NEW SIMILARITY MEASURES FOR LIGAND-BASED VIRTUAL SCREENING

MUBARAK HUSSEIN IBRAHIM HIMMAT

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

AUGUST 2017

I dedicate this work to my beloved parents, my wife, my brothers, my sisters, and to my lovely sons.

ACKNOWLEDGMENTS

In the Name of Allah, Most Gracious, Most Merciful. First and foremost, all praise and thanks to Allah (SWT), then I would like to extend thanks to the many people, who so generously contributed to the work presented in this thesis.

Firstly, I would like to express my sincere gratitude to my supervisor, Prof .Dr. Naomie Salim, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I greatly appreciate her vast knowledge and skill in many areas of research, she added considerably to my graduate experience; we learned from her how we could acquire knowledge in the better manner and without her assistance and her full supervision this work will never be a reality.

My sincere thanks also go to all members of our research group and especially chemoinformatics group, and many thanks to the University of Technology Malaysia and its community for their support and assistance during research years.

Finally, but by no means least, a gratitude thanks go to my all family especially my parents, my wife, my sons, my sisters and brothers for their patience, encouragement, and support, thank you so much for your Prayers throughout my Ph.D. and my life in general.

Last but not the least, I would like to thank my family: my parents and to my wife, brothers and sister for supporting me throughout my Ph.D. and my life in general.

ABSTRACT

The process of drug discovery using virtual screening techniques relies on "molecular similarity principle" which states that structurally similar molecules tend to have similar physicochemical and biological properties in comparison to other dissimilar molecules. Most of the existing virtual screening methods use similarity measures such as the standard Tanimoto coefficient. However, these conventional similarity measures are inadequate, and their results are not satisfactory to researchers. This research investigated new similarity measures. It developed a novel similarity measure and molecules ranking method to retrieve molecules more efficiently. Firstly, a new similarity measure was derived from existing similarity measures, besides focusing on preferred similarity concepts. Secondly, new similarity measures were developed by reweighting some bit-strings, where features present in the compared molecules, and features not present in both compared molecules were given strong consideration. The final approach investigated ranking methods to develop a substitutional ranking method. The study compared the similarity measures and ranking methods with benchmark coefficients such as Tanimoto, Cosine, Dice, and Simple Matching (SM). The approaches were tested using standard data sets such as MDL Drug Data Report (MDDR), Directory of Useful Decoys (DUD) and Maximum Unbiased Validation (MUV). The overall results of this research showed that the new similarity measures and ranking methods outperformed the conventional industry-standard Tanimoto-based similarity search approach. The similarity measures are thus likely to support lead optimization and lead identification process better than methods based on Tanimoto coefficients.

ABSTRAK

Proses penemuan ubat-ubatan menggunakan teknik pemeriksaan maya bergantung kepada " prinsip keserupaan molekul" yang menyatakan bahawa struktur molekul yang sama cenderung untuk mempunyai ciri-ciri fisiokimia dan biologi yang serupa, berbanding molekul yang lain. Kebanyakan kaedah pemeriksaan maya yang sedia ada menggunakan ukuran keserupaan seperti tahap pekali Tanimoto, tetapi langkah-langkah keserupaan konvensional ini masih tidak mencukupi dan tidak memuaskan hati penyelidik-penyelidik. Kajian ini mengkaji ukuran keserupaan baru yang ditemui dan membangunkan ukuran keserupaan baru serta kaedah penilaian molekul untuk melihat dan mendapatkan semula molekul yang lebih cekap. Pertama, ukuran keserupaan baru telah dibangunkan berdasarkan daripada ukuran keserupaan sedia ada, selain memberi tumpuan kepada konsep keserupaan terpilih. Kedua, ukuran keserupaan baru dibangunkan berdasarkan semakan pemberat pada rentetan-bit, di mana pertimbangan yang tinggi diberikan kepada ciri-ciri yang terdapat dalam kedua-dua molekul yang dibandingkan, dan ciri-ciri yang tidak terdapat dalam kedua-dua molekul dibandingkan. Pendekatan akhir mengkaji kaedah penilaian bagi membangunkan kaedah penilaian pengganti. Kajian ini membandingkan ukuran keserupaan dan kaedah penilaian dengan pekali penanda aras seperti Tanimoto, Cosine, Dice, dan Pemadanan Mudah (SM). Pendekatan ini menggunakan data ujian piawai seperti Laporan Data Ubat MDL (MDDR), Direktori Umpan Berguna (DUD), dan Pengesahan Saksama Maksimum (MUV). Keputusan keseluruhan kajian menunjukkan bahawa langkah-langkah persamaan yang dicadangkan dan kaedah penilaian mengatasi persamaan konvensional piawai industri yang berasaskan pendekatan Tanimoto. Persamaan yang dicadangkan dijangka dapat menyokong proses pengenalpastian dan pengoptimuman pendahulu ubatan dengan lebih baik berbanding kaedah berasaskan persamaan Tanimoto.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1D	-	One Dimension
2D	-	Tow Dimension
3D	-	Three Dimension
AIM	-	Atoms-in-Molecules
AM	-	Adjacency Matrix
ASMTP	-	Adapted Similarity Measure of Text Processing
AUC	-	Area Under the Curve
BEDROC	-	Boltzmann enhanced discrimination of receiver operating characteristic
CML	-	Chemical Markup Language
DUD	-	Directory of Useful Decoys
ECFC	-	Atom Type Extended-Connectivity Fingerprint
EEFC	-	Atom Type Atom Environment Fingerprint
EHFC	-	Atom Type Hashed Atom Environment Fingerprint
FCFC	-	Functional Class Extended-Connectivity Fingerprint
FEFC	-	Functional Class Atom Environment Fingerprint
FHFC	-	Functional Class Hashed Atom Environment Fingerprint
HTS	-	High-throughput Screening
IPRP	-	Interactive Probability Ranking Principle
k-NN	-	K-Nearest Neighbors
LBVS	-	Ligand-Based Virtual Screening

MCS	-	Maximal Common Substructure
MDDR	-	MDL Drug Data Report
MDL	-	Molecular Design Limited
MMR		Maximal Marginal Relevance
MUV	-	Maximum Unbiased Validation
PRP	-	Probability Ranking Principle
PT	-	Portfolio Theory
QBE	-	Query By Example
QSAR	-	Quantitative Structure-Activity Relationship
ROC	-	Receiver Operating Characteristic
SBVS	-	Structure-Based Virtual Screening
SMILES	-	Simplified Molecular Input Line System
SMTP	-	Similarity Measure for Text Classification and Clustering
SQB	-	Standard Quantum Based
SQL	-	Structured Query Language
SVM	-	Support Vector Machines
TAN	-	Tanimoto
TSS	-	Turbo Similarity Searching
VS	-	Virtual Screening
WLN	-	Wiswesser Line Notation

CHAPTER 1

INTRODUCTION

1.1 Introduction

In chemical and pharmaceutical research, computers have been used for many years to decrease the cost of drug discovery (Todeschini and Consonni, 2009). Many different computer techniques and methods have been applied, and the data mining methods and information retrieval methods have been widely used in chemical, biomedical, and other medical fields. The actual laboratory drug discovery process can take between 12 and 15 years and can cost approximately more than one million dollars (Rollinger *et al.,* 2008); for that, considerable effort has been made to cover research into this area. This has taken years and cost in excess of $1 billion. It is complex and costly and consumes a lot of time in laboratory experiments. These two above-mentioned reasons have attracted the attention of researchers in different aspects to solve and reduce the long drug discovery time and its high cost. One of the rich science areas within the last decades is chemoinformatics, which is a multi-disciplinary area that combines many older different disciplines such as computational chemistry, chemometrics and Quantitative Structure–Activity Relationship (QSAR). The term chemoinformatics has some synonyms in literature, as it is also known as Chemical Informatics and Chemical Information. Its general definition is "the use of computer and informational techniques applied to a range of problems in the field of chemistry" (Brown, 1998). Another definition is "The mixing of different information resources for the purpose of transforming 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" (Brown, 1998). Another general definition was given newly by Gasteiger

(2016)“chemoinformatics is the use of informatics methods to solve chemical problems”.

The process of discovering new drugs using computational screening methods is being continuously developed, and improved as it is one of the most important tools for drug discovery. Virtual screening now becomes an alternative to High-throughput Screening (HTS). HTS was considered the basic and main method for drug candidate development, but virtual screening (VS) with its various techniques and search methods is becoming a reliable method for drug discovery.

Virtual screening methods can be used in many aspects of chemistry, such as molecule ranking, clustering, docking and virtual screening; as a result, this is now used as a complementary tool to HTS in drug discovery, because the rational drug discovery requires fast and computationally straightforward methods that distinguish active ligands from inactive molecules in huge molecular databases. Huge databases can be screened easily and successfully in a short time. VS, or screening as described here, is the process of selecting molecules to help in bioactivity testing. This screening is applied automatically by computer methods that select molecules; this is generally referred to as VS, and the Ligand-based virtual screening extrapolates from known active compounds used as input information and aims at identify structurally diverse compounds having similar bioactivity, regardless of the methods that are applied.

The screening methods conducted by computers are employed to rank the molecules according to their structures and put the most promising structures at the top of the list(Brown, 1998; Chen and Reynolds, 2002); this gives a high ranking to those molecules with structures that may be similar to structures that have already been tested. The screening methods and concept of molecular similarity are closely related to those used in information retrieval. Researchers have found most of the existing ligand–based similarity methods and similarity measures to be unsatisfactory, and consider the Tanimoto as the better similarity measure (Dávid Bajusz, 2015). However, some new similarity measures for information retrival

have recently been proposed (Lin *et al.*, 2014; Todeschini *et al.*, 2012) as well as some proposed for virtual screening that outperformed the Tanimoto, refuting the claim that only Tanimoto could achieve better results (Al-Dabbagh *et al.*, 2015).

Our general hypothesis for this work is that although considerable enhancements could be achieved in ligand-based virtual screening, more effort needs to be provided to help accelerate the drug discovery process and some of its major pitfalls and challenges that still need to be solved in order to handle the exponentially increased volume of molecule data(Cereto-Massagué *et al.*, 2014; Muegge and Mukherjee, 2016). As mentioned above, the general belief is that the Tanimoto similarity measure is the best similarity measure for virtual screening in spite of many similarity measures that have been proposed and applied in other aspects of science. This belief has led researchers to ignore the recently proposed similarity measures, and at the same time reduce the determination of researchers in cheminformatics to use and modify the similarity measures that could outperform the existing similarity measures for virtual screening.

This thesis, primarily focus on ligand-based virtual screening. Different algorithms are proposed based on bit-strings and fragment-based that enhanced ligand-based virtual screenings. The rest of this chapter discusses the background of the problem, the importance of the study, the objectives and scope of this research. The last section will describe the organization and outline of the thesis**.**

1.2 Problem Background

Great efforts have been made to provide new drugs to the market, and there are considerable investments in the research regarding this issue. The development of a new drug consumes very long timeframes and high cost as mentioned earlier in this chapter. In chemoinformatics, researchers try to help the industry and chemists to make the drug discovery process less risky and less costly and accelerate the processing time, which takes years(DiMasi *et al.*, 2016; Wang *et al.*, 2016). Virtual

screening provides many tools and methods to provide considerable influence in drug discovery and in the process of obtaining a drug candidate. Recently, many new techniques have been proposed in chemoinformatics to be used as a substitute for old, traditional, synthesized laboratories testing a New Chemical Entity (NCE) approaches, high-throughput screening (HTS), combinatorial chemistry (CC)(Li *et al.*, 2016). With HTS screening, millions of chemical, pharmacological, or genetic tests could be conducted in a short time by using computer aids that could execute a million processes in a few seconds. Although there is no doubt that considerable progress has been made in the field of computational drug discovery and ligand prediction(Chen *et al.*, 2016; De Vivo *et al.*, 2016), the commonly used methodology is still far from perfect, and it needs more work to satisfy chemists. According to some studies, the estimated time to produce a new drug to the market is twelve years, at an estimated cost ranging from US$92 million to US$ 883 million (DiMasi *et al.*, 2016; Morgan *et al.*, 2011). Differences in methods, data sources, and timeframes explain some of the variation in estimates. As a result, the focus of most researchers in cheminformatics is twofold: reducing the cost and time of drug discovery process, and avoiding the failure rates in later stages of drug development. Hence, the time and cost of finding and testing new chemical entities can be considered the main objective in drug discovery. For virtual screening, researchers strive for ways to find new active compounds and to bring these compounds to the market as quickly as possible.

The huge chemical compound libraries provide a good source of new potential drugs that can be randomly or methodically tested or screened to find good drug compounds. It is now possible to test hundreds of thousands of compounds in a short time using high-throughput screening techniques. Therefore, virtual chemical libraries that are done by computer systems become useful supporters that aid this process of drug discovery (Xu and Hagler, 2002).

Chemists have always struggled with the difficult problem of deciding which chemical structures to synthesize among large numbers of compounds. However, this is still a small percentage of the total number that could be synthesized. Therefore, in recent years the techniques of chemical search have been

called virtual screening, which encompasses a variety of computational techniques that are used to test a large number of compounds by computer instead of experience(Bajorath, 2013; Muegge and Mukherjee, 2015; Stumpfe and Bajorath, 2011; Walters *et al.*, 1998). These computational methods can be used for searching chemical libraries to filter out the unwanted chemical compounds, and these methods allow chemists to reduce a huge virtual library, and make it more manageable size to assess the probability that each molecule will exhibit the same activities against a specific biological target. The approaches of virtual screening can be categorized into structure-based virtual screening (SBVS) approaches(Ono *et al.*, 2014; Vuorinen *et al.*, 2014), and ligand-based virtual screening (LBVS) approaches. The SBVS approaches can be used when the 3D structure of the biological target is available, such as ligand-protein docking and de novo design. The LBVS approaches are applicable in the case of absence of such structural information, such as machine learning methods and similarity methods.

The similarity methods may be the simplest and most widely-used tools for LBVS of chemical databases(Cereto-Massagué *et al.*, 2015a; Willett, 2009; Willett *et al.*, 1998).The increased importance of similarity searching applications is mainly due to its role in lead optimization in drug discovery programs, where the nearest neighbors for an initial lead compound are sought in order to find better compounds. There are many studies in the literature associated with the measurement of molecular similarity (Bender and Glen, 2004; Maldonado *et al.*, 2006; Nikolova and Jaworska, 2003).Similarity searching aims to search and scan chemical databases to identify those molecules that are most similar to a user-defined reference structure using some quantitative measures of intermolecular structural similarity. However, the most common approaches are based on 2D fingerprints, with the similarity between a reference structure and a database structure computed using association coefficients such as the Tanimoto coefficient (Dávid Bajusz, 2015; Deng *et al.*, 2015; Johnson and Maggiora, 1990; Todeschini *et al.*, 2012). The similarity measures methods play a significant role in detecting the rate for pairwise molecular similarity(Lynch and Ritland, 1999). These methods can be employed to find the most similar molecules among thousands of compounds, and then organize these similar molecules in decreasing order depending on the probability ranking

principle that only relies on the values of probability between the molecules and molecular target.

In general, the processes of a similarity measure for molecules have two stages, which are similarity stage and ranking stage. At similarity level, the performance of conventional similarity methods has been enhanced in various ways. Some studies have used the weighting scheme (Abdo and Salim, 2010; Ahmed *et al.*, 2012; Jaghoori *et al.*, 2015; Kar and Roy, 2013; Klinger and Austin, 2006), while others have employed the techniques of data fusion (Ahmed *et al.*, 2014; Salim *et al.*, 2003; Willett, 2013b). The relevant feedback has also been applied and used in LBVS to improve the performance of similarity methods (Abdo *et al.*, 2012; Abdo *et al.*, 2011). However, the effectiveness of any similarity method has been found to vary greatly from one biological activity to another in a way that is difficult to predict (Gasteiger, 2016; Sheridan and Kearsley, 2002). In addition, the use of any two methods has been found to retrieve different subsets of actives from the chemical library, so it is advisable to utilize several search methods where possible.

Considerable effort has been expended in finding the appropriate similarity measures in virtual screening among such available of choices of similarity measures, and this has attracted the attention of researchers from the early time of High Throughput Screening, and cheminformatics.

Many similarity measures have been applied in cheminformatics for virtual screening. These similarity measures have contributed in screening performance. Some other similarity measures have been adapted and derived from existing similarity measures and achieved good results in other areas, but haven't been applied in virtual screening. In addition, many similarity measures have been proposed for text (Lin *et al.*, 2014), and could be adapted for virtual screening due to many similar aspects between the text and chemical information retrieval. Thus, the algorithms that have been applied in text information retrieval can also be applied in chemical information retrieval (Obaid *et al.*, 2017; Willett, 2000a).

The fragment bases and bit-strings similarity method has gained attention from researchers in chemoinformatics and especially in virtual screening (Abdo and Salim, 2010; Ahmed *et al.*, 2012; Chen and Reynolds, 2002; Holliday *et al.*, 2002; Zoete *et al.*, 2009) , and many types of research are focused on it. The molecules databases (fingerprint) contain a large number of bit-strings that represent the molecules features (Bajorath, 2017; from Structure, 1997; Todeschini and Consonni, 2009; Todeschini *et al.*, 1994), and considering all these features as the same and giving them same weight features in similarity calculations is not fair. This is because most proposed methods usually assume that all molecular features are equal in importance. On the other hand, all weighting schemes calculate the weight for each feature independently with no relation to all other features, in general, The summarization of the all mentioned problem background are demonstrated in Table 1.1. For all these mentioned cases, in order to enhance the virtual screening effectiveness, feature reweighting using important bit-strings calculations can enhance the recall of similarity measure.

In order to enhance the effectiveness of the similarity measure, the primary aim of this research is to propose ligand-based similarity methods, and propose a ranking method based on bit-strings and fragment-based reweighting. Additional aims include adapting an existing similarity measure, adapting text similarity measure and proposing alternative ranking method to be used for ligand-based virtual screening.

Table 1.1: Summarization of problem background

Issue	What have been done in LR	Why not enough	Proposed method
Similarity Method Computational methods can be used for searching chemical libraries to filter out the unwanted chemical compounds, to reduce the cost and the time in drug discovery programs**.**	**Enhancement of similarity measures using:** Similarity coefficients (Consonni and Todeschini, 2012; Dávid Bajusz, 2015; Lin *et al.*, 2014; Rognan and Bonnet, 2014; Todeschini *et al.*, 2012). • Data Fusion (Chen *et al.*, 2010; Sastry *et al.*, 2013; Willett, 2013a). • Relevance feedback (Abdo *et al.*, 2012; Agarwal *et al.*, 2010; Chen *et al.*, 2009b). Weighting functions (Ahmed *et al.*, 2012; Arif *et al.*, 2010; Holliday *et al.*, 2013). Machine Learning (Cereto-Massagué *et al.*, 2015b; Durrant and Amaro, 2015; H Haga and Ichikawa, 2016; Lavecchia, 2015).	Although several similarity coefficients and techniques have been applied to enhance VS,but the area of VS still requires more investigation to determine whether other coefficients might yield a higher level of screening effectiveness than those which been used for virtual screening .	Enhance the effectiveness of the Ligand-based similarity searching method by adapting several similarity measures from information retrieval field. ***Adapted Similarity Measure of Text Processing (ASMTP)***
Fragment Reweighting The retrieval performance of the LBVS methods was observed to be improved significantly when chemical fragment weightings were used.	**Finding new weighting schemes or functions** (Abdo and Salim, 2010; Arif *et al.*, 2010; Holliday *et al.*, 2013).	There are other weighting features methods need to be investigating to assign more weights to the bit-strings for improving the effectiveness of LBVS.	Enhance the effectiveness of similarity measure by reweighting molecular bit strings. ***Adapted Simple Matching Similarity Coefficient (ASMSC),***
Molecular Ranking Principle Rank the active chemical compounds at higher ranking position than inactive ones. The most popular technique is probability ranking principle (PRP) has been used for molecular ranking can prioritize the molecules in decreasing order of value to the user's reference relying on the probability value of molecules.	• Enhancement of PRP (Text IR & Chemical IR) • Classification methods (Dörr 2015, Chen 2014, Rathke 2010) • Regression methods (Li 2011, Hasegawa 2010) • Data Fusion (Willet, 2013) • Alternative ranking approaches (Text IR) • QPRP Quantum probability ranking principle (Zuccon, 2012) • **IPRP Interactive Probability Ranking** Principle (Sheridan ,2008)	One of the key controversial issues of PRP is the independence among ranking compounds, which prevents molecule's ranking position from the effect of other molecules.	Enhance the effectiveness of similarity measure by using Maximal Marginal Relevance (MMR) ranking principle of molecules that is inspired from text and document retrieval domain. ***Maximal Marginal Relevance (MMR) for LBVS***

1.3 Problem Statement

In general, the aim of virtual screening is developing new drugs, in addition, its significance is to decrease the consumption of times and cost which is considered a big challenge in drug discovery process, where the estimated cost of drug discovery exceeding millions and years to discover new drugs, and virtual screening reduce this cost to be very low compared to conducting experiments in real laboratory screening.

By understanding the problem background that has been discussed in the previous section, it can be concluded that the needs of many chemical similarity search methods is considered one of the continuing challenges in cheminformatics (Sheridan and Kearsley, 2002) ,the ligand-based virtual similarity methods have been under development for decades, and the ligand-based virtual screening field still needs more investigation. In addition, in coming up with a new proposed similarity measure and a similar information retrieval field for improvement, there are limitations of the currently used similarity measures.

The aim of this study research is to develop a ligand-based similarity method based on developing algorithms that emphasize the common structural features (bit-strings) and give high priority in similarity calculations, and reweighting some bit-strings when conducting the search on chemical databases to retrieve the active compounds with the most similar biological activity to the specific reference structure.

Recently, many studies in text information retrieval have proved that retrieval models are based on some new similarity measures and have provided significant improvements in retrieval performance compared to conventional models, and this could be adapted for ligand-based virtual screening.

The all developed similarity methods as well as benchmark similarity coefficients have used the classical ranking approach when ranking the chemical structures, and this study will also investigate the most popular common ranking methods used in information retrieval and propose an alternative method to conventional probability ranking principle (PRP) (Robertson, 1977).

The proposed algorithms apply different approaches to fingerprint data fragment reweighting; this approach is based on fragment reweighting factors. Fragment reweighting here is the process of adding some constant weight to the original weight in order to improve retrieval performance in information retrieval systems. This approach has been derived from document retrieval filed.

The core of virtual screening is to develop anew drugs that decrease the consumption of times and cost .will help in development of representation of time spent on the virtual screening experiments is not taken as a big issue when it has been compared to the high cost and long duration of screening of molecules in a real laboratory. For that this research does not concern the time of virtual screening as an important factor.

1.4 Research Questions

Referring to the problem background, th**e** main questions of this research are:

- Can some similarity measures from document retrieval be adapted to improve ligand-based virtual screening?

- How can new similarity matrices be developed for virtual screening using some preferred similarity measure properties used in document retrieval areas?

- Can the ligand-based virtual screening performance be improved by reweighting some bit-strings of the features?

- How can other ranking method be proposed to improve the effectiveness of virtual screening?

1.5 Research Objectives

The main goal of this research is to develop a similarity-based virtual screening approach using reweighted fragments or the bit-strings, with the ability to improve the retrieval effectiveness and provide an alternative to existing tools for ligand-based virtual screening. Therefore, our general hypothesis for this study is How could constructing and adapting similarity measures and ranking methods from document retrieval can help improve the retrieval performance of molecular similarity? To achieve this goal, the following objectives have been set:

- To investigate some molecule features (bit-strings) to be reweighting for enhance retrieval effectiveness of VS.
- To formulate and adapt new similarity metric for ligand-based virtual screening. Virtual screening.
- To formulate a similarity-based virtual screening method for molecular similarity searching based on text and document retrieval similarity measure concepts.
- To formulate and develop alternative ranking method for ligand-based virtual screening instead of conventional probability ranking principle (PRP).

1.6 Importance of the Study

This study introduced some ligand-based virtual screening algorithms that incorporate adaptation and modification of some similarity measures in order to enhance the efficiency of ligand-based virtual screening. It is also suggested an alternative ranking method that could outperform probability-ranking principle (PRP), which is considered the most popular ranking theory for current similarity searching methods in LBVS. The study rely on the believe that some modification of the existing methods could provide valuable enhancement.

1.7 Scope of the Study

This study will focus on ligand-based virtual screening, especially on similarity-based virtual screening using 2D fingerprint representations of molecular structure. The 2D fingerprint is a vector that encodes the presence and absence of the topological structure that represents the typical atoms, bond, or ring-canter fragment. The proposed screening methods mentioned before will be used to quantify the degree of structural resemblance between a pair of molecules characterized by 2D fingerprints. Most methods are applied with both binary and non-binary 2D fingerprints descriptors. The study focuses on the fragment, bit-string and reweighting methods and similarity coefficients and ranking methods to present an enhancement of molecular retrieval. The bit-strings emphasize the common structural features (bit-strings) and give high priority in similarity calculations. The reweighting factor here will take some similarity concepts to reweight some bit-string values.

The proposed virtual screening enhancement solutions in this study have been evaluated by simulated virtual screening experiments that were conducted on large benchmark datasets which have been derived from MDL Drug Data Report (MDDR) database ("Symyx Technologies. MDL drug data report: Sci Tegic Accelrys Inc., the

MDL Drug Data Report (MDDR). Database is available at http://www.accelrys.com/,"), Maximum Unbiased Validation (MUV) (Rohrer and Baumann, 2009;), the MDL Drug Data Report (MDDR) and the Directory of Useful Decoys (DUD) (Rohrer and Baumann, 2009) where single and multiple reference structures are available. The performance of this method is evaluated against the performance of conventional 2D similarity measure Tanimoto.

1.8 Thesis Outline

This section describes the organization of the thesis. There are seven chapters in this thesis, which are:

Chapter 1, *Introduction*: this chapter gives a general introduction to chemoinformatics, drug discovery, and virtual screening topic of the proposed research work. There are brief overviews of some of the issues concerning the virtual screening research area, and it briefly discusses the following topics: problem background, the problem statement, objectives of the study, research scope, and significance of the study.

Chapter 2, *Molecular representations and Similarity concepts*: this chapter begins with an overview of computer representations of chemical structures and various types of searching mechanisms offered by chemical information systems. In the third section, we present molecular representations that can be employed for molecular similarity searching as well as for molecular analysis and clustering. The chapter describes in detail the 2D fingerprint-based similarity methods and different types of similarity coefficients. The chapter also briefly discusses the implementation of machine learning techniques to molecular similarity and similarity measures of text and document areas. At the end of the chapter there is a conclusion that summarizes the applicability of the discussed methods to molecular similarity searching and the best ways to improve the performance of these methods.

Chapter 3, *Research Methodology*: this chapter describes the overall methodology adopted in this research to achieve the objectives of this thesis; it presents the methodology used in this research. A methodology is generally a guideline for solving a research problem. It contains the generic framework of the research and the steps required to carry out the research systematically, and it discusses in detail the datasets that will be used to conduct the experiments of the proposed methods. This includes discussion on the research components such as the phases, techniques, and tools involved. At the end of the chapter, we will conclude with a summary.

Chapter 4, *Enhancing Ligand-based Virtual Screening Using Bit-strings Reweighting*: this chapter introduces the new ligand-based virtual screening ranking algorithm, called Adapted Simple Matching Similarity Coefficient (ASMSC) that emphasizes the common molecular structural features (bit-strings) to be given a high priority in similarity calculations. The chapter describes the construction of the algorithm and experiments done to evaluate the proposed coefficient. In the results and discussion section, the results are presented and discussed.

Chapter 5, *constructing new similarity metric and Adapting Document Similarity Measures for Ligand-based Virtual Screening*: the study investigates the newly documented similarity measure and adapts it for ligand-based virtual screening. The adapted SMTP algorithm focuses on the preferred selected similarity properties. In the results and discussion section, all experiments conducted on different datasets are discussed, and the chapter also discusses comparison of the achieved results with the standard coefficient of VS, and discusses the investigation of the effectiveness of proposed adapted similarity measure. At the end of the chapter we will conclude with a summary.

Chapter 6, *Using Maximal Marginal Relevance in Ligand-based Virtual Screening:* the chapter investigates the susceptibility of using the concepts of MMR in order to enhance the efficiency of ligand-based virtual screening. We will examine the use of MMR with different datasets to investigate its capability to

improve virtual screening. The chapter discusses some ranking methods that have been applied in information retrieval, and it covers a comparison of the achieved results with the standard coefficient of VS. It also discusses the investigation of the effectiveness of proposed adapted similarity ranking. At the end of the chapter, we will conclude with a summary.

Chapter 7, *Conclusion and Future Work:* this is the last chapter, and it provides a conclusion of the overall work of this thesis. It highlights the findings and contribution made by this study and provides suggestions and recommendations for future research.

1.9 Summary

In this chapter, we give a broad overview of the problems involved in the molecular similarity. This chapter serves as an introduction to the research problem set out earlier in this thesis. The goal, objectives, the scope and the outline of this thesis are also presented.

Baralis, Elena, Cagliero, Luca, Jabeen, Saima, Fiori, Alessandro, & Shah, Sajid. (2013). Multi-document summarization based on the Yago ontology. *Expert Systems with Applications, 40*(17), 6976-6984.

Barnard, John M. (1993). Substructure searching methods: old and new. *Journal of Chemical Information and Computer Sciences, 33*(4), 532-538.

Baroni-Urbani, Cesare, & Buser, Mauro W. (1976). Similarity of binary data. *Systematic Biology, 25*(3), 251-259.

Baulieu, FB. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification, 6*(1), 233-246.

Bender, Andreas. (2011). Bayesian methods in virtual screening and chemical biology *Chemoinformatics and Computational Chemical Biology* .pp. 175-196.

Bender, Andreas, & Glen, Robert C. (2004). Molecular similarity: a key technique in molecular informatics. *Organic & Biomolecular Chemistry, 2*(22), 3204-3218.

Bender, Andreas, Mussa, Hamse Y, Glen, Robert C, & Reiling, Stephan. (2004). Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *Journal of chemical information and computer sciences, 44*(5), 1708-1718.

Brown, Frank K. (1998). Chemoinformatics: what is it and how does it impact drug discovery. *Annual reports in medicinal chemistry, 33*, 375-384.

Brown, RD, Bures, MG, & Martin, YC. (1995). Similarity and cluster-analysis applied to molecular diversity. *AMERICAN CHEMICAL SOCIETY*, Vol. 209, pp. 3-COMP.

Brown, Robert D, & Martin, Yvonne 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.

Carbó, R., Besalú, E., Amat, Ll, & Fradera, X. (1995). Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationships (QSPR). *Journal of Mathematical Chemistry, 18*(2), 237-246.

Carbonell, Jaime, & Goldstein, Jade. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. *The 21st annual international ACM SIGIR conference on Research and development in information retrieval.Augest 01-02,* 1998. Carnegie Mellon University.1998. 335-336

Carhart, Raymond E, Smith, Dennis H, & Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences, 25*(2), 64-73.

Carpineto, C. and G. Romano (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)* 44(1): 1.

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. *Methods, 71,* pp.58-63.

Cereto-Massagué, Adrià, Ojeda, María José, Valls, Cristina, Mulero, Miquel, Garcia-Vallvé, Santiago, & Pujadas, Gerard. (2014). Molecular fingerprint similarity search in virtual screening. *Methods*. *71*, 56-68.

Cereto-Massagué, Adrià, Ojeda, María José, Valls, Cristina, Mulero, Miquel, Garcia-Vallvé, Santiago, & Pujadas, Gerard. (2015). Molecular fingerprint similarity search in virtual screening. *Methods, 71*, 58-63.

Ceroni, Alessio, Costa, Fabrizio, & Frasconi, Paolo. (2007). Classification of small molecules by two-and three-dimensional decomposition kernels. *Bioinformatics, 23*(16), 2038-2045.

Charifson, Paul S, Corkery, Joseph J, Murcko, Mark A, & Walters, W Patrick. (1999). Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of medicinal chemistry, 42*(25), 5100-5109.

Chen, Beining, Mueller, Christoph, & Willett, Peter. (2009). Evaluation of a Bayesian inference network for ligand-based virtual screening. *Journal of cheminformatics, 1*(1), 1-10.

Chen, Beining, Mueller, Christoph, & Willett, Peter. (2010). Combination Rules for Group Fusion in Similarity-Based Virtual Screening. *Molecular Informatics, 29*(6-7), 533-541.

Chen, Can, Wang, Ting, Wu, Fengbo, Huang, Wei, He, Gu, Ouyang, Liang, . . . Jiang, Qinglin. (2014). Combining structure-based pharmacophore modeling, virtual screening, and in silico ADMET analysis to discover novel tetrahydro-quinoline based pyruvate kinase isozyme M2 activators with antitumor activity. *Drug design, development and therapy, 8*, 1195.

Chen, Harr, & Karger, David R. (2006). *Less is more: probabilistic models for retrieving fewer relevant documents*. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, 2006 Aug 6. ACM.Washington, USA. (pp. 429-436).

Chen, Hong, Peng, Jiangtao, Zhou, Yicong, Li, Luoqing, & Pan, Zhibin. (2014). Extreme learning machine for ranking: Generalization analysis and applications. *Neural Networks, 53*, 119-126.

Chen, Jenny, Holliday, John, & Bradshaw, John. (2009). A machine learning approach to weighting schemes in the data fusion of similarity coefficients. *Journal of chemical information and modeling, 49*(2), 185-194.

Chen, Xin, & Reynolds, Charles H. (2002). Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *Journal of chemical information and computer sciences, 42*(6), 1407-1414.

Chen, Xing, Yan, Chenggang Clarence, Zhang, Xiaotian, Zhang, Xu, Dai, Feng, Yin, Jian, & Zhang, Yongdong. (2016). Drug–target interaction prediction: databases, web servers and computational models. *Briefings in bioinformatics, 17*(4), 696-712.

Choi, Seung-Seok, Cha, Sung-Hyuk, & Tappert, Charles C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics, 8*(1), 43-48.

Chowdhury, Gobinda. (2010). *Introduction to modern information retrieval*: London WCIE 7AE .Facet publishing.

Consonni, Viviana, & Todeschini, Roberto. (2012). New similarity coefficients for binary data. *Match-Communications in Mathematical and Computer Chemistry, 68*(2), 581.

Cooper, William S. (1971). The inadequacy of probability of usefulness as a ranking criterion for retrieval system output. *University of California, Berkeley*.

Corder, Gregory W, & Foreman, Dale I. (2014). Richardson, Alice M. Nonparametric Statistics: A Step by Step Approach. *International Statistical Review*. 83(1) 163-164.

Cramer III, Richard D, Redl, George, & Berkoff, Charles E. (1974). Substructural analysis. Novel approach to the problem of drug design. *Journal of Medicinal Chemistry, 17*(5), 533-535.

Dávid Bajusz, Anita Rácz, Károly Héberger. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics., 7:20*.

Davis, Charles H, & McKim, Geoffrey W. (1999). Systematic weighting and ranking: Cutting the Gordian knot. *Journal of the Association for Information Science and Technology, 50*(7), 626.

De Castro, Pablo AD, de França, Fabrício O, Ferreira, Hamilton M, Coelho, Guilherme Palermo, & Von Zuben, Fernando J. (2010). Query expansion using an immune-inspired biclustering algorithm. *Natural Computing, 9*(3), 579-602.

De Vivo, Marco, Masetti, Matteo, Bottegoni, Giovanni, & Cavalli, Andrea. (2016). Role of molecular dynamics and related methods in drug discovery. *Journal of medicinal chemistry, 59*(9), 4035-4061.

Deng, Guannan, Song, LianLian, Jiang, Yanli, & Fu, Jingchao. (2015). A kind of monotonic similarity measure of interval-valued fuzzy sets. *The Fuzzy Systems and Knowledge Discovery (FSKD), 12th International Conference*. 14 January 2016. Zhangjiajie, China.1-4.

Devillers, James, & Balaban, Alexandru T. (2000). *Topological indices and related descriptors in QSAR and QSPAR*: CRC Press.(1).Amisterrdam Ntherland.

Dice, Lee R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*(3), 297-302.

DiMasi, Joseph A, Grabowski, Henry G, & Hansen, Ronald W. (2016). Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of health economics, 47*, 20-33.

Dixon, Steven L, & Koehler, Ryan T. (1999). The hidden component of size in two-dimensional fragment descriptors: side effects on sampling in bioactive libraries. *Journal of medicinal chemistry, 42*(15), 2887-2900.

Dörr, Alexander, Rosenbaum, Lars, & Zell, Andreas. (2015). A ranking method for the concurrent learning of compounds with various activity profiles. *Journal of Cheminformatics*(1), 2.

Downs, Geoffrey M, & Willett, Peter. (1996). Similarity searching in databases of chemical structures. *Reviews in computational chemistry, 7*, 1-66.

Downs, Geoffrey M, Willett, Peter, & Fisanick, William. (1994). Similarity searching and clustering of chemical-structure databases using molecular property data. *Journal of Chemical Information and Computer Sciences, 34*(5), 1094-1102.

Drwal, Malgorzata N, & Griffith, Renate. (2013). Combination of ligand-and structure-based methods in virtual screening. *Drug Discovery Today: Technologies, 10*(3), e395-e401.

Durrant, Jacob D, & Amaro, Rommie E. (2015). Machine-Learning Techniques Applied to Antibacterial Drug Discovery. *Chemical biology & drug design, 85*(1), 14-21.

Dyke, H. Gordon. (1959). A figure-of-merit ordering system for a search output. *American Documentation, 10*(1), 85-86. doi: 10.1002/asi.5090100112

Ellis, David, Furner-Hines, Jonathan, & Willett, Peter. (1993). Measuring the degree of similarity between objects in text retrieval systems. *Perspectives in Information Management, 3*(2).

Faith, Daniel P. (1983). Asymmetric binary similarity measures. *Oecologia, 57*(3), 287-290.

Faith, Daniel P, Minchin, Peter R, & Belbin, Lee. (1987). Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio, 69*(1-3), 57-68.

Flower, Darren R. (1998). On the properties of bit string-based measures of chemical similarity. *Journal of chemical information and computer sciences, 38*(3), 379-386.

from Structure, Selecting Optimally Diverse Compounds. (1997). Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors Matter. *Journal of Medicinal Chemistry, 40*(8), 1219-1229.

Gasteiger, Johann. (2003). *Handbook of chemoinformatics* (Vol. 1): Wiley Online Library.

Gasteiger, Johann. (2016). Chemoinformatics: Achievements and challenges, a personal view. *Molecules, 21*(2), 151.

Gasteiger, Johann, & Engel, Thomas. (2006). *Chemoinformatics: a textbook*: John Wiley & Sons.

Gasteiger, Johann, & Zupan, Jure. (1993). Neural networks in chemistry. *Angewandte Chemie International Edition in English, 32*(4), 503-527.

Ginn, Claire MR, Willett, Peter, & Bradshaw, John. (2002). Combination of molecular similarity measures using data fusion *Virtual Screening: An Alternative or Complement to High Throughput Screening?* (pp. 1-16).Netherlands .Springer.

Girschick, Tobias, Puchbauer, Lucia, & Kramer, Stefan. (2013). Improving structural similarity based virtual screening using background knowledge. *Journal of cheminformatics, 5*, 50.

Gleason, Henry Allan. (1920). Some applications of the quadrat method. *Bulletin of the Torrey Botanical Club, 47*(1), 21-33.

Godden, Jeffrey W, Xue, Ling, & Bajorath, Jürgen. (2000). Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *Journal of Chemical Information and Computer Sciences, 40*(1), 163-166.

Grant, J Andrew, Haigh, James A, Pickup, Barry T, Nicholls, Anthony, & Sayle, Roger A. (2006). Lingos, finite state machines, and fast similarity searching. *Journal of chemical information and modeling, 46*(5), 1912-1918.

Gravetter, Frederick J, & Wallnau, Larry B. (2016). *Statistics for the behavioral sciences*: Cengage Learning.(10th ed).China.Cenge laerning Publisher.

Guo, Shengbo, & Sanner, Scott. (2010). *Probabilistic latent maximal marginal relevance.* Paper presented at the Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval.

H Haga, Jason, & Ichikawa, Kohei. (2016). Virtual screening techniques and current computational infrastructures. *Current pharmaceutical design, 22*(23), 3576-3584.

Han, LY, Ma, XH, Lin, HH, Jia, J, Zhu, F, Xue, Y, . . . Chen, YZ. (2008). A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor. *Journal of Molecular Graphics and Modelling, 26*(8), 1276-1286.

Harper, Gavin, Bradshaw, John, Gittins, John C, Green, Darren VS, & Leach, Andrew R. (2001). Prediction of biological activity for high-throughput screening using binary kernel discrimination. *Journal of Chemical Information and Computer Sciences, 41*(5), 1295-1300.

Hasegawa, Kiyoshi, & Funatsu, Kimito. (2010). Non-linear modeling and chemical interpretation with aid of support vector machine and regression. *Current computer-aided drug design, 6*(1), 24-36.

Hersey, Anne, Chambers, Jon, Bellis, Louisa, Bento, A Patrícia, Gaulton, Anna, & Overington, John P. (2015). Chemical databases: curation or integration by user-defined equivalence? *Drug Discovery Today: Technologies, 14*, 17-24.

Hert, Jérôme, Willett, Peter, Wilton, David J, Acklin, Pierre, Azzaoui, Kamal, Jacoby, Edgar, & Schuffenhauer, Ansgar. (2004). Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *Journal of chemical information and computer sciences, 44*(3), 1177-1185.

Hert, Jérôme, Willett, Peter, Wilton, David J, Acklin, Pierre, Azzaoui, Kamal, Jacoby, Edgar, & Schuffenhauer, Ansgar. (2005). Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information. *Journal of medicinal chemistry, 48*(22), 7049-7054.

Hert, Jérôme, Willett, Peter, Wilton, David J, Acklin, Pierre, Azzaoui, Kamal, Jacoby, Edgar, & Schuffenhauer, Ansgar. (2006). New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *Journal of chemical information and modeling, 46*(2), 462-470.

Hiemstra, Djoerd. (2001). *Using language models for information retrieval*: Taaluitgeverij Neslia Paniculata.(1 th ed). Netherlands. Taaluitgeverij Neslia Paniculata

Hill, T, & Lewicki, P. (2012). *Electronic statistics textbook.(5 th ed)*. Tulsa, *OK, USA*.

Holliday, John D, Hu, CY, & Willett, Peter. (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.

Holliday, John D, Salim, Naomie, Whittle, Martin, & Willett, Peter. (2003). Analysis and display of the size dependence of chemical similarity coefficients. *Journal of chemical information and computer sciences, 43*(3), 819-828.

Holliday, John D, & Willett, Peter. (1996). Definitions of" dissimilarity" for dissimilarity-based compound selection. *Journal of Biomolecular Screening, 1*(3), 145-151.

Holliday, John D, Willett, Peter, & Xiang, Hua. (2013). Interactions between weighting scheme and similarity coefficient in similarity-based virtual screening. *Methodologies and Applications for Chemoinformatics and Chemical Engineering*, pp. 310-321.

Horvath, D, & Jeandenans, C. (2000). Molecular similarity and virtual screening. In silico methods to retrieve active analogs in the context of discovering therapeutic compounds. *ACTUALITE CHIMIQUE*(9), 64-67.

Hu, Guoping, Kuang, Guanglin, Xiao, Wen, Li, Weihua, Liu, Guixia, & Tang, Yun. (2012). Performance evaluation of 2D fingerprint and 3D shape similarity methods in virtual screening. *Journal of chemical information and modeling, 52*(5), 1103-1113.

Huang, Ruili, Southall, Noel, Wang, Yuhong, Yasgar, Adam, Shinn, Paul, Jadhav, Ajit, . . . Austin, Christopher P. (2011). The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Science translational medicine, 3*(80), 80ps16-80ps16.

Hubalek, Zdenek. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biological Reviews, 57*(4), 669-689.

Hull, Richard D, Singh, Suresh B, Nachbar, Robert B, Sheridan, Robert P, Kearsley, Simon K, & Fluder, Eugene M. (2001). Latent semantic structure indexing (LaSSI) for defining chemical similarity. *Journal of medicinal chemistry, 44*(8), 1177-1184.

Jaccard, Paul. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist, 11*(2), 37-50.

Jaghoori, Mohammad Mahdi, Altena, Allard J, Bleijlevens, Boris, Ramezani, Sara, Font, Juan Luis, & Olabarriaga, Silvia D. (2015). A multi-infrastructure gateway for virtual drug screening. *Concurrency and Computation: Practice and Experience, 27*(16), 4478-4490.

Johnson, Mark A, & Maggiora, Gerald M. (1990).*Concepts and applications of molecular similarity*. (2 th ed). Los Angeles, Calif, Wiley.

Jones, Karen Sparck. (1971). A*utomatic keyword classification for information retrieval*.London.Butterworths

Jorissen, Robert N, & Gilson, Michael K. (2005). Virtual screening of molecular databases using a support vector machine. *Journal of chemical information and modeling, 45*(3), 549-561.

Kar, Supratik, & Roy, Kunal. (2013). How far can virtual screening take us in drug discovery? *Expert Opinion on Drug Discovery, 8*(3), 245-261.

Kitchen, Douglas B, Decornez, Hélène, Furr, John R, & Bajorath, Jürgen. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery, 3*(11), 935-949.

Klinger, S., & Austin, J. (2006). *Weighted superstructures for chemical similarity searching .The 9th Joint Conference on Information Sciences.* May 10, 2006.1-18.

Kriegel, Hans-Peter, Brecheisen, Stefan, Kröger, Peer, Pfeifle, Martin, & Schubert, Matthias. (2003). *Using sets of feature vectors for similarity search on voxelized CAD objects.* The Proceedings of the 2003 ACM SIGMOD international conference on Management of data. June 9-12, 2003, San Diego, CA. pp. 587-598.

Kubinyi, Hugo, Mannhold, Raimund, Timmerman, Hendrik, Böhm, Hans-Joachim, & Schneider, Gisbert. (2008). *Virtual screening for bioactive molecules* (Vol. 10): John .Weinheim Germany.Wiley & Sons.

Lajiness, Michael S. (1997). Dissimilarity-based compound selection techniques. *Perspectives in drug discovery and design, 7*, 65-84.

Lavecchia, Antonio. (2015). Machine-learning approaches in drug discovery: methods and applications. *Drug discovery today, 20*(3), 318-331.

Lavecchia, Antonio, & Di Giovanni, Carmen. (2013). Virtual screening strategies in drug discovery: a critical review. *Current medicinal chemistry, 20*(23), 2839-2860.

Lavrenko, Victor, & Croft, W Bruce. (2001). Relevance based language models. The Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 1 Sep 2001, 120-127.

Lee, Changki, & Lee, Gary Geunbae. (2006). Information gain and divergence-based feature selection for machine learning-based text categorization. *Information processing & management, 42*(1), 155-165.

Lew, Michael S, Sebe, Nicu, Djeraba, Chabane, & Jain, Ramesh. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2*(1), 1-19.

Li, Jiao, Zheng, Si, Chen, Bin, Butte, Atul J, Swamidass, S Joshua, & Lu, Zhiyong. (2016). A survey of current trends in computational drug repositioning. *Briefings in bioinformatics, 17*(1), 2-12.

Li, Liwei, Wang, Bo, & Meroueh, Samy O. (2011). Support Vector Regression Scoring of Receptor–Ligand Complexes for Rank-Ordering and Virtual Screening of Chemical Libraries. *Journal of Chemical Information and Modeling, 51*(9), 2132-2138.

Lin, Yung-Shen, Jiang, Jung-Yi, & Lee, Shie-Jue. (2014). A similarity measure for text classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on, 26*(7), 1575-1590.

Lv, Yuanhua, & Zhai, ChengXiang. (2009). *A comparative study of methods for estimating query language models with pseudo feedback.* Paper presented at the Proceedings of the 18th ACM conference on Information and knowledge management. 2-6 November 2009. Hong Kong, China .1895-1898.

Lv, Yuanhua, & Zhai, ChengXiang. (2010). Positional relevance model for pseudo-relevance feedback. *Proceedings of the 33rd international ACM SIGIR*

*conference on Research and development in information retrieval.*July 19 - 23, 2010. Geneva, Switzerland. 579-586.

Lynch, Michael, & Ritland, Kermit. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics, 152*(4), 1753-1766.

Maldonado, AnaG, Doucet, J. P., Petitjean, Michel, & Fan, Bo-Tao. (2006). Molecular similarity and diversity in chemoinformatics: From theory to applications. *Molecular Diversity, 10*(1), 39-79.

Manning, Christopher D, Raghavan, Prabhakar, & Schütze, Hinrich. (2008). *Introduction to information retrieval* (Vol. 1). Americas, New York:Cambridge university press. Cambridge.

Maron, Melvin Earl, & Kuhns, John L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM), 7*(3), 216-244.

Marrero-Ponce, Yovani, Castillo-Garit, Juan A, Olazabal, Ervelio, Serrano, Hector S, Morales, Alcidez, Castanedo, Nilo, . . . Torrens, Francisco. (2005). Atom, atom-type and total molecular linear indices as a promising approach for bioorganic and medicinal chemistry: theoretical and experimental assessment of a novel method for virtual screening and rational design of new lead anthelmintic. *Bioorganic & medicinal chemistry, 13*(4), 1005-1020.

Martin, Yvonne C, Kofron, James L, & Traphagen, Linda M. (2002). Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry, 45*(19), 4350-4358.

Mathews, Jonathan P, & Chaffee, Alan L. (2012). The molecular representations of coal–A review. *Fuel, 96*, 1-14.

Matter, Hans. (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.

McGaughey, Georgia B, Sheridan, Robert P, Bayly, Christopher I, Culberson, J Chris, Kreatsoulas, Constantine, Lindsley, Stacey, . . . Cornell, Wendy D. (2007). Comparison of topological, shape, and docking methods in virtual screening. *Journal of chemical information and modeling, 47*(4), 1504-1519.

McGill, Michael. (1979). *An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems*. National Science Foundation, Washington, DC

Melucci, Massimo. (2011). Can information retrieval systems be improved using quantum probability? *Advances in Information Retrieval Theory* (pp. 139-150): Springer.

Milton, J Susan, & Arnold, Jesse C. (2002). *Introduction to probability and statistics: principles and applications for engineering and the computing science(1 ed).*United States :McGraw-Hill, Inc.

Morgan, Steve, Grootendorst, Paul, Lexchin, Joel, Cunningham, Colleen, & Greyson, Devon. (2011). The cost of drug development: a systematic review. *Health Policy, 100*(1), 4-17.

Mori, Tatsunori, & Sasaki, Takuro. (2002). Information Gain Ratio meets Maximal Marginal Relevance. *Proceedings of the Third NTCIR Workshop*. 2 Feb 2002 .Tokyo, Japan. 101-118.

Muchmore, Steven W, Debe, Derek A, Metz, James T, Brown, Scott P, Martin, Yvonne C, & Hajduk, Philip J. (2008). Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *Journal of chemical information and modeling, 48*(5), 941-948.

Muegge, Ingo, & Mukherjee, Prasenjit. (2016). An overview of molecular fingerprint similarity search in virtual screening. *Expert opinion on drug discovery, 11*(2), 137-148.

Mulvihill, John, & Brenner, Everett H. (1968). Ranking boolean search output. *American Documentation, 19*(2), 204-205.

Nasr, Ramzi, Vernica, Rares, Li, Chen, & Baldi, Pierre. (2012). Speeding up chemical searches using the inverted index: the convergence of chemoinformatics and text search methods. *Journal of chemical information and modeling, 52*(4), 891-900.

Nikolova, Nina, & Jaworska, Joanna. (2003). Approaches to Measure Chemical Similarity – a Review. *QSAR & Combinatorial Science, 22*(9-10), 1006-1026.

Niwattanakul, Suphakit, Singthongchai, Jatsada, Naenudorn, Ekkachai, & Wanapu, Supachanun. (2013). Using of Jaccard coefficient for keywords similarity. *Proceedings of the International MultiConference of Engineers and Computer Scientists*. 13 March 2013. Hong Kong. Vol. 1, p. 6.

Obaid, Abdullah Y, Voleti, Sreedhara, Bora, Roop Singh, Hajrah, Nahid H, Omer, Abdulkader M Shaikh, Sabir, Jamal SM, & Saini, Kulvinder Singh. (2017).

Cheminformatics studies to analyze the therapeutic potential of phytochemicals from Rhazya stricta. *Chemistry Central Journal, 11*(1), 11.

Ono, Katsuki, Takeuchi, Koh, Ueda, Hiroshi, Morita, Yasuhiro, Tanimura, Ryuji, Shimada, Ichio, & Takahashi, Hideo. (2014). Structure-Based Approach To Improve a Small-Molecule Inhibitor by the Use of a Competitive Peptide Ligand. *Angewandte Chemie, 126*(10), 2635-2639.

Pilot, Scitegic Pipeline. (2008). Accelrys. *Inc.: San Diego, CA*.

Plewczynski, Dariusz, Spieser, Stéphane AH, & Koch, Uwe. (2006). Assessing different classification methods for virtual screening. *Journal of chemical information and modeling, 46*(3), 1098-1106.

Ponte, Jay M, & Croft, W Bruce. (1998). A language modeling approach to information retrieval. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 24 - 28, Aug 1998 . New York, NY, USA .pp. 275-281.

Rathke, Fabian, Hansen, Katja, Brefeld, Ulf, & Müller, Klaus-Robert. (2010). StructRank: a new approach for ligand-based virtual screening. *Journal of chemical information and modeling, 51*(1), 83-92.

Riniker, Sereina, & Landrum, Gregory A. (2013). Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of cheminformatics, 5*(1), 1-17.

Robertson, Stephen E. (1977). The probability ranking principle in IR. *Readings in information retrieval*, 281-286.

Robertson, Stephen E, & Jones, K Sparck. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science, 27*(3), 129-146.

Roche, Olivier, Trube, Gerhard, Zuegge, Jochen, Pflimlin, Pascal, Alanine, Alexander, & Schneider, Gisbert. (2002). A virtual screening method for prediction of the HERG potassium channel liability of compound libraries. *ChemBioChem, 3*(5), 455-459.

Rogers, David J, & Tanimoto, Taffee T. (1960). A computer program for classifying plants. *Science (New York, NY), 132*(3434), 1115-1118.

Rognan, Didier, & Bonnet, Pascal. (2014). [Chemical databases and virtual screening]. *Medecine sciences: M/S, 30*(12), 1152-1160.

Rohrer, Sebastian G, & Baumann, Knut. (2009). Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *Journal of chemical information and modeling, 49*(2), 169-184.

Rollinger, Judith M, Stuppner, Hermann, & Langer, Thierry. (2008). Virtual screening for the discovery of bioactive natural products .*Natural Compounds as Drugs Volume I* (pp. 211-249).

Salim, N., Holliday, J., & Willett, P. (2003). Combination of fingerprint-based similarity coefficients using data fusion. *J Chem Inf Comput Sci, 43*. doi: 10.1021/ci025596j

Salton, Gerard, Singhal, Amit, Mitra, Mandar, & Buckley, Chris. (1997). Automatic text structuring and summarization. *Information Processing & Management, 33*(2), 193-207.

Sastry, G Madhavi, Inakollu, VS Sandeep, & Sherman, Woody. (2013). Boosting Virtual Screening Enrichments with Data Fusion: Coalescing Hits from Two-Dimensional Fingerprints, Shape, and Docking. *Journal of chemical information and modeling, 53*(7), 1531-1542.

Schneider, Gisbert, & Böhm, Hans-Joachim. (2002). Virtual screening and fast automated docking methods. *Drug Discovery Today, 7*(1), 64-70.

Schucany, William R, & Frawley, William H. (1973). A rank test for two group concordance. *Psychometrika, 38*(2), 249-258.

Sheridan, Robert P, & Kearsley, Simon K. (2002). Why do we need so many chemical similarity search methods? *Drug discovery today, 7*(17), 903-911.

Sirci, Francesco, Goracci, Laura, Rodríguez, David, van Muijlwijk-Koezen, Jacqueline, Gutiérrez-de-Terán, Hugo, & Mannhold, Raimund. (2012). Ligand-, structure-and pharmacophore-based molecular fingerprints: a case study on adenosine A1, A2A, A2B, and A3 receptor antagonists. *Journal of computer-aided molecular design, 26*(11), 1247-1266.

Sneath, PHA. (1966). Relations between chemical structure and biological activity in peptides. *Journal of theoretical biology, 12*(2), 157-195.

Sokal, obert R., & Sneath, Peter Henry Andrews. (1963). *Principles of numerical taxonomy*: .NY .USA :WH Freeman.

Sokal, Robert R. (1958). A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull, 38*, 1409-1438.

Sokal, RR. (1985). The principles of numerical taxonomy: twenty-five years later. *Computer-assisted bacterial systematics*(15), 1.

Sørensen, Thorvald. (1948). {A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons}. *Biol. Skr., 5*, 1-34.

Stumpfe, Dagmar, & Bajorath, Jürgen. (2011). Applied Virtual Screening: Strategies, Recommendations, and Caveats *Virtual Screening* (pp. 291-318): Wiley-VCH Verlag GmbH & Co. KGaA.

Swamidass, S Joshua, & Baldi, Pierre. (2007). Mathematical correction for fingerprint similarity measures to improve chemical retrieval. *Journal of chemical information and modeling, 47*(3), 952-964.

Symyx Technologies. MDL drug data report: Sci Tegic Accelrys Inc., the MDL Drug Data Report (MDDR). Database is available at http://www.accelrys.com/. .

Taktak, Imen, Tmar, Mohamed, & Hamadou, Abdelmajid Ben. (2009). Query reformulation based on relevance feedback *Flexible query answering systems* .(pp. 134-144). .

Tao, Tao, & Zhai, ChengXiang. (2006). Regularized estimation of mixture models for robust pseudo-relevance feedback. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.* 06-11 August 2006. Seattle, Washington, USA. 162-169.

Todeschini, Roberto, & Consonni, Viviana. (2009). *Molecular Descriptors for Chemoinformatics, Volume 41 (2 Volume Set)* (Vol. 41): John Wiley & Sons.

Todeschini, Roberto, Consonni, Viviana, Xiang, Hua, Holliday, John, Buscema, Massimo, & Willett, Peter. (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.

Todeschini, Roberto, Lasagni, Marina, & Marengo, Emilio. (1994). New molecular descriptors for 2D and 3D structures. Theory. *Journal of chemometrics, 8*(4), 263-272.

Trinajstić, Nenad, Klein, Douglas J, & Randić, Milan. (1986). On some solved and unsolved problems of chemical graph theory. *International Journal of Quantum Chemistry, 30*(S20), 699-742.

Truchon, Jean-François, & Bayly, Christopher 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.

Varnek, Alexandre, & Tropsha, Alex. (2008). *Chemoinformatics approaches to virtual screening*.NC US :Royal Society of Chemistry.

Vuorinen, Anna, Engeli, Roger, Meyer, Arne, Bachmann, Fabio, Griesser, Ulrich J, Schuster, Daniela, & Odermatt, Alex. (2014). Ligand-based pharmacophore modeling and virtual screening for the discovery of novel 17β-hydroxysteroid dehydrogenase 2 inhibitors. *Journal of medicinal chemistry, 57*(14), 5995-6007.

Walters, W. Patrick, Stahl, Matthew T., & Murcko, Mark A. (1998). Virtual screening—an overview. *Drug Discovery Today, 3*(4), 160-178.

Wang, Binghe, & Ekins, Sean. (2006). *Computer applications in pharmaceutical research and development* (Vol. 2). ATLANTA GA US : John Wiley & Sons.

Wang, Binghe, Hu, Longqin, & Siahaan, Teruna J. (2016). *Drug delivery: principles and applications*.ATLANTA GA US : John Wiley & Sons.

Wang, Jun, & Zhu, Jianhan. (2009). Portfolio theory of information retrieval. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.* 19 - 23 July 2009 .Boston, MA, USA. 115-122.

Wassermann, Anne Mai, Geppert, Hanna, & Bajorath, Jürgen. (2009). Searching for target-selective compounds using different combinations of multiclass support vector machine ranking methods, kernel functions, and fingerprint descriptors. *Journal of chemical information and modeling, 49*(3), 582-592.

Weininger, David. (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.

Whittle, Martin, Willett, Peter, Klaffke, Werner, & van Noort, Paula. (2003). Evaluation of similarity measures for searching the dictionary of natural

products database. *Journal of chemical information and computer sciences, 43*(2), 449-457.

Widdows, Dominic, & Widdows, Dominic. (2004). *Geometry and meaning* (Vol. 773). CSLI ,US,publications Stanford.

Willett, John. (1987). *Similarity and clustering in chemical information systems.* New York, NY, USA: John Wiley & Sons, Inc.

Willett, P. (2003). Similarity-based approaches to virtual screening. *Biochemical Society Transactions, 31*(Pt 3), 603-606.

Willett, P. (2013). Combination of similarity rankings using data fusion. *J Chem Inf Model, 53*. doi: 10.1021/ci300547g

Willett, Peter. (2000). Textual and chemical information processing: different domains but similar algorithms. *Information Research, 5*(2).

Willett, Peter. (2006a). Enhancing the Effectiveness of Ligand-Based Virtual Screening Using Data Fusion. *QSAR & Combinatorial Science, 25*(12), 1143-1152.

Willett, Peter. (2006b). Similarity-based virtual screening using 2D fingerprints. *Drug discovery today, 11*(23), 1046-1053.

Willett, Peter. (2009). Similarity methods in chemoinformatics. *Annual review of information science and technology, 43*(1), 1-117.

Willett, Peter. (2011). Similarity searching using 2D structural fingerprints *Chemoinformatics and Computational Chemical Biology* (pp. 133-158): Springer.

Willett, Peter. (2013). Combination of similarity rankings using data fusion. *Journal of chemical information and modeling, 53*(1), 1-10.

Willett, Peter, Barnard, John M., & Downs, Geoffrey M. (1998). Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences, 38*(6), 983-996. doi: 10.1021/ci9800211

Willett, Peter, & Winterman, Vivienne. (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.

Willett, Peter, Winterman, Vivienne, & Bawden, David. (1986). Implementation of nearest-neighbor searching in an online chemical structure search system. *Journal of Chemical Information and Computer Sciences, 26*(1), 36-41.

Williams, Antony J, Ekins, Sean, & Tkachenko, Valery. (2012). Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. *Drug discovery today, 17*(13), 685-701.

Wilson, Gregory L, & Lill, Markus A. (2011). Integrating structure-based and ligand-based approaches for computational drug design. *Future medicinal chemistry, 3*(6), 735-750.

Wilton, David J, Harrison, Robert F, Willett, Peter, Delaney, John, Lawson, Kevin, & Mullier, Graham. (2006). Virtual screening using binary kernel discrimination: analysis of pesticide data. *Journal of chemical information and modeling, 46*(2), 471-477.

Xu, Jinxi, & Croft, W Bruce. (1996). Query expansion using local and global document analysis. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval.* 18 - 22 August 1996.Zurich, Switzerland.4-11.

Xu, Jun, & Hagler, Arnold. (2002). Chemoinformatics and drug discovery. *Molecules, 7*(8), 566-600.

Xue, Ling, Godden, Jeffrey W, & Bajorath, Jürgen. (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.

Yang, Lingpeng, Ji, Donghong, & Leong, Munkew. (2007). Document reranking by term distribution and maximal marginal relevance for chinese information retrieval. *Information processing & management, 43*(2), 315-326.

Yap, Chun Wei. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry, 32*(7), 1466-1474.

Zamora, Elena M, & Blower Jr, Paul E. (1984). Extraction of chemical reaction information from primary journal text using computational linguistics

techniques. 1. Lexical and syntactic phases. *Journal of chemical information and computer sciences, 24*(3), 176-181.

Zhai, ChengXiang, & Lafferty, John. (2006). A risk minimization framework for information retrieval. *Information Processing & Management, 42*(1), 31-55.

Zhao, Wei, Hevener, Kirk E, White, Stephen W, Lee, Richard E, & Boyett, James M. (2009). A statistical framework to evaluate virtual screening. *BMC bioinformatics, 10*(1), 1.

Zoete, Vincent, Grosdidier, Aurélien, & Michielin, Olivier. (2009). Docking, virtual high throughput screening and in silico fragment-based drug design. *Journal of cellular and molecular medicine, 13*(2), 238-248.

Zuccon, Guido, & Azzopardi, Leif. (2010). Using the Quantum Probability Ranking Principle to Rank Interdependent Documents. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger & K. van Rijsbergen (Eds.), *Advances in Information Retrieval* (Vol. 5993, pp. 357-369).

Zuccon, Guido, Azzopardi, Leif, & Rijsbergen, C.J. "Keith" van. (2010). Has portfolio theory got any principles*? Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 19-23 Julay 2010 .Geneva, Switzerland.755-756.

Zuccon, Guido, Azzopardi, Leif, & Rijsbergen, C.J. "Keith" van. (2011). The interactive PRP for diversifying document rankings. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 24-28 July 2011. Beijing, China.14-29.

Zuccon, Guido, Azzopardi, Leif, & van Rijsbergen, CJ. (2011a). The interactive PRP for diversifying document rankings. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.* 24-28 July 2011. Beijing, China.42-56.

Zuccon, Guido, Azzopardi, Leif, & Van Rijsbergen, CJ Keith. (2011b). An analysis of ranking principles and retrieval strategies *Advances in Information Retrieval Theory* (pp. 151-163).

Zuccon, Guido, Azzopardi, Leif, Zhang, Dell, & Wang, Jun. (2012). Top-k retrieval using facility location analysis *Advances in Information Retrieval* (pp. 305-316).

Zuccon, Guido, Azzopardi, LeifA, & van Rijsbergen, Keith. (2009). The Quantum Probability Ranking Principle for Information Retrieval. In L. Azzopardi, G. Kazai, S. Robertson, S. Rüger, M. Shokouhi, D. Song & E. Yilmaz (Eds.), *Advances in Information Retrieval Theory* (Vol. 5766, pp. 232-240).

LIST OF PUBLICATIONS

1. **Mubarak Himmat** , Naomie Salim , Mohammed Mumtaz Al-Dabbagh , Faisal Saeed and Ali Ahmed," *Adapting Document Similarity Measures for Ligand-Based Virtual Screening* ", Molecules 21 (4), 476 (2016) (**Indexed: Impact Factor 2.64 ,Q2**).
2. **Mubarak Himmat** , Naomie Salim , Mohammed Mumtaz Al-Dabbagh , Faisal Saeed and Ali Ahmed. "*An algorithm for similarity-based virtual screening*." Journal of Chemical & Pharmaceutical Research 7.4 (2015).(**Indexed: Scopus**).
3. **Mubarak Himmat** , Naomie Salim , Mohammed Mumtaz Al-Dabbagh , Faisal Saeed and Ali Ahmed,"*DATA FUSION APPROACHES IN LIGAND-BASED VIRTUAL SCREENING: RECENT DEVELOPMENTS OVERVIEW*", ARPN Journal of Engineering and Applied Sciences 10 (3), 1017-1022, (2015) .(Indexed: Scopus).
4. **Mubarak Himmat** , Naomie Salim , Mohammed Mumtaz Al-Dabbagh , Faisal Saeed and Ali Ahmed ,"*Data mining and fusion methods in ligand-based virtual screening*", Journal of Chemical and Pharmaceutical Sciences. (2015) (**Indexed: Scopus**).
5. Mohammed Mumtaz Al-Dabbagh , Naomie Salim, **Mubarak Himmat** , Faisal Saeed and Ali Ahmed", A quantum-based similarity method in virtual screening" , Molecules 20 (10), 18107(2015), **(Indexed: Impact Factor 2.64 ,Q2).**
6. **Mubarak Himmat** , Naomie Salim , Mohammed Mumtaz Al-Dabbagh , Faisal Saeed and Ali Ahmed," Enhancement of Virtual Screening using Bit-strings reweighting "**.Under review** .
7. **Mubarak Himmat** , Naomie Salim , Mohammed Mumtaz Al-Dabbagh , Faisal Saeed and Ali Ahmed," Applying Alternative Ranking Method to Enhance Ligand-Based Virtual Screening " **Under review** .