FUSION OF MOLECULAR REPRESENTATIONS AND PREDICTION OF BIOLOGICAL ACTIVITY USING CONVOLUTIONAL NEURAL NETWORK AND TRANSFER LEARNING

HENTABLI HAMZA

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

DECEMBER 2019

DEDICATION

I would like to dedicate this work

in the memory of my father and for my mother

my beloved wife and lovely kids

"Mohamed, Oussama & Saif-edine"

for being patience, supportive, and understanding.

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.

“If we knew what it was, we were doing, it would not be called research, would it?”

Albert Einstein

Research is to see what everybody else has seen, and to think what nobody else has

thought Albert Szent-Gyorgyi

ACKNOWLEDGEMENT

In the Name of Allah, Most Gracious, Most Merciful

All praise and thanks are due to Allah, and peace and blessings be upon his messenger, Mohammed (peace be upon him).

Alhamdulillah, all praises to Allah S.W.T., The Greatest and The Most Merciful for His guidance and blessing, because without him I can't finished this research. I also wish to express my gratitude to my thesis supervisor, Prof. Dr. Naomie Salim who introduced me to the field of chemoinformatics and for her enthusiastic guidance, valuable help, encouragement and patience for all the aspect of the thesis progress. Her numerous comments, criticisms and suggestion during the preparation of this project are greatly appreciated. Also, for her patience on problems that occurred during the process of completing this thesis.

Furthermore, special thanks go to Maged Nasser, for many days of scientific discussions, this work would not have been possible without his support. Many thanks go to the people who supported me in this research, especially Dr. Faisal Saeed and Noureddine. I also would like to thank all my friends that give me support and help during the writing of the thesis. Their support and help always give me the motivation and energy I need to finish the thesis. My appreciation also extended to all academic and non-academic member of the Faculty of Computing for their warm heart cooperation during my stay in Universiti Teknologi Malaysia.

My heartfelt acknowledgement is expressed to my family, especially My mother and Wife, my sisters Meriem, Rekia, Selma, Imane, Rachida, Assmaa, Hanan, Inass and for my brothers Mohamed, Abderzak and Sidali, without their guidance, support, encouragement and advises I may never have overcome this long journey in my studies. When I felt down, their love always gives me the strength to face the challenges of the research. I would also like to thank people that directly or indirectly help me in finishing the thesis. Thank you very much.

ABSTRACT

Basic structural features and physicochemical properties of chemical molecules determine their behaviour during chemical, physical, biological and environmental processes and hence need to be investigated for determining and modelling the actions of the molecule. Computational approaches such as machine learning methods are alternatives to predict physiochemical properties of molecules based on their structures. However, limited accuracy and error rates of these predictions restrict their use. This study developed three classes of new methods based on deep learning convolutional neural network for bioactivity prediction of chemical compounds. The molecules are represented as a convolutional neural network (CNN) with new matrix format to represent the molecular structures. The first class of methods involved the introduction of three new molecular descriptors, namely Mol2toxicophore based on molecular interaction with toxicophores features, Mol2Fgs based on distributed representation for constructing abstract features maps of a selected set of small molecules, and Mol2mat, which is a molecular matrix representation adapted from the well-known 2D-fingerprint descriptors. The second class of methods was based on merging multi-CNN models that combined all the molecular representations. The third class of methods was based on automatic learning of features using values within the neurons of the last layer in the proposed CNN architecture. To evaluate the performance of the methods, a series of experiments were conducted using two standard datasets, namely MDL Drug Data Report (MDDR) and Sutherland datasets. The MDDR datasets comprised 10 homogeneous and 10 heterogeneous activity classes, whilst Sutherland datasets comprised four homogeneous activity classes. Based on the experiments, the Mol2toxicophore showed satisfactory prediction rates of 92% and 80% for homogeneous and heterogeneous activity classes, respectively. The Mol2Fgs was better than Mol2toxicophore with prediction accuracy result of 95% for homogeneous and 90% for heterogeneous activity classes. The Mol2mat molecular representation had the highest prediction accuracy with 97% and 94% for homogeneous and heterogeneous datasets, respectively. The combined multi-CNN model leveraging on the knowledge acquired from the three molecular presentations produced better accuracy rate of 99% for the homogeneous and 98% for heterogeneous datasets. In terms of molecular similarity measure, use of the values in the neurons of the last hidden layer as the automatically learned feature in the multi-CNN model as a novel molecular learning representation was found to perform well with 88.6% in terms of average recall value in 5% structures most similar to the target search. The results have demonstrated that the newly developed methods can be effectively used for bioactivity prediction and molecular similarity searching.

ABSTRAK

Ciri-ciri struktur asas dan sifat fizikokimia molekul kimia boleh menentukan kelakuan molekul semasa proses kimia, fizikal, biologi dan persekitaran dan oleh itu ia perlu dikaji untuk menentukan dan memodelkan semua tindakan molekul. Walau bagaimanapun, ketepatan yang terhad dan kadar ralat ramalan yang tidak sekata mengehadkan penggunaannya. Kajian ini mencadangkan kaedah baru berdasarkan pembelajaran rangkaian neural konvolusi terhadap ramalan bioaktiviti sebatian kimia. Molekul ini diwakili sebagai rangkaian neural konvolusi (CNN) dengan format matriks baru untuk mewakili struktur molekul. Kaedah kelas pertama melibatkan pengenalan tiga deskriptor molekul baru, iaitu Mol2toxicophore, berdasarkan interaksi molekul dengan ciri-ciri toksikophores, Mol2Fgs, berdasarkan perwakilan teragih untuk membina peta ciri abstrak set molekul kecil yang terpilih, dan Mol2mat, yang merupakan matriks molekul perwakilan yang disesuaikan daripada deskriptor cap jari 2D yang terkenal. Kaedah kelas kedua adalah berdasarkan penggabungan model multi-CNN yang menggabungkan semua perwakilan molekul. Kaedah kelas ketiga didasarkan pada pembelajaran pemberat ciri secara automatik yang menggunakan nilai dalam neuron di lapisan terakhir dalam seni bina CNN yang dicadangkan. Untuk menilai prestasi kaedah tersebut, satu siri eksperimen dijalankan menggunakan dua set data standard, iaitu MDL Drug Data Report (MDDR) dan set data Sutherland. Dataset MDDR terdiri daripada 10 kelas aktiviti homogen dan 10 heterogen, sementara kumpulan Sutherland mengandungi empat kelas aktiviti homogen. Berdasarkan eksperimen, Mol2toxicophore menunjukkan kadar ramalan yang memuaskan sebanyak 92% dan 80% untuk kelas aktiviti homogen dan heterogen. Mol2Fgs lebih baik daripada Mol2toxicophore dengan hasil ketepatan ramalan 95% untuk homogen dan 90% untuk kelas aktiviti heterogen. Perwakilan molekul Mol2mat mempunyai ketepatan ramalan tertinggi dengan 97% dan 94% untuk dataset homogen dan heterogen. Model gabungan multi-CNN memanfaatkan pengetahuan yang diperoleh daripada tiga persembahan molekul menghasilkan kadar ketepatan yang lebih baik sebanyak 99% untuk homogen dan 98% untuk dataset heterogen. Dari segi ukuran kesamaan molekul, penggunaan nilai-nilai dalam neuron lapisan tersembunyi yang terakhir sebagai ciri yang dipelajari secara automatik dalam model multi-CNN sebagai perwakilan pembelajaran molekul baru didapati berfungsi dengan baik dengan 88.6% dari segi nilai penarikan balik purata dalam struktur 5% yang paling hampir sama dengan carian sasaran. Hasilnya telah menunjukkan bahawa kaedah yang baru dibangunkan dapat digunakan secara efektif untuk ramalan bioaktiviti dan pencarian kesamaan molekul.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	-	Artificial Neural Network
QSAR	-	Quantitative Structure-Activity Relationship
CAMD	-	Computer Aided Molecular Design
VS	-	Virtual Screening
QSTR	-	Quantitative Structure Toxicity Relationship
QSPR	-	Quantitative Structure Property Relationship
MNA	-	Multilevel Neighbourhoods of Atoms
BKD	-	Binary Kernel Discrimination
NBC	-	Naiïve Bayesian Classifier
SVM	-	Support Vector Machines
AUC	-	Area Under Curve
CNN	-	Convolutional Neural Network
DL	-	Deep Learning
ML	-	Machine Learning
MDDR	-	MDL Drug Data Report
SMILES	-	Simplified Molecular Input Line System
ECFP	-	Extended-Connectivity Fingerprints
TAN	-	Tanimoto Coefficient
PCA	-	Principal Component Analysis
AI	-	Artificial Intelligence
DNN	-	Deep Neural Network
FGs	-	Functional Groups
SAR	-	Structure-Activity Relationship
NLP	-	Natural Language Processing
MPS	-	Mean Pairwise Similarity

LIST OF SYMBOLS

δ	-	squared deviations
W	-	Kendall

CHAPTER 1

INTRODUCTION

1.1 Overview

The ability to store and search chemical structures and other related information in a computer database have prompted a huge increment in the quantity of chemical compounds and biological information that is accessible for discovery programs in pharmaceutical and agrochemical commercial ventures. Computer Aided Molecular Design (CAMD) it is a method that helps in studying the chemical properties of structural configurations which have been developed via software programs. The concept of rational molecular design uses synergy of chemical combinations and permutations via computer software and advanced computer technology so that new compounds can be developed. The important processes used in CAMD and computational chemistry are Quantitative Structure Activity Relationships (QSAR), molecular quantum mechanics, machine learning, analysis on basis of structural configuration, molecular graphics and illustration of data illustrating the binding of ligand to receptor and calculating the intermolecular bonds.

Many different sectors have benefitted from the use of CAMD like study of organic chemicals, development of new drugs, study of biochemical phenomenon occurring in nature, catalysts and solutions used in experiments. Other fields like agriculture, animal husbandry, medicine and material sciences (like study of compounds made up of different molecules, polymers, chemicals, semiconductors and nonlinear phenomenon) have also benefitted from these developments (Handa *et al.*, 2013; Barakat, 2014; Pérez-Sánchez *et al.*, 2014).

Any chemical compound is characterised by its biological activity, which helps the application of the compound in the agricultural chemistry, cosmetic, medicinal, and food industries (Wang *et al.*, 2014). Biological activity is described as the effect

that is noted when a chemical compound interacts with a biological system. Biological activity highlights the interaction between the chemical compounds and a biological system at any biological organisation level, right from the molecular to the organism level. Hence, the biological activities of the chemical compounds must be studied with the help of different testing systems, like the “in vitro” (i.e., cells, individual molecules or subcellular organelles), “ex vivo” (i.e. isolated tissues or organs), and “in vivo” (animal experiments at the preclinical stage or human trials at the clinical trial stage).

Drug design is based on the ligand structure, i.e., is based on the molecular similarities principal “molecules with a similar structure show similar biological activities”, it is used for analysing the characteristics of the biological activities of the compounds (Abdo et al., 2010). However, this statement is generally violated and the computer-based methods used for biological activity prediction, based on the pairwise structural similarity of the molecules, often provide an inaccurate estimate (Anusevicius et al., 2015). Despite this fact, many machine learning techniques are applied for deriving satisfactory results (Druzhilovskiy *et al.*, 2017; Ancuceanu *et al.*, 2019; Ballester, 2019), as they can help in estimating the biological activity profiles/ spectra of the chemical compounds, such as the diverse biological activities (specific toxicity, pharmacotherapeutic effects, mechanism of action, metabolism, effect of the gene expression, etc.) (Filimonov et al., 2014).

The quantitative structure-activity relationship (QSAR) takes a gander at the consequence of theoretical mixes in light of the examined aftereffect of beforehand blended results. It then associates attributes figured from the ligands structure with their measured results. The discovered results can be used to foresee the activities of hypothetical particles. The outcomes may, likewise, be significant in planning enhanced ligands. QSAR procedures contrast in the plan of particles that relate to the movement and in the systems for choosing the relationship.

In QSAR similarity searching strategy, the activities of unknown compounds (target) are predicted by comparing them with the known chemical compounds. Thereafter, the researcher assigns the activities of similar compounds to the target compounds as shown if Figure 1.1. Though many of the target prediction techniques

have been successful, some problems still exist. Researchers have applied different techniques for predicting different target subsets for the same molecule (Ding *et al.*, 2013).

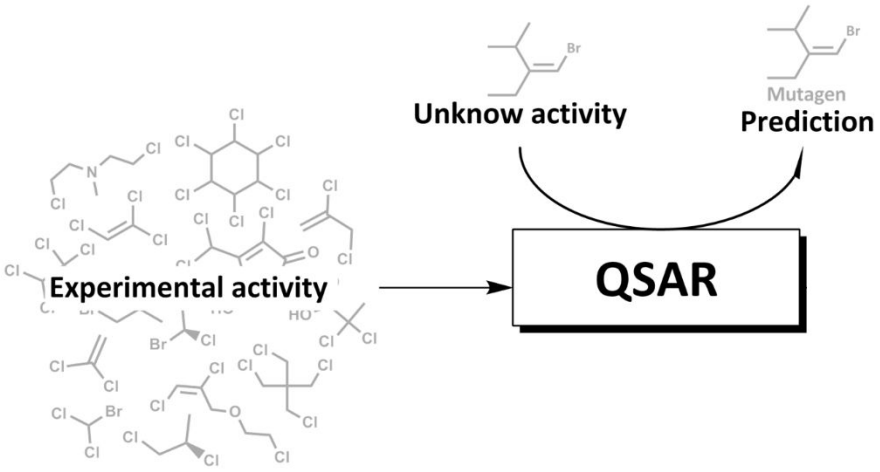

Figure 1.1 The prediction of biological activity for unknown molecular activity.

The popular Machine Learning ML algorithms, using the compound classification method for activity prediction (target), were Binary Kernel Discrimination (BKD) (Willett *et al.*, 2007), Genetic programming for QSAR investigation (Archetti *et al.*, 2010), Naïve Bayesian Classifier (NBC) (Xia *et al.*, 2004), hybrid soft-computing method (Pérez-Sánchez *et al.*, 2014), Artificial Neural Networks (ANNs) (Winkler and Burden, 2002) and Support Vector Machines (SVM) (Kawai *et al.*, 2008). The Bayesian belief network classifier was also used for predicting the ligand-based targets and their activities (Ammar *et al.*, 2014).

The key issue for the problems of Biological Activities Prediction of Chemical Compounds examined in this study is whether application of Deep learning approaches based on new sources of knowledge can improve the prediction accuracy. The aim of the proposed models in this research is to improve prediction using Deep learning Convolutional Neural Network. In this study, proposed and developed a novel machine learning process that was based on the Convolutional Neural Network (CNN) method,

which was considered as one of the best deep learning methods in the field of machine learning algorithms.

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

1.2 Problem Background

A few methods which are used for predicting the biological activity and determining the Structure-Activity Relationship (SAR) are based on the linear regression(Free and Wilson, 1964; Hansch, 1969) and other linear techniques like the partial least squares (Joereskog and Wold, 1982). It is noted that the real-life SARs are non-linear, especially across the diverse set of compounds with different chemical structures. For improving the flexibility and range of the SARs that can be modelled, several novel approaches have been developed based on those described in the machine learning literature (Bolis *et al.*, 1991; Burden, 1996; King and Srinivasan, 1997; Sadowski and Kubinyi, 1998). These models are generally flexible (i.e., a feed-forward neural network having linear output units along with a single hidden layer that can approximate a continuous function having an arbitrary accuracy (Cybenko, 1989)). However, some issues exist like sensitivity to the noise or over-fitting. Some other approaches were based on the estimation (either implicitly (Cramer *et al.*, 1974; Ormerod *et al.*, 1989) or explicitly (Gao *et al.*, 1999)) of the probability of the compound activity, based on the presumption that the descriptor variables were stochastically independent. This assumption is usually not valid, and these techniques show a low ability to model the complex relationships between a diverse set of molecules. In the past few years, some researchers have applied recursive partitioning techniques for problems related to a large number of diverse chemical compounds (Rusinko *et al.*, 1999; Cho *et al.*, 2000). The recursive partitioning processes are better

as they can be predicted using a set of rules, wherein one can easily reach any specific node in the resultant tree.

The similarity methods are seen to be a simple and popular tool for determining the biological activities of the various chemical compounds. This was because these techniques use a single known bioactive molecule (a target or reference molecule) as a start point for database search. The database structures were ranked in decreasing order of similarity with regards to the user-defined, active, reference (query) structure, based on the expectation that all the nearest neighbors display the activity like the query structure. All similarity searching techniques are categorised based on the dimensionality of the molecular structures used for determining the compound similarity, i.e., the 2D and 3D similarity methods.

Many studies published earlier were related to the measurement of molecular similarity for determining the biological activities of molecules (Ding *et al.*, 2013; Kowalski, 2013; Cereto-Massagué *et al.*, 2015; Schymanski *et al.*, 2017; Ancuceanu *et al.*, 2019). However, the popular approaches were based on the 2D fingerprints, wherein the similarities between the target and the database structure were computed based on the association coefficient like the Tanimoto coefficient (Carbó-Dorca and Mezey, 2013; Ding *et al.*, 2013; Ammar *et al.*, 2014; Cereto-Massagué *et al.*, 2015; Kumari *et al.*, 2018). Several similarity methods used for biological activities prediction have been described for computing the QSAR between all molecules. However, the effectiveness of a similarity technique varies according to the biological activities, which affects the prediction (Carbó-Dorca and Mezey, 2013; Simões *et al.*, 2018). Furthermore, two methods retrieve a different subset of active molecules from a chemical database, hence, it is better to use many search methods, if possible (Ancuceanu *et al.*, 2019; Ballester, 2019).

In order to enhance the effectiveness of the similarity methods and QSAR measurement, the aim of this research is to develop a novel method of biological activity prediction, based on deep learning convolutional neural network (CNN), which incorporates the molecule's substructural information which are identified as functional groups or toxicophores. In addition, the biological activities will also be

predicted using the molecules' distributed representation. This approach included the encoding and storage of information regarding the chemical compounds by establishing their interactions and similarities to the standard toxicophores and functional groups. Furthermore, this method should be able to introduce a novel molecular matrix representation for molecular bioactivity mapping using small molecules. Finally, the availability of the combine multi convolutional neural network and several fingerprint types when they are available can help enhance the effectiveness of the molecular representation and prediction method.

1.3 Problem Statement

Since the traditional QSAR methods still suffer from their poor prediction accuracy and sensitivity specially in heterogenous activity classes, more works are still required to develop new approaches for the area of QSAR measurement. Therefore, this research raises several challenges, such as improving the prediction accuracy and enhancing the molecular representation. Here the researcher put forward the Research Questions (RQ) that will be further investigated in this study.

(a) Can we improve the performance of biological activities prediction by utilizing the deep learning convolutional neural network?

(b) How can fusion of molecular descriptor improve the performance of molecular activities prediction of unknown molecules using CNN?

(c) How can convolutional neural network model and transfer learning strategy improve the ligand-based virtual screening and molecular similarity searching?

1.4 Research Objectives

By understanding the problem statements which has been discussed earlier, the QSAR try to measure the biological activity of chemical compound that the Drug design will be interested in it. Since the traditional QSAR methods still suffer from

their poor prediction accuracy and sensitivity, more works are still required to develop new approaches for the area of QSAR measurement. The main goal of this study is to develop high-accuracy QSAR models, relying on CNN technique through taking into account the useful merging many sources of information for the purpose of enhancing the prediction accuracy. Therefore, this research raises several challenges, such as improving the prediction accuracy and enhancing the molecular representation. Thus, to achieve the goal mentioned above, the following objectives have been set.

1. To incorporate the functional groups information, the fingerprint representations and the relationships of small molecules to biological activity into a new CNN-based molecular matrix representation for prediction the biological activity.

2. To combine multi molecular representations in one CNN-based model to improve the performance of the prediction of biological activities specifically in heterogenous activity classes.

3. To investigate whether the combined CNN model and transfer learning strategy can be a better alternative to improve the ligand-based virtual screening.

1.5 Research Scope

In order to achieve the objectives stated above, the scope of this research is limited to the following, the validation and evaluation of the quality of the prediction model proposed in this research will be tested on different datasets that have been used to validate the classification of molecules based on structure-activity relationship.

The databases aimed to be used in this study are only limited to chemical data from MDDR (*Sci Tegic Accelrys Inc*, no date) and four data sets were taken from Sutherland and Helma (Jeffrey J Sutherland *et al.*, 2003; Helma *et al.*, 2004; Sutherland *et al.*, 2004) literatures with compounds classified as active or inactive: cyclooxygenase inhibitors, ligands of the benzodiazepine receptor, dihydrofolate reductase inhibitors, ligands of the estrogen receptor (ER) and finally mutagens (MUT)

of molecular structures. These data sets have been used by literatures for validating prediction models.

The proposed code has been implemented in the Keras (Chollet, 2015), which is a public deep learning software, based on Theano (Bastien *et al.*, 2012) and Tensorflow (Abadi *et al.*, 2016). The weights in the neural networks were initialised according to the Keras settings. All layers in the deep network were initialised simultaneously with the ADADELTA (Zeiler, 2012). The complete network was trained using the Dell Precision T1700 CPU system with a 16GB memory and the professional-grade NVIDIA GeForce GTX 1060 6GB graphics. The next section we put forward the Significance of the Study.

1.6 Significance of the Study

The need to identify the biological activity of molecules is the foundation for the work presented here. It is strongly believed that complete automatic biological activity prediction system can improve and help in the drug development process. The motivation of conducting this PhD study is to propose new state-of-the-art, optimized and innovative techniques for the prediction of biological activity.

Proposed techniques should be capable to provide promising performance in an undesirable situation such as new presentation for molecules, and precise biological activity by reducing the prediction error. In light of the above-mentioned issues, the findings of this study do contribute meaningfully to what is currently known about prediction and estimate the biological activity of unknown molecule. Nonetheless, the significance of this study is not only limited to knowledge enrichment, but also to the development of a new method for future implementation and prediction of biological activity. The next section we put forward the describes the organization of the thesis.

1.7 Thesis Organization

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

Chapter 1, *Introduction*: this chapter gives a general introduction to the topic of the proposed research work. Brief overviews of some of the issues concerning the research are also mentioned in this chapter. Besides the problem background, this chapter also includes the problem statement, objectives of the study, research scope and the significance of the study.

Chapter 2, *Literature Review*: in this chapter, the researcher presents an overview of biological activity studies. It covers the basic approaches of biological activity prediction, the Computer-Aided Molecular Design and Quantitative Structure-Activity Relationships. Furthermore, this chapter also reviews the significant efforts which have been put in biological activity studies and provides the theoretical explanation and fundamental concepts related to it. Also, literature reviews on other concepts related to the current study; such as Molecular Representations, prediction models, Conditions for Applicability of QSAR and QSAR Origins and Evolution. Finally, the challenges that face the biological activity prediction.

Chapter 3, *Research Methodology*: this chapter presents the methodology used in this research. A Methodology is 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. This chapter includes discussion of the research components such as the phases, techniques, and tools involved.

Chapter 4, *Convolutional Neural Network Model Based on Toxicophores Features for Bioactivity Prediction*. described a novel molecular representation which could help in observing and characterising each molecule in the matrix based on its interaction with the toxicophores. This proposed matrix presents the molecular activity of the compound based on its toxic properties. Furthermore, this Mol2toxicophore showed a low overlap and segregated all biological activities.

Chapter 5, *Convolutional Neural Network Model Based on Small Molecules for Bioactivity Prediction*. showed that the distributed representation was able to construct the abstract features, needed for predicting the toxicity of the compounds. This chapter described a novel molecular matrix representation based on a select set of small molecules. Finally, the new matrix representation can easily highlight the biological activities of the unknown molecules.

Chapter 6, *Convolutional Neural Network Model Based on 2d Fingerprint for Bioactivity Prediction*. described the prediction of the biological activities of molecules using the molecular fingerprints in the CNN model. The researcher use 2D fingerprint descriptors as a new molecular matrix for representing the "Mol2mat" in the CNN model. After analysing the multi fingerprints, the researcher study all the probable combinations.

Chapter 7, *Convolutional Neural Network Model Based on Combination of all Representation for Bioactivity Prediction*. describes a new CNN architecture which applies all the knowledge derived from the 3 molecular representatives and combines them together to form one compact molecular descriptor. The researcher presents a combination of the multi-molecular representation with CNN, to predict the activities of the unknown compounds.

Chapter 8, *Adapting the Combination model for Implementing New Molecular Representation Scheme for Ligand-Based Virtual Screening*: the researcher use the same CNN architecture, describe in Chapter 7, for implementing a novel molecular descriptor, which could be used in the ligand-based virtual screening and molecular similarity measurements. This method used the values within the neurons of the CNN layer as a novel molecular learning representative. This method could be very effective for Ligand-Based Virtual Screening.

Chapter 9, *Conclusion and Future Work*: this chapter provides the conclusions of the research work discussed throughout this study. The chapter present and highlights the contributions of the research and put forward recommendations for future work.

REFERENCES

Abadi, M., Barham, P., Chen, J. and Chen, Z. (2016) 'TensorFlow: A System for Large-Scale Machine Learning', This paper is included in the Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16), pp. 265–283.

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), pp. 1012–1020.

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), pp. 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), pp. 1081–8.

Adl, A., Zein, M. and Hassanien, A. E. (2016) 'PQSAR: The membrane quantitative structure-activity relationships in cheminformatics', Expert Systems with Applications. Elsevier Ltd, 54, pp. 219–227.

Ahmed, A., Abdo, A. and Salim, N. (2011) 'Ligand-based Virtual screening using Fuzzy Correlation Coefficient', in International Journal of Computer Applications (0975, pp. 38–43.

Ahmed, A., Abdo, A. and Salim, N. (2012a) 'An enhancement of Bayesian inference network for ligand-based virtual screening using minifingerprints', Fourth International Conference on Machine Vision (ICMV 2011): Computer Vision and Image Analysis; Pattern Recognition and Basic Technologies, 8350(Icmv 2011), p. 83502U.

Ahmed, A., Abdo, A. and Salim, N. (2012b) 'Ligand-Based Virtual Screening Using Bayesian Inference Network and Reweighted Fragments', The Scientific World Journal, 2012, pp. 1–7.

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

Al-Dabbagh, M. M., Salim, N., Himmat, M., Ahmed, A. and Saeed, F. (2017) 'Quantum probability ranking principle for ligand-based virtual screening', Journal of Computer-Aided Molecular Design. Springer International Publishing, 31(4), pp. 365–378.

Ammar, A., Valérie, L., Philippe, J., Naomie, S. and Maude, P. (2014) 'Prediction of new bioactive molecules using a Bayesian belief network', Journal of Chemical Information and Modeling, 54(1), pp. 30–36.

Ancuceanu, R., Dinu, M., Neaga, I., Laszlo, F. G. and Boda, D. (2019) 'Development of QSAR machine learning-based models to forecast the effect of substances on malignant melanoma cells', Oncology Letters, 17(5), pp. 4188–4196.

Angermueller, C., Pärnamaa, T., Parts, L. and Stegle, O. (2016) 'Deep learning for computational biology', Molecular Systems Biology, 12(7), pp. 1–16.

Anusevicius, K., Mickevicius, V., Stasevych, M., Zvarych, V., Komarovska-Porokhnyavets, O., Novikov, V., Tarasova, O., Gloriozova, T. and Poroikov, V. (2015) 'Synthesis and chemoinformatics analysis of N-aryl-β-alanine derivatives', Research on Chemical Intermediates. Springer Netherlands, 41(10), pp. 7517–7540.

Archetti, F., Giordani, I. and Vanneschi, L. (2010) 'Genetic programming for QSAR investigation of docking energy', Applied Soft Computing Journal, 10(1), pp. 170–182.

Bai, L., Cui, L., Bai, X. and R. Hancock, E. (2018) 'Deep depth-based representations of graphs through deep learning networks', Neurocomputing. Elsevier B.V., (xxxx).

Bajusz, D., Rácz, A. and Héberger, K. (2015) 'Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?', Journal of Cheminformatics. Journal of Cheminformatics, 7(1), pp. 1–13.

Ballester, P. J. (2019) 'Machine learning for molecular modelling in drug design', Biomolecules, 9(6), pp. 10–12.

Banerjee, P., Siramshetty, V. B., Drwal, M. N. and Preissner, R. (2016) 'Computational methods for prediction of in vitro effects of new chemical structures', Journal of Cheminformatics. Springer International Publishing, 8(1), pp. 1–11.

Barakat, K. (2014) 'Computer-Aided Drug Design', Journal of Pharmaceutical Care & Health Systems, 1(4), pp. 1–2.

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Bouchard, N., Warde-Farley, D. and Bengio, Y. (2012) 'Theano: new features and speed improvements', pp. 1–10.

Bender, A., Jenkins, J. L., Scheiber, J., Sukuru, S. C. K., Glick, M. and Davies, J. W. (2009) 'How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space', Journal of Chemical Information and Modeling, 49(1), pp. 108–119.

Bengio, Y. (2009) Learning Deep Architectures for AI, Foundations and Trends® in Machine Learning.

Benigni, R., Giuliani, A., Franke, R. and Gruska, A. (2000) 'Quantitative structure-activity relationships of mutagenic and carcinogenic aromatic amines', Chemical Reviews, 100(10), pp. 3697–3714.

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Krüger, F. A., Light, Y., Mak, L., McGlinchey, S. and others (2014) 'The ChEMBL bioactivity database: an update', Nucleic acids research. Oxford University Press, 42(D1), pp. D1083--D1090.

Bobach, C., Böhme, T., Laube, U., Püschel, a. and Weber, L. (2012) 'Automated compound classification using a chemical ontology', Journal of Cheminformatics, 4(12), pp. 1–12.

Bolis, G., Di Pace, L. and Fabrocini, F. (1991) 'A machine learning approach to computer-aided molecular design.', Journal of computer-aided molecular design, 5(6), pp. 617–628.

Booth, B. and Zemmel, R. (2004) 'Prospects for productivity', Nature Reviews Drug Discovery, 3(5), pp. 451–456.

Bugmann, G. (1998) 'Normalized Gaussian radial basis function networks', Neurocomputing, 20(1–3), pp. 97–110.

Bullins, B., Princeton, G. A., Hazan, E., Kalai, A., Research Roi Livni, M., Garivier, A. and Kale, S. (2019) 'Generalize Across Tasks: Efficient Algorithms for Linear Representation Learning', Proceedings of Machine Learning Research, 98, pp. 1–12.

Burden, F. R. (1996) 'Using Artificial Neural Networks to Predict Biological Activity from Simple Molecular Structural Considerations', Quantitative Structure-Activity Relationships. WILEY-VCH Verlag, 15(1), pp. 7–11.

Carbó-Dorca, R. and Mezey, P. G. (2013) Fundamentals of molecular similarity. Springer Science & Business Media.

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(August), pp. 58–63.

Chen, Yu-Hsin and Krishna, Tushar and Emer, Joel and Sze, Vivienne (2016) 'Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks', in IEEE International Solid-State Circuits Conference, ISSCC 2016, Digest of Technical Papers, pp. 262–263.

Cheng, Y., Wang, F., Zhang, P. and Hu, J. (2016) 'Risk Prediction with Electronic Health Records: A Deep Learning Approach', Proceedings of the 2016 SIAM International Conference on Data Mining, pp. 432–440.

CHIH-CHUNG, C. (2011) 'LIBSVM : A library for support vector machines', ACM Transactions on Intelligent Systems and Technology, 2, pp. 27:1-27:27.

Cho, S. J., Shen, C. F. and Hermsmeier, M. a. (2000) 'Binary Formal Inference-Based Recursive Modeling Using Multiple Atom and Physicochemical Property Class Pair and Torsion Descriptors as Decision Criteria', Journal of Chemical Information and Modeling, 40(3), pp. 668–680.

Chollet, F. (2015) 'Keras Documentation', Keras.Io.

Cramer, R. D., Redl, G. and Berkoff, C. E. (1974) 'Substructural Analysis. Novel Approach to the Problem of Drug Design', Journal of Medicinal Chemistry, 17(5), pp. 533–535.

Cybenko, G. (1989) 'Degree of approximation by superpositions of a sigmoidal function', Mathematics of control, signals and systems, 9(3), pp. 303–314.

Dahl, G. E., Jaitly, N. and Salakhutdinov, R. (2014) 'Multi-task Neural Networks for QSAR Predictions', pp. 1–21.

Dahl, G. E., Sainath, T. N. and Hinton, G. E. (2013) 'Improving Deep Neural Networks for {LVCSR} Using Rectified Linear Units and Dropout', IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8609–8613.

Das, A., Ghosh, S., Sarkhel, R., Choudhuri, S., Das, N. and Nasipuri, M. (2019) 'Combining Multilevel Contexts of Superpixel Using Convolutional Neural Networks to Perform Natural Scene Labeling', Advances in Intelligent Systems and Computing, 740, pp. 297–306.

David S, T. and Dean A, P. (1989) ‘What’S Hidden in the Hidden Layer’, In Depth Neural Networks .Byte 14.8, pp. 227–233.

Ding, H., Takigawa, I., Mamitsuka, H. and Zhu, S. (2013) ‘Similarity-based machine learning methods for predicting drug-target interactions: a brief review.’, Briefings in bioinformatics, 15(5), pp. bbt056-.

Dolz, J., Desrosiers, C. and Ayed, I. Ben (2018) ‘IVD-Net: Intervertebral disc localization and segmentation in MRI with a multi-modal UNet’, Proceedings ofthe MICCAI 2018 IVD Challenge, (Ivd), pp. 1–7.

Druzhilovskiy, D. S., Rudik, a. V., Filimonov, D. a., Gloriozova, T. a., Lagunin, a. a., Dmitriev, a. V., Pogodin, P. V., Dubovskaya, V. I., Ivanov, S. M., Tarasova, O. a., Bezhentsev, V. M., Murtazalieva, K. a., Semin, M. I., Maiorov, I. S., Gaur, a. S., Sastry, G. N. and Poroikov, V. V. (2017) ‘Computational platform Way2Drug: from the prediction of biological activity to drug repurposing’, Russian Chemical Bulletin, 66(10), pp. 1832–1841.

Ertl, P. (2017) ‘An algorithm to identify functional groups in organic molecules’, Journal of Cheminformatics. Springer International Publishing, 9(1), pp. 1–7.

Feldman, H. J., Dumontier, M., Ling, S., Haider, N. and Hogue, C. W. V. (2005) ‘CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules’, FEBS Letters, 579(21), pp. 4685–4691.

Fernández-De Gortari, E., García-Jacas, C. R., Martinez-Mayorga, K. and Medina-Franco, J. L. (2017) ‘Database fingerprint (DFP): an approach to represent molecular databases’, Journal of Cheminformatics. Springer International Publishing, 9(1), pp. 1–9.

Filimonov, D. a, Lagunin, a a, Gloriozova, T. a, Rudik, a V, Druzhilovskii, D. S., Pogodin, P. V and Poroikov, V. V (2014) ‘Prediction of the biological activity spectra of organic compounds uing the PASSonline website resource’, Chemistry of Heterocyclic Compounds, 50(3), pp. 444–457.

Free, S. M. and Wilson, J. W. (1964) ‘a Mathematical Contribution To Structure-Activity Studies.’, Journal of medicinal chemistry, 7(4), pp. 395–399.

Gao, H., Williams, C., Labute, P. and Bajorath, J. (1999) ‘Binary quantitative structure-activity relationship (QSAR) analysis of estrogen receptor ligands.’, Journal of chemical information and computer sciences, 39(1), pp. 164–8.

Gatys, L. a, Ecker, A. S. and Bethge, M. (2015) ‘A Neural Algorithm of Artistic Style’, arXiv preprint, pp. 1–16.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. and Overington, J. P. (2012) 'ChEMBL: A large-scale bioactivity database for drug discovery', Nucleic Acids Research, 40(D1), pp. 1100–1107.

Ghasemi, F., Mehridehnavi, A., Fassihi, A. and Pérez-Sánchez, H. (2018) 'Deep neural network in QSAR studies using deep belief network', Applied Soft Computing Journal. Elsevier B.V., 62, pp. 251–258.

Goodfellow, I., Bengio, Y. and Courville, A. (2016) 'Book Review: Deep Learning', Deep Learning, 22(4), pp. 351–354.

Gupta, A., Wang, H. and Ganapathiraju, M. (2015) 'Learning structure in gene expression data using deep architectures, with an application to gene clustering', 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1328–1335.

GUPTA, V. (2017) Image Classification using Convolutional Neural Networks in Keras.

Handa, K., Nakagome, I., Yamaotsu, N., Gouda, H. and Hirono, S. (2013) 'Three-dimensional quantitative structure-activity relationship analysis of inhibitors of human and rat cytochrome P4503A enzymes', Drug Metabolism and Pharmacokinetics, 28(4), pp. 345–355.

Hansch, C. (1969) 'A Quantitative Approach to Biochemical Structure-Activity Relationships', Chemical Research, 2(4), pp. 232–239.

Hansen, K., Mika, S., Schroeter, T., Sutter, A., Laak, A. Ter, Thomas, S. H., Heinrich, N. and Müller, K. R. (2009) 'Benchmark data set for in silico prediction of Ames mutagenicity', Journal of Chemical Information and Modeling, 49(9), pp. 2077–2081.

He, M., Yang, Q., Norvil, A., Sherris, D. and Gowher, H. (2018) 'Characterization of Small Molecules Inhibiting the Pro-Angiogenic Activity of the Zinc Finger Transcription Factor Vezf1', Molecules, 23(7), p. 15.

Helma, C., Cramer, T., Kramer, S. and De Raedt, L. (2004) 'Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds', Journal of Chemical Information and Computer Sciences, 44(4), pp. 1402–1411.

Hentabli, H., Naomie, S. and Saeed, F. (2016) 'AN ACTIVITY PREDICTION MODEL USING SHAPE-BASED DESCRIPTOR METHOD', Jurnal Teknologi, 1, pp. 1–8.

Hentabli, H., Saeed, F., Abdo, A. and Salim, N. (2014) 'A new graph-based molecular descriptor using the canonical representation of the molecule', Scientific World Journal, 2014.

Hentabli, H., Salim, N., Abdo, A. and Saeed, F. (2012) 'LWDOSM : Language for Writing Descriptors', Advanced Machine Learning Technologies and Applications. Springer Berlin Heidelberg, pp. 247–256.

Hentabli, H., Salim, N., Abdo, A. and Saeed, F. (2013) 'LINGO-DOSM : LINGO for Descriptors of Outline', Intelligent Information and Database Systems. Springer Berlin Heidelberg, pp. 315–324.

Himmat, M., Salim, N., Al-Dabbagh, M. M. and Ahmed, A. (2015) 'An algorithm for similarity-based virtual screening.', Journal of Chemical and Pharmaceutical Research, 7(4), pp. 974–979.

Himmat, M., Salim, N., Al-Dabbagh, M. M., Saeed, F. and Ahmed, A. (2015) 'Data mining and fusion methods in ligand-based virtual screening', Journal of Chemical and Pharmaceutical Sciences, 8(4), pp. 964–969.

Joereskog, K. G. and Wold, H. O. A. (1982) Systems under indirect observation: Causality, structure, prediction. North Holland.

John, G. H. and Langley, P. (2013) 'Estimating Continuous Distributions in Bayesian Classifiers', pp. 338–345.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L. (2014) 'Large-scale Video Classification with Convolutional Neural Networks', in Proceedings of International Computer Vision and Pattern Recognition (CVPR 2014).

Kawai, K., Fujishima, S. and Takahashi, Y. (2008) 'Predictive Activity Profiling of Drugs by Topological-Fragment-Spectra-Based Support Vector Machines', Journal of chemical information and modeling, 48(6), pp. 1152–1160.

Kazius, J., McGuire, R. and Bursi, R. (2005) 'Derivation and validation of toxicophores for muta- genicity prediction', J. Med. Chem, 48, pp. 312–320.

King, R. D. and Srinivasan, a (1997) 'The discovery of indicator variables for QSAR using inductive logic programming.', Journal of computer-aided molecular design, 11(6), pp. 571–80.

Kowalski, B. R. (2013) Chemometrics: mathematics and statistics in chemistry. Springer Science & Business Media.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012) 'ImageNet Classification with Deep Convolutional Neural Networks', Advances In Neural Information Processing Systems, pp. 1–9.

Kumari, M., Tiwari, N., Chandra, S. and Subbarao, N. (2018) 'Comparative analysis of machine learning based QSAR models and molecular docking studies to screen potential anti-tubercular inhibitors against InhA of mycobacterium tuberculosis', International Journal of Computational Biology and Drug Design, 11(3), pp. 209–235.

LeCun, Y., Bottou, L., Orr, G. B. and Müller, K.-R. (1998) 'Effiicient BackProp', in Neural Networks: Tricks of the Trade, this book is an outgrowth of a 1996 NIPS workshop, pp. 9–50.

LeCun, Y., Yoshua, B. and Geoffrey, H. (2015) 'Deep learning', Nature, 521(7553), pp. 436–444.

Legendre, P. (2005) 'Species associations: The Kendall coefficient of concordance revisited', Journal of Agricultural, Biological, and Environmental Statistics, 10(2), pp. 226–245.

Lewis, R. a. and Wood, D. (2014) 'Modern 2D QSAR for drug discovery', Wiley Interdisciplinary Reviews: Computational Molecular Science, 4(6), pp. 505–522.

Lo, Y. C., Rensi, S. E., Torng, W. and Altman, R. B. (2018) 'Machine learning in chemoinformatics and drug discovery', Drug Discovery Today. Elsevier Ltd, 23(8), pp. 1538–1546.

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. and Svetnik, V. (2015) 'Deep neural nets as a method for quantitative structure-activity relationships', Journal of Chemical Information and Modeling, 55(2), pp. 263–274.

Martin, Y. C. (1991) 'Computer-assisted rational drug design.', Methods in enzymology, 203, p. 587.

Martin, Y. C., Kofron, J. L. and Traphagen, L. M. (2002) 'Do structurally similar molecules have similar biological activity?', Journal of Medicinal Chemistry, 45(19), pp. 4350–4358.

McCulloch, W. S. and Pitts, W. (1988) 'Neurocomputing: foundations of research', Cambridge, MA, USA, pp. 15–27.

McCulloch, W. S. and Pitts, W. H. (1943) ‘Alogical calculus of the idea immanent in nervous activity’, Bulletin of Mathematical Biophysics, 5, pp. 115–133.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) ‘Efficient Estimation of Word Representations in Vector Space’, pp. 1–12.

Mikolove, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) ‘Distributed Representations of Words and Phrases and their Compositionality’, pp. 1–9.

MLA, J. and Maggiora, G. M. (1990) ‘Concepts and Application of Molecular Similarity’, Wiley Interdisciplinary Reviews-Computational Molecular Science, 50, pp. 376–377.

Nair, V. and Hinton, G. E. (2010) ‘Rectified Linear Units Improve Restricted Boltzmann Machines’, Proceedings of the 27th International Conference on Machine Learning, (3), pp. 807–814.

Ormerod, A., Willett, P. and Bawden, D. (1989) ‘Comparison of Fragment Weighting Schemes for Substructural Analysis’, Quantitative Structure-Activity Relationships. WILEY-VCH Verlag, 8(2), pp. 115–129.

Park, E., Han, X., Berg, T. L. and Berg, A. C. (2016) ‘Combining Multiple Sources of Knowledge in Deep CNNs for Action Recognition’, 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 53, pp. 1–8.

Pérez-Sánchez, H., Cano, G. and García-Rodríguez, J. (2014) ‘Improving drug discovery using hybrid softcomputing methods’, Applied Soft Computing Journal. Elsevier B.V., 20, pp. 119–126.

Pradeep, P., Povinelli, R. J., White, S. and Merrill, S. J. (2016) ‘An ensemble model of QSAR tools for regulatory risk assessment’, Journal of Cheminformatics. Springer International Publishing, 8(1), pp. 1–9.

Qabaja, A., Alshalalfa, M., Alanazi, E. and Alhajj, R. (2014) ‘Prediction of novel drug indications using network driven biological data prioritization and integration’, Journal of Cheminformatics, 6(1), pp. 1–14.

Qiu, Z., Yao, T., Ngo, C.-W., Tian, X. and Mei, T. (2019) ‘Learning Spatio-Temporal Representation with Local and Global Diffusion’, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12056–12065.

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D. and Pande, V. (2015) ‘Massively Multitask Networks for Drug Discovery’, (Icml).

Rogers, D. and Hahn, M. (2010) ‘Extended-connectivity fingerprints’, Journal of chemical information and modeling. ACS Publications, 50(5), pp. 742–754.

Rosenblatt, F. (1958) ‘The perceptron: a probabilistic model for information storage and organization in the brain.’, Psychological review. American Psychological Association, 65(6), p. 386.

Rosenblatt, F. (1961) Principles of neurodynamics. perceptrons and the theory of brain mechanisms.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1988) ‘Neurocomputing: Foundations of research’, ch. Learning Representations by Back-propagating Errors, pp. 696–699.

Rusinko, a, Farmen, M. W., Lambert, C. G., Brown, P. L. and Young, S. S. (1999) ‘Analysis of a large structure/biological activity data set using recursive partitioning.’, Journal of chemical information and computer sciences, 39(6), pp. 1017–26.

Sadowski, J. and Kubinyi, H. (1998) ‘A scoring scheme for discriminating between drugs and nondrugs.’, Journal of medicinal chemistry, 41(18), pp. 3325–9.

Saeed, F. and Salim, N. (2013) ‘Using soft consensus clustering for combining multiple clusterings of chemical structures’, Jurnal Teknologi (Sciences and Engineering), 63(1), pp. 9–11.

Sainath, T., Mohamed, A. R., Kingsbury, B. and Ramabhadran, B. (2013) ‘Deep convolutional neural networks for LVCSR’, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8614–8618.

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), pp. 435–442.

Schreiber, S. L. (2005) ‘Small molecule, the missing link in the central dogma’, NATURE CHEMICAL BIOLOGY, 1(2), pp. 64–66.

Schymanski, E. L., Ruttkies, C., Krauss, M., Brouard, C., Kind, T., Dührkop, K., Allen, F., Vaniya, A., Verdegem, D., Böcker, S., Rousu, J., Shen, H., Tsugawa, H., Sajed, T., Fiehn, O., Ghesquière, B. and Neumann, S. (2017) ‘Critical Assessment of Small Molecule Identification 2016: automated methods’, Journal of Cheminformatics. Springer International Publishing, 9(1), pp. 1–21.

Sci Tegic Accelrys Inc (no date).

Shen, M.-Y., Su, B.-H., Esposito, E. X., Hopfinger, A. J. and Tseng, Y. J. (2011) ‘A Comprehensive SVM Binary hERG Classification Model Based on Extensive

but Biased Endpoint hERG Data Sets.', Chemical research in toxicology, pp. 934–949.

Sidney Siegel (1956) Nonparametric statistics for the behavioral sciences. New York, NY, England: Mcgraw-Hill Book Company.

Siegel, S. and Castellan, N. J. (1988) 'Nonparametric Statistics for the Behavioral Sciences', McGraw-HiU Book Company, New York.

Simões, R. S., Maltarollo, V. G., Oliveira, P. R. and Honorio, K. M. (2018) 'Transfer and multi-task learning in QSAR modeling: Advances and challenges', Frontiers in Pharmacology, 9(FEB), pp. 1–7.

Simonyan, K. and Zisserman, A. (2014) 'Two-Stream Convolutional Networks for Action Recognition in Videos', in Advances in Neural Information Processing Systems 27, pp. 568–576.

Snyder, R. D. and Smith, M. D. (2005) 'Computational prediction of genotoxicity : room for improvement chemical space considerations with associated non-causal activity correlations . REVIEWS', REVIEWSdrug discovery todaye BIOSILICO, 10(16), pp. 1120–1124.

Stevenson, J. M. and Mulready, P. D. (2003) 'Pipeline Pilot 2.1 By Scitegic, 9665 Chesapeake Drive, Suite 401, San Diego, CA 92123-1365.', Journal of the American Chemical Society. American Chemical Society, 125(5), pp. 1437–1438.

Su, B. H., Slien, M. Y., Esposito, E. X., Hopnnger, A. J. and Tseng, Y. J. (2010) 'In silico binary classification QSAR models based on 4D-fingerprints and MOE descriptors for prediction of hERG blockage', Journal of Chemical Information and Modeling, 50(7), pp. 1304–1318.

Su, H., Maji, S., Kalogerakis, E. and Learned-Miller, E. (2015) 'Multi-view Convolutional Neural Networks for 3D Shape Recognition', In Proceedings of the IEEE international conference on computer vision, 1(02), pp. 945–953.

Sutherland dataset (no date).

Sutherland, Jeffrey J, Brien, L. a O. and Weaver, D. F. (2003) 'Spline-Fitting with a Genetic Algorithm : A Method for Developing Classification Structure - Activity Relationships', Journal of Chemical Information and Modeling, pp. 1906–1915.

Sutherland, Jeffrey J., O'Brien, L. a. and Weaver, D. F. (2003) 'Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure-

Activity Relationships', Journal of Chemical Information and Computer Sciences, 43(6), pp. 1906–1915.

Sutherland, J. J., O'Brien, L. a. and Weaver, D. F. (2004) 'A comparison of methods for modeling quantitative structure-activity relationships', Journal of Medicinal Chemistry, 47(22), pp. 5541–5554.

Todeschini, R. and Consonni, V. (2000) 'Handbook of Molecular Descriptors'. John Wiley & Sons.

Unterthiner, T., Mayr, A., Klambauer, G. and Hochreiter, S. (2015) 'Toxicity Prediction using Deep Learning'.

Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Wegner, J. K. and Ceulemans, H. (2014) 'Deep Learning as an Opportunity in Virtual Screening', Deep Learning and Representation Learning Workshop: NIPS 2014, pp. 1–9.

Vaidya, A., Jain, Sourabh, Jain, Shweta, Jain, A. K. and Agrawal, R. K. (2014) 'Quantitative Structure-Activity Relationships: A Novel Approach of Drug Design and Discovery', Journal of Pharmaceutical Sciences and Pharmacology, 1(3), pp. 219–232.

Vane, J. R. (2000) 'The Mechanism of Action of Anti-Inflammatory Drugs', Advances in Eicosanoid Research, pp. 1–23.

Wang, H., Meghawat, A., Morency, L.-P. and Xing, E. P. (2016) 'Select-Additive Learning: Improving Cross-individual Generalization in Multimodal Sentiment Analysis', 1.

Wang, H. and Raj, B. (2017) 'On the Origin of Deep Learning', Arxiv, pp. 1–72.

Wang, X., Chen, H., Yang, F., Gong, J., Li, S., Pei, J., Liu, X., Jiang, H., Lai, L. and Li, H. (2014) 'iDrug: a web-accessible and interactive drug discovery and design platform.', Journal of cheminformatics, 6(1), p. 28.

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Zhou, Z., Han, L., Karapetyan, K., Dracheva, S., Shoemaker, B. A. and others (2011) 'PubChem's BioAssay database', Nucleic acids research. Oxford University Press, 40(D1), pp. D400--D412.

Weininger, D. (1988) 'SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules', Journal of Chemical Information and Computer Sciences, 28(1), pp. 31–36.

Weininger, D., Weininger, A. and Weininger, J. L. (1989) ‘SMILES. 2. Algorithm for Generation of Unique SMILES Notation’, Journal of Chemical Information and Computer Sciences, 29(2), pp. 97–101.

Werbos, P. (1974) Beyond Regression:" New Tools for Prediction and Analysis in the Behavioral Sciences, Ph. D. dissertation, Harvard University.

Willett, P. (2006) ‘Similarity-based virtual screening using 2D fingerprints’, Drug Discovery Today, 11(23–24), pp. 1046–1053.

Willett, P., Barnard, J. M. and Downs, G. M. (1998) ‘Chemical similarity searching’, Journal of Chemical Information and Computer Sciences, 38(6), pp. 983–996.

Willett, P., Wilton, D., Hartzoulakis, B., Tang, R., Ford, J. and Madge, D. (2007) ‘Prediction of ion channel activity using binary kernel discrimination’, Journal of Chemical Information and Modeling, 47(5), pp. 1961–1966.

Williams, J., Comanescu, R., Radu, O. and Tian, L. (2018) ‘DNN Multimodal Fusion Techniques for Predicting Video Sentiment’, In Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), pp. 64–72.

Winkler, D. a and Burden, F. R. (2002) ‘Application of neural networks to large dataset QSAR, virtual screening, and library design.’, Methods in molecular biology (Clifton, N.J.), 201, pp. 325–367.

Witten, I. H., Frank, E., Hall, M. A. and Pal, C. J. (2016) Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Wu, F. X. and Li, M. (2019) ‘Deep learning for biological/clinical data’, Neurocomputing. Elsevier B.V., 324, pp. 1–2.

Xia, X., Maliski, E. G., Gallant, P. and Rogers, D. (2004) ‘Classification of kinase inhibitors using a Bayesian model’, J.Med.Chem., 47, pp. 4463–4470.

Yuan, Y., Xun, G., Suo, Q., Jia, K. and Zhang, A. (2019) ‘Wave2Vec: Deep representation learning for clinical temporal data’, Neurocomputing. Elsevier B.V., 324, pp. 31–42.

Zeiler, M. D. (2012) ‘ADADELTA: An Adaptive Learning Rate Method’.

LIST OF PUBLICATIONS

Hamza, Hentabli, Naomie Salim, and Faisal Saeed. "An activity prediction model using shape-based descriptor method." *Jurnal Teknologi* 78, no. 6-12 (2016).

Hamza, Hentabli, Naomie Salim, and Faisal Saeed. "Quantitative Structure Activity Relationships In Computer Aided Molecular Design." *Jurnal Teknologi* 78, no. 9-3 (2016).

Hentabli, Hamza, Faisal Saeed, Ammar Abdo, and Naomie Salim. "A New Graph-Based Molecular Descriptor Using The Canonical Representation Of The Molecule." *The Scientific World Journal* 2014 (2014).

Hentabli, Hamza, Naomie Salim, Ammar Abdo, and Faisal Saeed. "LINGO-DOSM: LINGO for descriptors of outline shape of molecules." In *Asian Conference on Intelligent Information and Database Systems*, pp. 315-324. Springer, Berlin, Heidelberg, 2013.

Hentabli, Hamza, Naomie Salim, Ammar Abdo, and Faisal Saeed. "LWDOSM: language for writing descriptors of outline shape of molecules." In *International Conference on Advanced Machine Learning Technologies and Applications*, pp. 247-256. Springer, Berlin, Heidelberg, 2012.

Hentabli, Hamza, Naomie Salim, 2018, PATENT, *New Molecular representation based on Small Molecules for Bioactivity Prediction*, IP/PT/2018/0753, APPROVED.

Hentabli, Hamza, Naomie Salim, 2018, PATENT, *MeramalNet: A Deep Learning Convolutional Neural Network for Bioactivity Prediction in Structure-Based Drug Discovery*, IP/PT/2018/0757, APPROVED.

Hentabli, Hamza, Naomie Salim, 2018, PATENT, *Mol2Matrix : Matrix for Molecular representation based on toxicophores*, IP/PT/2018/0755, APPROVED.