MOLECULAR SIMILARITY SEARCHING BASED ON DEEP LEARNING FOR FEATURE REDUCTION

MAGED MOHAMMED SAEED NASSER

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

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

MAY 2022

DEDICATION

This thesis is dedicated to my father, who taught me that the best kind of knowledge to have is that which is learned for its own sake. It is also dedicated to my mother, who taught me that even the largest task can be accomplished if it is done one step at a time.

ACKNOWLEDGEMENT

First of all, all praise and thanks are due to Allah, and peace and blessings be upon His Messenger, Muhammad (Peace Be Upon Him). I would like to thank Allah for granting me this opportunity and providing me with the strength to complete this thesis successfully.

I also wish to express my sincere appreciation to my supervisor, Prof.Dr. Naomie Salim, for her kind advice, encouragement, and guidance. Her belief that it was, indeed, possible to finish kept me going. Without her valuable guidance this study could never have reached its completion. I am also very thankful to my brother Prof.Dr. Mohamed MS. Nasser for his unlimited support and precious advice.

Lastly and always, I wish to express my deepest gratitude and love for my beloved family members, especially my mother, father, brothers and sisters for their utmost support, patience and understanding throughout my PhD study. I also extend my heartfelt thanks to Dr. Faisal Saeed, Dr. Hamza Hentabli, Fuad A. Ghaleb, Dr. Saeed Hamood Alsamhi and Dr. Waleed Ali for the support and incisive comments in making this study a success. My sincere appreciation is also extended to all my colleagues. Thanks all for being my inspiration.

ABSTRACT

The concept of molecular similarity has been widely used in rational drug design, where structurally similar molecules are explored in molecular databases for retrieving functionally similar molecules. The most used conventional similarity methods are two-dimensional (2D) fingerprints to evaluate the similarity of molecules towards a target query. However, these descriptors include redundant and irrelevant features that might impact the effectiveness of similarity searching methods. Moreover, the majority of existing similarity searching methods often disregard the importance of some features over others and assume all features are equally important. Thus, this study proposed three approaches for identifying the important features of molecules in chemical datasets. The first approach was based on the representation of the molecular features using Autoencoder (AE), which removes irrelevant and redundant features. The second approach was the feature selection model based on Deep Belief Networks (DBN), which are used to select only the important features. In this approach, the DBN is used to find subset of features that represent the important ones. The third approach was conducted to include descriptors that complement to each other. Different important features from many descriptors were filtered through DBN and combined to form a new descriptor used for molecular similarity searching. The proposed approaches were experimented on the MDL Data Drug Report standard dataset (MDDR). Based on the test results, the three proposed approaches overcame some of the existing benchmark similarity methods, such as Bayesian Inference Networks (BIN), Tanimoto Similarity Method (TAN), Adapted Similarity Measure of Text Processing (ASMTP) and Quantum-Based Similarity Method (SQB). The results showed that the performance of the three proposed approaches proved to be better in term of average recall values, especially with the use of structurally heterogeneous datasets that could produce results than other methods used previously to improve molecular similarity searching.

ABSTRAK

Konsep keserupaan molekul telah digunakan secara meluas dalam reka bentuk ubat rasional, di mana molekul yang serupa secara struktur dicari dalam pangkalan data molekul untuk mendapatkan semula molekul yang serupa secara fungsi. Kaedah keserupaan konvensional yang paling banyak digunakan ialah cap jari dua dimensi (2D) untuk menilai keserupaan molekul dengan molekul sasaran. Walau bagaimanapun, pemerihal ini merangkumi ciri berlebihan dan tidak berkaitan yang mungkin mempengaruhi keberkesanan kaedah pencarian keserupaan. Selain itu, sebilangan besar kaedah pencarian keserupaan yang sedia ada sering mengabaikan kepentingan antara ciri berbanding yang lain dan menganggap semua ciri adalah sama penting. Oleh itu, kajian ini mencadangkan tiga pendekatan untuk mengenal pasti ciri penting molekul dalam set data kimia. Pendekatan pertama adalah berdasarkan perwakilan ciri molekul menggunakan Auto-Pengekod (AE) yang menyingkirkan ciri-ciri yang tidak relevan dan berlebihan. Pendekatan kedua adalah model pemilihan ciri berdasarkan Rangkaian Kepercayaan Mendalam (DBN) yang digunakan untuk memilih ciriciri penting sahaja. Dalam pendekatan ini, DBN digunakan sebagai pengenalpastian ciri subset yang mewakili ciri-ciri yang penting. Akhirnya, pendekatan ketiga telah dijalankan untuk memastikan pemerihal yang saling melengkapi. Ciri penting yang berbeza dari setiap pemerihal kemudiannya ditapis melalui DBN dan digabungkan untuk membentuk pemerihal baharu yang digunakan untuk pencarian keserupaan molekul. Pendekatan yang dicadangkan telah diuji pada set data piawaian Laporan Data Ubat MDL (MDDR). Berdasarkan keputusan ujian, tiga pendekatan yang dicadangkan mengatasi beberapa kaedah keserupaan penanda aras sedia ada seperti Rangkaian Inferens Bayesian (BIN), Kaedah Keserupaan Tanimoto (TAN), Pengukuran Keserupaan Terhadap Pemprosesan Teks (ASMTP) dan Kaedah Keserupaan Berasaskan Kuantum (SQB). Hasil kajian menunjukkan bahawa prestasi ketiga-tiga pendekatan yang dicadangkan terbukti lebih baik dari segi dapatan semula terutamanya dengan penggunaan set data yang mempunyai struktur pelbagai dapat menghasilkan keputusan yang lebih baik berbanding kaedah lain yang digunakan sebelum ini untuk meningkatkan pencarian keserupaan molekul.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1D	-	One Dimension
2D	-	Two Dimensions
3D	-	Three Dimensions
AE	-	Autoencoders
AI	-	Artificial Intelligence
AM	-	Adjacency Matrix
ASMTP	-	Adapted Similarity Measure of Text Processing
BDL	-	Bayesian Deep Learning
BIN	-	Bayesian Inference Networks
BP	-	Back Propagation
CC	-	Combinatorial Chemistry
CDK	-	Atom Type Chemistry Development Kit
CNN	-	Convolutional Neural Networks
DBM	-	Deep Boltzmann Machines
DBN	-	Deep Belief Networks
DILI	-	Predictive Modelling of Drug-Induced Liver Injury
DL	-	Deep learning
DNN	-	Deep Neural Networks
ECFC	-	Atom Type Extended Connectivity Fingerprints Counts
ECFP	-	Atom Type Extended-Connectivity Fingerprint
EEFC	-	Atom Type Atom Environment Fingerprints
EHFC	-	Atom Type Hashed Atom Environment Fingerprints
EPFC	-	Atom Type Connectivity Fingerprint Counts
FCFC	-	Functional Class Extended Connectivity Fingerprints Counts
FCFP	-	Functional Class Extended Connectivity Fingerprints
FEFC	-	Functional Class Atom Environment Fingerprints
FHFC	-	Atom Type Atom Environment Fingerprints
FPFC	-	Functional Class Daylight Path-Based Fingerprint Counts
FPFP	-	Functional Class Daylight Path-Based Fingerprints
GOFP	-	Graph Only Fingerprints

HTS	-	High-Throughput Biochemical Compound Screening
LBVS	-	Ligand-Based Virtual Screening
LCFC	-	Log P Types Extended Connectivity Fingerprint Counts
LCFP	-	Log P Extended Connectivity Fingerprint
LPFC	-	Log P Types Daylight Path-Based Fingerprint Counts
LPLP	-	Log P Types Daylight Path-Based Fingerprints
LWDOSM	-	Language for Writing Descriptors of Outline Shape of Molecules
MDDR	-	MDL Drug Data Report
ML	-	Machine Learning
MRF	-	Markov Random Field
MSE	-	Mean Squared Error
NN	-	Neural Networks
PCA	-	Principal component analysis
QSAR	-	Quantitative Structure–Activity Relationship
RBM	-	Restricted Boltzmann Machines
SBVS	-	Structure-Based Virtual Screening
SMILES	-	Simplified Molecular Input Line Entry Specification
SMTP	-	Similarity Measure for Text Classification and Clustering
SQB	-	Quantum-Based Similarity Measure
TAN	-	Tanimoto
TSS	-	Turbo Similarity Searching
VS	-	Virtual Screening

LIST OF SYMBOLS

w	-	Weight Matrix
b	-	Bias Vector
Φ	-	Activation Function
$\mathcal{L}$, e	-	Error Value
α, ε	-	Learning Rate
h	-	Hidden Layer
v	-	Visible Layer

CHAPTER 1

INTRODUCTION

1.1 Introduction

Over the years, computers have been utilized in chemical and pharmaceutical research with the aim to reduce the cost of drug discovery. There are various types of computer techniques and methods that have been employed throughout chemical, biomedical, and other medical fields especially data mining and information retrieval.

The actual laboratory drug discovery experimentation is bound to take between 12 to 15 years with an estimated cost of over one billion dollars (Hughes *et al.*, 2011). Because of that, extensive work has been conducted in this research area. Many researchers try to resolve the time-consuming drug discovery and its high-cost issues. Over the last decade, chemoinformatics has been known as one of the richest scientific areas where it offers a multi-disciplinary area that incorporates various disciplines which includes computational chemistry, chemometrics, and Quantitative Structure Activity Relationship (QSAR).

Chemoinformatics is also known as Chemical Informatics and Chemical Information. It combines the computer science and chemistry discipline to retrieve information of chemical compounds (Begam and Kumar, 2012; Engel, 2006). The first definition of chemoinformatics was pioneered by Brown in the Annual Reports of Medicinal Chemistry which deliberated on the role and impact of Chemoinformatics in Drug Discovery. Chemoinformatics definition was presented by Brown (Brown, 1998) as the following:

“Chemoinformatics is the combination of those information resources to transform data into information and information into knowledge with the aim to offer a prompt and better decision in the area of drug leads to identification and organization".

The definition varied among other researchers. According to Paris (1999): "Chemoinformatics is a generic word which integrates design, creation, organization, storage, management, retrieval, analysis, dissemination, visualization and chemical information application based on its unique terms and as a surrogate or index for other data and knowledge".

The use of computational screening methods advances as it becomes one of the most significant techniques available for drug discovery. Besides, it is also believed to be a substitute for high-throughput biochemical compound screening (HTS). The HTS is formerly known to be the fundamental approach and most important method for developing drug candidates. However, virtual screening with the variations techniques and search methods offered better reliability for drug discovery.

Virtual Screening (VS) is one of the most widely used computational methods for searching small molecules libraries in drug discovery. The VS is frequently used in protein to identify structures that are most likely to bind to a drug target (Rollinger *et al.*, 2008b). Unlike high-throughput screening (HTS), which requires that a compound exist physically, the main advantage of virtual screening is that, it was carried out using computational techniques that allowed researchers to screen and search extremely large parts of the chemical space and massive number of molecules in a short period of time with minimal risk and cost.

There are two approaches known in virtual screening which are ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS) methods. One of LBVS approaches is similarity searching, which aims to search and scan chemical databases for molecules that are most identical to a user-defined reference structure by implementing a quantitative measurement of intermolecular structural similarity. LBVS methods search for those compounds that are similar to the ligand and requires a known active input. The second approach, SBVS searches for compounds that fits the target binding site and it requires the structure of the target protein (Gonczarek *et al.*, 2016).

Although significant improvements in ligand-based virtual screening could be made, further effort is necessary to promote a rapid drug discovery process and address several of the main issues as well as challenges to manage the exponentially growing amount of molecular data (Cereto-Massagué *et al.*, 2015; Muegge and Mukherjee, 2016). The main aim of this proposed study is to identify whether the concepts of deep learning approaches based on new sources of knowledge can improve the performance of molecular similarity searching.

This chapter proceeds as follows: Section 1.2 reviews the problem background; Section 1.3 explains the problem statements. Section 1.4 presents the research questions of the study; Section 1.5 describes on the objectives of the study. Section 1.6 mentions the scope of the study; Section 1.7 explains the Significance of this Study and lastly Section 1.8 shows the organization of the thesis.

1.2 Problem Background

As mentioned earlier in this chapter, the average cost to bring a drug to market is more than USD 1 billion and the average time taken from the initial phase of drug discovery until the drug can reach the market is between 12-15 years. With the use of chemoinformatics, researchers aim to provide a better solution to making the drug development process less risky and costly (DiMasi *et al.*, 2016; Wang *et al.*, 2016).

Computer aided drug design has been used to aid in the process of drug discovery. It optimizes and reduces the time and cost of discovering and developing new drugs. In a computer-aided drug discovery program, virtual screening methods have been used to select compounds for testing or to design combinatorial libraries (Li *et al.*, 2016).

The use of HTS screening has proven to help perform millions of chemicals, pharmacological, or genetic tests over a short period of time with the help of computer that is able to execute millions of processes within seconds. Despite significant advances in the field of computational drug discovery and ligand prediction (Chen *et al.*, 2016; De Vivo *et al.*, 2016), the most extensively employed approach is yet far from ideal, and greater effort is needed in meeting chemists expectations. There are two main objectives in drug discovery which are addressing issues of long-time processing and the high cost of manufacturing for the discovery of new chemical entities. With virtual screening, researchers aim to identify alternative approaches to discover new active compounds and deliver these compounds to market in a timely manner.

It is a hassle for chemists to deal with the problem of selecting chemical structures to synthesize from a vast variety of compounds. But this only deal with a small percentage number of molecules that could be synthesized. Hence, chemical search techniques are known as virtual screening that covers various computational techniques which are developed to evaluate a huge number of compounds using computers rather than experiences (Bajorath, 2013; Muegge and Mukherjee, 2016). These computational approaches can be applied to search for chemical libraries and remove out those of unnecessary chemical compounds.

In cheminformatics, various similarity measures have been used for virtual screening which helped to improve the screening results. There are several similarity measures that have been developed and derived from current similarity measures which have proven to be effective in other domains but have yet to be used in virtual screening. Text information retrieval algorithms can be applied to chemical information retrieval (Obaid *et al.*, 2017). Bayesian Inference Network is one of the techniques that have been applied in text for several types of research domains and it has been used extensively in virtual screening as an alternative similarity searching method which outperformed the conventional similarity approaches (Abdo et al., 2010; Abdo et al., 2014; Abdo et al., 2012; Abdo et al., 2011). There are several similarity measures that have been recently developed for virtual screening that outperformed the Tanimoto coefficient such as quantum-based similarity measure

(SQB) (Al-Dabbagh *et al.*, 2015), and adapting text similarity measures (ASMTP) (Himmat *et al.*, 2016) which has been derived from a similarity measure of text processing and ideal for virtual screening.

Many researchers in chemoinformactics have taken great interest in the fragment bases and bit-strings similarity method particularly, virtual screening (Abdo and Salim, 2010; Ahmed *et al.*, 2012a) in which variations of study have been conducted within the domain area. The molecular databases (fingerprint) consists of massive amount of bit-strings that describes the molecules features (Bajorath, 2017a; Todeschini and Consonni, 2009). Many of the classical similarity approaches assume that molecular features/ fragments that do not relate to biological activity carry the same weight as the important ones. Generally, chemists may consider some fragments to be more important than others based on the chemical structure diagrams, such as functional groups. For this reason, the weight for each fragment in chemical structure compounds has been investigated by giving more weight to those fragments that are more important. Therefore, a match between two molecules on highly weighted features contributes more to a total similarity than a match based on less important features (Arif *et al.*, 2016). Various weighting functions have been introduced for a new fragment weighting scheme for Bayesian Inference Network in Ligand-Based Virtual Screening (Abdo and Salim, 2010). The fragment reweighting approach has been developed by applying reweighting variables and relevance feedback for a better performance of retrieval recall of Bayesian Inference Network (Ahmed *et al.*, 2012b).

The application of data fusion techniques has contributed to major improvement on the overall performances of conventional similarity methods (Ahmed *et al.*, 2014a; Willett, 2013). The combination of multiple data sources that is translated into a single source in which, the fused source result is expected to be more informative compared to the individual input sources (Liggins II *et al.*, 2017). The concept of multiple information combination sources has been adequately applied (Willett, 2013) and recent researchers have found that in terms of similarity, more actives among top ranking molecules can be obtained using fusion of several similarity coefficients compared to individual coefficients (Brey *et al.*, 2002). The molecular representations, query molecules, docking scores and similarity coefficients have been combined mostly using linear combination techniques (Salim *et al.*, 2003). In many fusion experiments, the use of fusion source over a single source has shown better results either in text retrieval or chemical compound retrieval. In order to achieve the best retrieval performance through data fusion, these two requirements must be taken into measure; the accurateness of each individual source and the independence of sources relative to one another (Saeed *et al.*, 2014).

Most of the similarity-based virtual screening techniques deal with large amounts of data that contain redundant and irrelevant features. The current molecule's fingerprint is often made up of multiple properties. Furthermore, since their importance levels differ, removing some features may improve the recall of the similarity measure (Vogt *et al.*, 2010). Data with irrelevant and redundant features can affect the results of the virtual screening and make it harder to interpret (Liu and Motoda, 2007; Solorio-Fernández *et al.*, 2020). Features selection can enhance the recall of similarity measure and allow important features to be given more weight while ignoring the unimportant features (Vogt *et al.*, 2010). Many types of currently used fingerprints are highly complex, with many features, therefore many bit positions, often exceeding 1000 features or fragments.

Several deep learning techniques and each with their own advantages offered have been widely applied in various types of fields and it is believed that the performance of the proposed deep learning in these fields were the best compared to all previously work done for addressing similar problems. Autoencoders (AE) is a powerful deep learning technique that essentially used in a situation where complex data such as image and video are involved. AE is good in handling low dimensional feature representation from the inputs based on the unsupervised learning (Liu *et al.*, 2017; Strub and Mary, 2015; Zhang *et al.*, 2019). The AE has the advantage of establishing a functional link between the high-dimensions and low-dimensions representations and vice versa. The AE establishes efficient functional linkages between the high-dimensions and low- dimensions representations and is compelled by using the nonlinear distance metric-based cost function to provide a meaningful points arrangement in the low- dimensions representation (Lemke and Peter, 2019). In chemoinformatics, one of the major drawbacks of chemical fingerprints in virtual screening is that the fingerprint descriptors often consist of irrelevant and redundant features. Therefore, by removing some of these features can improve the recall of the similarity measure performance (Vogt et al., 2010). Thus, this research proposed a new approach for identifying the important features of molecules in chemical datasets based on the representation of the molecular features using Autoencoder (AE), with the aim of removing irrelevant and redundant features for improving the performance of the similarity searching measures.

Deep belief network (DBN) is another technique that has been effectively applied to further investigate on feature abstraction and image reconstruction features (Peng *et al.*, 2016; Semwal *et al.*, 2017; Suk *et al.*, 2016; Zou *et al.*, 2015). DBN has been effectively used for multi-level feature selection for the selection of least number of the most discriminative genes in order to improve sample classification accuracy (Ibrahim *et al.*, 2014), as well as feature selection for remote sensing scene classification (Zou *et al.*, 2015). The reconstruction in the DBN can be achieved by applying layer-wise weights on the input features, called reconstruction weights. Given all layer-wise reconstruction weights, a reconstruction error can be calculated for each input feature. The variation of features would commonly make the difference of reconstruction errors. Intuitively, a feature with a lower reconstruction error is more re-constructible. In the feature learning procedure of the DBN, the more re-constructible features are more prone to hold the feature intrinsic. Due to these, the DBN is proposed in this research to investigate whether some features are more important than others, through molecular structure. Moreover, the importance of each feature is taken into consideration, by selecting the features that are more re-constructible as the discriminative features, which leads to a new feature-selection method for ligand-based virtual screening.

1.3 Problem Statement

As discussed in the problem background, it can be inferred that there are two implicit problems relating to chemical similarity search methods which are first, it assumes all molecular features are as equivalent in significance, therefore all molecular features are utilized into similarity measure calculation. Second, each weighting schemes calculates each feature’s weight independently with zero connection to all other features (Vogt *et al.*, 2010). The effectiveness of a retrieval can be improved with the help of features reweighting and feature selection by improving the recall of similarity measure.

Several reweighting methods such as features reweighting, features selection, mini-fingerprint, fuzzy correlation coefficient have been used to improve the performance of Bayesian inference network similarity method (Ahmed *et al.*, 2011a, 2011b; Ahmed *et al.*, 2013), but still the performance for highly diverse dataset is low and require more enhancement (Ahmed *et al.*, 2012a). Additionally, the effectiveness of the performance of similarity methods and retrieval of chemical structures may be improved by importing techniques from different fields (Al-Dabbagh *et al.*, 2015; Himmat *et al.*, 2016).

The chemical datasets are represented using different descriptors to convert molecules into numerical values whereby each descriptor has different important features compared to others (Fouaz *et al.*, 2019). In this research, features selection is proposed to all molecular fingerprints descriptors and only the important features are selected from each descriptor. They are combined to form a new descriptor which then will be used to obtain the improvement of the molecules similarity searching performance for chemical databases.

1.4 Research Questions

The main research question is:

How do representation of molecular features and feature selection positively effect and improve the recall of molecular similarity searching?

To answer the above-mentioned research question, this thesis must address the following issues:

How can a new dimension reduction based on Autoencoder be adopted in LBVS for removing redundant and irrelevant molecular features and improve the effectiveness of the molecular similarity searching?

How can deep belief networks (DBN) be adopted in ligand based virtual screening for molecular reconstruction of features weight?

How to design a new feature selection model for LBVS based on reconstruction features weight to improve the recall of molecular similarity searching?

Can feature selection model be used with multiple descriptors to select and combine the important features from multiple descriptors to improve the effectiveness of molecular similarity searching?

1.5 Research Objectives

The aim of this study is to develop a ligand-based similarity methods based on new molecular features representation, features reweighting and feature selection methods to improve the effectiveness of molecular similarity searching, by reconstructing features weight based on deep learning techniques and selecting only the important features with lowest error value. Selecting a subset of the original feature set will save time and contributes to rapid search on vast chemical databases to retrieve compounds with the most identical biological activity to the reference structure.

Thus, the following objectives have been set with the intent of achieving this goal:

(a) To investigate a new LBVS dimension reduction based on Autoencoder deep learning for a new representation of the molecules features with low dimensions to improve the molecular similarity searching.

(b) To design a new feature selection model for LBVS based deep belief networks to select the important features which can be used to improve the molecular similarity searching.

(c) To improve the molecular similarity searching based on combining the important subset features from multiple descriptors based on feature selection model.

1.6 Research Scope

The focus of this study will be on similarity methods that are based on 2D fingerprints. These similarity methods can be used to measure the extent of molecule pairs which are characterised by 2D fingerprints and would structurally resemble each other. The applications of these methods are conducted using 2D fingerprints that are binary and non-binary.

In addition, this study focuses on different approaches which are based on deep learning that were used to enhance the effectiveness of molecular retrieval. The first approach is based on the representation of the molecular features using Autoencoder (AE), where it aims to remove irrelevant and redundant features and presented a new molecular representation to enhance the recall of the similarity searching. The second approach is features selection model based on deep belief networks (DBN) which is used to calculate the molecular reconstruction features weight and select only the important features according to the least reconstruction features error value that will be used later in similarity calculation to enhance the performance of molecular similarity searching by rapidly screening very large chemical datasets with millions of compounds in a short period of time. The proposed approaches based on deep learning were used MDL Drug Data Report (MDDR) datasets for the training.

The MDDR datasets have been represented by several 2D fingerprints such as atom type extended-connectivity fingerprints (ECFP), atom type extended connectivity fingerprints counts (ECFC), functional class daylight path-based fingerprint counts (FPFC), functional class extended connectivity fingerprints counts (FCFC), atom type connectivity fingerprints counts (EPFC), functional class daylight path-based fingerprints (FPFP), A Log P types extended connectivity fingerprint counts (LCFC), A Log P types daylight path-based fingerprint counts (LPFC), functional class extended connectivity fingerprints (FCFP), A Log P types daylight path-based fingerprints (LPLP) and A Log P extended connectivity fingerprint (LCFP) (Chen and Reynolds, 2002; Klon *et al.*, 2004; Sakkiah *et al.*, 2014); in which all these fingerprints have different important features and different molecular representations. The concept of data fusion with deep learning techniques have been implemented with some of these molecular fingerprints and only the important selected features from each fingerprint were combined to form a new descriptor which will be used to obtain performance improvement of molecular similarity searching for chemical databases.

In this study, the similarity approaches will assess a large dataset that is obtained from the MDDR database. This database makes single and multiple reference structures accessible. The comparison between the performance of this method and the performance of traditional 2D similarity methods will be presented, including Tanimoto Similarity Method (TAN), Bayesian Inference Networks (BIN), Adapted Similarity Measure of Text Processing (ASMTP) and Quantum-Based Similarity Method (SQB).

1.7 Significance of the Study

The primary aim of this research is to improve the effectiveness of molecular similarity searching ligand-based virtual screening by utilizing deep learning techniques for molecular representation features and feature selection for selecting the most essential features for the application of molecular similarity searching calculation. At present, the average cost of discovering and developing a new drug into the market is very expensive, which approximately reaching to $1 billion or more and it is time consuming (Morgan *et al.*, 2011). A huge portion of this cost is estimated due to high failure rates of molecules that seems to be convincing drug candidates during initial stages of screening. This often happened because there are millions of chemical compounds needed to be screened, however it is very difficult to test all these overwhelming numbers of compounds in chemical databases within a short period of time.

Particularly, this study proposes a new molecular representation based on deep learning techniques using multiple levels of abstractions and non-linear transformation so that the features of molecules that are most optimized for virtual screening purposes can be learned and optimized, layer by layer.

Autoencoder has proven to be extremely effective at reducing data dimensionality while preserving significant underlying features. The Autoencoder has been employed to remove the molecules irrelevant and redundant features as well as produce a new molecular representation with low dimension. This new representation based on the low dimension is regarded as a new descriptor and used to enhance the recall of the similarity searching measures.

The most distinguishable representations of molecules have been trained and learned based on beep belief networks to reconstruct the molecules features weight and select only important features for similarity searching. The selection of subset molecular features can help in shortening the time needed for screening hundreds of thousands of compounds. Thus, making it necessary for computer-based methods for compound selection and evaluation. The proposed hybrid deep learning techniques is expected to enhance the performance of the similarity searching to discover more novel drug compounds by proposing new ways to select the important features that helped to enhance the performance of the molecular similarity searching.

1.8 Thesis Organization

The outline of the thesis is presented in this section. There are seven chapters included in this thesis, which are organized as the followings:

Chapter 1, Introduction: This chapter provides an overview of the proposed research work in chemoinformatics, drug discovery, and virtual screening topics, as well as a brief review on the problem background, the problem statement, the objectives of the study, the research scope, and the significance of the study.

Chapter 2, Literature Review: This chapter presents an overview of the area of chemoinformatics which contains the chemical structure representations, chemical descriptors, molecular similarity searching methods, and molecular similarity coefficients. Moreover, this chapter covers the ways used for improving the molecular similarity searching. The overview of deep learning concept and the deep learning use in others research fields and in chemoinformatics has been presented in this chapter. A brief description for the proposed deep learning techniques that will be used in this research is also included. This chapter is concluded with a discussion and overview of the relevance techniques applied to molecular similarity searching, as well as the most effective strategies to enhance these methods performances.

Chapter 3, Research Methodology: This chapter discusses on the methodologies used for achieving the aims presented in this thesis and a general illustration of the experimental designs is also presented.

Chapter 4, Similarity Searching Based on Deep Learning Autoencoder for Molecular Features Representation: This chapter describes and develop a new LBVS dimensionality reduction for low dimensional feature representation proposed to improve the performance of similarity searching measures. The motivation behind the importance of low dimensional feature representation of molecules was discussed in this chapter. This chapter is concluded with results findings and a brief conclusion.

Chapter 5, Features Reweighting and Selection in Ligand-Based Virtual Screening for Molecular Similarity Searching: This chapter describes a new feature selection model for LBVS called DBN-FS. Here in this chapter, the reconstructed features weight for molecules are presented based on deep belief networks (DBN). A new feature selection model is proposed based on the reconstructed features weight to select only the important features those have low error value. The new selected features are then used as a new descriptor to improve the performance of the similarity measure performance. The DBN-FS model and experimental design for this approach is presented in this chapter and the end of the chapter focuses on the results and discussion of the work.

Chapter 6, Combination Selected Important Features from Multiple Descriptors: This chapter provides a new approach of the combination of important features from multiple descriptors and produced a new descriptor based on the selected features to use it for improving the performance of the similarity searching measures. The motivation behind the importance of the combining features from several descriptors was discussed in this chapter. An analysis of the outcomes of this approach is provided at the conclusion of this chapter, and it is compared to all prior suggested approaches as well as standard similarity measures.

Chapter 7, Conclusion and Future Work: The discussion and conclusion on the all-inclusive work of the thesis is presented in this chapter. This chapter emphasizes on the results and its contributions, as well as providing suggestions and recommendations for future work.

REFERENCES

Abdo, A., Chen, B., Mueller, C., Salim, N., and Willett, P. (2010). Ligand-based virtual screening using bayesian networks. *Journal of chemical information and modeling, 50*(6), 1012-1020.

Abdo, A., Leclère, V., Jacques, P., Salim, N., and Pupin, M. (2014). Prediction of new bioactive molecules using a bayesian belief network. *Journal of chemical information and modeling, 54*(1), 30-36.

Abdo, A., Saeed, F., Hamza, H., Ahmed, A., and Salim, N. (2012). Ligand expansion in ligand-based virtual screening using relevance feedback. *Journal of Computer-Aided Molecular Design, 26*(3), 279-287.

Abdo, A., and Salim, N. (2009a). Similarity-based virtual screening using bayesian inference network. *Chemistry Central Journal, 3*(1), 1-12.

Abdo, A., and Salim, N. (2009b). Similarity-based virtual screening using bayesian inference network. *Chemistry Central Journal, 3*(S1), P44.

Abdo, A., and Salim, N. (2009c). Similarity-Based Virtual Screening Using Bayesian Inference Network: Enhanced Search Using 2D Fingerprints and Multiple Reference Structures. *QSAR & Combinatorial Science, 28*(6-7), 654-663.

Abdo, A., and Salim, N. (2010). New fragment weighting scheme for the bayesian inference network in ligand-based virtual screening. *Journal of chemical information and modeling, 51*(1), 25-32.

Abdo, A., and Salim, N. (2011). New fragment weighting scheme for the bayesian inference network in ligand-based virtual screening. *Journal of chemical information and modeling, 51*(1), 25-32.

Abdo, A., Salim, N., and Ahmed, A. (2011). Implementing relevance feedback in ligand-based virtual screening using Bayesian inference network. *Journal of Biomolecular Screening, 16*(9), 1081-1088.

Ahmed, A., Abdo, A., and Salim, N. (2011a). An enhancement of Bayesian inference network for ligand-based virtual screening using minifingerprints. *Paper presented at the Fourth International Conference on Machine Vision* (ICMV 11), 83502U-83502U-83505.

Ahmed, A., Abdo, A., and Salim, N. (2011b). Ligand-based virtual screening using fuzzy correlation coefficient. *International Journal of Computer Applications, 19*(9), 38-43.

Ahmed, A., Abdo, A., and Salim, N. (2012a). Ligand-based virtual screening using Bayesian inference network and reweighted fragments. *The Scientific World Journal, 2012*.

Ahmed, A., Abdo, A., and Salim, N. (2012b). Ligand-based Virtual screening using Bayesian inference network and reweighted fragments. *The Scientific World Journal*, 7.

Ahmed, A., Saeed, F., Salim, N., and Abdo, A. (2014a). Condorcet and borda count fusion method for ligand-based virtual screening. *Journal of cheminformatics, 6*(1), 19.

Ahmed, A., Saeed, F., Salim, N., and Abdo, A. (2014b). Condorcet and borda count fusion method for ligand-based virtual screening. *J. Cheminformatics, 6*, 1.

Ahmed, A., Salim, N., and Abdo, A. (2013). Fragment Reweighting in Ligand-Based Virtual Screening. *Advanced Science Letters, 19*(9), 2782-2786.

Al-Dabbagh, M. M., Salim, N., Himmat, M., Ahmed, A., and Saeed, F. (2015). A quantum-based similarity method in virtual screening. *Molecules, 20*(10), 18107-18127.

Alsenan, S. A., Al-Turaiki, I., and Hafez, A. (2020). Chemoinformatics for Data Scientists: an Overview. *Paper presented at the Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services*, 456-461.

Arif, S. M., Holliday, J. D., and Willett, P. (2015a). The use of weighted 2D fingerprints in similarity-Based virtual screening. In *Advances in Mathematical Chemistry and Applications*, 92-112.

Arif, S. M., Holliday, J. D., and Willett, P. (2015b). *The use of weighted 2D fingerprints in similarity-Based virtual screening* (Vol. 1).

Arif, S. M., Holliday, J. D., and Willett, P. (2016). The Use of Weighted 2D Fingerprints in Similarity-Based Virtual Screening. *Paper presented at the Elsevier Inc*, 92-112.

Bachmann, R., Mizrahi, D., Atanov, A., and Zamir, A. (2022). MultiMAE: Multi-modal Multi-task Masked Autoencoders. *arXiv preprint arXiv:2204.01678*.

Bagherian, M., Sabeti, E., Wang, K., Sartor, M. A., Nikolovska-Coleska, Z., and Najarian, K. (2021). Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Briefings in bioinformatics,* 22(1), 247-269.

Bajorath, J. (2013). Virtual Screening Methods Ch 15. 483-505.

Bajorath, J. (2017a). Compound data mining for drug discovery. In *Bioinformatics,* 247-256.

Bajorath, J. (2017b). Molecular Similarity Concepts for Informatics Applications. *Bioinformatics: Volume II: Structure, Function, and Applications*, 231-245.

Bajorath, J. (2017c). Molecular similarity concepts for informatics applications. In *Bioinformatics,* 231-245.

Bajusz, D., Rácz, A., and Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics, 7*(1), 20.

Baoyi Wang, S. S., Shaomin Zhang. (2015). Research on Feature Selection Method of Intrusion Detection Based on Deep Belief Network. *Proceedings of the 2015 3rd International Conference on Machinery, Materials and Information Technology Applications*, 556-561.

Bawden, D. (1985). Communication, storage and retrieval of chemical information. Pp. 297. Ellis Horwood, Chichester. 1985.

Begam, B. F., and Kumar, J. S. (2012). A Study on Cheminformatics and its Applications on Modern Drug Discovery. *Procedia Engineering, 38,* 1264-1275.

Beltrán, N. H., Duarte-Mermoud, M. A., Salah, S., Bustos, M., Peña-Neira, A. I., Loyola, E., et al. (2005). Feature selection algorithms using Chilean wine chromatograms as examples. *Journal of food engineering, 67*(4), 483-490.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and trends® in Machine Learning, 2*(1), 1-127.

Bero, S. A., Muda, A. K., Choo, Y.-H., Muda, N. A., and Pratama, S. F. (2017). Similarity Measure for Molecular Structure: A Brief Review. *Paper presented at the 6th International Conference on Computer Science and Computational Mathematics (ICCSCM)*, Langkawi, malaysia, volume 892.

Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J., and Chen, H. (2018). Application of generative autoencoder in de novo molecular design. *Molecular informatics, 37*(1-2), 1700123.

Blum, A. L., and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence, 97*(1), 245-271.

Boureau, Y.-l., and Cun, Y. L. (2008). Sparse feature learning for deep belief networks. *Paper presented at the Advances in neural information processing systems*, 1185-1192.

Breuleux, O., Bengio, Y., and Vincent, P. (2011). Quickly generating representative samples from an rbm-derived process. *Neural computation, 23*(8), 2058-2073.

Brey, R. L., Holliday, S., Saklad, A., Navarrete, M., Hermosillo–Romo, D., Stallworth, C., et al. (2002). Neuropsychiatric syndromes in lupus: prevalence using standardized definitions. *Neurology, 58*(8), 1214-1220.

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

Byvatov, E., Fechner, U., Sadowski, J., and Schneider, G. (2003). Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of chemical information and computer sciences, 43*(6), 1882-1889.

Cai, C., Gong, J., Liu, X., Gao, D., and Li, H. (2013). Molecular similarity: methods and performance. *Chinese Journal of Chemistry, 31*(9), 1123-1132.

Capecchi, A., Probst, D., and Reymond, J.-L. (2020). One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of cheminformatics, 12*(1), 1-15.

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

Chen, B., Mueller, C., and Willett, P. (2010). Combination rules for group fusion in similarity-based virtual screening. *Molecular informatics, 29*(6-7), 533-541.

Chen, D., Huang, X., and Fan, Y. (2021). Thermodynamics-based model construction for the accurate prediction of molecular properties from partition coefficients. *Frontiers in chemistry*, 9.

Chen, X., and Reynolds, C. H. (2002). Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *J Chem Inf Comput Sci,* 42.

Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., et al. (2016). Drug--target interaction prediction: databases, web servers and computational models. *Briefings in bioinformatics, 17*(4), 696-712.

Choi, S.-S., Cha, S.-H., and Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics, 8*(1), 43-48.

Chowdhury, G. (2010). Introduction to modern information retrieval, 3rd Edition. *Australian Academic and Research Libraries, 41*(4), 305-306.

Ciregan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. *Paper presented at the Computer Vision and Pattern Recognition (CVPR)*, 3642-3649.

Coates, A., Ng, A., and Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. *Paper presented at the Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 215-223.

Coley, C. W., Rogers, L., Green, W. H., and Jensen, K. F. (2017). Computer-assisted retrosynthesis based on molecular similarity. *ACS central science, 3*(12), 1237-1245.

Collobert, R., and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *Paper presented at the Proceedings of the 25th international conference on Machine learning*, 160-167.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research, 12*(Aug), 2493-2537.

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

Dahl, G. E., Jaitly, N., and Salakhutdinov, R. (2014). Multi-task neural networks for QSAR predictions. *arXiv preprint arXiv:1406.1231*.

Dahl, G. E., Yu, D., Deng, L., and Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing, 20*(1), 30-42.

David, L., Thakkar, A., Mercado, R., and Engkvist, O. (2020). Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics, 12*(1), 1-22.

de Castro, P. A., de França, F. O., Ferreira, H. M., Coelho, G. P., and Von Zuben, F. J. (2010a). Query expansion using an immune-inspired biclustering algorithm. *Natural Computing, 9*(3), 579-602.

de Castro, P. A., Fran\ca, F. O. d., Ferreira, H. M., Coelho, G. P., and Zuben, F. J. V. (2010b). Query expansion using an immune-inspired biclustering algorithm. *Natural Computing, 9*(3), 579-602.

de Sousa Luis, J. A., Barros, R. P., de Sousa, N. F., Muratov, E., Scotti, L., and Scotti, M. T. (2021). Virtual screening of natural products database. *Mini Reviews in Medicinal Chemistry, 21*(18), 2657-2730.

De Vivo, M., Masetti, M., Bottegoni, G., and Cavalli, A. (2016). Role of molecular dynamics and related methods in drug discovery. *Journal of medicinal chemistry, 59*(9), 4035-4061.

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

Dong, J., Cao, D.-S., Miao, H.-Y., Liu, S., Deng, B.-C., Yun, Y.-H., et al. (2015). ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. *Journal of cheminformatics, 7*(1), 1-10.

Dong, S., Wang, P., and Abbas, K. (2021). A survey on deep learning and its applications. *Computer Science Review, 40*, 100379.

Duan, J., Dixon, S. L., Lowrie, J. F., and 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.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., et al. (2015). Convolutional networks on graphs for learning molecular fingerprints. *Paper presented at the Advances in neural information processing systems*, 2224-2232.

Eckert, H., and Bajorath, J. (2006). Design and evaluation of a novel class-directed 2D fingerprint to search for structurally diverse active compounds. *Journal of chemical information and modeling, 46*(6), 2515-2526.

Ellis, D., Furner-Hines, J., and Willett, P. (1993). Measuring the degree of similarity between objects in text retrieval systems. *perspectives in information management-annual review-, 3*, 128-128.

Engel, T. (2006). Basic Overview of Chemoinformatics. *Journal of Chemical Information and Modeling, 46*(6), 2267-2277.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb), 625-660.

Fouaz, B., Hacene, B., Hamza, H., and Saeed, F. (2019). Similarity searching in ligand-based virtual screening using different fingerprints and different similarity coefficients. *International Journal of Intelligent Systems Technologies and Applications, 18*(4), 405-425.

Ghasemi, F., Mehridehnavi, A., Perez-Garrido, A., and Perez-Sanchez, H. (2018). Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. *Drug Discov. Today, 23*(10), 1784-1790.

Gillet, V., and Willett, P. (2007). Compound selection using measures of similarity and dissimilarity. In *Comprehensive Medicinal Chemistry II* (Vol. 4), 167-191.

Glorot, X., and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Paper presented at the Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249-256.

Gonczarek, A., Tomczak, J. M., Zaręba, S., Kaczmar, J., Dąbrowski, P., and Walczak, M. J. (2016). Learning Deep Architectures for Interaction Prediction in Structure-based Virtual Screening. *arXiv preprint arXiv:1610.07187*.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *Paper presented at the Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, 6645-6649.

Grebner, C., Matter, H., and Hessler, G. (2022). Artificial Intelligence in Compound Design. In *Artificial Intelligence in Drug Design*, 349-382.

Hentabli, H., Salim, N., Abdo, A., and Saeed, F. (2012). LWDOSM: language for writing descriptors of outline shape of molecules. *Paper presented at the*

International Conference on Advanced Machine Learning Technologies and Applications, 247-256.

Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., et al. (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, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., et al. (2005). Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information. *Journal of medicinal chemistry, 48*(22), 7049-7054.

Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., et al. (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.

Himmat, M., Salim, N., Al-Dabbagh, M. M., Saeed, F., and Ahmed, A. (2016). Adapting document similarity measures for ligand-based virtual screening. *Molecules, 21*(4), 476.

Hinton, G. (2010). A practical guide to training restricted Boltzmann machines. *Momentum, 9*(1), 926.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine, 29*(6), 82-97.

Hinton, G. E. (2012). A practical guide to training restricted Boltzmann machines. In *Neural networks: Tricks of the trade,* 599-619.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation, 18*(7), 1527-1554.

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science, 313*(5786), 504-507.

Holliday, J. D., Hu, C., and 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.

Holliday, J. D., Salim, N., Whittle, M., and 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.

Hong, H., Xie, Q., Ge, W., Qian, F., Fang, H., Shi, L., et al. (2008). Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *Journal of chemical information and modeling, 48*(7), 1337-1344.

Hu, Y., Stumpfe, D., and Bajorath, J. (2017). Recent advances in scaffold hopping. *J. Med. Chem, 60*(4), 1238-1246.

Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L. (2011). Principles of early drug discovery. *British journal of pharmacology, 162*(6), 1239-1249.

Ibrahim, R., Yousri, N. A., Ismail, M. A., and El-Makky, N. M. (2014). Multi-level gene/MiRNA feature selection using deep belief nets and active learning. *Paper presented at the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 3957-3960.

Jaitly, N., Nguyen, P., Senior, A., and Vanhoucke, V. (2012). Application of pretrained deep neural networks to large vocabulary speech recognition. *Paper presented at the Thirteenth Annual Conference of the International Speech Communication Association.* 2577-2580.

Jasial, S., Hu, Y., Vogt, M., and Bajorath, J. (2016). Activity-relevant similarity values for fingerprints and implications for similarity searching. *F1000Research, 5*.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. (2014). Caffe: Convolutional architecture for fast feature embedding. *Paper presented at the Proceedings of the 22nd ACM international conference on Multimedia*, 675-678.

Kearsley, S. K., Sallamack, S., Fluder, E. M., Andose, J. D., Mosley, R. T., and Sheridan, R. P. (1996). Chemical similarity using physiochemical property descriptors. *Journal of Chemical Information and Computer Sciences, 36*(1), 118-127.

Kim, Y., Lee, H., and Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. *Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 3687-3691.

Kimber, T. B., Chen, Y., and Volkamer, A. (2021). Deep learning in virtual screening: recent applications and developments. *International Journal of Molecular Sciences, 22*(9), 4435.

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

Klon, A. E., Glick, M., Thoma, M., Acklin, P., and Davies, J. W. (2004). Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. *Journal of medicinal chemistry, 47*(11), 2743-2749.

Kochev, N., Monev, V., and Bangov, I. (2018). 7 Searching Chemical Structures. *Chemoinformatics: Basic Concepts and Methods*, 231.

Kogej, T., Engkvist, O., Blomberg, N., and Muresan, S. (2006). Multifingerprint based similarity searches for targeted class compound selection. *Journal of chemical information and modeling, 46*(3), 1201-1213.

Kombo, D. C., Tallapragada, K., Jain, R., Chewning, J., Mazurov, A. A., Speake, J. D., et al. (2013). 3D molecular descriptors important for clinical success. *Journal of chemical information and modeling, 53*(2), 327-342.

Konda, K. R. (2016). Unsupervised Relational Feature Learning for Vision. Univ.-Bibliothek Frankfurt am Main.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks.* Paper presented at the Advances in neural information processing systems, 1097-1105.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., and Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. *Paper presented at the Proceedings of the 24th international conference on Machine learning*, 473-480.

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. *Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 8595-8598.

Le Roux, N., and Bengio, Y. (2008). Representational power of restricted Boltzmann machines and deep belief networks. *Neural computation, 20*(6), 1631-1649.

Le Roux, N., and Bengio, Y. (2010). Deep belief networks are compact universal approximators. *Neural computation, 22*(8), 2192-2207.

Leach, A. R., and Gillet, V. J. (2007). An introduction to chemoinformatics.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444.

Legendre, P. (2005). Species associations: the Kendall coefficient of concordance revisited. *Journal of agricultural, biological, and environmental statistics, 10*(2), 226.

Lemke, T., and Peter, C. (2019). EncoderMap: dimensionality reduction and generation of molecule conformations. *Journal of chemical theory and computation, 15*(2), 1209-1215.

Li, J., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., and Lu, Z. (2016). A survey of current trends in computational drug repositioning. *Briefings in bioinformatics, 17*(1), 2-12.

Li, S., Kawale, J., and Fu, Y. (2015). Deep collaborative filtering via marginalized denoising auto-encoder. *Paper presented at the Proceedings of the 24th ACM international on conference on information and knowledge management,* 811-820.

Liggins II, M., Hall, D., and Llinas, J. (2017). Handbook of multisensor data fusion: theory and practice.

Lin, Y.-S., Jiang, J.-Y., and Lee, S.-J. (2014a). A similarity measure for text classification and clustering. *IEEE transactions on knowledge and data engineering, 26*(7), 1575-1590.

Lin, Y. S., Jiang, J. Y., and Lee, S. J. (2014b). A similarity measure for text classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on, 26*(7), 1575-1590.

Liu, H., and Motoda, H. (2007). Computational methods of feature selection.

Liu, W., Wang, X., Zhou, X., Duan, H., Zhao, P., and Liu, W. (2020). Quantitative structure-activity relationship between the toxicity of amine surfactant and its molecular structure. *Science of The Total Environment, 702*, 134593.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing, 234*, 11-26.

Luo, L., Zhang, S., Wang, Y., and Peng, H. (2018). An alternate method between generative objective and discriminative objective in training classification Restricted Boltzmann Machine. *Knowledge-Based Systems, 144*, 144-152.

Lusci, A., Pollastri, G., and Baldi, P. (2013). Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *Journal of chemical information and modeling, 53*(7), 1563-1575.

Lv, Y., and Zhai, C. (2009). A comparative study of methods for estimating query language models with pseudo feedback. *Paper presented at the International Conference on Information and Knowledge Management, Proceedings*, 1895-1898.

Lv, Y., and Zhai, C. (2010). Positional relevance model for pseudo-relevance feedback. *Paper presented at the SIGIR 2010 Proceedings - 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 579-586.

Maggiora, G. M. (2014). Introduction to molecular similarity and chemical space. In *Foodinformatics*, 1-81).

Marrero-Ponce, Y., Castillo-Garit, J. A., Olazabal, E., Serrano, H. S., Morales, A., Castanedo, N., et al. (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.

Masui, H., and Yoshida, M. (1996). SPECTRA: a spectral information management system featuring a novel combined search function. *Journal of chemical information and computer sciences*, *36*(2), 294-298.

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

Matter, H., and Pötter, T. (1999). Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *Journal of chemical information and computer sciences*, *39*(6), 1211-1225.

Mauri, A., Consonni, V., and Todeschini, R. (2016). Molecular Descriptors. *Handbook of Computational Chemistry*, 1-29.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Paper presented at the Advances in neural information processing systems*, 3111-3119.

Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C., and Greyson, D. (2011). The cost of drug development: a systematic review. *Health policy, 100*(1), 4-17.

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

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. *Paper presented at the Proceedings of the 28th international conference on machine learning (ICML-11)*, 689-696.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., and Ogata, T. (2015). Audio-visual speech recognition using deep learning. *Applied Intelligence, 42*(4), 722-737.

Obaid, A. Y., Voleti, S., Bora, R. S., Hajrah, N. H., Omer, A. M. S., Sabir, J. S., et al. (2017). Cheminformatics studies to analyze the therapeutic potential of phytochemicals from Rhazya stricta. *Chemistry Central Journal, 11*, 1.

Pacheco, A. G., Krohling, R. A., and da Silva, C. A. (2018). Restricted Boltzmann machine to determine the input weights for extreme learning machines. *Expert Systems with Applications, 96*, 77-85.

Paris, G. (1999). Meeting of the American Chemical Society http://www.warr.com/warrzone2000.html.

Pathirage, C. S. N., Li, J., Li, L., Hao, H., Liu, W., and Wang, R. (2019). Development and application of a deep learning–based sparse autoencoder framework for structural damage identification. *Structural Health Monitoring, 18*(1), 103-122.

Peng, Z., Li, Y., Cai, Z., and Lin, L. (2016). Deep Boosting: Joint feature selection and analysis dictionary learning in hierarchy. *Neurocomputing, 178*, 36-45.

Pipeline Pilot Software : SciTegic Accelrys Inc. San Diego Accelrys Inc; 2008: http://www.accelrys.com/.

Polanski, J., and Gasteiger, J. (2016). Computer representation of chemical compounds. *Handbook of Computational Chemistry; Leszczynski, J., Puzyn, T., Eds*, 1-43.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., et al. (2018). A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR), 51*(5), 1-36.

Rada, J., and Cruz, R. (2014). Vertex-degree-based topological indices over graphs. *MATCH Commun. Math. Comput. Chem, 72*(3), 603-616.

Raevsky, O. A. (2004). Physicochemical descriptors in property-based drug design. *Mini reviews in medicinal chemistry, 4*(10), 1041-1052.

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V. (2015). Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*.

Robertson, S. E., Walker, S., and Hancock-Beaulieu, M. M. (1995). Large test collection experiments on an operational, interactive system: Okapi at TREC. *Information Processing & Management, 31*(3), 345-360.

Rollinger, J. M., Stuppner, H., and Langer, T. (2008a). Virtual screening for the discovery of bioactive natural products, *Progress in Drug Research* (Vol. 65), 211-249.

Rollinger, J. M., Stuppner, H., and Langer, T. (2008b). Virtual screening for the discovery of bioactive natural products. In *Natural Compounds as Drugs,* Volume I, 211-249.

Saeed, F., Salim, N., and Abdo, A. (2014). Combining multiple clusterings of chemical structures using cluster-based similarity partitioning algorithm. *International journal of computational biology and drug design, 7*(1), 31-44.

Saeed, F., Salim, N., and Shamsir, M. S. (2015). Consensus methods for virtual screening and clustering of chemical structure databases. *Journal of Chemical and Pharmaceutical Sciences, 8*(1), 112-116.

Sakkiah, S., Arooj, M., Lee, K. W., and Torres, J. Z. (2014). Theoretical approaches to identify the potent scaffold for human sirtuin1 activator: Bayesian modeling and density functional theory. *Medicinal Chemistry Research, 23*(9), 3998-4010.

Salim, N. (2002). *Analysis and comparison of molecular similarity measures.* University of Sheffield.

Salim, N., Holliday, J., and Willett, P. (2003). Combination of fingerprint-based similarity coefficients using data fusion. *Journal of chemical information and computer sciences, 43*(2), 435-442.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks, 61*, 85-117.

Schneider, G., and Bhm, H. J. (2002). Virtual screening and fast automated docking methods. *Drug Discovery Today, 7*(1), 64-70.

Schuffenhauer, A., Floersheim, P., Acklin, P., and Jacoby, E. (2003). Similarity metrics for ligands reflecting the similarity of the target proteins. *Journal of chemical information and computer sciences, 43*(2), 391-405.

Semwal, V. B., Mondal, K., and Nandi, G. C. (2017). Robust and accurate feature selection for humanoid push recovery and classification: deep learning approach. *Neural Computing and Applications, 28*(3), 565-574.

Sirci, F., Goracci, L., Rodrguez, D., Muijlwijk-Koezen, J. v., Gutirrez-de-Tern, H., and Mannhold, R. (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.

Socher, R., Lin, C. C., Manning, C., and Ng, A. Y. (2011). Parsing natural scenes and natural language with recursive neural networks. *Paper presented at the Proceedings of the 28th international conference on machine learning (ICML-11),* 129-136.

Solorio-Fernández, S., Carrasco-Ochoa, J. A., and Martínez-Trinidad, J. F. (2020). A review of unsupervised feature selection methods. *Artificial Intelligence Review, 53*(2), 907-948.

Strub, F., and Mary, J. (2015). Collaborative filtering with stacked denoising autoencoders and sparse inputs. *Paper presented at the NIPS workshop on machine learning for eCommerce*.

Suk, H.-I., Lee, S.-W., Shen, D., and Initiative, A. s. D. N. (2016). Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis. *Brain Structure and Function, 221*(5), 2569-2587.

Sun, Y., Wang, X., and Tang, X. (2013). Deep convolutional network cascade for facial point detection. *Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition*, 3476-3483.

Sundararajan, K., and Woodard, D. L. (2018). Deep learning for biometrics: A survey. *ACM Computing Surveys (CSUR), 51*(3), 1-34.

Suzuki, Y., and Ozaki, T. (2017). Stacked denoising autoencoder-based deep collaborative filtering using the change of similarity. *Paper presented at the 2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA),* 498-502.

Svensson, F., Karlén, A., and Sköld, C. (2012). Virtual screening data fusion using both structure-and ligand-based methods. *Journal of chemical information and modeling, 52*(1), 225-232.

Swersky, K., Chen, B., Marlin, B., and De Freitas, N. (2010). A tutorial on stochastic approximation algorithms for training restricted Boltzmann machines and deep belief nets. *Paper presented at the Information Theory and Applications Workshop (ITA)*, 2010, 1-10.

Syuib, M., Arif, S. M., and Malim, N. (2013). Comparison of similarity coefficients for chemical database retrieval. *Paper presented at the Artificial Intelligence, Modelling and Simulation (AIMS), 2013 1st International Conference on,* 129-133.

Taktak, I., Tmar, M., and Hamadou, A. B. (2009). Query reformulation based on relevance feedback, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 5822), 134-144.

Tharwat, A. (2016). Principal component analysis-a tutorial. *International Journal of Applied Pattern Recognition, 3*(3), 197-240.

Todeschini, R., and Consonni, V. (2009). *Molecular descriptors for chemoinformatics, volume 41 (2 volume set)* (Vol. 41).

Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., and 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.

Tompson, J. J., Jain, A., LeCun, Y., and Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. *Paper presented at the Advances in neural information processing systems*, 1799-1807.

Unterthiner, T., Mayr, A., Klambauer, G., and Hochreiter, S. (2015). Toxicity prediction using deep learning. *arXiv preprint arXiv:1503.01445*.

Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Wegner, J. K., Ceulemans, H., et al. (2014). Deep learning as an opportunity in virtual screening. *Paper presented at the Proceedings of the Deep Learning Workshop at NIPS*.

Van Rijsbergen, C. J. (2004). The geometry of information retrieval.

Vázquez, J., López, M., Gibert, E., Herrero, E., and Luque, F. J. (2020). Merging ligand-based and structure-based methods in drug discovery: An overview of combined virtual screening approaches. *Molecules, 25*(20), 4723.

Vogt, M., and Bajorath, J. (2012). Chemoinformatics: a view of the field and current trends in method development. *Bioorganic & medicinal chemistry, 20*(18), 5317-5323.

Vogt, M., Wassermann, A. M., and Bajorath, J. (2010). Application of information—Theoretic concepts in chemoinformatics. *Information, 1*(2), 60-73.

Walters, W. P., Stahl, M. T., and Murcko, M. A. (1998). Virtual screening---an overview. *Drug Discovery Today, 3*(4), 160-178.

Wang, B., Hu, L., and Siahaan, T. J. (2016). Drug delivery: principles and applications.

Wang, H., and Raj, B. (2017). On the origin of deep learning. *arXiv preprint arXiv:1702.07800*.

Wang, H., and Yeung, D.-Y. (2016). Towards Bayesian deep learning: A survey. *arXiv preprint arXiv:1604.01662*.

Wang, N., DeLisle, R. K., and Diller, D. J. (2005). Fast small molecule similarity searching with multiple alignment profiles of molecules represented in one-dimension. *Journal of medicinal chemistry, 48*(22), 6980-6990.

Wang, Z., Sun, H., Shen, C., Hu, X., Gao, J., Li, D., et al. (2020). Combined strategies in structure-based virtual screening. *Physical Chemistry Chemical Physics, 22*(6), 3149-3159.

Wani, M. A., Bhat, F. A., Afzal, S., and Khan, A. I. (2020). Introduction to deep learning. In *Advances in Deep Learning,* 1-11.

Whittle, M., Willett, P., Klaffke, W., and van Noort, P. (2003). Evaluation of similarity measures for searching the dictionary of natural products database. *Journal of chemical information and computer sciences, 43*(2), 449-457.

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

Willett, P. (2006a). Enhancing the effectiveness of ligand-based virtual screening using data fusion. *QSAR & Combinatorial Science, 25*(12), 1143-1152.

Willett, P. (2006b). Similarity-based virtual screening using 2D fingerprints. *Drug discovery today, 11*(23-24), 1046-1053.

Willett, P. (2010a). Similarity searching using 2D structural fingerprints. *Chemoinformatics and computational chemical biology,* 133-158.

Willett, P. (2010b). Similarity searching using 2D structural fingerprints. In *Chemoinformatics and computational chemical biology,* 133-158.

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

Willett, P. (2020). The literature of chemoinformatics: 1978–2018 (Vol. 21, pp. 5576).

Willett, P., Barnard, J. M., and Downs, G. M. (1998). Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences, 38*(6), 983-996.

Willett, P., and 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.

Willighagen, E. L., Mayfield, J. W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., et al. (2017). The Chemistry Development Kit (CDK) v2. 0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of cheminformatics, 9*(1), 1-19.

Wu, L., Cui, P., Pei, J., Zhao, L., and Song, L. (2022). Graph Neural Networks. In *Graph Neural Networks: Foundations, Frontiers, and Applications,* 27-37.

Xie, L., Xu, L., Kong, R., Chang, S., and Xu, X. (2020). Improvement of prediction performance with conjoint molecular fingerprint in deep learning. *Frontiers in pharmacology,* 2148.

Xu, J., and Croft, W. B. (2017). Quary Expansion Using Local and Global Document Analysis. *Paper presented at the ACM SIGIR Forum,* 168-175.

Xu, Y., Dai, Z., Chen, F., Gao, S., Pei, J., and Lai, L. (2015). Deep learning for drug-induced liver injury. *Journal of chemical information and modeling, 55*(10), 2085-2093.

Xue, L., Godden, J. W., and Bajorath, J. (2000). Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *Journal of chemical information and computer sciences, 40*(5), 1227-1234.

Xue, L., Godden, J. W., Stahura, F. L., and Bajorath, J. (2003). Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *Journal of chemical information and computer sciences, 43*(4), 1218-1225.

Xue, L., Stahura, F. L., Godden, J. W., and Bajorath, J. (2001). Mini-fingerprints detect similar activity of receptor ligands previously recognized only by three-dimensional pharmacophore-based methods. *Journal of chemical information and computer sciences, 41*(2), 394-401.

Yu, D., and Deng, L. (2011). Deep learning and its applications to signal and information processing [exploratory dsp]. *IEEE Signal Processing Magazine, 28*(1), 145-154.

Zeng, N., Li, H., and Peng, Y. (2021). A new deep belief network-based multi-task learning for diagnosis of Alzheimer’s disease. *Neural Computing and Applications*, 1-12.

Zhang, C., Patras, P., and Haddadi, H. (2019). Deep learning in mobile and wireless networking: A survey. *IEEE Communications surveys & tutorials, 21*(3), 2224-2287.

Zhao, L., Guanhua, N., Hui, W., Qian, S., Gang, W., Bingyou, J., et al. (2020). Molecular structure characterization of lignite treated with ionic liquid via FTIR and XRD spectroscopy. *Fuel, 272*, 117705.

Zou, Q., Cao, Y., Li, Q., Huang, C., and Wang, S. (2014). Chronological classification of ancient paintings using appearance and shape features. *Pattern Recognition Letters, 49*, 146-154.

Zou, Q., Ni, L., Zhang, T., and Wang, Q. (2015). Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters, 12*(11), 2321-2325.

LIST OF PUBLICATIONS

Indexed Journal with Impact Factors

1. **Maged Nasser,** Naomie Salim, Faisal Saeed, Shadi Basurra, Idris Rabiu, Hentabli Hamza, Muaadh A Alsoufi, Feature Reduction for Molecular Similarity Searching Based on Autoencoder Deep Learning, *Biomolecules*, 2022, 12(4), 508 (**Q2, IF = 4.879**).

2. **Maged Nasser,** Naomie Salim, Hentabli Hamza, Faisal Saeed, and Idris Rabiu, Improved deep learning based method for molecular similarity searching using stack of deep belief networks. *Molecules,* 2021. **26**(1): p. 128. (**Q2, IF =4.412**).

3. **Maged Nasser,** Naomie Salim, Hentabli Hamza, Faisal Saeed, and Idris Rabiu, Features Reweighting and Selection in ligand-based Virtual Screening for Molecular Similarity Searching Based on Deep Belief Networks. *Advances in Data Science and Adaptive Analysis*, 2020. 12(03n04): p. 2050009. (**ISI Indexed)**.

4. **Maged Nasser,** Hentabli Hamza, Naomie Salim, and Faisal Saeed, The Measurement of Molecular Biological Activity based on Quantitative Structure Activity Relationships. *International Journal of Innovative Computing*, 2018. 8(3). (**Google Scholar Indexed**).

Indexed Conference Proceedings

1. **Maged Nasser,** Naomie Salim, and Hentabli Hamza. Molecular Similarity Searching Based on Deep Belief Networks with Different Molecular Descriptors. in Proceedings of the 2020 2nd International Conference on Big Data Engineering and Technology. 2020. p. 18-24. (**Scopus indexed**).

2. **Maged Nasser,** Naomie Salim, Hentabli Hamza, and Faisal Saeed. Deep Belief Network for Molecular Feature Selection in Ligand-Based Virtual

Screening. in International Conference of Reliable Information and Communication Technology. 2018. Springer: p. 3-14. (**ISI Indexed**).

3. Hentabli Hamza, **Maged Nasser,** Naomie Salim, and Faisal Saeed. Bioactivity prediction using convolutional neural network. in International Conference of Reliable Information and Communication Technology. 2019. Springer: p. 341-351. (**ISI Indexed**).

4. Hentabli Hamza, Naomie Salim, **Maged Nasser,** and Faisal Saeed, Meramalnet: a deep learning convolutional neural network for bioactivity prediction in structure-based drug discovery, international conference on big data. In: 8th International Conference on Signal, Image Processing and Pattern Recognition, 21 March 2020 - 22 March 2020, Vienna, Austria. (**Scopus Indexed**).