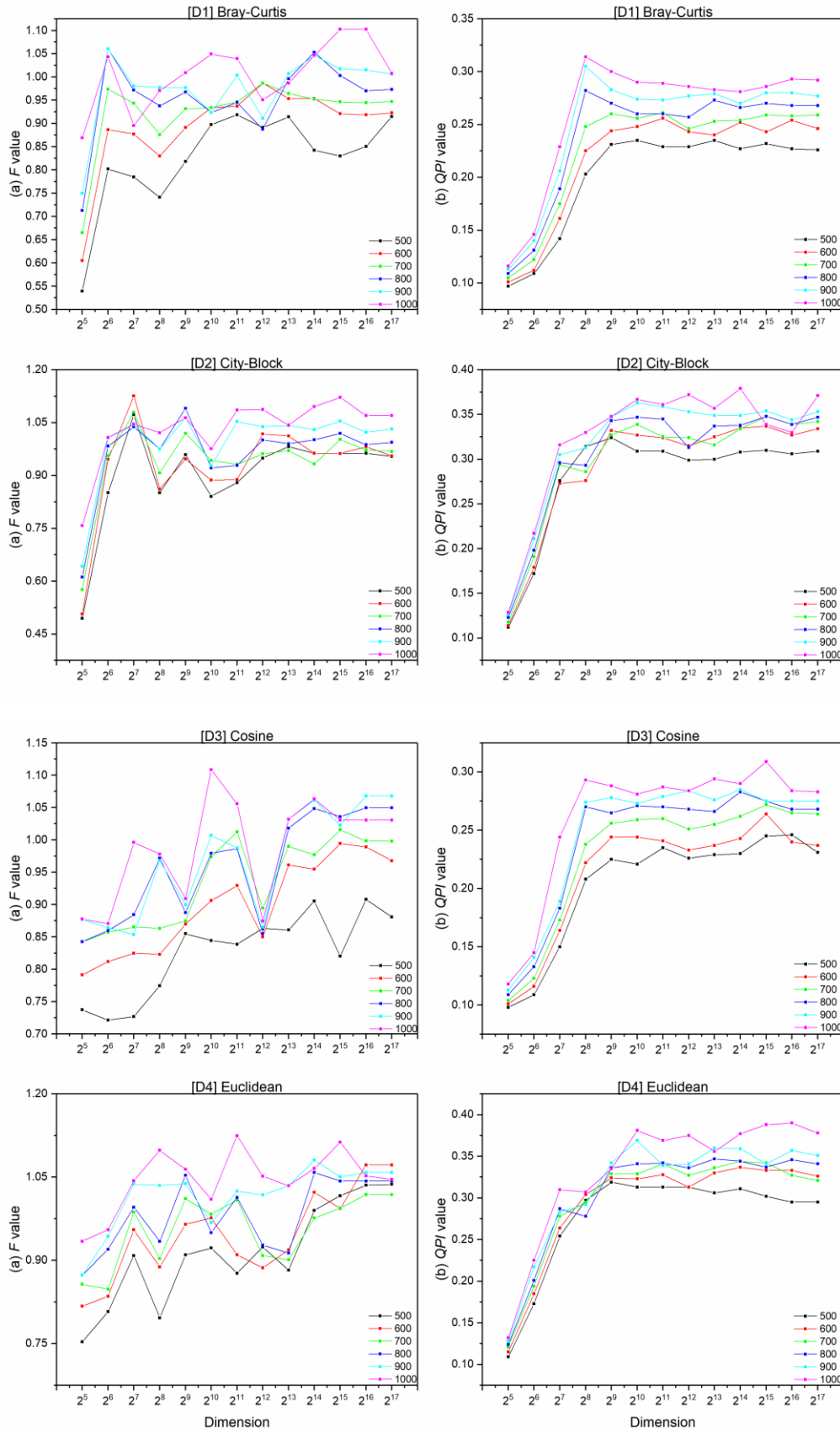


Figure B-1 Effects of dimensionality on Group Average clustering measured by (a) F -measure and (b) QPI -measure for MDDR dataset using various distance coefficients (Refer to Table B-1 for detail values)

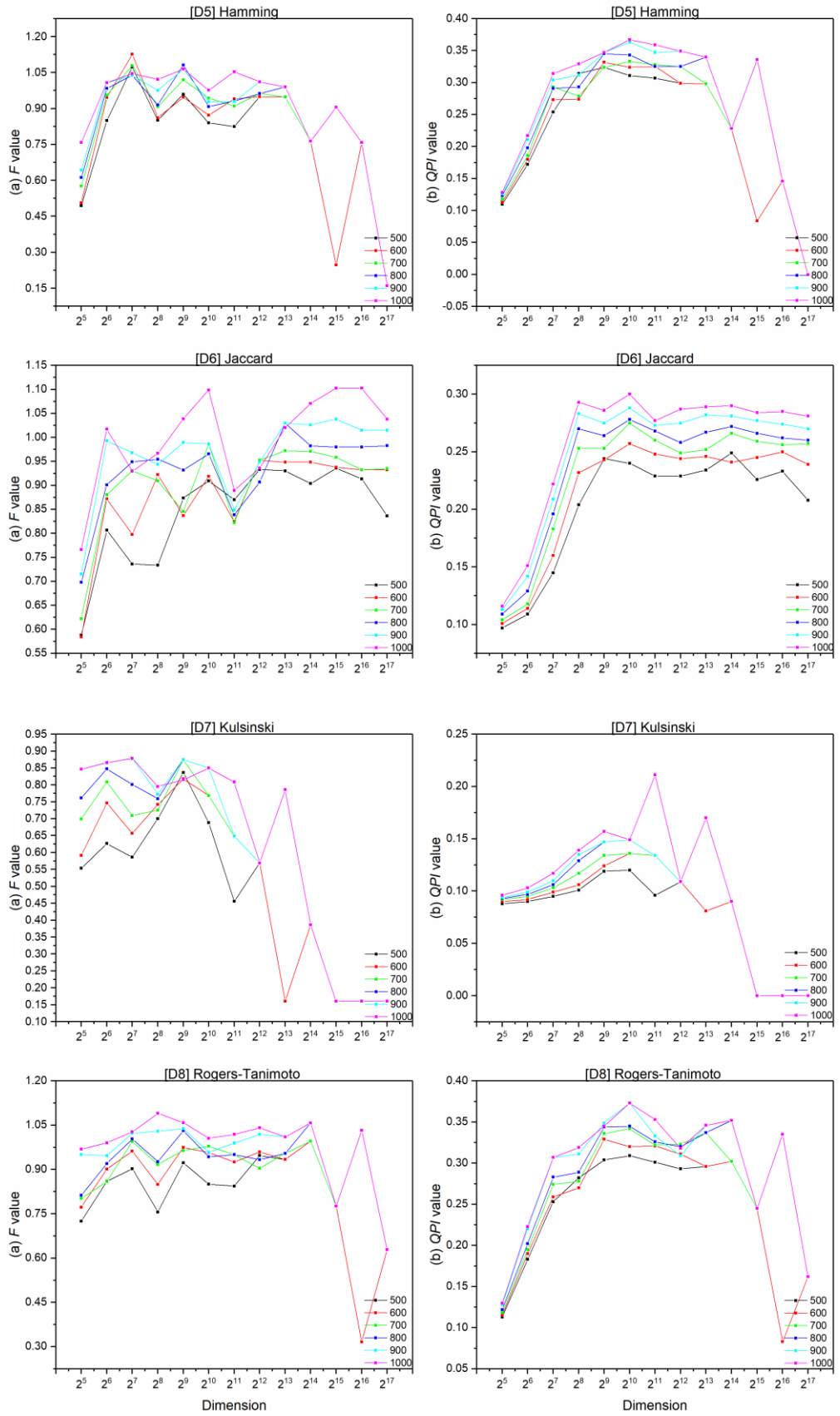


Figure B-1 (continued)

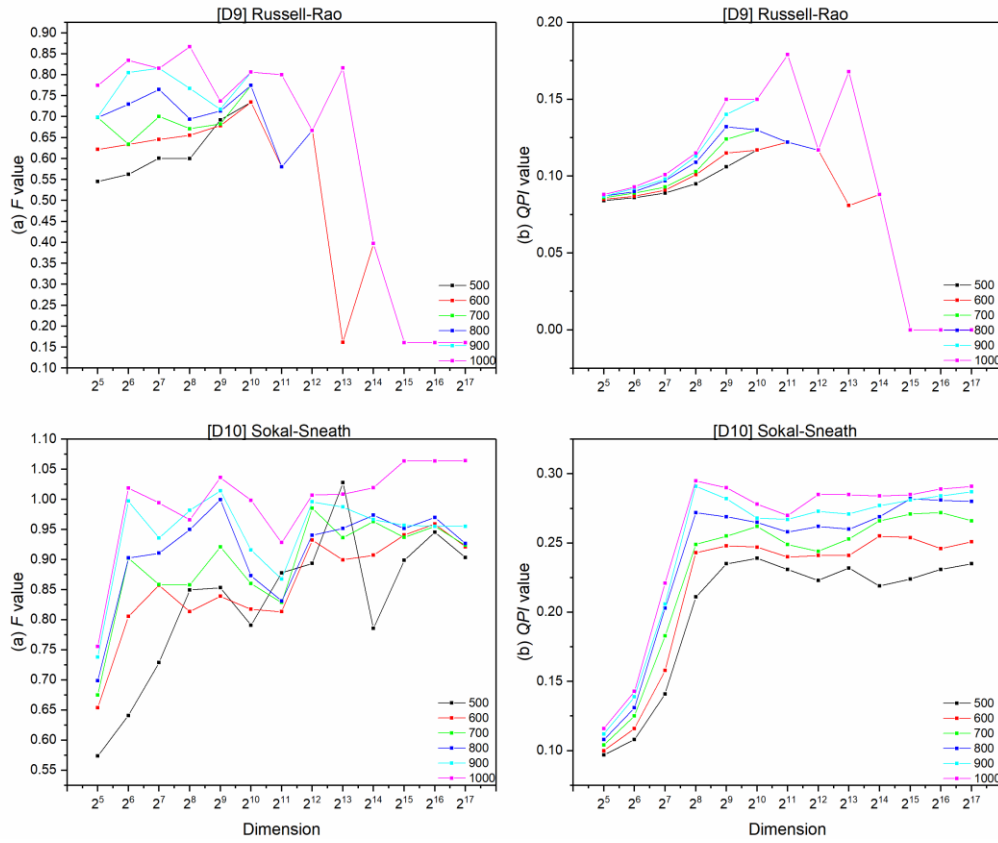


Figure B-1 (continued)

Table B-2 Effectiveness value of Ward’s clustering measured by (a) F -measure and (b) QPI -measure for the WOMBAT dataset using various distance coefficients and fingerprint dimensions. The range of the standard deviation, σ , for the mean F is between 0.055 and 0.336

Distance Coefficients	Fingerprint Dimensions	Partition											
		(a) F -Measure					(b) QPI -Measure						
		500	600	700	800	900	1000	500	600	700	800	900	1000
[D1] Bray-Curtis	2^5	0.430	0.449	0.494	0.530	0.552	0.555	0.103	0.110	0.116	0.119	0.123	0.126
	2^6	0.774	0.788	0.796	0.802	0.918	0.918	0.196	0.205	0.224	0.234	0.243	0.251
	2^7	0.779	0.808	0.846	0.866	0.891	0.931	0.263	0.271	0.272	0.287	0.307	0.320
	2^8	0.862	0.915	0.925	0.946	0.976	0.994	0.280	0.296	0.311	0.337	0.346	0.360
	2^9	0.954	0.974	0.968	0.973	0.974	0.988	0.269	0.279	0.306	0.331	0.357	0.370
	2^{10}	0.870	0.886	0.917	0.941	0.972	1.020	0.282	0.282	0.319	0.342	0.363	0.373
	2^{11}	0.899	0.932	0.958	0.969	1.004	1.004	0.275	0.289	0.306	0.333	0.349	0.374
	2^{12}	0.885	0.905	0.952	0.952	0.979	0.986	0.285	0.304	0.322	0.344	0.364	0.369
	2^{13}	0.920	0.949	0.955	0.967	0.987	0.999	0.267	0.277	0.301	0.323	0.343	0.354
	2^{14}	0.870	0.891	0.910	0.922	0.945	0.945	0.277	0.298	0.325	0.335	0.347	0.366
	2^{15}	0.963	0.983	0.951	0.951	0.987	0.987	0.297	0.303	0.326	0.330	0.339	0.361
	2^{16}	0.964	0.985	0.951	0.951	0.987	0.987	0.299	0.312	0.333	0.337	0.348	0.363
	2^{17}	0.964	0.985	0.942	0.951	0.987	0.987	0.311	0.329	0.348	0.341	0.347	0.363
[D2] City-Block	2^5	0.520	0.521	0.536	0.590	0.600	0.603	0.116	0.119	0.121	0.129	0.133	0.135
	2^6	0.887	0.908	0.927	0.999	1.049	1.068	0.192	0.206	0.219	0.234	0.248	0.260
	2^7	0.832	0.884	0.903	0.915	0.919	0.955	0.284	0.295	0.315	0.332	0.339	0.350
	2^8	0.890	0.920	0.935	0.940	0.990	0.998	0.296	0.318	0.321	0.334	0.343	0.368
	2^9	0.926	0.958	0.982	1.040	1.040	1.040	0.279	0.303	0.322	0.335	0.342	0.361
	2^{10}	0.916	0.953	0.971	1.040	1.040	1.049	0.303	0.313	0.344	0.345	0.351	0.379
	2^{11}	0.904	0.934	0.971	1.005	1.005	1.005	0.310	0.317	0.340	0.355	0.360	0.377
	2^{12}	1.033	1.017	1.022	1.051	1.101	1.101	0.325	0.354	0.351	0.356	0.384	0.399
	2^{13}	1.002	1.004	1.016	1.039	1.039	1.039	0.294	0.311	0.346	0.355	0.372	0.386
	2^{14}	0.955	1.011	1.011	1.041	1.041	1.041	0.297	0.311	0.313	0.342	0.367	0.397
	2^{15}	0.973	1.021	1.021	1.039	1.039	1.039	0.310	0.324	0.338	0.351	0.357	0.374
	2^{16}	0.962	1.021	1.031	1.039	1.039	1.039	0.301	0.323	0.329	0.344	0.350	0.375
	2^{17}	0.976	1.021	1.021	1.039	1.039	1.039	0.310	0.329	0.320	0.337	0.358	0.376
[D3] Cosine	2^5	0.472	0.530	0.562	0.596	0.669	0.677	0.111	0.117	0.122	0.124	0.127	0.128
	2^6	0.802	0.820	0.844	0.875	0.967	1.039	0.193	0.210	0.216	0.229	0.235	0.240
	2^7	0.856	0.899	0.873	0.873	0.910	0.910	0.253	0.275	0.306	0.310	0.322	0.328
	2^8	0.849	0.916	0.936	0.960	0.979	0.983	0.271	0.299	0.307	0.336	0.350	0.373
	2^9	0.903	0.956	0.920	0.931	0.932	0.932	0.292	0.302	0.316	0.349	0.371	0.386
	2^{10}	0.892	0.950	0.951	0.970	1.005	1.005	0.280	0.293	0.323	0.334	0.357	0.381
	2^{11}	0.864	0.944	0.944	0.957	0.995	0.995	0.264	0.277	0.298	0.317	0.343	0.367
	2^{12}	0.906	0.926	0.963	0.977	0.983	0.988	0.281	0.308	0.327	0.351	0.362	0.375
	2^{13}	0.847	0.888	0.891	0.906	0.947	0.947	0.295	0.301	0.312	0.337	0.361	0.380
	2^{14}	0.824	0.884	0.901	0.913	0.939	0.955	0.266	0.295	0.313	0.336	0.360	0.374
	2^{15}	0.872	0.928	0.944	0.968	0.991	0.995	0.281	0.303	0.318	0.333	0.350	0.365
	2^{16}	0.873	0.945	0.944	0.968	0.991	0.995	0.280	0.302	0.313	0.333	0.352	0.370
	2^{17}	0.876	0.940	0.939	0.968	0.991	0.995	0.278	0.292	0.311	0.318	0.338	0.353
[D4] Euclidean	2^5	0.520	0.563	0.589	0.641	0.640	0.753	0.120	0.125	0.128	0.129	0.132	0.135
	2^6	0.995	1.083	1.086	1.137	1.140	1.149	0.207	0.231	0.248	0.260	0.267	0.272
	2^7	0.901	0.925	0.925	0.932	0.940	0.974	0.257	0.271	0.294	0.324	0.335	0.338
	2^8	0.930	0.951	0.958	0.978	0.986	0.994	0.293	0.302	0.326	0.340	0.359	0.379
	2^9	1.009	0.973	0.995	0.986	0.990	1.040	0.294	0.323	0.355	0.362	0.378	0.380
	2^{10}	0.874	0.923	0.939	0.939	0.939	0.942	0.297	0.310	0.336	0.341	0.358	0.379
	2^{11}	0.928	0.950	0.967	0.987	0.987	1.028	0.300	0.318	0.348	0.360	0.368	0.384
	2^{12}	0.916	0.962	0.967	0.967	1.017	1.020	0.287	0.304	0.334	0.354	0.367	0.404
	2^{13}	0.917	0.966	0.986	1.041	1.041	1.040	0.311	0.332	0.342	0.365	0.369	0.383
	2^{14}	0.901	0.966	0.981	1.035	1.035	1.034	0.316	0.340	0.357	0.365	0.370	0.395
	2^{15}	0.925	0.965	0.986	1.041	1.044	1.040	0.308	0.318	0.345	0.357	0.362	0.389
	2^{16}	0.936	0.982	0.986	1.041	1.041	1.040	0.304	0.334	0.344	0.361	0.365	0.400
	2^{17}	0.898	0.928	0.971	1.025	1.025	1.024	0.290	0.314	0.332	0.346	0.351	0.386

The best-performing fingerprint dimension in each column of the table is italicised, bold-faced and marked in red for ease of reference.

Table B-2 (continued)

Distance Coefficients	Fingerprint Dimensions	(a) <i>F</i> -Measure						Partition						(b) <i>QPI</i> -Measure						
		500	600	700	800	900	1000	500	600	700	800	900	1000	500	600	700	800	900	1000	
[D5] Hamming	2 ⁵	0.520	0.521	0.536	0.590	0.600	0.603	0.116	0.119	0.121	0.129	0.133	0.135							
	2 ⁶	0.887	0.908	0.927	0.999	1.049	1.068	0.192	0.206	0.219	0.234	0.248	0.260							
	2 ⁷	0.832	0.884	0.903	0.915	0.919	0.955	0.286	0.294	0.311	0.332	0.339	0.349							
	2 ⁸	0.890	0.920	0.935	0.940	0.990	0.998	0.296	0.317	0.321	0.331	0.343	0.368							
	2 ⁹	0.926	0.958	0.982	1.040	1.040	1.040	0.279	0.303	0.322	0.335	0.342	0.361							
	2 ¹⁰	0.916	0.953	0.971	1.040	1.040	1.049	0.300	0.313	0.342	0.345	0.351	0.372							
	2 ¹¹	0.904	0.934	0.971	1.005	1.005	1.005	0.308	0.316	0.340	0.354	0.358	0.376							
	2 ¹²	1.033	1.017	1.013	1.025	1.101	1.101	0.322	0.354	0.356	0.361	0.379	0.398							
	2 ¹³	1.002	0.965	1.016	1.039	1.039	1.039	0.284	0.311	0.345	0.356	0.375	0.384							
	2 ¹⁴	0.936	0.997	0.997	1.041	1.041	1.041	0.299	0.309	0.309	0.335	0.335	0.371							
	2 ¹⁵	0.899	0.973	0.973	0.973	1.039	1.039	0.269	0.318	0.318	0.318	0.358	0.358							
	2 ¹⁶	0.896	0.896	1.031	1.031	1.031	1.031	0.245	0.245	0.323	0.323	0.323	0.323							
	2 ¹⁷	0.939	0.939	0.939	0.939	0.939	0.939	0.299	0.299	0.299	0.299	0.299	0.299							
	[D6] Jaccard	2 ⁵	0.500	0.530	0.534	0.559	0.564	0.586	0.108	0.113	0.117	0.120	0.122	0.126						
2 ⁶		0.979	1.006	1.067	1.068	1.068	1.050	0.200	0.211	0.214	0.226	0.238	0.241							
2 ⁷		0.890	0.931	0.944	0.962	1.010	1.010	0.256	0.279	0.285	0.305	0.321	0.329							
2 ⁸		0.908	0.942	0.955	0.982	0.988	0.992	0.285	0.301	0.312	0.332	0.350	0.359							
2 ⁹		0.943	0.964	0.964	0.996	1.000	1.000	0.275	0.301	0.323	0.344	0.350	0.368							
2 ¹⁰		0.950	0.950	0.977	1.000	1.000	1.000	0.283	0.290	0.312	0.330	0.340	0.359							
2 ¹¹		0.950	0.956	0.975	0.982	0.995	0.995	0.277	0.302	0.336	0.357	0.361	0.375							
2 ¹²		0.959	0.978	1.017	0.999	0.999	0.999	0.281	0.307	0.324	0.354	0.362	0.372							
2 ¹³		0.941	0.954	0.981	0.998	0.998	0.999	0.279	0.300	0.325	0.349	0.365	0.371							
2 ¹⁴		0.939	0.960	0.986	0.998	1.002	1.002	0.282	0.304	0.329	0.338	0.351	0.366							
2 ¹⁵		0.924	0.965	1.002	1.002	1.002	1.002	0.282	0.297	0.316	0.340	0.357	0.372							
2 ¹⁶		0.945	0.973	1.002	1.002	1.002	1.002	0.293	0.296	0.316	0.336	0.346	0.358							
2 ¹⁷		0.967	0.997	1.002	1.002	1.002	1.002	0.287	0.315	0.324	0.348	0.363	0.367							
[D7] Kulsinski		2 ⁵	0.427	0.454	0.510	0.601	0.606	0.669	0.113	0.117	0.121	0.125	0.128	0.131						
	2 ⁶	0.833	0.935	0.972	1.043	1.043	1.049	0.200	0.210	0.226	0.237	0.247	0.251							
	2 ⁷	0.921	0.953	1.000	1.000	0.996	0.996	0.292	0.301	0.307	0.320	0.327	0.337							
	2 ⁸	0.950	0.962	0.983	0.990	0.990	0.990	0.286	0.308	0.315	0.321	0.337	0.361							
	2 ⁹	0.944	0.952	0.985	0.990	0.990	0.990	0.284	0.297	0.328	0.336	0.345	0.355							
	2 ¹⁰	0.885	0.935	0.951	0.951	0.951	0.951	0.294	0.308	0.320	0.329	0.342	0.359							
	2 ¹¹	0.925	0.983	0.990	0.990	0.990	0.999	0.306	0.315	0.325	0.330	0.343	0.364							
	2 ¹²	0.878	0.933	0.949	0.949	0.949	0.949	0.334	0.349	0.341	0.365	0.367	0.381							
	2 ¹³	0.928	0.978	1.003	1.003	1.003	1.002	0.296	0.314	0.324	0.338	0.345	0.355							
	2 ¹⁴	0.975	0.980	1.002	1.002	1.002	1.002	0.283	0.294	0.319	0.327	0.352	0.352							
	2 ¹⁵	0.940	0.940	1.003	1.003	1.003	1.002	0.296	0.296	0.329	0.329	0.329	0.366							
	2 ¹⁶	0.883	0.972	0.972	0.972	0.972	0.972	0.246	0.334	0.334	0.334	0.334	0.334							
	2 ¹⁷	0.784	0.784	0.784	0.784	1.036	1.036	0.239	0.239	0.239	0.239	0.335	0.335							
	[D8] Rogers-Tanimoto	2 ⁵	0.484	0.493	0.529	0.593	0.624	0.630	0.115	0.118	0.122	0.125	0.128	0.133						
2 ⁶		0.940	1.025	1.021	1.024	1.039	1.108	0.214	0.229	0.239	0.252	0.263	0.277							
2 ⁷		0.845	0.901	0.925	0.941	0.981	0.996	0.253	0.286	0.310	0.323	0.328	0.343							
2 ⁸		0.921	0.948	0.972	0.975	0.978	0.990	0.311	0.321	0.322	0.324	0.347	0.378							
2 ⁹		0.927	0.953	1.007	1.029	1.029	1.029	0.285	0.316	0.336	0.355	0.362	0.377							
2 ¹⁰		0.879	0.916	0.937	0.995	0.995	1.006	0.312	0.324	0.347	0.356	0.349	0.359							
2 ¹¹		0.914	0.967	0.971	1.039	1.039	1.039	0.302	0.334	0.339	0.344	0.346	0.376							
2 ¹²		1.006	0.995	1.006	1.041	1.091	1.091	0.305	0.331	0.353	0.361	0.378	0.393							
2 ¹³		0.998	1.002	1.021	1.039	1.039	1.039	0.291	0.340	0.352	0.367	0.383	0.388							
2 ¹⁴		0.965	0.978	1.011	1.039	1.039	1.039	0.290	0.298	0.308	0.342	0.368	0.388							
2 ¹⁵		0.951	1.016	1.016	1.029	1.029	1.039	0.305	0.320	0.320	0.341	0.341	0.361							
2 ¹⁶		0.914	0.976	0.976	0.976	1.039	1.039	0.273	0.312	0.312	0.312	0.352	0.352							
2 ¹⁷		0.842	0.842	1.021	1.021	1.021	1.021	0.258	0.258	0.319	0.319	0.319	0.319							

The best-performing fingerprint dimension in each column of the table is italicised, bold-faced and marked in red for ease of reference.

Table B-2 (continued)

Distance Coefficients	Fingerprint Dimensions	(a) <i>F</i> -Measure								Partition					
		500	600	700	800	900	1000	500	600	(b) <i>QPI</i> -Measure					
											700	800	900	1000	
[D9] Russell-Rao	2 ⁵	0.468	0.483	0.487	0.552	0.585	0.596	0.105	0.108	0.110	0.114	0.116	0.119		
	2 ⁶	0.919	1.021	1.010	1.010	0.998	1.002	0.190	0.207	0.218	0.228	0.234	0.246		
	2 ⁷	0.927	0.946	0.950	0.997	0.997	1.004	0.267	0.278	0.295	0.307	0.325	0.333		
	2 ⁸	0.955	0.959	0.987	1.005	0.994	0.994	0.268	0.285	0.309	0.335	0.349	0.348		
	2 ⁹	0.952	0.975	0.985	0.994	0.994	0.994	0.290	0.330	0.363	0.355	0.373	0.388		
	2 ¹⁰	0.932	0.979	1.002	1.002	1.002	1.002	0.282	0.327	0.345	0.361	0.374	0.386		
	2 ¹¹	0.923	0.995	1.002	1.002	1.040	1.040	0.303	0.312	0.320	0.337	0.353	0.367		
	2 ¹²	0.891	0.915	0.956	0.956	1.006	1.052	0.299	0.309	0.317	0.341	0.365	0.368		
	2 ¹³	0.952	0.969	1.003	1.003	1.003	1.002	0.309	0.332	0.326	0.331	0.336	0.338		
	2 ¹⁴	0.904	0.924	1.002	1.002	1.002	1.002	0.287	0.284	0.298	0.321	0.344	0.344		
	2 ¹⁵	0.941	0.941	1.002	1.002	1.002	1.002	0.314	0.314	0.333	0.333	0.333	0.365		
	2 ¹⁶	0.864	0.953	0.953	0.953	0.953	0.953	0.268	0.330	0.330	0.330	0.330	0.330		
	2 ¹⁷	0.839	0.839	0.839	0.839	1.036	1.036	0.231	0.231	0.231	0.231	0.342	0.342		
	[D10] Sokal-Sneath	2 ⁵	0.488	0.567	0.570	0.592	0.611	0.661	0.115	0.119	0.124	0.127	0.133	0.135	
		2 ⁶	0.920	1.021	1.047	1.117	1.146	1.146	0.208	0.228	0.243	0.257	0.259	0.264	
		2 ⁷	0.941	0.941	0.974	0.992	0.999	1.016	0.249	0.264	0.287	0.310	0.327	0.350	
		2 ⁸	0.950	0.991	1.050	1.020	1.020	1.020	0.307	0.308	0.322	0.336	0.335	0.364	
2 ⁹		0.952	0.984	1.002	1.002	1.002	1.002	0.268	0.279	0.299	0.309	0.332	0.350		
2 ¹⁰		0.923	0.969	0.990	0.990	0.990	0.990	0.292	0.339	0.334	0.335	0.349	0.364		
2 ¹¹		0.888	0.934	0.951	0.951	0.951	0.951	0.256	0.296	0.315	0.318	0.330	0.353		
2 ¹²		0.959	0.964	1.002	1.002	1.002	1.002	0.319	0.333	0.347	0.353	0.357	0.373		
2 ¹³		0.981	1.034	1.043	1.034	1.034	1.034	0.320	0.342	0.354	0.365	0.377	0.390		
2 ¹⁴		0.976	0.978	1.002	1.002	1.002	1.002	0.300	0.318	0.348	0.354	0.355	0.357		
2 ¹⁵		0.976	1.000	1.002	1.002	1.002	1.002	0.303	0.321	0.332	0.334	0.347	0.359		
2 ¹⁶		0.975	1.000	1.002	1.002	1.002	1.002	0.346	0.360	0.351	0.358	0.372	0.380		
2 ¹⁷		0.975	1.012	1.002	1.002	1.002	1.002	0.312	0.356	0.356	0.361	0.363	0.394		

The best-performing fingerprint dimension in each column of the table is italicised, bold-faced and marked in red for ease of reference.

Appendix B Additional Results of Chapter 6

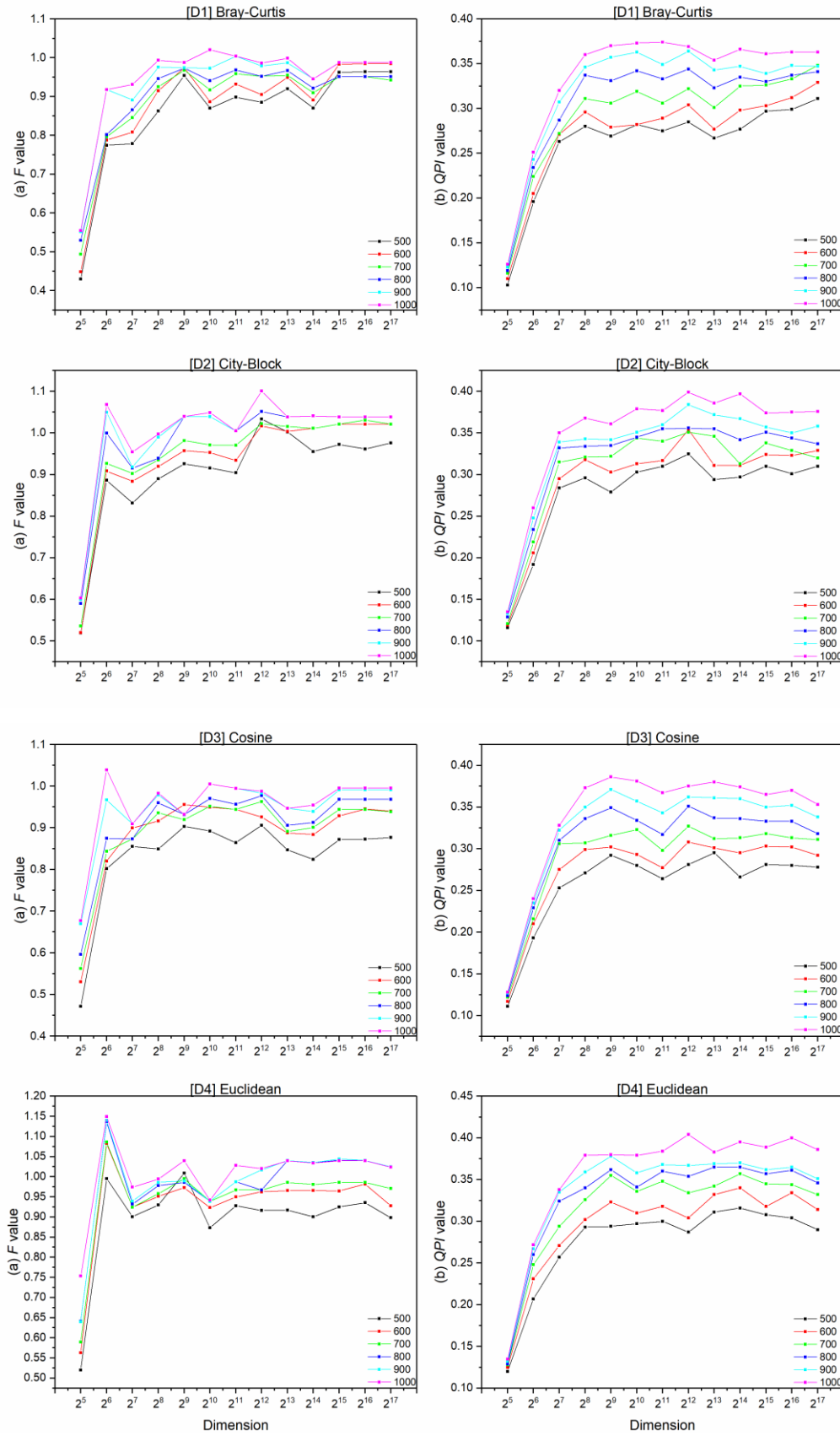


Figure B-2 Effects of dimensionality on Ward's clustering measured by (a) F -measure and (b) QPI -measure for WOMBAT dataset using various distance coefficients (Refer to Table B-2 for detail values)

Appendix B Additional Results of Chapter 6

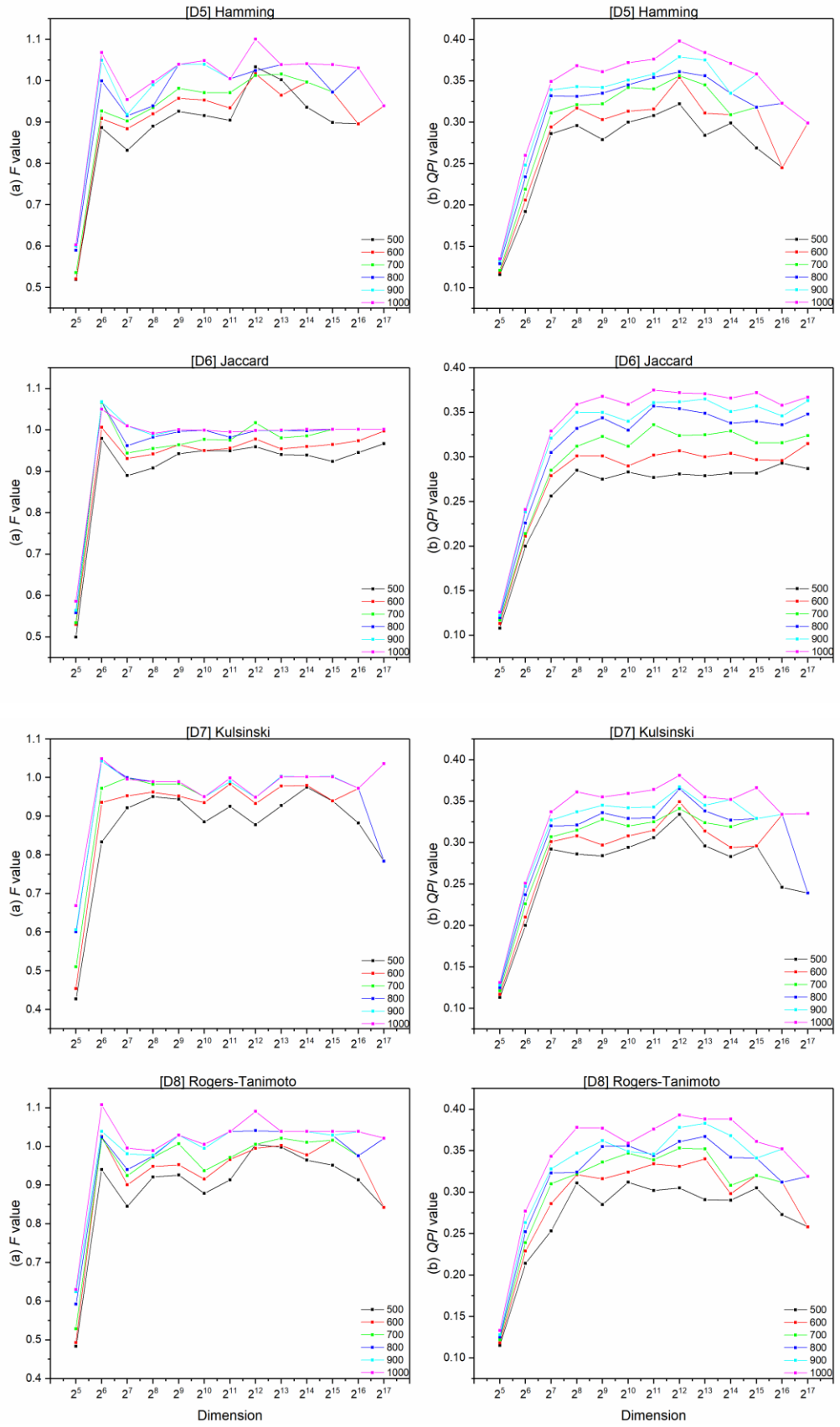


Figure B-2 (continued)

Appendix B Additional Results of Chapter 6

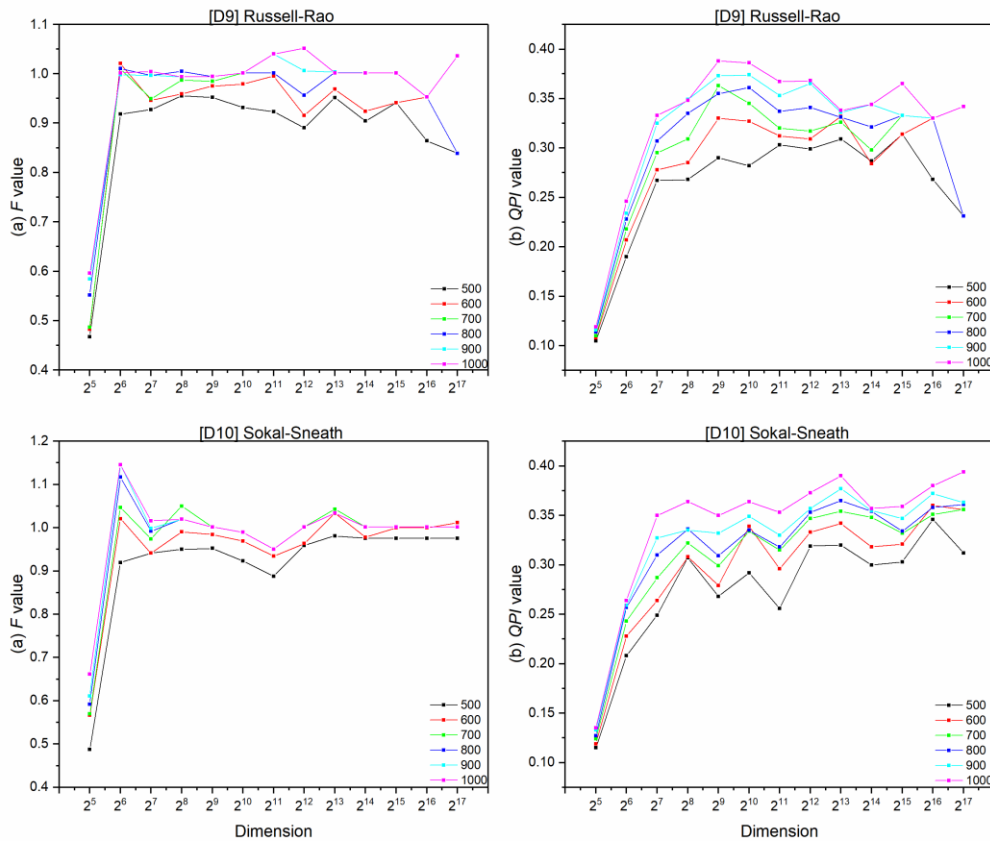


Figure B-2 (continued)

Table B-3 Effectiveness value of Group Average clustering measured by (a) *F*-measure and (b) *QPI*-measure for the WOMBAT dataset using various distance coefficients and fingerprint dimensions. The range of the standard deviation, σ , for the mean *F* is between 0.000 and 0.466

Distance Coefficients	Fingerprint Dimensions	Partition											
		(a) <i>F</i> -Measure						(b) <i>QPI</i> -Measure					
		500	600	700	800	900	1000	500	600	700	800	900	1000
[D1] Bray-Curtis	2 ⁵	0.393	0.400	0.495	0.514	0.560	0.578	0.076	0.078	0.081	0.084	0.086	0.089
	2 ⁶	0.434	0.460	0.600	0.644	0.669	0.688	0.088	0.092	0.101	0.106	0.113	0.123
	2 ⁷	0.469	0.533	0.624	0.621	0.683	0.718	0.131	0.135	0.155	0.190	0.194	0.191
	2 ⁸	0.630	0.795	0.785	0.817	0.897	0.917	0.166	0.180	0.191	0.209	0.226	0.242
	2 ⁹	0.748	0.802	0.843	0.856	0.869	0.894	0.190	0.214	0.222	0.250	0.250	0.257
	2 ¹⁰	0.725	0.753	0.805	0.835	0.909	0.909	0.186	0.224	0.245	0.244	0.264	0.273
	2 ¹¹	0.767	0.914	0.903	0.903	0.929	0.957	0.211	0.215	0.236	0.246	0.258	0.311
	2 ¹²	0.768	0.854	0.874	0.915	0.937	0.963	0.205	0.229	0.241	0.257	0.319	0.326
	2 ¹³	0.917	0.932	0.893	0.898	0.941	0.966	0.205	0.228	0.241	0.254	0.307	0.313
	2 ¹⁴	0.788	0.846	0.901	0.914	0.922	0.966	0.207	0.234	0.234	0.270	0.310	0.323
	2 ¹⁵	0.784	0.847	0.899	0.912	0.914	0.961	0.215	0.228	0.234	0.276	0.310	0.321
	2 ¹⁶	0.775	0.847	0.899	0.912	0.936	0.961	0.208	0.223	0.234	0.273	0.310	0.321
	2 ¹⁷	0.733	0.843	0.911	0.912	0.936	0.961	0.199	0.238	0.232	0.272	0.307	0.317
[D2] City-Block	2 ⁵	0.407	0.480	0.549	0.557	0.614	0.648	0.089	0.091	0.096	0.097	0.099	0.101
	2 ⁶	0.712	0.721	0.773	0.838	0.901	0.927	0.178	0.190	0.188	0.194	0.202	0.214
	2 ⁷	0.922	0.985	0.999	1.004	0.999	0.999	0.267	0.281	0.264	0.302	0.315	0.332
	2 ⁸	0.850	0.905	0.973	0.974	1.005	1.013	0.238	0.254	0.303	0.391	0.388	0.402
	2 ⁹	0.792	0.910	0.976	0.979	0.942	0.997	0.246	0.238	0.269	0.294	0.313	0.344
	2 ¹⁰	0.821	0.881	0.925	0.946	0.966	0.987	0.253	0.263	0.280	0.280	0.304	0.345
	2 ¹¹	0.826	0.858	0.937	0.951	0.964	0.983	0.246	0.306	0.324	0.321	0.328	0.338
	2 ¹²	0.840	0.874	0.905	0.973	0.987	0.981	0.251	0.313	0.312	0.325	0.330	0.355
	2 ¹³	0.841	0.857	0.911	0.949	0.975	0.976	0.242	0.273	0.320	0.331	0.350	0.350
	2 ¹⁴	0.805	0.864	0.926	0.926	0.987	0.936	0.241	0.303	0.310	0.280	0.356	0.380
	2 ¹⁵	0.845	0.921	0.939	0.959	0.987	0.937	0.241	0.294	0.323	0.292	0.357	0.380
	2 ¹⁶	0.916	0.921	0.941	0.945	0.966	0.922	0.231	0.288	0.327	0.287	0.347	0.355
	2 ¹⁷	0.875	0.920	0.936	0.940	0.966	0.922	0.236	0.320	0.328	0.280	0.346	0.358
[D3] Cosine	2 ⁵	0.306	0.365	0.414	0.467	0.496	0.541	0.076	0.079	0.081	0.083	0.086	0.089
	2 ⁶	0.435	0.483	0.569	0.643	0.669	0.691	0.087	0.095	0.100	0.106	0.114	0.121
	2 ⁷	0.549	0.570	0.610	0.637	0.680	0.743	0.123	0.136	0.150	0.179	0.193	0.197
	2 ⁸	0.643	0.742	0.821	0.839	0.926	0.927	0.171	0.181	0.200	0.217	0.227	0.241
	2 ⁹	0.750	0.781	0.810	0.821	0.851	0.890	0.179	0.198	0.221	0.218	0.247	0.249
	2 ¹⁰	0.728	0.832	0.827	0.850	0.906	0.939	0.215	0.220	0.240	0.249	0.274	0.283
	2 ¹¹	0.818	0.891	0.899	0.906	0.920	0.960	0.204	0.232	0.229	0.240	0.258	0.318
	2 ¹²	0.854	0.921	0.933	0.944	0.944	0.957	0.210	0.234	0.243	0.271	0.275	0.354
	2 ¹³	0.843	0.810	0.874	0.907	0.921	0.956	0.195	0.217	0.222	0.250	0.260	0.345
	2 ¹⁴	0.819	0.821	0.866	0.904	0.923	0.956	0.199	0.228	0.240	0.244	0.267	0.329
	2 ¹⁵	0.798	0.845	0.862	0.905	0.909	0.957	0.195	0.224	0.234	0.248	0.281	0.350
	2 ¹⁶	0.728	0.906	0.925	0.961	0.961	0.987	0.203	0.237	0.247	0.248	0.267	0.348
	2 ¹⁷	0.793	0.906	0.925	0.961	0.961	0.987	0.200	0.243	0.254	0.251	0.267	0.350
[D4] Euclidean	2 ⁵	0.448	0.535	0.559	0.588	0.645	0.674	0.088	0.090	0.095	0.097	0.098	0.102
	2 ⁶	0.652	0.707	0.781	0.847	0.858	0.979	0.180	0.188	0.201	0.208	0.213	0.226
	2 ⁷	0.940	0.979	0.959	0.958	0.958	0.962	0.224	0.253	0.262	0.310	0.304	0.316
	2 ⁸	0.868	0.901	0.942	0.981	0.989	0.990	0.258	0.282	0.279	0.313	0.362	0.370
	2 ⁹	0.836	0.886	0.940	0.946	0.947	0.983	0.249	0.261	0.252	0.285	0.307	0.338
	2 ¹⁰	0.817	0.857	0.925	0.944	0.973	0.975	0.219	0.262	0.277	0.268	0.284	0.320
	2 ¹¹	0.832	0.873	0.917	0.917	0.966	1.008	0.234	0.297	0.315	0.306	0.332	0.338
	2 ¹²	0.883	0.936	0.955	0.960	0.964	0.973	0.242	0.283	0.312	0.297	0.302	0.361
	2 ¹³	0.877	0.901	0.945	0.930	0.936	0.984	0.255	0.329	0.343	0.299	0.314	0.364
	2 ¹⁴	0.865	0.892	0.930	0.915	0.970	0.922	0.246	0.280	0.312	0.307	0.348	0.348
	2 ¹⁵	0.847	0.913	0.925	0.928	0.980	0.937	0.252	0.282	0.329	0.322	0.344	0.337
	2 ¹⁶	0.833	0.882	0.910	0.919	0.972	0.928	0.233	0.286	0.316	0.322	0.334	0.348
	2 ¹⁷	0.841	0.877	0.923	0.924	0.984	0.934	0.231	0.287	0.311	0.315	0.338	0.354

The best-performing fingerprint dimension in each column of the table is italicised, bold-faced and marked in red for ease of reference.

Table B-3 (continued)

Distance Coefficients	Fingerprint Dimensions	(a) <i>F</i> -Measure						(b) <i>QPI</i> -Measure						
		500	600	700	800	900	1000	500	600	700	800	900	1000	
[D5] Hamming	2 ⁵	0.407	0.480	0.529	0.557	0.614	0.648	0.089	0.091	0.095	0.097	0.099	0.101	
	2 ⁶	0.707	0.721	0.759	0.838	0.901	0.927	0.176	0.190	0.189	0.194	0.201	0.214	
	2 ⁷	0.916	0.979	0.999	1.004	0.999	0.999	0.267	0.281	0.264	0.302	0.316	0.332	
	2 ⁸	0.820	0.905	0.957	0.974	1.003	1.013	0.223	0.250	0.301	0.391	0.387	0.402	
	2 ⁹	0.792	0.910	0.954	0.979	0.942	0.979	0.246	0.238	0.257	0.277	0.310	0.340	
	2 ¹⁰	0.788	0.881	0.888	0.946	0.966	0.987	0.234	0.262	0.282	0.282	0.297	0.344	
	2 ¹¹	0.826	0.858	0.925	0.943	0.963	0.963	0.246	0.303	0.310	0.296	0.329	0.329	
	2 ¹²	0.772	0.772	0.875	0.973	0.973	0.973	0.228	0.228	0.318	0.290	0.290	0.290	
	2 ¹³	0.743	0.858	0.858	0.858	0.858	0.975	0.154	0.250	0.250	0.250	0.250	0.346	
	2 ¹⁴	0.619	0.619	0.619	0.926	0.926	0.926	0.133	0.133	0.133	0.309	0.309	0.309	
	2 ¹⁵	0.805	0.805	0.805	0.805	0.805	0.805	0.214	0.214	0.214	0.214	0.214	0.214	
	2 ¹⁶	0.467	0.467	0.467	0.467	0.467	0.467	0.091	0.091	0.091	0.091	0.091	0.091	
	2 ¹⁷	0.126	0.126	0.126	0.126	0.126	0.126	0.000	0.000	0.000	0.000	0.000	0.000	
	[D6] Jaccard	2 ⁵	0.388	0.427	0.460	0.491	0.586	0.618	0.075	0.078	0.080	0.083	0.085	0.088
		2 ⁶	0.408	0.444	0.582	0.706	0.732	0.745	0.085	0.092	0.100	0.105	0.114	0.120
		2 ⁷	0.460	0.516	0.675	0.716	0.691	0.795	0.121	0.137	0.149	0.164	0.196	0.196
		2 ⁸	0.627	0.796	0.790	0.803	0.832	0.918	0.172	0.179	0.198	0.203	0.223	0.238
2 ⁹		0.780	0.821	0.840	0.874	0.878	0.903	0.205	0.216	0.222	0.242	0.251	0.262	
2 ¹⁰		0.745	0.828	0.832	0.848	0.958	0.958	0.214	0.237	0.248	0.257	0.271	0.279	
2 ¹¹		0.734	0.867	0.891	0.892	0.918	0.943	0.202	0.219	0.231	0.241	0.257	0.271	
2 ¹²		0.806	0.903	0.874	0.887	0.920	0.929	0.204	0.233	0.245	0.249	0.268	0.306	
2 ¹³		0.896	0.896	0.903	0.903	0.933	0.939	0.220	0.219	0.245	0.256	0.306	0.312	
2 ¹⁴		0.771	0.812	0.881	0.904	0.928	0.930	0.202	0.222	0.233	0.275	0.296	0.314	
2 ¹⁵		0.845	0.868	0.894	0.908	0.924	0.950	0.206	0.221	0.230	0.248	0.292	0.295	
2 ¹⁶		0.855	0.858	0.894	0.908	0.929	0.950	0.214	0.231	0.233	0.253	0.259	0.291	
2 ¹⁷		0.855	0.858	0.894	0.908	0.924	0.950	0.212	0.232	0.232	0.285	0.287	0.291	
[D7] Kulsinski		2 ⁵	0.327	0.391	0.425	0.502	0.520	0.556	0.068	0.069	0.070	0.071	0.073	0.074
		2 ⁶	0.318	0.337	0.389	0.435	0.462	0.462	0.070	0.071	0.073	0.075	0.077	0.079
		2 ⁷	0.325	0.416	0.441	0.432	0.438	0.500	0.073	0.077	0.081	0.084	0.087	0.091
		2 ⁸	0.331	0.334	0.379	0.427	0.513	0.538	0.081	0.083	0.092	0.099	0.105	0.109
	2 ⁹	0.408	0.579	0.618	0.618	0.720	0.730	0.093	0.100	0.110	0.110	0.118	0.128	
	2 ¹⁰	0.429	0.429	0.515	0.515	0.586	0.586	0.100	0.100	0.118	0.118	0.131	0.131	
	2 ¹¹	0.287	0.475	0.475	0.475	0.475	0.608	0.082	0.120	0.120	0.120	0.120	0.156	
	2 ¹²	0.350	0.350	0.350	0.350	0.350	0.350	0.098	0.098	0.098	0.098	0.098	0.098	
	2 ¹³	0.127	0.127	0.127	0.531	0.531	0.531	0.064	0.064	0.064	0.131	0.131	0.131	
	2 ¹⁴	0.163	0.163	0.163	0.163	0.163	0.163	0.073	0.073	0.073	0.073	0.073	0.073	
	2 ¹⁵	0.126	0.126	0.126	0.126	0.126	0.126	0.000	0.000	0.000	0.000	0.000	0.000	
	2 ¹⁶	0.126	0.126	0.126	0.126	0.126	0.126	0.000	0.000	0.000	0.000	0.000	0.000	
	2 ¹⁷	0.126	0.126	0.126	0.126	0.126	0.126	0.000	0.000	0.000	0.000	0.000	0.000	
	[D8] Rogers-Tanimoto	2 ⁵	0.414	0.438	0.513	0.559	0.653	0.674	0.085	0.089	0.094	0.098	0.102	0.103
		2 ⁶	0.646	0.701	0.769	0.872	0.878	0.913	0.168	0.178	0.184	0.195	0.205	0.211
		2 ⁷	0.869	0.966	0.932	0.942	0.949	0.943	0.227	0.236	0.263	0.301	0.284	0.294
		2 ⁸	0.901	0.948	0.956	1.042	1.063	1.058	0.262	0.273	0.283	0.273	0.300	0.320
2 ⁹		0.811	0.888	0.971	0.975	0.982	0.997	0.254	0.259	0.280	0.292	0.301	0.328	
2 ¹⁰		0.832	0.861	0.894	0.954	0.944	0.976	0.310	0.268	0.269	0.270	0.286	0.305	
2 ¹¹		0.832	0.838	0.925	0.937	0.973	0.992	0.233	0.268	0.290	0.278	0.319	0.338	
2 ¹²		0.738	0.837	0.914	0.985	0.948	0.948	0.234	0.258	0.309	0.298	0.324	0.324	
2 ¹³		0.810	0.810	0.916	0.939	0.939	0.939	0.240	0.240	0.278	0.289	0.289	0.289	
2 ¹⁴		0.759	0.868	0.868	0.868	0.868	0.891	0.151	0.256	0.256	0.256	0.256	0.355	
2 ¹⁵		0.586	0.586	0.586	0.931	0.931	0.931	0.134	0.134	0.134	0.306	0.306	0.306	
2 ¹⁶		0.822	0.822	0.822	0.822	0.822	0.822	0.218	0.218	0.218	0.218	0.218	0.218	
2 ¹⁷		0.466	0.466	0.466	0.466	0.466	0.466	0.091	0.091	0.091	0.091	0.091	0.091	

The best-performing fingerprint dimension in each column of the table is italicised, bold-faced and marked in red for ease of reference.

Table B-3 (continued)

Distance Coefficients	Fingerprint Dimensions	Partition											
		(a) <i>F</i> -Measure				(b) <i>QPI</i> -Measure							
		500	600	700	800	900	1000	500	600	700	800	900	1000
[D9] Russell-Rao	2 ⁵	0.355	0.356	0.381	0.381	0.420	0.459	0.066	0.066	0.067	0.067	0.068	0.069
	2 ⁶	0.275	0.328	0.367	0.397	0.437	0.461	0.067	0.068	0.069	0.069	0.071	0.071
	2 ⁷	0.291	0.348	0.435	0.470	0.535	0.535	0.070	0.070	0.073	0.074	0.078	0.078
	2 ⁸	0.327	0.364	0.397	0.471	0.483	0.498	0.074	0.076	0.081	0.087	0.089	0.093
	2 ⁹	0.511	0.633	0.646	0.675	0.748	0.811	0.080	0.089	0.094	0.100	0.104	0.117
	2 ¹⁰	0.325	0.538	0.538	0.601	0.601	0.601	0.087	0.102	0.102	0.118	0.118	0.136
	2 ¹¹	0.222	0.402	0.402	0.402	0.402	0.603	0.078	0.107	0.107	0.107	0.107	0.137
	2 ¹²	0.338	0.338	0.338	0.338	0.338	0.338	0.093	0.093	0.093	0.093	0.093	0.093
	2 ¹³	0.127	0.127	0.127	0.619	0.619	0.619	0.064	0.064	0.064	0.138	0.138	0.138
	2 ¹⁴	0.178	0.178	0.178	0.178	0.178	0.178	0.075	0.075	0.075	0.075	0.075	0.075
	2 ¹⁵	0.126	0.126	0.126	0.126	0.126	0.126	0.000	0.000	0.000	0.000	0.000	0.000
	2 ¹⁶	0.126	0.126	0.126	0.126	0.126	0.126	0.000	0.000	0.000	0.000	0.000	0.000
	2 ¹⁷	0.126	0.126	0.126	0.126	0.126	0.126	0.000	0.000	0.000	0.000	0.000	0.000
[D10] Sokal-Sneath	2 ⁵	0.365	0.420	0.422	0.474	0.546	0.570	0.075	0.077	0.079	0.082	0.084	0.087
	2 ⁶	0.425	0.531	0.699	0.733	0.737	0.739	0.088	0.092	0.102	0.107	0.117	0.121
	2 ⁷	0.476	0.587	0.653	0.675	0.735	0.802	0.127	0.140	0.155	0.185	0.187	0.196
	2 ⁸	0.637	0.767	0.796	0.808	0.908	0.912	0.167	0.185	0.194	0.207	0.228	0.233
	2 ⁹	0.758	0.808	0.827	0.874	0.884	0.910	0.200	0.210	0.233	0.248	0.267	0.265
	2 ¹⁰	0.735	0.744	0.785	0.832	0.917	0.917	0.210	0.229	0.246	0.260	0.276	0.286
	2 ¹¹	0.723	0.816	0.850	0.853	0.875	0.923	0.215	0.223	0.232	0.238	0.252	0.265
	2 ¹²	0.855	0.903	0.869	0.871	0.908	0.918	0.205	0.231	0.244	0.257	0.323	0.332
	2 ¹³	0.862	0.888	0.897	0.910	0.912	0.940	0.197	0.260	0.280	0.315	0.311	0.320
	2 ¹⁴	0.817	0.824	0.908	0.909	0.926	0.932	0.203	0.219	0.238	0.248	0.306	0.315
	2 ¹⁵	0.893	0.858	0.908	0.909	0.926	0.940	0.206	0.214	0.233	0.305	0.304	0.315
	2 ¹⁶	0.783	0.860	0.907	0.909	0.926	0.940	0.204	0.224	0.233	0.244	0.306	0.316
	2 ¹⁷	0.777	0.860	0.907	0.909	0.926	0.940	0.195	0.220	0.238	0.307	0.305	0.315

The best-performing fingerprint dimension in each column of the table is italicised, bold-faced and marked in red for ease of reference.

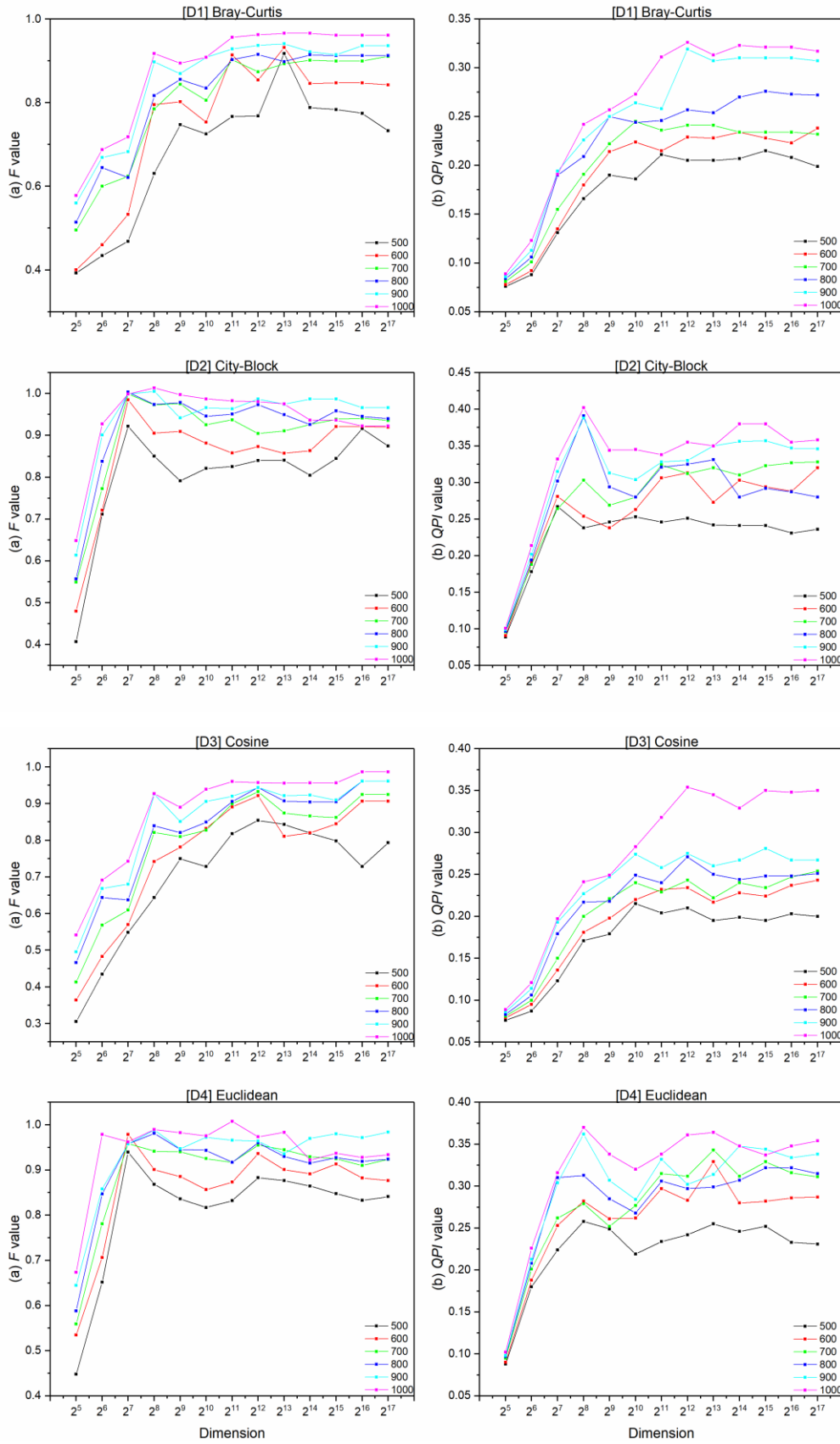


Figure B-3 Effects of dimensionality on Group Average clustering measured by (a) F -measure and (b) QPI -measure for WOMBAT dataset using various distance coefficients (Refer to Table B-3 for detail values)

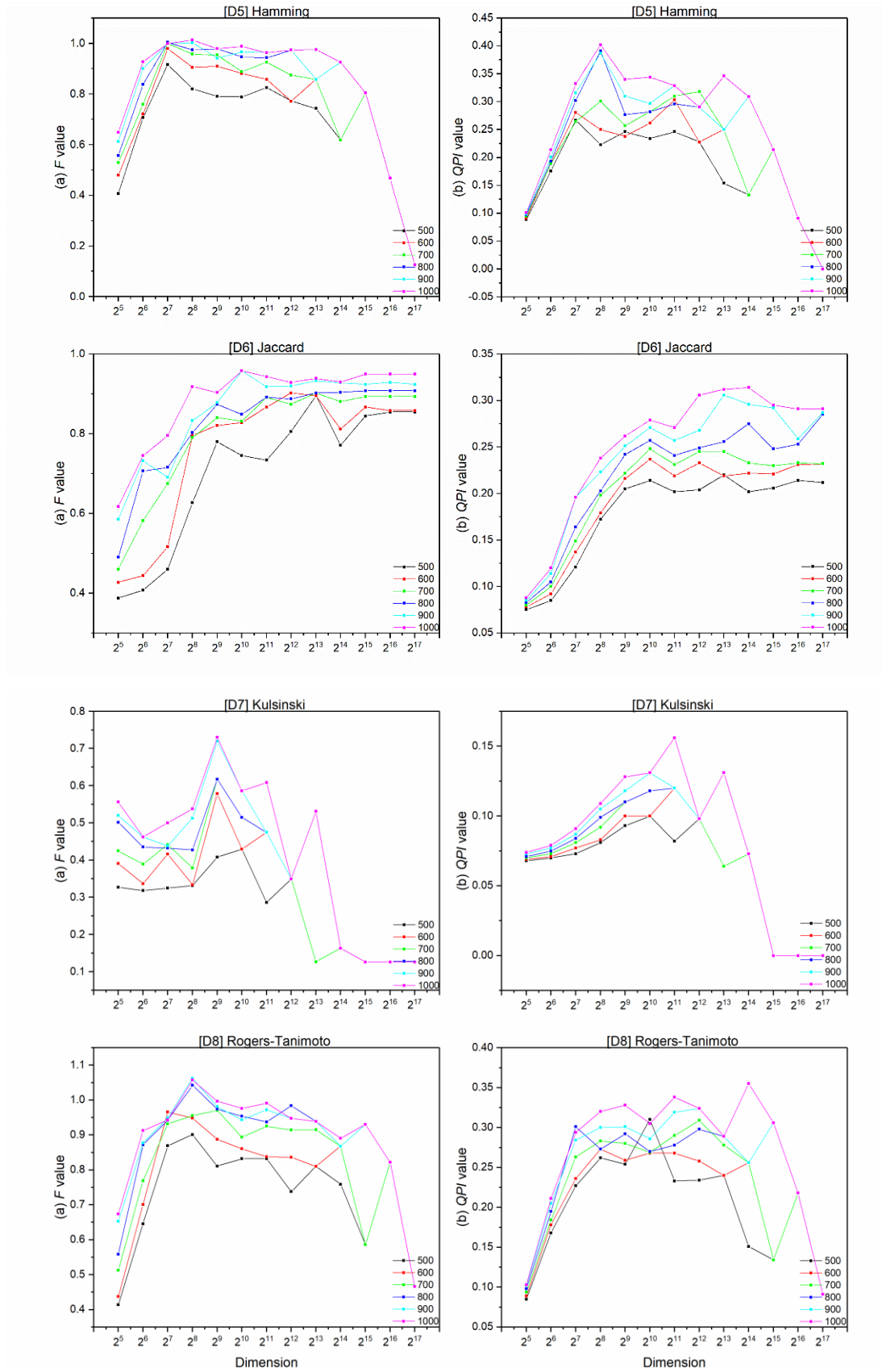


Figure B-3 (continued)

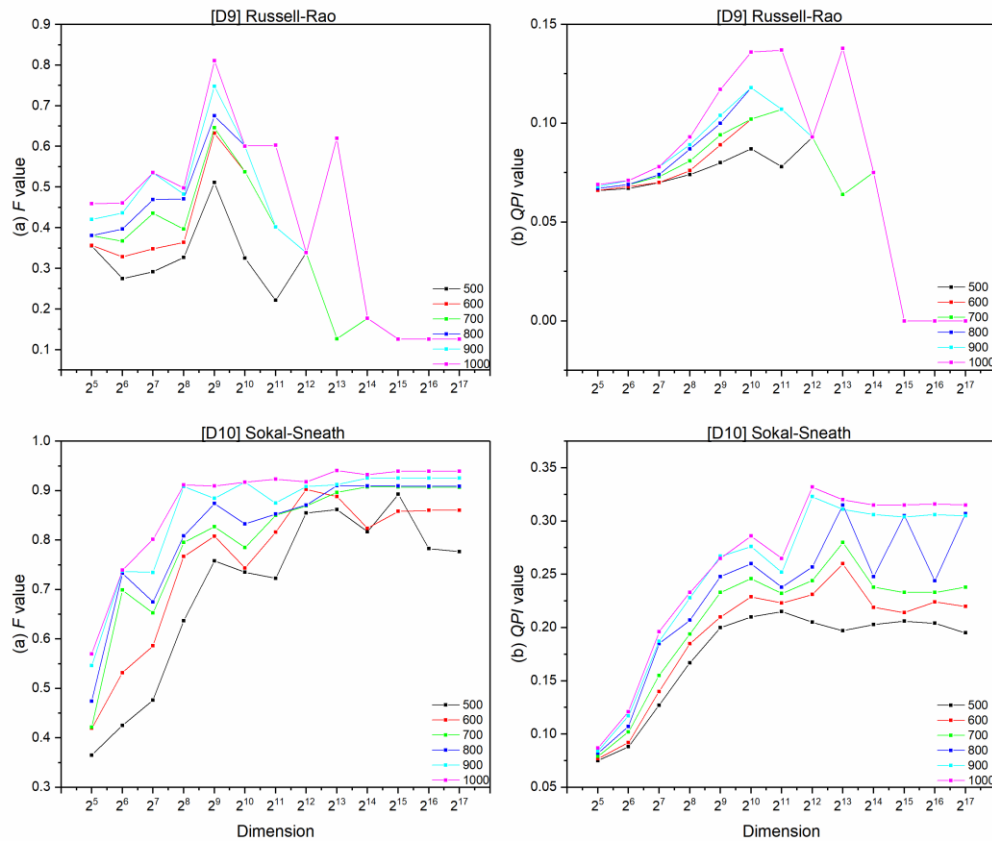


Figure B-3 (continued)

Appendix C Additional Results of Chapter 7

Table C-1 Variance estimation of similarity search components (3 level cross-classified model) for 150 reference structures

Model No.	Intercept (Mean EF)	Effect L3 Variance (Fingerprint)	Effect L2 Variance (Similarity Coefficient)	Effect L1 Variance (Residual Error)
5HT (MPS = 0.34)				
1	4.437	13.365	0.668	0.969
2	5.465	1.729	2.484	1.334
3	1.594	3.117	0.139	0.698
4	4.243	3.286	1.029	1.463
5	3.105	2.390	0.363	0.946
6	7.702	7.574	2.303	2.437
7	7.007	5.408	2.389	1.887
8	9.180	9.165	4.587	3.092
9	6.155	3.582	1.130	4.438
10	2.324	2.299	0.136	0.571
5HT1A (MPS = 0.37)				
11	1.637	0.249	0.033	0.141
12	7.914	9.349	1.391	2.242
13	3.931	7.408	0.762	1.871
14	5.794	9.715	1.973	2.758
15	7.080	22.093	3.138	6.911
16	7.854	21.028	3.435	6.326
17	3.225	5.609	0.598	0.980
18	12.045	39.632	5.434	9.414
19	12.546	50.677	5.781	9.008
20	12.042	51.400	5.649	8.431
5HT3 (MPS = 0.35)				
21	4.403	5.371	0.602	2.634
22	1.776	1.662	0.153	0.407
23	12.831	84.702	3.821	15.112
24	2.674	6.270	0.150	1.249
25	3.579	3.664	0.024	2.342
26	7.221	22.094	1.387	7.777
27	27.357	50.438	9.619	23.631
28	26.556	42.110	10.397	25.137
29	27.782	56.359	14.899	21.284
30	25.515	73.787	8.988	13.962
AChE (MPS = 0.36)				
31	8.322	23.009	1.374	3.979
32	7.933	19.183	1.618	2.741
33	8.884	19.328	2.311	2.855
34	9.524	10.567	0.525	3.505
35	9.521	14.459	1.123	3.209
36	1.079	0.300	0.005	0.163
37	11.836	6.023	0.691	1.543
38	22.586	20.633	10.385	11.419
39	10.106	20.699	2.134	4.118
40	6.760	5.878	1.476	3.314
AT1 (MPS = 0.52)				
41	6.430	42.396	2.355	11.679
42	34.558	87.156	66.798	82.178
43	30.310	111.689	53.158	62.094
44	30.327	108.162	46.212	57.823
45	28.705	137.240	41.022	80.110
46	30.157	104.919	44.730	74.675
47	46.634	114.361	98.825	130.524
48	36.678	95.607	33.328	95.962
49	27.345	162.659	36.485	83.347
50	6.720	27.571	3.036	9.350

The grey box indicates larger variance when compared between the variance estimated for L3 and L2 while the italic and bold faced indicate largest variance when compared between the variance estimated for L3, L2 and the residual error within the same reference compound.

Table C-1 (continued)

Model No.	Intercept (Mean EF)	Effect L3 Variance (Fingerprint)	Effect L2 Variance (Similarity Coefficient)	Effect L1 Variance (Residual Error)
COX (MPS = 0.28)				
51	10.036	10.325	1.299	3.362
52	7.153	5.564	2.593	3.153
53	4.918	11.729	0.113	4.067
54	4.413	3.443	0.317	1.515
55	14.601	7.234	6.561	4.396
56	13.195	24.094	6.503	13.357
57	7.181	5.455	3.648	3.727
58	8.251	14.597	3.822	7.486
59	5.354	16.845	1.373	7.636
60	7.260	12.167	3.219	4.945
D2 (MPS = 0.35)				
61	3.583	5.366	0.372	1.051
62	5.007	6.779	0.539	0.909
63	4.897	7.621	0.482	0.913
64	1.490	0.266	0.006	0.235
65	13.933	42.849	4.272	11.753
66	5.064	6.886	0.554	1.627
67	5.432	8.589	0.493	1.692
68	10.869	22.406	3.082	4.722
69	6.495	37.475	0.468	2.964
70	13.869	32.004	4.890	15.634
FXA (MPS = 0.39)				
71	6.776	3.452	1.251	1.941
72	6.494	2.196	1.052	1.445
73	5.915	6.396	1.315	4.555
74	6.898	8.463	2.421	5.680
75	5.187	2.692	0.974	2.162
76	5.280	2.733	1.276	1.956
77	5.941	5.840	6.072	6.477
78	2.286	0.884	0.913	1.124
79	2.812	1.915	1.237	1.881
80	17.355	7.787	14.204	11.082
HIVP (MPS = 0.43)				
81	18.610	30.666	6.086	14.873
82	9.257	1.925	1.107	1.585
83	8.800	4.822	1.722	2.226
84	20.425	29.896	7.982	12.380
85	20.864	19.862	7.302	12.413
86	20.445	46.468	6.787	20.709
87	20.140	6.458	7.616	11.354
88	12.194	2.981	2.465	2.283
89	16.418	18.318	7.626	9.326
90	16.472	47.030	6.982	16.631
MMP1 (MPS = 0.40)				
91	31.471	127.166	17.645	24.619
92	29.731	117.006	11.234	22.563
93	30.987	94.372	13.847	16.106
94	12.477	4.003	10.833	7.599
95	15.573	54.817	7.320	10.008
96	20.196	147.754	15.280	23.406
97	19.727	35.765	6.438	11.552
98	13.109	11.696	4.842	6.540
99	30.706	137.315	12.970	21.345
100	16.369	47.924	5.037	7.557

The grey box indicates larger variance when compared between the variance estimated for L3 and L2 while the italic and bold faced indicate largest variance when compared between the variance estimated for L3, L2 and the residual error within the same reference compound.

Table C-1 (continued)

Model No.	Intercept (Mean EF)	Effect L3 Variance (Fingerprint)	Effect L2 Variance (Similarity Coefficient)	Effect L1 Variance (Residual Error)
PDE4 (MPS = 0.31)				
101	5.729	37.694	1.112	3.171
102	1.188	1.229	0.031	0.912
103	6.408	27.319	1.662	4.372
104	9.770	38.629	3.287	6.561
105	10.078	45.134	2.995	7.247
106	8.523	48.335	2.325	6.009
107	14.236	77.454	4.414	17.463
108	10.611	18.559	0.346	2.898
109	10.981	20.105	0.954	3.403
110	11.451	25.748	1.256	3.408
PKC (MPS = 0.42)				
111	12.451	1.904	1.367	2.470
112	53.141	193.366	8.106	28.016
113	5.715	26.661	0.738	4.396
114	4.642	8.877	0.342	1.336
115	13.390	1.898	2.861	4.751
116	11.165	1.574	0.802	1.921
117	11.695	1.323	1.119	2.037
118	13.673	0.534	1.386	2.959
119	13.929	0.101	0.680	1.826
120	13.767	0.428	1.093	1.982
Renin (MPS = 0.45)				
121	6.444	8.269	2.352	3.888
122	11.987	8.010	6.077	3.241
123	7.199	7.062	2.959	1.672
124	9.749	20.062	8.930	8.556
125	0.840	0.084	0.157	0.574
126	1.014	0.172	0.073	0.303
127	1.020	0.113	0.423	1.094
128	9.366	15.843	2.785	2.967
129	9.459	2.145	3.732	2.589
130	12.490	41.097	13.101	12.682
SubP (MPS = 0.43)				
131	37.015	258.794	57.147	87.199
132	14.757	49.111	6.762	14.710
133	10.854	25.991	7.177	13.103
134	6.106	4.971	1.275	6.796
135	37.747	225.417	82.373	114.124
136	15.063	76.002	11.654	15.141
137	36.127	336.515	76.101	125.003
138	36.635	232.465	76.079	112.295
139	14.947	128.740	9.467	25.072
140	11.632	29.336	7.215	10.475
Thrombin (MPS = 0.35)				
141	15.477	57.288	9.861	10.508
142	18.917	48.530	19.911	15.043
143	15.496	19.484	12.267	6.558
144	3.092	1.810	0.687	1.498
145	8.578	6.184	3.030	4.882
146	19.844	15.377	11.024	9.058
147	21.456	10.008	12.515	9.481
148	17.983	10.948	10.229	6.131
149	14.071	28.830	9.949	11.604
150	16.754	8.991	8.067	10.047

The grey box indicates larger variance when compared between the variance estimated for L3 and L2 while the italic and bold faced indicate largest variance when compared between the variance estimated for L3, L2 and the residual error within the same reference compound.

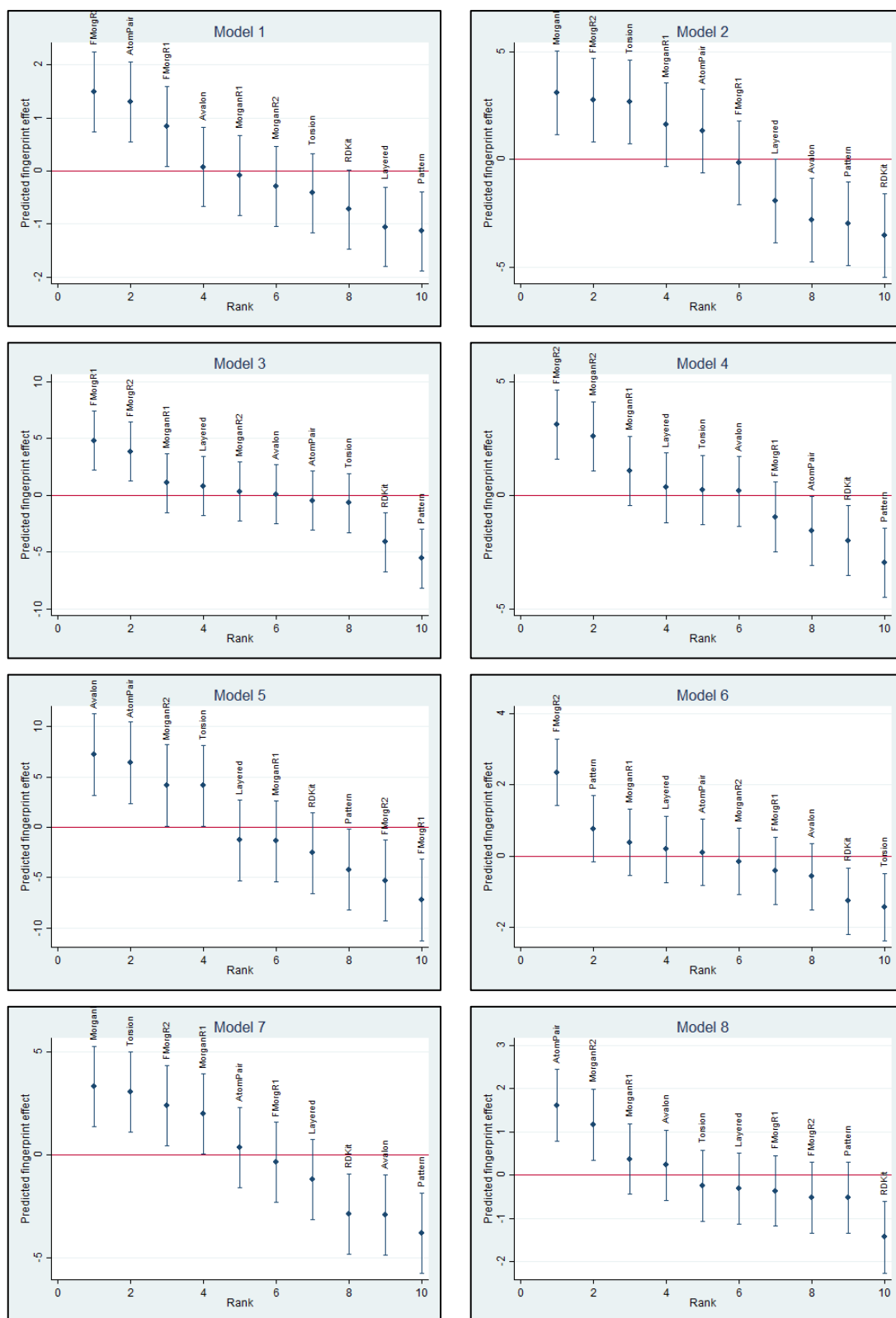


Figure C-1 Caterpillar plot of the fingerprint-level residuals with 95% Bayesian credible intervals for 15 activity classes of ChEMBL dataset

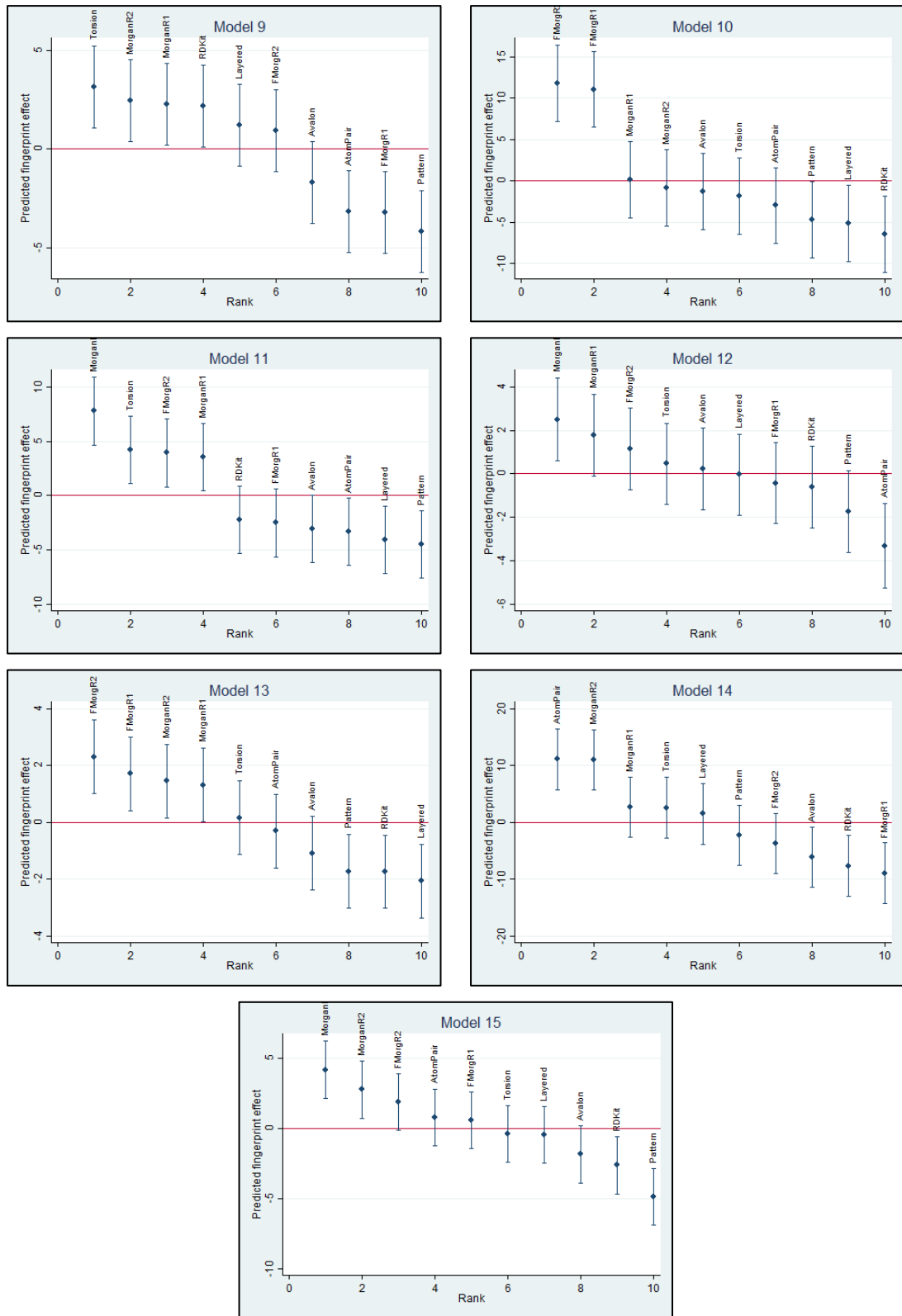


Figure C-1 (continued)

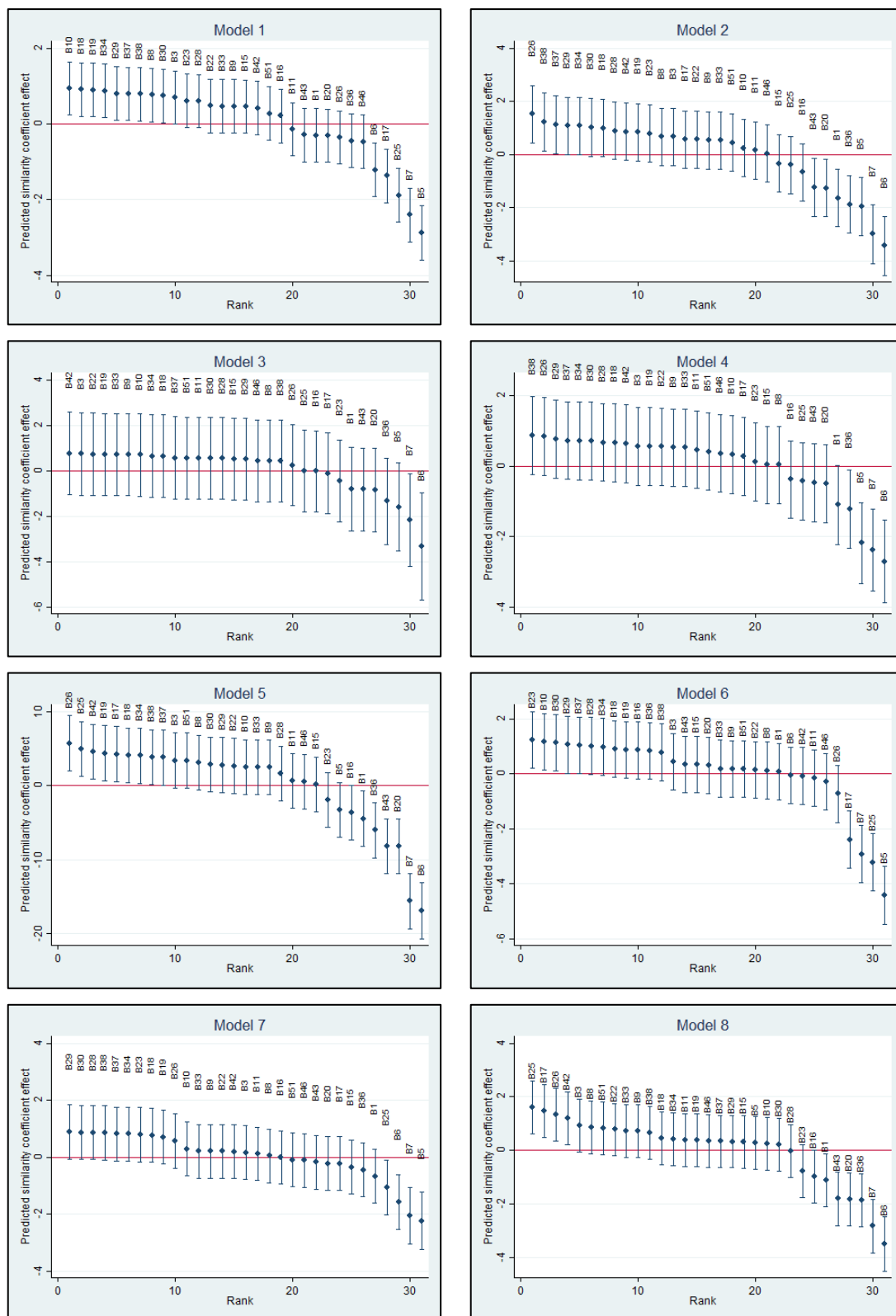


Figure C-2 Caterpillar plots of the similarity coefficient-level residuals with 95% Bayesian credible intervals for 15 activity classes of ChEMBL dataset

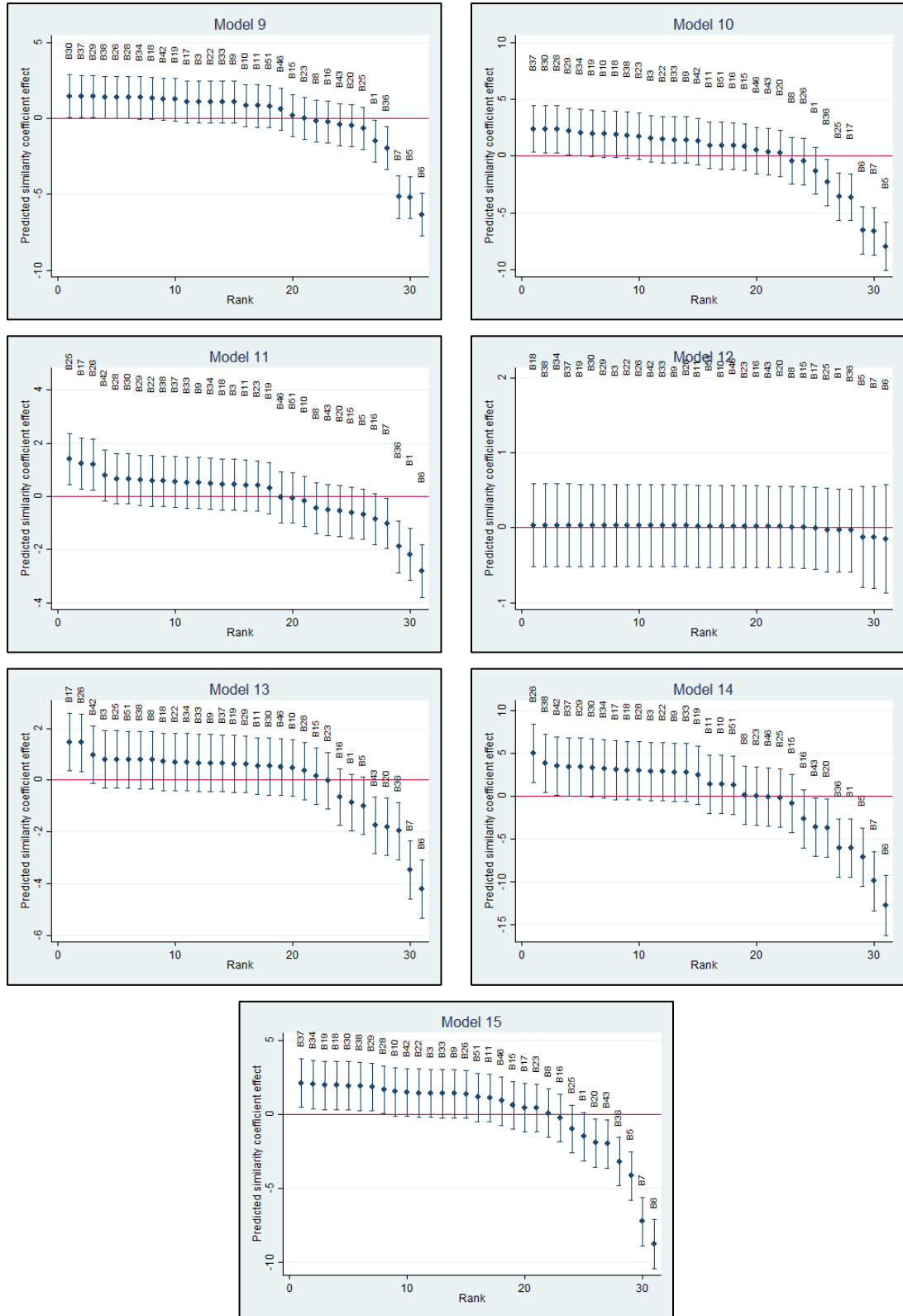


Figure C-2 (continued)