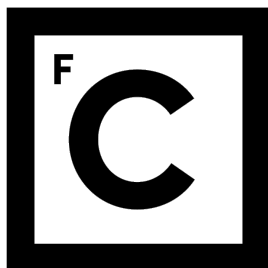


UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS



Ciências
ULisboa

**Computational approaches to virtual screening in human central nervous system
therapeutic targets**

“Documento Definitivo”

Doutoramento em Biologia

Especialidade de Biologia de Sistemas

Samina Kausar

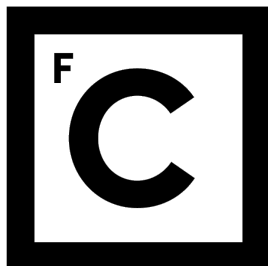
Tese orientada por:

Prof. Doutor Andre Falcao e Prof. Doutora Rita C. Guedes

Documento especialmente elaborado para obtenção do grau de doutor

2019

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS



**Ciências
ULisboa**

**Computational approaches to virtual screening in human central nervous system
therapeutic targets**

Doutoramento em Biologia

Especialidade de Biologia de Sistemas

Samina Kausar

Tese orientada por:

Prof. Doutor Andre Falcao e Prof. Doutora Rita C. Guedes

Júri

Presidente:

- Doutor Rui Manuel dos Santos Malhó, Professor Catedrático e Presidente do Departamento de Biologia Vegetal da Faculdade de Ciências da Universidade de Lisboa.

Vogais:

- Doutor João Montargil Aires de Sousa, Professor Auxiliar com Agregação, Faculdade de Ciências da Universidade Nova de Lisboa;
- Doutor Cláudio Manuel Soares, Professor Associado, Instituto de Tecnologia Química e Biológica Antonio Xavier da Universidade Nova de Lisboa;
- Doutor Daniel Pedro de Jesus Faria, Investigador Pós Doutoramento, Instituto Gulbenkian de Ciência;
- Doutor André Osório e Cruz de Azeredo Falcão, Professor Auxiliar, Faculdade de Ciências da Universidade de Lisboa (orientador);
- Doutora Cátia Luísa Santana Calisto Pesquita, Professor Auxiliar, Faculdade de Ciências da Universidade de Lisboa;

Financiado pela Fundação para a Ciência e Tecnologia (SFRH/BD/106083/2015)

Documento especialmente elaborado para obtenção do grau de doutor

2019

Dedication

I dedicate this dissertation to my mother.

Colours of dark grey and black fill the world in which I live
No other feeling could possibly be worse than this
Where once was a room filled with laughter and cheer
Now stands loneliness, emptiness, and despair.

No one could ever fathom how wretchedly my heart aches
And how I greatly regret that I lost a last chance to meet you.
If I could go back in time, and change only one wrong that I've done
I'd go back to the hour, to the second, on the day I lost you.

May Allah (SWT) have mercy on her and on us (Amin). May we all join her in Jannah Firdaws Insha'Allah to restart an everlasting life without any fear of loss of my dearest ones.

Acknowledgements

All praises for **Almighty God**, Lord of the worlds, the merciful and compassionate, who has created a fascinating world and gifted human beings with knowledge and wisdom to explore the closed domains and veiled legends of the nature. His help and bounteous blessing enabled me to perceive and pursue higher ideas of life. All praises and respect are for my ideal and role model the **Holy Prophet Hazrat Muhammad (Sallallah-o-Allah-e-Waallehe Wasalam)**, who taught me to see beyond the illusionary world and enlightened every atom of my body with the love of my creator and humanity. He is the greatest social reformer, an eternal source of knowledge and torch of guidance in both materialistic and mystical worlds.

I gratefully acknowledge the Fundação para a Ciência e Tecnologia for a doctoral Grant (SFRH/BD/111654/2015), MIMED Project Funding (PTDC/EEI-ESS/4923/2014) and the LASIGE Research Unit, ref. UID/CEC/00408/2019 for providing the infrastructure.

No acknowledgment could express my heartiest gratitude and deep sense of devotion to my admirable, cheerful, and respectable supervisor, **Dr. Andre Falcao**. His enthusiastic guidance, kind attitude and continuous encouragement at every step during this whole PhD enabled me to achieve my goal. I am fortunate to have a chance to say him; you are not only a wonderful teacher but also a great source of inspiration to others. The tough problem is not in identifying winners: it is in making winners out of ordinary people. Thank you for doing such a tough job; thanks for your patience and great efforts you put into training me in the scientific field. I wish you everlasting happiness, deep contentment, peace of mind, and the special warmth of giving and sharing.

With a deep sense of honor, I also wish to express my gratitude to my co-advisor **Dr. Rita Guedes**. Thanks for your warm encouragement, beneficial suggestions and supervision for refining my concepts and improving my work in the last and important phase of my research work to make this thesis possible. There is nothing I can do in return of her sincere guidance except best wishes. May her heart and home fill with love, joy, patience, and understanding.

I am also grateful to my thesis advisory committee members **Dr. Isabel A.C.** and **Dr. Daniel Faria** for their valuable feedback and suggestions during my qualification step.

I would like to give my gratitude to my dearest friends **André Lamúrias**, **Ana Rita** and **Catarina Ventura** for the time with laughter, mutual encouragement, love and care. I would also like to express my sincere thanks to my all friends in **Molecular Modeling Group of RitaGuedes Lab** and my **PhD colleagues** for sharing their time and moral support.

I feel pleasure in transcribing my whole hearted thanks to my loving brothers, **Dr. Fayyaz Ahmad** and **Dr. Shamshad Ahmad** for their caring attitude and father-like support. They both are a motivation and inspiration when I think of my goals and future.

I owe great debt of prayers and very sincere thanks to my all sisters, my dearest younger brother and family-in-law for their prayers, affection and cooperation that can never be paid back. They are a great gift of God, the special part of my life. I am incomplete without their love and warmth. I do prayer for them with my every breath.

My proud, my husband: I extend my deepest appreciation and sincerest thanks to my best friend and soul mate, **Dr. Muhammad Asif**. A motivational husband is like a blessing to a wife. He always motivated me to overcome my fear and to face the world with confidence. I respect and appreciate him for never trying to push me back from what I desire in my life. Thanks for adding wings to my dreams, helping me to achieve my goals and say “Good Bye” to nightmares.

Words are lacking to express my deep love and respect to my affectionate **parents**, for their love and support, encouragement, interest and unceasing prayers. They are like a guiding star that showed me the path to success and when I triumph, they cherished my achievements. I never thought when I will be writing this part of acknowledgement, I will be missing my beloved mother who has gone forever away from my loving eyes and just left a void that can never fill. Though her life was short, I will make sure her memory lives on at each step as long as I shall live. May God give long happy and healthy life to my father and rest my mother in peace forever. Amin

Samina Kausar

Abstract

In the past several years of drug design, advanced high-throughput synthetic and analytical chemical technologies are continuously producing a large number of compounds. These large collections of chemical structures have resulted in many public and commercial molecular databases. Thus, the availability of larger data sets provided the opportunity for developing new knowledge mining or virtual screening (VS) methods. Therefore, this research work is motivated by the fact that one of the main interests in the modern drug discovery process is the development of new methods to predict compounds with large therapeutic profiles (multi-targeting activity), which is essential for the discovery of novel drug candidates against complex multifactorial diseases like central nervous system (CNS) disorders. This work aims to advance VS approaches by providing a deeper understanding of the relationship between chemical structure and pharmacological properties and design new fast and robust tools for drug designing against different targets/pathways.

To accomplish the defined goals, the first challenge is dealing with big data set of diverse molecular structures to derive a correlation between structures and activity. However, an extendable and a customizable fully automated *in-silico* Quantitative-Structure Activity Relationship (QSAR) modeling framework was developed in the first phase of this work. QSAR models are computationally fast and powerful tool to screen huge databases of compounds to determine the biological properties of chemical molecules based on their chemical structure. The generated framework reliably implemented a full QSAR modeling pipeline from data preparation to model building and validation. The main distinctive features of the designed framework include a) efficient data curation b) prior estimation of data modelability and, c) an-optimized variable selection methodology that was able to identify the most biologically relevant features responsible for compound activity. Since the underlying principle in QSAR modeling is the assumption that the structures of molecules are mainly responsible for their pharmacological activity, the accuracy of different structural representation approaches to decode molecular structural information largely influence model predictability. However, to find the best approach in QSAR modeling, a comparative analysis of two main categories of molecular representations that included descriptor-based (vector space) and distance-based (metric space) methods was carried out. Results obtained from five QSAR data sets showed that distance-based method was

superior to capture the more relevant structural elements for the accurate characterization of molecular properties in highly diverse data sets (remote chemical space regions). This finding further assisted to the development of a novel tool for molecular space visualization to increase the understanding of structure-activity relationships (SAR) in drug discovery projects by exploring the diversity of large heterogeneous chemical data. In the proposed visual approach, four non-linear DR methods were tested to represent molecules lower dimensionality (2D projected space) on which a non-parametric 2D kernel density estimation (KDE) was applied to map the most likely activity regions (activity surfaces). The analysis of the produced probabilistic surface of molecular activities (PSMAs) from the four datasets showed that these maps have both descriptive and predictive power, thus can be used as a spatial classification model, a tool to perform VS using only structural similarity of molecules.

The above QSAR modeling approach was complemented with molecular docking, an approach that predicts the best mode of drug-target interaction. Both approaches were integrated to develop a rational and re-usable polypharmacology-based VS pipeline with improved hits identification rate. For the validation of the developed pipeline, a dual-targeting drug designing model against Parkinson's disease (PD) was derived to identify novel inhibitors for improving the motor functions of PD patients by enhancing the bioavailability of dopamine and avoiding neurotoxicity. The proposed approach can easily be extended to more complex multi-targeting disease models containing several targets and anti/off-targets to achieve increased efficacy and reduced toxicity in multifactorial diseases like CNS disorders and cancer.

This thesis addresses several issues of cheminformatics methods (e.g., molecular structures representation, machine learning, and molecular similarity analysis) to improve and design new computational approaches used in chemical data mining. Moreover, an integrative drug-designing pipeline is designed to improve polypharmacology-based VS approach. This presented methodology can identify the most promising multi-targeting candidates for experimental validation of drug-targets network at the systems biology level in the drug discovery process.

Keywords: Virtual screening; Systems biology; Polypharmacology; Systems pharmacology; Machine Learning, QSAR modeling, Chemical space visualization, Molecular docking

Resumo

Triagem virtual (VS – de “Virtual Screening”) refere-se a uma variedade de métodos in-silico que servem como análogos computacionais de triagem biológica de alto rendimento no processo moderno de descoberta de fármacos. O objetivo de VS é pesquisar em grandes bases de dados de pequenas moléculas para selecionar as estruturas químicas que possuem maior probabilidade de atividade em testes biológicos num programa de descoberta principal. Assim, o VS pode reduzir o custo total do desenvolvimento de fármacos, selecionando números razoáveis de moléculas candidatas/principais. Uma molécula principal é um composto que possui as propriedades farmacológicas necessárias e, portanto, é usado como um bom ponto de partida para a descoberta de fármacos. O processo de VS depende da disponibilidade e da quantidade de informações estruturais e bioatividade de outros compostos sobre os mesmos alvos.

Nos últimos anos o processo de design de fármacos tem usado tecnologias químicas sintéticas e analíticas avançadas de alto rendimento que estão continuamente uma grande quantidade de informação sobre a interação entre alvos terapêuticos e moléculas candidatas. Simultaneamente o grande desenvolvimento das técnicas de síntese orgânica tem potenciado um aumento exponencial das estruturas conhecidas. Estas grandes coleções de estruturas químicas resultam em muitos conjuntos de dados moleculares públicos e comerciais. Assim, a disponibilidade de conjuntos de dados maiores levou à oportunidade de desenvolver novos métodos de prospecção de dados usando técnicas de aprendizagem automática em VS. Este trabalho de investigação é motivado pelo facto que um dos principais interesses no processo moderno de descoberta de fármacos é o desenvolvimento de novos métodos para prever compostos com perfis terapêuticos com atividade multi-alvo, essencial. Este é um factor importante na descoberta de novos candidatos a fármacos contra doenças multifatoriais complexas, como distúrbios do sistema nervoso central (SNC). Este trabalho visa avançar abordagens de VS, proporcionando uma compreensão mais profunda da relação entre estrutura química e propriedades farmacológicas e desenhar novas ferramentas rápidas e robustas para design de fármacos para alvos terapêuticos distintos.

Para atingir os objetivos definidos, o primeiro desafio é lidar com um grande conjunto de dados de diferentes estruturas moleculares para encontrar correlações entre estruturas e atividades. Para tal, no decorrer deste trabalho uma framework automatizada de modelação extensível e personalizável in-silico para

modelação da Relação de Atividade de Estrutura Quantitativa (QSAR “Quantitative structure Activity Relationship”) foi desenvolvida. O objetivo principal das ferramentas baseadas em modelos QSAR é o desenvolvimento de modelos de aprendizagem automática sobre bases de dados de pequenas moléculas devidamente anotadas com informação sobre a sua actividade, para posteriormente encontrar compostos promissores com os efeitos biológicos desejados, executando esses modelos em grandes repositórios de estruturas moleculares. A modelação QSAR é uma aplicação de abordagens de aprendizagem automática que é cada vez mais utilizada pelas maiores empresas do sector farmacêutico, sobretudo nas fases iniciais do desenvolvimento de fármacos. A aprendizagem automática no processo de descoberta de fármacos é tipicamente usada para produzir modelos QSAR robustos, capazes de prever com confiança a atividade farmacológica de compostos com base nas suas informações estruturais moleculares, assumindo uma forte correlação entre estruturas e atividade biológica. Assim, os modelos QSAR associam quantitativamente a atividade biológica de moléculas (ligandos) com a sua estrutura tipificadas sob a forma de características químicas e propriedades moleculares. A framework gerada implementou de forma fiável um processo completo de modelação QSAR, desde a obtenção e preparação de dados até a construção e validação dos modelos. As principais características distintivas da estrutura projetada incluem: a) curação eficiente de dados; b) estimativa prévia da modelabilidade dos dados; c) metodologia otimizada de seleção de variáveis que foi capaz de identificar as características biologicamente mais relevantes responsáveis pela atividade de compostos. O desempenho da metodologia implementada de modelação QSAR foi testado em trinta conjuntos de dados de diferentes alvos terapêuticos do SNC. A análise dos resultados obtidos mostrou que o procedimento de seleção de variáveis desenvolvidas no fluxo de trabalho de modelação QSAR automatizado foi capaz de remover 62-99 % de dados redundantes e procedeu de forma consistente com conjuntos de dados de alta dimensão (1141 preditores). A seleção da melhor representação molecular para decodificar eficientemente as informações das estruturas moleculares em formatos legíveis por computador ainda é uma tarefa desafiante na informática. A representação numérica de estruturas é utilizada como matrizes de dados de entrada para modelar e compreender relações quantitativas entre estruturas e atividade biológica na modelação de QSAR. Para encontrar a melhor abordagem na modelação QSAR, foi realizada uma análise comparativa de duas categorias principais de representações moleculares que incluíram métodos baseados em descritores (espaço vetorial) e baseados em distância (espaço métrico). Uma representação de espaço vetorial ou espaço linear ocorre quando o conjunto de instâncias de modelação é representado como um vetor, com suas características medidas em relação a algum referencial e, portanto, há uma noção de magnitude e direção a partir da origem. Na maioria dos estudos de modelação de QSAR,

o espaço vetorial é a representação mais usada, onde cada estrutura química é traduzida usando um conjunto de descritores moleculares. Isso geralmente é chamado de 'espaço dos descritores moleculares', que representa diferentes características / propriedades estruturais. A representação do espaço métrico, por outro lado, é construída usando as distâncias medidas entre um conjunto de instâncias que queremos modelar, usando qualquer métrica.

Os resultados obtidos a partir de cinco conjuntos de dados QSAR mostraram que o método baseado em distâncias foi superior para capturar os elementos estruturais mais relevantes para a caracterização precisa de propriedades moleculares em conjuntos de dados altamente diversificados (regiões de espaço químico remoto). A descoberta a partir de análise comparativa de representações moleculares ajudou ainda no desenvolvimento de uma nova ferramenta para a visualização do espaço molecular para aumentar a compreensão das relações estrutura-atividade em projetos de descoberta de drogas, explorando a diversidade de grandes dados químicos heterogêneos. Moléculas dentro de espaços químicos teóricos/conceituais de alta dimensão são consideradas objetos e as distâncias entre elas são usadas para extrapolar atividade ou propriedades biológicas. O tamanho do espaço químico é enorme e não tem um número bem definido. Uma pequena fração de espaço químico indefinido, variando de milhares a milhões de compostos, está disponível em pequenas bases de dados moleculares que são usados para explorar e visualizar a complexidade das estruturas químicas durante o processo de drug development. Os métodos de visualização do espaço químico combinam o conceito de estrutura molecular e similaridade de atividade. No entanto dependem criticamente das representações moleculares e do modo de quantificação da similaridade, que é posteriormente usado para calcular a representação espacial métrica. Na abordagem visual proposta, quatro métodos de projeção de espaços métricos num referencial vetorial em 2 dimensões foram testados para representar uma menor dimensionalidade de moléculas (espaço projetado 2D) em que uma estimativa da densidade de probabilidade não-paramétrica foi aplicada para mapear as regiões de atividade mais prováveis (superfícies de atividade). A análise destas superfícies probabilísticas de atividades moleculares (PSMA de "probabilistic surface map of activity") de quatro conjuntos de dados mostrou que estes mapas possuem poder descritivo e preditivo, podendo ser utilizados como modelos de classificação, como ferramenta para realizar VS utilizando apenas a similaridade estrutural de moléculas sem necessidade de qualquer parametrização ou ajustamento adicional.

A abordagem de modelação QSAR acima referida foi complementada e comparada com a utilização dos métodos docking molecular, uma abordagem que prevê o modo de interação molécula-alvo terapêutico. Ambas as abordagens foram integradas para desenvolver uma pipeline de VS Para validar o

pipeline desenvolvido, um modelo de design de fármacos de duplo alvo para a doença de Parkinson (PD) foi desenvolvido por forma a identificar novos inibidores para melhorar as funções motoras dos pacientes com PD, aumentando a biodisponibilidade de dopamina mas evitando a neurotoxicidade. A abordagem proposta pode ser facilmente adaptada para outros modelos de doença multi-alvo mais complexo, contendo vários alvos e anti-alvos para alcançar maior eficácia e redução da toxicidade em doenças multifatoriais, como outras patologias do SNC ou cancro. Em suma, este trabalho aborda várias questões de métodos de quiminformática (por exemplo, representação de estruturas moleculares, aprendizagem automática e análise de similaridade molecular) para melhorar e projetar novas abordagens computacionais usadas na prospecção de dados químicos. Além disso, um pipeline integrativo de design de fármacos foi projetado para melhorar a abordagem VS baseada em polifarmacologia. A metodologia apresentada pode identificar as moléculas candidatas mais promissoras para múltiplos alvos.

Palavras Chave: Triagem virtual; Biologia de Sistemas; Polifarmacologia; Farmacologia de sistemas; Aprendizado de Máquina; modelagem QSAR; Visualização de espaço químico; docking molecular

Contents

Contents	xxii
List of Figures	xxv
List of Tables	xxvi
List of Abbreviations	xxvii
1 Introduction	1
1.1 Problem statement and the aims of the study	5
1.2 General methodology	7
1.3 Overview of the document	10
1.4 Publications and participation in academic activities	12
1.4.1 Papers in scientific peer-reviewed journals	12
1.4.2 Participation in conferences	13
1.4.3 Participation in academic competitions in science	14
2 Background/state of the art	21
2.1 Virtual screening approaches	21
2.1.1 Cheminformatics applications in virtual screening	24
2.1.1.1 Molecular structure representation/transformation methods	24
2.1.1.2 Molecular similarity concepts	28
2.1.1.3 Quantitative Structure-Activity Relationship (QSAR): machine learning	34
2.1.2 Molecular docking analysis and structure-based virtual screening .	36
2.1.2.1 Target structure analysis and selection for docking	36

CONTENTS

2.1.2.2	Docking software selection and parameters optimization	36
2.1.2.3	Ligands database preparation for virtual screening	37
2.1.3	Large compound repositories for virtual screening	38
3	An automated framework for QSAR model building	51
SAMINA KAUSAR AND ANDRE O FALCAO		
3.1	Introduction	53
3.1.1	Background	53
3.1.2	Objectives	56
3.2	Automated model building	58
3.2.1	Architecture	60
3.2.2	Data access and processing	60
3.2.3	Descriptors calculation	64
3.2.4	Data transformation and data partitioning	64
3.2.5	Data set modelability estimation	65
3.2.6	Feature selection	66
3.2.7	Model building	69
3.2.7.1	Model without feature selection	69
3.2.7.2	Model with feature selection	69
3.2.8	External validation and model applicability domain	70
3.2.9	Extensibility	71
3.3	Results	72
3.3.1	Workflow implementation	72
3.3.1.1	Input data parameters	72
3.3.1.2	Input data set options	73
3.3.1.3	Data set retrieval and data pre-processing	74
3.3.1.4	From data to validated models	75
3.3.2	Real world cases	76
3.3.2.1	Data sets description	76

CONTENTS

3.3.2.2	Data preparation and variable scaling	77
3.3.2.3	Data set modelability measure	79
3.3.2.4	Feature ranking by Random Forest	80
3.3.2.5	Stepwise estimation models and feature selection	80
3.3.2.6	Model results	81
3.3.2.7	Model applicability domain analysis	84
3.3.2.8	Predictive performance comparison with published QSAR model	85
3.4	Discussion	86
3.5	Conclusion	90
4	Analysis and comparison of vector space and metric space representations in QSAR modeling	105
SAMINA KAUSAR AND ANDRE O FALCAO		
4.1	Introduction	107
4.1.1	Molecular similarity and metric space representation	108
4.1.2	Metric spaces vs vector spaces	110
4.2	Methodology	111
4.2.1	Overview of the methodology	111
4.2.2	Vector space representation	112
4.2.2.1	Descriptor based representations	112
4.2.2.2	Fingerprint based representations	113
4.2.3	Metric space representation	114
4.2.3.1	Fingerprint-based similarity	115
4.2.3.2	NAMS-based similarity	116
4.2.4	Model building	116
4.2.4.1	Feature reduction with PCA	118
4.2.4.2	Feature selection with Random Forests	119
4.2.4.3	Support Vector Machine	120
4.2.4.4	Model evaluation and external validation	120

CONTENTS

4.3	Data	121
4.3.1	Data preparation for vector and metric space representations	123
4.4	Results	124
4.4.1	Implementation of analysis	124
4.4.2	Results of generated models	124
4.4.2.1	Is metric space representation as good as the most common vector space based approaches?	126
4.4.2.2	Which similarity representation carries the maximum chemical/structural information content to establish the best relationship between local similarities and activity?	129
4.4.2.3	How effective is using a reduced dimensionality of the metric/vector space with Principal Components, replacing explicit descriptors/fingerprints in QSAR modeling?	129
4.4.2.4	Is there any solution that is globally better on a variety of difficult problems?	130
4.5	Discussion	131
4.5.1	Computation time	134
4.6	Conclusions	135
5	A visual approach for analysis and inference of molecular activity spaces	151
SAMINA KAUSAR AND ANDRE O FALCAO		
5.1	Introduction	153
5.2	Methodology	157
5.2.1	Overview of the methodology	157
5.2.2	Molecular dis/similarity quantification	158
5.2.3	From similarity to distance	160
5.2.4	Dimensionality reduction	162
5.2.4.1	Principal Coordinates Analysis (PCoA)	164
5.2.4.2	Kruskal Multidimensional Scaling (KMDS)	164
5.2.4.3	Sammon mapping	165
5.2.4.4	t-Distributed Stochastic Neighbor Embedding (t-SNE)	166

CONTENTS

5.2.5	Probabilities density estimation	167
5.2.5.1	2D kernel density estimation	168
5.2.6	Defining active probability regions	169
5.2.7	Test set embedding and model validation	171
5.3	Data	173
5.4	Implementation	174
5.5	Results and discussion	175
5.6	Conclusion	180
6	Comparative analysis of QSAR modeling and molecular docking: a rational approach in polypharmacology	193
SAMINA KAUSAR, RITA C. GUEDES AND ANDRE O FALCAO		
6.1	Introduction	193
6.2	Targets selection for Parkinson's disease	196
6.2.1	Dual-targeting of COMT and GSK3B	198
6.2.1.1	Catechol-O-methyltransferase	198
6.2.1.2	Glycogen synthase kinase-3B	200
6.3	Overview of virtual screening methodology	202
6.3.1	QSAR-based virtual screening	204
6.3.2	Molecular docking-based virtual screening	207
6.3.3	Ligands database preparation for screening	212
6.4	Preliminary results and discussion	213
6.4.1	QSAR binary classification models	213
6.4.2	Molecular docking models	214
6.4.3	NCI database screening and hit selection	216
6.4.4	In-vitro validation results of COMT specific inhibitors	220
6.4.5	Future perspectives	222
7	General discussion and conclusions	245
7.1	Contribution	245

CONTENTS

7.1.1	Automation of quantitative structure-activity relationship	246
7.1.2	Molecular structural representation	248
7.1.3	Chemical space visualization	249
7.1.4	Polypharmacology based virtual screening	250
7.2	Limitations and future work	252

List of Figures

1.1	Applications of polypharmacology in systems pharmacology	2
1.2	The growth of the total number of publications in the field of systems biology, systems pharmacology and poly-pharmacology in the PubMed database.	3
1.3	Graphical representation of the thesis objectives and the methodology	8
2.1	Overview of virtual screening approaches	22
2.2	Overview of QSAR modeling	34
3.1	Overview of automated QSAR modeling workflow	59
3.2	Automated QSAR modeling methodology	61
3.3	Input data set options.	62
3.4	Input parameters.	73
3.5	Comparison of models with and without feature selection.	87
3.6	Size of the problems and predictive power of fitted models.	88
3.7	Models over-fitting analysis.	89
3.8	$MODI_{ss}R^2$ versus QSAR_PVE for 30 datasets.	90
4.1	Vector space vs metric space	109
4.2	QSAR modeling approaches	112
4.3	Comparisons of QSAR models's predictive performance using IVS.	126
4.4	Friedman's test results and interquartile ranges of tested models	127
4.5	(A) Boxplots of the three modeling approaches grouped by the different data sets; (B) Groups and interquartile ranges of the medians of tested models from Friedman's test post-hoc analysis	128
4.6	Overall performance of similarity representation using PCA on metric space based QSAR modeling approach	130

LIST OF FIGURES

4.7	Overall performance of metric space representation after removing nearest neighbours in PCA on metric space based QSAR modeling approach	133
5.1	Overview of the methodology.	158
5.2	Distance functions for similarity to distance transformations.	163
5.3	Test set projection over map surface (PSMA) with PCooA.	176
5.4	Test set projection over map surface of selected PSMA with highest performance.	177
5.5	HERG Shepard plot for PCooA, KMDS, SM and t-SNE	179
5.6	Test set projection over 2D probability map of selected models with highest performance.	181
6.1	Pathway model of dual-targeting of COMT and GSK3 β in Parkinson's disease.	199
6.2	Methodology overview of QSAR and molecular docking comparative analysis.	203
6.3	Superposition of docked (red) and experimental poses (green) of (A) GSK3 β protein complex (PDB:1Q41) and (B) COMT protein complex (PDB:3BWY)	215
6.4	Virtual screening results post-processing workflow	217
6.5	Comparative analysis of QSAR and molecular docking based predicted hits.	218
6.6	Mode of the interaction of a ligand (NCI ID: 7434146) (A) GSK3 β protein complex (PDB:1Q41) and (B) COMT protein complex (PDB:3BWY)	221
6.7	<i>In-vitro</i> validation results of COMT specific inhibitors	223

List of Tables

2.1	Types of molecular descriptors according to dimensions	27
2.2	Software programs for calculating different molecular descriptors.	29
2.3	Types of molecular fingerprint used for virtual screening	30
3.1	Description of selected problems	78
3.2	QSAR models based on all descriptors (RDKit descriptors and Morgan fingerprints) datasets.	82
3.3	Comparison of performance of QSAR models (with and without feature selection)	83
4.1	Data set description	122
4.2	Data size before and after removing nearest neighbours - Thr - Similarity threshold; N - New data set size	132
5.1	Data set description	174
5.2	Results on validation set ((*) – best model). Abbreviations: Principal Coordinates Analysis (PCoA), Kruskal Multidimensional Scaling (KMDS), Sammon mapping (SM), and t-Distributed Stochastic Neighbor Embedding (t-SNE)	178
6.1	COMT and GSK3 β QSAR binary classification models results. MCC: Matthews Correlation Criterion	214
6.2	Dual-targeting compounds of COMT and GSK3 β . Red text: QSAR best hits, Green text: Best docked hits	219

List of Abbreviations

VS	Virtual Screening
SVM	Support Vector Machines
RF	Random Forest
DR	Dimension Reduction
CNS	Central Nervous System
QSAR	Quantitative-Structure Activity Relationship
NAMS	Non-contiguous Atom Matching structure Similarity
SAR	Structure-Activity Relationships
KDE	Kernel Density Estimation
PSMA	Probabilistic Surface of Molecular Activity
PD	Parkinson's Disease
PCoA	Principal Coordinates Analysis
KMDS	Kruskal Multidimensional Scaling
SM	Sammon mapping
t-SNE	t-Distributed Stochastic Neighbor Embedding
Tc	Tanimoto coefficient
REACH	Registration, Evaluation, Authorization, and Restriction of Chemicals
OECD	Organization for Economic Co-operation and Development
KNIME	KoNstanz Information MinEr
IVS	Independent Validation Set
MODI	MODeability Index
GA	Genetic Algorithms
VI	Variable Importance
PVE	Proportion of the Variance Explained
RMSE	Root Mean Squared Error
AD	Applicability Domain
NNRTIs	Non-Nucleoside Analogue Reverse-Transcriptase Inhibitors
PCA	Principal Component Analysis
PCs	Principal Components
FS	Feature Selection
OPC-model	Optimized number of PC model
SF-Model	Model having selected number of features
AUC	Area Under Curve
MCC	Matthews Coefficient Correlation
NLM	Non-Linear Mapping

List of Abbreviations

PDF Probability Density Function
KDM Kernel Density Map
ROC Receiver Operating Characteristic
TPR True Positive Rate
FPR False Positive Rate
COMT Catechol-O-MethylTransferase
GSK3B Glycogen Synthase Kinase-3 Beta
L-DOPA L-3,4-Dihydroxyphenylalanine
LID L-DOPA-Induced Dyskinesia
SAM S-adenosylmethionine
DA Dopamine
6-OHDA Neurotoxin 6-hydroxydopamine
ROS Reactive Oxygen Species
SN Substantia Nigra
3-OMD 3-O-methyldopa
CV Cross-Validation
PAINS Pan Assay Interference Compounds
NCI National Cancer Institute

1

Introduction

Today drug design and discovery has moved from the molecular/single target to the target networks (polypharmacology) and systems-biology-oriented level (systems pharmacology)[1]. According to the National Institute of General Medical Sciences “*Systems biology seeks to predict the quantitative behaviour of an in-vivo biological process under realistic perturbation*” [2]. To address the current challenges in biomedical science, systems biology incorporates information from various levels (genomics, transcriptomics, proteomics, metabolomics, and etc.) of biological systems. Instead of considering only local states of a system, the systems biology based approaches also accounts for the complex and highly precise interactions of current state with other explicit systems components. Consequently, systems biology methods combined with experimental validations are being used by researchers to provide a valuable understanding of the multifactorial and complicated cellular mechanisms in complex diseases like central nervous system (CNS) disorders or cancer [3, 4]. The primary purpose of these approaches is the integration of quantitative information about multiple molecular- and cellular-level components (omics data) for investigating biological

1. INTRODUCTION

networks and predicting drug targets and their role in pathophysiology (Figure 1.1) [5, 6, 3, 4, 7].

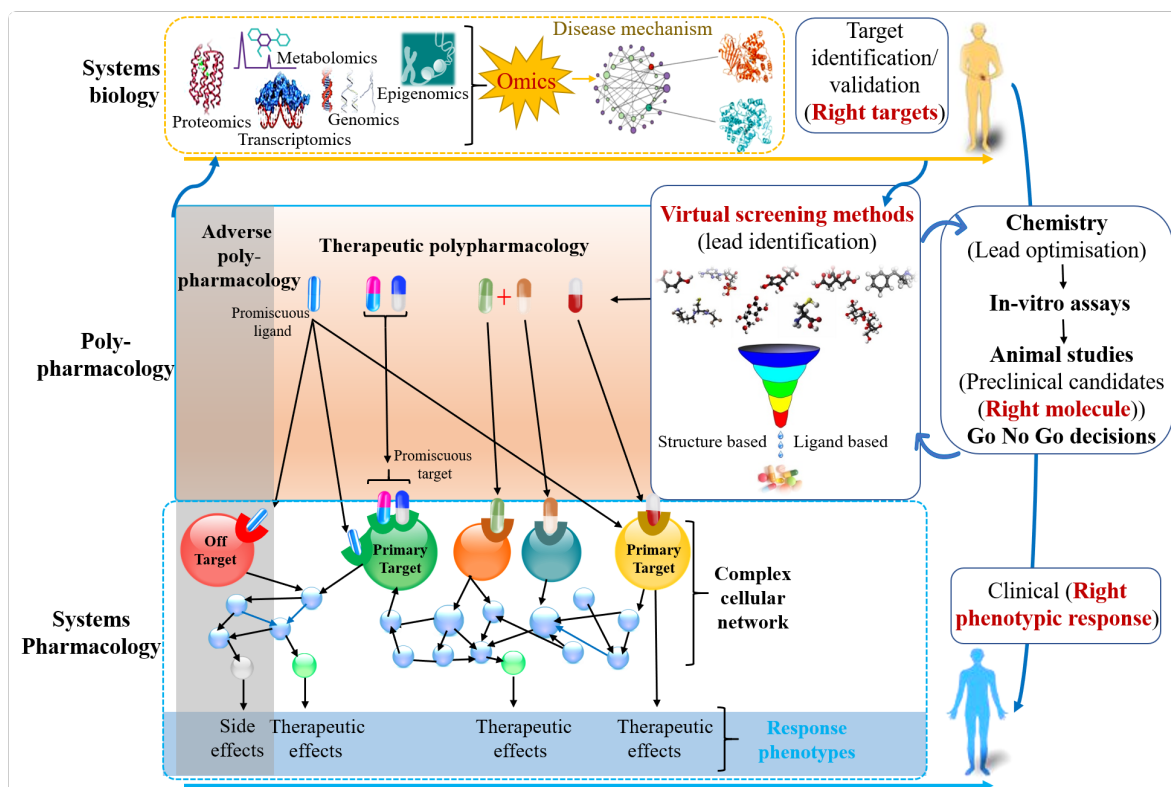


Figure 1.1: Applications of polypharmacology in systems pharmacology

Recent advances in omics technologies and the availability of the resulting data in public databases significantly contributed to increasing success rates in systems biology and its applications in drug discovery or pharmacology [5, 8]. Systems pharmacology inherits methods from systems biology to integrate drug discovery with biological information and then uses this integrated representation to elucidate the drug action mechanism. (Figure 1.1).

In the new systems pharmacology paradigm, to understand the complex binding profile of drugs molecules at the network level, the traditional single-target drug discovery process is being replaced with multi-target drug designing (polypharmacology) to predict the promis-

cuous (multi-targeting) behaviour of drugs which provide an opportunity not only to discover new uses for already known compounds, but also to increase the efficacy of already known drugs, and avoiding affinity to related off-targets (Figure 1.1). Thus, the identification of new drug-target interactions appears as the key to finding new targets for old drugs and new drug candidates for known targets [9, 10, 11].

Systems pharmacology that combines methods from systems biology and poly pharmacology has become the current state-of-the-art method for drug designing. The figure 1.2, plotted by considering the number of publications for “systems biology”, “poly-pharmacology”, and “systems pharmacology” terms that were searched in PubMed clearly shows a paradigm shift towards systems pharmacology for drug discovery.

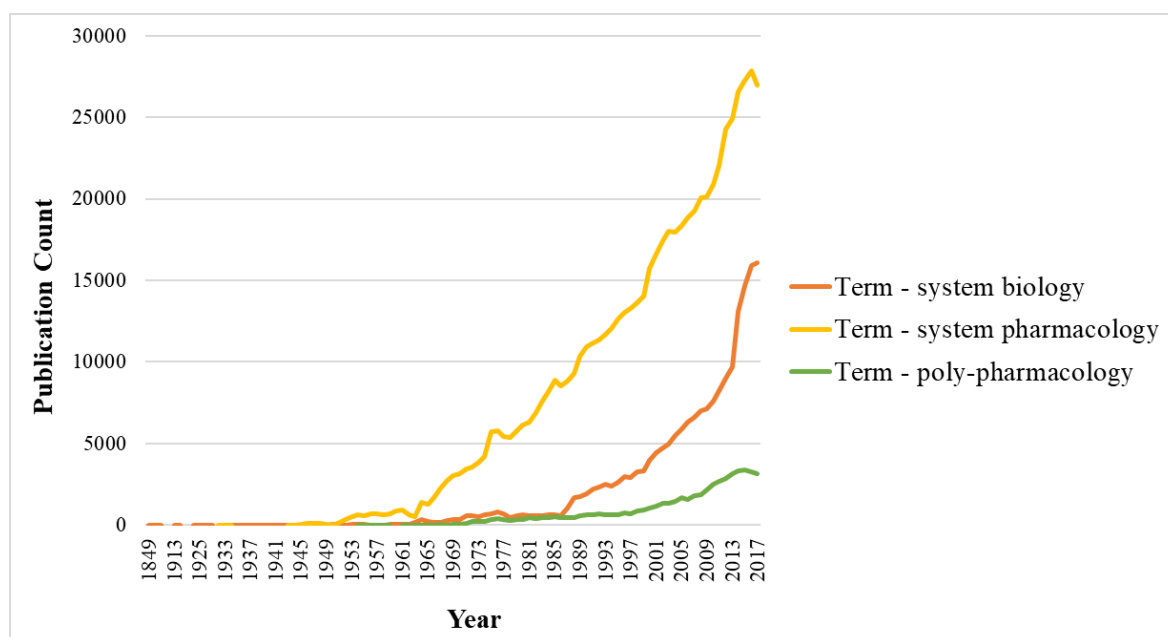


Figure 1.2: The growth of the total number of publications in the field of systems biology, systems pharmacology and poly-pharmacology in the PubMed database.

Nonetheless, the success of drug designing for complex diseases depends on an interdisciplinary multistep process which mainly involves: a) understanding of disease mechanism (system biology level) for the identification of clinically and biologically validated targets

1. INTRODUCTION

(right targets), b) screening of exogenous entities (right drug molecule) from the large chemical space that can manipulate required biological targets for normalizing several disease associated pathways and networks and finally, c) analysis of the mechanism of action of the drug in complex biological systems (system pharmacology level) to test and achieve required therapeutic effects (right phenotypic effect) (Figure 1.1).

Several omics studies clearly indicate that CNS disorders such as Parkinson's disease, Alzheimer's disease, Huntington's disease, or amyotrophic lateral sclerosis are multifactorial in origin and have complex pathomechanisms [12, 13]. Systems-biology-oriented analyses of complex disease phenotypes in CNS disorders have been helping in narrowing down biological networks into a fewer number of relevant disease-causing targets [14, 15, 16]. Better biology of such complex diseases provides both opportunities and challenges in the discovery and development of novel medicines for their treatment [17, 18]. In such highly heterogeneous diseases where many potential defects in the structure, function, or regulation of the cells are involved, single-target medications have failed [17, 19]. Thus, there is a mounting need to bring new drugs into clinical practice for the rebalancing of the several proteins or events, that contribute to the etiology, pathogenesis, and progression of diseases.

Due to the complexity of the large network of disease-causing factors in CNS pathology, in recent years, *in-silico* methods for polypharmacology are being promoted in pharmaceutical industries and academia to achieve a desired multi-target activity profiles of drugs [17, 19, 20]. Polypharmacology approaches have discovered many drugs with enlarged therapeutic ranges including single drug-multiple targets interaction (promiscuous ligands) and also multiple drugs binding to one target (promiscuous targets) (Figure 1.1). The purpose of predicting drug promiscuity is twofold: firstly, for screening all possible off-target proteins that can detract drug action towards unwanted side effects and, secondly, the promiscuous behaviour of drugs that can also be exploited for multi-target drug design [19]. Thus, many of the modern drug design methods are mainly focused on drug repurposing (i.e., new targets for known/old drugs) [17, 21]. Drug repurposing can significantly reduce the cost of

1.1 Problem statement and the aims of the study

an expensive and time-consuming process to bring a drug to market with improved clinical efficacy and safety [17, 22, 23].

Virtual screening (VS) has been typically considered an area of computer-aided drug discovery where computational approaches including ligand-based VS, or structure-based VS aim for predicting drug-target(s) binding affinities to reach desired therapeutic effects and off-target(s) binding for avoiding possible adverse effects. In the absence of the target protein structure, ligand-based VS methods are used for identifying new molecular structures (ligands) based on the principle that similar molecules are likely to have similar properties [24, 25]. Commonly used ligand-based VS approaches (e.g., similarity searching [25], pharmacophore mapping [26] and machine learning methods [27, 28, 29]) analyse large molecular databases on the basis of compounds chemical and biological properties, structure, shape, bioactivity, and *in-silico* descriptors or computed properties to identify ligands likely to have similar properties to the known actives [25, 30, 31, 32, 33]. On the other hand, structure-based VS approaches use experimentally known structure of the target proteins to apply well-known molecular docking to discover small molecules that mimic the binding interaction of ligands into the active site with a high predicted binding affinity (scoring) [34, 35].

1.1 Problem statement and the aims of the study

Given the complexity and multifactorial nature of CNS diseases, efforts have been made by academic researchers, medicinal chemists, and pharmaceutical industries for the development of polypharmacological approaches as a promising therapeutic strategy to find multi-targeting drug molecules [17, 5, 19, 20]. However, the reported sensitivity of these methods is around 50% [36, 19] which must be improved to get the desired predictive performance. The possible reason for lower or reduced predictability for multi-targeting drugs is a global assumption behind ligand-based VS and structure-based VS methods that are mostly adopted

1. INTRODUCTION

for polypharmacology techniques [19]. Ligand-based VS focus on known active molecules to find similar molecules considering the assumption that structurally similar molecules have the tendency to bind to similar targets [24] and structure-based VS uses structures of target proteins following the supposition that proteins with similar binding sites will bind similar ligands [37]. Nonetheless, to work at the level of systems pharmacology, the main challenges are: a) at molecular level (i.e., dealing with big data sets of diverse molecular structures and their biological properties for the understanding of the structure-activity relationship), b) at cellular level for exploring complex cellular networks, and c) finally at the organism level to model whole biological system for the understanding of big systems-based picture. On the molecular level, polypharmacology approaches are adapted for computationally fast large-scaled screenings of millions of compounds against all of the desired proteins using their known molecules as a query and the most promising molecules are selected for further chemical synthesis and experimental (*in-vitro* and *in-vivo*) testing [17, 19, 20, 8].

It is then possible to clearly define the goal of this work, which can thus be formulated as:

Can existing modeling approaches be improved so as to advance virtual screening in multi-target drug design?

Keeping in mind this goal, the presented document covers the design and implementation of cheminformatics methods to tackle different chemical data manipulation and modeling problems in medicinal chemistry, with particular emphasis on machine learning applications in QSAR modeling, molecular structural representation methods, and molecular similarity analysis and its applications in chemical space visualization (Figure 1.1). Moreover, this thesis further compares the performance of molecular docking and QSAR modeling and integrates the advantages of these two different methods of VS in polypharmacology, an important task in systems pharmacology. Such efforts can play a significant role in computational chemical data mining and understanding of the relationship between chemical

structure and desired properties to identify and prioritize promising multi-targeting candidate molecules for experimental validation at the systems level in the drug discovery process. A more specific overview of the objectives of this study is represented in Figure 1.3 including the following tasks:

- **Objective-1:** Automation and validation of the a state-of-the-art machine learning approaches to develop a fully automated and reliable QSAR model building platform. This platform would provide an efficient data curation process for preparing a good quality data by directly accessing online manually curated databases to build global models. Moreover, to check the processed data quality the automated workflow will include prior estimation of data modelability to avoid time-consuming modeling trials.
- **Objective-2:** Extensive analysis of molecular representation methods to assess different data analysis and modeling methods. This objective was divided into two tasks. First task would be a comparative analysis of the most widely used molecular structural transformation methods used to convert compounds structures into machine readable formats, which is required for fitting machine learning models in QSAR problems. Second tasks is designing a novel tool for visual characterization and diversity analysis of chemical data by representing high dimensional molecular spaces into reduced dimensionality.
- **Objective-3:** To optimize and establish a rational and re-usable polypharmacology-based drug designing pipeline by integrating molecular docking and QSAR modeling approaches.

1.2 General methodology

An overview of the designed and implemented methodology to achieve the above defined objectives is explained as follows.

1. INTRODUCTION

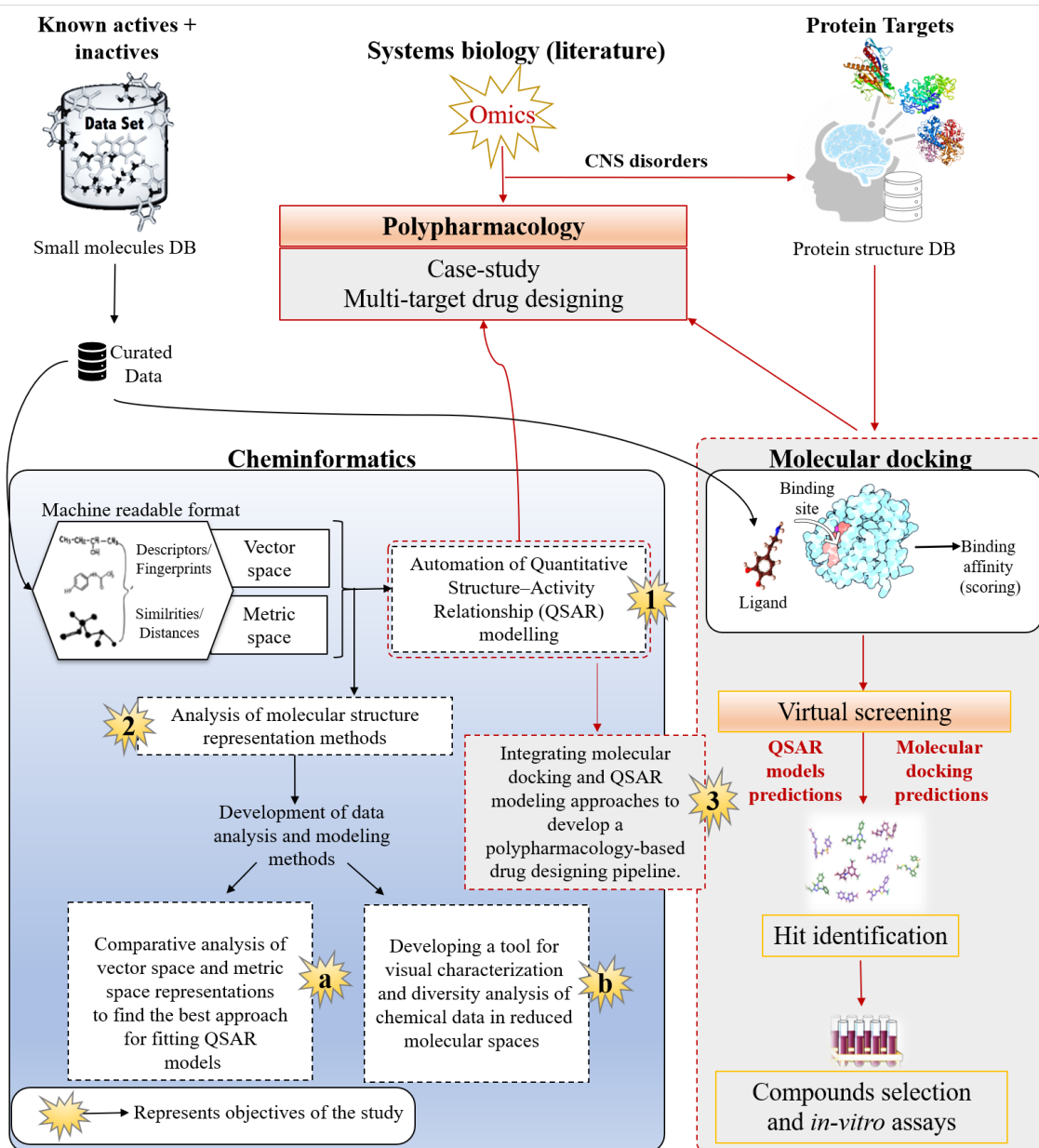


Figure 1.3: Graphical representation of the thesis objectives and the methodology

1.2 General methodology

In the first objective (Figure 1.3), an extendable and highly customizable automated QSAR modeling framework was developed that can be used as a QSAR modelling pipeline to build robust predictive models. The workflow, given a target or problem, automatically collect and curate molecular structures and corresponding biological activity data for a specified target. Furthermore, to quantify various features of molecular structures a variety of chemical descriptors are computed. Before entering in the laborious process of feature selection, model building and validation, data modelability evaluation is performed for data quality validation. Difficult or poor-quality data sets are not recommended for modeling. After data quality assessment, QSAR modeling framework follows a feature selection process to select an optimal set of features by using a state-of-the-art methodology and follows an unbiased standard protocol of QSAR model building with external and internal retrospective validation.

The first task of the second objective (Figure 1.3) of this document details a thorough evaluation of existing molecular structure representations as the accurate characterization of biological molecular properties in QSAR modeling largely depends on the relevance of the selected molecular representation. To accomplish this task seven different molecular representations from two main categories of molecular representation (vector space and metric space) were used in five distinct QSAR data sets. For completion two different dimensionality reduction methods (principal components reduction and feature selection) were tested, thus, in total eighteen different approaches for each data set were implemented. The number of features was selected using cross-validation, and, each final model was assessed against an independent validation set randomly selected from the initial data set, which was never used in any step of the model fitting phase.

It was further developed a reliable pipeline that can efficiently be used to build probabilistic surfaces of molecular activity (PSMAs) for a visual characterization of molecules in molecular activity spaces. This approach is, to my knowledge, new and allows building a non-parametric classification model out of raw data of molecular similarity. Visualization

1. INTRODUCTION

of such high dimensional metric space data is a difficult challenge in many different domains of data analysis, as it demands efficient and robust techniques to adequately represent data variability in lower dimensions (2D or 3D). In this study, four different dimension reduction algorithms were tested to transform structural distance matrices in 2D to generate a topographical map, which should be able to present a visual analysis of the diversity of large heterogeneous chemical data, from the projected space (X, Y coordinates). The generated activity probability maps were assessed and validated as a classification model for four different QSAR datasets.

Finally, it was aimed to integrate the knowledge of molecular docking with the state-of-the-art QSAR modeling methodology (established in previous objectives) to develop a rational drug designing methodology pipeline. The developed approach was used to perform fast and computationally efficient virtual screening of large databases to predict compounds that bind multiple targets against Parkinson's disease chosen as a case study. The identified hits were ranked/sorted according to three criteria including a) best docking predictions only, b) best QSAR predictions and c) both (consensus) best docking and QSAR model predictions. Compounds from these three categories were considered for testing in *in-vitro* assays to compare the robustness of best performing understudy approaches. The purpose of this task is the development of novel and re-usable polypharmacology method for rational design of multi-targeting ligands.

1.3 Overview of the document

Other chapters of this document are organized as follows:

- Chapter 2 gives a background on several concepts to explain this work. The chapter starts with an introduction to VS and provides an overview of the most common computational approaches used to perform VS. Two different domains of VS methods in-

cluding cheminformatics applications in ligand-based VS and structure based VS using molecular docking are explained. Background on cheminformatics covers several areas including molecular structure representation/transformation methods, molecular similarity applications in similarity-based VS and molecular spaces visualization, and this section ends with the introduction of QSAR modeling. In the section of molecular structure-based VS, overview of the important tasks to perform molecular docking analysis is presented, namely target structure analysis and selection for docking, docking software selection and parameters optimization and, ligands database preparation for virtual screening. In the last section of this chapter large compound repositories for VS are briefly explained.

- Chapter 3 pinpoints the need of automation of the QSAR modeling process and highlights the advantages of automation of repetitive tasks in the laborious drug discovery process. This chapter provide a review of QSAR/QSPR modeling “life cycle” some standard steps, critical for reliable model building, and presents a fully automated QSAR modeling platform.
- Chapter 4 focuses on a thorough comparative analysis of molecular structures representation (vector space and metric space) methods to find the best suitable approach to describe molecular structural information, which is used for predicting relationships between biological activity (response variable) and structural information (predictors) in QSAR modeling.
- Chapter 5 presents a methodology that is able to integrate the advantages of the different methods including (a) molecular space representation, (b) non-linear dimension reduction methods and (c) Non-parametric 2D kernel density estimation (KDE) function to build a probabilistic surface of molecular activity (PSMA). The activity probability maps characterize actives and inactives into lower dimensional metrics spaces that is useful for chemical activity spaces visualization and can also serve as spatial

1. INTRODUCTION

classification model with clear predictive properties.

- Chapter 6 describes the problem of low sensitivity of polypharmacological approaches. It also highlights the applications of VS methods in the identification of chemical structures with multi-targeting activity. Additionally, chapter 6 covers all the concepts such as the implementation and validation of molecular docking and QSAR methods for hit identification, and presents a polypharmacology-based VS methodology that integrates molecular docking and QSAR modeling approaches. The purpose of integrating these two approaches was to identify promising inhibitors against PD by introducing a dual-targeting drug designing.
- Chapter 7 Presents an overview of each objective, concludes their corresponding contributions and discusses limitations and future directions of the present work.

1.4 Publications and participation in academic activities

1.4.1 Papers in scientific peer-reviewed journals

1. Kausar S, Falcao AO (2018). An automated framework for QSAR model building. *Journal of Cheminformatics*. 16;10(1):1. doi: 10.1186/s13321-017-0256-5.

- QSAR modeling workflow zipped source file and all generated models with their completely curated data sets are available at: <https://github.com/Saminakausar/Automated-framework-for-QSAR-model-building>

2. Kausar S, Falcao AO (2019). Analysis and comparison of vector space and metric space representations in QSAR modeling. *Molecules*. 24(9). doi: <https://doi.org/10.3390/molecules24091698>

1.4 Publications and participation in academic activities

- Supplementary data (Additional file 1) contains three supplementary tables including: Table S1: List of RDkit 2D and 3D descriptors, Table S2: 5-fold cross-validation results, Table S3: External validation results and a Figure S1: Selection of optimized number of PCs: PVE vs. number of PCs plot from PCA on metric space are available at: <https://www.mdpi.com/1420-3049/24/9/1698>.
3. Kausar S, Falcao AO (2019). A visual approach for analysis and inference of molecular activity spaces. *Journal of Cheminformatics*. **(First revision submitted)**
 - All data sets and R source code (PSMA.Rmd and PSMA.html) for analysis and inference of molecular activity spaces are available at: <https://github.com/Saminakausar/A-visual-approach-for-analysis-and-inference-of-molecular-activity-spaces>
 4. Kausar S, Guedes RC, Falcao AO (2019). Comparative analysis of QSAR modelling and molecular docking: a rational approach in polypharmacology. **(In progress)**
 - COMT and GSK3B QSAR classification models are available at: <https://github.com/Saminakausar/Automated-framework-for-QSAR-model-building>

1.4.2 Participation in conferences

1. Kausar S, Falcao AO (2015). Predicting amyotrophic lateral sclerosis progression: a simple Bayesian model for longitudinal and time-to-event clinical trial data. Poster presented at the LaSIGE workshop, Department of Informatics, Faculty of Science, University of Lisbon. November 7, 2015

1. INTRODUCTION

2. Kausar S, Falcao AO (2017). An open-source platform for automated processing and integration of data in pharmacological activity modelling. Poster presentation at a conference “From Single- to Multiomics: Applications and Challenges in Data Integration”, EMBL Heidelberg, Germany. November 12th-14th 2017.
3. Kausar S, Falcao AO (2018). Analysis and inference within the molecular space: A visual approach using NAMS and multidimensional scaling. Poster presentation at a conference “11th International Conference on Chemical Structures”, Noordwijkerhout The Netherlands. May 27th-31st 2018.
4. Kausar S, Falcao AO (2018). Selective modelling of MAO-B inhibitors for neurodegenerative disorders’ *in-silico* molecular screening. Poster presentation at workshop “WORKSHOP ON INTEGRATIVE APPROACHES IN NEURODEGENERATION”, Faculty of Sciences University of Lisbon, Lisbon, Portugal. Jun 21st-23rd 2018.
5. Franco C, Kausar S, Brito MA, Guedes R, Falcão A. Computational modeling in glioblastoma: a comprehensive approach to overcome the blood-brain barrier and target EGFR and PI3K signaling. Annual Blood-Brain Barrier Consortium Meeting. Portland, OR, USA, March 7-8th, 2019, [P11].

1.4.3 Participation in academic competitions in science

1. Participated in a Multi-Targeting Drug Dream Challenge where anticipated outcomes was: 1) novel and re-usable methods for rational design of multi-targeting compounds and 2) a benchmark standard for assessing multi-targeting compounds. Under this competition, a polypharmacology-based virtual screening approach was developed to predict the structures of candidates that can bind to and inhibit the activity of multiple independent targets for two biological problems including medullary thyroid carcinoma and neurodegenerative model of tauopathies.

References

- [1] Antonio Lavecchia and Carmen Cerchia. ‘In silico methods to address polypharmacology: current status, applications and future perspectives’. In: *Drug Discovery Today* 21.2 (Feb. 2016), pp. 288–298. ISSN: 13596446. DOI: 10.1016/j.drudis.2015.12.007.
- [2] Anderson J. *National institute of general medical sciences centers for systems biology*. 2003. URL: <http://www.nigms.nih.gov/funding/systems.html>.
- [3] Mete Civelek and Aldons J. Lusic. ‘Systems genetics approaches to understand complex traits.’ In: *Nature reviews. Genetics* 15.1 (Jan. 2014), pp. 34–48. ISSN: 1471-0064. DOI: 10.1038/nrg3575.
- [4] Ale Prokop and Béla Csukás. *Systems Biology*. Ed. by Aleš Prokop and Béla Csukás. Dordrecht: Springer Netherlands, 2013. ISBN: 978-94-007-6802-4. DOI: 10.1007/978-94-007-6803-1.
- [5] Eugene C. Butcher, Ellen L. Berg and Eric J. Kunkel. ‘Systems biology in drug discovery’. In: *Nature Biotechnology* 22.10 (Oct. 2004), pp. 1253–1259. ISSN: 1087-0156. DOI: 10.1038/nbt1017.
- [6] Han-Yu Chuang, Matan Hofree and Trey Ideker. ‘A decade of systems biology.’ In: *Annual review of cell and developmental biology* 26 (2010), pp. 721–44. ISSN: 1530-8995. DOI: 10.1146/annurev-cellbio-100109-104122.
- [7] Trey Ideker, Timothy Galitski and Leroy Hood. ‘A new approach to decoding life: systems biology.’ In: *Annual review of genomics and human genetics* 2.1 (Sept. 2001), pp. 343–72. ISSN: 1527-8204. DOI: 10.1146/annurev.genom.2.1.343.

1. INTRODUCTION

- [8] Andrew M. Stern et al. ‘A Perspective on Implementing a Quantitative Systems Pharmacology Platform for Drug Discovery and the Advancement of Personalized Medicine’. In: *Journal of Biomolecular Screening* 21.6 (July 2016), pp. 521–534. ISSN: 1087-0571. DOI: 10.1177/1087057116635818.
- [9] Yunan Luo et al. ‘A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information’. In: *Nature Communications* (2017). ISSN: 20411723. DOI: 10.1038/s41467-017-00680-8.
- [10] Tapio Pahikkala et al. ‘Toward more realistic drug-target interaction predictions.’ In: *Briefings in bioinformatics* 16.2 (Mar. 2015), pp. 325–37. ISSN: 1477-4054. DOI: 10.1093/bib/bbu010.
- [11] Monica Schenone et al. ‘Target identification and mechanism of action in chemical biology and drug discovery.’ In: *Nature chemical biology* 9.4 (Apr. 2013), pp. 232–40. ISSN: 1552-4469. DOI: 10.1038/nchembio.1199.
- [12] SL Budd Haeberlein and TJR Harris. ‘Promising Targets for the Treatment of Neurodegenerative Diseases’. In: *Clinical Pharmacology & Therapeutics* 98.5 (Nov. 2015), pp. 492–501. ISSN: 00099236. DOI: 10.1002/cpt.195.
- [13] Sean Quinlan et al. ‘MicroRNAs in Neurodegenerative Diseases’. In: *International Review of Cell and Molecular Biology*. 2017, pp. 309–343. ISBN: 1015-6305. DOI: 10.1016/bs.ircmb.2017.04.002.
- [14] Jianguo Xia and David S. Wishart. ‘MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data.’ In: *Nucleic acids research* 38.Web Server issue (July 2010), W71–7. ISSN: 1362-4962. DOI: 10.1093/nar/gkq329.

REFERENCES

- [15] Marc-Emmanuel Dumas and Laetitia Davidovic. ‘Metabolic Profiling and Phenotyping of Central Nervous System Diseases: Metabolites Bring Insights into Brain Dysfunctions’. In: *Journal of Neuroimmune Pharmacology* 10.3 (Sept. 2015), pp. 402–424. ISSN: 1557-1890. DOI: 10.1007/s11481-014-9578-5.
- [16] Tamás Korcsmáros et al. ‘How to design multi-target drugs.’ In: *Expert opinion on drug discovery* 2.6 (June 2007), pp. 799–808. ISSN: 1746-0441. DOI: 10.1517/17460441.2.6.799.
- [17] Andrew Anighoro, Jürgen Bajorath and Giulio Rastelli. ‘Polypharmacology: Challenges and Opportunities in Drug Discovery’. In: *Journal of Medicinal Chemistry* 57.19 (Oct. 2014), pp. 7874–7887. ISSN: 0022-2623. DOI: 10.1021/jm5006463.
- [18] Yannis Karamanos and Gwënaël Pottiez. ‘Proteomics and the blood–brain barrier: how recent findings help drug development’. In: *Expert Review of Proteomics* 13.3 (Mar. 2016), pp. 251–258. ISSN: 1478-9450. DOI: 10.1586/14789450.2016.1143780.
- [19] Violeta I Pérez-Nueno. ‘Using quantitative systems pharmacology for novel drug discovery’. In: *Expert Opinion on Drug Discovery* 10.12 (Dec. 2015), pp. 1315–1331. ISSN: 1746-0441. DOI: 10.1517/17460441.2015.1082543.
- [20] Alan Talevi. ‘Multi-target pharmacology: possibilities and limitations of the “skeleton key approach” from a medicinal chemist perspective’. In: *Frontiers in Pharmacology* 6.SEP (Sept. 2015), pp. 1–7. ISSN: 1663-9812. DOI: 10.3389/fphar.2015.00205.
- [21] Weilin Zhang, Jianfeng Pei and Luhua Lai. ‘Computational Multitarget Drug Design.’ In: *Journal of chemical information and modeling* 57.3 (2017), pp. 403–412. ISSN: 1549-960X. DOI: 10.1021/acs.jcim.6b00491.
- [22] Nicola Nosengo. ‘Can you teach old drugs new tricks?’ In: *Nature* 534.7607 (June 2016), pp. 314–316. ISSN: 0028-0836. DOI: 10.1038/534314a.

1. INTRODUCTION

- [23] Curtis R Chong and David J Sullivan. ‘New uses for old drugs.’ In: *Nature* 448.7154 (Aug. 2007), pp. 645–6. ISSN: 1476-4687. DOI: 10.1038/448645a.
- [24] Mark A Johnson and Gerald M Maggiora. *Concepts and applications of molecular similarity*. 1990. ISBN: 0471621757.
- [25] Ingo Muegge and Prasenjit Mukherjee. ‘An overview of molecular fingerprint similarity search in virtual screening’. In: *Expert Opinion on Drug Discovery* 11.2 (2016), pp. 137–148. ISSN: 1746-0441. DOI: 10.1517/17460441.2016.1117070.
- [26] Hongmao Sun. ‘Pharmacophore-Based Virtual Screening’. In: *Current Medicinal Chemistry* 15.10 (Apr. 2008), pp. 1018–1024. ISSN: 09298673. DOI: 10.2174/092986708784049630.
- [27] Yu-Chen Lo et al. ‘Machine learning in chemoinformatics and drug discovery’. In: *Drug Discovery Today* 23.8 (Aug. 2018), pp. 1538–1546. DOI: 10.1016/j.drudis.2018.05.010.
- [28] Bruno J. Neves et al. ‘QSAR-based virtual screening: Advances and applications in drug discovery’. In: *Frontiers in Pharmacology* 9.NOV (2018), pp. 1–7. ISSN: 16639812. DOI: 10.3389/fphar.2018.01275.
- [29] James Melville, Edmund Burke and Jonathan Hirst. ‘Machine Learning in Virtual Screening’. In: *Combinatorial Chemistry & High Throughput Screening* 12.4 (May 2009), pp. 332–343. ISSN: 13862073. DOI: 10.2174/138620709788167980.
- [30] Jürgen Bajorath. *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*. Ed. by Jürgen Bajorath. Vol. 275. *Methods in Molecular Biology*. Totowa: Humana Press, 2004. ISBN: 978-1-58829-261-2. DOI: 10.1385/1592598021.
- [31] Roberto Todeschini and Viviana Consonni. *Molecular Descriptors for Chemoinformatics*. Ed. by Roberto Todeschini and Viviana Consonni. *Methods and Principles in Medicinal Chemistry*. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, July 2009. ISBN: 9783527628766. DOI: 10.1002/9783527628766.

REFERENCES

- [32] Nina Nikolova and Joanna Jaworska. ‘Approaches to Measure Chemical Similarity—a Review’. In: *QSAR & Combinatorial Science* 22.910 (2003), pp. 1006–1026. ISSN: 1611-020X. DOI: 10.1002/qsar.200330831.
- [33] Alan R. Katritzky, Victor S. Lobanov and Mati Karelson. ‘QSPR: the correlation and quantitative prediction of chemical and physical properties from structure’. In: *Chemical Society Reviews* 24.4 (1995), pp. 279–87. ISSN: 0306-0012. DOI: 10.1039/cs9952400279.
- [34] Xuan-Yu Meng et al. ‘Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery’. In: *Current Computer Aided-Drug Design* 7.2 (June 2011), pp. 146–157. ISSN: 15734099. DOI: 10.2174/157340911795677602.
- [35] Nataraj S. Pagadala, Khajamohiddin Syed and Jack Tuszynski. ‘Software for molecular docking: a review.’ In: *Biophysical reviews* 9.2 (2017), pp. 91–102. ISSN: 1867-2450. DOI: 10.1007/s12551-016-0247-1.
- [36] Eugen Lounkine et al. ‘Large-scale prediction and testing of drug activity on side-effect targets’. In: *Nature* 486.7403 (June 2012), pp. 361–367. ISSN: 0028-0836. DOI: 10.1038/nature11159.
- [37] V. J. Haupt and Michael Schroeder. ‘Old friends in new guise: repositioning of known drugs with structural bioinformatics’. In: *Briefings in Bioinformatics* 12.4 (July 2011), pp. 312–326. ISSN: 1467-5463. DOI: 10.1093/bib/bbr011.

2

Background/state of the art

2.1 Virtual screening approaches

Virtual screening (VS) refers to a range of *in-silico* methods that serve as computational analogues of biological high-throughput screening in modern drug discovery process [1]. The aim of VS is to search large small-molecule databases to select the chemical structures that have the largest probability of activity in biological testing in a “lead discovery programme”. Thus, VS may reduce the total cost of drug development by screening the experimentally manageable numbers of candidate/lead molecules [1, 2, 3, 4, 5]. A lead molecule is a compound that has the required pharmacological properties and thus is used as a good starting point for drug discovery. VS procedures (Figure 2.1) depending on the availability of the amount of targets structural and bioactivity information use one or a combination of computational methods which can be classified as ligand-based VS and structure-based VS [6].

Structure-based VS, such as molecular (protein–ligand) docking can be implemented if

2. BACKGROUND/STATE OF THE ART

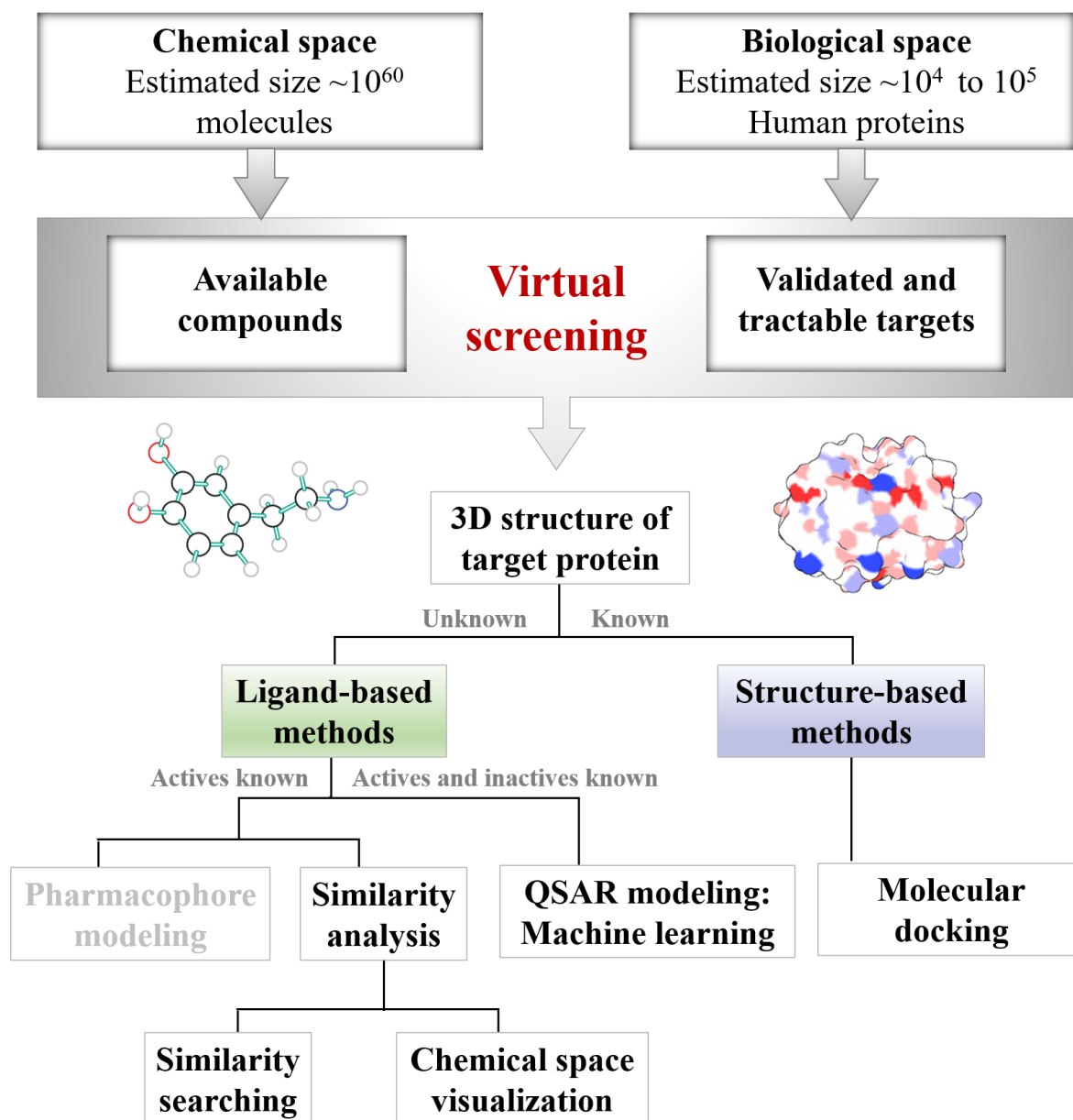


Figure 2.1: Overview of virtual screening approaches

2.1 Virtual screening approaches

the crystallographic structure of the biological target is known [7]. While in the absence of a 3D structural model of target, different ligand-based VS that incorporate various level of chemical information and representations (three-, two-, and even one dimensional) [8, 9, 10] are preferred. Examples of ligand-based approaches include: a) similarity searching (e.g., shape-based [11], and fingerprint-based similarity [12, 8] is commonly used when single or just a few chemical molecules are known, b) Pharmacophore-based mapping/matching [13] can be implemented when many actives and inactive are available, and c) machine learning methods are well suited for larger amount of active and inactive ligand molecules to derive a quantitative structure-activity relationship (QSAR) [14, 5, 15]. Moreover, similarity analysis has further applications in the visualization of the characteristics of big chemical space which can also serve in predicting biological activity [16, 17, 18]. However, ligand-based VS methods identify molecules from large chemical libraries that share some structural/biological activity similarity with the active ligand molecules at hand which have been identified as potential leads. Ligand-based VS approaches are being addressed mainly under cheminformatics and considered as a fast and powerful virtual screening methodology to deal with big chemical space comparatively to traditional and relatively computationally "inefficient" molecular docking and scoring approaches (Figure 2.1).

The scope of this thesis focuses on molecular docking studies as mainstream structure-based VS method and advanced cheminformatics applications for VS particularly chemical data transformation methods, similarity analysis and, QSAR modeling. Pharmacophore methods that extract common 3D pharmacophoric pattern (features important for the molecule to be active) from known actives and later used to a 3D database searching are out of the scope of the presented study.

2. BACKGROUND/STATE OF THE ART

2.1.1 Cheminformatics applications in virtual screening

Cheminformatics (also known as chemoinformatics and chemical informatics) is a broad field that was originally developed with the goal for computationally fast and accurate searching of large compound databases for chemical information retrieval and extraction using computer science and information technology [6, 19, 20]. The term chem[o]informatics was first time defined as:

“mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization” [21].

Cheminformatics techniques soon became a critical part of the drug discovery process to tackle other chemical problems including molecular graph processing, descriptor and fingerprint construction for transforming chemical structure into chemical information (compounds physical, chemical, and biological properties) for further applications in similarity searching and chemical space exploration, pharmacophore and scaffold analysis, library design, and data mining using machine learning methods to correlating structural features with biological properties [14, 6]. However, such approaches that correlate structure-activities are among the most focused areas in pharmaceutical companies and lie at the heart of virtual screening, the most important current application of cheminformatics in the process of drug designing [14].

2.1.1.1 Molecular structure representation/transformation methods

One of the major tasks in cheminformatics is the transformation of chemical structures into computer-readable formats. Molecules are complex entities that have atoms connected with covalent bonds and also chemical reactions can convert them from one form to another form. Thus, the conversion of molecular structures to information/properties is not a simple

2.1 Virtual screening approaches

process but involved multilayer computational processing. In the first step, a molecular structure is represented as molecular graph/connection table and then features vectors (known as descriptors and fingerprints) are calculated that stores different form of information including chemical properties, geometries, interactions and reactions. Since, each representation does not include all information about all structural elements, which arises a question of selecting the best structural transformation each time for characterizing compounds under specific problems. Hence, chemical structural representation is not always explicit and unique [22]. Nowadays, a large variety of methods have been developed for encoding a compound as a feature vector to represent and to mine the molecular information [23, 6].

In this section, the basic concept of graph theory and molecular representation like molecular descriptors and fingerprints are described. These methods are most commonly used for performing several operations (e.g., storage/retrieval, identity, substructure/superstructure relationships, similarity and multivariate relationships).

Molecular/chemical graph theory

Molecular/chemical graph theory is a topology branch of mathematical chemistry which is important to understand the structures (specified by their graph representations) containing the chemical information that influences their biological activities and is necessary to solve molecular problems [24]. A molecular graph (also known as a “chemical graph” or “structural graph”) is a simple graph represents structural formula of a molecule having nodes to represent atoms and edges that represent covalent bonds between the corresponding atoms. A molecular graph represents only the topology of chemical structure and has no information about their 3D arrangement. Therefore, molecular graphs can only distinguish molecules isomeric forms (structurally distinct compounds (non-isomorphic graph) but same molecular formula) but cannot discriminate stereoisomers or conformational isomers. Different ways for constructing chemical graphs have been proposed [25, 26, 6] and these variations

2. BACKGROUND/STATE OF THE ART

of chemical graphs representation may not obviously correspond to a “standard” chemical compound. Graph representation reduces the complexity of chemical systems due to loss of some structural information (properties derived using molecular geometry, stereochemistry or 3D conformation). But atomic connectivity information from chemical graphs using a bond adjacency matrix, or topological distance matrix supports the calculation of several molecular descriptors treated as the molecular signatures that are useful for cheminformatics modeling [27]. It is necessary to have efficient methods to convert the molecular graph to computer-readable format for further application of chemical graphs (e.g., chemometric analysis for comparing and quantify chemical diversity) [27, 28].

Molecular descriptors

Molecular structural descriptors are numerical vectors of features describe the information encoded in chemical structures [29]. These descriptors are derived with mathematical formulae obtained from Chemical Graph Theory, Information Theory, Quantum Mechanics, etc., while others directly illustrate some relevant feature of the molecules (Table 2.1) [30, 31]. Molecular descriptors are typically defined as:

”The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment” [31].

Molecular descriptors can be divided into 4 broader categories depending on the degree/-level of structural information required to compute them [32, 22, 6, 31]. Thus, molecular descriptors can be constitutional (0 or 1 dimensional (D)), topological (2D), geometrical (3D) and physico-chemical properties-based (4D) (Table 2.1) [33, 34, 22, 31]. 2D descriptors including topological indices, molecular profiles and 2D autocorrelation descriptors are the most commonly used types of descriptors reported in the literature [33]. 2D descriptors give equally good or even better performance than the other higher dimensional types and save the

2.1 Virtual screening approaches

Table 2.1: Types of molecular descriptors according to dimensions

Type	Dimension	Derived from	Description
Constitutional (Theoretical descriptors)	0D, 1D	Molecular formula and chemical graph	Reflect the chemical information of molecules without considering atom connectivity information. Example: Molecular weights, atom counts, bond counts, fragment counts, functional group counts etc.
Topological (Theoretical descriptors)	2D	Structural topology	Use the information of connectivity of atoms in molecules without their geometric information. Example: Balaban, weiner, zagreb and randic connectivity indices, BCUTS etc.
Geometrical (Theoretical descriptors)	3D	Structural geometry	Represent the 3D information regarding the molecular size, shape, and position of the atoms in space. Example: GETAWAY, autocorrelation WHIM, and 3D-MORSE etc.
Physicochemical (Theoretical + experimental descriptors)	4D	Chemical behaviour (thermodynamics), steric properties, and electronic properties	Used to describe molecular properties from conformational behaviour or observed chemical interactions with the surrounding space Example: Different constants (partition coefficients (logP), hydrophobic substituent (p), acid dissociation, Hammett, taft steric, and Charton's constants), molRef (Molar refractivity), Volsurf, raptor, GRID etc.

computational time required for a laborious process of 3D structural features (e.g., autocorrelation descriptors, substituent constants, surface: volume descriptors and quantum–chemical descriptors) or 4D molecular properties or structural conformations prediction [33, 10, 35]. Description and examples of each type of descriptors are described in Table 2.1 [36, 31]. Different types of descriptors are used for various tasks in cheminformatics e.g., molecular data mining/VS, compound diversity analysis and activity prediction [36, 31].

As several types of representations of molecular structures and descriptors calculations

2. BACKGROUND/STATE OF THE ART

programs (free and commercial) are available that automatically extract thousands of descriptors from different structural representations and differ in their computation time and the complexity of the encoded information [22, 31]. Software programs for generating different molecular descriptors are listed in Table 2.2.

Molecular fingerprints

Molecular fingerprints are high-dimensional vectors (binary bit strings) that encode a fragment or characteristic of a given molecule. Fingerprints are well-known molecular representation commonly used in molecular diversity analysis, similarity-based virtual screening, and in the clustering of chemical databases [37, 14, 6, 22]. Different settings (e.g., generation method, length, size of patterns and number of bits activated by each pattern etc.) are used to encode molecular information in the binary vectors of fingerprints. Each bit in the bit strings or fingerprints represents the absence (0) or presence (1) molecular structural feature or chemical descriptor value [12, 38]. However, molecule fingerprints serve as the simpler form of molecular representation that is used for assessing molecular similarity through their fast and computationally easier comparisons and are also deployed as descriptors for predictive modelling to estimate the biological activities [39, 40, 41, 23, 42]. In the past years, the definitions and classification of molecular fingerprints and their applications have been reviewed in detail [43, 38, 44, 45]. Molecular fingerprints the most popular for VS can be grouped into different classes depending on the methods used to compute them (Table 2.3) [12, 38]. Table 2.3 summarises several types of fingerprint and softwares have been developed to generate them by mapping molecular structures to bit strings [38].

2.1.1.2 Molecular similarity concepts

Molecular similarity analysis has numerous applications including molecular structure superposition, common substructure searching, similarity-based virtual screening (similar-

2.1 Virtual screening approaches

Table 2.2: Software programs for calculating different molecular descriptors.

Software	Total descriptors	Type of descriptors	Operating system/ Platforms	home page
Commercial				
DRAGON	5270	Constitutional, topological, 2D-autocorrelations, geometrical, WHIM, GETAWAY, RDF, functional groups, properties etc.	Windows/ Unix/Linux	Link
MOE	over 300	Topological, physical properties, structural keys, etc.	Windows/Linux/ SGI /MAC /Sun	Link
MOLCONN-Z	over 40	Molecular Connectivity, Shape, and Information Indices (topological)	Windows /Unix /Linux /MAC	Link
ADMET predictor	297 (2D: 266 and 3D: 31)	constitutional, functional group counts, topological, E-state, molecular patterns, electronic properties, 3D descriptors, hydrogen bonding, acid-base ionization etc.	Windows	Link
ACD/labs	-	Predict molecular properties (molecular weight, density, log P, log S, log D, and pKa, etc.)	-	Link
ADRIANA. Code	1244	Global physicochemical descriptors, atom property-weighted 2D- and 3D-autocorrelations and RDF, surface property-weighted autocorrelations.	Windows /Unix	Link
CODESSA	1500	Constitutional, topological, geometrical, charge-related, semi-empirical, thermodynamical	Windows	Link
ADAPT	260	Topological, geometrical, electronic, physicochemical	Unix /Linux	Link
MOLGEN-QSPR	707	Topological, constitutional, geometrical etc.	Unix /Linux /Sun	Link
Freeware				
CDK	over 200	Topological, geometrical, electronic, constitutional	Windows /Unix /MacOS	Link
ALOGPS2.1	-	log P, log S, and pKa etc.	Web-based	Link
ChemDes	3679	Molecular descriptors from several open source packages (Chemopy, CDK , RDKit , Pybel ,BlueDesc , and PaDEL etc.)	Web-based	Link
E-DRAGON	Over 1,600	Molecular descriptors	Web-based	Link
JOELib	over 40	Topological, geometrical, properties etc.	Windows/Unix/Linux	Link
MODEL	3778	Constitutional, electronic descriptors, topological, and quantum chemistry etc.	Web-based	Link
MOLD2	779	one and two-dimensional molecular descriptors	-	Link
PADEL	863 (1D and 2D: 729 and 3D: 134)	Constitutional, WHIM, topological etc.	Java JRE	Link
PreADMET	955	Constitutional, topological, geometrical, physicochemical etc.	Windows	Link
PowerMV	over 1000	Constitutional, BCUT etc.	Windows	Link

2. BACKGROUND/STATE OF THE ART

Table 2.3: Types of molecular fingerprint used for virtual screening

Type	Description	Name of fingerprints	Notable software for fingerprints
Topological	Derived from the molecular graph by capturing paths of molecular features (e.g., molecular size, shape, branching, presence of heteroatoms and number of connecting bonds) and encode features information as numeric form into bit string patterns.	Daylight and AtomPairs	RDKit, CDK, OEChem TK, Open Babel, and jCompoundMapper etc.
Structural keys	Set each specific bit of the bit string depending on the presences or absence of certain features including functional groups, substructure motifs, or structural fragments from a given list of structural keys.	MACCS, PubChem, and BCI	OEChem TK, RDKit, MOE, Pipeline Pilot, Open Babel, CDK ChemFP, and BCI toolkits etc.
Circular fingerprints	are hashed topological fingerprints that record radial environment of each atom instead of considering molecular paths. Each individual bit in these fingerprints has no specific meaning, therefore, they cannot be suitable for substructure queries in VS.	Molprint2D and ECFP2-6	Open Babel, jCompoundMapper, Pipeline Pilot, Chemaxon's JChem, CDK and RDKit etc.
Pharmacophore fingerprints	Encode the information of features and interactions important for biological activity (pharmacophore). In pharmacophoric fingerprints, a list of distance ranges (patterns) is calculated using three-(or four-) point combination of features and distances (topological in 2D (number of bond lengths), Euclidean in 3D (space)) and generated patterns are then stored in a bit string. Consequently, pharmacophore fingerprints can encode 3D information are usually quite sparse.	Pharmacophore fingerprints in 2D/3D	MOE, jCompoundMapper and Canvas etc.
Hybrid fingerprints	Combine the same bits string bits set using different approaches	UNITY 2D	SYBYL-X

ity searching in chemical databases), diversity selection in virtual combinatorial libraries, chemical/activity spaces visualization and QSAR/QSPR modeling [19, 46, 47].

At a qualitative level, calculation of molecular similarity is a central task in cheminformatics and also played a fundamental role in medicinal chemistry [46, 48, 49, 45, 50]. The underlying idea of molecular similarity bases on the cornerstone *Similar Property Principle*, which states similar compounds should have similar properties [46]. Since molecular similarity is typically evaluated as an indicator of activity similarity, biological activity is extrapolated from the calculated similarity in current studies. Ideally, according to this principle, small structural changes of compounds behave proportionally to all physical and biological properties or in other words increasing structural similarity between two compounds correlates with an increased likelihood to share the same activity. In practice, it is usually observed in the ligand optimization that progressive small structural changes mostly disrupt or destroy

2.1 Virtual screening approaches

compounds biological activity. Hence, overall similarity is not always crucial for the similar activity, whereas, sometimes, the local similarity of the molecules (specific active regions) give rise to related activities. This exception of local similarity concepts falls outside the applicability domain of the *Similar Property Principle*.

Despite the apparent simplicity, the *Similar Property Principle* does not guide methodologically for defining and calculating the structural similarity of two molecules. A variety of methods, structural descriptors/fingerprints (explained earlier in previous sections), and similarity functions/coefficients have been introduced [29, 46, 48, 49, 45, 51, 52, 53, 54]. Irrespective to the specific analysis, molecular similarity values (always lies between 0 to 1) largely depend upon an appropriate combination of two basic components including (a) a molecular structural representation to find the overlapped or similar features and (b) similarity function/coefficient to quantify the similarity between them. Also, sometimes a weighting scheme is also applied if differential weighting is required for important structural features for similarity assessment [29, 46, 48, 49, 45, 51, 52, 53, 54].

Similarity-based virtual screening

Similarity-based VS is also known as ligand-based VS [55, 56]. As it has been explained earlier (Section: molecular similarity) molecular similarity based on *Similar Property Principle*, thus by holding this principle in the similarity-based VS method any library/database of compounds with unknown activity is compared to a set of known active molecules (reference or target structure), and the molecules that share similarity above a certain threshold are considered likely to be active. Although screening can be performed using several other methods or their combination, the simpler and straightforward approach is similarity-based VS [43]. But, in VS, quantification of structural similarity between two compounds is itself complex and bound with a problem of numerical representation of molecular structures, which is still a challenge in cheminformatics [38]. Hence, some level of

2. BACKGROUND/STATE OF THE ART

simplification is required to make a comparison between molecular representations fast and computationally easier. The most popular methods to represent molecular structural features in similarity comparison can be divided into two broader categories including descriptors-dependent methods and descriptors-independent methods. Descriptors-dependent methods include structural descriptors and molecular fingerprints. Descriptors-independent methods include molecular graph matching approaches, which use graph theory to represent a molecule with a labelled graph where vertices correspond to the atoms and edges represent covalent bonds (see the section: Molecular/chemical graph theory). Several molecular graph matching approaches with some advantages and limitations are available to compare labelled graphs [57, 58, 59, 60].

The most commonly used structural representation for comparing molecules is molecular fingerprints (Table 2.3). Different types of molecular fingerprints have different lengths (number of bits) and encode chemical information ranging from simple 2D to complicated higher dimensions (3D/4D). Nonetheless, 2D fingerprints are usually preferred because of the simplicity and high computational efficiency of methods used to calculate them. Moreover, in VS simple 2D chemical information show relatively better performance than the higher dimensions (complex chemical features) that also have been explored for quantifying similarity between pairs of molecules [35].

The molecular comparison is finally assessed to quantify molecular similarity. A variety of (dis)similarity functions/coefficients have been introduced that return a molecular similarity score to indicate the level of similarity between molecules under comparison [48, 53, 61, 62, 54]. In cheminformatics, for molecular fingerprints, the most popular, fast and easily implemented similarity function is Tanimoto coefficient (T_c) [54, 62]. T_c compares binary vectors (any type of fingerprint) of two molecules and quantifies the fraction of the number of “on” bits (feature present) common in the pair of molecules to the total number “on” bits in both molecules under comparison.

2.1 Virtual screening approaches

Similarity-based VS methods have several advantages, e.g., these are (a) adaptable because they can consider any molecular representation that is supportive to similarity comparisons and (b) efficient because do not require models fitting or complex parametrization required for QSAR modelling.

Molecular similarity and chemical/molecular space visualization

An important aspect of molecular similarity is the definition of high dimensional theoretical/ conceptual chemical spaces where molecules are considered as an object and distances between them used to extrapolate biological activity or properties [54]. Size of the chemical space is huge and has no well-defined number. A small fraction of undefined chemical space ranging from thousands to millions of compounds is available in small molecular databases that are used for exploring and visualizing the complexity of chemical structures during drug designing [17, 16, 63]. Chemical space visualization methods combine the concept of molecular structure and activity similarity; however, they critically depend on molecular representations and the way of similarity quantification (discussed earlier: section “molecular similarity”) that is further used to compute metric space (spatial) representation. In a metric space, a molecule is defined as a set or vector of distances measured from the similarity between that molecule to all the other molecules in a given chemical data set. A large number of similarity computations methods are available some of them use molecular descriptors/fingerprints as input molecular representation [29, 46, 48, 49, 45, 51, 52, 53, 54] and others directly use molecular graph matching [57, 58, 59, 60] (see section “molecular structure representation”).

As in chemical space small intermolecular distances represents high structural similarity and similar activity (also known as activity spaces), interactive analysis and visualization of it can describe the molecular diversity of all possible potential biologically active molecules and; thus, can serve as a strong virtual screening tool (see thesis chapter 5).

2. BACKGROUND/STATE OF THE ART

2.1.1.3 Quantitative Structure-Activity Relationship (QSAR): machine learning

The main objective of *in-silico* quantitative structure-activity relationship (QSAR) models-based tools is fast and accurate mining of the large small-molecules databases to find promising lead compounds with desired biological effects. QSAR modeling is as an application of machine learning approaches that have become a trend in the early stages of drug development [64, 65, 66]. Machine learning in drug discovery cycle is used to produce a robust QSAR model, capable of trustful prediction of the pharmacological activity of compounds based on their molecular structural information assuming a strong correlation between structures and biological activity. Thus, QSAR models quantitatively link the variations of the biological activity of molecules (ligands) with changes in their structures (molecular characteristics/properties) [67, 19, 68] (Figure 2.2).

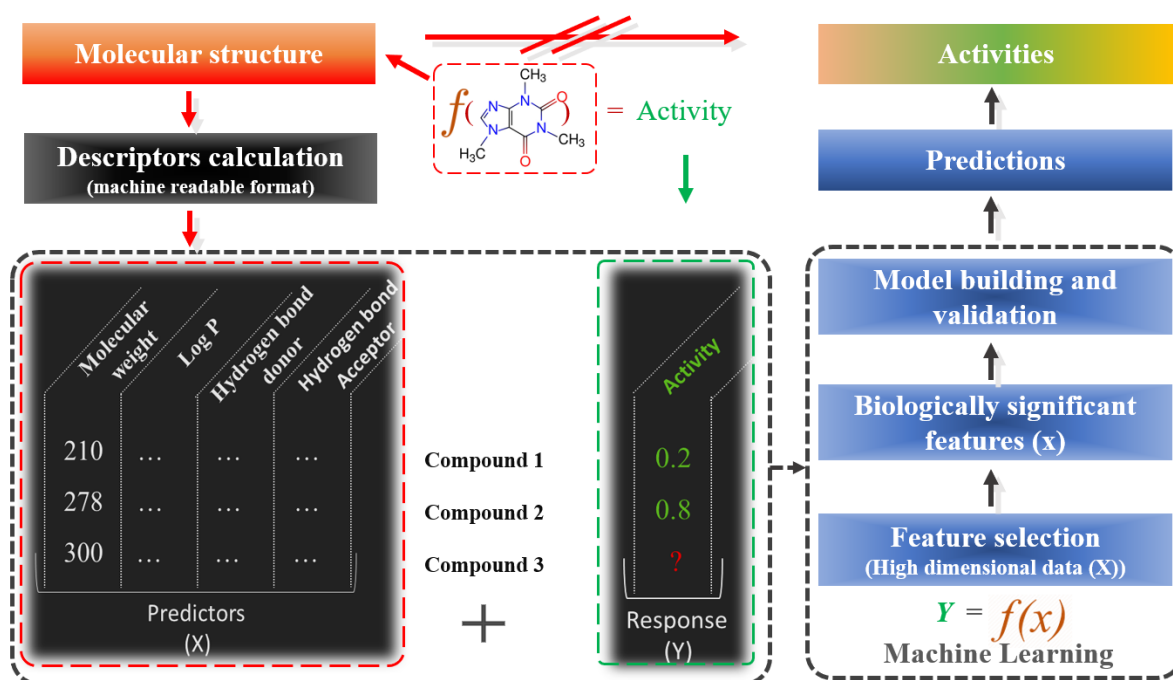


Figure 2.2: Overview of QSAR modeling

Many successful studies have been reported in the literature that show the productiv-

2.1 Virtual screening approaches

ity of machine learning approaches combined with traditional practices to study medicinal chemistry challenges [69]. Moreover, it is accepted that it is not feasible to develop a model providing reliable estimates for all possible compounds. Classical QSAR methodologies have many limitations, specifically (1) the model's predictive power dependency on the variety of highly correlated predictor variables; (2) the molecular diversity and distribution of the molecules in the training set; and (3) model's requirement to be retrained every time with addition/removal of new compounds [70, 71, 72, 73, 74, 75, 76]. However, the reliable predictability of a QSAR model for large and diverse datasets depends on multiple steps involved in the process of model building. Accuracy or mistakes in each preceding step of QSAR modelling cycle may affect the subsequent ones and eventually the overall performance of produced QSAR models [76, 77, 78, 79].

The most critical standard steps for reliable QSAR model building include a) the selection of the chemical dataset for bioactivity type of interest, b) data curation, c) choice of appropriate molecular representation to describe the information encoded in the given structures to generate the input data suitable to the machine learning, d) selection of optimal set of biologically relevant variables (feature selection), e) the selection of algorithms for interpretable model building by the application of one or several machine learning approaches, f) the validation of the built models with an internal test set to assess their quality and for final model selection and, g) an external validation of the selected model with an external test set (Independent Validation Set) to ascertain its predictability for unseen data that never used in model training process [79, 69] (see chapter 3).

Similarity-based VS methods (discussed above) are unable to capture the specific features that drive critical similarities/differences because they treat similarity as a scalar quantity and do not have sufficient data granularity [51]. Hence, to have a granular insight into structure and activity relationships QSAR modeling is an alternative way to find a quantitative relation of sub-structural features to biological activities [80].

2. BACKGROUND/STATE OF THE ART

2.1.2 Molecular docking analysis and structure-based virtual screening

Molecular docking is a computational structure-based VS technique that is used to modeling and predicting best fitting conformation of a ligand into the active/binding site of the 3D structure of a protein (intermolecular complex) [81, 82]. The crucial factor necessary for molecular docking-based hit identification are a) the target structure analysis and selection, b) the software selection and parameters optimization, and c) the ligand database preparation for VS (see thesis chapter 6). Each of these factors is explained in the following sections.

2.1.2.1 Target structure analysis and selection for docking

Analysis and pre-processing of the 3D structure of a target protein is required to get a good quality structure before performing docking [83, 84]. For example, for 3D structures that are produced using X-ray crystallography method, low R-factor and high resolution are used as indicators of good quality structures. Moreover, crystallographic structures due to different experimental conditions may contain salts and other molecules, and also some additives. Structures must be cleaned from these molecules prior to docking analysis. There are many other factors including protonation states, coordinating metal ions, co-factors, water molecules, conformational states (active/inactive), and mutations. The most crucial step to define search boundaries (search space) in docking search algorithm is the binding site selection. Big search space not only increase the computational cost but may reduce the accuracy (high false positive rates) [83, 84, 85, 86, 2, 87, 4].

2.1.2.2 Docking software selection and parameters optimization

Docking software selection and parameters optimization: Docking algorithms in different docking software have two basic components like search algorithm and scoring function [83, 84, 85, 86, 2, 87, 4]. Search algorithms generate “poses” (protein-ligand geometries) of

2.1 Virtual screening approaches

the ligand within the selected active site of the target protein. Scoring function identifies the best position, conformation and orientation of ligand using defined search space boundaries and estimate THE binding affinity (minimum energy-most stable and strong binding). Different docking software (AutoDock, GOLD, and DOCK etc.) use different scoring functions to rank the most likely ligands in VS [88, 6, 3, 83, 84, 85, 86, 2, 87, 4]. Several comparative analysis of docking programs have shown that no software is the absolute best choice across all protein structures [89, 88, 90]. However, if the known binders (ligands with experimental coordinates) are available, several docking softwares should be tested to optimize or to select scoring functions and selection of software that is able to reproduce the experimental pose and affinity of known molecules. The docking results of known molecules-target interactions can also be used as reference score (threshold) for further ranking and evaluating the VS hits [83, 84, 85, 86, 2, 87, 4].

2.1.2.3 Ligands database preparation for virtual screening

Ligands database preparation for VS: The choice of compounds in the database to be screened in VS is the most essential step for a successful finding of potential ligands. Secondly, the compounds structures in the selected database should be cleaned up to prepare their physical states (e.g., the correct protomers, tautomers, and enantiomers) or 3D geometries compatible to the docking program being used. Many other factors are considered depending on the objective of the VS to filter ligands with undesirable physicochemical properties. The most important filters include a) apply properties (molecular weight, logP, polar surface area, and number of hydrogen bond donors and acceptors) cut-offs using the Lipinski rules to choose drug-like compounds, b) lead-like compounds can be selected by eliminating bigger compounds, unattractive as leads for optimization, c) compounds that contain functional groups (e.g., nitro groups) linked with toxicity or the ones whose structures can interfere with the pharmacological assay and also are more prone to aggregation should be removed [91, 2, 86].

2. BACKGROUND/STATE OF THE ART

2.1.3 Large compound repositories for virtual screening

In the past several years of the drug discovery process, the advance high-throughput synthetic and analytical chemical technologies continuously producing a large number of compounds. As a result, the amount of synthesized and known chemical data is exponentially growing with the passage of time. Nonetheless, in recent years, cheminformatics data collection methods made it possible to store the huge heterogeneous experimental testing data in chemical databases that are being available to the research community [41]. However, the availability of large collections of chemical structures in many public (PubChem, ChEMBL, and ZINC etc.) and commercial (ChemDiv, Specs, and Enamine etc.) molecular databases [85, 2] providing the opportunity for developing computational knowledge mining/VS methods. These methods serve as fast and robust tools for the discovery of novel drug candidates against new targets/pathways and also useful for a deep understanding of the relationship between chemical structure and pharmacological properties.

Many researchers usually use in-house virtual libraries for screening while freely available repositories such as ZINC database (www.zinc.docking.org) [92] contains approximately 35M high-quality ligand structures ready to be used in ligand-based or structure-based VS. ZINC database gives several options of compounds selection like custom filtering to getting desired ligands or choosing the pre-defined subsets of ligands e.g., drug-like, lead-like, and fragment-like, ready-to-dock, in-stock categories. Different physicochemical property cut-offs, structural forms/3D formats and availability restrictions are used to prepare the pre-defined subsets of compounds in ZINC.

References

- [1] Florence Stahura and Jurgen Bajorath. ‘Virtual Screening Methods that Complement HTS’. In: *Combinatorial Chemistry & High Throughput Screening* 7.4 (June 2004), pp. 259–269. ISSN: 13862073. DOI: 10.2174/1386207043328706.
- [2] Xavier Fradera and Kerim Babaoglu. ‘Overview of Methods and Strategies for Conducting Virtual Small Molecule Screening’. In: *Current protocols in chemical biology* 9.3 (2017), pp. 196–212. ISSN: 21604762. DOI: 10.1002/cpch.27.
- [3] Milena Lazarova. ‘Virtual Screening – Models , Methods and Software Systems’. In: *International Scientific Conference Computer Science* (2008).
- [4] Campbell McInnes. ‘Virtual screening strategies in drug discovery’. In: *Current Opinion in Chemical Biology* 11.5 (2007), pp. 494–502. ISSN: 13675931. DOI: 10.1016/j.cbpa.2007.08.033.
- [5] Bruno J. Neves et al. ‘QSAR-based virtual screening: Advances and applications in drug discovery’. In: *Frontiers in Pharmacology* 9.NOV (2018), pp. 1–7. ISSN: 16639812. DOI: 10.3389/fphar.2018.01275.
- [6] Andrew R. Leach and Valerie J. Gillet. *An Introduction To Chemoinformatics*. Dordrecht: Springer Netherlands, 2007. ISBN: 978-1-4020-6290-2. DOI: 10.1007/978-1-4020-6291-9.
- [7] David Wilton et al. ‘Comparison of Ranking Methods for Virtual Screening in Lead-Discovery Programs’. In: *Journal of Chemical Information and Computer Sciences* 43.2 (Mar. 2003), pp. 469–474. ISSN: 0095-2338. DOI: 10.1021/ci025586i.
- [8] Robert D. Brown and Yvonne C. Martin. ‘The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding’. In: *Journal of Chemical Information and Computer Sciences* 37.1 (Jan. 1997), pp. 1–9. ISSN: 0095-2338. DOI: 10.1021/ci960373c.

2. BACKGROUND/STATE OF THE ART

- [9] Hans Matter. ‘Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors.’ In: *Journal of medicinal chemistry* 40.8 (Apr. 1997), pp. 1219–29. ISSN: 0022-2623. DOI: 10.1021/jm960352+.
- [10] Steven L. Dixon and Kenneth M. Merz. ‘One-Dimensional Molecular Representations and Similarity Calculations: Methodology and Validation’. In: *Journal of Medicinal Chemistry* 44.23 (Nov. 2001), pp. 3795–3809. ISSN: 0022-2623. DOI: 10.1021/jm010137f.
- [11] Johannes Kirchmair et al. ‘How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information’. In: *Journal of Chemical Information and Modeling* 49.3 (Mar. 2009), pp. 678–692. ISSN: 1549-9596. DOI: 10.1021/ci8004226.
- [12] Ingo Muegge and Prasenjit Mukherjee. ‘An overview of molecular fingerprint similarity search in virtual screening’. In: *Expert Opinion on Drug Discovery* 11.2 (2016), pp. 137–148. ISSN: 1746-0441. DOI: 10.1517/17460441.2016.1117070.
- [13] Hongmao Sun. ‘Pharmacophore-Based Virtual Screening’. In: *Current Medicinal Chemistry* 15.10 (Apr. 2008), pp. 1018–1024. ISSN: 09298673. DOI: 10.2174/092986708784049630.
- [14] Yu-Chen Lo et al. ‘Machine learning in chemoinformatics and drug discovery’. In: *Drug Discovery Today* 23.8 (Aug. 2018), pp. 1538–1546. DOI: 10.1016/j.drudis.2018.05.010.
- [15] James Melville, Edmund Burke and Jonathan Hirst. ‘Machine Learning in Virtual Screening’. In: *Combinatorial Chemistry & High Throughput Screening* 12.4 (May 2009), pp. 332–343. ISSN: 13862073. DOI: 10.2174/138620709788167980.
- [16] Jean Louis Reymond et al. ‘Chemical space as a source for new drugs’. In: *Med-ChemComm* 1.1 (2010), pp. 30–38. ISSN: 20402503. DOI: 10.1039/c0md00020e.

REFERENCES

- [17] Mahendra Awale et al. 'Chemical Space: Big Data Challenge for Molecular Diversity'. In: *CHIMIA International Journal for Chemistry* 71.10 (2017), pp. 661–666. ISSN: 0009-4293. DOI: [10.2533/chimia.2017.661](https://doi.org/10.2533/chimia.2017.661).
- [18] Natália Aniceto et al. 'A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: Reliability-density neighbourhood'. In: *Journal of Cheminformatics* 8.1 (2016), pp. 1–20. ISSN: 17582946. DOI: [10.1186/s13321-016-0182-y](https://doi.org/10.1186/s13321-016-0182-y).
- [19] Johann Gasteiger. *Handbook of Chemoinformatics : From Data to Knowledge*. Vol. 1-4. May. Weinheim: Wiley-VCH, 2008, pp. 1–1870. ISBN: 9783527618279. DOI: [10.1002/9783527618279](https://doi.org/10.1002/9783527618279).
- [20] Alexandre Varnek and Igor I. Baskin. *Chemoinformatics as a theoretical chemistry discipline*. 2011. DOI: [10.1002/minf.201000100](https://doi.org/10.1002/minf.201000100).
- [21] Frank K. Brown. 'Chapter 35 - Chemoinformatics: What is it and How does it Impact Drug Discovery.' In: ed. by James A. Bristol. Vol. 33. *Annual Reports in Medicinal Chemistry*. Academic Press, 1998, pp. 375–384. DOI: [https://doi.org/10.1016/S0065-7743\(08\)61100-8](https://doi.org/10.1016/S0065-7743(08)61100-8).
- [22] Johann Gasteiger. *Handbook of Chemoinformatics*. Ed. by Johann Gasteiger. Vol. 1-4. Weinheim, Germany: Wiley-VCH Verlag GmbH, Aug. 2003, pp. 1–1870. ISBN: 9783527618279. DOI: [10.1002/9783527618279](https://doi.org/10.1002/9783527618279).
- [23] Jürgen Bajorath. *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*. Ed. by Jürgen Bajorath. Vol. 275. *Methods in Molecular Biology*. Totowa: Humana Press, 2004. ISBN: 978-1-58829-261-2. DOI: [10.1385/1592598021](https://doi.org/10.1385/1592598021).
- [24] D. Bonchev. *Chemical Graph Theory: Introduction and Fundamentals*. CRC Press, 2018. ISBN: 9781351461597. URL: <https://books.google.pt/books?id=5X1aDwAAQBAJ>.

2. BACKGROUND/STATE OF THE ART

- [25] Ramón García-Domenech et al. ‘Some New Trends in Chemical Graph Theory’. In: *Chemical Reviews* 108.3 (Mar. 2008), pp. 1127–1169. ISSN: 0009-2665. DOI: 10.1021/cr0780006.
- [26] Jürgen Bajorath. *Cheminformatics and Computational Chemical Biology*. Ed. by Jürgen Bajorath. Vol. 672. Methods in Molecular Biology. Totowa, NJ: Humana Press, 2011. ISBN: 978-1-60761-838-6. DOI: 10.1007/978-1-60761-839-3.
- [27] Thomas H. Cormen et al. *Introduction to Algorithms, 3rd Edition*. MIT Press, 2009. ISBN: 978-0-262-03384-8. URL: <http://mitpress.mit.edu/books/introduction-algorithms>.
- [28] Denis Fourches and Alexander Tropsha. ‘Using graph indices for the analysis and comparison of chemical datasets’. In: *Molecular Informatics* 32.9-10 (2013), pp. 827–842. ISSN: 18681751. DOI: 10.1002/minf.201300076.
- [29] Nina Nikolova and Joanna Jaworska. ‘Approaches to Measure Chemical Similarity—a Review’. In: *QSAR & Combinatorial Science* 22.910 (2003), pp. 1006–1026. ISSN: 1611-020X. DOI: 10.1002/qsar.200330831.
- [30] Alan R. Katritzky, Victor S. Lobanov and Mati Karelson. ‘QSPR: the correlation and quantitative prediction of chemical and physical properties from structure’. In: *Chemical Society Reviews* 24.4 (1995), pp. 279–87. ISSN: 0306-0012. DOI: 10.1039/cs9952400279.
- [31] Roberto Todeschini and Viviana Consonni. *Molecular Descriptors for Chemoinformatics*. Ed. by Roberto Todeschini and Viviana Consonni. Methods and Principles in Medicinal Chemistry. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, July 2009. ISBN: 9783527628766. DOI: 10.1002/9783527628766.
- [32] Jean-Loup Faulon and Andreas Bender. *Handbook of Chemoinformatics Algorithms*. Vol. 20103144. Chapman & Hall/CRC Mathematical & Computational Biology.

REFERENCES

- Chapman and Hall/CRC, Apr. 2010. ISBN: 978-1-4200-8292-0. DOI: 10.1201/9781420082999.
- [33] Jürgen Bajorath. ‘Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening’. In: *Journal of Chemical Information and Computer Sciences* 41.2 (Mar. 2001), pp. 233–245. ISSN: 0095-2338. DOI: 10.1021/ci0001482.
- [34] Carolina H Andrade et al. ‘4D-QSAR: Perspectives in Drug Design’. In: *Molecules* 15.5 (May 2010), pp. 3281–3294. ISSN: 1420-3049. DOI: 10.3390/molecules15053281.
- [35] Hans Matter and Thorsten Pötter. ‘Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets’. In: *Journal of Chemical Information and Computer Sciences* 39.6 (Nov. 1999), pp. 1211–1225. ISSN: 0095-2338. DOI: 10.1021/ci980185h.
- [36] Asad U Khan. ‘Descriptors and their selection methods in QSAR analysis : paradigm for drug design’. In: 21.8 (2016), pp. 1291–1302.
- [37] John W. Raymond and Peter Willett. ‘Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases.’ In: *Journal of computer-aided molecular design* 16.1 (Jan. 2002), pp. 59–71. URL: <http://www.ncbi.nlm.nih.gov/pubmed/12197666>.
- [38] Adria Cereto-Massagué et al. ‘Molecular fingerprint similarity search in virtual screening’. In: *Methods* 71 (Jan. 2015), pp. 58–63. ISSN: 10462023. DOI: 10.1016/j.ymeth.2014.08.005.
- [39] Kunal Roy, Supratik Kar and Rudra Narayan Das. *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*. Elsevier, 2015. ISBN: 9780128015056. DOI: 10.1016/C2014-0-00286-9.

2. BACKGROUND/STATE OF THE ART

- [40] Alexander Tropsha and Alexander Golbraikh. 'Predictive QSAR modeling workflow, model applicability domains, and virtual screening.' In: *Current pharmaceutical design* 13.34 (2007), pp. 3494–504. ISSN: 1873-4286. DOI: 10.2174/13816120778279425
- [41] Alexandre Varnek and Alex Tropsha. *Chemoinformatics Approaches to Virtual Screening*. Ed. by Alexandre Varnek and Alex Tropsha. Cambridge: Royal Society of Chemistry, 2008. ISBN: 978-0-85404-144-2. DOI: <http://dx.doi.org/10.1039/9781847558879>.
- [42] C. James, D. Weininger and J. Delaney. *Daylight Theory Manual version 4.9*. 2011.
- [43] P Willett. 'Similarity-based approaches to virtual screening.' In: *Biochemical Society transactions* 31.Pt 3 (2003), pp. 603–606. ISSN: 0300-5127. DOI: 10.1042/.
- [44] Peter Willett. 'Similarity-based virtual screening using 2D fingerprints'. In: *Drug Discovery Today* 11.23-24 (Dec. 2006), pp. 1046–1053. ISSN: 13596446. DOI: 10.1016/j.drudis.2006.10.005.
- [45] Gerald Maggiora et al. 'Molecular Similarity in Medicinal Chemistry'. In: *Journal of Medicinal Chemistry* 57.8 (Apr. 2014), pp. 3186–3204. ISSN: 0022-2623. DOI: 10.1021/jm401411z.
- [46] Mark A Johnson and Gerald M Maggiora. *Concepts and applications of molecular similarity*. 1990. ISBN: 0471621757.
- [47] R Carbo-Dorca and P G Mezey. *Advances in Molecular Similarity*. Advances in Molecular Similarity v. 2. Elsevier Science, 1999, p. 296. ISBN: 9780080552262.
- [48] Peter Willett, John M. Barnard and Geoffrey M. Downs. 'Chemical Similarity Searching'. In: *Journal of Chemical Information and Computer Sciences* 38.6 (Nov. 1998), pp. 983–996. ISSN: 0095-2338. DOI: 10.1021/ci9800211.
- [49] Andreas Bender and Robert C Glen. 'Molecular similarity: a key technique in molecular informatics.' In: *Organic & biomolecular chemistry* 2.22 (2004), pp. 3204–3218. ISSN: 1477-0520. DOI: 10.1039/b409813g.

REFERENCES

- [50] Jens Auer and Jürgen Bajorath. ‘Molecular Similarity Concepts and Search Calculations’. In: *Methods in Molecular Biology*. 2008, pp. 327–347. ISBN: 9781603274289. DOI: 10.1007/978-1-60327-429-6_17.
- [51] Hanna Eckert and Jürgen Bajorath. ‘Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches’. In: *Drug Discovery Today* 12.5-6 (2007), pp. 225–233. ISSN: 13596446. DOI: 10.1016/j.drudis.2007.01.011.
- [52] Dagmar Stumpfe and Jürgen Bajorath. ‘Similarity searching’. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.2 (Mar. 2011), pp. 260–282. ISSN: 17590876. DOI: 10.1002/wcms.23.
- [53] Gerald M Maggiora and Veerabahu Shanmugasundaram. ‘Molecular Similarity Measures’. In: *Methods in molecular biology (Clifton, N.J.)* 2004, pp. 1–50. DOI: 10.1385/1-59259-802-1:001.
- [54] Jürgen Bajorath. ‘Molecular Similarity Concepts for Informatics Applications’. In: *Bioinformatics: Volume II: Structure, Function, and Applications*. Ed. by Jonathan M. Keith. New York, NY: Springer New York, 2017, pp. 231–245. ISBN: 978-1-4939-6613-4. DOI: 10.1007/978-1-4939-6613-4_13.
- [55] Brian K. Shoichet. ‘Virtual screening of chemical libraries’. In: *Nature* 432.7019 (2004), pp. 862–865. ISSN: 0028-0836. DOI: 10.1038/nature03197.
- [56] Patrick W. Walters, Matthew T. Stahl and Mark A. Murcko. ‘Virtual screening—an overview’. In: *Drug Discovery Today* 3.4 (1998), pp. 160–178. ISSN: 13596446. DOI: 10.1016/S1359-6446(97)01163-X.
- [57] Ana L Teixeira and Andre O Falcao. ‘Noncontiguous atom matching structural similarity function’. In: *Journal of Chemical Information and Modeling* 53.10 (2013), pp. 2511–2524. ISSN: 15499596. DOI: 10.1021/ci400324u.

2. BACKGROUND/STATE OF THE ART

- [58] Hans-Christian Ehrlich and Matthias Rarey. ‘Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review’. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.1 (Jan. 2011), pp. 68–79. ISSN: 17590876. DOI: 10.1002/wcms.5.
- [59] John W. Raymond and Peter Willett. ‘Maximum common subgraph isomorphism algorithms for the matching of chemical structures.’ In: *Journal of computer-aided molecular design* 16.7 (July 2002), pp. 521–33. ISSN: 0920-654X. DOI: 10.1023/A:1021271615909.
- [60] John M. Barnard. ‘Substructure searching methods: Old and new’. In: *Journal of Chemical Information and Modeling* 33.4 (July 1993), pp. 532–538. ISSN: 1549-9596. DOI: 10.1021/ci00014a001.
- [61] Amos Tversky. ‘Features of similarity.’ In: *Psychological Review* 84.4 (1977), pp. 327–352. ISSN: 0033-295X. DOI: 10.1037/0033-295X.84.4.327.
- [62] Dávid Bajusz, Anita RÁCz and Károly Héberger. ‘Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?’ In: *Journal of Cheminformatics* 7.1 (2015), pp. 1–13. ISSN: 17582946. DOI: 10.1186/s13321-015-0069-3.
- [63] Christopher M. Dobson. ‘Chemical space and biology’. In: *Nature* 432.7019 (Dec. 2004), pp. 824–828. ISSN: 0028-0836. DOI: 10.1038/nature03192.
- [64] Shivani Agarwal, Deepak Dugar and Shiladitya Sengupta. ‘Ranking Chemical Structures for Drug Discovery : A New Machine Learning Approach’. In: (2010), pp. 716–731.
- [65] Kun Yi Hsin, Samik Ghosh and Hiroaki Kitano. ‘Combining machine learning systems and multiple docking simulation packages to improve docking prediction reliability for network pharmacology’. In: *PLoS ONE* 8.12 (2013). ISSN: 19326203. DOI: 10.1371/journal.pone.0083922.

REFERENCES

- [66] Atsushi Matsumoto, Shin Aoki and Hayato Ohwada. ‘Comparison of Random Forest and SVM for Raw Data in Drug Discovery : Prediction of Radiation Protection and Toxicity Case Study’. In: 6.2 (2016), pp. 145–148. DOI: 10.18178/ijmlc.2016.6.2.589.
- [67] Roberto Todeschini and Viviana Consonni. *Handbook of Molecular Descriptors*. Vol. 11. July. Wiley-VCH Verlag GmbH, Weinheim, Germany, 2008, p. 688. ISBN: 9783527613106. DOI: 10.1002/9783527613106.
- [68] Noel M. O’Boyle and Roger A Sayle. ‘Comparing structural fingerprints using a literature-based similarity benchmark’. In: *Journal of Cheminformatics* 8.1 (Dec. 2016), p. 36. ISSN: 1758-2946. DOI: 10.1186/s13321-016-0148-0.
- [69] Angélica Nakagawa Lima et al. ‘Use of machine learning approaches for novel drug discovery.’ In: *Expert opinion on drug discovery* 11.3 (2016), pp. 225–239. ISSN: 1746-045X. DOI: 10.1517/17460441.2016.1146250.
- [70] Igor V. Tetko et al. ‘Accurate In Silico log P_{ij}/i_i Predictions: One Can’t Embrace the Unembraceable’. In: *QSAR & Combinatorial Science* 28.8 (2009), pp. 845–849. ISSN: 1611020X. DOI: 10.1002/qsar.200960003.
- [71] Qiang Zhang and Ingo Muegge. ‘Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting, and Consensus Scoring’. In: *Journal of Medicinal Chemistry* 49.5 (Mar. 2006), pp. 1536–1548. ISSN: 0022-2623. DOI: 10.1021/jm050468i.
- [72] Georgia Melagraki et al. ‘In silico exploration for identifying structure-activity relationship of MEK inhibition and oral bioavailability for isothiazole derivatives’. In: *Chemical Biology and Drug Design* 76.5 (2010), pp. 397–406. ISSN: 17470277. DOI: 10.1111/j.1747-0285.2010.01029.x.

2. BACKGROUND/STATE OF THE ART

- [73] Anderson Coser Gaudio and Eliana Zandonade. 'PROPOSITION, VALIDATION AND ANALYSIS OF QSAR MODELS'. In: *Quim Nova. SBQ* 24.5 (2001), pp. 658–671. ISSN: 01004042. DOI: 10.1590/S0100-40422001000500013.
- [74] M. M C Ferreira. 'Multivariate QSAR'. In: *Journal of the Brazilian Chemical Society*. Vol. 13. 6. 2002, pp. 742–753. ISBN: 0103-5053. DOI: 10.1590/S0103-50532002000600004.
- [75] Douglas M. Hawkins. 'The Problem of Overfitting'. In: *Journal of Chemical Information and Computer Sciences* 44.1 (2004), pp. 1–12. ISSN: 00952338. DOI: 10.1021/ci0342472.
- [76] Artem Cherkasov et al. 'QSAR Modeling: Where Have You Been? Where Are You Going To?' In: *Journal of Medicinal Chemistry* 57.12 (June 2014), pp. 4977–5010. ISSN: 0022-2623. DOI: 10.1021/jm4004285.
- [77] Igor V. Tetko et al. 'Critical assessment of QSAR models of environmental toxicity against tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection'. In: *Journal of Chemical Information and Modeling* 48.9 (2008), pp. 1733–1746. ISSN: 15499596. DOI: 10.1021/ci800151m.
- [78] Douglas Young et al. 'Are the chemical structures in your QSAR correct?' In: *QSAR and Combinatorial Science* 27.11-12 (2008), pp. 1337–1345. ISSN: 1611020X. DOI: 10.1002/qsar.200810084.
- [79] Alexander Tropsha. 'Best practices for QSAR model development, validation, and exploitation'. In: *Molecular Informatics* 29.6-7 (2010), pp. 476–488. ISSN: 18681743. DOI: 10.1002/minf.201000061.
- [80] Spencer M. Free and James W. Wilson. 'A Mathematical Contribution to Structure-Activity Studies'. In: *Journal of Medicinal Chemistry* 7.4 (1964), pp. 395–399. ISSN: 15204804. DOI: 10.1021/jm00334a001.

REFERENCES

- [81] Xuan-Yu Meng et al. 'Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery'. In: *Current Computer Aided-Drug Design* 7.2 (June 2011), pp. 146–157. ISSN: 15734099. DOI: 10.2174/157340911795677602.
- [82] Nataraj S. Pagadala, Khajamohiddin Syed and Jack Tuszynski. 'Software for molecular docking: a review.' In: *Biophysical reviews* 9.2 (2017), pp. 91–102. ISSN: 1867-2450. DOI: 10.1007/s12551-016-0247-1.
- [83] Andrew M Davis, Stephen A St-Gallay and Gerard J Kleywegt. 'Limitations and lessons in the use of X-ray structural information in drug design.' In: *Drug discovery today* 13.19-20 (Oct. 2008), pp. 831–41. ISSN: 1359-6446. DOI: 10.1016/j.drudis.2008.06.006.
- [84] David R Cooper et al. 'X-ray crystallography: assessment and validation of protein–small molecule complexes for drug discovery'. In: *Expert Opinion on Drug Discovery* 6.8 (Aug. 2011), pp. 771–782. ISSN: 1746-0441. DOI: 10.1517/17460441.2011.585154.
- [85] Maria Kontoyianni. 'Docking and Virtual Screening in Drug Discovery'. In: *Proteomics for Drug Discovery: Methods and Protocols*. Ed. by Iulia M. Lazar, Maria Kontoyianni and Alexandru C. Lazar. New York, NY: Springer New York, 2017, pp. 255–266. ISBN: 978-1-4939-7201-2. DOI: 10.1007/978-1-4939-7201-2_18.
- [86] Stefano Forli. 'Charting a Path to Success in Virtual Screening'. In: *Molecules* 20.10 (Oct. 2015), pp. 18732–18758. ISSN: 1420-3049. DOI: 10.3390/molecules201018732.
- [87] Qingliang Li and Salim Shah. 'Structure-Based Virtual Screening'. In: *Methods in Molecular Biology*. 2017, pp. 111–124. DOI: 10.1007/978-1-4939-6783-4_5.

2. BACKGROUND/STATE OF THE ART

- [88] Renxiao Wang, Yipin Lu and Shaomeng Wang. ‘Comparative Evaluation of 11 Scoring Functions for Molecular Docking’. In: *Journal of Medicinal Chemistry* 46.12 (June 2003), pp. 2287–2303. ISSN: 0022-2623. DOI: 10.1021/jm0203783.
- [89] Douglas B. Kitchen et al. ‘Docking and scoring in virtual screening for drug discovery: methods and applications’. In: *Nature Reviews Drug Discovery* 3.11 (Nov. 2004), pp. 935–949. ISSN: 1474-1776. DOI: 10.1038/nrd1549.
- [90] Sheng-You Huang, Sam Z. Grinter and Xiaoqin Zou. ‘Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions’. In: *Physical Chemistry Chemical Physics* 12.40 (2010), p. 12899. ISSN: 1463-9076. DOI: 10.1039/c0cp00151a.
- [91] Jürgen Bajorath. ‘Activity artifacts in drug discovery and different facets of compound promiscuity’. In: *F1000Research* 3 (Oct. 2014), p. 233. ISSN: 2046-1402. DOI: 10.12688/f1000research.5426.1.
- [92] John J. Irwin and Brian K. Shoichet. ‘ZINC—a free database of commercially available compounds for virtual screening.’ In: *Journal of chemical information and modeling* 45.1 (2005), pp. 177–82. ISSN: 1549-9596. DOI: 10.1021/ci049714+.

3

An automated framework for QSAR model building

SAMINA KAUSAR AND ANDRE O FALCAO

Abstract

Background: *In-silico* Quantitative-Structure Activity Relationship (QSAR) models based tools are widely used to screen huge databases of compounds in order to determine the biological properties of chemical molecules based on their chemical structure. With the passage of time, the exponentially growing amount of synthesized and known chemicals data demands computationally efficient automated QSAR modeling tools, available to researchers that may lack extensive knowledge of machine learning modeling. Thus, a fully automated and advanced modeling platform can be an important addition to the QSAR community. **Results:** In the presented workflow the process from data preparation to model building and valida-

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

tion has been completely automated. The most critical modeling tasks (data curation, data set characteristics evaluation, variable selection and validation) that largely influence the performance of QSAR models were focused. It is also included the ability to quickly evaluate the feasibility of a given data set to be modeled. The developed framework is tested on data sets of thirty different problems. The best-optimized feature selection methodology in the developed workflow is able to remove 62% to 99% of all redundant data. On average, about 19% of the prediction error (RMSE) was reduced by using feature selection producing an increase of 49% in the percentage of variance explained (PVE) compared to models without feature selection. Selecting only the models with a modelability score above 0.6, average PVE scores were 0.71. A strong correlation was verified between the modelability scores and the PVE of the models produced with variable selection. **Conclusions:** We developed an extendable and highly customizable fully automated QSAR modeling framework. This designed workflow does not require any advanced parameterization nor depends on users decisions or expertise in machine learning/programming. With just a given target or problem, the workflow follows an unbiased standard protocol to develop reliable QSAR models by directly accessing online manually curated databases or by using private data sets. The other distinctive features of the workflow include prior estimation of data modelability to avoid time-consuming modeling trials for non modelable data sets, an efficient variable selection procedure and the facility of output availability at each modeling task for the diverse application and reproduction of historical predictions. The results reached on a selection of thirty QSAR problems suggest that the approach is capable of building reliable models even for challenging problems.

Keywords: Quantitative-Structure Activity Relationship (QSAR); Machine Learning; Feature selection; Variable importance; Random forests; Support Vector Machines; KNIME; Data set modelability

3.1 Introduction

3.1.1 Background

The advantages of automation of repetitive tasks in the laborious drug discovery process are numerous and include increased research quality by reducing error along with significant time saving, boosted up productivity, and capacity to name a few. In this era where large amounts of data are produced every day and large computational resources are available, the introduction of machine learning approaches has significantly automated the drug discovery procedure and provides a faster alternative for ultrahigh-throughput screening of large databases of chemical molecules against a biological target [1, 2, 3].

Machine learning approaches are being applied in the drug discovery cycle to produce a robust model, capable of empirical predictions of biological properties of candidate compounds for new therapeutic molecules. Many successful studies have been reported in the literature which attests the importance of machine learning approaches combined with traditional practices to approach medicinal chemistry challenges [4]. In traditional lab work methodologies, many expensive tests are often required which many times include animal testing to provide information about human safety for suggested chemicals. The legislation does not support such frequent experiments on laboratory animals, but rather promotes the sharing of data to the use of integrated alternative *in-vitro* and *in-silico* strategies of toxicokinetics [5, 6, 7]. Currently the Avicenna Research and Technological Roadmap, funded by the European Commission, strongly suggests the use of *in-silico* techniques coupled with clinical trials [8]. This framework describes strategic priorities to establish the safety assessment of new medical interventions and at the same time minimizes the ethically concerned activities such as the animal or human experimentation.

Several available *in-silico* QSAR models based tools are widely used to screen very large databases of compounds in order to determine toxicity or any desired biological effects of

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

chemical molecules based on their chemical structure [9, 10]. The well-characterized internationally accepted validation principles for creating validated models have been used by regulatory agencies of United States (US) and gaining a boost in the European Union (EU) too [11, 12, 8, 13]. In the EU, the standard recommendations of chemicals risk assessment by regulatory QSAR models has been set by the REACH (Registration, Evaluation, Authorization, and Restriction of Chemicals)[14] and the Organization for Economic Co-operation and Development (OECD)[15]. The progress of such projects highlights the increased importance of productivity gains from fully accessible automation in the drug discovery and QSAR modeling fields.

These days, the aim of pharmaceutical projects is the integration of complex non-homogeneous data to build global models intended to be applicable within wide ranges of chemical space. However, with the passage of time, there is an exponentially growing amount of synthesized and known chemical compounds data being added to the many existing molecule databases, public or private. This rise of available data is producing new opportunities to build models with broader applicability domains while at the same time challenging the existing models, as wider data sets allow for a more extensive testing and validation of previous *in-silico* screening efforts. From these databases, data can easily be explored to build QSAR models based on available structural properties of the compounds that correlate with their biological activity [16, 17, 18]. These models can also be used as an efficient tool to improve the understanding of biological processes. Also, well-trained and properly validated models are reliable for automated prediction of physiological characteristics of new compounds to assist the experimental drug discovery process by decreasing the time of the initial screening stages [19, 20, 21, 22].

The QSAR/QSPR modeling “life cycle” involves some standard steps, critical for reliable model building. These steps include 1) model building by the application of one or several machine learning approaches, 2) model validation with an internal test set to assess its quality 3) model selection according to the results of the internal validation procedure,

and 4) model validation with an external test set (Independent Validation Set) to ascertain its predictability of the properties of compounds never tested in model building and thus giving a more reliable measure of the selected model quality [22, 4]. It is also important to consider model updating as new data may become available. This repetitive nature of QSAR/QSPR modeling “life cycle” highlights a fundamental requirement of automation of critical steps with well-defined input, outputs, and success criteria in both the drug discovery industry and biomedical research. To achieve this objective, it is fundamental to have a scrutinizable procedure for applying to a variety of problems. Automating such procedures in the form of a reusable workflow is a reachable goal with current technology, provided that a reliable method is extant and applicable to a wide range of problems. Such automation would reduce the necessary and often tedious labor of model building, while at the same time guaranteeing that, for the available data, a quality model is reached.

Over the past decade, attempts have been made to attract the attentions towards the need of automation of the QSAR modelling process. More recently, Dixon et al. [23] developed a machine-learning application (AutoQSAR) for automated QSAR modeling. It is unable to access data directly from online repositories and users required deep understanding to prepare a curated and standardized data set before modeling by AutoQSAR. eTOXlab [24] which is another framework allows automated QSAR mainly by a command line interface. Python programming skills are necessary to work with eTOXlab. An interesting alternative of integrated solution for fully automated modeling is OCHEM [25] but it's online nature makes it unsuitable for using it with private/sensitive data sets, which demand better privacy and safety independent of third party. Cox et al. [26] designed a Pipeline Pilot web application (QSAR Workbench). This application makes the built models available to all users in Pipeline Pilot [27], which is not freely available to the vast scientific community. The Automated Predictive Modeling, another modeling system [28], demands expert technical skills and significant resources for model development and maintenance.

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

3.1.2 Objectives

Some of the major pinpointed gaps in the above discussed software packages include lack of fully automated process, require that users have a thorough understanding of the data and modeling problems and several require computer programming and/or machine-learning knowledge, complex parameterization to customize complex modeling algorithms, and most do not give full access to view the intermediate results at each step of the modeling. Also to the best of our knowledge, none of these packages provide a facility to check overall data quality/feasibility to produce a robust QSAR model (data modelability), which can be an important measure to minimize time and computational cost. In the current work, we developed an open source automated QSAR modeling system that addresses these issues by providing better solutions for expert and non-expert users. The key ideas behind structuring the presented automated QSAR modeling workflow platform are:

- It should be freely available and support any operating system with easy installation
- Should be easily be applied for fully automated QSAR modeling by directly accessing up to date data from online molecules databases or by using private data sets
- Provide automated data curation facility including removal of irrelevant data by selecting only the bioactivity type of interest, filtering out missing data, handling of duplicates (e.g. same or two experimental records: same structure) and dealing with several forms of the same molecule (including salt groups)
- Reliably perform most critical tasks of QSAR modeling including descriptor/fingerprints calculation, feature selection, model building, validation, and prediction
- Make a prior estimation of the feasibility of any given data set to produce a predictive QSAR model before the time-consuming process of feature selection, model building and validation

- It should adopt the best optimized feature selection methodology to select the adequate features for each problem. This is a critical task necessary to avoid over-fitting and to have a better understanding of the data, the model and the factors involved.
- The application must follow the same protocol of training series to re-train and update models with new molecules as they become available and to make external predictions
- For different applications and reproduction of historical predictions, all outputs of intermediate tasks and each previous version of models must be stored on local machines.
- Regarding extensibility, the framework should provide useful starting points for performing customization to modify and further extend the existing workflow by domain specific interests.

Many research labs aim to develop their own complete workflow by using workflow automation tools for a broader domain of related biological problems [29, 30, 31, 32]. Some of the more popular workflow frameworks include Taverna [33], Pipeline Pilot [27], Galaxy[34], Kepler [35], Loni Pipeline [36] or the KoNstanz Information MinEr (KNIME) [37]. These well-deployed workflows with graphical user interface provide a clear view of the running process rather than working as a black box, or with complex and opaque code. Moreover, it is an efficient way to manage complex chemical data to help standardize procedures, automate laborious procedures, and assist in data analysis [29]. For the current study, we have selected KNIME, an open source data-mining framework developed by the Nycomed Chair for Bioinformatics and Information Mining at the University of Konstanz to manipulate and analyze data with a strong emphasis on chemical manipulation and information management. KNIME has made it easy to perform the calculation of molecular descriptors to quantify molecular structures, evaluation of chemical similarity and other cheminformatics problems [CDK [38], RDKit [39], Schrodinger [40, 41], ChEMBL [42], OpenPHACTS [43], BioSolveIT (<http://www.biosolveit.de/KNIME>).

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

The developed open source automated QSAR modeling KNIME workflow embeds all tools necessary to perform all steps of the QSAR life cycle by following best practicing methods [22, 44]. This designed workflow can easily be applied to build the predictive QSAR models reliably by directly accessing online manually curated databases or using users own private data without having expertise in machine learning/programming. In this work, we illustrate and describe a model building workflow with an optimized feature selection methodology and show its application in real world examples, by directly fetching binding data for thirty different QSAR problems from an online manually curated database (ChEMBL [42]) and building models using runtime prepared processed data. The workflow, given a target or problem, automatically accesses and processes molecular data, calculates descriptors and fingerprints, evaluates data set modelability, selects optimized set of features by using an established methodology [45] and follows an unbiased standard protocol [22, 44] of QSAR model building by external and internal validation. The objective of this work is not to highlight the predictive power of the presented models but rather to elaborate a reliable methodology to automate the production of models with good predictive qualities for very difficult problems. Nonetheless, the quality of the results suggests that the approach is capable of building reliable models for a large variety of problems.

3.2 Automated model building

The main focus of the current work is to present an implementation of a well-defined and efficient modeling procedure capable of building robust and reliable models and validate them both internally and externally. To accomplish this it was necessary to address two critical issues in QSAR modeling. The first one is to know how to deal with high dimensional data by identifying and selecting the subset of descriptors sufficient to predict the desired biochemical property. The second aspect in a modeling workflow is model validation, so that the model results can be unbiasedly assessed. This will ultimately qualify the applicability

3.2 Automated model building

of the model for activity prediction of external compounds in drug discovery processes [22]. An overview of the standard protocol of automated QSAR modeling workflow is shown in Figure 3.1.

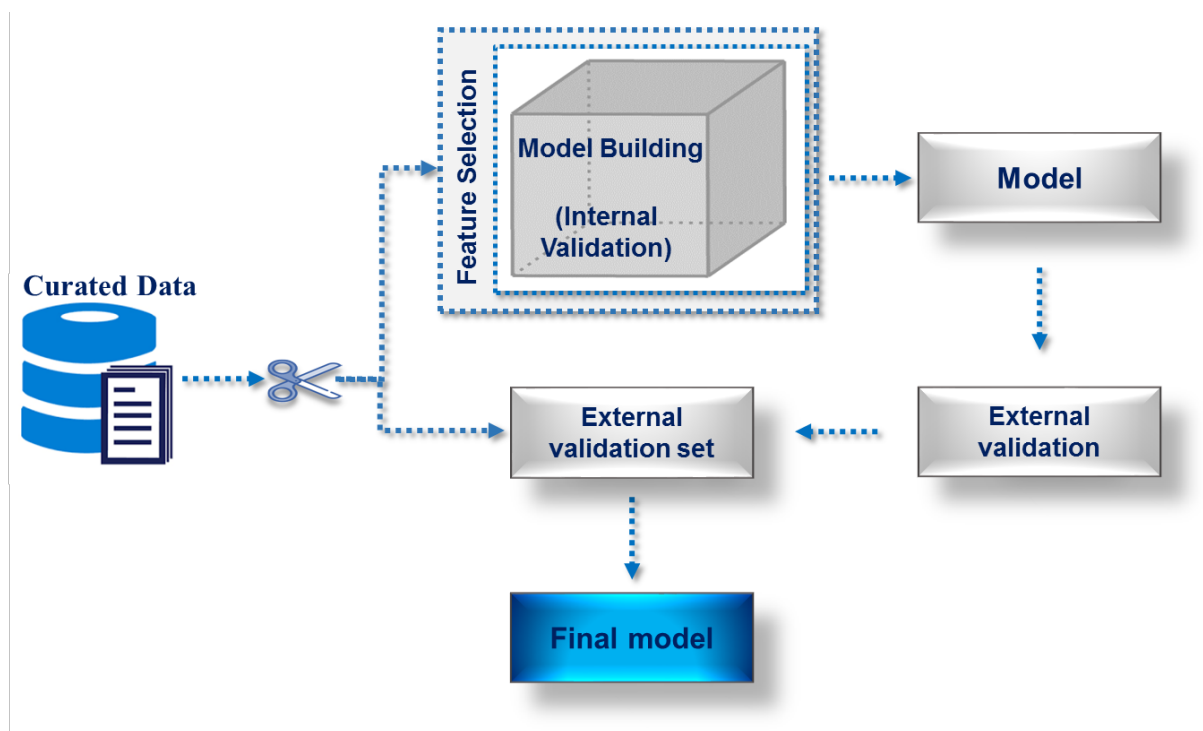


Figure 3.1: Overview of automated QSAR modeling workflow

This workflow starts with data preparation and data quality validation, data curation that includes gathering molecular structures and corresponding biological activity data for a specified target. Furthermore, to quantify various features of molecular structures a variety of chemical descriptors are computed. Before proceeding to the time-consuming trials of feature selection, model building and validation, data modelability evaluation is performed. Difficult data sets will not be recommended to model. After this step, the feature selection process follows, so as to identify an optimized non-redundant set of variables that can lead to best models. This critical step not only provides a better understanding of generated data but also improves the prediction performance of relevant predictors [45]. This latter phase

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

typically involves extensive testing of different models with an increasing set of variables. Finally, when a relevant and reduced set of variables has been determined, it can be used to develop the final QSAR model by following a rigorous internal and external validation process without compromising model quality assessment.

3.2.1 Architecture

This QSAR modeling workflow uses several customized nodes of (KNIME version 3.2) and is able to access online databases with millions of bioactive compounds. KNIME nodes can perform an extensive set of functions for many different tasks such as read/write data files, data processing, statistical analysis, data mining, and graphical visualization. Moreover, to reduce the complexity of large complicated workflow, a particular part of the workflow (sub-workflows) can be isolated in meta-nodes. The developed workflow aims at the simplification and automation of the QSAR model building. An overview of the implemented methodology is shown in Figure 3.2.

The complete process is divided into several systematic tasks of QSAR modeling including a) data access and processing, b) descriptors calculation, c) data set modelability estimation d) feature selection, e) model building and f) validation, along with adequate data visualization. Each of these subtasks is enclosed within the KNIME meta-nodes that are isolated from the rest of workflow enabling easy parameterization with a user-friendly configuration interface. The details of each task are covered in the following sections.

3.2.2 Data access and processing

There are typically two different alternatives for data set construction in model building,

3.2 Automated model building

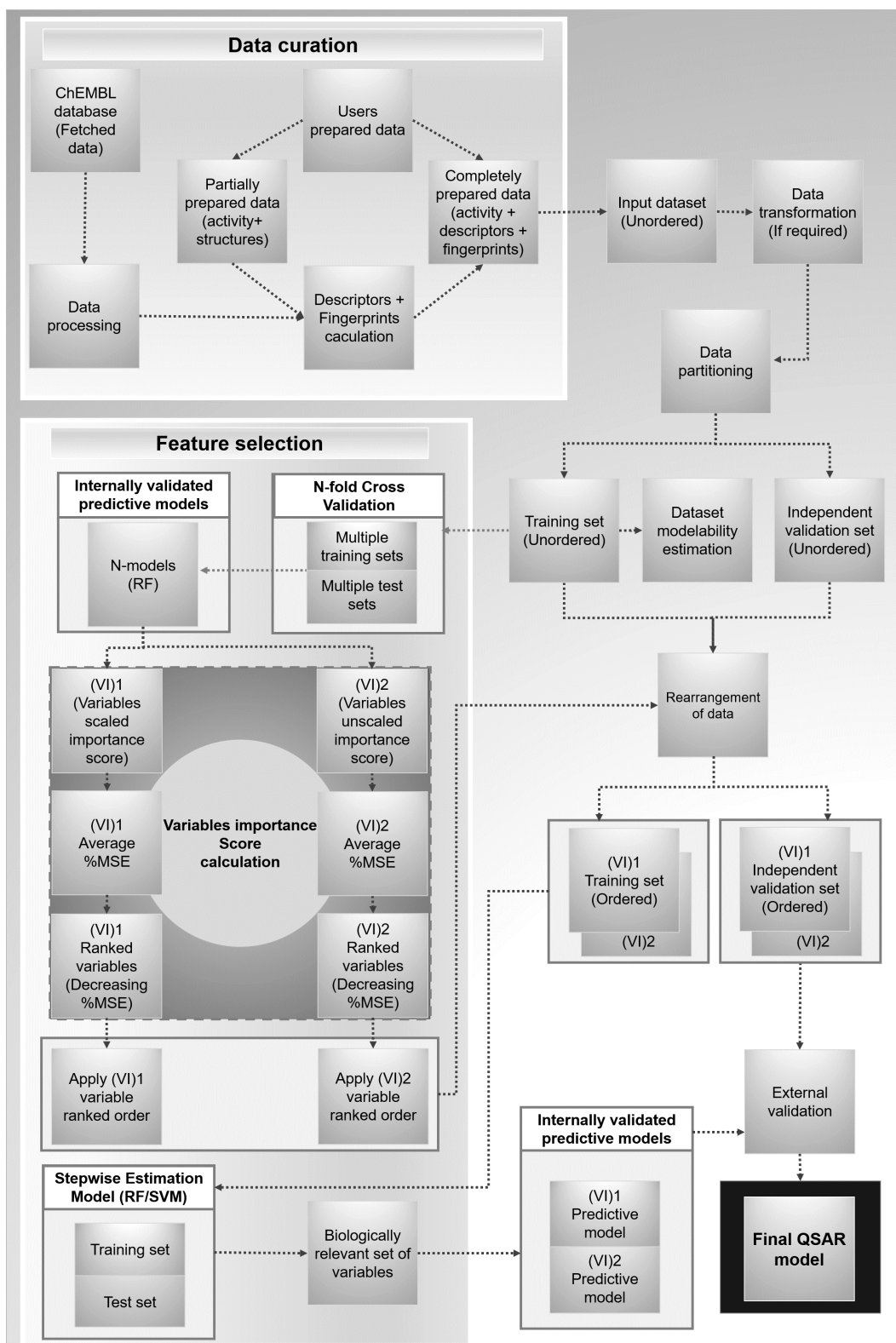


Figure 3.2: Automated QSAR modeling methodology

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

either the user has its own private data set with measurements curated from different sources or measured in the lab, or else retrieves the information from an available online data repository, that is continuously being updated by dedicated teams. The proposed workflow is able to encompass both approaches, giving the user the ability to use its own data set (with optional structural and descriptor calculation) or use an online repository (Figure 3.3).

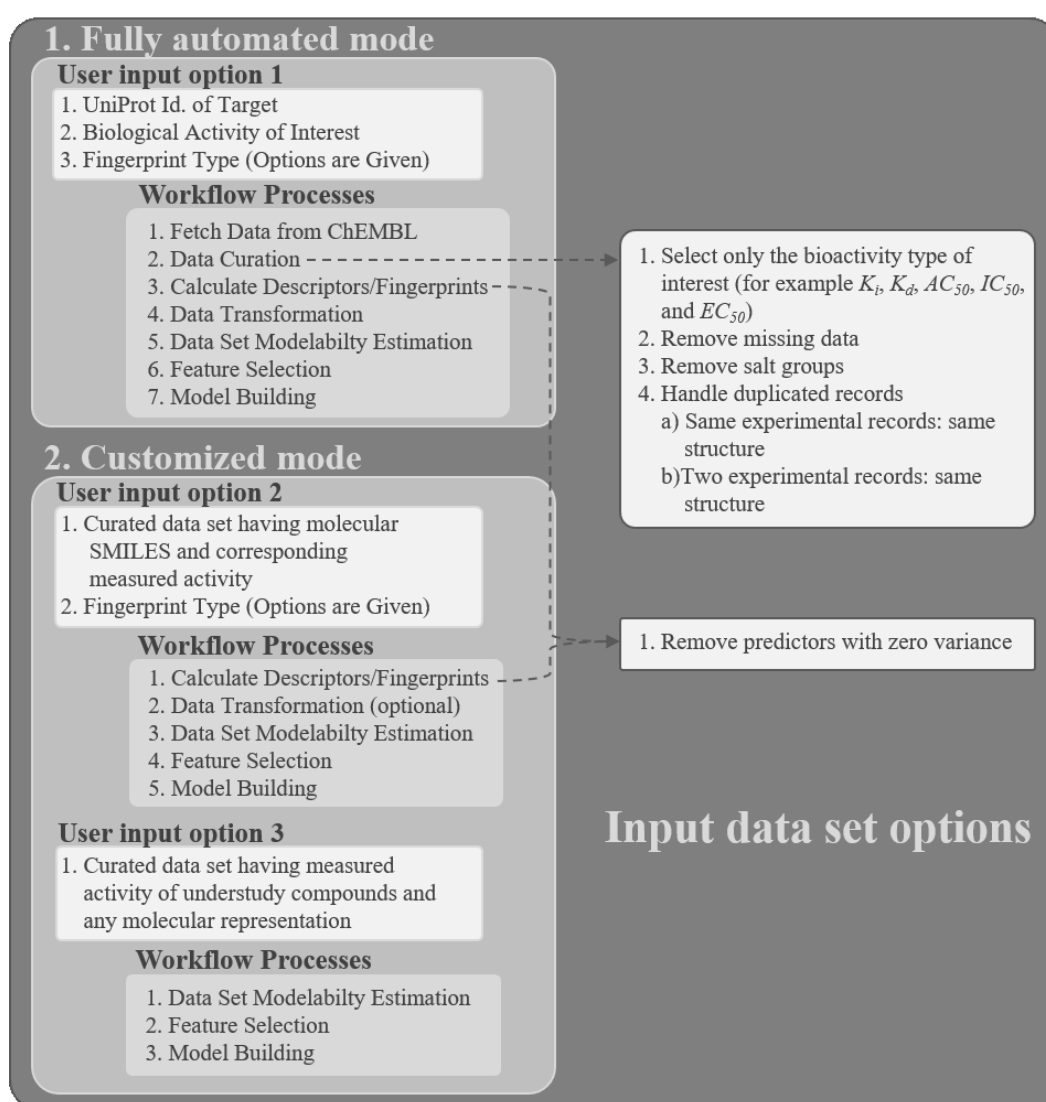


Figure 3.3: Input data set options. Overview of possible ways to submit input data to the automated QSAR modeling workflow

3.2 Automated model building

Nowadays, there are several large open source databases with annotated bioactivities for small molecules, with comprehensive information on biological properties of millions of chemicals. This wide data availability is one of driving forces beneath this effort. Most popular molecular databases like PubChem [46, 47], PDSP Ki [48], and ChEMBL [42] have become leading cheminformatics resources. The “Fully Automated” mode focus on ChEMBLdb by taking advantage of KNIME facility to access ChEMBL data. KNIME provides two built in nodes “ChEMBLdb Connector” and “ChEMBLdb Connector Input” to interact with RESTful and XML web services of ChEMBLdb. This facility for other chemical databases is not available yet. However, the ChEMBL database of more than 1.5 million bioactive compounds and 9,000 biological targets is capable to provide an ample variety of problems. In KNIME, the “ChEMBL database” meta-node encapsulates a complete workflow to access data from ChEMBLdb, data processing, and descriptor and fingerprint calculation. Hence, users can quickly access ChEMBLdb chemicals data for any target of interest by just a simple query of the desired UniProt ID and associated biological activity. The data obtained from ChEMBL may contain information related to all available biological activities extant for a given biological target (for example K_i , K_d , AC_{50} , IC_{50} , and EC_{50}). This retrieved data is processed by retaining only the user’s requested biological activity type records, and other relevant information related to chemical structures and assays. As the objective is to quantify a ligand-target interaction (activation or inhibition of the target), therefore any activity value can be utilized to count data related to the hypothesis. Overall data curation also includes the identification of missing data and duplicates (current year records are considered in two experimental records for same molecular structure) and dealing with several forms of the same molecule (including salt groups).

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

3.2.3 Descriptors calculation

The usage of descriptors and other computational representations of molecular structures is one of the principal methods applied to screen the new active molecules. The current workflow automatically calculates several molecular descriptors and structural characteristics for the retrieved molecules.

Along with this facility of online data access, users can also submit their fully prepared data file by using other input data set options with any types of descriptors calculated elsewhere. The workflow is able to use RDKit for descriptor calculations and can compute as well nine different fingerprint types, including Morgan, FeatMorgan, AtomPair, Torsion, RDKit, Avalon, Layered, MACCS and Pattern [39]

3.2.4 Data transformation and data partitioning

Scaling/transformation of the response variables (associated bioactivities) can be performed to standardize highly varying values in raw data for proper training of predictive model, where often data is transformed with a logarithmic function. This transformation can be skipped if data is already normalized. For the assessment of the applicability (prediction error) of the developed QSAR model, at this stage, the submitted data (either by automated retrieval from an online source, or by direct loading from a private data set), is divided into training set and Independent Validation Set (IVS) through a random partition. The training set is further used in N-fold cross validation process for internal model evaluation and selection while the IVS data is used to perform an unbiased model validation after the best model is built and selected. The latter is never used for any feature selection or model training procedure. So as not to bias the results.

3.2.5 Data set modelability estimation

Predictive performance of QSAR models highly depends upon different characteristics (e.g., size, chemical diversity, activity distribution or presence of activity cliffs) of various data sets [49, 50, 51]. It may not be always possible to build reliable QSAR models for certain data sets. To identify difficult problems, recent studies have introduced the concept of “data set modelability” meaning a prior estimate of the feasibility to obtain robust QSAR models by using a given descriptor space for data set of molecules [52, 53, 54]. The key idea behind this concept is based on the similarity principle that states that ‘similar compounds typically exhibit similar activity’ [55]. However, For every compound in a given data set, the nearest neighbors, i.e., compounds with the smallest distance from a given compound should possess similar activity. If the target property values for highly similar compounds are significantly different, then it means that the problem is probably hard to solve and most approaches will not be able to model it.

In the presented workflow, we followed a well established k-nearest neighbors approach based criteria, the modelability index (MODI) [53]. Golbraikh et al. [53] proposed several statistical criteria for estimating the feasibility of classification (e.g., data set diversity (*MODI_DIV*), activity cliff indices (*MODI_ACI*), correct classification rate (*MODI_CCR*)) and regression (similarity search coefficient of determination (*MODI_q²* and *MODI_ssR²*)). MODI is calculated as the Leave-One-Out (LOO) cross validation coefficient of determination of a simple k-Nearest Neighbours approach for data classification or regression over the training set, where k is typically either 3 or 5. MODI is fast to compute and helps modelers to quickly evaluate whether any given chemical compound data set can be modelled, giving an estimation of the predictability of the computed models before the actual modeling takes place. Data sets with very low MODI index are not recommended for model building, as a low MODI index informs the user that additional data processing and manual curation

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

may be required. However, according to the suggested MODI score for regression problems [53], in the automated QSAR modeling workflow, we suggest a MODI score to be >0.45 for reaching a model with acceptable predictability ($PVE \geq 0.60$).

3.2.6 Feature selection

The goal of QSPR/QSAR models is to correlate the molecular structures with their physiochemical/biological properties [20, 21, 22]. There are three main difficulties to achieve this task: 1) how to quantify molecular structure; 2) identify which are the relevant structural descriptors (or structure derived) that are the most adequate for the problem at hand; and 3) how to actually map the descriptors selected to the property being modeled [56, 20, 57, 21, 22]. Molecular descriptors can approximate most structural properties and a huge corpus of literature is extant on this subject [58]. Currently the number of chemical descriptors is so large that one of the biggest problems is selecting the most adequate features for each problem [58, 59]. Several issues typically need to be addressed in feature selection when the number of available variables is very large [60, 61, 62]. Some of the typical problems are:

- (a) Some descriptors appear highly correlated.
- (b) In several biological contexts no hypothesis is available about target structure for inferring binding activity.
- (c) Having many descriptors many times just do not improve the model quality, as the number of features advances, the number of spurious correlations increases as well and adding redundant or irrelevant variables to the model do not increase the model predictive abilities.
- (d) Sometimes the given descriptors are not, by themselves, able to contribute to modeling activity, but by combining them with other available descriptors, may

sometimes increase the model prediction capabilities.

- (e) The identification of a limited set of descriptors from the available list is many times necessary to avoid over-fitting, allow the desired physicochemical property to be adequately predicted by the constructed model and to have a better understanding of the models and the factors involved.

For the purpose of feature selection, several statistical and non-linear machine learning methods have been employed in QSPR/QSAR modeling as filter techniques. Some direct feature filtering approaches includes correlation matrix, Fisher's weight, Principal Components Analysis or Weighted Principal Components Analysis or Partial Least Squares (PCA/WPCA/PLS) loadings, regression coefficients, variable importance in PLS projections [VIP]) and Random Forest (RF). Some other are iterative methods for example, Ordered Predictor Selection-Partial Least Squares (OPS-PLS), Sequential Forward/Backward Selection, randomized methods that combine PLS with Genetic Algorithms (GA) or Monte-Carlo algorithms [45, 63, 64, 65, 66]. The direct filter methods are simpler and faster selecting variables, since they require only a metric calculation (a coefficient or weight) and the application of a cut-off value to determine the rejection of some variables due to the low importance to the model construction. Iterative methods have high computational cost, since most of them use filter methods in iterative ways or in combination with machine learning techniques. However, to deal with high dimensional data, the best-optimized methodology is always required to select the minimum subset of descriptors to predict a certain property with a good performance, less computational/time cost and in a more robust way. The application of non-linear machine learning algorithms to explore the non-linear relationships between descriptors and biological activities is increasing within the QSAR community [67, 68]. For feature selection in predictive models, we implemented a RFs voting procedure that can be used for the variable rankings according to their importance in RFs models [69, 65, 45]. In this ensemble method, each variables importance score is calculated by several available

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

variable importance's (VI) measures . One of the widely used VI measure in the regression problems is increase in mean of the error of a tree “Mean Squared Error (MSE)”, which explains how much prediction error increases with the random permutation of given variable while keeping all others unchanged in a node of a tree [70, 71, 69, 65]. Moreover, RF provides two options to fetch the VI score, which includes scaled and unscaled importance score. The scaled importance (also called z-score) is the default output of the randomForest function, which is obtained by division of the raw/unscaled importance by its standard error.

$$Z_j = \frac{(VI(X_j))}{\frac{\hat{\sigma}}{\sqrt{ntree}}} \quad (3.1)$$

However, some studies indicate that the unscaled importance VI (X_j) has better statistical properties and recommended for regression problems [72, 72, 73].

$$VI(X_j) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(X_j)}{ntree} \quad (3.2)$$

The current workflow followed the best performing RF based feature selection method, which is a hybrid approach [45]. The principle of this hybrid technique is to get: (1) possible set of variables, most relevant to the property of interest by using the variable importance (VI) function of RFs and (2) obtain the minimal set of features with a possibly best predictive performance along with unfavorable ratio between the number of predictors and number of observations. Practically, this approach counts variable importance by calculating the average mean squared error (MSE) provided by RF from a series of runs as a tool to rank the predictors. Hence, the VI based ranked variables can be feed to any machine learning

algorithm to build the stepwise predictive models to find a better balance between the biologically relevant set of features and prediction error (RMSE).

3.2.7 Model building

3.2.7.1 Model without feature selection

To verify performance of the applied feature selection method, it is necessary to assess model predictive behavior without any feature selection. Hence, developed QSAR modeling workflow, build a model with whole set of descriptors to confirm that elimination of irrelevant or non informative variables is improving predictive power of given model.

3.2.7.2 Model with feature selection

Automated QSAR modeling workflow follows a RF based feature selection method and provide ranked order of variables without eliminating any variable. These ranked variables are sequentially added to the learning algorithm to find the most relevant set of predictors leading to the model of smallest error rates.

The most employed machine learning approaches used in *in-silico* drug design are artificial neural networks (ANN), support vector machines (SVM), decision trees (DT), random forests (RF) and k-nearest neighbors (KNN) [63, 4]. Among the mentioned methods, the use of SVM to build QSAR models has become very popular in the last years [74, 75, 76, 77]. Moreover, many studies also explain the suitability of RF for high dimensional QSAR/QSPR datasets [70, 78, 45]. Hence, SVM [79] and RF [70], non-linear supervised learning methods are made available in the QSAR modeling workflow. This is mainly due to the fact that

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

these methods are robust in finding good modeling approaches in complex situations where the number of variables is very large and the number of instances is typically small. In such situations, many other machine learning methods (decision trees, neural nets, or linear models) can easily over fit, producing models unable to generalize outside the training space. Nonetheless, other algorithms can easily be used within KNIME, either through its customized nodes or by linking KNIME to R modules where most modeling approaches have been implemented.

To evaluate models predictability, data is split into training and test set to generate and validate stepwise estimation model by sequentially feeding ranked variables. The best features based internally validated model is finally presented for external validation.

3.2.8 External validation and model applicability domain

It is crucial to define the applicability domains of developed models by a critical step of external validation by using an IVS, which is not used in any part of the training process. In the developed workflow, a stringent protocol [22] of model validation is followed to ensure robustness and predictive power of the constructed models. The evaluation of the models' fitness is performed by comparing the proportion of the variance explained (PVE) by the predictive model, and the root mean squared error (RMSE) [80](see Eqs. 3.1 and 3.2). Externally evaluated final models can be used as a tool for external prediction and virtual screening.

$$PVE = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (3.3)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (3.4)$$

In Eqs. 3.3 and 3.4, y_i and \hat{y}_i are the measured and predicted biologically associated values for compound i , respectively, and \bar{y} is the mean of all activities from the compounds in the data set.

Nevertheless, in external predictions, the new data has molecules not present in the training set, therefore some predictions made with the model can be unreliable. This issue may be addressed by training models with a larger size and increased diversity, which many times is not an option in QSAR studies, or to circumscribe the model by defining its applicability domain (AD) in the chemical space [81, 82]. In the model AD, a similarity threshold between the training and validation set is established to flag the newly encountered compounds for which predictions may be unreliable. If the similarity between the training and validation set or new chemical is beyond the defined similarity threshold, the new compound is accounted to be outside the AD and the prediction is considered unreliable [81, 82]. In this QSAR modeling workflow, a well-established method [82] is used to define the domain of applicability of the built models based on the Euclidean distances among the training data and IVS.

3.2.9 Extensibility

The main modeling workflow is subdivided into several tasks. Each subtask is performed by small workflows that are developed and encapsulated within meta-nodes to establish independent processing and analysis. The subdivision of the complete modeling process in

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

QSAR modeling workflow architecture provides several advantages including a) it reduces the complexity of modeling framework 2) improves the understanding of the implemented machine learning procedure and c) increases the flexibility for future modification of the workflow. Hence, users can easily modify and further extend the presented workflow by domain-specific interests to add new features.

3.3 Results

3.3.1 Workflow implementation

Each task during drug designing from data preparation to model development and validation is critical to the accuracy of the predictive power of QSAR models [22]. The first stage of data preparation includes data collection, data cleaning by removing unwanted data, and appropriate molecular representation of underlying chemical compounds. In the second step the curated data is evaluated by data modelability criteria to check either given data set is reasonable to generate a QSAR model with significant predictive power. The third step includes extraction of more relevant biological features entitles as feature selection. Finally, model development and validations emphasize on a standardized process of internal and external model validation. QSAR modeling workflow is developed especially focusing on these mentioned major tasks to develop best-established methodology based framework.

3.3.1.1 Input data parameters

To run automated QSAR modeling workflow, simple settings of “Input Parameter” meta-node (Figure 3.4), like the choice of the target protein (name and UniProt ID.), molecular fingerprints, nfold value, working directory path and the type of activity measures are required

to build the best possible predictive model in very short time. No parameter is required to get RDKit descriptors for the given target; these are calculated by using the RDKit nodes embedded inside “ChEMBL Database” meta-node. Optional parameters node “Machine learning algo” provide the choice of machine learning algorithm (by default = SVM).

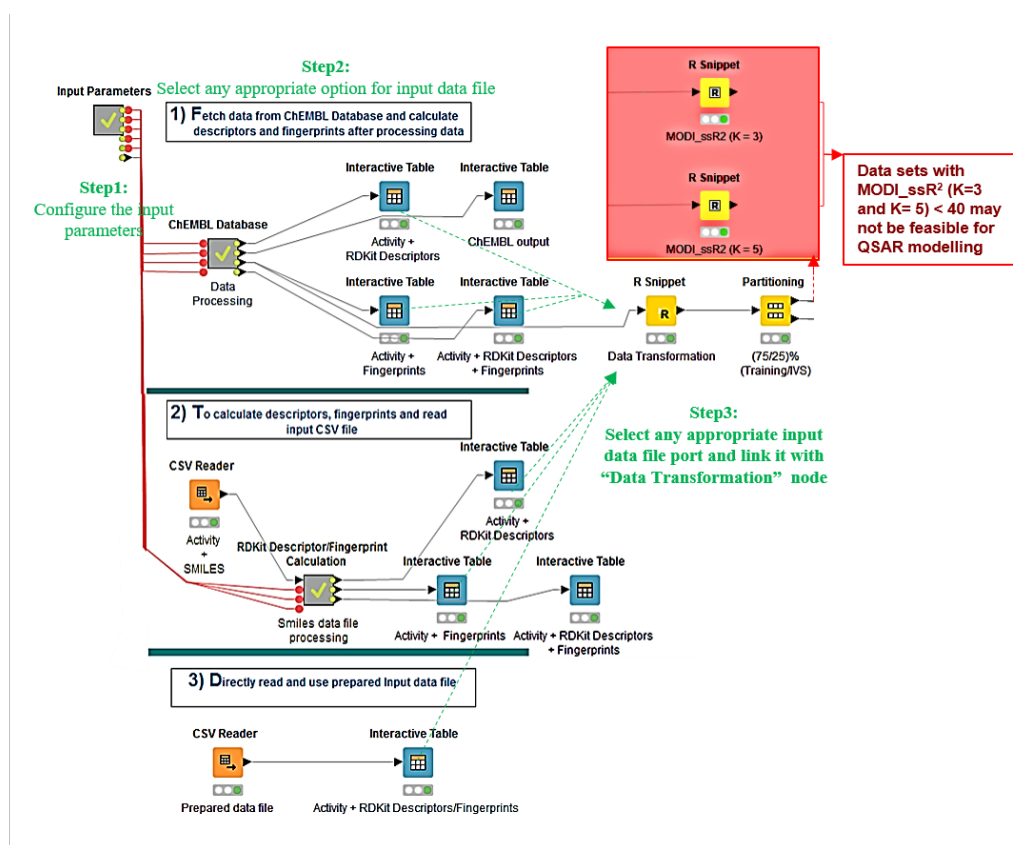


Figure 3.4: Input parameters. Input configurations required before to run the workflow

3.3.1.2 Input data set options

Automated QSAR modeling workflow provides three options to take input data files (Figure 3.3). The first option provides a “Fully Automated” mode, which directly accesses data from ChEMBL database with a simple query of UniProt accession number of a target

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

protein and associated bioactivity type. No deep understanding of data is required for the first option.

There are two other alternatives for modelling within “Customized” mode, if the user wants to work with personal data sets, and none of them requires ChEMBL data retrieval. Within “Customized” mode, the two alternatives deal with different available structural and descriptors-based information within the data sets, as the user is able to provide most of the data. Users with preliminary knowledge of their understudy problems can choose option 1 of “Customized” mode to process the known list of curated molecules. In the case of a thorough understanding of given modeling problem, where the user has previously computed the necessary molecular representation (with chemical descriptors or other structural information) the “Customized” mode option 2 bypasses all the descriptor computation phases and proceeds directly to model building. Hence, by adding flexibility in the way the user is able to provide input data, this constructed framework is able to cover some of the most common needs of modelers.

3.3.1.3 Data set retrieval and data pre-processing

In the “Fully Automated” mode to fetch data from ChEMBL the “ChEMBL Database” meta-node is developed in a given workflow (Figure 3.4). This meta-node can automatically prepare standard input data sets to explore a ChEMBLdb reported compounds-chosen receptor interaction by quantification of bioactivity of molecules.

In ChEMBLdb, different measures for binding affinities have been standardized, some of them remain more used like the half-maximal effective concentration (EC_{50}), the half-maximal inhibitory concentration (IC_{50}) and the inhibitory constant (K_i). EC_{50} value represents the molar concentration ($M = \text{mol/L}$) of an agonist that produces half of the maximal possible effect of that agonist. The simple definition of IC_{50} is a molar concentration of an antagonist that reduces the response to an agonist by 50%. Moreover, it can be explained as

the molar concentration of an unlabeled agonist or antagonist that inhibits the binding of a radio-ligand by 50%; or can be considered as the molar concentration of an inhibitory agonist that reduces a response by 50% of the maximal attainable inhibition [83, 84]. K_i value is used to quantify a ligand-receptor interaction based on the equilibrium dissociation constant (K). Hence, smaller the K_i value is associated with higher ligand-receptor binding affinities [68, 85].

In this machine learning pipeline, the focus is to set a standard protocol of regression problem based on any measure to predict the tendency of chemical molecules to either activate (K_i , K_d , AC_{50} , or EC_{50}) or inhibit (e.g., those with IC_{50} values/ K_i values) a selected target. The “ChEMBL Database” meta-node returns ChEMBL retrieved data (ChEMBL ID., reference, bioactivity type, assay description, activity value, and smiles strings), the calculated descriptors, and fingerprints data sets. Both the data sets of descriptors and the fingerprints can be used for further processing and modeling.

3.3.1.4 From data to validated models

Data pre-processing occupies a large time cost in QSAR modeling process. Many nodes are available in KNIME for data manipulation including row/column filtration, merging, splitting, concatenation and joining, type conversion and data transformation, row grouping and aggregation, and data table pivoting. Moreover, to process and handle large amount of data on a standard computer, KNIME also provides efficient memory management architecture. Hence, developed automated QSAR modeling workflow incorporates these all advantages of data processing and handling. It automatically fetches and processes data in an efficient way with the combinations of KNIME built in nodes with in this workflow. Data processing time depends upon the size of problems, while hardly one minute is required for small problems with less than 500 observations.

After data preparation, the next important task is fitting an appropriate machine learn-

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

ing algorithm to build a predictive model. For this purpose KNIME contains model building nodes for almost all options of machine learning and predictive models including most popular algorithms such as Bayes models, fuzzy rules, fuzzy c-means, k-means, neural networks, decision tree models, hierarchical and the self-organizing tree algorithms, linear and polynomial regression models, support vector machines, and supervised machine learning.

Nonetheless, along with simple statistical analysis and mathematical operations facilities, nodes to perform cross validation and bagging are also available. In addition, to integrate large number of statistical and graphical libraries, R [86] package is supported by KNIME to cover advanced data manipulation and modeling.

Automated QSAR modeling workflow can easily be customized to embed any of the mentioned algorithms. The implemented methodology in the current workflow combines series of R nodes to read data (R Source node), to draw plots (R View node), to train and build model (R Learner and R Predictor nodes) to perform additional tasks by personalized code (R Snippet node). However, major tasks of feature selection by RF and model building by SVM are performed with the help of inter-connected R nodes. Finally, the developed models are saved by model writer node in the user defined directory that can easily be read by model read node to make new predictions.

3.3.2 Real world cases

3.3.2.1 Data sets description

We tested the proposed QSAR modeling workflow on datasets of different members of protein families. These proteins include glutamate [NMDA] receptor, sigma non-opioid intracellular receptor (Sigma), beta-adrenergic receptor (ADRB), alpha-adrenergic receptor, histamine receptor (HRH), Potassium voltage-gated channel subfamily H member, dopam-

ine (DA-Rs) and serotonin (5-HT) receptors (Table 3.1).

The selection of these thirty different target proteins is independent of any hypothesis. Here, our emphasis is to examine the performance of applied strategy of QSAR modeling to solve diverse issues rather than to produce the best predictive model for each problem. To run the workflow, an initial configuration of “Input Parameter” meta-node is required to set the values of given parameters including number of folds for cross-validation (nfold), target protein name and UniProt accession number, working directory path, fingerprints and associated bioactivity. Hence, to prepare datasets for given problems “Input Parameter” meta-node was configured by providing name and UniProt accession number (Homo sapiens specific) of selected receptors, the associated bioactivity type (Table 3.1), Morgan fingerprints and “nfold” value was specified to perform tenfold cross validation (nfold = 10).

3.3.2.2 Data preparation and variable scaling

A subset of any data set from ChEMBL Database is passed through the R Snippet node (Data Transformation) (Figure 3.4). Variables scaling/transformation is important to standardize the range of independent feature to normalize the highly varying values in raw data for proper functionality of many machine learning algorithms. Recently, ChEMBLdb introduced pChEMBL value, which is an approach to standardize different activity types/values/units. pChEMBL is defined as a negative logarithm of molar IC_{50} , XC_{50} , EC_{50} , AC_{50} , K_i , K_d or Potency [42]. Some other methods to normalize widely varying ranges of activity values are also reported in the literature. For example, pK_i values are the negative logarithm to base 10 of the equilibrium dissociation constant, which allows an easier comparison of binding affinities. Thus, standard deviations are symmetrical for pK_i values but not for K_i values [84]. A generic formula was applied to convert values into scaled values (sp(Activity value))

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

Table 3.1: Description of selected problems

Uniprot ID.	Target Protein Name	Associated Bioactivities (Y)	Total Number of Observations (N-retrieved)	Total Number of Observations (N-Processed)
Q05586	Glutamate [NMDA] receptor	IC50	512	320
Q99720	Sigma non-opioid intracellular receptor 1 (Sigma1R)	IC50	1895	762
Q99720	Sigma non-opioid intracellular receptor 1 (Sigma1R)	Ki	2584	1465
CHEMBL613288 (Uniprot ID NA)	Sigma non-opioid intracellular receptor 2 (Sigma2R)	Ki	553	497
P08588	Beta-1 adrenergic receptor (ADRB1)	IC50	1471	599
P07550	Beta-2 adrenergic receptor (ADRB2)	IC50	1424	554
P13945	Beta-3 adrenergic receptor (ADRB3)	EC50	1478	1227
P35348	Alpha-1A adrenergic receptor	Ki	1650	1260
P35368	Alpha-1b adrenergic receptor	Ki	1567	1260
P25100	Alpha-1D adrenergic receptor	Ki	2076	1060
P35367	Histamine H1 receptor (HRH1)	Ki	2239	1222
P25021	Histamine H2 receptor (HRH2)	Ki	1218	385
Q9Y5N1	Histamine H3 receptor (HRH3)	Ki	3799	3101
Q9H3N8	Histamine H4 receptor (HRH4)	Ki	1486	1095
Q12809	Potassium voltage-gated channel subfamily H member 2 (HERG)	Ki	2539	1481
P21728	D(1A) dopamine receptor (DRD1)	Ki	2244	1087
P14416	D(2) dopamine receptor (DRD2)	IC50	1667	725
P35462	D(3) dopamine receptor (DRD3)	IC50	1174	326
P21917	D(4) dopamine receptor (DRD4)	Ki	3409	1900
P21918	D(1B) dopamine receptor (DRD5)	Ki	529	341
P47898	5-hydroxytryptamine receptor 5A	Ki	382	302
P50406	5-hydroxytryptamine receptor 6	Ki	4084	2632
P46098	5-hydroxytryptamine receptor 3A	Ki	517	432
P28222	5-hydroxytryptamine receptor 1B	Ki	1129	938
P41595	5-hydroxytryptamine receptor 2B	Ki	2034	1149
P28335	5-hydroxytryptamine receptor 2C	Ki	3433	2157
P28221	5-hydroxytryptamine receptor 1D	Ki	1153	973
P08908	5-hydroxytryptamine receptor 1A	Ki	4008	3244
Q13639	5-hydroxytryptamine receptor 4	Ki	540	422
P34969	5-hydroxytryptamine receptor 7	Ki	1753	1438

within “Data Transformation” node according to the following rules:

$$\begin{aligned} &\text{If Activity value} \geq 10000, \text{ sp(Activity value)} = 0 \\ &\quad \text{If } 10000 > \text{Activity value} > 1, \\ \text{sp(Activity value)} &= \frac{(4 - \log_{10}(\text{Activity value}))}{4} \quad (3.5) \\ &\text{If } 1 \geq \text{Activity value}, \text{ sp(Activity value)} = 1 \end{aligned}$$

Where sp(Activity value) represents the scaled activity value

Finally, after normalization of response variables (bioactivities) data is divided by random sampling into 75% training set and 25% independent validation set that will not be used in any training process (Figure 3.4).

3.3.2.3 Data set modelability measure

As stated, before the modeling phase of the thirty selected problems, the “modelability index” (MODI) is calculated [53]. MODI requires that the activities of compounds in all data sets and their distribution in the descriptor space (predictors) must range in the interval [0,1]. Biological activities were scaled according to Eq. 3.5, while descriptors were processed using a simple [0,1] scaling (Eq. 3.6).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (3.6)$$

where x is the original descriptor and x' is the scaled result of that variable.

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

3.3.2.4 Feature ranking by Random Forest

Data sets of all descriptors (descriptors and fingerprints) was used to consider high dimensional data sets for unbiased implementation of developed workflow to build an robust model based on best relevant features from highly redundant data.

This framework identify the most important features the ones that are responsible for the relevant molecular activity. Feature selection is a crucial step to reduce computation time and storage, improve model interpretability, understanding, performance, and remove irrelevant features (noisy data) to avoid over fitting [87]. Hence, we followed a strong method of RF based feature selection with a particular emphasis to generate more reliable, predictable, and generalized QSAR models [45]. QSAR modeling workflow finds the ranked ordered list of variables (descriptors and fingerprints) according to both scaled ((VI)1) and unscaled ((VI)2) importance scores.

Due to the stochastic nature of the RF algorithm, nfold cross validation was performed to fit RF models, and the importance of variables was recorded for each run. In the end, variables were ranked by sorting average variable importance scores in descending order. The process of features ranking is performed by two kinds of meta-nodes including “Model Validation” and “mean(%MSE) Calculator”. Hence, the output of these two meta-nodes is a processed input data rearranged by two kinds of variable rankings methods, first by scaled variable importance based ranked order, and second by unscaled importance based variables ranking.

3.3.2.5 Stepwise estimation models and feature selection

The produced ordered training data with more relevant to less important variables was further processed by meta-node “Build Model by Adding Ranked Variables”, which firstly

splits data into training and test set and introduces each ranked variable into a new SVMs fitted models. Each new model is validated by test set, and the statistical results of these stepwise estimation models are recorded to find the best set of features with minimum predictive error (RMSE). The results of the selected features based models (SF-models) of all target proteins clearly indicate large reduction of the total number of features (F) into more relevant features (SF) in all data sets. In the given problems, the maximum reduction of the features is 1037 to 9 variables ranked by scaled importance approach and 1079 to 29 variables in the case of unscaled importance. Similarly, the minimum reduction is 1134 to 470 variables and 1132 to 432 variables by scaled and unscaled importance methods respectively. Hence, on average applied methodology of feature selection performs adequate dimensionality reduction that is an important task to improve the quality of the predictive model.

3.3.2.6 Model results

After selecting the predictive model with best set of features (SF-model), the model's final assessment was performed using of the IVS. External validation is a critical step to make sure unbiased evaluation of developed model [20, 22, 44]. The IVS considered for external validation was never used in feature reduction and model training processes . On average, the difference between predictive performance of internally and externally validated SF-models is not large with optimally fitted models (Table 3.2).

SF-models of three receptors including Sigma1R (bioactivity dataset of IC_{50}), 5-HT2B and 5-HT4 showed poor generalization due to over-fitting in both methods of feature selection. In the other cases some SF-models like Sigma1R (bioactivity data set of K_i), 5-HT1A, 5-HT3A, 5-HT5A, 5-HT1D, ADRB1, DRD4 and DRD5 performed even better for external predictions.

To validate the efficiency of the implemented methodology, a model was also developed without feature selection (full-model). The external validation score of full-model is also

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

Table 3.2: QSAR models based on all descriptors (RDKit descriptors and Morgan fingerprints) datasets.

Target Protein Name	Total Number of Observations (N-Processed)		Total Number of Features (F)	Feature Selection By Scaled Variables Importance (VI)1					Feature Selection By unscaled Variables Importance (VI)2				
	Training Set	IVS		Selected Variable (SF)	SF-Model (test Set)		Final Model (SF-Model (IVS))		Selected Variable (SF)	SF-Model (test Set)		Final Model (SF-Model (IVS))	
					PVE	RMSE	PVE	RMSE		PVE	RMSE	PVE	RMSE
Glutamate [NMDA] receptor	240	80	949	120	0.78	0.12	0.69	0.17	120	0.79	0.12	0.73	0.16
Sigma non-opioid intracellular receptor 1 (Sigma1R)	572	190	1079	220	0.68	0.15	0.47	0.19	29	0.62	0.16	0.40	0.20
Sigma non-opioid intracellular receptor 1 (Sigma1R)	1099	366	1117	111	0.64	0.17	0.60	0.18	116	0.59	0.17	0.61	0.17
Sigma non-opioid intracellular receptor 2 (Sigma2R)	373	124	875	201	0.71	0.11	0.57	0.14	234	0.66	0.13	0.61	0.14
Beta-1 adrenergic receptor (ADRB1)	450	149	1040	150	0.70	0.14	0.72	0.13	180	0.80	0.12	0.71	0.13
Beta-2 adrenergic receptor (ADRB2)	416	138	1032	133	0.76	0.13	0.70	0.16	76	0.75	0.13	0.69	0.16
Beta-3 adrenergic receptor (ADRB3)	921	306	1093	310	0.64	0.15	0.56	0.17	170	0.57	0.19	0.55	0.18
Alpha-1A adrenergic receptor	945	315	1108	206	0.69	0.16	0.67	0.18	170	0.73	0.16	0.66	0.18
Alpha-1b adrenergic receptor	945	315	1106	275	0.71	0.15	0.65	0.15	115	0.69	0.15	0.62	0.16
Alpha-1D adrenergic receptor	795	265	1109	270	0.69	0.16	0.65	0.17	370	0.68	0.16	0.66	0.17
Histamine H1 receptor (HRH1)	917	305	1116	76	0.79	0.15	0.72	0.17	237	0.79	0.14	0.76	0.16
Histamine H2 receptor (HRH2)	289	96	1037	9	0.30	0.11	0.32	0.13	180	0.62	0.07	0.33	0.13
Histamine H3 receptor (HRH3)	2326	775	1134	397	0.62	0.16	0.63	0.16	282	0.66	0.16	0.63	0.16
Histamine H4 receptor (HRH4)	822	273	1075	123	0.63	0.18	0.56	0.18	330	0.63	0.17	0.55	0.18
Potassium voltage-gated channel subfamily H member 2 (HERG)	1111	370	1132	120	0.69	0.12	0.54	0.15	160	0.64	0.12	0.55	0.15
D(1A) dopamine receptor (DRD1)	816	271	1118	118	0.73	0.15	0.68	0.17	219	0.75	0.15	0.70	0.16
D(2) dopamine receptor (DRD2)	544	181	1092	91	0.66	0.16	0.63	0.18	150	0.71	0.16	0.62	0.19
D(3) dopamine receptor (DRD3)	245	81	1054	36	0.58	0.21	0.58	0.19	195	0.66	0.18	0.61	0.18
D(4) dopamine receptor (DRD4)	1425	475	1124	368	0.60	0.18	0.63	0.17	395	0.60	0.18	0.62	0.17
D(1B) dopamine receptor (DRD5)	256	85	957	135	0.68	0.18	0.76	0.15	142	0.75	0.17	0.77	0.15
5-hydroxytryptamine receptor 5A	227	75	980	140	0.83	0.13	0.87	0.12	38	0.81	0.14	0.84	0.13
5-hydroxytryptamine receptor 6	1974	658	1132	320	0.72	0.15	0.68	0.16	432	0.69	0.17	0.67	0.16
5-hydroxytryptamine receptor 3A	324	108	1045	150	0.69	0.19	0.71	0.19	230	0.62	0.21	0.71	0.19
5-hydroxytryptamine receptor 1B	704	234	1103	255	0.79	0.15	0.75	0.16	145	0.79	0.15	0.76	0.15
5-hydroxytryptamine receptor 2B	862	287	1130	101	0.51	0.18	0.37	0.19	110	0.57	0.15	0.39	0.19
5-hydroxytryptamine receptor 2C	1618	539	1135	263	0.67	0.16	0.62	0.18	244	0.64	0.18	0.62	0.17
5-hydroxytryptamine receptor 1D	730	243	1112	120	0.82	0.15	0.76	0.18	250	0.76	0.19	0.77	0.18
5-hydroxytryptamine receptor 1A	2433	811	1134	470	0.61	0.19	0.65	0.17	360	0.59	0.19	0.66	0.17
5-hydroxytryptamine receptor 4	317	105	948	203	0.80	0.16	0.66	0.22	280	0.83	0.15	0.71	0.20
5-hydroxytryptamine receptor 7	1079	359	1122	210	0.65	0.16	0.59	0.18	290	0.66	0.16	0.61	0.17

calculated to compare the performance with final predictive model with selected features (SF-model). The comparison of the performance of externally validated full model and externally validated final SF-model clearly confirms the effectiveness of the feature selection method. The results from all thirty different data sets show a significant increase in predictive power (PVE) and reduction in prediction error (RMSE) by removing the noisy data and considering the most relevant features (Table 3.3).

In the developed QSAR models of selected problems, the PVE score of the full-model ranges 0.13-0.59 while in the SF-model PVE ranges between 0.32-0.87 and 0.33-0.84 from scaled importance ((VI)1) and unscaled importance ((VI)2) methods respectively. However, an average PVE increase in both methods, ((VI)1) and ((VI)2) is almost 49% of the PVE of the full-model. The number of features in SF-models ranges between 0.0079%-16% of the

3.3 Results

Table 3.3: Comparison of performance of QSAR models (with and without feature selection)

Target Protein Name	Total Number of Observations (N-Processed)		Total Number of Features (F)	PVE (IVS)			RMSE (IVS)		
	Training Set	IVS		Full Model without Feature Selection	Final Model with Feature Selection		Full Model without Feature Selection	Final Model with Feature Selection	
				full-Model	SF-Model (VI)1	SF-Model (VI)2	full-Model	SFModel (VI)1	SF-Model (VI)2
Glutamate [NMDA] receptor	240	80	949	0.30	0.69	0.73	0.25	0.17	0.16
Sigma non-opioid intracellular receptor 1 (Sigma1R)	572	190	1079	0.31	0.47	0.40	0.21	0.19	0.20
Sigma non-opioid intracellular receptor 1 (Sigma1R)	1099	366	1117	0.45	0.60	0.61	0.21	0.18	0.17
Sigma non-opioid intracellular receptor 2 (Sigma2R)	373	124	875	0.46	0.57	0.61	0.16	0.14	0.14
Beta-1 adrenergic receptor (ADRB1)	450	149	1040	0.41	0.72	0.71	0.19	0.13	0.13
Beta-2 adrenergic receptor (ADRB2)	416	138	1032	0.46	0.70	0.69	0.21	0.16	0.16
Beta-3 adrenergic receptor (ADRB3)	921	306	1093	0.37	0.56	0.55	0.21	0.17	0.18
Alpha-1A adrenergic receptor	945	315	1108	0.53	0.67	0.66	0.21	0.18	0.18
Alpha-1b adrenergic receptor	945	315	1106	0.48	0.65	0.62	0.18	0.15	0.16
Alpha-1D adrenergic receptor	795	265	1109	0.47	0.65	0.66	0.21	0.17	0.17
Histamine H1 receptor (HRH1)	917	305	1116	0.59	0.72	0.76	0.21	0.17	0.16
Histamine H2 receptor (HRH2)	289	96	1037	0.13	0.32	0.33	0.14	0.13	0.13
Histamine H3 receptor (HRH3)	2326	775	1134	0.46	0.63	0.63	0.19	0.16	0.16
Histamine H4 receptor (HRH4)	822	273	1075	0.34	0.56	0.55	0.22	0.18	0.18
Potassium voltage-gated channel subfamily H member 2 (HERG)	1111	370	1132	0.42	0.54	0.55	0.17	0.15	0.15
D(1A) dopamine receptor (DRD1)	816	271	1118	0.50	0.68	0.70	0.21	0.17	0.16
D(2) dopamine receptor (DRD2)	544	181	1092	0.51	0.63	0.62	0.21	0.18	0.19
D(3) dopamine receptor (DRD3)	245	81	1054	0.32	0.58	0.61	0.24	0.19	0.18
D(4) dopamine receptor (DRD4)	1425	475	1124	0.47	0.63	0.62	0.20	0.17	0.17
D(1B) dopamine receptor (DRD5)	256	85	957	0.56	0.76	0.77	0.20	0.15	0.15
5-hydroxytryptamine receptor 5A	227	75	980	0.58	0.87	0.84	0.22	0.12	0.13
5-hydroxytryptamine receptor 6	1974	658	1132	0.48	0.68	0.67	0.20	0.16	0.16
5-hydroxytryptamine receptor 3A	324	108	1045	0.41	0.71	0.71	0.27	0.19	0.19
5-hydroxytryptamine receptor 1B	704	234	1103	0.45	0.75	0.76	0.23	0.16	0.15
5-hydroxytryptamine receptor 2B	862	287	1130	0.31	0.37	0.39	0.20	0.19	0.19
5-hydroxytryptamine receptor 2C	1618	539	1135	0.48	0.62	0.62	0.21	0.18	0.17
5-hydroxytryptamine receptor 1D	730	243	1112	0.49	0.76	0.77	0.24	0.18	0.18
5-hydroxytryptamine receptor 1A	2433	811	1134	0.43	0.65	0.66	0.21	0.17	0.17
5-hydroxytryptamine receptor 4	317	105	948	0.35	0.66	0.71	0.26	0.22	0.20
5-hydroxytryptamine receptor 7	1079	359	1122	0.43	0.59	0.61	0.21	0.18	0.17

total number of processed features considered in full models, which contain 1135 variables. The average reduction in the number of features is 83% of the total number. Moreover, error analysis of all predictive models shows an average RMSE of the full-model is 0.21 and in the case of SF-model the average RMSE is 0.17 in both methods. Hence, an average error decrease is 19% of the RMSE of the full-model. The large improvement of SF-models predictive performance and decrease in error rate exhibit the strength of unbiased methodology followed in automated QSAR modeling workflow.

All intermediate results can be visualized by interactive tables and graphical outputs from

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

data visualization layers. After completion of the QSAR model building workflow, outputs of each task are saved in the user's defined working directory. The availability of these intermediate data in the end of each task is useful to restore historical predictions and the given processed data with filtered features can further be used in any other application.

3.3.2.7 Model applicability domain analysis

For all thirty problems, feature selection and model development was carried out using the training set; however, model applicability to external compounds depends on the structural similarity between the chemicals in the IVS and the training set molecules. Model predictability is considered more reliable if the IVS chemicals fall within the AD. We used a KNIME node "Domain-Similarity" [82, 88] to analyze the AD of the models developed by the presented workflow. "Domain-Similarity" node uses similarity measurements to define the AD using Euclidean distances among all training compounds and the test or IVS compounds. The prediction may be unreliable if the distance of an external set compound to its nearest neighbor in the training set is higher than defined AD (out of AD).

In majority of the thirty selected problems compounds within the IVS were inside the AD, with the exception of six problems where some instances were outside the AD. These are the D(1A) dopamine receptor (3 molecules outside the AD), D(2) dopamine receptor (2 molecules), D(3) dopamine receptor (2 molecules), Sigma non-opioid intracellular receptor 1 with activity K_i (1 molecule), HRH2 (1 molecule), and 5-hydroxytryptamine receptor 1D (1 molecule). As the IVS should be a data set not controlled by the modellers, this QSAR modeling workflow does not remove these molecules and the decision is left to the users on how to handle the more prediction-error prone instances of the IVS.

3.3.2.8 Predictive performance comparison with published QSAR model

In the above analysis of the selected thirty problems, “Fully Automated” mode was tested where all processes from data retrieval to model building are completely automated (Figure 3.3). We further used “Customized” mode, of the workflow (Figure 3.3), to demonstrate the efficiency of implemented methodology in the developed automated QSAR model by comparing its performance to the published solutions of scientific problems. For this purpose, we selected one very recent example on antiviral binding affinity data for non-nucleoside analogue reverse-transcriptase inhibitors (NNRTIs) from the QsarDB repository [89]. The same training (31 molecules) and external validation (8 molecules) datasets of chemical compounds with their corresponding scaled bioactivity (pK_i) were taken from the published work [90] for model building in this workflow. The curated dataset of NNRTIs with the 39 ligands in SMILES format and their computed pK_i was submitted in “Customized” mode option 1 (Figure 3.3). As K_i values were already scaled [90], so we skipped the “Data Transformation” node and adjusted the data partitioning node for the simple division of reported 31 training and 8 IVS molecules (Figure 3.4). RDKit descriptors and fingerprints were computed automatically for this given input dataset of NNRTIs. MODI scores for the first three options of fingerprints (Morgan, FeatMorgan, AtomPair) in the “Input Parameter” meta-node (Figure 3.4) were lower than the threshold (MODI >0.45). Thus, we skipped these 3 fingerprints and continued the modeling process using RDKit descriptors and torsion fingerprints for which MODI score was greater than the threshold (for K3, MODI = 0.46 and for K5, MODI = 0.48).

Performance of automated QSAR modeling workflow based SF-models in antiviral binding affinity prediction on external validation set or IVS for NNRTIs was markedly better in both options (scaled and unscaled variable importance) of feature selection than the published [90] QSAR model. The PVE score of the SF-model((VI)1) is 0.81 and for SF-

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

model((VI)2) is 0.82 while the published solution showed 0.725 scores of the squared coefficient of correlation (R^2) for the same IVS. In the same way, the RMSE score of the SF-model((VI)1) is 0.34 and for SF-model((VI)2) is 0.33 while the published solution showed 0.2230 (RMSE= 0.47) score of squared standard error of the regression (S^2) for the same IVS. All the molecules of the IVS were found within the AD; thus predictions can be considered reliable.

3.4 Discussion

In the current work, an extendable platform was designed that can be used as a QSAR modeling pipeline to get an optimized predictive model. The performance of the presented automated QSAR modeling workflow was assessed for thirty different data sets of size ranging from 300 to 3200 molecules and the features set of 1141 descriptors (RDKit descriptors and fingerprints). We have further compared the results obtained from our workflow with a published QSAR modeling problem and the results obtained were significantly better than the original authors efforts, even though the approach followed was mostly unsupervised.

Comparison of all constructed full-models and SF-models revealed improved predictive power with a small set of biologically relevant variables (Figure 3.5). Hence, feature selection methodology was found efficient to deal with high dimensional data by selecting adequate features for each problem to predict a certain property with a good performance, less computational/time cost. For regression problems, compelling evidences exists for the robustness of RF unscaled variable importance measure $VI(X_j)$ because of its statistical properties [72, 72, 73]. Consistent with literature, overall performance of selected sub-set of variables by RF unscaled importance measure ((VI)2) was better than scaled importance measure ((VI)1).

To explore the role of the training data sets size in determining the performance of pre-

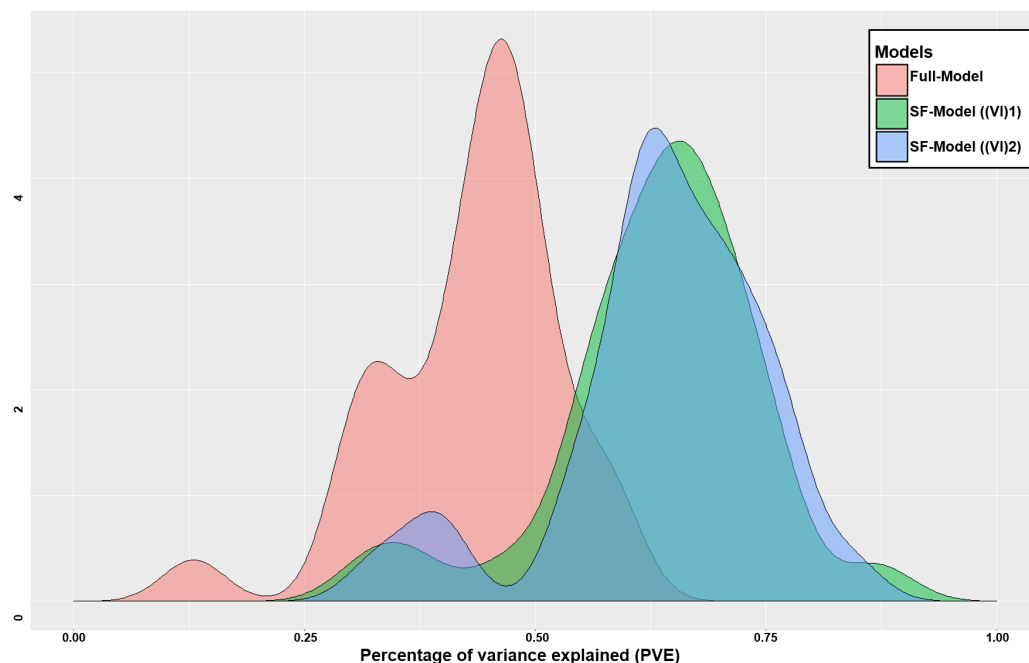


Figure 3.5: Comparison of models with and without feature selection. Pink color represents the full-model without feature selection (with all variables (F)), green color is for SF-model ((VI)1) contains predefined set of features (SF) identified by scaled permutation importance, and blue color represents SF-model ((VI)2) having selected features (SF) by unscaled variable importance measure.

dictive models, PVE for each model was compared with data set size (Figure 3.6). Models trained with data sets less than 1500 molecules showed quite diverse predictive performance. The data set size of the best performing model of the receptor 5-HT_{5A} with PVE value 0.87 is 302 molecules and least performing model of the receptor HRH₂ with PVE value 0.32 has 385 molecules. The models performance was stable in larger data sized problems. Possible reasons for these variations in performance is may be the complex nature of the problem and the size limitations [44]. Hence, availability of more data may help to find real trends in data with a satisfactory solution.

In regression modeling, one of the most critical problem is over-fitting of a model which results into poor generalization and reduced performance on unseen data. One widely accep-

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

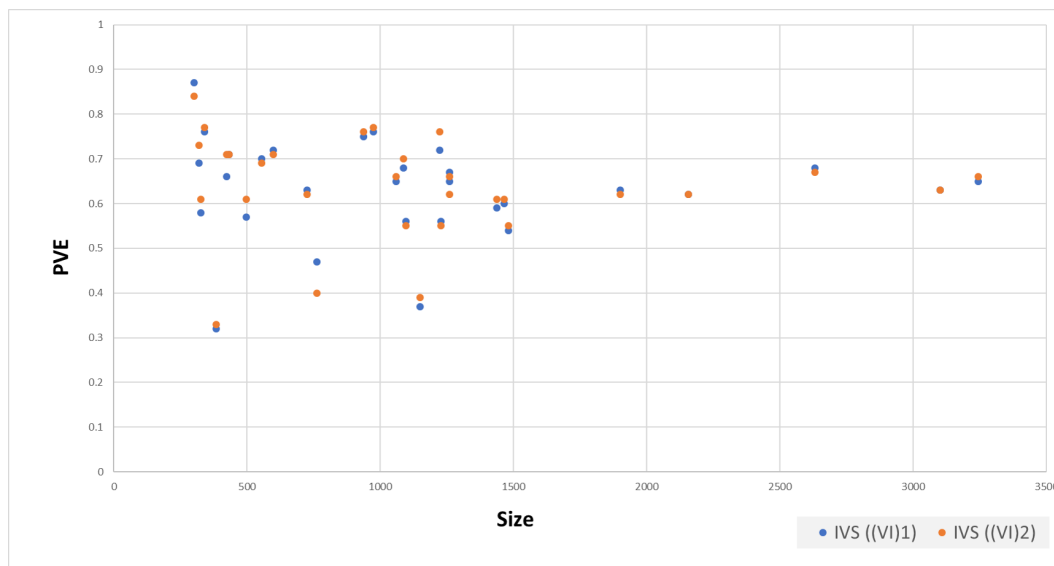


Figure 3.6: Size of the problems and predictive power of fitted models. Blue dots represent externally validated models with feature selection by scaled importance, and golden yellow color denotes externally validated models with feature selection by unscaled importance measure.

ted measure for testing over-fitting is to observe performance over independent validation data set [22, 4]. Hence, SF-model's final assessment was performed using of the independent validation set (IVS). The internal (test set) and external (IVS) prediction results of the SF-models were compared to identify the over-fitted models (Figure 3.7) in both methods of feature selection like the scaled (Figure 3.7A) and unscaled importance (Figure 3.7B). In both feature selection methods, none of both is completely superior to the other one. For example, problem Histamine H2 receptor (HRH2) is a worst generalized model constructed by unscaled importance based feature selection, but was optimally fitted by the scaled importance based set of features. Hence, our focus was on the problems that were failed in both feature selection methods. Out of thirty problems, three models were found over-fitted in both methods.

Worst cases include 5-hydroxytryptamine receptor 2B (5-HT2B), 5-hydroxytryptamine receptor 4 (5-HT4) and Sigma non-opioid intracellular receptor 1 (Sigma1R) that are over-

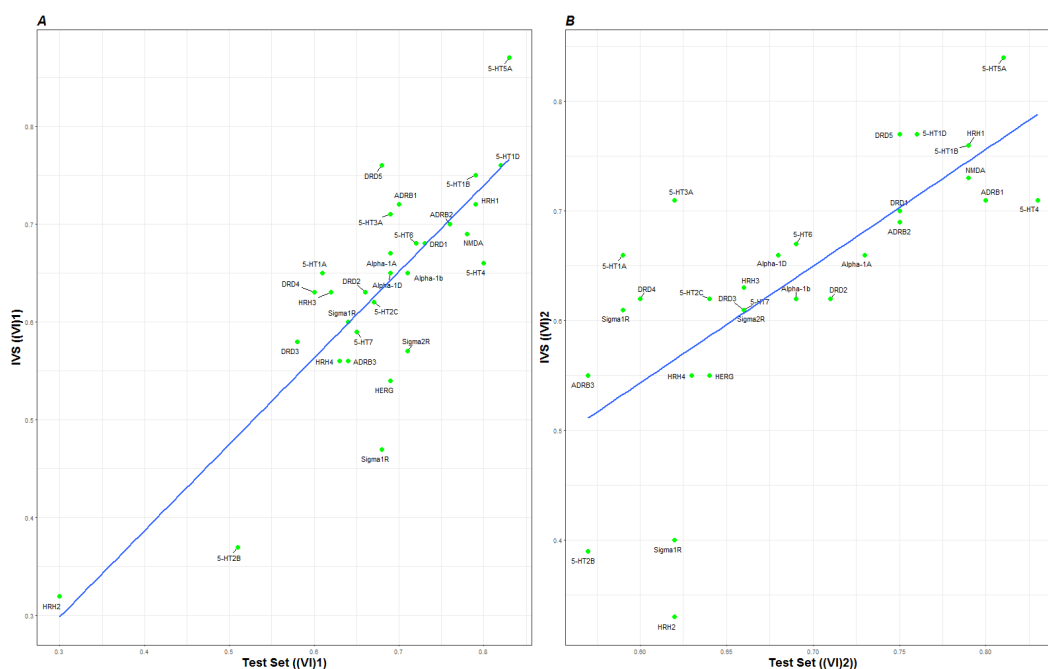


Figure 3.7: Models over-fitting analysis. Models with a predefined set of features identified by scaled variable importance (A) and unscaled variable importance (B).

fitted in both variable selection processes. Comparison of experimental and predicted activity values was carried out to analyze poor prediction of particular activity value points. The over-fitted models were unable to accurately predict the response variable at extreme values and large errors were observed near the upper and lower extreme of the experimental range. These mispredictions may result from data sets with very few measured instances with values near the experimental range. However, insufficient patterns of predictors may reduce the model coverage and lead to poor generalization [44].

In the end, PVE scores (QSAR_PVE(IVS)) of full-models and final SF-models were compared with their corresponding $MODI_{ss}R^2$ scores (Figure 3.8). Results showed significant correlation between the PVE for the IVS in SF-models and $MODI_{ss}R^2$ (correlation=0.76 for $MODI_{ss}R^2$ with $K = 3$ and correlation=0.78 for $MODI_{ss}R^2$ with $K = 5$) (Figure 3.8 A and B). This is consistent with the published work [53], which suggests that the $MODI_{ss}R^2$ score should be ≥ 0.46 for 3 nearest neighbors and ≥ 0.47 for 5 nearest

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

neighbors. The correlation between the full-models PVE and $MODI_{ss}R^2$ was not as significant. This weaker correlation was expected as full-models may contain irrelevant and highly correlated variables which directly influence the models predictive power by causing them to over fit the training sets. Hence, the implemented feature selection approach has an efficient role for achieving robust models with reliable predictive performance.

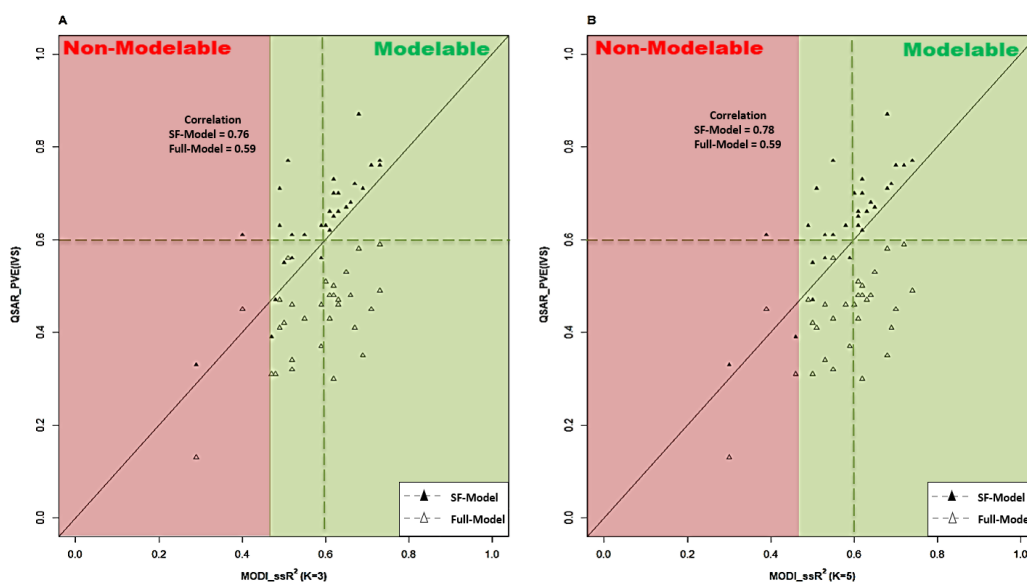


Figure 3.8: $MODI_{ss}R^2$ versus QSAR_PVE for 30 datasets. K is the number of nearest neighbors. (A) $K = 3$ and (B) $K = 5$. QSAR_PVE(IVS) is PVE score of externally validated models without feature selection (Full-model) and with selected features (SF-model). High correlation with SF-models QSAR_PVE suggests $MODI_{ss}R^2$ is good modelability criteria. Weaker correlation between Full-model QSAR_PVE and $MODI_{ss}R^2$ emphasize the importance of feature selection to obtain actual and reliable predictive performance of QSAR model.

3.5 Conclusion

The developed QSAR modeling workflow is a fully automated QSAR pipeline to assist all users including those are not expert in machine learning and have less knowledge of

available data. Creation of an optimal predictive model demands many critical and time-consuming steps, including data collection and processing, appropriate data representation (descriptors and fingerprints calculation), evaluation of the data set modelability, best predictors selection, machine learning models fitting and validation. QSAR modeling workflow completely automates the laborious and iterative process of modeling to tackle different problems. Following are the key advantages of proposed QSAR modeling workflow:

- It automatically fetches high-quality compounds data set from continuously improving and growing curated databases (e.g. ChEMBL). Hence, the potential of direct access of the online data sets enables to this fully automated framework a widely used platform for QSAR model building.
- Important aspects of the data processing by selecting only the bioactivity type of interest, dealing with duplicates, removing missing data and salt groups, descriptors calculation, and data normalization are handled in a very flexible and consistent manner.
- Prior estimate of data set modelability can reduce modelers efforts by focusing in the most promising problems or identifying the challenging ones that may require more data, more descriptor variability or different strategies.
- Best practice feature selection and an exhaustive validation procedure are followed in the presented workflow in order to ensure minimal bias in model development and evaluation. The analysis of the obtained results of thirty different target-drug interaction predictive models concludes that the developed feature selection methodology performs consistently well for high-dimensional data by removing 62% to 99% redundant data. This large reduction of irrelevant variables minimizes the computational/time cost, improves the predictive power of model and provides a better understanding of the underlying relationship between the property of interest and the relevant features.

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

- The automated QSAR modeling framework is not a black-box prediction system, rather it is an extensible and highly customizable tool to develop the robust predictive models and provide the output of all modeling task for the diverse application and reproduction of historical predictions. Moreover, it ensures that the same protocol is used for updating models with new molecules as they become available.
- It is worth mentioning that the generated workflow feeds the selected feature-matrix to SVM models but these variables can be used as input for any other non-linear machine learning method which can be easily implemented in the framework.

In conclusion, with the above mentioned adopted features of the presented open source automated QSAR modeling framework, it is hoped to guarantee that the most important aspects of QSAR modeling are addressed and consistently applied. This framework has been tested against thirty data sets, some very difficult, and generally as produced robust results; this has been achieved without any need of users thorough understanding of data, computer programming and/or machine-learning knowledge and complex parameterization to customize the complex modeling algorithms and procedures.

Acknowledgements

The authors gratefully acknowledge Fundação para a Ciência e Tecnologia for a doctoral grant (SFRH/BD/111654/2015), MIMED project funding (PTDC/EEI-ESS/4923/2014) and UID/CEC/00408/2013 (LaSIGE) for providing the infrastructure.

References

- [1] Shivani Agarwal, Deepak Dugar and Shiladitya Sengupta. 'Ranking Chemical Structures for Drug Discovery : A New Machine Learning Approach'. In: (2010), pp. 716–731.
- [2] Kun Yi Hsin, Samik Ghosh and Hiroaki Kitano. 'Combining machine learning systems and multiple docking simulation packages to improve docking prediction reliability for network pharmacology'. In: *PLoS ONE* 8.12 (2013). ISSN: 19326203. DOI: 10.1371/journal.pone.0083922.
- [3] Atsushi Matsumoto, Shin Aoki and Hayato Ohwada. 'Comparison of Random Forest and SVM for Raw Data in Drug Discovery : Prediction of Radiation Protection and Toxicity Case Study'. In: 6.2 (2016), pp. 145–148. DOI: 10.18178/ijmlc.2016.6.2.589.
- [4] Angélica Nakagawa Lima et al. 'Use of machine learning approaches for novel drug discovery.' In: *Expert opinion on drug discovery* 11.3 (2016), pp. 225–239. ISSN: 1746-045X. DOI: 10.1517/17460441.2016.1146250.
- [5] E Mantus. 'Toxicity Testing in the 21st Century'. In: *Alttox.Org* (2007). DOI: 10.17226/11970.
- [6] Thomas Hartung. 'Toxicology for the twenty-first century'. In: *Nature* 460.7252 (2009), pp. 208–212. ISSN: 0028-0836. DOI: 10.1038/460208a.
- [7] Grace Patlewicz et al. 'Proposing a scientific confidence framework to help support the application of adverse outcome pathways for regulatory purposes'. In: *Regulatory Toxicology and Pharmacology* 71.3 (2015), pp. 463–477. ISSN: 10960295. DOI: 10.1016/j.yrtph.2015.02.011.

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

- [8] Marco Viceconti, Andriano Henney and Edwin Morley-Fletcher. 'In silico clinical trials: how computer simulation will transform the biomedical industry.' In: *Avicenna Coordination Support Action* (2016). DOI: 10.13140/RG.2.1.2756.6164.
- [9] Joanna Jaworska, Tom Aldenberg and Nina Nikolova. 'Review of methods for QSAR applicability domain estimation by the training set'. In: *Technical report. The European Commission—Joint Research Centre Institute for Health & Consumer Protection-ECVAM* (2005).
- [10] Rositsa Serafimova, Mojca Fuart Gatnik and Andrew Worth. 'Review of QSAR Models and Software Tools for Predicting Genotoxicity and Carcinogenicity'. In: *Publications Office of the European Union. JRC Scientific and technical reports* (2010). DOI: 10.2788/26123.
- [11] M Zeeman et al. 'U . S . EPA Regulatory Perspectives on the Use of QSAR for New and Existing Chemical Evaluations'. In: *SAR and QSAR in Environmental Research* 3:3.December 2014 (1995), pp. 179–201. DOI: 10.1080/10629369508234003.
- [12] Luis G Valerio. 'In silico toxicology models and databases as FDA Critical Path Initiative toolkits'. In: *HUMAN GENOMICS* 5.3 (2011), pp. 200–207.
- [13] T. Martin. 'User ' s Guide for T . E . S . T . (version 4 . 2) (Toxicity Estimation Software Tool)'. In: (2016).
- [14] Christina Rudén and Sven Ove Hansson. 'Registration , Evaluation , and Authorization of Chemicals (REACH) Is but the First Step — How Far Will It Take Us ? Six Further Steps to Improve the European Chemicals Legislation'. In: 1 (2010), pp. 6–10. DOI: 10.1289/ehp.0901157.
- [15] Environment Directorate et al. 'OECD Environment Health and Safety Publications Series on Testing and Assessment'. In: *Assessment* (2004), pp. 20–21. DOI: 10.1787/9789264079151-en.

REFERENCES

- [16] A.R. Katritzky et al. 'Structurally Diverse Quantitative Structure-Property Relationship Correlations of Technologically Relevant Physical Properties'. In: *Journal of Chemical Information and Modeling* 40.1 (2000), pp. 1–18. ISSN: 1549-9596. DOI: 10.1021/ci9903206.
- [17] Alan R Katritzky et al. 'The present utility and future potential for medicinal chemistry of QSAR/QSPR with whole molecule descriptors.' In: *Current topics in medicinal chemistry* 2.12 (2002), pp. 1333–1356. ISSN: 15680266. DOI: 10.2174/1568026023392922.
- [18] Jean Pierre Doucet and Annick Panaye. 'Three Dimensional QSAR: Applications in Pharmacology and Toxicology'. In: (2010), p. 575.
- [19] Scott Doniger, Thomas Hofmann and Joanne Yeh. 'Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms.' In: *Journal of computational biology : a journal of computational molecular cell biology* 9.6 (2002), pp. 849–864. ISSN: 1066-5277. DOI: 10.1089/10665270260518317.
- [20] Alexander Tropsha and Alexander Golbraikh. 'Predictive QSAR modeling workflow, model applicability domains, and virtual screening.' In: *Current pharmaceutical design* 13.34 (2007), pp. 3494–504. ISSN: 1873-4286. DOI: 10.2174/138161207782794257.
- [21] T Puzyn, J Leszczynski and M T Cronin. *Recent Advances in QSAR Studies: Methods and Applications (Challenges and Advances in Computational Chemistry and Physics)*. Ed. by Mark T. D. Cronin Tomasz Puzyn (Editor), Jerzy Leszczynski (Editor). 2010 editi. Springer, 2009. ISBN: 9781402097829.
- [22] Alexander Tropsha. 'Best practices for QSAR model development, validation, and exploitation'. In: *Molecular Informatics* 29.6-7 (2010), pp. 476–488. ISSN: 18681743. DOI: 10.1002/minf.201000061.

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

- [23] Steven L Dixon et al. 'AutoQSAR: an automated machine learning tool for best-practice QSAR modeling'. In: *Future medicinal chemistry* (2016).
- [24] Pau Carrió et al. 'eTOXlab , an open source modeling framework for implementing predictive models in production environments'. In: (2015). DOI: 10.1186/s13321-015-0058-6.
- [25] Kumar Pandey and Matthias Rupp. 'Online chemical modeling environment (OCHEM): web platform for data storage , model development and publishing of chemical information'. In: (2011), pp. 533–554. DOI: 10.1007/s10822-011-9440-2.
- [26] Richard Cox et al. 'QSAR workbench : automating QSAR modeling to drive compound design'. In: (2013), pp. 321–336. DOI: 10.1007/s10822-013-9648-4.
- [27] J M Stevenson and P D Mulready. 'Pipeline pilot 2.1'. In: *Journal Of The American Chemical Society* 125.5 (2003), pp. 1437–1438.
- [28] DVS Green et al. 'Automated predictive modelling: modeller's utopia or fools' gold?' In: (2008).
- [29] Michael P Mazanetz et al. 'Drug discovery applications for KNIME: an open source data mining platform.' In: *Current topics in medicinal chemistry* 12.18 (2012), pp. 1965–79. ISSN: 1873-4294. DOI: 10.2174/156802612804910331.
- [30] Claire L. Mellor, Fabian P. Steinmetz and Mark T D Cronin. 'Using Molecular Initiating Events to Develop a Structural Alert Based Screening Workflow for Nuclear Receptor Ligands Associated with Hepatic Steatosis'. In: *Chemical Research in Toxicology* 29.2 (2016), pp. 203–212. ISSN: 15205010. DOI: 10.1021/acs.chemrestox.5b00480.
- [31] Yocheved Gilad, Katalin Nadassy and Hanoch Senderowitz. 'A reliable computational workflow for the selection of optimal screening libraries'. In: *Journal of Cheminformatics* (2015), pp. 1–17. ISSN: 1758-2946. DOI: 10.1186/s13321-015-0108-0.

REFERENCES

- [32] George Nicola et al. 'Connecting proteins with drug-like compounds : Open source drug discovery workflows with BindingDB and KNIME'. In: (2015), pp. 1–22. DOI: 10.1093/database/bav087.
- [33] Dunca Hull et al. 'Taverna: A tool for building and running workflows of services'. In: *Nucleic Acids Research* 34.WEB. SERV. ISS. (2006), pp. 729–732. ISSN: 03051048. DOI: 10.1093/nar/gkl320.
- [34] Belinda Giardine et al. 'Galaxy: A platform for interactive large-scale genome analysis'. In: *Genome Research* 15.10 (2005), pp. 1451–1455. ISSN: 10889051. DOI: 10.1101/gr.4086505.
- [35] I Altintas et al. 'Kepler: an extensible system for design and execution of scientific workflows. 16th International Conference on Scientific and Statistical Database Management.' In: *Petros Nomikos Conference Center, Santorini Island, Greece I* (2004), pp. 423–424. ISSN: 1099-3371. DOI: 10.1109/SSDM.2004.1311241.
- [36] David E. Rex, Jeffrey Q. Ma and Arthur W. Toga. 'The LONI Pipeline Processing Environment'. In: *NeuroImage* 19.3 (2003), pp. 1033–1048. ISSN: 10538119. DOI: 10.1016/S1053-8119(03)00185-X.
- [37] Michael R. Berthold et al. 'KNIME - The Konstanz Information Miner'. In: *SIGKDD Explorations* 11.1 (2009), pp. 26–31. ISSN: 19310145. DOI: 10.1145/1656274.1656280.
- [38] Christoph Steinbeck et al. 'The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics'. In: *Journal of Chemical Information and Computer Sciences* 43.2 (2003), pp. 493–500. ISSN: 00952338. DOI: 10.1021/ci025584y.
- [39] Greg Landrum. *RDKit Documentation*. 2018. (Visited on 03/09/2018).

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

- [40] Richard A. Friesner et al. 'Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy'. In: *Journal of Medicinal Chemistry* 47.7 (2004), pp. 1739–1749. ISSN: 00222623. DOI: 10.1021/jm0306430.
- [41] Steven L. Dixon, Alexander M. Smondryev and Shashidhar N. Rao. 'PHASE: A novel approach to pharmacophore modeling and 3D database searching'. In: *Chemical Biology and Drug Design* 67.5 (2006), pp. 370–372. ISSN: 17470277. DOI: 10.1111/j.1747-0285.2006.00384.x.
- [42] A. Patrícia Bento et al. 'The ChEMBL bioactivity database: An update'. In: *Nucleic Acids Research* 42.D1 (2014), pp. 1083–1090. ISSN: 03051048. DOI: 10.1093/nar/gkt1031.
- [43] Antony J. Williams et al. 'Open PHACTS: Semantic interoperability for drug discovery'. In: *Drug Discovery Today* 17.21-22 (2012), pp. 1188–1198. ISSN: 13596446. DOI: 10.1016/j.drudis.2012.05.016.
- [44] Artem Cherkasov et al. 'QSAR Modeling: Where Have You Been? Where Are You Going To?' In: *Journal of Medicinal Chemistry* 57.12 (June 2014), pp. 4977–5010. ISSN: 0022-2623. DOI: 10.1021/jm4004285.
- [45] Ana L. Teixeira, João P. Leal and Andre O. Falcao. 'Random forests for feature selection in QSPR models - An application for predicting standard enthalpy of formation of hydrocarbons'. In: *Journal of Cheminformatics* 5.2 (2013), p. 1. ISSN: 17582946. DOI: 10.1186/1758-2946-5-9.
- [46] Yanli Wang et al. 'PubChem : a public information system for analyzing bioactivities of small molecules'. In: 37.June (2009), pp. 623–633. DOI: 10.1093/nar/gkp456.
- [47] Yanli Wang et al. 'PubChem ' s BioAssay Database'. In: 40.December 2011 (2012). DOI: 10.1093/nar/gkr1132.

REFERENCES

- [48] Bryan L Roth et al. ‘The Multiplicity of Serotonin Receptors : Uselessly Diverse Molecules or an Embarrassment of Riches ?’ In: (2000).
- [49] D. Fourches, E. Muratov and a. Tropsha. ‘Trust but verify: on the importance of chemical structure curation in chemoinformatics and QSAR modeling research’. In: *J. Chem. Inf. Model.* 50.7 (2010), pp. 1189–1204.
- [50] Douglas Young et al. ‘Are the chemical structures in your QSAR correct?’ In: *QSAR and Combinatorial Science* 27.11-12 (2008), pp. 1337–1345. ISSN: 1611020X. DOI: 10.1002/qsar.200810084.
- [51] Denis Fourches and Alexander Tropsha. ‘Using graph indices for the analysis and comparison of chemical datasets’. In: *Molecular Informatics* 32.9-10 (2013), pp. 827–842. ISSN: 18681751. DOI: 10.1002/minf.201300076.
- [52] Alexander Golbraikh et al. ‘Data set modelability by QSAR’. In: *Journal of Chemical Information and Modeling* 54.1 (2014), pp. 1–4. ISSN: 15499596. DOI: 10.1021/ci400572x.
- [53] Alexander Golbraikh et al. *Modelability Criteria: Statistical Characteristics Estimating Feasibility to Build Predictive QSAR Models for a Dataset*. Ed. by Jerzy Leszczynski and Manoj K. Shukla. Boston, MA: Springer US, 2014, pp. 187–230. ISBN: 978-1-4899-7445-7. DOI: 10.1007/978-1-4899-7445-7_7.
- [54] Gilles Marcou, Dragos Horvath and Alexandre Varnek. ‘Kernel Target Alignment Parameter: A New Modelability Measure for Regression Tasks’. In: *Journal of Chemical Information and Modeling* 56.1 (2016), pp. 6–11. ISSN: 15205142. DOI: 10.1021/acs.jcim.5b00539.
- [55] Mark A Johnson and Gerald M Maggiora. *Concepts and applications of molecular similarity*. 1990. ISBN: 0471621757.

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

- [56] A Yasri and D Hartsough. 'Toward an Optimal Procedure for Variable Selection and QSAR Model Building'. In: *Journal of Chemical Information and Computer Sciences* 41.5 (2001), pp. 1218–1227. ISSN: 0095-2338. DOI: 10.1021/ci010291a.
- [57] J C Dearden, M T D Cronin and K L E Kaiser. 'How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR)'. In: *SAR and QSAR in environmental research* 20.December 2012 (2009), pp. 241–266. ISSN: 1062-936X. DOI: 10.1080/10629360902949567.
- [58] Roberto Todeschini and Viviana Consonni. *Handbook of Molecular Descriptors*. Vol. 11. July. Wiley-VCH Verlag GmbH, Weinheim, Germany, 2008, p. 688. ISBN: 9783527613106. DOI: 10.1002/9783527613106.
- [59] Mati Karelson. 'Molecular Descriptors in QSAR/ QSPR'. In: March (2000), p. 35168. ISSN: 1433-7851. DOI: 10.1002/1521-3773(20010316)40:6<1136::AID-ANIE1136>3.0.CO;2-M.
- [60] Anderson Coser Gaudio and Eliana Zandonade. 'PROPOSITION, VALIDATION AND ANALYSIS OF QSAR MODELS'. In: *Quim Nova. SBQ* 24.5 (2001), pp. 658–671. ISSN: 01004042. DOI: 10.1590/S0100-40422001000500013.
- [61] M. M C Ferreira. 'Multivariate QSAR'. In: *Journal of the Brazilian Chemical Society*. Vol. 13. 6. 2002, pp. 742–753. ISBN: 0103-5053. DOI: 10.1590/S0103-50532002000600004.
- [62] Douglas M. Hawkins. 'The Problem of Overfitting'. In: *Journal of Chemical Information and Computer Sciences* 44.1 (2004), pp. 1–12. ISSN: 00952338. DOI: 10.1021/ci0342472.
- [63] Peixun Liu and Wei Long. 'Current mathematical methods used in QSAR/QSPR studies'. In: *International Journal of Molecular Sciences* 10.5 (2009), pp. 1978–1998. ISSN: 14220067. DOI: 10.3390/ijms10051978.

REFERENCES

- [64] Maykel Pérez González et al. ‘Variable selection methods in QSAR: an overview.’ In: *Current topics in medicinal chemistry* 8.18 (2008), pp. 1606–1627. ISSN: 15680266. DOI: 10.2174/156802608786786552.
- [65] Robin Genuer, Jean-michel Poggi and Christine Tuleau-malot. ‘Variable selection using Random Forests’. In: *Pattern Recognition Letters* 31.14 (2012), pp. 2225–2236. DOI: <https://doi.org/10.1016/j.patrec.2010.03.014>.
- [66] Matthias Dehmer et al. *Statistical Modelling of Molecular Descriptors in QSAR / QSPR*. February. Weinheim, Germany: Wiley-VCH Verlag GmbH, 2012, p. 32434. ISBN: 9783527324347.
- [67] Joelle Gola et al. ‘ADMET property prediction: The state of the art and current challenges’. In: *QSAR and Combinatorial Science* 25.12 (2006), pp. 1172–1180. ISSN: 1611020X. DOI: 10.1002/qsar.200610093.
- [68] A Z Dudek, T Arodz and J Galvez. ‘Computational methods in developing quantitative structure-activity relationships (QSAR): a review’. In: *Comb Chem High Throughput Screen* 9.3 (2006), pp. 213–228. ISSN: 13862073. DOI: 10.2174/138620706776055539.
- [69] Robin Genuer, J-M Poggi and C Tuleau. ‘Random Forests : some methodological insights’. In: *Inria* 6729 (2008), p. 32.
- [70] Leo Breiman. ‘Random Forests’. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.
- [71] Gérard Biau. ‘Analysis of a Random Forests Model’. In: *Journal of Machine Learning Research* 13 (2012), pp. 1063–1095. ISSN: 1532-4435.
- [72] Carolin Strobl et al. ‘Conditional Variable Importance for Random Forests for Random Forests’. In: 23 (2008).
- [73] Kristin K Nicodemus et al. ‘The behaviour of random forest permutation-based variable importance measures under predictor correlation’. In: (2010).

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

- [74] Liew Chin Yee and Yap Chun Wei. 'Current Modeling Methods Used in QSAR/QSPR'. In: *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, Mar. 2012, pp. 1–31. ISBN: 9783527324347. DOI: 10.1002/9783527645121.ch1.
- [75] Alexandre Varnek and Igor Baskin. 'Machine Learning Methods for Property Prediction in Chemoinformatics'. In: *Journal of Chemical Information and Modeling* 52.6 (2012), pp. 1413–1437. ISSN: 1549-9596. DOI: 10.1021/ci200409x.
- [76] J C Gertrudes et al. 'Machine learning techniques and drug design.' In: *Current medicinal chemistry* 19.25 (2012), pp. 4289–97. ISSN: 1875-533X. DOI: 10.2174/092986712802884259.
- [77] Dimitar Dobchev, Girinath Pillai and Mati Karelson. 'In silico machine learning methods in drug development'. In: *Current Topics in Medicinal Chemistry* 14.16 (2014), pp. 1913–1922. ISSN: 15680266. DOI: 10.2174/1568026614666140929124203.
- [78] A Statnikov, L Wang and C Aliferis. 'A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification'. In: *BMC Bioinformatics* 9.1 (2008), p. 319. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-319.
- [79] Corinna Cortes and Vladimir Vapnik. 'Support-Vector Networks'. In: *Machine Learning* 20.3 (1995), pp. 273–297. ISSN: 15730565. DOI: 10.1023/A:1022627411411.
- [80] Andrej-Nikolai Spiess and Natalie Neumeyer. 'An evaluation of R² as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach.' In: *BMC pharmacology* 10 (2010), p. 6. ISSN: 1471-2210. DOI: 10.1186/1471-2210-10-6.
- [81] Qiang Zhang and Ingo Muegge. 'Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting, and Consensus Scoring'. In:

REFERENCES

- Journal of Medicinal Chemistry* 49.5 (Mar. 2006), pp. 1536–1548. ISSN: 0022-2623. DOI: 10.1021/jm050468i.
- [82] Georgia Melagraki et al. ‘In silico exploration for identifying structure-activity relationship of MEK inhibition and oral bioavailability for isothiazole derivatives’. In: *Chemical Biology and Drug Design* 76.5 (2010), pp. 397–406. ISSN: 17470277. DOI: 10.1111/j.1747-0285.2010.01029.x.
- [83] Antoni Cortes et al. ‘Inhibition and Types of Inhibition : New Ways of Analysing Data’. In: 268 (2001), pp. 263–268.
- [84] Richard R Neubig et al. ‘International Union of Pharmacology Committee on Receptor Nomenclature and Drug Classification. XXXVIII. Update on terms and symbols in quantitative pharmacology.’ In: *Pharmacological reviews* 55.4 (2003), pp. 597–606. ISSN: 0031-6997. DOI: 10.1124/pr.55.4.4.
- [85] Craig L. Brace et al. ‘Contemporary QSAR classifiers compared’. In: *Journal of Chemical Information and Modeling* 47.1 (2007), pp. 219–227. ISSN: 15499596. DOI: 10.1021/ci600332j.
- [86] R Development Core Team. ‘R: A Language and Environment for Statistical Computing’. In: (2011).
- [87] Asad U Khan. ‘Descriptors and their selection methods in QSAR analysis : paradigm for drug design’. In: 21.8 (2016), pp. 1291–1302.
- [88] Antreas Afantitis et al. ‘Ligand - Based virtual screening procedure for the prediction and the identification of novel β -amyloid aggregation inhibitors using Kohonen maps and Counterpropagation Artificial Neural Networks’. In: *European Journal of Medicinal Chemistry* 46.2 (2011), pp. 497–508. ISSN: 02235234. DOI: 10.1016/j.ejmech.2010.11.029.
- [89] Birgit Viira, Alfonso T. García-Sosa and Uko Maran. ‘QDB archive #202’. In: *QsarDB repository* (2017). ISSN: 18734243.

3. AN AUTOMATED FRAMEWORK FOR QSAR MODEL BUILDING

- [90] Birgit Viira, Alfonso T. García-Sosa and Uko Maran. ‘Chemical structure and correlation analysis of HIV-1 NNRT and NRT inhibitors and database-curated, published inhibition constants with chemical structure in diverse datasets’. In: *Journal of Molecular Graphics and Modelling* 76 (2017), pp. 205–223. ISSN: 18734243. DOI: 10.1016/j.jmgm.2017.06.019.

4

Analysis and comparison of vector space and metric space representations in QSAR modeling

SAMINA KAUSAR AND ANDRE O FALCAO

Abstract

The performance of quantitative structure-activity relationship (QSAR) models largely depends on the relevance of the selected molecular representation, used as input data matrices. This work presents a thorough comparative analysis of two main categories of molecular representations (vector-space and metric-space) for fitting robust machine learning models in QSAR problems. For the assessment of these methods, seven different molecular representations that included RDKit descriptors, five different fingerprints types (MACCS, PubChem,

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

FP2-based, Atom Pair and ECFP4) and a graph matching approach (Non-contiguous atom matching structure similarity; NAMS) in both vector space and metric space, were subjected to state-of-art machine learning methods that included different dimensionality reduction methods (feature selection and linear dimensionality reduction). Five distinct QSAR data sets were used for direct assessment and analysis. Results show that, in general, metric-space and vector-space representations are able to produce equivalent models, but there are significant differences between individual approaches. The NAMS-based similarity approach consistently outperformed most fingerprint representations in model quality closely followed by AtomPair fingerprints. To further verify these findings the metric-space based models were fitted to the same data sets with the closest neighbors removed. These latter results further strengthened the above conclusions, metric space graph-based approach appeared significantly superior to the other representations, albeit at a significant computational cost.

4.1 Introduction

In the past 50 years, quantitative structure-activity relationship (QSAR) has become a powerful tool for drug design and discovery. The underlying principle in QSAR modeling is the assumption that molecular structure information is sufficient to model and predict biological or pharmacological activity. Hence, in QSAR studies, different molecular representations have been used to describe the information encoded in molecular structures so as to predict the quantitative relationships between biological activity (response-variable) and structural information (predictors) [1, 2, 3, 4, 5].

The performance of QSAR models for accurate characterization of biological molecular properties largely depends on the relevance of the selected molecular representation. Such representations can be divided into two broad categories of methods, namely, vector space and metric space representations [6]. A vector space or linear space representation occurs when the set of modeling instances is represented as a vector, with its characteristics measured relative to some reference frame and thus have a notion of magnitude and direction from the origin. In most QSAR modeling studies, vector space is the most common representation used, where each chemical structure is translated using a set of molecular descriptors. This is generally referred as the ‘chemical feature space’, which represents different structural characteristics/properties [5, 7, 8]. Nevertheless, vector space based QSAR modeling has two major modeling issues: Firstly, it is the determination of the set of features capable of structural representation and, secondly, the identification of the subset of features that more significantly are able to predict the desired property [9, 10, 11, 12, 13]. Metric space representation, on the other hand, is built on the principle of measured distances between a set of instances that we want to model. As sometimes it is difficult to identify specific features of a real world entity as a molecule, many times it is easier to quantify its distance or similarity to other instances. A typical case for using metric space representations is in protein functional annotation; while it is quite hard to define a set of features that characterize a protein, the

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

similarity between proteins (whether structural or sequence based) is commonly used to assign its function as it is known that, above a given similarity threshold, proteins maintain its function [14, 15]. In *in silico* screening the similarity principle leads to the simplest database screening methods, if a seed molecule has been determined experimentally as active, the first approach to find other actives is to identify similar molecules, as the probability of finding other actives increases with the proximity to the base molecule [16, 17]. QSAR metric space modeling is hampered also by two different issues. In the first place we need to determine how to measure similarity between molecules - for which there are currently several and conflicting approaches - and secondly, it is necessary to compute the distance of each molecule to all the molecules in the training sets, which may entail difficult computational problems. Distance matrices as they are quadratic to the number of instances of the data set add difficulties to the modeling effort and do not scale well, even with increased computational power available today. Any vector space is a metric space, as it is possible to compute the distances between instances using any common distance metric as the Euclidean distance. On the other hand, there are some data sets for which no vector representation is known (e.g. proteins), while on the other hand, it is possible to compute their distance. Thus, all vector spaces are metric spaces, but the reverse is not true (Figure 4.1).

4.1.1 Molecular similarity and metric space representation

Molecular similarity largely depends upon an appropriate combination of two basic components including (a) a molecular structural representation to find the overlapped or similar features and (b) similarity function/coefficient to quantify the similarity between them [18, 19, 20, 21, 22, 23, 24, 25, 26]. By far, the most commonly used structural representation for comparing molecules is the use of two-dimensional (2D) molecular fingerprints. Fingerprints are a sort of binary fragment descriptors, where each bit represents the hashing product of the possible chemical fragments of a molecule. There are currently several widely used

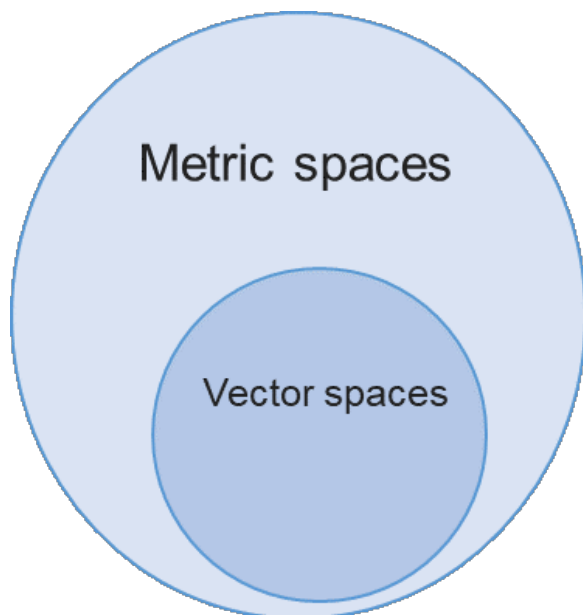


Figure 4.1: Vector space vs metric space

fingerprints that differ on the form that a molecule is decomposed, the size of the representation and the hashing algorithm [27]. Some other descriptor independent methods are also available for molecular similarity comparisons, which include molecular graph matching approaches[28, 29, 30, 31]. To quantify molecular similarity, the most common method used is the Tanimoto (Jaccard) similarity coefficient [32, 33], however, there are many other similarity/distance methods [20, 25, 34, 33, 26]. Also, the one-complement D of the Tanimoto/Jaccard coefficient, where $D = 1 - J$, has been proven to be a real metric, satisfying all the known properties of distance measures [35]. Comparatively to vector space based methods there is limited research work reported in the literature to explore the quantitative relationship between computed molecular similarity-activity in QSAR/QSPR modeling [36, 37, 38, 39, 40, 41, 7, 19, 42, 43, 16, 44, 45].

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

4.1.2 Metric spaces vs vector spaces

With all the aforementioned concerns the main question that we want to address in this study is to know if either a metric space or a vector space modeling approach outperforms the other in QSAR regression problems. Therefore, in this work, we have carried out a comparative analysis of molecular structural representation using some of the most commonly used vector and metric space based methodologies and compare its results. Overall we seek to answer the following four questions:

- (a) Is metric space representation as good as the most common vector space based approaches?
- (b) Which similarity representation carries the maximum chemical/structural information content to establish the best relationship between structural similarity and activity?
- (c) How effective is the use of reduced dimensionality of the feature space with principal components, by replacing explicit descriptors/fingerprints in QSAR modeling?
- (d) Is there any molecular structures representations method that is generally better than the others?

To accomplish these goals the following work was performed: Five distinct data sets with distinct modelability characteristics were selected and curated from ChEMBL23. Then several modeling efforts were applied systematically to all selected data sets, namely i) a typical vector space representation of molecules was performed by using an extensive set of chemical descriptors then used for model fitting in a QSAR optimization framework that includes automated data processing, descriptors/fingerprints computation and feature selection; ii) similarity matrices were computed for all data sets using a variety of methods (five

fingerprint-based and one graph-based), these similarity matrices were then used for modeling by using their principal components as model components; iii) the fingerprint-based representations, as they actually also represent molecular features, have been further used in a vector-based model, using the same linear dimensionality reduction method. For all these three different modeling choices, the number of features (or principal components) used on each model was selected by using 5-fold cross validation, and, each final model was assessed against an independent validation set randomly selected from the initial data set, which was never used in any step of the model fitting phase.

4.2 Methodology

4.2.1 Overview of the methodology

We collected and curated the molecular data for each biological target from ChEMBL23 [46], then all molecules of each data set were represented using different fingerprint models and molecular descriptors and separated into different modeling problems. To perform all the analyses, initially each data set was randomly split into training and independent validation sets (IVS), the former used for training and model selection, and the latter for the final evaluation of the model. A state-of-the-art QSAR modeling approach [47] was performed to build a predictive model using an optimized feature selection procedure. The other models for the same data sets required first the computation of five different fingerprint sets, these were used for additional vector space modeling and for the computation of similarity matrices between all molecules of each data set. Additionally one graph based structural similarity approach (NAMS) was used for making one further similarity matrix for metric-space modeling. PCA was applied to both the similarity matrices and the bare fingerprints so as to create and evaluate models by iteratively increasing the number of principal components. The predictive performance of all data representations was assessed using the IVS,

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

which was never used during feature/PC selection (Figure 4.2). The details of each step of the followed methodology are covered in the following sections.

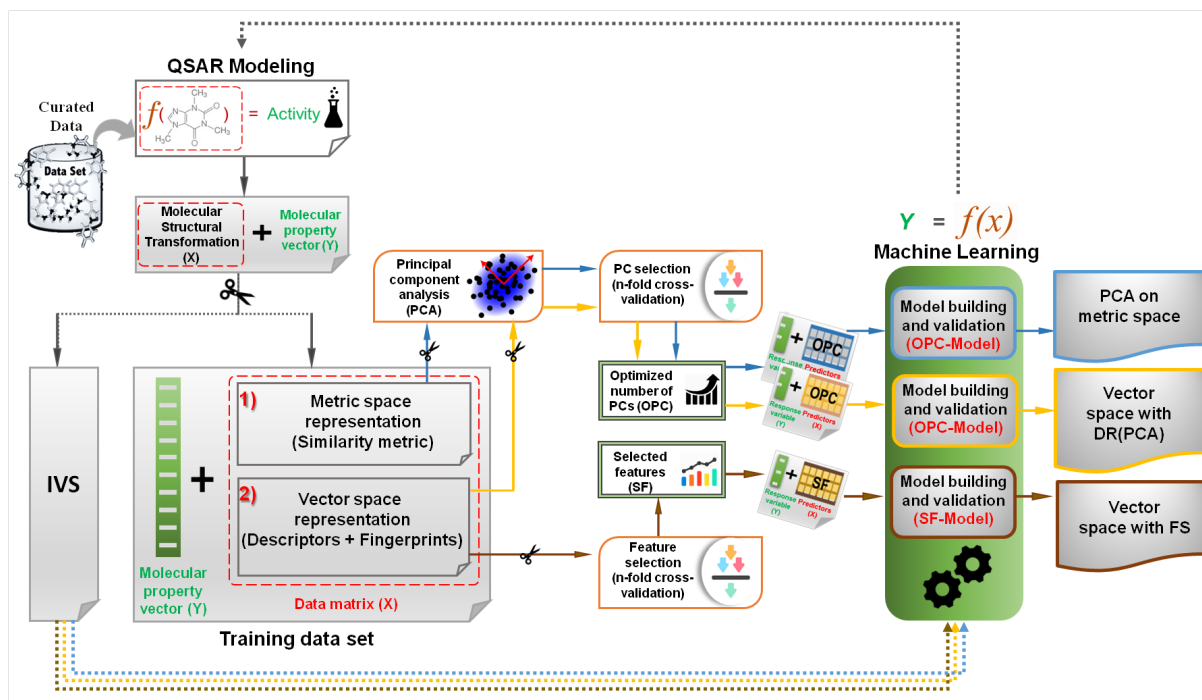


Figure 4.2: QSAR modeling approaches

4.2.2 Vector space representation

In a vector space, each molecule is represented by using a feature vector that contains several molecular properties (descriptors) or structural features represented using a binary array of fixed size (fingerprints) [48, 27].

4.2.2.1 Descriptor based representations

Molecular descriptors aim to selectively describe the information encoded in the structure [48]. Some molecular descriptors are derived with mathematical formulae obtained from

Chemical Graph Theory, Information Theory, Quantum Mechanics among other methods that directly illustrate some relevant features of the molecules [49, 48]. Molecular descriptors can be divided into 4 broader categories: constitutional (1D), topological (2D), geometrical (3D) and physico-chemical properties-based (4D) descriptors [50, 48]. 2D descriptors are the most commonly used types of descriptors.

4.2.2.2 Fingerprint based representations

Another well-known molecular representation is molecular fingerprints, which are fixed-length bit-strings where each bit encoding a fragment or characteristic of a given molecule [27]. Molecular fingerprints are often very different in length and their complexity ranging from 2D/simple representations of relevant structural features to 3D/complicated pharmacophore arrangements. Thus, many types of fingerprints have been generated with different settings (generation method, length, size of patterns and number of bits activated by each pattern etc.) and are further deployed as descriptors for predictive modeling to estimate the biological activities [51, 27, 52, 12, 53, 54].

In principle, 3D representation should have higher information content than 2D, but surprisingly, higher complexity is often more error-prone and less robust in performance [55, 56, 57, 58, 26]. 2D fingerprints can encode different structural information. For example, molecular fragments and structural patterns, topological pathways through compounds, or topological atom environments either as bit strings or feature sets. Numerous software packages have been developed to generate several types of fingerprint for drug discovery applications [54]. Moreover, the basic principle of fingerprints generating algorithms and their comparative performance in a variety of problems has been extensively studied in many reported works [59, 8, 26, 54]. The preferred molecular fingerprints can be grouped into the following three classes:

- (a) Topological/path-based fingerprints (e.g., Daylight like RDkit[27, 60], Atom Pairs[61])

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

capture the paths between atom types by describing their different combinations and always assign same bits position to same substructures within the compared molecules that sometimes results into bit collisions but also useful for clustering compounds.

- (b) Circular fingerprints (e.g., ECFP[62]) record circular atom environments that grow radially from the central atom connections. In topological and circular fingerprints individual bit has no definite meaning.
- (c) Structural keys fingerprints (e.g., MACCS [63], PubChem [64]) each specific bit position represent the presences (1) or absence (0) of predefined functional groups, substructure motifs, or fragments.

2D fingerprints can be easily be calculated by specialized, open-source, and readily available software packages (e.g. OpenBabel [65] or RDkit [60]). 2D fingerprint-based similarity analysis is most widely used methodology in ligand-based virtual screening, clustering and diversity analysis [66, 59, 24, 26, 59, 67].

4.2.3 Metric space representation

A molecule in metric space is defined only as its relation (distance or similarity) to all other molecules in the data set. Technically a metric space is computed using distances between all the elements of a data set creating a distance matrix which can then be used in a variety of modeling techniques, as hierarchical agglomerative clustering or k-Nearest Neighbours models [68, 69]. There is a variety of ways to transform similarities into distances [16, 54], however as all the methodologies for comparing molecules produce similarity matrices, it was deemed unnecessary to transform the similarities into distances and use similarity matrices directly for modeling, as this extra transformation would introduce one further step in the data preparation procedure with no clear advantage.

In descriptors-independent methods, graph matching approaches have been used. In these methods graph theory is used to represent molecules as labelled graphs whose vertices correspond to the atoms and edges correspond to the covalent bonds. Several techniques with some advantages and limitations are available to compare labelled graphs[29]. In the descriptor-independent methods, many advancements have been introduced to improve the sensitivity of graph matching methodology to find consistent and reliable molecular similarity results. One of these methods is the non-contiguous atom matching structural similarity (NAMS) , which has shown modeling advantages over other structural methods[28], although the computational cost of its application can be high.

4.2.3.1 Fingerprint-based similarity

Many types of 2D and 3D molecular fingerprints have been generated to code chemical structures/properties into bit string representations [20, 70, 71]. Molecular fingerprints representation allows for easy comparison of molecules by identifying and quantifying the amount of overlapped elements between them. The applications of molecular fingerprints has been broadly reviewed and used in the literature [72, 54, 70, 22]. There is a large variety of similarity and distance functions that have been introduced and return a molecular similarity score [59, 54]. In cheminformatics the prevalent approach is the use of the Tanimoto coefficient Tc over molecular fingerprints [26, 33]. In the case of 2D fingerprints comparison, for binary vectors of fingerprint representing two molecules A and B, Tc is defined as:

$$Tc(A, B) = \frac{A \cap B}{A \cup B} = \frac{c}{a + b - c} \quad (4.1)$$

In Eq. 4.1, a corresponds to the number of bits set to 1 in molecule A, b is the number of bits set to 1 in molecule B, while c is the number of common set bits in both molecules. As

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

referred above, $1 - T_c$ is an actual distance measure, encompassing all 4 for properties of distance measures.

4.2.3.2 NAMS-based similarity

NAMS is a graph matching algorithm, which uses a new atom alignment method to quantify the structural similarity between compared molecules[28]. NAMS breaks complex molecular structures into simpler parts to reduce molecule to atom-bond-atom structures and calculates global structural similarity score from the best optimal alignment between the atoms of compared molecules. This algorithm has shown an higher discriminant power for biological activity than other structural or graph matching approaches. One of the reasons is that the applied atom matching methodology is able to consider important characteristics of atoms and bonds such as chirality and double bond stereo-isomerism that are many times ignored in other approaches.

Given the structural representation of any two molecules, NAMS is able to compute its similarity score. NAMS can be fine tuned with several parameters that allow users to increase the importance of any specific molecular characteristics (atom or bond similarities, and atomic characteristics like atom stereo isomerism or double bond cis-trans isomerisms). Changing the parameters will change the resulting molecular similarities, but the overall results of comparing large and diverse data sets are not very much changed. For the current work, only the parameters were used.

4.2.4 Model building

In QSAR Modelling, the most well-known machine learning approaches include neural networks (ANN), support vector machines (SVM), decision trees, random forests (RF) and k-Nearest Neighbours [73, 74]. Since the last few years, RF [75] and SVM [76], two non-

linear supervised learning methods, have become the most prevalent algorithms in QSAR studies [77, 75, 78, 79, 80, 81, 82]. One of the biggest advantages of SVM is its ability to deal high dimensional and duplicated data with lower risk of model over fitting [79, 80, 81, 82], while, on the other hand, RF are considered specially robust in complex situations of high dimensional QSAR/QSPR data sets [77, 75, 78]. Hence, RF and SVM are the basic algorithms used in the learning phase of the current work.

As stated, one of the most prevalent issues in QSAR modeling approaches is variable redundancy or colinearity with complex correlation patterns between descriptors or the presence of irrelevant features in the data set, and the present of irrelevant features that may reduce the quality of the produced models. These are consequences of the high dimensionality of such problems. Such issues are aggravated by the fact that in QSAR studies there are many times much more predictors than the number of actual instances to fit. [12, 10, 13, 11, 9], which will make more difficult to find adequate fitted models. Several approaches have been followed in the literature to solve the descriptor selection problem in QSAR modeling [73, 83, 84, 77, 85]. These approaches can be roughly divided into different categories: feature reduction and feature selection. In feature reduction the main purpose is to algebraically combine sets of features into statistically independent new components. There are several methods that purport to accomplish these goals, among which is principal component analysis (PCA), singular value decomposition or kernel PCA[86]. PCA is by far the most used method in feature reduction while Kernel based PCA is beginning to get some traction in the literature[87]. Feature selection, on the other hand, is a more complex problem, and in essence can be resumed to find and select the smallest set of features that are capable of producing the best model. Methods to address this problem include the identification of linear correlations between all variables, bootstrapping methods capable of voting which variables have the highest impact on model quality, or the use of optimization meta-heuristics like genetic algorithms [73, 83, 84, 77, 85].

In this work, we used two of the most common methods for feature reduction. PCA

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

was used in the metric space data produced from the similarity matrices and fingerprint data, while random forests were used to identify the most relevant features capable of producing the highest scoring models.

4.2.4.1 Feature reduction with PCA

Principal Component Analysis (PCA) is a linear reduction method to calculate the most meaningful basis to re-express high dimensional data into a reduced space. However, PCA is a useful tool in QSAR modeling for dealing with the problem of data high dimensionality and collinearity [68, 4]. In typical QSAR studies, PCA is used to analyze original the data matrix in which molecules are represented by several types of predictors variables (molecular descriptors/fingerprints). PCA performs dimensionality reduction by transforming original descriptors space into linear orthogonal combinations of original variables named as principal components (PCs). The generated PCs are uncorrelated and always ranked according to the decreasing data variance of the original variables [68]. As the first components contain the highest amount of data variance, models can be fit to data by gradually incrementing the components in the model. One first model will use only the first component, a second model will use the first two components, and so on, checking which of these models with reduced dimensions is capable of producing the least amount of error in k-fold cross validation. Since each PC is an independent source of the original data variance, PCs have been used as a model input mainly when data high dimensionality is a big issue and most models are sensitive to the number of variables to use [68]. Several studies are reported in the literature where PCA is applied for dimensionality reduction in QSPR/QSAR problems [88, 89, 90, 4].

In this study, we performed PCA in both vector space representations (descriptors and fingerprints data matrices) and metric space representations (fingerprint-based similarity data matrices and NAMS-based similarity data metric). The generated PCs were used to build

QSAR models with dimensionality reduction (DR). We compared the predictive performance of QSAR models generated by the reduced dimensionality of metric space with typical PCA-based QSAR models where vector space is reduced by PCA.

4.2.4.2 Feature selection with Random Forests

A Random Forest (RF) is an ensemble supervised nonlinear machine learning algorithm for classification or regression [75]. This algorithm generates a set of weakly independent Decision Trees that are built using randomly selected subsets of the data. Each generated tree is produced by randomly selecting a set of predictors from the full set and by sampling with replacement instances from the same data pool. This will create a set of randomly generated trees (a forest) each one created from different data and variable partitions. The RF algorithm then uses a consensus voting procedure to combine the predictions from all randomly generated weak models to make more robust predictions. One of the consequences of this bootstrap procedure is that it is possible to assess the power that each variable has in the final predictions. The trees that include such variables will typically have higher prediction power, and as such it is possible to rank each variable in terms of its overall importance to the model quality. Many studies showed that RFs voting procedure can be used for feature selection by ranking and selecting each variable according to its importance in RFs models [91, 85, 77]. In this ensemble method, each variable's importance score is calculated by several variable importance (VI) measures. In regression problems, an increase in the mean squared error of a tree is one of the widely used VI measures, which explains how much prediction error increases with the random permutation of any given variable while keeping all others unchanged in a node of a tree [75, 91, 85, 92]. In this work, we followed the Random Forest (RF) based feature selection method [77] to rank features in high dimensional vector space according to their importance score that are later used in the feedforward feature selection procedure (Figure 4.2).

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

4.2.4.3 Support Vector Machine

An SVM [76] is a supervised machine learning algorithm that has been widely used for classification and regression-based data analysis in many fields, including QSAR studies [79, 80, 81, 82, 77]. For a given set of data instances, a discriminative SVM algorithm focuses on the identification of support vectors (data instances) to draw a decision hyperplane in a high dimensional space that best separate data instances with maximum margins. SVM uses different kernel functions for data transformation in a new hyperplane; these can be linear, radial basis functions, sigmoid or polynomial, which are generally considered good choices for a majority of problems. The discovery of support vectors highly depends on the selected kernel function. Differently from other methodologies where there is a learning phase that heuristically searches thorough the multidimensional feature space, in SVM learning this search procedure is a mathematical optimization procedure and it is guaranteed that an optimal solution can be found in polynomial time. This also implies that, as no random component, is involved, the same solution model will be produced for each model. In this work, we used SVM in the process of feedforward feature selection where PCs from vector/metric reduced dimensionality space and RF importance score based ranked variables from features/vector space were stepwise subjected to the SVM, and final QSAR models were developed with an optimized set of selected dimensions (Figure 4.2). For the current work, for all problems, the radial basis function was selected.

4.2.4.4 Model evaluation and external validation

N-fold cross-validation or model internal validation is the simplest approach, where the training data set is randomly divided into N parts (folds), and each part is used as an external set for the validation of the predictive model, which was fitted by using the remaining compounds in the other $N-1$ partitions. Cross-validation is essential to optimize modeling parameters, variable selection and to verify the internal predictive power and robustness of

the QSAR model [89]. In our analysis, we performed N-fold cross-validation to find an optimized number of most relevant variables (variable/PCs selection). For this purpose, a feed forward approach was used to generate estimation models by sequentially adding the RF importance score based ranked variables (more relevant to least significant) and PCs extracted from vector space and metric spaces as an input in SVM algorithm. The internal predictive performance of each model was assessed by computing the score of the Percentage of Variance Explained (PVE) and Root Mean Squared Error (RMSE) of each predictive model in cross-validation [93]. As the cross-validation may result into a different number of best performing variables for different folds, an average of the PVE score was recorded across all folds each time. Finally, the set of dimensions that lead to the smallest average score of predictive error in all folds was considered as the selected number of descriptors/fingerprints/PCs. After performing all this feature optimization, the whole training data set was reused to develop a model with the selected features to perform a blind external prediction using the independent validation set.

4.3 Data

We tested the proposed QSAR modeling methodology on five data sets for common human biological targets, retrieved from ChEMBL23[46]. These were selected independently of any previous hypothesis (Table 4.1). We used an automated QSAR modeling workflow [47] to collect and curate data for each selected target. The bioactivity data of the selected problems was retrieved using the UniProt accession number (Table 4.1).

Moreover, missing data, salt groups, mixtures (e.g., in unconnected molecules smaller fragments were excluded) were removed. In duplicated data, if more than one record was present for the same compound, the one kept would be its most recent measurement, according to the publication year. All data sets feature K_i as the bioactivity measure. However, the logarithm of K_i is more typically used for modeling and makes more biological sense.

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

Table 4.1: Data set description

Uniprot ID.	Gene Name	Target Protein Name	Associated Bioactivities (Y)	Total Number of Observations (N-Processed)
P35367	HRH1	Histamine H1 receptor	Ki	1222
Q99720	SIGMAR1	Sigma non-opioid intracellular receptor 1	Ki	226
Q12809	HERG	Potassium voltage-gated channel subfamily H member 2	Ki	1481
P35462	DRD3	D(3) dopamine receptor	Ki	2902
P28223	HTR2A	5-hydroxytryptamine receptor 2A	Ki	2088

Also, to encompass several problems of the more extreme values, it was decided to clamp the values between an interval, so that very weak or possibly inactive molecules receive the same low score, while it is many times unnecessary to discriminate results with $Ki \leq 1nM$, as these are very active molecules. Thus the following expression (Eq. 4.2) was used for all data sets to transform Ki into $spKi$ (scaled and clamped pKi):

$$spKi = \begin{cases} 0, & \text{if } Ki \geq 10,000 nM, \\ \frac{4 - \log_{10}(Ki)}{4}, & \text{if } 1nM < Ki < 10,000 nM, \\ 1, & \text{if } Ki \leq 1 nM \end{cases} \quad (4.2)$$

$spKi$ values are thus clamped between 0 and 1, the most active compounds having the values closer or equal to 1, and the lesser actives or inactives will have values of zero. This clamping assumes that Ki values below 1 nM are considered extremely active compounds, while molecules with Ki values above 10,000 nM will be considered as very weak or inactives.

4.3.1 Data preparation for vector and metric space representations

For each problem, molecules were represented in metric and in vector spaces by using 3 different approaches: a) common vector space methods, using molecular descriptors or fingerprints named as vector space with FS (feature selection); b) Principal components over the similarity matrices categorized as PCA on metric space; and c) principal components over molecular descriptors and fingerprints placed in vector space with DR (PCA) (Figure 4.2).

For vector space representation, we used 1348 descriptors (2D and 3D) calculated for each selected problem with RDKit [60] toolkit. Separate modeling efforts were performed by testing separately five different types of fingerprints, which include ECFP6 (Circular), PubChem (Substructure keys) computed using CDK [94] toolkit and MACCS (Substructure keys), RDkit (Path-based) and, Atom Pairs (Path-based) generated by using RDKit[60]. The data preparation for principal components over metric space representation involved the computation of the similarity matrices between all elements of the training set and computing the distances of the IVS to those of the training set. Similarity matrices using the Tanimoto index were obtained for each of the five fingerprints adding the NAMS graph based molecular matching algorithm. Models generated using dimensionality reduction of metric and vector spaces were named ‘optimized number of PC models’ (OPC), as the procedure emphasizes selecting the best number of PCs, capable of producing the more reliable models. Predictive models built using vector space with FS were named as SF-Model (model having selected number of features) (Figure 4.2).

Thus, a total of eighteen different molecular representations were used in this study, and served as input data to machine learning algorithm for the generation of ninety regression models for five selected problems.

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

4.4 Results

4.4.1 Implementation of analysis

All molecular descriptors and fingerprints used in this study were calculated using CDK[94] and RDKit[60] built in nodes of an open source data-mining framework KNIME (version 3.2)[95]. All analysis was performed using R (version 3.4.4) [96] on a desktop workstation powered by a 6th-generation Core i7 Processor (3.41 GHz) with 16 GB RAM. Package e1071 [97] was used for SVM algorithm and an R library RandomForest (RF) [98] for RF. Both SVM and RF algorithms were implemented with the default parameters. R Package factoextra was used for dimension reduction using PCA[99]. This is noteworthy that in the PCA based QSAR modeling, orthogonal projections/PCs for test sets in N-fold CV and IVS were calculated by using R's `PCA predict ()` function.

4.4.2 Results of generated models

OPC-Models and SF-Models were fitted with the whole training data sets of all selected targets. For all data sets, the training data was used to evaluate and select the model that was able to produce the smallest RMSE or PVE (ratio of the variance explained). Typically this involved selecting models with a reduced number of features or PCs. The final models after feature selection were validated using the same IVS for each problem set.

The first aspect that stands out from these results is that the most relevant factor for explaining model quality is the nature of the data itself. Predictive performance of QSAR models highly depends upon different characteristics of the data set (e.g., size, chemical diversity, and presence of activity cliffs) [100, 101, 102, 103, 104, 105, 106]. As an example the *HERG* data set can be easily seen as a difficult problem independently of the approach followed to model it (Figure 4.3), on the other hand, the human Histidine Receptor 1 (*HRH1*)

appears as generally more easily modelable, while the remaining three problems (*SIGMARI*, *DRD3* and *HTR2A*) show intermediate modelability characteristics. Secondly, with some relevant cases noted afterwards, no single method appears uniformly above the others, and each method's performance seems to be heavily dependent upon the data set characteristics.

To have a more encompassing view of the the produced results we performed a Friedman ranked test [107] ; this is a non-parametric test used to assess different treatments applied to different test situations, as is the current case. In the present situation a modeling approach is considered a treatment which is evaluated by its results for the different data sets. Each model is then ranked according to its performance, where the best models have a lower rank and vice-versa. The Friedman test is able then to evaluate each performance according to its rank in all data sets thus effectively providing a performance value for each modeling approach. Another advantage of the Friedman test is that it allows for a post-hoc analysis that is able to better qualify the differences verified between treatments, for instance by grouping similar models with similar performance values. For each modeling data set, the rank in PVE of each modeling approach was calculated in R's agricolae package (Figure 4.4)[108]. The test results showed that there were significant differences between treatments with a Chi -squared test of 38.44 with 17 degrees of freedom, giving a p-value of 2.2×10^{-3} which strongly suggests that there are statistically significant differences between the different modeling approaches.

The post-hoc analysis of the Friedman test allows groupings of statistically indistinct treatments under the same grouping [107]. A treatment can belong to several groups. In Figure 4.4, it is highlighted to which groups each model belongs. The discriminating alpha used was 0.05. It can clearly be seen the the only elements that belong to group *e* - the one with model rankings consistently lower (thus indicating higher quality modeling approaches) are NAMS metric space PCA and Atom Pairs fingerprints with classical feature selection; on the other hand the use of RDkit fingerprints both with metric space representation and PCA dimensionality reduction appear consistently in the highest positions (worst models).

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

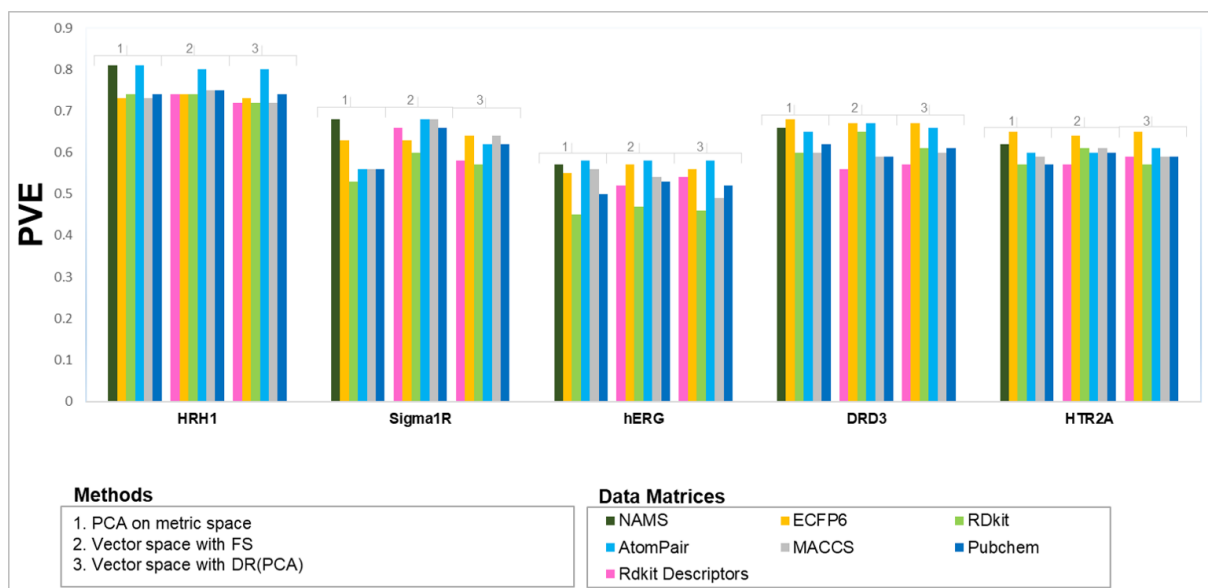


Figure 4.3: Comparisons of QSAR models's predictive performance using IVS. PVE - Percentage of Variance Explained by the model

We further dissected these individualized results according to the four major questions that were the main objectives of our analysis. These questions are addressed one by one in the following sections.

4.4.2.1 Is metric space representation as good as the most common vector space based approaches?

To answer this question, the results of all three different approaches (simple feature selection, PCA dimensionality reduction in both vector spaces and metric spaces) was analyzed. A comparison of OPC-Models generated using PCA on metric and vector spaces and SF-Models built using vector spaces with FS showed that the predictive performance of each QSAR model was influenced by the selected type of molecular structural representation (Figure 4.5), which was expected and consistent with the literature [50, 109, 103]. We performed a similar analysis using the Friedman test over the ranks of the median values of each data

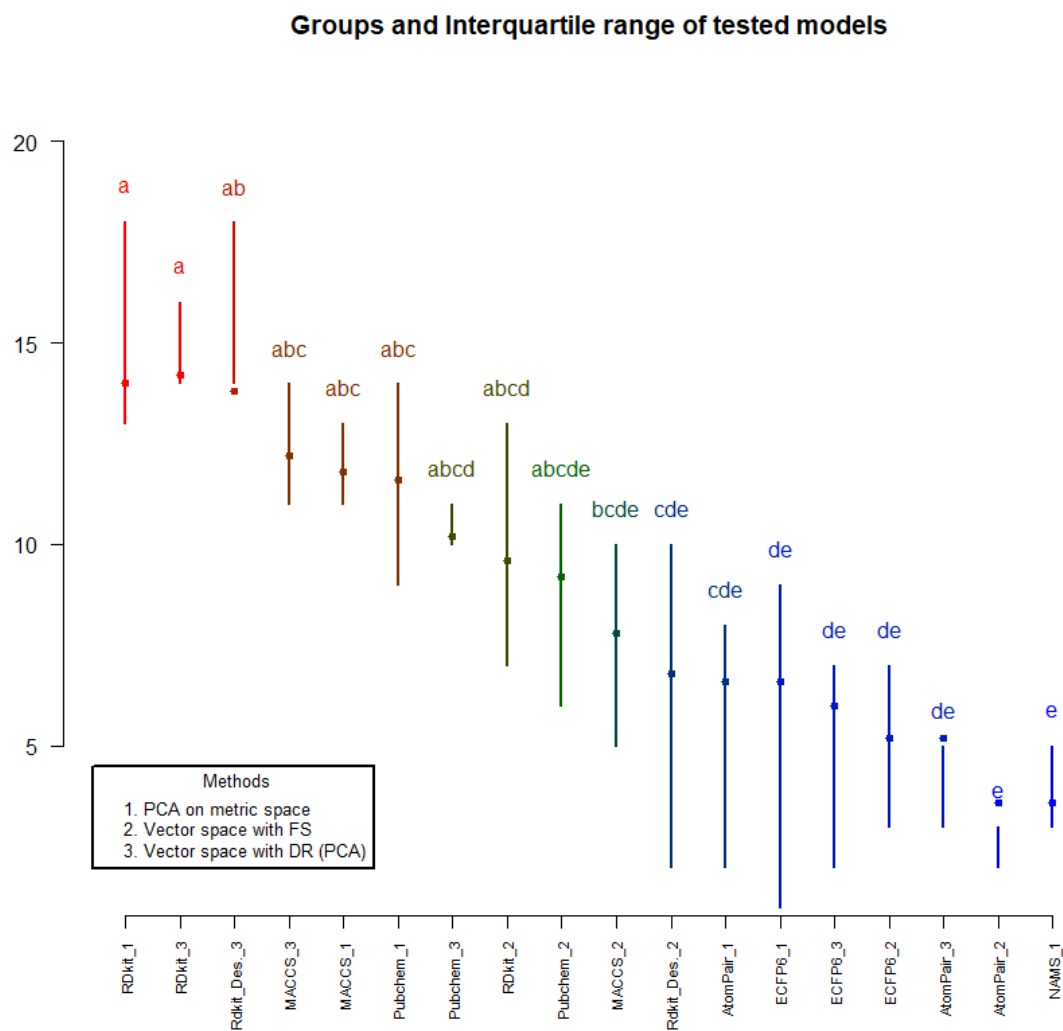


Figure 4.4: Friedman's test results and interquartile ranges of tested models

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

modeling approach from the explained variance (PVE) of the fitted models each respective IVS (Figure 4.5)

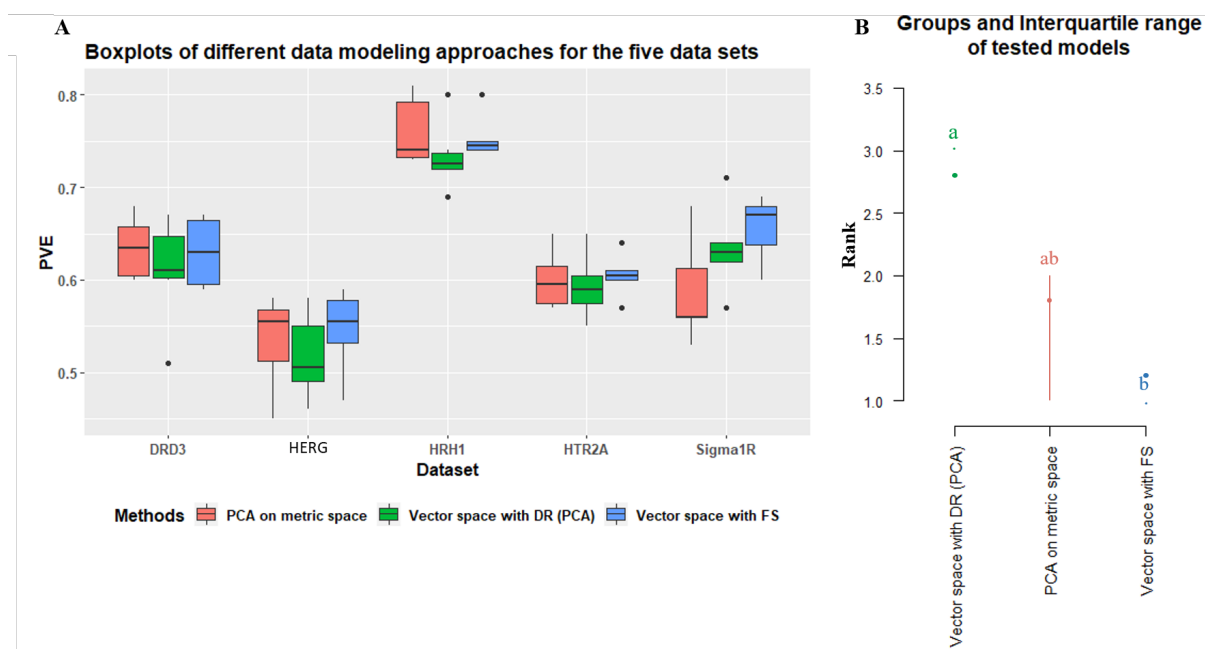


Figure 4.5: (A) Boxplots of the three modeling approaches grouped by the different data sets; (B) Groups and interquartile ranges of the medians of tested models from Friedman's test post-hoc analysis

Feature selection over vector spaces has proven to be globally the most reliably modeling approach and appears to be significantly better relatively to the use of PCA over the same data. Metric Space PCA, appears as somewhere in between, closer to the feature selection approach. The Friedman test for this data yield a Chi-squared value of 6.0, which corresponded to a p-value of 0.049, just below the 0.05 threshold. With such results it is fair to conclude that the usage of metric space data may compromise the quality of the models produced when comparing results to traditional vector-space feature selection models, yet it clearly outperforms vector space PCA based approaches. It is nonetheless striking that the highest ranking method from the overall assessment is NAMS, a Metric space based approach, which may allow us to suggest that possibly the other methods for calculating

molecular similarities may be the culprit for this decreased performance, and may not be as adequate to compute molecular distances, being nonetheless quite effective as descriptor producers for modeling.

4.4.2.2 Which similarity representation carries the maximum chemical/structural information content to establish the best relationship between local similarities and activity?

To analyze which similarity representation contributed more significantly to reliable predictive modeling the overall performance of generated OPC-Models using six similarity data matrices (NAMS, ECFP6, RDkit, Atom Pairs, MACCS, and PubChem based similarities) was evaluated again using Friedman test (Figure 4.6). The ranking of each metric space based approach was assessed for each data set and the overall quality of each model quantified from the use of the Friedman test and respective post-hoc analysis. For the present case NAMS appears clearly as the best approach followed closely by Atom Pairs and ECFP6 fingerprints, the former appearing in the same group as NAMS. The Chi-squared test for the metric-space based approaches ranked comparison was 15.2, with 5 degrees of freedom, which corresponds to a p-value of 9.5×10^{-3} . Thus, test results again suggest that NAMS molecular similarity is able to more reliably capture important structural information, which eventually generates a better quantitative relationship between the local similarities and compound activity.

4.4.2.3 How effective is using a reduced dimensionality of the metric/vector space with Principal Components, replacing explicit descriptors/fingerprints in QSAR modeling?

This question can actually be answered by observing the previous results. It seems clear that when directly comparing PCA over direct feature selection (Figure 4.5), the latter pro-

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

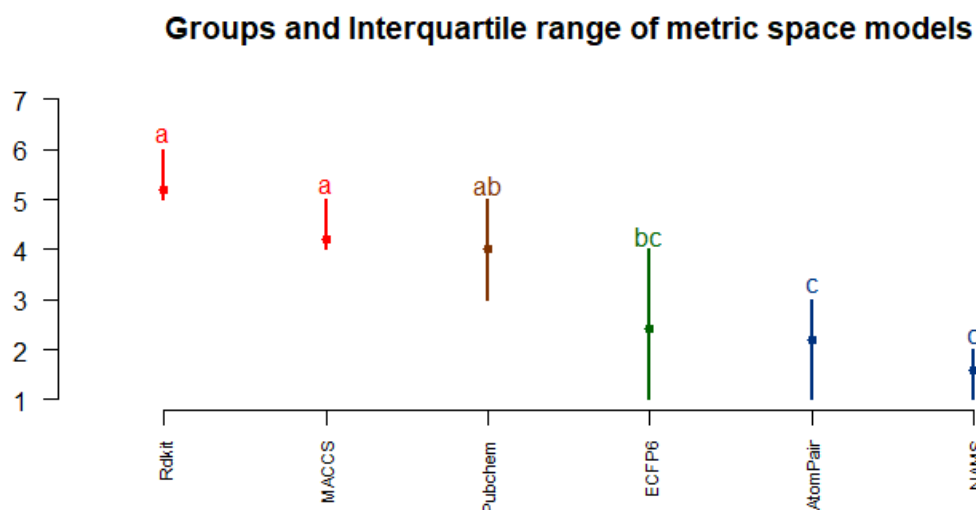


Figure 4.6: Overall performance of similarity representation using PCA on metric space based QSAR modeling approach

duces markedly better results, which strongly suggests that the dimensionality reduction achieved with PCA is a poor proxy for a better structured search for the most relevant descriptors in a modeling problem. Nonetheless, using PCs from the similarity matrix allows us to capture the same information available from vector space modeling. These results also highlight the capability of fingerprints for producing high quality models, without the need for other chemical descriptors. Furthermore, the fingerprint generating method appears critical for producing the most reliable models. As it is patent from the above results, Atom Pairs and ECFP6 fingerprints appear as the best fingerprint-based similarity approaches, while the RDkit and PubChem fingerprints appear consistently lagging behind all other models.

4.4.2.4 Is there any solution that is globally better on a variety of difficult problems?

From the above results it is clear that there is not a single best approach for dealing with complex QSAR problems. Although metric-space based NAMMS and Atom Pairs come

most of the times on the top places, they are not consistent for all data sets. For instance, Atom Pairs fingerprint representation perform poorly for the HTR2A model, while NAMS does not appear on top for the DRD3 data set. Similarly, as referred above, there does not appear any intrinsic advantage in changing from a fingerprint vector-space based approach to similarity based metric-space modeling. The most consistent result was that the use of PCA with descriptor data was generally a poor modeling approach. PCA can be used nonetheless with distance matrices being able to capture reliable information for modeling.

4.5 Discussion

Many studies have demonstrated that the selection of different types of molecular structural representation has a larger impact on the predictability of QSAR models than the choice of model optimization methods [109, 110, 59, 8, 26, 54, 67]. Our results confirm these findings further suggesting that reduced metric space representation using NAMS-based similarity and Atom Pairs fingerprints with feature selection were the methods that more consistently address a variety of modeling problems.

Nonetheless one further concern over such studies is how much novel information is actually being discovered from those models as it is a known fact that similar molecules tend to have similar biological properties. Therefore, a distinct possibility is that the usage of similarity matrices for inference may be able to make reliable predictions only when very similar molecules to the training data set are present. Thus, one further test for these modeling approaches is to understand how reliable are these methods for making models where all very similar molecules have been removed, and no molecule, either in the training set or the IVS, has a high similarity to any other. This would allow to evaluate the capability of each approach for making inference when very diverse compounds are fed into the model. Therefore, to check the robustness of the tested methodologies, the five data sets were manipulated by converting them into harder problems with only structurally diverse molecules,

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

making certain that no molecules within a given similarity threshold are present in each data set. Accordingly, five new data sets were created based on the initial ones but where no molecule would be present if it was similar within a given threshold, to the others already present. As different similarity methods produce different scores for the same molecules, the thresholds were adjusted for each similarity methods, to make sure that model would be trained with a similar number of instances (4.2). This complementary analysis relates obviously only to metric-space modeling, as such the following results will only focus on this modeling approach.

Table 4.2: Data size before and after removing nearest neighbours - Thr - Similarity threshold; N - New data set size

Target Protein Name	Data size without removing nearest neighbours	NAMS		ECFP6		RDkit		Atom Pairs		MACCS		Pubchem	
		Thr	N	Thr	N	Thr	N	Thr	N	Thr	N	Thr	N
Histamine H1 receptor (HRH1)	1222	0.80	379	0.55	378	0.80	371	0.67	376	0.84	379	0.87	391
Sigma non-opioid intracellular receptor 1 (Sigma1R)	226	0.87	312	0.61	310	0.89	305	0.75	309	0.92	311	0.94	321
Potassium voltage-gated channel subfamily H member 2 (HERG)	1481	0.80	397	0.54	394	0.82	392	0.69	395	0.83	395	0.86	403
D(3) dopamine receptor (DRD3)	2902	0.80	478	0.52	481	0.77	470	0.67	480	0.87	484	0.86	484
5-hydroxytryptamine receptor 2A	2088	0.80	432	0.47	432	0.78	424	0.63	426	0.83	429	0.85	437

After removing the nearest neighbours, all data sets were again randomly split into training and independent validation sets and the same data processing procedures were repeated for these new more challenging data sets. Also the same modeling principles were repeated by training the models with the training set, while simultaneously selecting the best feature set, and finally validating the best model with the corresponding IVS. The overall performance of the same models over these new data sets was assessed, nonetheless because the number of instances present in all new problems are different, both RMSE and PVE were used to adequately assess each model's performance (Figure 4.7).

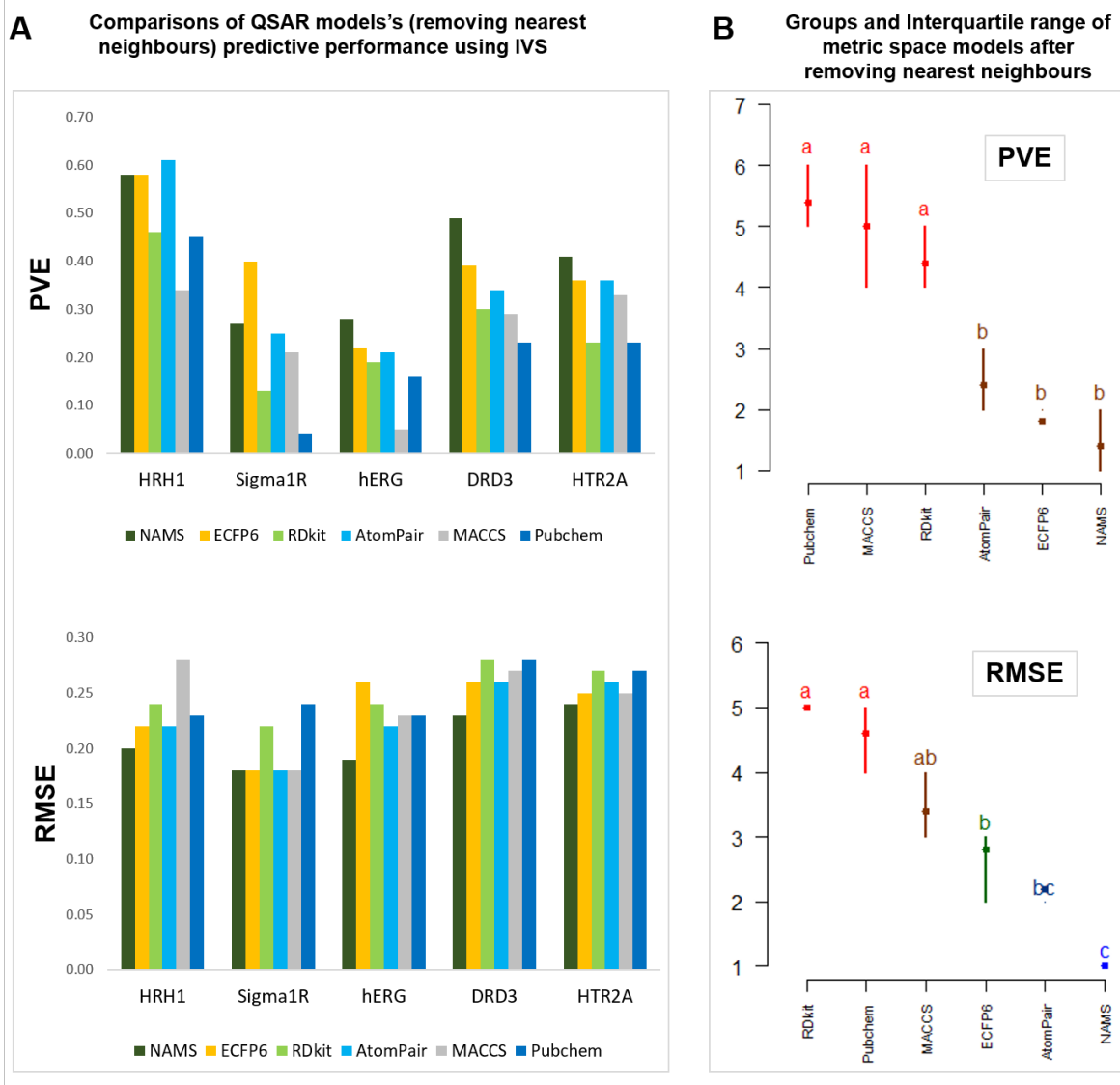


Figure 4.7: Overall performance of metric space representation after removing nearest neighbours in PCA on metric space based QSAR modeling approach

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

As it can be seen, with such hampered data the performance of QSAR models has naturally dropped, with a decrease in PVE ranging from 0.15 to 0.52 (Figure 4.7A). This finding is consistent with the literature[8], in that similar molecules present in models tend to inflate results statistics. It can also now be promptly evaluated that the differences between the different models are now amplified, being clearly easier to visually identify which approaches distinguish themselves from all others. Nonetheless the overall model ranking was not significantly changed. Thus NAMS similarity representation was, for these data sets, clearly the highest performing model, achieving for all cases the lowest RMSE scores. Using the Friedman's interquartile range graph (Figure 4.7B) using grouped together with Atom Pairs or ECPF6, depending on using PVE or RMSE as the performance score. All other fingerprint approaches were not up to the referred methods in these more difficult challenges. The Friedman test for the PVE had a Chi-squared value of 21.1 p-value 7.8×10^{-10} .

4.5.1 Computation time

The execution time of QSAR models built from reduced dimensionality of metric space ranged between 60.61 to 48.88 minutes and for vector space 52.53 to 15.34 minutes while vector space with FS computational time ranges between 860 minutes (DRD3) to 17 minutes (Sigma1R). Comparative analysis of computational time showed that reduced dimensionality significantly reduced the complexity of the problem in hands and then eventually computational time cost is also decreased.

The computation time is an important issue when comparing different modeling approaches, especially when the use of metric space methods is being evaluated as the use of a full similarity matrix is required for each data set. Furthermore, metric space modeling requires that one of the steps for inference is that for each new molecule its distance to all of the molecules in the training set is assessed. This is not typically a problem for academic studies but may put a large computational burden for actual screening efforts when several

millions of molecules are being evaluated. This problem is aggravated in the case of the specific non-fingerprint approach we tested (NAMS). Although apparently able to produce a more accurate distance that translates into better prediction models overall, it does so at a much higher computational cost. With the current common hardware, the average computational cost to compute the similarity of two molecules is 12 ms, which for many problems may be too high. As an example, for computing the similarity of one new molecule to a training set of 1000 molecules it will require 12 Secs. Such computational costs (although the problem is trivially parallelizable), for very large data sets, may collocate unacceptable computational costs.

4.6 Conclusions

In this study we compared different molecular representation approaches for input into QSAR machine learning methods. These approaches were divided into vector-space and metric-space based, the former, where each molecule is represented as a vector of different characteristics and the latter where a molecule is represented with its distance or similarity to others of known activity. We have tested 5 different fingerprint types (RDKit-FP2-based, MACCS, PubChem, Atom Pairs and Morgan's ECFP6) both as vectors of descriptors and, in metric space approaches, with Tanimoto scores computed for similarity. One exclusively vector-space approach was also tested, where common chemical descriptors were computed and used in vector-space modeling as well as a pure metric-space method with a molecular graph-based similarity (NAMS). We also tested whether it was more adequate to use dimensionality reduction methods (as with PCA) or a more computer-intensive feature selection procedure. These representation and dimensionality reduction methods were tested over five different data sets of different modelability, and analyzed through Friedman's test for ranking models. Results showed that, the choice of molecular representations to compute molecular similarity is more important than the modeling approach followed, thus certain

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

methods produced consistently better results. ECFP6 and Atom Pairs fingerprints were the clear best approaches for modeling in vector spaces, surpassing all other methods. Classic molecular descriptors do not show any advantage for any of the data sets in this study. Regarding dimensionality reduction methods, the use of principal components appeared to be inferior to the use of random forest-based feature selection. The former method, albeit faster to process, produced in general results not on par to the latter.

In his study, metric-space modeling by itself, did not appear clearly superior to a vector-space based approach and, for the same representation, using Fingerprints as descriptors tended to produce better results than using molecular distances from those same fingerprints. However when using metric-space representations it becomes even more clear the differences between similarity methods, where NAMS and Atom Pairs fingerprints appear objectively better than all other representations. Finally to verify whether metric-space based representations can be used for more remote inference, where the chemical space is evaluated in regions distant from from the training data, the above conclusions regarding metric-space modeling appear amplified with a larger distance between similarity methods, where NAMS and Atom Pairs fingerprints appear clearly separated from the others.

Finally, metric-space based methods are more computationally expensive, requiring, for each new molecule, the computation of molecular similarities to each instance of the training set. This is a particularly severe cost for the graph-based similarity algorithm used (NAMS), where the computation cost is a serious factor that may hamper its applicability in a real world virtual screening approach, despite being overall the method that is more consistently capable of producing high quality QSAR models.

Acknowledgements

The authors gratefully acknowledge Fundação para a Ciência e Tecnologia for a doctoral grant (SFRH/BD/111654/2015), MIMED project funding (PTDC/EEI-ESS/4923/2014) and LASIGE Research Unit, ref. UID/CEC/00408/2019 for providing the infrastructure.

References

- [1] Artem Cherkasov et al. ‘QSAR Modeling: Where Have You Been? Where Are You Going To?’ In: *Journal of Medicinal Chemistry* 57.12 (June 2014), pp. 4977–5010. ISSN: 0022-2623. DOI: 10.1021/jm4004285.
- [2] A Z Dudek, T Arodz and J Galvez. ‘Computational methods in developing quantitative structure-activity relationships (QSAR): a review’. In: *Comb Chem High Throughput Screen* 9.3 (2006), pp. 213–228. ISSN: 13862073. DOI: 10.2174/138620706776055539.
- [3] Corwin Hansch et al. ‘Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients’. In: *Nature* 194.4824 (1962), pp. 178–180. ISSN: 0028-0836. DOI: 10.1038/194178b0.
- [4] Changkyoo Yoo and Mohsen Shahlaei. ‘The applications of PCA in QSAR studies: A case study on CCR5 antagonists.’ In: *Chemical biology & drug design* (2017). ISSN: 1747-0285. DOI: 10.1111/cbdd.13064.
- [5] Roberto Todeschini and Viviana Consonni. *Handbook of Molecular Descriptors*. Vol. 11. July. Wiley-VCH Verlag GmbH, Weinheim, Germany, 2008, p. 688. ISBN: 9783527613106. DOI: 10.1002/9783527613106.
- [6] Edgar Chávez et al. ‘Searching in Metric Spaces’. In: *ACM Comput. Surv.* 33.3 (Sept. 2001), pp. 273–321. ISSN: 0360-0300. DOI: 10.1145/502807.502808. URL: <http://doi.acm.org/10.1145/502807.502808>.

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

- [7] Johann Gasteiger. *Handbook of Chemoinformatics : From Data to Knowledge*. Vol. 1-4. May. Weinheim: Wiley-VCH, 2008, pp. 1–1870. ISBN: 9783527618279. DOI: 10.1002/9783527618279.
- [8] Noel M. O’Boyle and Roger A Sayle. ‘Comparing structural fingerprints using a literature-based similarity benchmark’. In: *Journal of Cheminformatics* 8.1 (Dec. 2016), p. 36. ISSN: 1758-2946. DOI: 10.1186/s13321-016-0148-0.
- [9] A Yasri and D Hartsough. ‘Toward an Optimal Procedure for Variable Selection and QSAR Model Building’. In: *Journal of Chemical Information and Computer Sciences* 41.5 (2001), pp. 1218–1227. ISSN: 0095-2338. DOI: 10.1021/ci010291a.
- [10] T Puzyn, J Leszczynski and M T Cronin. *Recent Advances in QSAR Studies: Methods and Applications (Challenges and Advances in Computational Chemistry and Physics)*. Ed. by Mark T. D. Cronin Tomasz Puzyn (Editor), Jerzy Leszczynski (Editor). 2010 editi. Springer, 2009. ISBN: 9781402097829.
- [11] J C Dearden, M T D Cronin and K L E Kaiser. ‘How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR)’. In: *SAR and QSAR in environmental research* 20.December 2012 (2009), pp. 241–266. ISSN: 1062-936X. DOI: 10.1080/10629360902949567.
- [12] Alexander Tropsha and Alexander Golbraikh. ‘Predictive QSAR modeling workflow, model applicability domains, and virtual screening.’ In: *Current pharmaceutical design* 13.34 (2007), pp. 3494–504. ISSN: 1873-4286. DOI: 10.2174/13816120778279425
- [13] Alexander Tropsha. ‘Best practices for QSAR model development, validation, and exploitation’. In: *Molecular Informatics* 29.6-7 (2010), pp. 476–488. ISSN: 18681743. DOI: 10.1002/minf.201000061.
- [14] Arthur M Lesk. *Introduction to Bioinformatics, Fourth Edition*. Oxford, United Kingdom : Oxford University Press, 2014, p. 400. ISBN: 9780199651566.

REFERENCES

- [15] Christine A. Orengo and Alex Bateman. *Protein Families: Relating Protein Sequence, Structure, and Function*. Ed. by Christine Orengo and Alex Bateman. Hoboken, NJ, USA: John Wiley & Sons, Inc., Dec. 2013, p. 552. ISBN: 9781118743089. DOI: 10.1002/9781118743089.
- [16] Ana L Teixeira and Andre O Falcao. ‘Structural similarity based kriging for quantitative structure activity and property relationship modeling.’ In: *Journal of chemical information and modeling* 54.7 (2014), pp. 1833–1849. ISSN: 1549-960X. DOI: 10.1021/ci500110v.
- [17] Yvonne C. Martin, James L. Kofron and Linda M. Traphagen. ‘Do Structurally Similar Molecules Have Similar Biological Activity?’ In: *Journal of Medicinal Chemistry* 45.19 (Sept. 2002), pp. 4350–4358. ISSN: 0022-2623. DOI: 10.1021/jm020155c.
- [18] Nina Nikolova and Joanna Jaworska. ‘Approaches to Measure Chemical Similarity—a Review’. In: *QSAR & Combinatorial Science* 22.910 (2003), pp. 1006–1026. ISSN: 1611-020X. DOI: 10.1002/qsar.200330831.
- [19] Mark A Johnson and Gerald M Maggiora. *Concepts and applications of molecular similarity*. 1990. ISBN: 0471621757.
- [20] Peter Willett, John M. Barnard and Geoffrey M. Downs. ‘Chemical Similarity Searching’. In: *Journal of Chemical Information and Computer Sciences* 38.6 (Nov. 1998), pp. 983–996. ISSN: 0095-2338. DOI: 10.1021/ci9800211.
- [21] Andreas Bender and Robert C Glen. ‘Molecular similarity: a key technique in molecular informatics.’ In: *Organic & biomolecular chemistry* 2.22 (2004), pp. 3204–3218. ISSN: 1477-0520. DOI: 10.1039/b409813g.
- [22] Gerald Maggiora et al. ‘Molecular Similarity in Medicinal Chemistry’. In: *Journal of Medicinal Chemistry* 57.8 (Apr. 2014), pp. 3186–3204. ISSN: 0022-2623. DOI: 10.1021/jm401411z.

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

- [23] Hanna Eckert and Jürgen Bajorath. ‘Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches’. In: *Drug Discovery Today* 12.5-6 (2007), pp. 225–233. ISSN: 13596446. DOI: 10.1016/j.drudis.2007.01.011.
- [24] Dagmar Stumpfe and Jürgen Bajorath. ‘Similarity searching’. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.2 (Mar. 2011), pp. 260–282. ISSN: 17590876. DOI: 10.1002/wcms.23.
- [25] Gerald M Maggiora and Veerabahu Shanmugasundaram. ‘Molecular Similarity Measures’. In: *Methods in molecular biology (Clifton, N.J.)* 2004, pp. 1–50. DOI: 10.1385/1-59259-802-1:001.
- [26] Jürgen Bajorath. ‘Molecular Similarity Concepts for Informatics Applications’. In: *Bioinformatics: Volume II: Structure, Function, and Applications*. Ed. by Jonathan M. Keith. New York, NY: Springer New York, 2017, pp. 231–245. ISBN: 978-1-4939-6613-4. DOI: 10.1007/978-1-4939-6613-4_13.
- [27] C. James, D. Weininger and J. Delaney. *Daylight Theory Manual version 4.9*. 2011.
- [28] Ana L Teixeira and Andre O Falcao. ‘Noncontiguous atom matching structural similarity function’. In: *Journal of Chemical Information and Modeling* 53.10 (2013), pp. 2511–2524. ISSN: 15499596. DOI: 10.1021/ci400324u.
- [29] Hans-Christian Ehrlich and Matthias Rarey. ‘Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review’. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.1 (Jan. 2011), pp. 68–79. ISSN: 17590876. DOI: 10.1002/wcms.5.
- [30] John W. Raymond and Peter Willett. ‘Maximum common subgraph isomorphism algorithms for the matching of chemical structures.’ In: *Journal of computer-aided molecular design* 16.7 (July 2002), pp. 521–33. ISSN: 0920-654X. DOI: 10.1023/A:1021271615909.

REFERENCES

- [31] John M. Barnard. ‘Substructure searching methods: Old and new’. In: *Journal of Chemical Information and Modeling* 33.4 (July 1993), pp. 532–538. ISSN: 1549-9596. DOI: 10.1021/ci00014a001.
- [32] D.R. Flower. ‘On the Properties of Bit String-Based Measures of Chemical Similarity’. In: *Journal of Chemical Information and Modeling* 38.3 (1998), pp. 379–386. ISSN: 1549-9596. DOI: 10.1021/ci970437z.
- [33] Dávid Bajusz, Anita Rácz and Károly Héberger. ‘Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?’ In: *Journal of Cheminformatics* 7.1 (2015), pp. 1–13. ISSN: 17582946. DOI: 10.1186/s13321-015-0069-3.
- [34] Amos Tversky. ‘Features of similarity.’ In: *Psychological Review* 84.4 (1977), pp. 327–352. ISSN: 0033-295X. DOI: 10.1037/0033-295X.84.4.327.
- [35] Jure Leskovec, Anand Rajaraman and Jeffrey David Ullman. *Mining of Massive Datasets*. 2nd. New York, NY, USA: Cambridge University Press, 2014. ISBN: 1107077230, 9781107077232.
- [36] R Benigni et al. ‘Molecular similarity matrices and quantitative structure-activity relationships: a case study with methodological implications’. In: *Journal of Medicinal Chemistry* 38 (1995), pp. 629–635.
- [37] Sung-Sau So and Martin Karplus. ‘Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 2. Applications’. In: *Journal of medicinal chemistry* 40.26 (1997), pp. 4360–4371. ISSN: 0022-2623. DOI: S0022-2623(97)00488-3.
- [38] David Robert, Lluís Amat and Ramon Carbó-Dorca. ‘Quantum similarity QSAR: Study of inhibitors binding to thrombin, trypsin, and factor Xa, including a comparison with CoMFA and CoMSIA methods’. In: *International Journal of Quantum*

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

- Chemistry* 80.3 (2000), pp. 265–282. ISSN: 00207608. DOI: 10.1002/1097-461X(2000)80:3<265::AID-QUA1>3.0.CO;2-K.
- [39] Xavier Gironés and Ramon Carbó-Dorca. ‘Molecular quantum similarity-based QSARs for binding affinities of several steroid sets’. In: *Journal of Chemical Information and Computer Sciences* 42.5 (2002), pp. 1185–1193. ISSN: 00952338. DOI: 10.1021/ci0202842.
- [40] Emili Besalú et al. ‘Molecular quantum similarity and the fundamentals of QSAR’. In: *Accounts of Chemical Research* 35.5 (2002), pp. 289–295. ISSN: 00014842. DOI: 10.1021/ar010048x.
- [41] R. Carbo-Dorca. ‘About the prediction of molecular properties using the fundamental Quantum QSPR (QQSPR) equation†’. In: *SAR and QSAR in Environmental Research* 18.3-4 (May 2007), pp. 265–284. ISSN: 1062-936X. DOI: 10.1080/10629360701304113.
- [42] R Carbo-Dorca and P G Mezey. *Advances in Molecular Similarity*. Advances in Molecular Similarity v. 2. Elsevier Science, 1999, p. 296. ISBN: 9780080552262.
- [43] Manuel Urbano Cuadrado, Irene Luque Ruiz and Miguel Ángel Gómez-Nieto. ‘A Steroids QSAR Approach Based on Approximate Similarity Measurements’. In: *Journal of Chemical Information and Modeling* 46.4 (July 2006), pp. 1678–1686. ISSN: 1549-9596. DOI: 10.1021/ci0600511.
- [44] Tobias Girschick et al. ‘Similarity boosted quantitative structure-activity relationship - A systematic study of enhancing structural descriptors by molecular similarity’. In: *Journal of Chemical Information and Modeling* (2013). ISSN: 15499596. DOI: 10.1021/ci300182p.
- [45] I. Luque Ruiz and M. Á. Gómez-Nieto. ‘QSAR classification and regression models for β -secretase inhibitors using relative distance matrices’. In: *SAR and QSAR in*

REFERENCES

- Environmental Research* 29.5 (May 2018), pp. 355–383. ISSN: 1062-936X. DOI: 10.1080/1062936X.2018.1442879.
- [46] Anna Gaulton et al. ‘The ChEMBL database in 2017’. In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D945–D954. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1074.
- [47] Samina Kausar and Andre O. Falcao. ‘An automated framework for QSAR model building’. In: *Journal of Cheminformatics* 10.1 (Dec. 2018), p. 1. ISSN: 1758-2946. DOI: 10.1186/s13321-017-0256-5.
- [48] Roberto Todeschini and Viviana Consonni. *Molecular Descriptors for Chemoinformatics*. Ed. by Roberto Todeschini and Viviana Consonni. Methods and Principles in Medicinal Chemistry. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, July 2009. ISBN: 9783527628766. DOI: 10.1002/9783527628766.
- [49] Alan R. Katritzky, Victor S. Lobanov and Mati Karelson. ‘QSPR: the correlation and quantitative prediction of chemical and physical properties from structure’. In: *Chemical Society Reviews* 24.4 (1995), pp. 279–87. ISSN: 0306-0012. DOI: 10.1039/cs9952400279.
- [50] Johann Gasteiger. *Handbook of Chemoinformatics*. Ed. by Johann Gasteiger. Vol. 1-4. Weinheim, Germany: Wiley-VCH Verlag GmbH, Aug. 2003, pp. 1–1870. ISBN: 9783527618279. DOI: 10.1002/9783527618279.
- [51] Jürgen Bajorath. *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*. Ed. by Jürgen Bajorath. Vol. 275. Methods in Molecular Biology. Totowa: Humana Press, 2004. ISBN: 978-1-58829-261-2. DOI: 10.1385/1592598021.
- [52] Kunal Roy, Supratik Kar and Rudra Narayan Das. *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*. Elsevier, 2015. ISBN: 9780128015056. DOI: 10.1016/C2014-0-00286-9.

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

- [53] Alexandre Varnek and Igor I. Baskin. *Chemoinformatics as a theoretical chemistry discipline*. 2011. DOI: 10.1002/minf.201000100.
- [54] Adrià Cereto-Massagué et al. ‘Molecular fingerprint similarity search in virtual screening’. In: *Methods* 71 (Jan. 2015), pp. 58–63. ISSN: 10462023. DOI: 10.1016/j.ymeth.2014.08.005.
- [55] Georgia B. McGaughey et al. ‘Comparison of Topological, Shape, and Docking Methods in Virtual Screening’. In: *Journal of Chemical Information and Modeling* 47.4 (July 2007), pp. 1504–1519. ISSN: 1549-9596. DOI: 10.1021/ci700052x.
- [56] Ingo Muegge. ‘Synergies of Virtual Screening Approaches’. In: *Mini-Reviews in Medicinal Chemistry* 8.9 (Aug. 2008), pp. 927–933. ISSN: 13895575. DOI: 10.2174/138955708785132792.
- [57] Robert P. Sheridan and Simon K. Kearsley. ‘Why do we need so many chemical similarity search methods?’ In: *Drug Discovery Today* 7.17 (Sept. 2002), pp. 903–911. ISSN: 13596446. DOI: 10.1016/S1359-6446(02)02411-X.
- [58] Qiang Zhang and Ingo Muegge. ‘Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting, and Consensus Scoring’. In: *Journal of Medicinal Chemistry* 49.5 (Mar. 2006), pp. 1536–1548. ISSN: 0022-2623. DOI: 10.1021/jm050468i.
- [59] Ingo Muegge and Prasenjit Mukherjee. ‘An overview of molecular fingerprint similarity search in virtual screening’. In: *Expert Opinion on Drug Discovery* 11.2 (2016), pp. 137–148. ISSN: 1746-0441. DOI: 10.1517/17460441.2016.1117070.
- [60] Greg Landrum. *RDKit Documentation*. 2018. (Visited on 03/09/2018).
- [61] Raymond E. Carhart, Dennis H. Smith and R. Venkataraghavan. ‘Atom pairs as molecular features in structure-activity studies: definition and applications’. In: *Journal of Chemical Information and Modeling* 25.2 (May 1985), pp. 64–73. ISSN: 1549-9596. DOI: 10.1021/ci00046a002.

REFERENCES

- [62] David Rogers and Mathew Hahn. ‘Extended-Connectivity Fingerprints’. In: *Journal of Chemical Information and Modeling* 50.5 (May 2010), pp. 742–754. ISSN: 1549-9596. DOI: 10.1021/ci100050t.
- [63] Joseph L. Durant et al. ‘Reoptimization of MDL Keys for Use in Drug Discovery’. In: *Journal of Chemical Information and Computer Sciences* 42.6 (Nov. 2002), pp. 1273–1280. ISSN: 0095-2338. DOI: 10.1021/ci010132r.
- [64] *PubChem Substructure Fingerprint*. (Visited on).
- [65] Noel M. O’Boyle et al. ‘Open Babel: An open chemical toolbox’. In: *Journal of Cheminformatics* 3.1 (2011), p. 33. ISSN: 1758-2946. DOI: 10.1186/1758-2946-3-33.
- [66] Peter Willett. ‘The Calculation of Molecular Structural Similarity: Principles and Practice’. In: *Molecular Informatics* 33.6-7 (June 2014), pp. 403–413. ISSN: 18681743. DOI: 10.1002/minf.201400024.
- [67] Swarit Jasial et al. ‘Activity-relevant similarity values for fingerprints and implications for similarity searching’. In: *F1000Research* 5.0 (2016), p. 591. ISSN: 2046-1402. DOI: 10.12688/f1000research.8357.2.
- [68] Jiawei Han, Micheline Kamber and Jian Pei. *Data Mining: Concepts and Techniques*. Elsevier, 2012. ISBN: 978-0-12-381479-1. DOI: 10.1016/B978-0-12-381479-1.00001-0.
- [69] Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. New York, NY: Springer New York, 2009. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7.
- [70] Peter Willett. ‘Similarity-based virtual screening using 2D fingerprints’. In: *Drug Discovery Today* 11.23-24 (Dec. 2006), pp. 1046–1053. ISSN: 13596446. DOI: 10.1016/j.drudis.2006.10.005.

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

- [71] Martin Vogt et al. 'Scaffold Hopping Using Two-Dimensional Fingerprints: True Potential, Black Magic, or a Hopeless Endeavor? Guidelines for Virtual Screening'. In: *Journal of Medicinal Chemistry* 53.15 (Aug. 2010), pp. 5707–5715. ISSN: 0022-2623. DOI: 10.1021/jm100492z.
- [72] P Willett. 'Similarity-based approaches to virtual screening.' In: *Biochemical Society transactions* 31.Pt 3 (2003), pp. 603–606. ISSN: 0300-5127. DOI: 10.1042/.
- [73] Peixun Liu and Wei Long. 'Current mathematical methods used in QSAR/QSPR studies'. In: *International Journal of Molecular Sciences* 10.5 (2009), pp. 1978–1998. ISSN: 14220067. DOI: 10.3390/ijms10051978.
- [74] Angélica Nakagawa Lima et al. 'Use of machine learning approaches for novel drug discovery.' In: *Expert opinion on drug discovery* 11.3 (2016), pp. 225–239. ISSN: 1746-045X. DOI: 10.1517/17460441.2016.1146250.
- [75] Leo Breiman. 'Random Forests'. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.
- [76] Corinna Cortes and Vladimir Vapnik. 'Support-Vector Networks'. In: *Machine Learning* 20.3 (1995), pp. 273–297. ISSN: 15730565. DOI: 10.1023/A:1022627411411.
- [77] Ana L. Teixeira, João P. Leal and Andre O. Falcao. 'Random forests for feature selection in QSPR models - An application for predicting standard enthalpy of formation of hydrocarbons'. In: *Journal of Cheminformatics* 5.2 (2013), p. 1. ISSN: 17582946. DOI: 10.1186/1758-2946-5-9.
- [78] A Statnikov, L Wang and C Aliferis. 'A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification'. In: *BMC Bioinformatics* 9.1 (2008), p. 319. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-319.

REFERENCES

- [79] Liew Chin Yee and Yap Chun Wei. 'Current Modeling Methods Used in QSAR/QSPR'. In: *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, Mar. 2012, pp. 1–31. ISBN: 9783527324347. DOI: 10.1002/9783527645121.ch1.
- [80] Alexandre Varnek and Igor Baskin. 'Machine Learning Methods for Property Prediction in Chemoinformatics'. In: *Journal of Chemical Information and Modeling* 52.6 (2012), pp. 1413–1437. ISSN: 1549-9596. DOI: 10.1021/ci200409x.
- [81] J C Gertrudes et al. 'Machine learning techniques and drug design.' In: *Current medicinal chemistry* 19.25 (2012), pp. 4289–97. ISSN: 1875-533X. DOI: 10.2174/092986712802884259.
- [82] Dimitar Dobchev, Girinath Pillai and Mati Karelson. 'In silico machine learning methods in drug development'. In: *Current Topics in Medicinal Chemistry* 14.16 (2014), pp. 1913–1922. ISSN: 15680266. DOI: 10.2174/1568026614666140929124203.
- [83] Maykel Pérez González et al. 'Variable selection methods in QSAR: an overview.' In: *Current topics in medicinal chemistry* 8.18 (2008), pp. 1606–1627. ISSN: 15680266. DOI: 10.2174/156802608786786552.
- [84] Matthias Dehmer et al. *Statistical Modelling of Molecular Descriptors in QSAR / QSPR*. February. Weinheim, Germany: Wiley-VCH Verlag GmbH, 2012, p. 32434. ISBN: 9783527324347.
- [85] Robin Genuer, Jean-michel Poggi and Christine Tuleau-malot. 'Variable selection using Random Forests'. In: *Pattern Recognition Letters* 31.14 (2012), pp. 2225–2236. DOI: <https://doi.org/10.1016/j.patrec.2010.03.014>.
- [86] Mohammed; J. Zaki and Wagner; Meira. *Data Mining and Analysis: fundamental concepts and algorithms*. New York, NY, USA: Cambridge University Press, 2014. ISBN: 9780521766333.

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

- [87] John Aldo Lee and Michel Verleysen. *Nonlinear Dimensionality Reduction*. Ed. by John A. Lee and Michel Verleysen. Information Science and Statistics. New York, NY: Springer New York, 2007. ISBN: 978-0-387-39350-6. DOI: 10.1007/978-0-387-39351-3.
- [88] Lennart Eriksson et al. ‘Megavariate analysis of environmental QSAR data. Part I – A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD)’. In: *Molecular Diversity* 10.2 (July 2006), pp. 169–186. ISSN: 1381-1991. DOI: 10.1007/s11030-006-9024-6.
- [89] Paola Gramatica. ‘Principles of QSAR models validation: Internal and external’. In: *QSAR and Combinatorial Science* 26.5 (2007), pp. 694–701. ISSN: 1611020X. DOI: 10.1002/qsar.200610151.
- [90] Alan R. Katritzky et al. ‘Interpretation of Quantitative Structure-Property and -Activity Relationships’. In: *Journal of Chemical Information and Computer Sciences* 41.3 (May 2001), pp. 679–685. ISSN: 0095-2338. DOI: 10.1021/ci000134w.
- [91] Robin Genuer, J-M Poggi and C Tuleau. ‘Random Forests : some methodological insights’. In: *Inria* 6729 (2008), p. 32.
- [92] Gérard Biau. ‘Analysis of a Random Forests Model’. In: *Journal of Machine Learning Research* 13 (2012), pp. 1063–1095. ISSN: 1532-4435.
- [93] Andrej-Nikolai Spiess and Natalie Neumeyer. ‘An evaluation of R² as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach.’ In: *BMC pharmacology* 10 (2010), p. 6. ISSN: 1471-2210. DOI: 10.1186/1471-2210-10-6.
- [94] Christoph Steinbeck et al. ‘The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics’. In: *Journal of Chemical Information and*

REFERENCES

- Computer Sciences* 43.2 (2003), pp. 493–500. ISSN: 00952338. DOI: 10.1021/ci025584y.
- [95] Michael R. Berthold et al. ‘KNIME - The Konstanz Information Miner’. In: *SIGKDD Explorations* 11.1 (2009), pp. 26–31. ISSN: 19310145. DOI: 10.1145/1656274.1656280.
- [96] R R Development Core Team. *R: A Language and Environment for Statistical Computing*. 2011. DOI: 10.1007/978-3-540-74686-7.
- [97] David Meyer et al. *Misc functions of the Department of Statistics (e1071), TU Wien*. 2014. DOI: citeulike-article-id:9958545.
- [98] Andy Liaw and Matthew Wiener. ‘Classification and Regression by randomForest’. In: *R News* 2.3 (2002), pp. 18–22.
- [99] Alboukadel Kassambara and F Mundt. *Package ‘factoextra’ for R: Extract and Visualize the Results of Multivariate Data Analyses*. 2017.
- [100] Jaroslaw Polanski et al. ‘Modeling robust QSAR’. In: *Journal of Chemical Information and Modeling* 46.6 (2006), pp. 2310–2318. ISSN: 15499596. DOI: 10.1021/ci050314b.
- [101] D. Fourches, E. Muratov and a. Tropsha. ‘Trust but verify: on the importance of chemical structure curation in chemoinformatics and QSAR modeling research’. In: *J. Chem. Inf. Model.* 50.7 (2010), pp. 1189–1204.
- [102] Denis Fourches and Alexander Tropsha. ‘Using graph indices for the analysis and comparison of chemical datasets’. In: *Molecular Informatics* 32.9-10 (2013), pp. 827–842. ISSN: 18681751. DOI: 10.1002/minf.201300076.
- [103] Douglas Young et al. ‘Are the chemical structures in your QSAR correct?’ In: *QSAR and Combinatorial Science* 27.11-12 (2008), pp. 1337–1345. ISSN: 1611020X. DOI: 10.1002/qsar.200810084.

4. ANALYSIS AND COMPARISON OF VECTOR SPACE AND METRIC SPACE REPRESENTATIONS IN QSAR MODELING

- [104] Alexander Golbraikh et al. 'Data set modelability by QSAR'. In: *Journal of Chemical Information and Modeling* 54.1 (2014), pp. 1–4. ISSN: 15499596. DOI: 10.1021/ci400572x.
- [105] Alexander Golbraikh et al. *Modelability Criteria: Statistical Characteristics Estimating Feasibility to Build Predictive QSAR Models for a Dataset*. Ed. by Jerzy Leszczynski and Manoj K. Shukla. Boston, MA: Springer US, 2014, pp. 187–230. ISBN: 978-1-4899-7445-7. DOI: 10.1007/978-1-4899-7445-7_7.
- [106] Gilles Marcou, Dragos Horvath and Alexandre Varnek. 'Kernel Target Alignment Parameter: A New Modelability Measure for Regression Tasks'. In: *Journal of Chemical Information and Modeling* 56.1 (2016), pp. 6–11. ISSN: 15205142. DOI: 10.1021/acs.jcim.5b00539.
- [107] Myles Hollander, Douglas A. Wolfe and Eric Chicken. *Nonparametric Statistical Methods*. Third Edit. Wiley Series in Probability and Statistics. Wiley, July 2015. ISBN: 9780470387375. DOI: 10.1002/9781119196037.
- [108] Felipe de Mendiburu. *Agricolae: Statistical Procedures for Agricultural Research. R Package Version 1.2-8*. 2017.
- [109] Igor V. Tetko et al. 'Critical assessment of QSAR models of environmental toxicity against tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection'. In: *Journal of Chemical Information and Modeling* 48.9 (2008), pp. 1733–1746. ISSN: 15499596. DOI: 10.1021/ci800151m.
- [110] Hao Zhu et al. 'Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis'. In: *Journal of Chemical Information and Modeling* 48.4 (2008), pp. 766–784. ISSN: 15499596. DOI: 10.1021/ci700443v.

5

A visual approach for analysis and inference of molecular activity spaces

SAMINA KAUSAR AND ANDRE O FALCAO

Abstract

Background: Molecular space visualization can help to explore the diversity of large heterogeneous chemical data, which ultimately may increase the understanding of structure-activity relationships (SAR) in drug discovery projects. Visual SAR analysis can therefore be useful for library design, chemical classification for their biological evaluation and virtual screening for the selection of compounds for synthesis or in vitro testing. As such, computational approaches for molecular space visualization have become an important issue in cheminformatics research. The proposed approach uses molecular similarity as the sole input for computing a probabilistic surface of molecular activity (PSMA). This similarity matrix

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

is transformed in 2D using different dimension reduction algorithms (Principal Coordinates Analysis (PCoA), Kruskal multidimensional scaling, Sammon mapping and t-SNE). From this projection, a kernel density function is applied to compute the probability of activity for each coordinate in the new projected space.

Results: This methodology was tested over four different quantitative structure-activity relationship (QSAR) binary classification data sets and the PSMA were computed for each. The generated maps showed internal consistency with active molecules grouped together for all data sets and all dimensionality reduction algorithms. To validate the quality of the generated maps, the 2D coordinates of test molecules were computed into the new reference space using a data transformation matrix. In total sixteen PSMA were built, and their performance was assessed using the Area Under Curve (AUC) and the Matthews Coefficient Correlation (MCC). For the best projections for each data set, AUC testing results ranged from 0.87 to 0.98 and the MCC scores ranged from 0.33 to 0.77, suggesting this methodology can validly capture the complexities of the molecular activity space. All four mapping functions provided generally good results yet the overall performance of PCoA and t-SNE was slightly better than Sammon mapping and Kruskal multidimensional scaling.

Conclusions: Our result showed that by using an appropriate combination of metric space representation and dimensionality reduction applied over metric spaces it is possible to produce a visual PSMA for which its consistency has been validated by using this map as a classification model. The produced maps can be used as prediction tools as it is simple to project any molecule into this new reference space as long as the similarities to the molecules used to compute the initial similarity matrix can be computed.

Keywords: Structure Activity Relationship (SAR; Molecular/chemical space; Two dimensional kernel density estimation; Noncontiguous Atom Matching Structural Similarity Function (NAMS; t-SNE; PCoA; non-metric MDS; Sammon mapping

5.1 Introduction

Chemical/molecular space reflects high dimensional conceptual spaces that describe the structural diversity of all possible potential pharmacologically active molecules. The size of molecular space is not well defined, yet a fraction of it ranging from thousands to millions of compounds is stored in small molecule databases. Consequently, a part of the huge molecular space is mainly focused to explore the complexity of a relevant small set of chemical structures in many different problems during drug design [1, 2, 3]. Nonetheless, molecular space interactive analysis and visualization can serve as a strong tool to explore the diversity of millions of compounds stored in public databases and can increase the performance of drug discovery process. For example, nearest neighbour searches in various defined property regions in molecular space (activity space map) can identify interesting similar molecules (potent analogues) with similar properties [1, 2, 4, 5].

Molecular space visualization methods require that molecules are projected into a reduced set of dimensions (most of the times, two or three) in such a way that the relative distances between molecules are better preserved in this new projected space. As distances should be preserved, molecules with similar activity profiles should appear clustered together. [1, 6]. Thus, molecular space visual analysis combines the concept of molecular structure and activity similarity [7, 6]. Since molecular dis/similarity is defined through pairwise distances between projected molecules in reference space, an appropriate choice of a molecular metric space (spatial) representation is crucial for reliable application of molecular spacial analysis. A molecule in metric space is defined as a set of distances computed from the similarity between that molecule to all the other molecules in a given chemical data set. For this purpose, many methods are available in literature to compute dis/similarity. A variety of methods uses either molecular descriptors or fingerprints, which represent different physico-chemical or structural characteristics [8, 9, 10, 11, 12, 13, 14, 15, 16]. These approaches entail that each molecule is initially reduced into a vector space by computing

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

a set of attributes, that can be used to infer distance, yet this is not always required as other independent approaches like molecular graph matching approaches can also be used for a direct assessment of structural similarity [17, 18, 19, 20].

In metric space representation, a set of M molecules is represented in M dimensions, as the distance to all the other elements of the set (including itself) must be present. As such, the visualization of this M -dimensional metric space in reduced spatial dimensionality is a challenge in data diversity analysis [7, 21, 22]. To address this issue many linear and non-linear approaches have been developed to reduce the dimensionality and complexity of molecular space [1, 21, 6, 23]. In all dimension reduction (DR) methods, the most important characteristic is the optimization of the criterion that guides dimensionality reduction. Since the concept of DR is mainly based on data geometrical representation where data is interpreted as discrete points/objects, the main objective to explore or analyze such geometrical spaces is to discover the relationships between the points within this complex structure of data (manifold) [22]. The main criterion that needs to be optimized in DR algorithms for metric space data is the approximation of the original intermolecular distances (proximity relationships) in the new projection space; DR approaches that are based on optimization of this dimensionality reduction criterion in a linear/non-linear way are collectively referred as distance-preserving approaches [22]. Principal component analysis (PCA) [24], is by far the most common method [1, 25, 26] used in DR, yet it does not fall into this category, as the main purpose of PCA is to represent in less dimensions the linear components that maximize the data variance, not necessarily preserving the distances between data. On the other hand Principal Coordinates Analysis (PCoA) [27, 28], Sammon mapping [29], self-organizing maps [30], stochastic neighbor embedding [31] or stochastic proximity embedding [32], to name but a few are distance-preserving DR algorithms and been used in cheminformatics [23, 33]. Most of the times, non-linear methods are usually preferred because linear algorithms may be limited to linear projection functions and therefore may not adequately handle complex associations that may be present in such problems [22].

Distance preserving DR methods can then make it possible to project molecules into a 2D reduced molecular space while preserving the original proximity (distances) of molecules as best as possible, assuming that there is always going to be a loss of information as the original molecular space should have a much higher dimensionality. To establish a structure-activity relationship, molecular activity surface maps mostly referred as “activity landscapes” are generated from 2D projected space (reference space) of molecules by adding a property of each molecule as a third dimension[6, 34, 23]. In such projections, the activity of molecules added as third dimension in the projected molecular space is the basis for fitting a generated surface that represents the activity magnitude. Since data is largely scattered in projected space, an interpolation algorithm [35, 33] is required to make a coherent surface onto this 2D projected map. Ideally, structurally similar molecules should appear grouped together in well-separated clusters and each group should have similar properties. This property may not always hold, and that is the case of “activity cliffs”, projected regions that exhibit similar molecules with largely varying activity very close together[35, 36, 9, 16]. Despite these challenges, such analysis may provide a global picture of the spatial characteristics of a given data set.

The descriptive and predictive accuracy of molecular space visualization approaches largely depends upon three main issues, including a) a choice of a molecular space representation, b) the accuracy of DR methods and c) the performance of the interpolation algorithm to generate well estimated activity surface from sparsely projected molecules. To this end, in our approach for visual characterization of molecules in conceptual spaces, a reliable pipeline is generated that can efficiently be used to build a probabilistic surface of molecular activity (PSMA), which can help to understand SAR in different situations. We have thus integrated the advantages of the following different methods in the proposed molecular space mapping approach:

- Choice of molecular space representation: Molecular pairwise similarity was quanti-

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

fied using a graph matching algorithm: The Non-contiguous Atom Matching Structural similarity (NAMS) [17]. This algorithm has a high discriminative power for very similar molecules over other structural or graph matching approaches. However, any other similarity computation method can be used.

- DR methods: We applied four non-linear DR methods including Principal Coordinates Analysis (PCoA) [27], Kruskal Multidimensional Scaling (KMDS) [28], Sammon mapping (SM), [29] and t-Distributed Stochastic Neighbor Embedding (t-SNE) [37].
- PSMA: Non-parametric 2D kernel density estimation (KDE) function [38] created within a Bayesian framework was used to map the most likely activity regions (activity surfaces) from sparsely distributed active and inactive compounds.

This approach is, to our knowledge, new and allows building a non-parametric model out of raw similarity data, which is useful for visualization and has clear predictive properties. Furthermore, t-SNE applications in molecular space diversity analysis are not a common practice in cheminformatics. A survey of recent literature showed only one work to visualize molecular space using this algorithm [39]. However, under this particular domain, this is a first effort to build activity spatial classification model using this algorithm by comparing its performance with other commonly used tools. Another novel point the present approach tried to address was the use of 2 dimensional KDE for model making. KDE is considered a powerful tool in statistics for truthful assessment of data actual distribution/characteristics [38]. In cheminformatics literature KDE has been used as a robust method to define applicability domains of quantitative structure-activity relationship (QSAR) predictive models [40, 41, 42]. Applicability domains are used to define a boundaries in molecular space within which new predictions of QSAR models are considered reliable [43]. We extended the same concept to computing probability density function for active and inactive molecules within 2D projected space and surface was generated from the 2D map containing high promising regions of active molecules. In the presented methodology, integration of KDE in SAR spa-

tial visualization is a new addition in the efforts of molecular space analysis. It must be made clear that, despite the fact that we are using a Bayesian approach to compute the PSMA, our method has no relation to any naive-Bayes implementation, as we are computing the full 2D probability map and not the individual probability distribution functions of each coordinate axes as is the case in the naive-Bayes algorithm. The complexity of the resulting maps clearly show that a naive-Bayes approach would be inadequate for this type of modeling.

5.2 Methodology

5.2.1 Overview of the methodology

The basic idea of this study is to capture the measured molecular distances according to any proven method and try to represent those molecules in a reduced reference space for analysis and visualization. Many dimensionality reduction methods are extant, [21, 22] and some of the more popular are PCoA, KMDS, SM, and t-SNE [39, 23]. The procedure to create a PSMA can summarily be described in the following steps. First, a full similarity matrix of a molecular data set is computed. Secondly, similarities are transformed into distances and projected into a 2-Dimensional (2D) space using one of the above mapping functions. Finally, the probabilities of this reduced space are computed using a 2D KDE function within a Bayesian perspective [44] to produce a probability map of a projected molecule for all classes. The generated 2D probability map should show the density distribution of training data by mapping the locations of the most likely activity regions of the projected molecular space. Such interactive class probability topographical map (PSMA) can serve as classification model. To project new molecules into the new reference space, a data transformation matrix can be used for embedding test molecules in the reference activity space. To classify each new projected molecule the generated PSMA is used to calculate their prob-

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

ability of belonging to either class. Models performance was assessed using test molecules predictions (Figure 5.1).

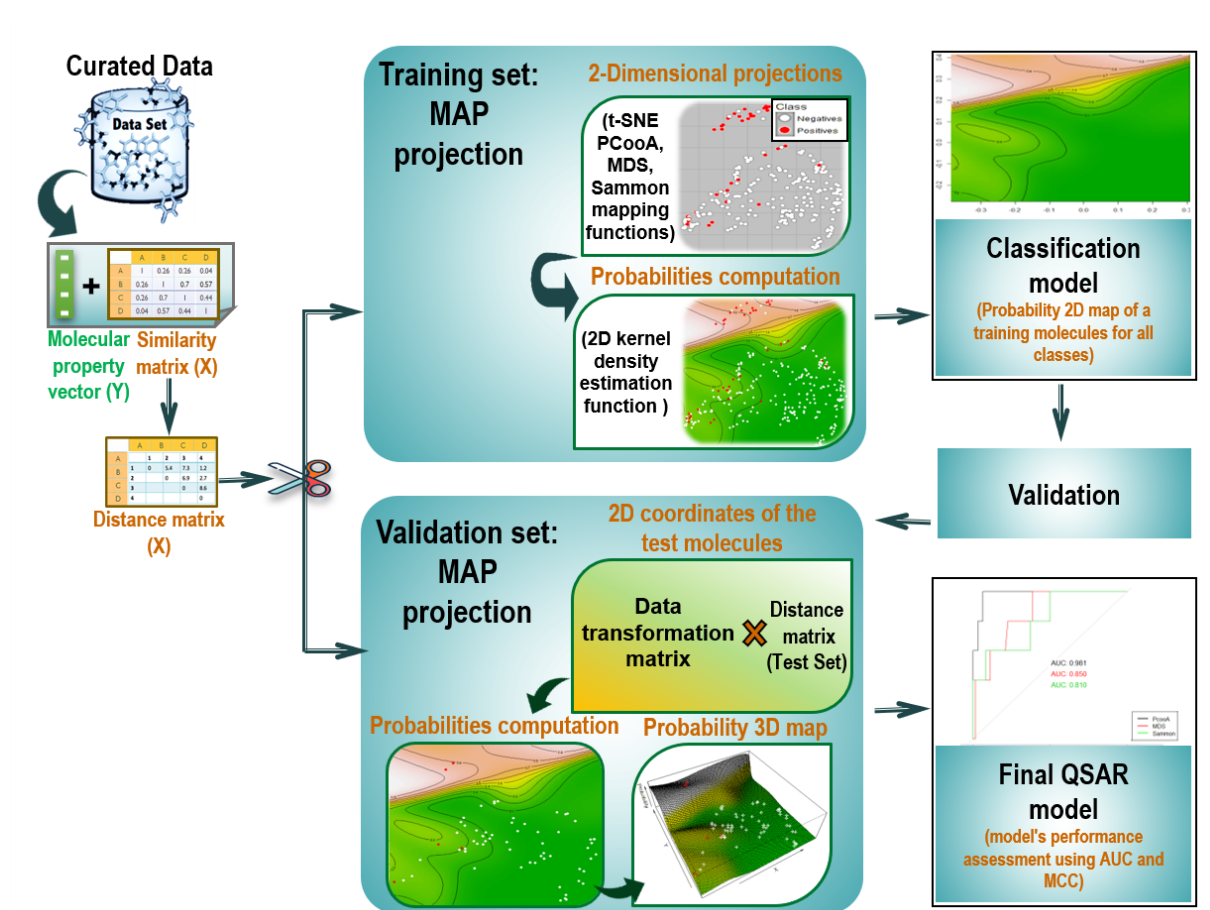


Figure 5.1: Overview of the methodology.

5.2.2 Molecular dis/similarity quantification

Chemical space analysis based on nearest neighbour searches in which molecular similarity analysis is a central task that is based on *Similar Property Principle* [7]. According to this similarity principle, globally similar compounds should have similar properties [9].

Since intermolecular distances between projected molecules are a measure of their molecular similarity or dissimilarity, its quantification must be robust for meaningful spatial/metric space representations, so that they may be able to map similar compounds in contiguous regions, a fundamental aspect for reliable property prediction [7, 6].

For similarity quantification, molecules are translated into numeric data using various molecular representations including structural descriptors and molecular fingerprints [45, 46, 47]. Molecular descriptors contain information of structural relevant features of molecules at different levels including constitutional (1D), topological (2D), geometrical (3D) and physico-chemical properties-based (4D) [45, 46]. Molecular fingerprints encode molecular structural information in a bit-string where each bit represents the presence (1) or absence (0) of a structural feature (e.g., chemical substructure, sub-graph, or 2D or 3D pharmacophore). 2D fingerprints are commonly used molecular representations for dis/similarity quantification because comparing bit-string is fast and easy [48, 49, 14, 16, 49, 50].

There are some conventional distance metrics like Euclidean, Hamming, Manhattan distance that measure the distance between compounds represented by using descriptors/fingerprints [7, 51]. Some other similarity coefficients are available for binary data (e.g. Tanimoto, Sorensen-Dice, cosine or Tversky) [52, 53]. Of those, the Tanimoto coefficient (T_c) is extensively applied in literature to compute similarity between molecules using molecular fingerprints [54, 51]. T_c compares two fingerprints and counts the number of on-bits (1) common in both with respect to the total number of on-bits (1) in each fingerprint. There are several other approaches that assess similarity using different algorithms based on superposition, molecular graph representations [55] histogram comparisons [56, 57] and Brownian processing of molecules [58]. In this study, we used for molecular similarity assessment a graph matching algorithm, the Non-contiguous Atom Matching Structural similarity (NAMS) [17]. This algorithm uses an atom alignment method to adequately quantify the structural similarity and has a high discriminative power for very similar molecules over other structural or graph matching approaches. NAMS breaks complex molecular structures

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

into simpler parts to reduce molecule to atoms and calculates global structural similarity score from the best alignment between the atoms of compared molecules. NAMS follows an atom matching methodology, which is able to consider the important characteristics of the atoms and bonds such as the chirality and the double bond stereoisomerism. These features are usually ignored in other approaches.

5.2.3 From similarity to distance

As stated above, since molecular similarity is measured by a distance between a pair of molecules in the chosen reference space, a distance function known as metric is mainly required to calculate distances between molecules in metric space representation. A dissimilarity function or a distance function $d(x, y)$ between the instances x and y must satisfy the following three basic properties:

$$\text{(Property 1) } d(x, x) = 0$$

$$\text{(Property 2) } d(x, y) \geq 0$$

$$\text{(Property 3) } d(x, z) \leq d(x, y) + d(y, z)$$

Which essentially state that a distance between an instance and itself should always be zero, any distance between any instances should never be negative and that the distance between 2 points should respect the triangle inequality. A function that transforms similarity into distance should accordingly be monotonically decreasing and intersect the X-axis precisely at $x = 1$. Using these principles similarities and distances can be inter-converted i.e. every similarity metric correspond to a distance metric and vice versa. If similarity function $s(x, y)$ is normalized $0 \leq s(x, y) \leq 1$ and $s(x, x) = 1$ for all $x, y \in X$ then similarity matrix can be transform into distance matrix with a simple distance functions (see Eq. 5.1 [59] and

5.2 [60])

$$d(x, y) = 1 - s(x, y) \quad (5.1)$$

$$s(x, y) = \frac{1}{1 + d(x, y)} \iff d(x, y) = \frac{1}{s(x, y)} - 1 \quad (5.2)$$

Other complying transformations can also be applied like the negative of the natural logarithm (Eq. 5.2)[33].

$$d(x, y) = -\ln(s(x, y)) \quad (5.3)$$

These last two eqs. show the property that similarity values of zero imply an infinite distance, so, for those extreme values, some clamping to a maximum distance may be necessary.

Within the molecular space a distance function should be modulated to set a particular meaning out of similarity measures. It can be readily observed that the last two functions appear concave (Figure 5.2), meaning that near the regions that have the lowest similarity, the impact on the resulting distance is the highest, which is counter-intuitive, as typically the conservation of activity for similar molecules is only verified at the highest levels of structural similarity. Such transformation functions may further increase the projection distortion,

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

as most algorithms will tend to minimize the error between the projected distances and the actual distances. A convex curve may solve this problem, by inflating the distances of very similar molecules but, on the other hand, if two molecules are very unrelated, the impact on the transformed distance will appear small. As such we propose the use of the following transformation which uses a parameter k that controls the convexity.

$$d(x, y) = 1 - \frac{k \times s(x, y)}{1 + k - s(x, y)} \quad (5.4)$$

In Eq. 5.4, small positive values of k entail extremely convex functions, while on the other hand, very high values approach $d(x, y) = 1 - s(x, y)$ (Figure 5.2). Empirically and visually we have determined that values of k ranging from 0.3 to 0.5 provide not too abrupt transitions, and a value of 0.382 was used in all problems ($0.382 \approx \phi - 1$, where ϕ is the Golden Ratio)

5.2.4 Dimensionality reduction

As stated, the visualization of metric space data is a difficult challenge in many different domains of data analysis, as it demands efficient and robust techniques to adequately represent in 2 or 3 dimensions the data variability present in an intrinsic multidimensional problem [21, 22]. The objective of metric space visualization is to generate a topographical map, which should be able to present a visual characterization of molecules by grouping them together on the basis of their structural similarity. As referred, a metric space is an $M \times M$ dimensional distance matrix where M compounds are represented each by M intermolecular distances. However, it is not trivial to graph the diversity of such high dimensional metric space. As referred, the main objective of dimensionality reduction (DR) in metric spaces is

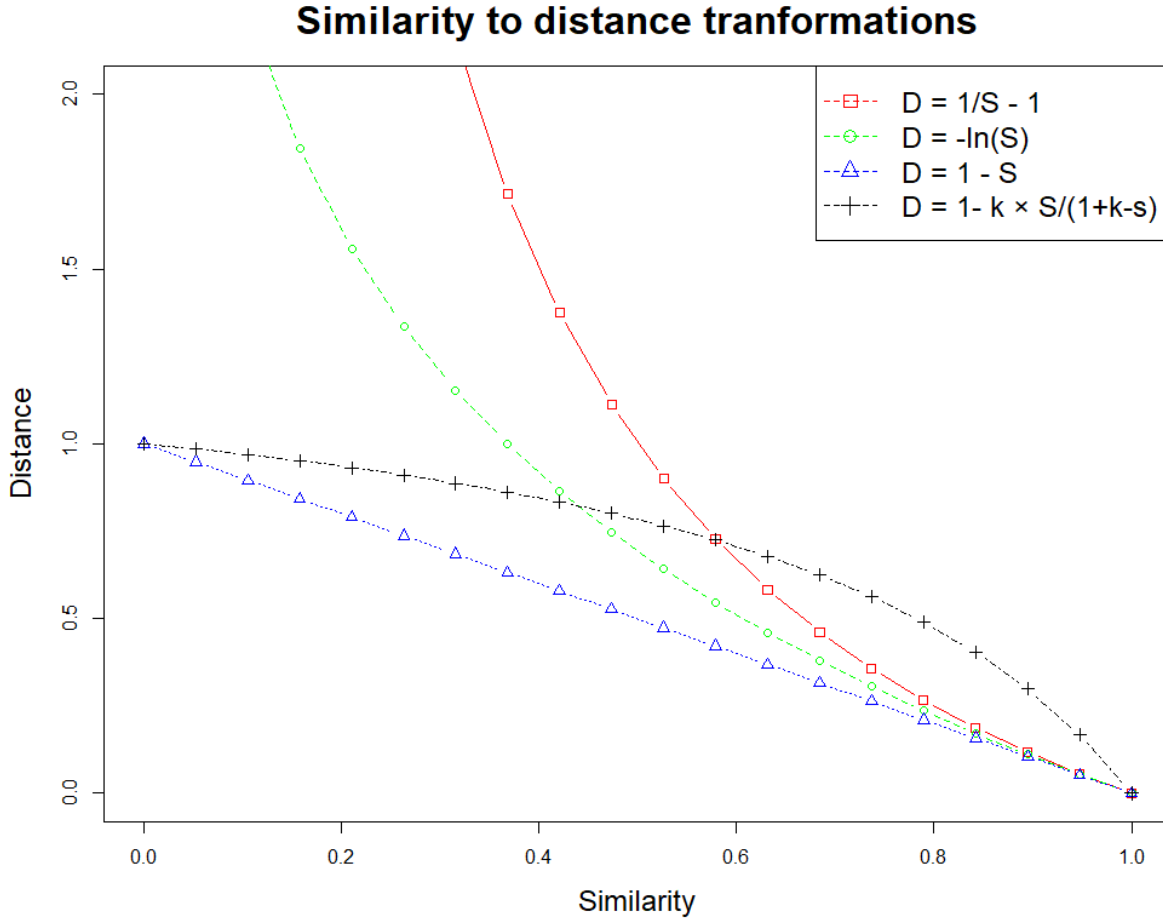


Figure 5.2: Distance functions for similarity to distance transformations.

the distance-preservation in original high dimensional space to reduce dimensions. These transformations can be linear or non-linear. The distance-preservation criteria is that any manifold complex geometrical structure of data can be projected into reduced number of dimensions, and the quality of such transformation can be measured by the difference between the original and the projected distances in the new space. A large number of nonlinear DR approaches are available that aim to preserve the local structure of data [21, 22]. In this work we used four of the most widely used DR distance-preserving techniques, namely, Principal Coordinates Analysis (PCoA), Kruskal Multidimensional Scaling (KMDS), Sammon

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

mapping (SM), and t-Distributed Stochastic Neighbor Embedding (t-SNE), for reducing the molecules' distance matrix in 2D and allow the visualization of the data. After DR, the newly projected instances were divided according to their activity class and a probability function assigned by using a kernel density function to each element of each class.

5.2.4.1 Principal Coordinates Analysis (PCoA)

Principal Coordinates Analysis (PCoA) [27] also known as metric multidimensional scaling (MDS). PCoA relies on a simple generative model possessing all the advantages and drawbacks of Principal components analysis, although its goal is to preserve distances, while PCA aims at preserving the data variance. However, differently from PCA which generally is performed by computing the eigenstructure of the covariance matrix of the data, in PCoA, the basic input is the distance matrix. To the squared of the distance matrix, each element is double centered and, to the resulting matrix, an eigendecomposition is performed. The eigenscaled coordinates of the first N eigenvectors are the projected coordinates resulting from this transformation. There are no tuning parameters for PCoA, however, results may vary depending on the distance function used for data metric space representation. It is important to notice that many implementations of KMDS or SM use the results of PCoA transformation as a starting point

5.2.4.2 Kruskal Multidimensional Scaling (KMDS)

Non-metric multidimensional scaling was developed by Kruskal [28] for resolving problems related to the linear multidimensional scaling algorithms, like PCoA. The KMDS is based on numerical optimization methods. This method uses ordinal information (i.e., proximity ranks) and then calculates the scaled proximities using monotonic transformation to determine the high-dimensional structure of data set. Finally, to visualize data in low di-

mensional features space, KMDS finds the best possible projections with minimum squared differences between the initial distances and the scaled ranking of the distances. Thus, in contrast to PCoA, KMDS does not attempt to directly preserve distances between the data points in the initial space but rather its order, or ranking, of the distances between objects [22]. KMDS optimizes the following stress function or error function (Eq. 5.5) to estimate the preservation of the pairwise distances (goodness of fit).

$$\text{Kruskal's stress} = \sqrt{\frac{\sum_{i,j} (d(i,j) - \hat{d}(i,j))^2}{\sum_{i,j} d^2(i,j)}} \quad (5.5)$$

where $d(i,j)$ are the collected proximities and $\hat{d}(i,j)$ is the distance measured between the i^{th} and j^{th} objects in low-dimensional representations

5.2.4.3 Sammon mapping

In 1969 Sammon [29] developed a non-linear variant of MDS, which is referred as Sammon mapping, Sammon's nonlinear mapping and NLM (Non-Linear Mapping). The word "mapping" used to represent the main objective of the method, which was to establish a mapping between a high-dimensional metric space and a lower-dimensional feature space. But, to some extent the 'mapping' word is misleading as it does not exactly generate a continuous mapping between these two spatial representations. The main goal of Sammon's algorithm is a dimensionality reduction of a finite set of objects/points by following the same basic principle of MDS algorithm. The main modification is its efficient optimization technique to minimize the Sammon's stress function (Eq. 5.6) by calculating its normalized value by the initial space distances. Sammon's algorithm does not require any parameter optimization,

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

but results may vary depending on the chosen different dissimilarity measures.

$$\text{Sammon's stress} = \frac{1}{\sum_{i < j} d(i, j)} \sum_{i < j} \frac{(d(i, j) - \hat{d}(i, j))^2}{d(i, j)} \quad (5.6)$$

where $d(i, j)$ are original distances and $\hat{d}(i, j)$ are distances between the i^{th} and j^{th} objects in reduced space

5.2.4.4 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-Distributed Stochastic Neighbor Embedding (t-SNE) [37] is a variant of Stochastic Neighbor Embedding (SNE) and was developed to solve two basic problems of the SNE algorithm including difficult optimization of a cost function and a problem referred as “crowding problem”. The main objective of both methods SNE and t-SNE is similar to MDS to project objects in reduced space, such that the pairwise distances between projected objects reflect the original distances between objects as good as possible; although this distance preservation is achieved in a non-linear way. t-SNE algorithm focuses on local data structures, to generate well-separated clusters. One of the key characteristics of this method is that the new distances of objects in the reduced feature space are determined probabilistically with close objects having a much higher probability of staying together in the new space than distant objects. In contrast to SNE, t-SNE does not compute Gaussian “induced” probabilities between each pair of points in embedded space instead it uses a heavy-tailed Student’s t-distribution for the same purpose to avoid projection of points at the same place (crowding effect). This method consequently allows efficient visualization of moderate distances in the initial space by larger distances in graphical configuration of projected space. Differently from the other methods, t-SNE is a probabilistic approach, thus different runs may produce different maps

5.2.5 Probabilities density estimation

The probability density function (PDF) is an informal way to explore and analyse the properties of any given quantitative variable. The PDF gives a natural description of the distribution of any random variable by specifying its probability for all values of its range. Since robust estimation of the probability density can be used to solve regression and classification problems, PDF is a fundamental concept in data analysis [38, 44]

The PDF for any given variable can be estimated using either parametric methods that assume the density function has a standard distribution function. As an example, if we assume a continuous variable has a normal distribution, then it is possible to compute the full PDF of this variable if the mean and the variance of the data are known and confidently mirror that of the original population. Non-parametric methods, on the other hand, are free of any assumptions and estimate probability density solely from data. One of the most common methods in one dimensional variables is to use a gaussian kernel function applied to each observation, and using the scaled sum of each kernel for each point within the defined range of the data. Non-parametric PDF estimation is an extensive research area in field of data exploration[38, 44]. Most of the existing techniques focus on low dimensional densities estimation (1 to 3D) because uni/bivariate PDF is relatively easy; however investigating PDF of data in higher dimensions (multivariate) is difficult and computational expensive. In lower dimensions histograms can be constructed that generate a non smooth representation of the PDF. But for smooth PDF estimation, the usage of kernel density estimation (KDE) is a common method, used in visual data exploraration [38, 44]. The multivariate KDE algorithm has been introduced to deal with high dimensional data with improved accuracy and speed [61, 44]. In our analysis we have used a bivariate KDE applied to the chemical data projected in 2D.

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

5.2.5.1 2D kernel density estimation

A kernel density estimation function generates an actual distribution of the data by calculating the probability of each data point in the given data without using any reference point [40] or prior assumptions. Kernel probability density function computes the PDF of the projected 2D space by summing up M-dimensional kernels placed on every projected coordinate. The basic kernel estimator can be expressed as

$$\hat{f}(x) = \frac{1}{nh} \sum_{j=1}^n K\left(\frac{x - x_j}{h}\right) \quad (5.7)$$

where \mathbf{K} is a fixed kernel and h is the calculated bandwidth for sample x_1, \dots, x_n . Commonly available kernel functions are Gaussian (normal), uniform, cosine, triangle, Epanechnikov, quartic (biweight), and tricube (triweight). The bandwidth, h , is a smoothing parameter that influences the width of PDF estimates. Choosing a bandwidth is a compromise between very smooth estimates (large h values) to remove insignificant bumps and wiggly estimates to find out real peaks (small h values). In this study, we applied a two-dimensional KDE with a Gaussian kernel [44] to calculate densities in the two-dimensional reduced space. It is defined as

$$f(x, y) = \frac{\sum_s \phi\left(\frac{(x-x_s)}{h_x}\right) \phi\left(\frac{(y-y_s)}{h_y}\right)}{nh_x h_y} \quad (5.8)$$

For determining the bandwidth (h_n), we used Silverman's heuristic approach [38] (h_n) for the Gaussian kernel function (Eq. 5.6) [44].

$$h_n \approx 1.06 \min \left(\hat{\sigma}, \frac{R}{1.34} \right) M^{-\frac{1}{5}} \quad (5.9)$$

where $\hat{\sigma}$ is the standard deviation of the reference coordinate, R , the difference between the 2nd and 3rd quartile and M the number of projected points.

In QSAR modelling, KDE is usually explored as an interpolation method to define the applicability domain of generated classification models [43]. Among the most widely used multivariate (high dimensional metric space) interpolation approaches (e.g., range-based, distance-based, geometrical), KDE is considered as one of the more advanced and accurate methods for calculating the applicability domain [40, 41, 42]. However, to the best of our knowledge, KDE is not used for visualization nor data classification over 2D activity landscapes.

5.2.6 Defining active probability regions

In the available literature, several other methods have been used for data visualization of the molecular space. Yet, in all cases each projected point is associated with its measured activity value and surfaces are generated according to the activity magnitude of each molecule or colour codes are used to differentiate different activities [35, 1, 23, 6]. In all these approaches, along with all referred issues in data visualizing methods most implemented interpolation methods are not adequate as classification tools.

To clearly identify the spatial regions where is a higher probability of finding active

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

compounds, to the 2D projected molecules, training data was divided into two classes of active and inactive molecules according to a predefined activity threshold. For both partitions a kernel density map (KDM) is computed, using a common bandwidth, previously computed with all the data. Each KDM can be seen as a measure of the likelihood of a molecule being a negative or a positive depending on its position on the 2D space, as each KDM is an actual probability function, with an integral summing to one. To compute the posteriors of both KDMs it is necessary to accommodate the data priors. Following Bayes' theorem [62], the posterior probability density (likelihood/probability of a randomly projected new molecule to be in positive class) can be calculated by normalizing the product of the conditional density probability (projected KDM) with the prior probability density of the given partition (positive or negative). Thus to identify whether or not a molecule in (x, y) coordinates being active it is necessary to evaluate each of both Eqs.(5.10 and 5.11)

$$P(M_+ | (x, y)) = \frac{P((x, y) | M_+)P(M_+)}{P((x, y))} \quad (5.10)$$

$$P(M_- | (x, y)) = \frac{P((x, y) | M_-)P(M_-)}{P((x, y))} \quad (5.11)$$

where M_+ and M_- , stand for active (positive) and non-active (negative) molecules. $P((x, y) | M_+)$ is the actual value of the KDM of positive molecules (the likelihood of being positive) and $P(M_+)$ the prior probability of the molecule being active. This illustrates that in the end it should be possible to compute the posterior probability $P(M_+ | (x, y))$. The corresponding meanings stand for Eq. 5.11 that quantify the likelihood, prior and posterior probabilities for the inactive molecules

These observations show that it is possible to compute an activity probability surface using the 2D coordinates of the projected molecules. This surface can therefore be visualized and it should be able to capture the more promising activity regions in the chemical landscape. Furthermore as this surface corresponds to an actual activity posterior map, this visualization tool could be used as a classifier, an actual spatial classification model.

5.2.7 Test set embedding and model validation

The creation of 2D surfaces from the original data will necessarily cause some loss of information. It is thus required to verify if the activity maps constructed are valid in the face of new observations. Therefore to assess model quality, each data set was randomly split into training and test sets. The training set was used to create the model surface and the test set molecules were later embedded, using a linear projection function. A distance transformation matrix (Eq. 5.13) was calculated to transform pairwise distances ($m \times m$) between test set molecules and the training set, into projected coordinates to the given reference molecular space (an $N \times 2$ matrix). If we assume that the original distance matrix is D , to transform each molecule to a new reference coordinates C , we would require a linear transformation T , thus

$$D \cdot T = C \tag{5.12}$$

As D is extant, and C is the result of the dimensionality reduction procedure into a new

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

reference coordinates, we can solve it for T :

$$T = D^{-1} \cdot C \quad (5.13)$$

Where D^{-1} is the inverse of the Distance matrix. T will then be a projection matrix that given the distances of any new molecule to each molecule of the training set, will project it into the new reduced space.

As the projected coordinates of each molecule into the new reference space, it should be easy to compute its activity probability (Eqs. 5.10 and 5.11). As the result is a probability function, the model's performance was assessed using AUC, which measures the entire two-dimensional area underneath the entire receiver operating characteristic (ROC) curve created by plotting the sensitivity/recall/true positive rate (TPR) against the false positive rate (FPR). (Eqs. 5.14 and 5.15).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (5.14)$$

$$FPR = 1 - \text{Specificity} = 1 - \frac{TN}{TN + FP} \quad (5.15)$$

where, for both eqs., TP are the true positives, TN, the true negatives, FP the false positives and FN, the false negatives. For AUC computation, the positive accepting threshold

is changed, and thus the values of these quantities will change accordingly. and provide the data for building the ROC curve.

A second, more stringent criterion is the use of the Matthews Correlation Criterion (MCC) [63], which encompasses the quantities defined above into one statistic that has been widely used for assessing the quality of binary classification models (Eq. 5.16). Differently from the AUC, the MCC will consider as positives only the instances where $P(M_+ | (x, y) > P(M_- | (x, y)))$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5.16)$$

5.3 Data

The designed methodology was tested over four human protein targets (Table 5.1), retrieved from ChEMBL23 [64]. We have looked for data sets for which biological activity was measured as K_i as it quantifies a ligand-receptor interaction based on the equilibrium dissociation constant (K) where smaller value corresponds higher ligand-receptor binding affinities and vice versa [65]. The selected data sets were curated using an automated QSAR modelling workflow [66] and divided into two classes using a cut-off activity value (K_i) to separate highly active molecules ($K_i \leq 10.0$) as positives and less active and non-active molecules ($K_i > 10.0$) as negatives. This resulted in unbalanced data sets with a much larger number of negatives than positives, which is a known characteristic of most problem sets in

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

Table 5.1: Data set description

Target Protein Name	Uniprot ID	Training set		Test set	
		Positives	Negatives	Positives	Negatives
Sigma non-opioid intracellular receptor 1 (SIGMAR1)	Q99720	46	135	10	35
Histamine H1 receptor (HRH1)	P35367	184	783	46	195
Potassium voltage-gated channel subfamily H member 2 (HERG)	Q12809	39	1142	12	283
D(1B) dopamine receptor (DRD5)	P21918	41	231	5	62

QSAR modeling.

5.4 Implementation

All analysis was implemented using R software (version 3.4.4) [67] on a PC desktop with a Core i7 Processor (3.41 GHz) and 16 GB RAM. Data sets of all selected problems were divided into a training set and test set using a random partition with (20/80)% ratio. Similarity matrices ($M \times M$ matrix containing intermolecular similarities) of all data sets were computed using NAMS [17] that allows for the computation of pairwise similarities between all molecules within a database. All the other NAMS parameters were left as default. Similarity matrices were converted into dissimilarity matrices (metric space representation) using eq. 4, with $k = 0.382$ for all data sets. For the DR processes, R `cmdscale` function [68] was used for PCoA, two functions from R package MASS [44] including `isoMDS` and `sammon` was used for KMDS and SM respectively. We used the t-SNE implementation from R library `Rtsne` [37]. Finally, for computing the kernel density map in 2D, the `kde2d` function from R package MASS [44] was used. The bandwidths for the positive and negative maps were calculated beforehand using the `bandwidth.nrd` function (see Eq. 5.9).

5.5 Results and discussion

We generated the activity (K_i) probability maps (PSMA) for four different problems (*SIGMARI*, *HRHI*, *HERG*, and *DRD5*) using 4 different DR methods: PCooa, KMDS, SM, and t-SNE. For each we have produced the probability of activity surface maps. These PSMA typically show consistency, and the regions with the highest probability of activity appear most of the times well differentiated from the negative regions. Figure 5.3 shows the results of PCooA projection for the 4 data sets, computed solely from the 80% training data. In these probability maps, surface height mirrors the kernel density distribution of active molecules (positive class) and the colour represents higher probability locations (most likely activity regions).

To check the quality of the produced probability surface maps, the test set molecules were projected into the new reference plane. The performance of all sixteen generated PSMA was assessed using AUC and MCC. AUC testing results range from 0.77 to 0.98 and MCC score ranges between 0.18 to 0.77 (Table 5.2). In two data sets (*SIGMARI* and *DRD5*), PCooA performance was better than the other DR methods while for *HRHI* and *HERG* t-SNE outperformed the others. All DR approaches provided generally good AUC results. The overall performance of PCooA and t-SNE was roughly the same (average AUC = 0.86) in all four problems with a slightly (and not statistically significant) more positive outcome for PCooA with the MCC score.

The test set projections over the best PSMA for each data set shows ground truth active molecules (as red circles) typically within the highest probability of activity regions (Figure 5.4). In the present analysis is several cases, the MCC was low, albeit always showing clear discriminant power. On the other hand, the AUC score was consistently high.

Since dimensionality reduction is one of the important task in data visualization where

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

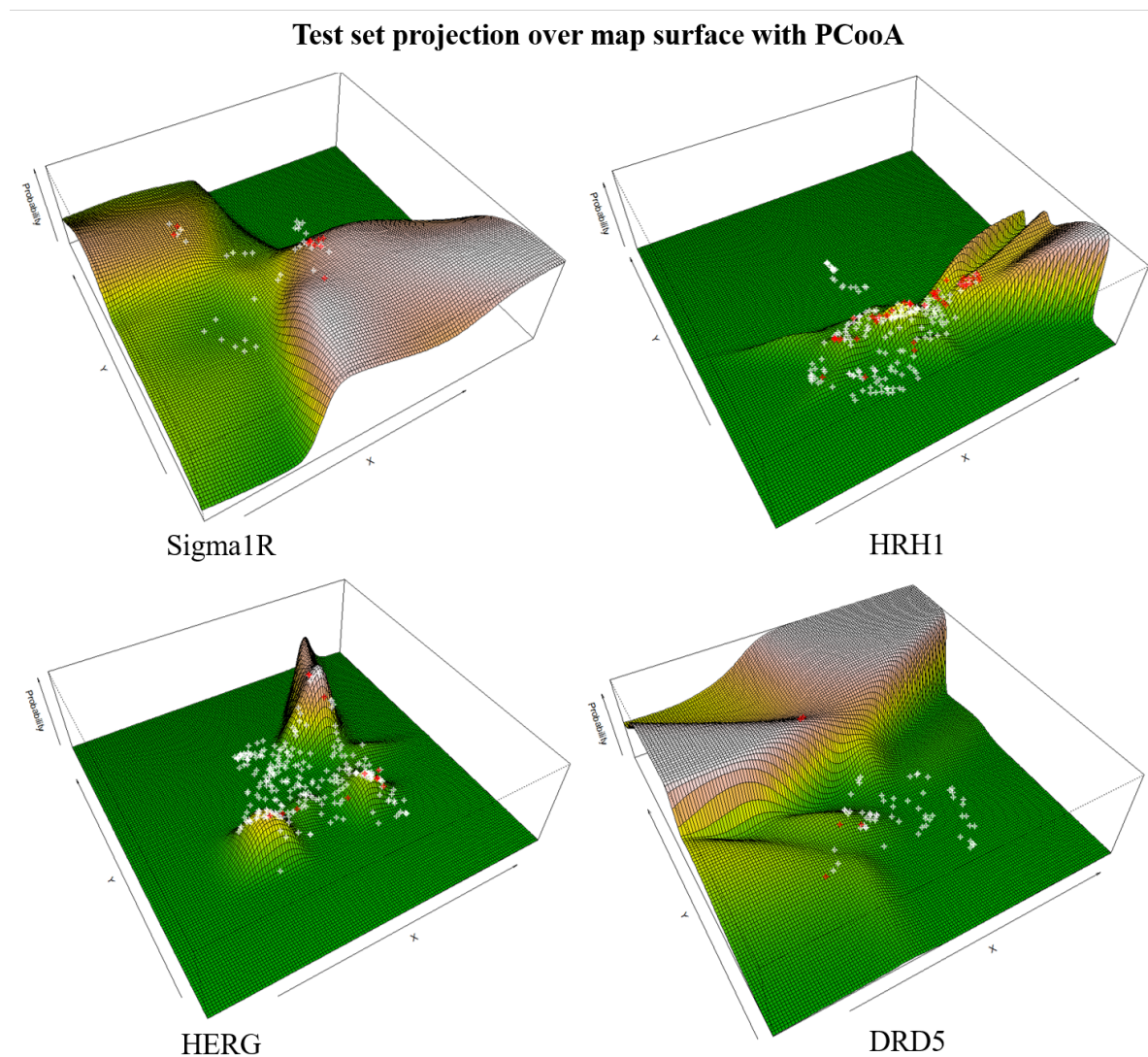


Figure 5.3: Test set projection over map surface (PSMA) with PCooA. Surfaces represents higher probability locations. red – circles are ground truth positives, white are ground truth negatives.

5.5 Results and discussion

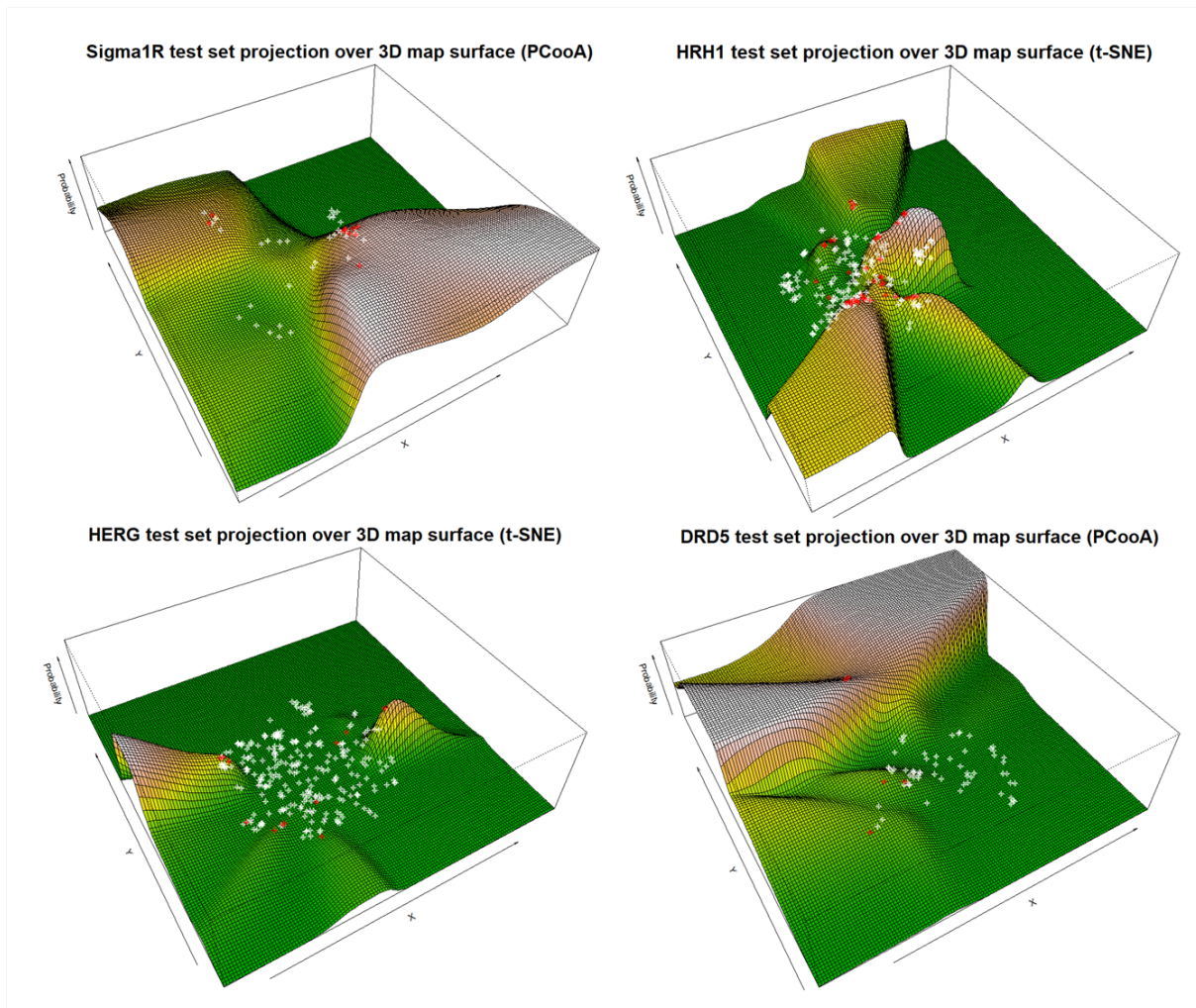


Figure 5.4: Test set projection over map surface of selected PSMAs with highest performance. Surfaces represents higher probability locations. red – circles are ground truth positives, white are ground truth negatives.

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

Table 5.2: Results on validation set ((* – best model). Abbreviations: Principal Co-ordinates Analysis (PCooA), Kruskal Multidimensional Scaling (KMDS), Sammon mapping (SM), and t-Distributed Stochastic Neighbor Embedding (t-SNE)

Target Protein Name	PCooA		KMDS		SM		t-SNE	
	AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC
Sigma non-opioid intracellular receptor 1 (Sigma1R)	0.87(*)	0.63	0.80	0.60	0.79	0.55	0.79	0.47
Histamine H1 receptor (HRH1)	0.80	0.45	0.83	0.43	0.78	0.36	0.87(*)	0.54
Potassium voltage-gated channel subfamily H member 2 (HERG)	0.80	0.18	0.77	0.24	0.80	0.25	0.89(*)	0.33
D(1B) dopamine receptor (DRD5)	0.98(*)	0.77	0.86	0.32	0.80	0.42	0.90	0.41
Overall performance (average score)	0.86	0.51	0.82	0.40	0.79	0.40	0.86	0.44

it is really necessary to capture the maximum original data information in the new reduced space, Shepard plots [44] were generated to analyze how much molecular initial proximity relationship remained intact. In Shepard plots the original distances are plotted against the projected distances and, ideally, the points (both distances) should lie on a straight line, which would indicate zero distortion in the projection function. The Shepard plot for the hERG data set, for all projection functions is shown (Figure 5.5). The 2D projections, for all approaches, showed a similar pattern, in which it can be seen that many large distances in the initial space fail to maintain that separation in the projected space, however, in all cases, very close molecules will always appear close, which shows that locality factors were preserved in all projections, which contributes to explain the quality of the classification models. Nonetheless, the projection of dissimilar molecules in the vicinity of similar molecules can generate noise in visualization of the real pattern of the data distribution. This is probably the cause for having low MCC scores in some data sets.

To verify whether the quality of the projection influences the classification results, the R^2 coefficient that measures how the projected distances measure against the original distances was calculated (Figure 5.5). It is apparent that KMDS shows the highest scores while Sammon mapping features the lowest values for all 4 test cases. It is therefore striking that KMDS although always able to produce consistently good models was never the projection

5.5 Results and discussion

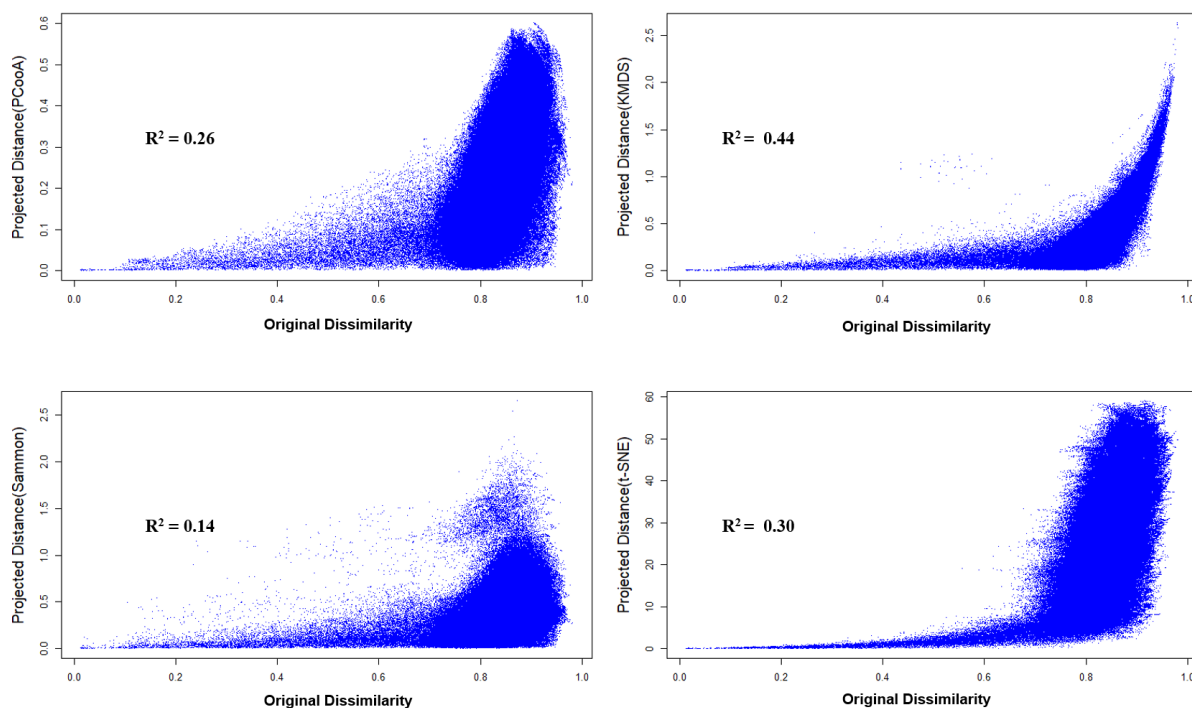


Figure 5.5: HERG Shepard plot for PCoA, KMDS, SM and t-SNE

that yielded the best results. This may suggest that, on this reduced dimension space, other factors rather than stricter distance preservation may be relevant for accurate model building, and the nonlinear optimization performed by KMDS actually hampers the projection quality for classification purposes.

To have a more detailed appraisal of the quality of the test set projections, the 2D molecular structures of top 6 test molecules with higher probability of being actives (predicted positives) are shown within the 2D probability map of the best 2D projections for each data set (Table 5.2), along with their ChEMBL IDs (Figure 5.6). It can be seen that, with only one exception, in all 4 data sets, all molecules were strong actives, although some not within the strict activity criterion ($K_i \leq 10$). For the Sigma 1 Receptor, there were 3 correctly predicted positives, and the three negatives incorrectly predicted had in fact very low K_i values, all of

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

them below 30 nM. For the Histamine 1 receptor, the 5 more likely molecules to be active were all correctly predicted as positives, which is striking as this data set is one of the hardest, with low classification results. The only miss is one molecule (CHEMBL1767152) with a $K_i = 31.62$, therefore with strong activity as well. The human hERG is the hardest problem, as the number of negative molecules largely outnumbers the positives, nonetheless for the 6 molecules with highest probability, 4 were correctly predicted only two were misses. As in the previous cases, both molecules (CHEMBL1086480 and CHEMBL1085091) are also strong actives, with $K_i \leq 50nM$. The last test set (Dopamine 5 receptor), is the one with the more striking situation, as this was the data set that had the highest classification performance. The two misses, the first molecule had a $K_i = 10.4$, thus clearly a borderline molecule. The compound CHEMBL595720, was the only one that on ChEMBL was a clear inactive with a measured $K_i \geq 10,000$. It can be pointed out that, on this specific problem, that molecule is outside the most active region which appears clearly marked on the upper region of the map, with an activity cliff crossing the full surface, identifying the most promising region for finding very active molecules.

5.6 Conclusion

This study aimed initially at presenting a visualization method that is able to capture the highest probability regions for molecules being active. To reach this goal, the molecular spaces of four data sets, captured as similarity matrices, were reduced into two a new reference space in 2D using four different algorithms. The X, Y coordinates generated from each DR methods were used by a 2D kernel density function to generate their corresponding activity probability maps (PSMAs). These PSMAs were able to depict the most likely activity regions, and appear consistent, with active molecules clearly grouping together. The analysis of the produced PSMAs from the 4 data sets showed the reliability of the proposed

5.6 Conclusion

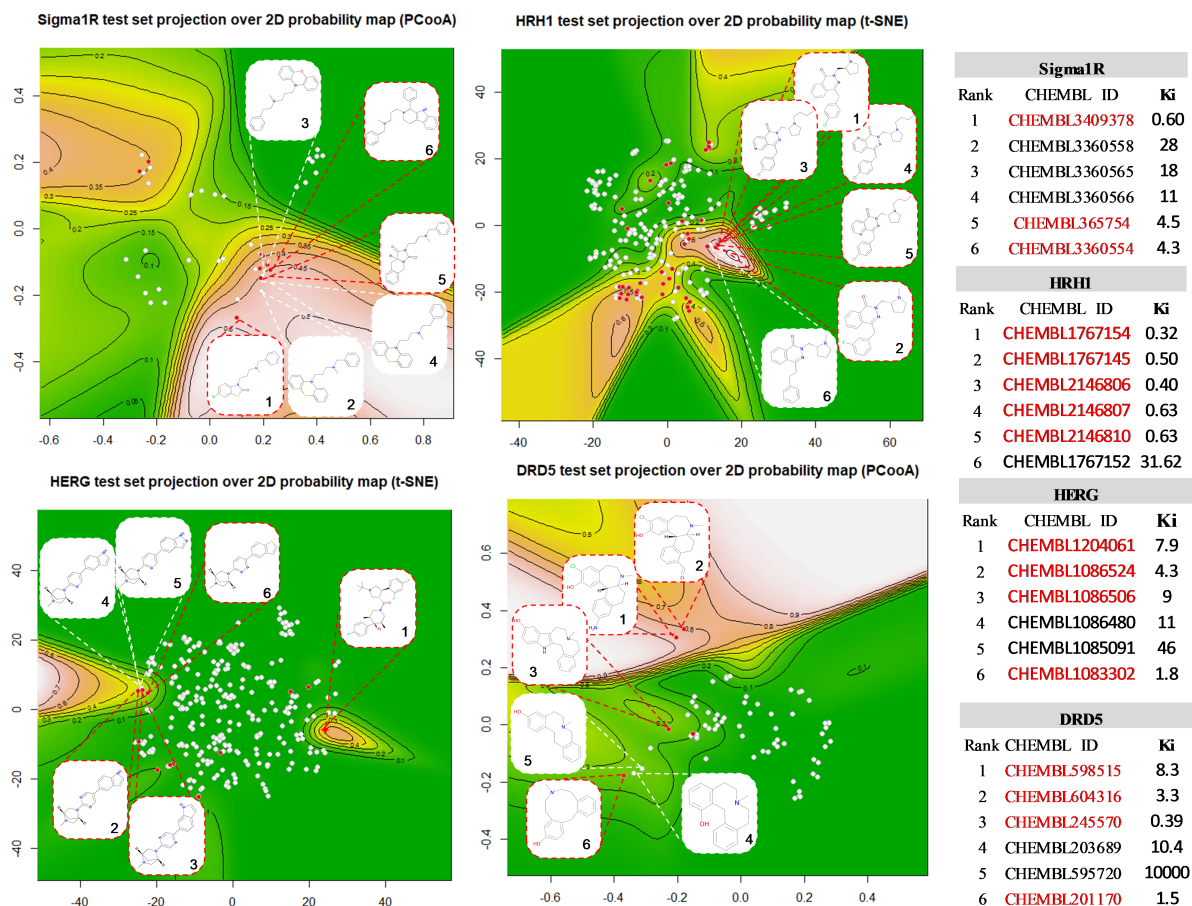


Figure 5.6: Test set projection over 2D probability map of selected models with highest performance. Contour lines represent 2D kernel density distribution of active molecules (positive class) and the colour other than green represents higher probability locations. red – circles are ground truth positives, white are ground truth negatives. ChEMBL IDs. in red color text (2D structures within red lined box) are true positives and other are false positives (2D structures within white lined box).

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

methodology as it can efficiently produce visual cues as to where the more promising regions of the molecular space are located. The presented approach allows for the projection of new molecules into the new projected space, thus allowing for model assessment with external data. Accordingly, to validate the quality of this 2D representation as a classification model, independent validation sets were projected over the generated PSMA, and the results were consistently good with AUC values, for the highest scoring projections, ranging from 0.87 to 0.98 and MCC scores ranging from 0.33 to 0.77. Although the followed approach did not aim at optimizing models for getting high classification accuracies, these results are strongly suggestive that it actually is capturing a large part of the modelable aspects of these SAR problems. This approach therefore uses only the 2D structural similarity between molecules to produce a non-parametric model that is both visually informative and shows demonstrable quality as a classification model.

The predictability of the presented spatial classification model (PSMA) is thus an attractive feature for virtual screening using only structural similarity of molecules. The applicability domain of such visual approaches can be vastly increased using larger data sets for any single or multiple targets. Comparatively to traditional QSAR models with a limited applicability domain, this activity space visualization directly uses structural similarity and thus may enhance SAR visualization within large activity spaces.

Acknowledgements

The authors gratefully acknowledge Fundação para a Ciência e Tecnologia for a doctoral grant (SFRH/BD/111654/2015), MIMED project funding (PTDC/EEI-ESS/4923/2014) and UID/CEC/00408/2013 (LaSIGE) for providing the infrastructure.

References

- [1] Mahendra Awale et al. 'Chemical Space: Big Data Challenge for Molecular Diversity'. In: *CHIMIA International Journal for Chemistry* 71.10 (2017), pp. 661–666. ISSN: 0009-4293. DOI: 10.2533/chimia.2017.661.
- [2] Jean Louis Reymond et al. 'Chemical space as a source for new drugs'. In: *MedChemComm* 1.1 (2010), pp. 30–38. ISSN: 20402503. DOI: 10.1039/c0md00020e.
- [3] Christopher M. Dobson. 'Chemical space and biology'. In: *Nature* 432.7019 (Dec. 2004), pp. 824–828. ISSN: 0028-0836. DOI: 10.1038/nature03192.
- [4] Pavel Sidorov et al. 'QSAR modeling and chemical space analysis of antimalarial compounds'. In: *Journal of Computer-Aided Molecular Design* 31.5 (2017), pp. 441–451. ISSN: 15734951. DOI: 10.1007/s10822-017-0019-4.
- [5] Jeremy Ash and Denis Fourches. 'Characterizing the Chemical Space of ERK2 Kinase Inhibitors Using Descriptors Computed from Molecular Dynamics Trajectories'. In: *Journal of Chemical Information and Modeling* 57.6 (2017), pp. 1286–1299. ISSN: 15205142. DOI: 10.1021/acs.jcim.7b00048.
- [6] Martin Vogt. 'Progress with modeling activity landscapes in drug discovery'. In: *Expert Opinion on Drug Discovery* 13.7 (2018), pp. 605–615. ISSN: 1746045X. DOI: 10.1080/17460441.2018.1465926.
- [7] Alexandre Varnek and Alex Tropsha. *Cheminformatics Approaches to Virtual Screening*. Ed. by Alexandre Varnek and Alex Tropsha. Cambridge: Royal Society of Chemistry, 2008. ISBN: 978-0-85404-144-2. DOI: <http://dx.doi.org/10.1039/9781847558879>.
- [8] Nina Nikolova and Joanna Jaworska. 'Approaches to Measure Chemical Similarity—a Review'. In: *QSAR & Combinatorial Science* 22.910 (2003), pp. 1006–1026. ISSN: 1611-020X. DOI: 10.1002/qsar.200330831.

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

- [9] Mark A Johnson and Gerald M Maggiora. *Concepts and applications of molecular similarity*. 1990. ISBN: 0471621757.
- [10] Peter Willett, John M. Barnard and Geoffrey M. Downs. ‘Chemical Similarity Searching’. In: *Journal of Chemical Information and Computer Sciences* 38.6 (Nov. 1998), pp. 983–996. ISSN: 0095-2338. DOI: 10.1021/ci9800211.
- [11] Andreas Bender and Robert C Glen. ‘Molecular similarity: a key technique in molecular informatics.’ In: *Organic & biomolecular chemistry* 2.22 (2004), pp. 3204–3218. ISSN: 1477-0520. DOI: 10.1039/b409813g.
- [12] Gerald Maggiora et al. ‘Molecular Similarity in Medicinal Chemistry’. In: *Journal of Medicinal Chemistry* 57.8 (Apr. 2014), pp. 3186–3204. ISSN: 0022-2623. DOI: 10.1021/jm401411z.
- [13] Hanna Eckert and Jürgen Bajorath. ‘Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches’. In: *Drug Discovery Today* 12.5-6 (2007), pp. 225–233. ISSN: 13596446. DOI: 10.1016/j.drudis.2007.01.011.
- [14] Dagmar Stumpfe and Jürgen Bajorath. ‘Similarity searching’. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.2 (Mar. 2011), pp. 260–282. ISSN: 17590876. DOI: 10.1002/wcms.23.
- [15] Gerald M Maggiora and Veerabahu Shanmugasundaram. ‘Molecular Similarity Measures’. In: *Methods in molecular biology (Clifton, N.J.)* 2004, pp. 1–50. DOI: 10.1385/1-59259-802-1:001.
- [16] Jürgen Bajorath. ‘Molecular Similarity Concepts for Informatics Applications’. In: *Bioinformatics: Volume II: Structure, Function, and Applications*. Ed. by Jonathan M. Keith. New York, NY: Springer New York, 2017, pp. 231–245. ISBN: 978-1-4939-6613-4. DOI: 10.1007/978-1-4939-6613-4_13.

REFERENCES

- [17] Ana L Teixeira and Andre O Falcao. 'Noncontiguous atom matching structural similarity function'. In: *Journal of Chemical Information and Modeling* 53.10 (2013), pp. 2511–2524. ISSN: 15499596. DOI: 10.1021/ci400324u.
- [18] Hans-Christian Ehrlich and Matthias Rarey. 'Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review'. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.1 (Jan. 2011), pp. 68–79. ISSN: 17590876. DOI: 10.1002/wcms.5.
- [19] John W. Raymond and Peter Willett. 'Maximum common subgraph isomorphism algorithms for the matching of chemical structures.' In: *Journal of computer-aided molecular design* 16.7 (July 2002), pp. 521–33. ISSN: 0920-654X. DOI: 10.1023/A:1021271615909.
- [20] John M. Barnard. 'Substructure searching methods: Old and new'. In: *Journal of Chemical Information and Modeling* 33.4 (July 1993), pp. 532–538. ISSN: 1549-9596. DOI: 10.1021/ci00014a001.
- [21] H.A. Gaspar, I.I. Baskin and A. Varnek. 'Visualization of a multidimensional descriptor space'. In: *ACS Symposium Series* 1222 (2016). ISSN: 19475918. DOI: 0.1021/bk-2016-1222.ch012.
- [22] John Aldo Lee and Michel Verleysen. *Nonlinear Dimensionality Reduction*. Ed. by John A. Lee and Michel Verleysen. Information Science and Statistics. New York, NY: Springer New York, 2007. ISBN: 978-0-387-39350-6. DOI: 10.1007/978-0-387-39351-3.
- [23] Dagmar Stumpfe and Jürgen Bajorath. 'Recent developments in SAR visualization'. In: *MedChemComm* 7.6 (2016), pp. 1045–1055. ISSN: 20402511. DOI: 10.1039/c6md00108d.
- [24] Colin Goodall and I T Jolliffe. 'Principal Component Analysis'. In: *Technometrics* 30.3 (Aug. 1988), p. 351. ISSN: 00401706. DOI: 10.2307/1270093.

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

- [25] Lars Ruddigkeit, Lorenz C. Blum and Jean-Louis Reymond. ‘Visualization and Virtual Screening of the Chemical Universe Database GDB-17’. In: *Journal of Chemical Information and Modeling* 53.1 (Jan. 2013), pp. 56–65. ISSN: 1549-9596. DOI: 10.1021/ci300535x.
- [26] Mahendra Awale, Ruud van Deursen and Jean-Louis Reymond. ‘MQN-Mapplet: Visualization of Chemical Space with Interactive Maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13’. In: *Journal of Chemical Information and Modeling* 53.2 (Feb. 2013), pp. 509–518. ISSN: 1549-9596. DOI: 10.1021/ci300513m.
- [27] Warren S. Torgerson. ‘Multidimensional scaling: I. Theory and method’. In: *Psychometrika* 17.4 (Dec. 1952), pp. 401–419. ISSN: 0033-3123. DOI: 10.1007/BF02288916.
- [28] J. B. Kruskal. ‘Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis’. In: *Psychometrika* 29.1 (Mar. 1964), pp. 1–27. ISSN: 0033-3123. DOI: 10.1007/BF02289565.
- [29] J.W. Sammon. ‘A Nonlinear Mapping for Data Structure Analysis’. In: *IEEE Transactions on Computers* C-18.5 (May 1969), pp. 401–409. ISSN: 0018-9340. DOI: 10.1109/T-C.1969.222678.
- [30] T. Kohonen. ‘The self-organizing map’. In: *Proceedings of the IEEE* 78.9 (1990), pp. 1464–1480. ISSN: 00189219. DOI: 10.1109/5.58325.
- [31] Geoffrey E Hinton and Sam T. Roweis. ‘Stochastic Neighbor Embedding’. In: *Advances in Neural Information Processing Systems 15*. Ed. by S. Becker, S. Thrun and K. Obermayer. MIT Press, 2003, pp. 857–864.
- [32] Dimitris K. Agrafiotis. ‘Stochastic proximity embedding’. In: *Journal of Computational Chemistry* 24.10 (July 2003), pp. 1215–1221. ISSN: 0192-8651. DOI: 10.1002/jcc.10234.

REFERENCES

- [33] Ana L Teixeira and Andre O Falcao. ‘Structural similarity based kriging for quantitative structure activity and property relationship modeling.’ In: *Journal of chemical information and modeling* 54.7 (2014), pp. 1833–1849. ISSN: 1549-960X. DOI: 10.1021/ci500110v.
- [34] Anne Mai Wassermann, Mathias Wawer and Jürgen Bajorath. ‘Activity landscape representations for structure-activity relationship analysis’. In: *Journal of Medicinal Chemistry* 53.23 (2010), pp. 8209–8223. ISSN: 00222623. DOI: 10.1021/jm100933w.
- [35] Lisa Peltason, Preeti Iyer and Jürgen Bajorath. ‘Rationalizing Three-Dimensional Activity Landscapes and the Influence of Molecular Representations on Landscape Topology and the Formation of Activity Cliffs’. In: *Journal of Chemical Information and Modeling* 50.6 (June 2010), pp. 1021–1033. ISSN: 1549-9596. DOI: 10.1021/ci100091e.
- [36] Dagmar Stumpfe et al. ‘Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry’. In: *Journal of Medicinal Chemistry* 57.1 (Jan. 2014), pp. 18–28. ISSN: 0022-2623. DOI: 10.1021/jm401120g.
- [37] Laurens van der Maaten and Geoffrey Hinton. ‘Visualizing Data using t-SNE’. In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.
- [38] Bw. Silverman. ‘Density estimation for statistics and data analysis’. In: *Chapman and Hall* 37.1 (1986), pp. 1–22. ISSN: 00359254. DOI: 10.2307/2347507.
- [39] Abraham Yosipof, Rita C Guedes and Alfonso T García-Sosa. ‘Data Mining and Machine Learning Models for Predicting Drug Likeness and Their Disease or Organ Category.’ In: *Frontiers in chemistry* 6.May (2018), p. 162. ISSN: 2296-2646. DOI: 10.3389/fchem.2018.00162.
- [40] Joanna Jaworska, Tom Aldenberg and Nina Nikolova. ‘Review of methods for QSAR applicability domain estimation by the training set’. In: *Atla* 33 (2005), pp. 445–459.

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

- [41] Faizan Sahigara et al. ‘Comparison of different approaches to define the applicability domain of QSAR models’. In: *Molecules* 17.5 (2012), pp. 4791–4810. ISSN: 14203049. DOI: 10.3390/molecules17054791.
- [42] Natália Aniceto et al. ‘A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: Reliability-density neighbourhood’. In: *Journal of Cheminformatics* 8.1 (2016), pp. 1–20. ISSN: 17582946. DOI: 10.1186/s13321-016-0182-y.
- [43] Alexander Tropsha and Alexander Golbraikh. ‘Predictive QSAR modeling workflow, model applicability domains, and virtual screening.’ In: *Current pharmaceutical design* 13.34 (2007), pp. 3494–504. ISSN: 1873-4286. DOI: 10.2174/13816120778279425
- [44] William N Venables and Brian D Ripley. *Modern Applied Statistics with S*. Springer-Verlag, 2002. ISBN: 0-387-95457-0. DOI: 10.1016/j.electacta.2013.08.022..
- [45] Johann Gasteiger. *Handbook of Chemoinformatics*. Ed. by Johann Gasteiger. Vol. 1-4. Weinheim, Germany: Wiley-VCH Verlag GmbH, Aug. 2003, pp. 1–1870. ISBN: 9783527618279. DOI: 10.1002/9783527618279.
- [46] Roberto Todeschini and Viviana Consonni. *Molecular Descriptors for Chemoinformatics*. Ed. by Roberto Todeschini and Viviana Consonni. Methods and Principles in Medicinal Chemistry. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, July 2009. ISBN: 9783527628766. DOI: 10.1002/9783527628766.
- [47] C. James, D. Weininger and J. Delaney. *Daylight Theory Manual version 4.9*. 2011.
- [48] Peter Willett. ‘The Calculation of Molecular Structural Similarity: Principles and Practice’. In: *Molecular Informatics* 33.6-7 (June 2014), pp. 403–413. ISSN: 18681743. DOI: 10.1002/minf.201400024.

REFERENCES

- [49] Ingo Muegge and Prasenjit Mukherjee. ‘An overview of molecular fingerprint similarity search in virtual screening’. In: *Expert Opinion on Drug Discovery* 11.2 (2016), pp. 137–148. ISSN: 1746-0441. DOI: 10.1517/17460441.2016.1117070.
- [50] Swarit Jasial et al. ‘Activity-relevant similarity values for fingerprints and implications for similarity searching’. In: *F1000Research* 5.0 (2016), p. 591. ISSN: 2046-1402. DOI: 10.12688/f1000research.8357.2.
- [51] Dávid Bajusz, Anita Rácz and Károly Héberger. ‘Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?’ In: *Journal of Cheminformatics* 7.1 (2015), pp. 1–13. ISSN: 17582946. DOI: 10.1186/s13321-015-0069-3.
- [52] Choi Seung-Seok, Cha Sung-Hyuk and Charles C Tappert. ‘A Survey of Binary Similarity and Distance Measures.’ In: *Journal of Systemics, Cybernetics & Informatics* (2010). ISSN: 16904524. DOI: 10.1.1.352.6123.
- [53] J W Johnston. ‘Similarity indices I: What do they measure?’ In: *Battelle: Pacific Northwest Laboratories, Richland, Washington* (1976).
- [54] D.R. Flower. ‘On the Properties of Bit String-Based Measures of Chemical Similarity’. In: *Journal of Chemical Information and Modeling* 38.3 (1998), pp. 379–386. ISSN: 1549-9596. DOI: 10.1021/ci970437z.
- [55] Valerie J. Gillet, Peter Willett and John Bradshaw. ‘Similarity Searching Using Reduced Graphs †’. In: *Journal of Chemical Information and Computer Sciences* 43.2 (Mar. 2003), pp. 338–345. ISSN: 0095-2338. DOI: 10.1021/ci025592e.
- [56] Robert P. Sheridan and Simon K. Kearsley. ‘Why do we need so many chemical similarity search methods?’ In: *Drug Discovery Today* 7.17 (Sept. 2002), pp. 903–911. ISSN: 13596446. DOI: 10.1016/S1359-6446(02)02411-X.

5. A VISUAL APPROACH FOR ANALYSIS AND INFERENCE OF MOLECULAR ACTIVITY SPACES

- [57] José Batista, Jeffrey W. Godden and Jürgen Bajorath. ‘Assessment of Molecular Similarity from the Analysis of Randomly Generated Structural Fragment Populations’. In: *Journal of Chemical Information and Modeling* 46.5 (Sept. 2006), pp. 1937–1944. ISSN: 1549-9596. DOI: 10.1021/ci0601261.
- [58] Daniel J. Graham, Christopher Malarkey and Matthew V. Schulmerich. ‘Information Content in Organic Molecules: Quantification and Statistical Structure via Brownian Processing’. In: *Journal of Chemical Information and Computer Sciences* 44.5 (Sept. 2004), pp. 1601–1611. ISSN: 0095-2338. DOI: 10.1021/ci0400213.
- [59] M. Thorrington-Smith. ‘West Indian Ocean phytoplankton: a numerical investigation of phytohydrographic regions and their characteristic phytoplankton associations’. In: *Marine Biology* 9.2 (May 1971), pp. 115–137. ISSN: 0025-3162. DOI: 10.1007/BF00348251.
- [60] R. Todeschini et al. ‘CAIMAN (Classification And Influence Matrix Analysis): A new approach to the classification based on leverage-scaled functions’. In: *Chemo-metrics and Intelligent Laboratory Systems* 87.1 (May 2007), pp. 3–17. ISSN: 01697439. DOI: 10.1016/j.chemolab.2005.11.001.
- [61] AG Gray and AW Moore. ‘Proceedings of the 2003 SIAM International Conference on Data Mining’. In: *Nonparametric Density Estimation: Toward Computational Tractability*. Ed. by Daniel Barbara and Chandrika Kamath. Philadelphia, PA: Society for Industrial and Applied Mathematics, May 2003. ISBN: 978-0-89871-545-3. DOI: <https://doi.org/10.1137/1.9781611972733.19>.
- [62] Richard O. Duda, Peter E. Hart and David G. Stork. *Pattern Classification (2Nd Edition)*. New York, NY, USA: Wiley-Interscience, 2000. ISBN: 0471056693.
- [63] Pierre Baldi and Søren Søren Brunak. ‘Bioinformatics: The Machine Learning Approach’. In: *MIT Press* (2001), IXXI, 1452. ISSN: 0269-8889. DOI: 10.1017/S0269888904220161.

REFERENCES

- [64] Anna Gaulton et al. 'The ChEMBL database in 2017'. In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D945–D954. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1074.
- [65] A Z Dudek, T Arodz and J Galvez. 'Computational methods in developing quantitative structure-activity relationships (QSAR): a review'. In: *Comb Chem High Throughput Screen* 9.3 (2006), pp. 213–228. ISSN: 13862073. DOI: 10.2174/138620706776055539.
- [66] Samina Kausar and Andre O. Falcao. 'An automated framework for QSAR model building'. In: *Journal of Cheminformatics* 10.1 (Dec. 2018), p. 1. ISSN: 1758-2946. DOI: 10.1186/s13321-017-0256-5.
- [67] R R Development Core Team. *R: A Language and Environment for Statistical Computing*. 2011. DOI: 10.1007/978-3-540-74686-7.
- [68] J. C. Gower. 'Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis'. In: *Biometrika* 53.3/4 (Dec. 1966), pp. 325–328. ISSN: 00063444. DOI: 10.2307/2333639.

6

Comparative analysis of QSAR modeling and molecular docking: a rational approach in polypharmacology

SAMINA KAUSAR, RITA C. GUEDES AND ANDRE O FALCAO

6.1 Introduction

Drug discovery is a laborious and an interdisciplinary endeavour that relies on the advancement of multi-disciplinary (quantum physics and chemistry, molecular biology, bioinformatics and information technologies etc.) high-tech investigations. Given a validated target, drug discovery cycle starts with hit and lead identification, followed by lead optimization and *in-vitro* and *in-vivo* analysis (pre-clinical trials) for validating the desired activity profiles before entering into clinical trials [1, 2, 3]. In the recent years, the increasing number

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

of publications showing the current trends in modern drug discovery methodology has shifted from the traditional single target ('magic bullets') towards multi-targeting drug designing or polypharmacology ('magic shotguns') [4, 5]. Formally, polypharmacology approaches predict the promiscuous (single drug molecule activity toward multiple targets) behaviour of drugs in a single or multiple disease pathways. In the case of complex multifactorial disease mechanisms, such as central nervous system (CNS) disorders or cancers, polypharmacology is the preferred protocol that involve a trade-off between the compound's activity toward required therapeutic targets (specificity) and nontherapeutic targets (promiscuity) to achieve a molecule with desired activity profiles related to the efficacy, safety, and adverse toxic effects [6, 7]. Today, sufficient amount of data of small molecules, validated targets and drug-target interactions is available; however, using this prior knowledge of biological information a large range of *in-silico* approaches have been developed for addressing both therapeutic and adverse aspects of polypharmacology [8, 9, 10]. These computational methodologies refer to unique virtual screening (VS) strategies that are adapted for polypharmacology applications. VS strategies can be broadly categorized as structure-based and ligand-based methods [11].

Structure-based VS depends on the availability of a 3D structure of the target protein as their underlying hypothesis is that structurally similar proteins (specifically similar binding pockets) are likely to exhibit similar functions/activity and thus, can bind to similar compounds. Structure-based approaches for polypharmacology depend on algorithms that either assess similarity between the binding pockets of different targets [12] or uses molecular docking for automatic evaluation of multi-targeting drugs [13]. Molecular (protein-ligand) docking has become a mainstream structure-based VS method searching for small molecules that mimic the binding interaction of ligands into the active site and rank them according to decreasing predicted binding affinity (scoring) [14, 15]. The continuous development in refinements and optimization of the docking algorithms has significantly improved the success rate in hits identification and has become one of the major sources of finding novel lead molecules (scaffold) that are used as a starting point in the drug discovery process [7, 16, 15,

17, 18, 19, 3].

Ligand-based methods such as e.g., similarity searching [20, 21, 22], pharmacophore mapping [23] and quantitative structure-activity relationships (QSAR) modeling are alternative options in the absence of a 3D structural model of target protein [9, 24, 25]. Ligand-based VS largely depends on the availability of the activity profiles among molecules (known actives and inactives) and targets. These approaches incorporate different chemical information (chemical and biological properties, structure, shape, and bioactivity etc.) and have been actively used for analysing large molecular databases to identify ligands likely to have similar properties to the known actives [21, 26, 27, 28, 29]. Among the Ligand-based VS methods, QSAR modeling is the most powerful tool due to its high and fast computational efficiency and a good hit rate [24]. In QSAR modeling, machine learning methods are used that rely on quantitative properties and bioactivity profiles of known actives and inactives for deriving correlations between molecular structural/property features and pharmacological activity [9, 24, 25]. Thus, ligand-based approaches are data-driven and largely depend on quality and the amount of prior knowledge of compounds' activity [30, 31, 32, 33, 34]. Polypharmacological applications of these methods are often limited to targets whose ligands have well-documented activity profiles [7].

VS is a computational analogue of biological high-throughput screening (HTS), an automated plate based experimental assays technique for rapid identification of best hits (compounds with desired activity) by screening a large collection (10^5 – 10^6) of chemical molecules [35, 24, 36]. VS has emerged as a reliable, fast and cost-effective technique for *in-silico* screening of large small-molecule databases for the discovery of lead compounds (drug candidates) [37, 1, 35, 3, 2, 19, 24]. Moreover, in the several comparative studies [36, 38, 39] of hit identification performance (true positive hit rate) of HTS and VS methods including docking and QSAR-based screening, hits rate from HTS was found 0.021% comparatively to docking hits rate 34.8% [39], while QSAR and HTS comparative studies showed HTS had hits rate ranges between 0.2-0.94% and QSAR had 3.6-28.2% [36, 38, 40]. However, the

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

overall reported observed range of hits rate from a validated VS methods is between 1-40% while for the HTS is 0.01-0.1% [36, 38, 40, 39, 41, 42]. Therefore, VS campaigns enrich the hits rate due to a higher rate of actives with desired activity and consequently, reduce the drug discovery cost that is much lower than HTS. Polypharmacology-based VS is a promising tool for finding drugs with a multi-target activity profile against complex diseases like CNS disorders [43, 44, 45, 7, 13]. But the advancement of multi-targeting VS is still required to deal with large amounts of chemical data for the screening or identification of chemical structures with direct binding and inhibitory activity against multiple targets [46, 44].

Aiming to find a rigorous method for VS with highest hits rate in polypharmacology, comparative analysis of both molecular docking and QSAR modeling approaches was performed. The objective of this study is twofold, firstly to provide a comparative view of predictability/hits rate of each method and secondly, development of a rational polypharmacology-based drug designing methodology by integrating the knowledge of molecular docking and QSAR modeling approaches. Integration of both strategies was an effort to combine their corresponding advantages to find novel therapeutic molecules using all available data of proteins crystallographic structures and their binders (ligand) structural and biological activity.

The developed pipeline was used to identify dual-targeting hits (inhibitors) of catechol-O-methyltransferase (COMT) and glycogen synthase kinase-3 beta (GSK3 β) against Parkinson's disease (PD) chosen as a case study.

6.2 Targets selection for Parkinson's disease

PD is the second-most common progressive neurodegenerative condition, including striatal dopamine deficiency due to dopaminergic neurons loss in specific areas of the substantia nigra and widespread accumulation Lewy bodies, aggregates of intraneuronal protein (α -synuclein) [47, 48, 49]. In the last 200 years, efforts and progress in PD research revealed

6.2 Targets selection for Parkinson's disease

that PD underlying neuropathology and progression involves multiple pathways including α synuclein proteostasis, mitochondrial function, oxidative stress, calcium homeostasis, axonal transport and neuroinflammation [47, 49]. Striatal dopamine reduction leads to the disorder of movement with different motor symptoms including a) slower movements (bradykinesia), b) partial or complete loss of body movements (hypokinesia), c) loss of voluntary muscles movement (akinesia), and d) muscles stiffness/rigidity and tremors [50].

Currently, none of any available anti-Parkinson drugs can prevent the progressive degeneration of dopaminergic neurons and only provide symptomatic treatment. In dopaminergic therapies, most of the drugs substitute striatal dopamine deficiency by a) stimulating dopamine receptors (i.e., dopamine agonists), b) increasing the dopamine biosynthesis (i.e., *L*-3,4-*Dihydroxyphenylalanine* (L-DOPA) administration), and c) decreasing dopamine degradation or metabolism (i.e., COMT or Monoamine oxidase B (MAO-B) inhibitors) [51, 52]. Systemic administration of L-DOPA, dopamine precursor amino acid is considered the most effective treatment for managing motor symptoms in all PD patients [53, 54]. But, prolong administration of L-DOPA can be neurotoxic and may cause L-DOPA-induced dyskinesia (LID). Although, the LID underlying mechanism is still incompletely understood but many studies suggested that two interacting factors like striatal dopaminergic denervation and pulsatile L-DOPA treatment possibly generate maladaptive neuronal responses [55] including dysregulation of dopamine transmission, and striatal neurons intracellular signalling cascades abnormalities, altered gene expression, and corticostriatal synaptic plasticity [56].

It is well demonstrated that dysregulation of the enzyme GSK3 β , regulate the glycogen synthesis being associated with diverse cellular processes related to cell survival and apoptosis and being involved in neurodegenerative diseases [57, 58, 59] including L-DOPA neurotoxicity in PD pathogenesis [60, 61]. However, considering the consensus finding of several research efforts, GSK3 β can be a potential target and development of new therapeutic inhibitors can be a promising strategy for preventing L-DOPA neurotoxicity and LID in PD [59, 61].

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

6.2.1 Dual-targeting of COMT and GSK3B

All available knowledge of promising and validated therapeutic targets of PD [47, 48, 49] was used to select a pair of targets (COMT and GSK3 β). Then the followed drug designing methodology was tested to identify target-specific and also dual-targeting inhibitors for COMT (inhibition increases the half-life of L-DOPA) and GSK3 β (inhibition prevents L-DOPA neurotoxicity and LID)(Figure 6.1).

However, the objective is to find the novel promising inhibitors that should be able to effectively block the multiple pathways for improving the motor functions of PD patients by enhancing the bioavailability of dopamine and avoiding the reported side effects of long exposure of L-DOPA [61, 59, 47].

6.2.1.1 Catechol-O-methyltransferase

Catechol-O-methyltransferase (COMT) is an important magnesium and S-adenosylmethionine (SAM) dependent enzyme, catalyzes the transfer of the methyl group of SAM to the hydroxyl group of both endogenous and exogenous catechols substrate (dopamine, norepinephrine, and epinephrine etc.) [51, 62]. COMT gene encodes two isoforms including a soluble form “S-COMT” (consists of 221 residues), and a membrane-bound form “MB-COMT”, contains an extended 50 residues at the N-terminus in human [63]. COMT has a single domain structure consisting of 8 α helices that are disposed around a central β sheet and its active site contains one S-COMT catalytic site with a conserved Mg^{2+} , important for the catalytic activity and a co-factor SAM binding site (similar to a Rossmann fold) [64].

Several studies have reported the role of COMT as a major catabolic regulator of catecholamine neurotransmitters in brain and coadministration of COMT inhibitors with L-DOPA for increasing L-DOPA half-life has become effective treatment in PD [51, 47, 65, 66]. Clinical studies have shown that COMT inhibition reduces synthesis of L-DOPA meta-

6.2 Targets selection for Parkinson's disease

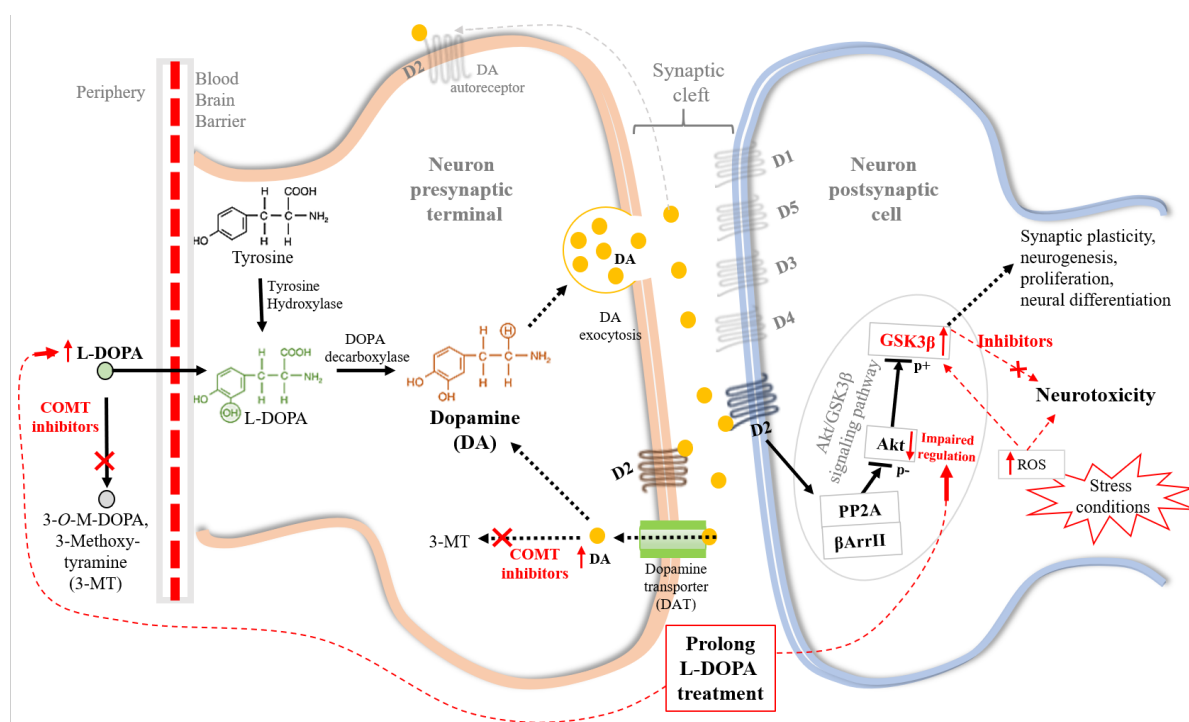


Figure 6.1: Pathway model of dual-targeting of COMT and GSK3 β in Parkinson's disease. Catechol-O-methyltransferase (COMT) inhibitors prevent peripheral metabolism of L-DOPA and enhance its availability, absorption and transport across blood–brain barrier (dashed red coloured line) in the presynaptic neuron. Increased level of L-DOPA promotes the synthesis of dopamine (DA) and thus, contributes in the enhanced availability of DA in synaptic cleft for post-synaptic neuron. Moreover, COMT inhibitors also reduce the degradation of DA, reabsorbed from the synaptic cleft through dopamine transporter (DAT). In post-synaptic neuron, DA receptor D2 activation leads to Akt inactivation in Akt/ GSK3 β signaling pathway that is dysregulated in stress conditions and L-DOPA treatment. Prolong administration of L-DOPA linked with neurotoxicity, where one possible reported factor is dysregulation of a multifunctional kinase enzyme Glycogen synthase kinase-3 beta (GSK3 β) (increased activity) due to imbalance/altered expression (decreased activity) of protein kinase B (Akt), negatively regulates GSK3 β by phosphorylation. Inactivation of GSK3 β regulates expression of different transcription factors that are involved in cell survival and activated GSK3 β can cause apoptosis (neuronal cell death). However, GSK3 β inhibitors may play important role to control neurotoxicity. L-DOPA treatment increase the release of a neurotoxin 6-hydroxydopamine (6-OHDA) that damage substantia nigra (SN) cells and oxidized form of 6-OHDA generates toxic reactive oxygen species (ROS) under oxidative stress. Green upward arrow: increased activity and green downward arrow: decreased activity, Protein phosphatase 2A (PP2A); β -arrestin-2 (β ArrII)

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

bolite 3-O-methyldopa (3-OMD) and prolong L-DOPA availability that contributes to a better absorption and blood brain barrier (BBB) transfer [51, 65, 66](Figure 6.1). However, over the last 50 years, COMT is considered as an attractive target for the development of new fast-acting anti-PD drugs to improve PD patients by delaying and preventing motor symptoms like motor fluctuations and dyskinesia [67, 65]. Tolcapone ($C_{14}H_{11}NO_5$) and entacapone ($C_{14}H_{15}N_3O_5$) are two potent COMT inhibitors reported for significant anti-PD effects. Although tolcapone is more efficacious/potent than entacapone, tolcapone is also linked with hepatotoxic side effects [68, 69]. Nonetheless, long-term follow-up studies are necessary for monitoring the non-neurological side effects of COMT inhibitors.

6.2.1.2 Glycogen synthase kinase-3B

Glycogen synthase kinase-3 (GSK3) is a multifunctional kinase enzyme responsible for the phosphorylation of glycogen synthase and regulates glycogen synthesis in glucose metabolism [70, 71]. GSK3 plays an important regulatory role in several cellular functions including signal transduction [72], division [73], differentiation, proliferation and growth [74, 75] and apoptosis [76]. GSK3 has two closely related serine/threonine (Ser/Thr) kinase isoforms GSK3 α and GSK3 β that are encoded by different genes. Both isoforms are negatively regulated by protein kinase B (Akt) mediated phosphorylation [77]. Experimental evidences have shown the role of GSK3 different process of neural development including receptors trafficking, neurogenesis, proliferation, neural differentiation, and synaptic plasticity [78, 79]. More importantly dysregulated GSK3 β is mainly considered a principal pathogenic enzyme due to its association with several neurodegenerative diseases such as Alzheimer's disease (AD), Amyotrophic lateral sclerosis (ALS), stroke, L-DOPA-induced dyskinesia and neurotoxicity in PD [61].

In PD, striatal dopamine deficiency stimulates the neuronal cells adaptive mechanisms for stabilizing the stress conditions by decreasing dopamine inactivation and increasing L-DOPA

6.2 Targets selection for Parkinson's disease

uptake, conversion to dopamine and releasing it into synaptic cleft [80]. Thus, depletion of dopaminergic neurons in PD largely disrupt the presynaptic control of dopamine release and clearance; also, L-DOPA dosing cycles generate large fluctuations in the extracellular concentration of dopamine [56]. Moreover, the sensitivity of post-synaptic dopamine receptors and their downstream signal transduction is also affected due to dopamine denervation and L-DOPA treatment, an important determinant for inducing striatal Akt/GSK3 β signalling imbalance that may lead to LID and neurotoxicity (Figure 6.1) [61]. Dopamine D2 receptor inhibits Akt by mediating a signaling complex formation consisting of protein phosphatase 2A (PP2A), β -arrestin-2 and Akt where PP2A dephosphorylate Akt during complex formation [81]. D2 mediated Akt inhibition causes activation of GSK3 β signaling. Akt/GSK3 β signaling pathways is one of the most critical pathways under dopamine impairment [81]. L-DOPA treatment or increased bioavailability of dopamine, result into a decreased activity of Akt and increased activation of GSK3 β in the striatum (causal relationship). This impaired regulation of Akt/GSK3 β signaling is aggravated under stress conditions such as increased release of a neurotoxin 6-hydroxydopamine (6-OHDA) and reactive oxygen species (ROS) formed from oxidised 6-OHDA. Stimulation of GSK3 β signaling alters the expression of its downstream substrates (transcription factors) that can induce neuronal cell apoptosis. However, GSK3 β inhibitors may be effective anti-PD drugs for controlling L-DOPA neurotoxicity [60, 61, 59].

GSK3 β contains three domains including (1) the N-terminal domain of 23–133 residues consisting of an incomplete β -barrel structure with seven antiparallel β -strands, (2) the α -helical C-terminal domain of 137–343 amino acids length, and (3) a small extra-domain subsequent to the C-terminal domain is made of 344–388 residues. GSK3 β has an ATP-binding site, magnesium-binding site and substrate-binding site [82]. ATP-binding site is the main active centre where Acceptor–Donor–Acceptor motif of Asp133–Val135 is critical for the identification of novel ATP-mimetic inhibitors possessing a complementary motif Donor–Acceptor–Donor (D–A–D) for proper inactivation of GSK3 β . In addition, Lys85

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

side chain and conserved structural water (near to back pocket) are important to make a key hydrogen bond with ligands [82, 83, 84].

6.3 Overview of virtual screening methodology

The main objective of this work is a comparative study and integration of existing VS methods to develop a robust and re-usable polypharmacology-based approach for the identification of the most promising multi-targeting hits by predicting the biological activity of millions of molecules against targets of interest. An overview of the standard VS protocol for automatic evaluation of multi-targeting drugs using QSAR modeling and molecular docking is shown in figure 6.2. The full process is divided into two sub-workflows including:

- **QSAR modeling:** QSAR-based VS workflow consists of multiple steps such as chemical data collection and curation, molecular descriptors calculation for decoding molecular structural information into a proper input data format required for the machine learning algorithm, feature selection to identify non-redundant and biologically significant set of variables using feed-forward variable selection which encompass testing of stepwise predictive models by feeding increasing set of variables. Finally, in model learning and validation phase, the optimised set of important features (predictors) is used to train final QSAR model by following an unbiased protocol of internal and external validation for model quality assessment.
- **Molecular docking:** Docking-based VS protocol implemented in the proposed drug design pipeline includes the assessment of two main steps that are critical for docking performance and success. The first step is the target structure analysis that is required to select a good quality structure, which is important for a robust predictive performance of docking studies. Target structure analysis depends on different parameters/criteria that deal with structure quality assessment, structural data pre-processing and

6.3 Overview of virtual screening methodology

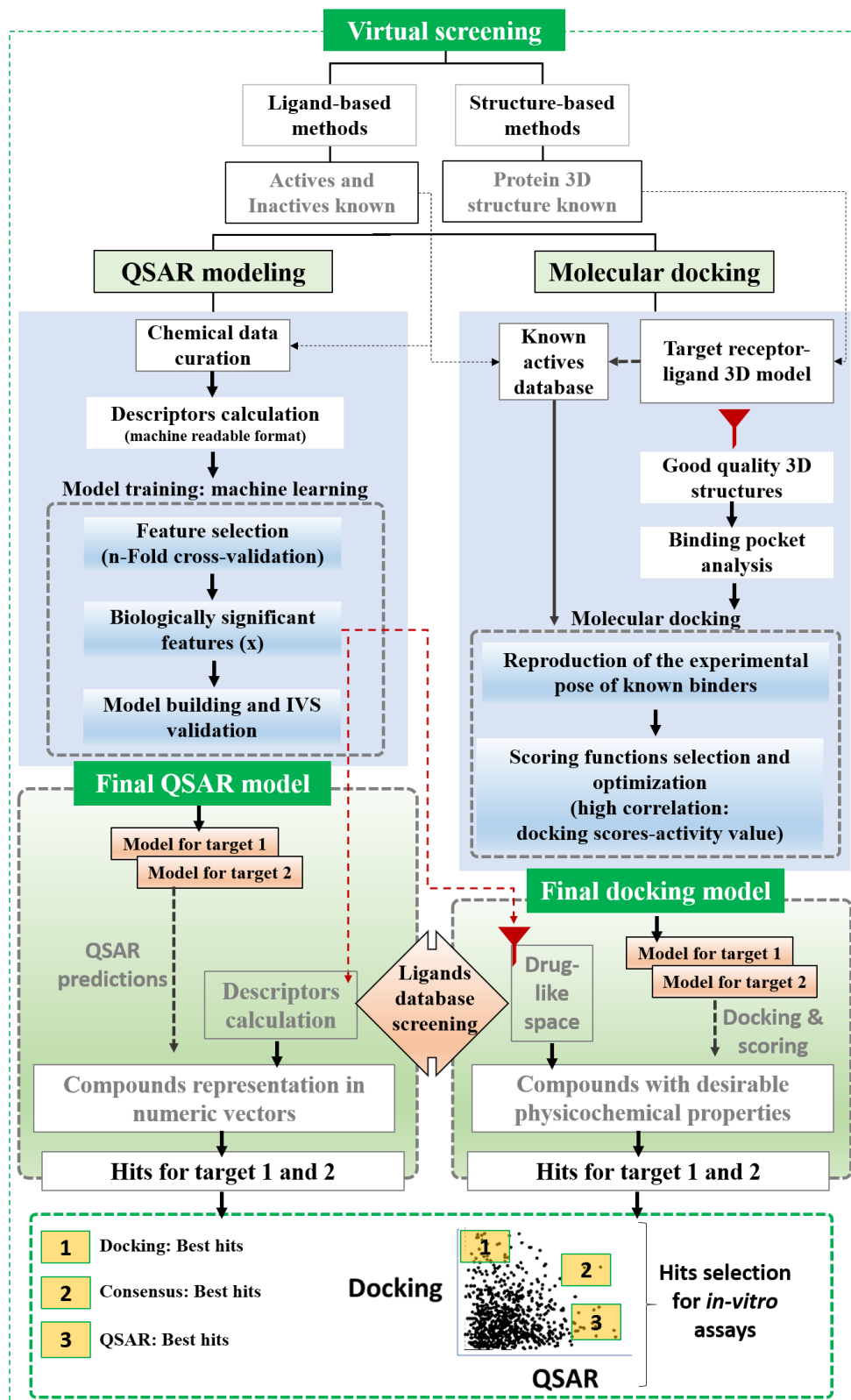


Figure 6.2: Methodology overview of QSAR and molecular docking comparative analysis.

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

cleaning, protonation states, coordinating water molecules, metal ions and co-factors, conformational states, and binding pocket analysis and selection. Second most critical docking parameter is a selection of the best scoring function that was optimised by comparing docked and X-ray ligand poses and evaluating the correlation between docking scores and biological activity.

Knowledge of both selected approaches was integrated to establish a rational drug designing methodology that was used for the screening of large compounds libraries (Figure 6.2). Resultant hits from molecular docking and QSAR were ranked and categorised into three groups: a) best-docked hits b) consensus predictions, and c) QSAR best hits. Top-ranked hits from each group was selected for further *in-vitro* validation. Each step of both docking and QSAR-based VS protocols is explained in the following sections.

6.3.1 QSAR-based virtual screening

In past 50 years, QSAR modeling [85] has been adopted as an efficient ligand-based drug designing method for predicting the biological activity/properties of new molecules. QSAR is as an application of machine learning approaches applied to derive a mathematical model, which attempts to finding an optimized correlation between molecular structure (predictors) and activity (response), which can be continuous (e.g., pIC_{50} , pEC_{50} , K_i) in regression modeling or categorical/binary (e.g., active, inactive, toxic, nontoxic) in classification problems [31, 86, 87, 88, 89]. Various machine learning algorithms including random forest (RF), support vector machines (SVM), decision trees (DT), k-nearest neighbour, naive Bayesian models, and artificial neural networks [9, 24, 25] has been developed for classification or regression problems. Since QSAR models quantitatively link the variations of the biological activity of molecules to their structural variations, these models are widely used in an initial and crucial step of lead identification and optimization. Nowadays, QSAR modeling has become a state-of-the-art method for accurate and fast VS of huge data repositories of diverse

6.3 Overview of virtual screening methodology

chemical structures [31, 24, 9, 90, 91].

The general process of QSAR modeling consists of several steps involving cheminformatics and machine learning techniques [92]. The systematized and crucial steps of QSAR model building protocol (Figure 6.2) are described as follow:

Data curation/pre-processing:

A preliminary step of data curation is a crucial task to enable reliable and rigorous QSAR model development [31, 93, 32, 92]. Several elements are considered to curate, clean and standardize chemical data including a) removing mixtures (handling of unconnected molecules) and missing data, b) splitting and eliminating salt groups, c) handling of duplicates (same experimental records: same structure and two experimental records: same structure) d) geometric optimization of the collected molecules and e) molecular structural normalization [32].

Molecular decoding/representation:

The performance of QSAR models largely depends on the relevance of the selected molecular representation (descriptors/fingerprints), describe the information (chemical features and properties) encoded in the given structures [31, 27, 92, 94, 95]. These derived structural features are used as input data (predictors) to the machine learning algorithms. For this study, a combined input dataset containing RDKit molecular descriptors and the ECF6, circular fingerprints were used for COMT and GSK3 β binary model building.

Feature selection:

Many machine learning approaches are applied to identify the most relevant structural features, responsible for the relevant pharmacological activity. This step also reduces the

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

dimensionality of the feature vector and reduce the chances of model overfitting [96, 97, 98, 99, 100, 33, 92].

Model learning:

In this step interpretable model is fitted/trained (optimal mapping between the input selected features and the activity responses) applying one or several machine learning approaches [33, 9, 101, 102, 103, 104]. Model learning phase includes model's performance internal assessment using N-fold cross-validation (CV) where dataset is randomly divided into N-parts (folds), and each part is used as a test set to verify the internal predictive power of the generated model and the remaining N-1 partitions are used as training data [92, 105]. Nonetheless, model training phase deal with the optimization of all modeling parameters and model internal validation [92]. The internal predictive performance of each model is assessed by computing the score of the Percentage of Variance Explained (PVE) and Root Mean Squared Error (RMSE) in regression modeling [106]. While classification models are evaluated using area-under-curve (AUC) [107], F-score [108] and Matthews Correlation Criterion (MCC) [109] measures.

Model validation:

This is usually the last phase of QSAR modeling process which includes a strict protocol of model validation [92]. After performing all this feature optimization in model learning phase, a final model is generated re-using the whole training data set and presented for final external validation. Model external validation is a blind prediction of the properties of unseen compounds in the external set or independent validation set (IVS) that should not be used in the process of model training. Thus, the whole data set is split into training and IVS before entering model training process in which only training data is used for model learning and internal validation. This stringent step of model external validation is helpful for evaluating

6.3 Overview of virtual screening methodology

the robustness of the generated QSAR model [92].

Accuracy or mistakes in each of these steps may affect the subsequent modeling process and thus, may cause a decrease in the overall performance of final QSAR model [92]. In this work, COMT and GSK3 β binary classification models were built by using a freely available automated QSAR modeling workflow [33], which is an extendable and highly customizable framework that follows an unbiased standard protocol to develop reliable QSAR models. Automated QSAR modeling framework [33] is only for regression problems, therefore, its several nodes were customised for classification problems in this study.

6.3.2 Molecular docking-based virtual screening

Molecular docking is an underlying and extensively studied method of structure-based VS technique. Docking is also called an *in-silico* HTS as it is a complementary tool to HTS and used to virtually placing (docking) and predicting best fitting confirmation of millions of compounds into the active/binding site of the 3D structure of a protein (intermolecular complex) [14]. All the solutions per docked molecule are evaluated based on a specified scoring scheme that measures the tightness of the fit (binding affinity) and returns the top-ranked solutions. The overall docking protocol is very straightforward, but some parameters may customise according to specific applications or interests.

The crucial factors necessary for implementing and validating molecular docking-based hit identification protocol are described as follow:

Target structure analysis:

The first main requirement of docking-based VS is the 3D structures of the target proteins [110, 111]. All the experimentally resolved structures (Nuclear magnetic resonance (NMR) or X-ray crystallography) are stored in Protein Databank (PDB: www.rcsb.org),

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

a primary 3D structural database where each 3D model has unique PDB ID. Since these 3D structures are static snapshots and have no dynamic behaviour, thus, their quality should be carefully evaluated for identifying true positive hits in the VS campaign. Following pre-processing and analytic criteria are mainly evaluated for selecting a crystallographic structure of targets before performing docking [110, 111].

Structure quality assessment: 3D structures that are produced using X-ray crystallography method, low residual factor (R-factor) and high resolution are used as indicators of overall good quality structures. R-factor has no detailed information of specific regions [111]. Another measure, B-factor (or temperature factor, or Debye-Waller factor), provides the more accurate information about the static/dynamic behaviour of each single atoms or groups of atoms and shows the erroneous regions in the structure. Lower the value of the B factor reflects more certain atomic positions and good quality model [17, 112].

Structural data pre-processing and cleaning: Structures should be carefully assessed for the several artifacts including overall conformation of the side chains and backbone, and inappropriate bond angles and lengths. Moreover, prior to docking different pre-processing steps are required to clean some elements (salts, molecules from buffer solutions, or crystallization compounds) that are often present in crystallographic structures. Also, hydrogen atoms must be added before running docking because hydrogens are not present in crystal structures. Special considerations are required for selecting target structures wild-type or mutated forms [17]. Wild-type protein 3D model is different from mutated forms where protein sequence modifications are applied for analysing the influence of sequence variations under different objectives [17, 16]

Protonation states: Protonation states and tautomers in the target structure play an important role for screening different inhibitors or binders that prefer different states [113,

6.3 Overview of virtual screening methodology

114]. Thus, if more than one key amino acids can be present in multiple states, protomer and tautomer ensembles are preferred over using only the most probable state for docking [17].

Coordinating water molecules, metal ions and co-factors: Crystal structures many times have structural features including structurally conserved water molecules, catalytic metal ions and co-factors. In common docking protocols “dry” model is used where all waters are removed from the structure. But if information is available about the possible role of water in ligand binding or in the mechanism of action of the enzyme, dry-model docking is not an optimal choice. Thus, in this case, docking programs that explicitly handle water molecules should be preferred [17, 16, 15, 2]. In the same way metals and co-factors are important for enzyme’s activity e.g., for metalloenzyme in which metal ions play a catalytic role, so these must be considered using docking programs that have been validated for the modeling of ligand-metal coordination [115, 116, 117]. Co-factor is a non-protein chemical compound (“helper molecules”) that can be bound to some receptors for assisting in biochemical transformations. Nonetheless, co-factors should be conserved with structure because their removal leaves some favourable cavities exposed for ligands, therefore, may add a bias towards docking results [17, 16, 118, 119].

Conformational states: Biological targets in solution are dynamic in nature or can adopt many conformational states including a) discrete active/inactive states, associated with proteins (kinases [120, 121], and GPCRs [122, 123] etc.) functions, b) some complicated dynamic transitions between states due to energetic components [124], c) induced fit models (holo structures) involves ligands, which induce conformational changes (stabilization of existing structures or small re-arrangements in active site) [125]. Experimental methods that are used to resolve a structure just return snapshots, a biased representation which is lacking dynamic behaviour. However, conformational variability must be considered depending on understudy goals that may be either a study of the most conserved structural regions [126]

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

or the analysis of some specific conformation [121, 17, 16]

Binding pocket analysis and selection: Binding site selection is a crucial step, which is necessary to define and limit the size of search space, a required parameter of the search algorithm in docking programs. Size of search space is critical as bigger searching boundaries not only increase the computational cost of the searching algorithm but may reduce the accuracy (high false positive rates) [17, 16, 15, 2]. In the crystal structure of ligand-receptor complex, the area occupied by small molecule represents binding pocket. However, search boundaries to run dockings encompass the region around the known binders. Blind docking (search space includes whole structure) is an alternative solution when no information of binding site is available and known structures are without bound ligand (apo-structures). Also, many binding pocket prediction tools have been developed for locating energetically favourable sites for ligand interactions [127].

Selection of best scoring function:

Docking algorithms have two basic components like search algorithm and scoring function [128, 129, 130]. Search algorithms generate “poses” (protein-ligand geometries) of the ligand within a user-defined search space (Explained earlier). While scoring function estimates binding affinity on the basis of the best position, conformation and orientation of ligand. Highest scoring hits are the ligands that are most tightly fit into the active site at minimum energy (most stable binding). Different docking software (Gold, Glide, Dock, AutoDock, FlexX, Fred, etc.) use different scoring functions to rank the most likely ligands in docking-based VS [129, 11, 2, 110, 111, 16, 17, 3, 18, 19]. Several comparative analysis of docking programs have shown that no software is the absolute best choice across all protein structures [128, 129, 130]. For any software, the following basic criteria can help to optimise scoring function and to validate the docking protocol:

6.3 Overview of virtual screening methodology

Comparison of docked and X-ray ligand poses: One of the good criteria for evaluating and ensure the robustness of VS protocol before screening the huge libraries of compounds is testing the implemented method using known binders (ligands with experimental coordinates). The idea behind is to assess if the implemented methodology is able to reproduce the experimental poses of known molecules. For this purpose, it is suggested to replace generalized scoring functions with target-specific models to be considered in the screening for validating the search parameters and best target 3D model, likely to produce better results [131]. Docking results of known molecules-target interactions serve as reference score (threshold) for ranking and evaluating the best hits [16, 17, 3, 18]. RMSD (Root-mean square deviation), a measure that quantify the similarity between an experimental and a docked poses is computed to find the best poses (RMSD < 2 Å).

Correlation between docking scores and biological activity: Correlation between predicted affinity (docking score) and the experimental bioactivities (IC_{50} , K_i , K_d , AC_{50} , or EC_{50} etc.) is commonly used criterion for evaluating the optimal choice of scoring function against each tested protein model [130, 132, 133]. Generally, Person's correlation [134] is calculated using normalized values of both experimentally determined activity e.g., negative log of inhibition or activation constants and ligand efficiency metric [135] computed from docking scores that remove bias towards larger molecules are considered for correlation analysis [136, 137].

In this work, all docking analysis were performed using GOLD 5.2 program [138], which uses a genetic algorithm for finding optimal conformations of ligand-protein interaction and evaluates poses with four available options of fitness function (ASP, ChemPLP, ChemScore and GoldScore). All the crystallographically determined protein-ligand complex structures for COMT (10 complexes: resolution between 1.3-2.8 Å) and GSK3 β (76 complexes: resolution ranges between 1.6-3.2 Å) were retrieved from PDB. Structures with resolution < 2.5 Å were tested to optimise searching parameters. Binding pocket was analysed using Molecu-

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

lar Operating Environment (MOE) tool [139] . Also, all information of the experimentally validated important ligand-protein interaction residues and other structural features (water molecules, metal ion and co-factor) for COMT [62] and GSK3 β [83] was included in docking studies. The choices for each of the above discussed docking-based VS requirements for target structure analysis and selection of best scoring function were made carefully according to the standard protocol to make the hit identification as efficient and accurate as possible.

6.3.3 Ligands database preparation for screening

Ligand database preparation for selecting the most appropriate compounds in the library to be screened is the most essential step for a successful finding of potential ligand. Nonetheless, depending on the objective of the VS several critical filters can be applied for eliminating compounds with undesirable physicochemical properties or unlikely to be active. These filters include a) sets of empirical rules (e.g., Lipinski's rule of five (RO-5) [140] and quantitative estimate of drug-likeness (QED) [141]) to define drug-like space, b) eliminating bigger compounds, inappropriate for small binding pockets and unattractive leads for optimization, c) using Pan Assay Interference Compounds (PAINS) [142] filters for removing promiscuous compounds, containing structural elements linked with toxicity or that can interfere with the pharmacological assay, d) chemical similarity cut-off for removing known actives or their highly similar molecules, and e) chemical feasibility and/or purchasability [143, 3, 17]. As ligands are treated as flexible molecules in docking programs, chemical libraries must be pre-processed for preparing correct physical states (e.g., the protomers, tautomers, and enantiomers) or for 3D geometries of ligands. Several softwares are freely available for generating conformations and other structural optimizations [144].

Publicly available National Cancer Institute (NCI) (\sim 275,000 compounds) library of the Developmental Therapeutics Program NCI/NIH (<http://dtp.nci.nih.gov>) was prepared for docking- and QSAR-based VS in this study. MOE energy minimization was

applied on full-atom ligand structures and protonated at a pH value of 7.4.

6.4 Preliminary results and discussion

6.4.1 QSAR binary classification models

GSK3 β and COMT binary classification models were build using molecules inhibitory activity (IC_{50}) classification data (compounds binding at $\leq 10 \mu\text{M}$ were labelled as active), collected from ExCAPE-DB [145], a repository of an integrated large-scale chemical dataset from publicly available databases (PubChem and ChEMBL) and comprises over 70 million molecules. The dataset used for COMT contains 191 molecules (146 actives and 45 inactives) and for GSK3 β collected data has 303,520 molecules (3334 actives and 300,186 inactives). COMT classification model was built using automated QSAR modelling framework [33] that includes data curation, descriptors calculation (RDKit descriptors and ECFP6 fingerprints), RF-based feature selection using 5-fold CV and final model building using SVM algorithm and external validation (Figure 6.2). Since GSK3 β is a complex problem having quite big data, therefore feature selection step was skipped due to high computational cost and model parameters were optimised using 5-fold CV in SVM algorithm and final model was generated using whole dataset (303,520 molecules).

The predictive performance of both QSAR classification models of COMT (IVS validation) and GSK3 β (average score of 5-fold CV) was evaluated using MCC score [109] with other well-known metrics of performance measure including recall, precision, sensitivity, specificity, and F-score [108] (Table 6.1). Both classification models were independently used for VS to predict the activity of NCI selected molecules.

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

Table 6.1: COMT and GSK3 β QSAR binary classification models results. MCC: Matthews Correlation Criterion

Target protein	Class	Recall	Precision	Sensitivity	Specificity	F-score	MCC (overall performance)
Catechol-O-methyltransferase	Active	0.94	0.94	0.94	0.82	0.94	0.76
Glycogen synthase kinase-3 beta	Active	0.74	0.94	0.74	1.00	0.83	0.83

6.4.2 Molecular docking models

To optimize the docking protocol two GOLD 5.2 scoring functions (ChemPLP and ChemScore) were tested using the X-ray structures 3BWY and 4XUE for COMT and 1Q41, 4PTE and 5F94 for GSK3 β . After several docking efforts, a combination of GOLD scoring function ChemPLP with crystallographic structure PDB entries 3BWY for COMT and 1Q41 for GSK3 β were found to be able to reproduce their corresponding experimental poses (Figure 6.3). Moreover, maximum correlation, 0.61 and 0.65 between predicted affinity (ChemPLP score) and experimental bioactivities ($\log(IC_{50})$) was observed with COMT:3BWY and GSK3 β :1Q41 docking models respectively. As COMT active site contains a catalytic site with one conserved Mg^{2+} and a co-factor SAM in the binding site [64], final docking model (PDB: 3BWY) performed best with these structural elements present and without water molecules. The binding pocket of COMT (PDB: 3BWY) was defined using about 15 Å centred from Lys144 atom “NZ” (2227). GSK3 β final docking model (PDB: 1Q41) performed well with two structurally important water molecules (positions HOH515 and HOH630 near to back pocket) using a binding pocket of about 15 Å centred from Asp133 atom “O” (1601). Both COMT:3BWY and GSK3 β :1Q41 models were independently used for predicting the binding affinities of NCI molecules.

6.4 Preliminary results and discussion

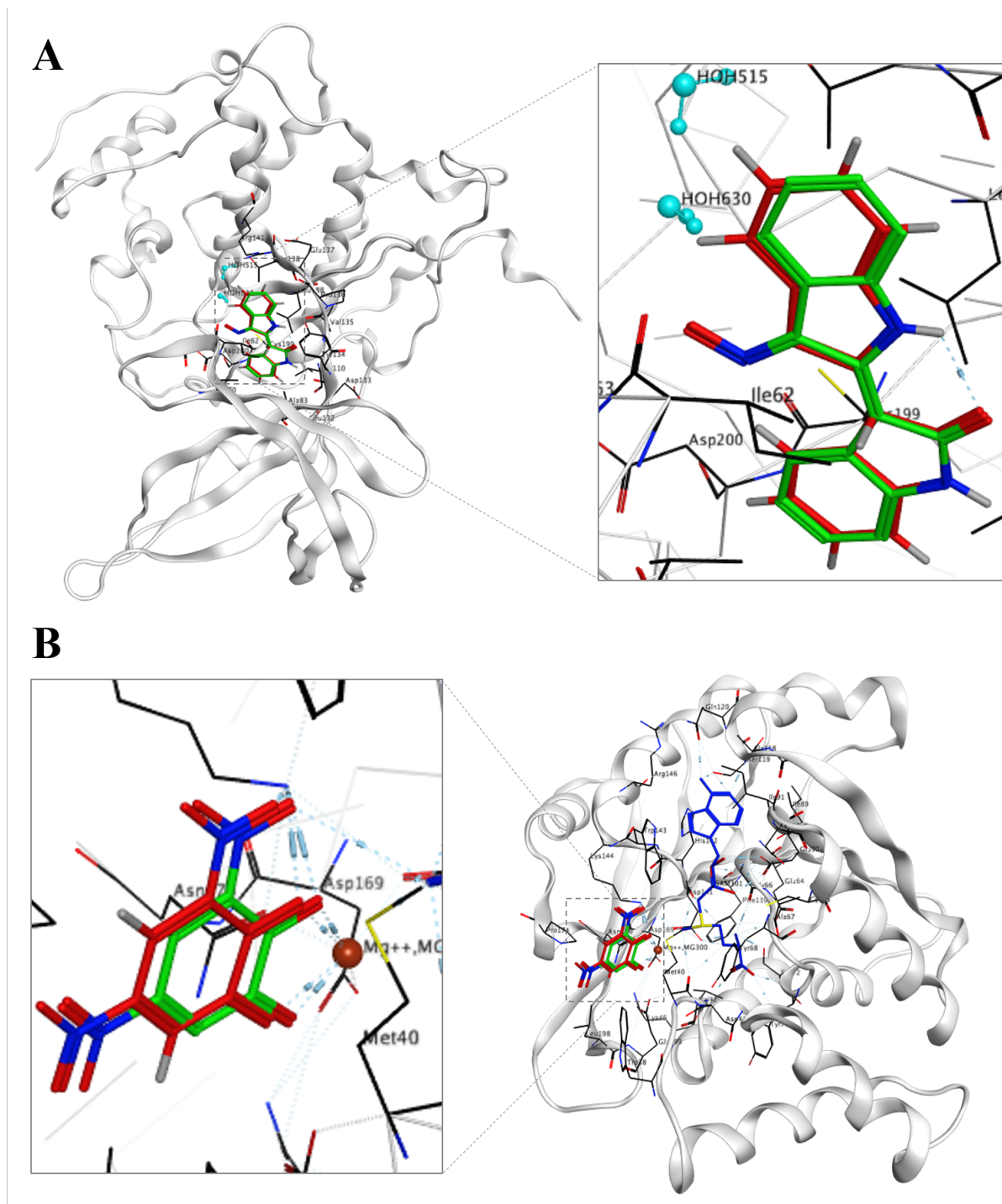


Figure 6.3: Superposition of docked (red) and experimental poses (green) of (A) GSK3 β protein complex (PDB:1Q41) and (B) COMT protein complex (PDB:3BWY)

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

6.4.3 NCI database screening and hit selection

The information of the most significant features ranked using RF-based feature selection process in QSAR modeling and other standard principles of ligand database preparation (explained in methodology) were used for defining drug-like space in NCI database. For COMT, NCI compounds were minimally filtered applying molecular descriptors ranges between 150-650 Da for molecular mass, 0-12 for *peoe_VSA10*, 0-30 for *peoe_VSA12*, 5-40 for *smr_VSA9*, and 0-6 for *slogP* and a total 184927 molecules were selected for VS. For *GSK3 β* drug-like space was defined using molecular descriptors filters including molecular mass ranges between 200-650 Da, *peoe_VSA2* \leq 28, *slogp_VSA8* \leq 38, TPSA ranges between 23-140, and 1-6 cut-off for *slogP* and a total 105851 molecules were included in VS.

QSAR- and docking-based prediction results were processed for the final ranking and selection of the best hits. This type of processing of results is known as post-processing/filtering of VS results that was enabled using a data-pipelining facility of KoNstanz Information MinEr (KNIME) [146]. KNIME is an open source data-mining framework that provides all facilities for chemical data analysis and manipulation. In the generated VS-post-processing workflow (Figure 6.4), docking results files were processed by applying QSAR model predictions and similarity scores between NCI screened molecules and training data was calculated and merged with the final output files. Similarities between molecules were calculated using Non-contiguous atom matching structure similarity (NAMS), a graph-based molecular matching algorithm [147].

Comparative analysis of QSAR modeling and molecular docking showed that both approaches, having totally different basic principles, were able to identify quite similar predicted actives. It can be seen in the presented density plots (Figure 6.5) that when a molecule predicted as active by QSAR models there is higher likelihood to have higher ChemPLP score. A majority of QSAR-based predicted actives were observed > 70 ChemPLP score.

6.4 Preliminary results and discussion

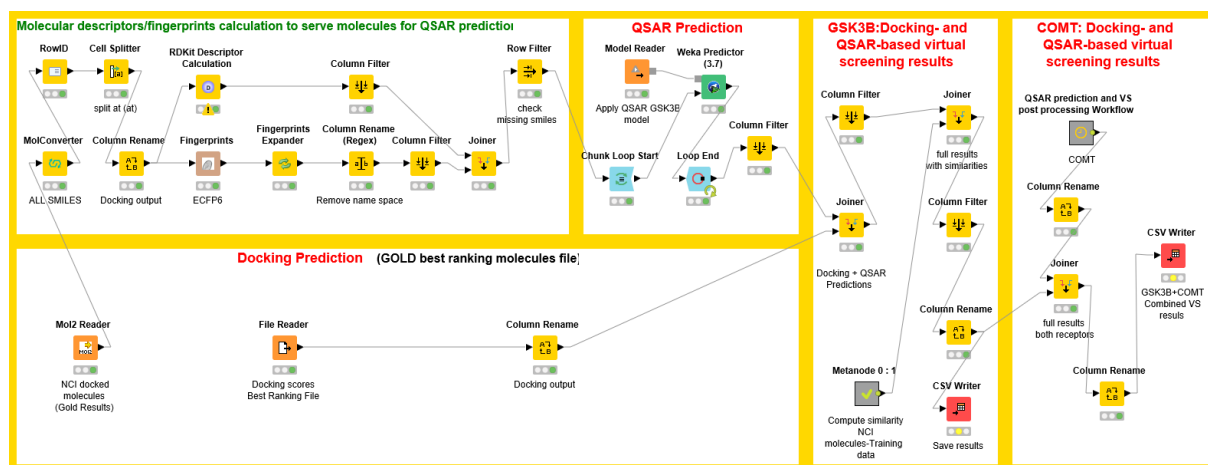


Figure 6.4: Virtual screening results post-processing workflow

Molecular docking predictions also suggested quite different hits from QSAR-based predictions; however, their combination may be helpful for identifying novel molecules (Figure 6.5).

In the case of molecular docking, predicted molecules were ranked according to their ChemPLP scores. All the molecules with a score > 70 and establishing key interactions with the residues in the binding pockets were selected as best docked hits. While in the case of QSAR, molecules predicted as “actives” were further sorted and ranked as best molecules that have probability > 90 , computed using SVM prediction function from R package e1071 [148]. But for common hits or dual-targeting compounds of both targets COMT and GSK3 β only 16 compounds were found where none of them passed the best ranking score criteria of both methods (Table 6.2) collectively. Some molecules were satisfying only docking cut-off while other were good in QSAR predictions. Thus, scoring criteria was compromised in the selection of dual-targeting compounds. Common hits sharing ≥ 0.50 NAMS similarity with the nearest neighbour in their corresponding targets training data were chosen for *in-vitro* testing (Table 6.2). In the top 16 dual-targeting compounds, the mode of the interaction of one compound NCI ID: 7434146 with both targets is shown in figure 6.6. Moreover,

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

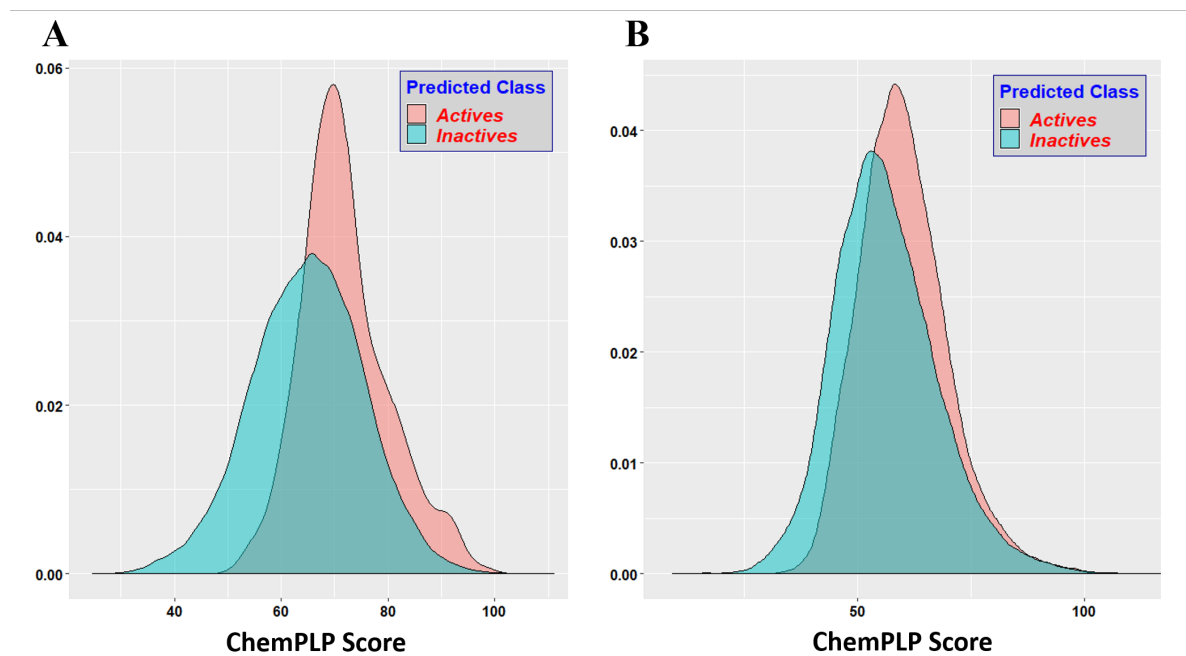
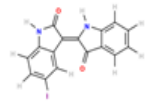
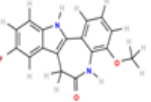
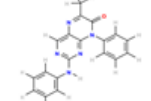
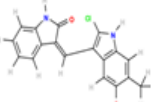
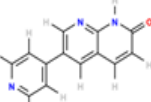
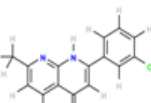
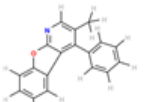
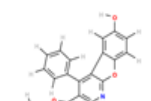
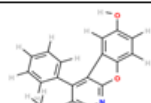


Figure 6.5: Comparative analysis of QSAR and molecular docking based predicted hits. NCI database screening results using QSAR classification and molecular docking models of (A) GSK3 β and (B) COMT

for comparing the success rate of both methods, the predicted hits specific for both targets were ranked/sorted according to three criteria including a) best-docked hits, b) consensus predictions, and c) QSAR best hits. 20 molecules from best-docked hits and QSAR top hits (ten for each) and 20 molecules from the consensus predictions were selected, however, a total 80 molecules for both targets were purchased for further testing in *in-vitro* assays.

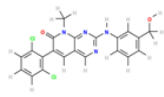
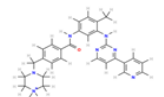
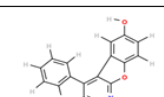
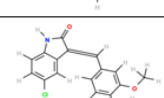
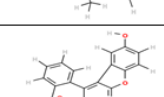
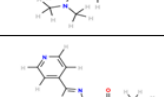
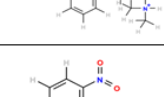
6.4 Preliminary results and discussion

Table 6.2: Dual-targeting compounds of COMT and GSK3 β . Red text: QSAR best hits, Green text: Best docked hits

NCI IDs.	Molecules	COMT: Prob. (Actives)	COMT: ChemPLP Score	COMT: Similarity with nearest neighbour in training data	GSK3B: Prob. (Actives)	GSK3B: ChemPLP Score	GSK3B: Similarity with nearest neighbour in training data
717816		0.81	50.84	0.53	0.88	77.69	0.95
702373		0.98	47.85	0.53	0.80	72.49	0.95
86929		0.95	61.53	0.55	0.86	83.93	0.84
717201		0.96	59.6	0.52	0.80	69.72	0.76
291813		0.97	49.15	0.62	1.00	61.39	0.69
679037		0.75	58.88	0.65	0.77	67.14	0.69
724552		0.82	54.01	0.62	0.87	70.91	0.68
720029		0.95	59.46	0.57	0.77	77.84	0.68
720028		0.96	54.57	0.62	0.98	72.58	0.67

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

Table 6.2: Continue...

NCI IDs.	Molecules	COMT: Prob. (Actives)	COMT: ChemPLP Score	COMT: Similarity with nearest neighbour in training data	GSK3B: Prob. (Actives)	GSK3B: ChemPLP Score	GSK3B: Similarity with nearest neighbour in training data
735424		0.78	73.24	0.54	0.98	76	0.67
743414		0.76	89.72	0.47	1.00	85.5	0.67
731012		0.82	53.82	0.60	0.93	70.55	0.66
736798		0.84	56.62	0.55	0.86	61.82	0.63
731011		0.95	58.01	0.55	0.95	72.25	0.63
678932		0.87	60.49	0.46	0.95	79.26	0.62
667710		0.91	53.95	0.49	0.81	62.68	0.60

6.4.4 In-vitro validation results of COMT specific inhibitors

Thirty-six COMT specific hits were purchased initially for *in-vitro* or prospective validation. Out of thirty-six compounds eleven were selected from best-docked hits, ten from QSAR best hits and fifteen from consensus predictions (Figure 6.7). Six compounds were found actives in *in-vitro* drug assay. A graphical abstract view of the results is shown in the figure 6.7. In the identified actives, four compounds were from consensus predictions, one

6.4 Preliminary results and discussion

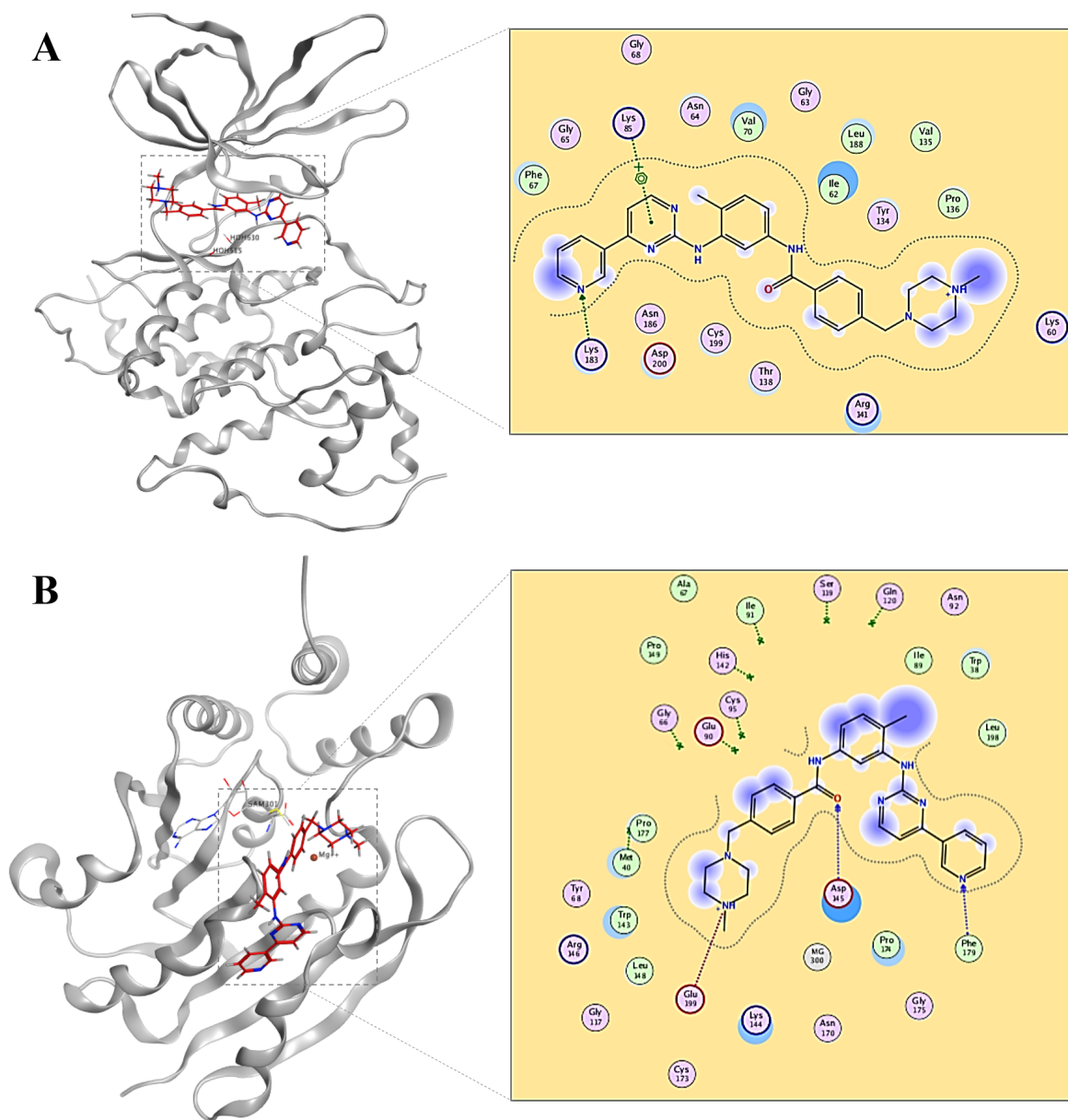


Figure 6.6: Mode of the interaction of a ligand (NCI ID: 7434146) (A) GSK3 β protein complex (PDB:1Q41) and (B) COMT protein complex (PDB:3BWY)

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

from best-docked hits and one from QSAR best hits. According to these results, integration of both QSAR modeling and molecular docking was a better approach for hits identification (Figure 6.7).

6.4.5 Future perspectives

Experimental validations would be completed in future for the rest of the selected hits including GSK3 β selective and dual-targeting compounds. Concerning comparative research of predictive performance of QSAR- and docking-based VS, a freely available larger molecular database such as ZINC [115] that contains approximately 35M high-quality ligand structures will also be screened to find more common hits. Experimental validations would be completed for all the selected hits including target selective and dual-targeting compounds. Expectedly the confirmed experimental hits will present new dual-targeting chemical scaffolds against GSK3 β and COMT targets and could be promising multi-targeting starting points for the development of new anti-PD drugs. Additionally, the comparative viewpoint of QSAR modeling and molecular docking hits rate would be a valuable support for optimising a rational drug designing pipeline. This proposed methodology can serve as enhanced polypharmacology approach to identify drug-target/s interactions with a promising potential to facilitate the process of drug discovery, drug side-effect prediction, and drug re-purposing.

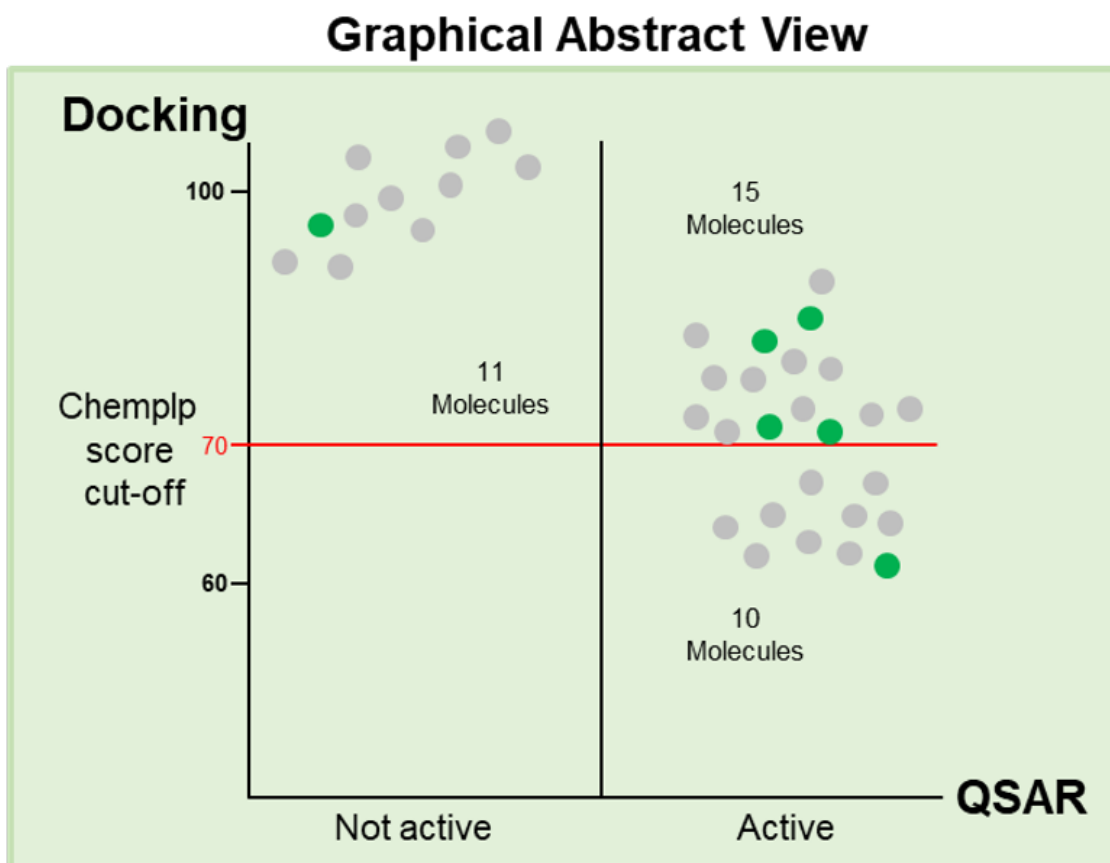


Figure 6.7: *In-vitro* validation results of COMT specific inhibitors. Green dots represent positive hits and gray dots represent negative hits

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

References

- [1] Patrick W. Walters, Matthew T. Stahl and Mark A. Murcko. ‘Virtual screening—an overview’. In: *Drug Discovery Today* 3.4 (1998), pp. 160–178. ISSN: 13596446. DOI: 10.1016/S1359-6446(97)01163-X.
- [2] Milena Lazarova. ‘Virtual Screening – Models , Methods and Software Systems’. In: *International Scientific Conference Computer Science* (2008).
- [3] Xavier Fradera and Kerim Babaoglu. ‘Overview of Methods and Strategies for Conducting Virtual Small Molecule Screening’. In: *Current protocols in chemical biology* 9.3 (2017), pp. 196–212. ISSN: 21604762. DOI: 10.1002/cpch.27.
- [4] Bryan L. Roth, Douglas J. Sheffler and Wesley K. Kroeze. ‘Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia’. In: *Nature Reviews Drug Discovery* 3.4 (Apr. 2004), pp. 353–359. ISSN: 1474-1776. DOI: 10.1038/nrd1346.
- [5] Klaus Strebhardt and Axel Ullrich. ‘Paul Ehrlich’s magic bullet concept: 100 years of progress’. In: *Nature Reviews Cancer* 8.6 (June 2008), pp. 473–480. ISSN: 1474-175X. DOI: 10.1038/nrc2394.
- [6] Jens-Uwe Peters. ‘Polypharmacology – Foe or Friend?’ In: *Journal of Medicinal Chemistry* 56.22 (Nov. 2013), pp. 8955–8971. ISSN: 0022-2623. DOI: 10.1021/jm400856t.
- [7] Balaguru Ravikumar and Tero Aittokallio. ‘Improving the efficacy-safety balance of polypharmacology in multi-target drug discovery’. In: *Expert Opinion on Drug Discovery* 13.2 (2018), pp. 179–192. ISSN: 1746045X. DOI: 10.1080/17460441.2018.1413089.

REFERENCES

- [8] Antonio Lavecchia. ‘Machine-learning approaches in drug discovery: methods and applications’. In: *Drug Discovery Today* 20.3 (Mar. 2015), pp. 318–331. ISSN: 13596446. DOI: 10.1016/j.drudis.2014.10.012.
- [9] Yu-Chen Lo et al. ‘Machine learning in chemoinformatics and drug discovery’. In: *Drug Discovery Today* 23.8 (Aug. 2018), pp. 1538–1546. DOI: 10.1016/j.drudis.2018.05.010.
- [10] Jianling Wang and Laszlo Urban. *Predictive ADMET*. Ed. by Jianling Wang and Laszlo Urban. Hoboken, NJ, USA: John Wiley & Sons, Inc., Mar. 2014. ISBN: 9781118783344. DOI: 10.1002/9781118783344.
- [11] Andrew R. Leach and Valerie J. Gillet. *An Introduction To Chemoinformatics*. Dordrecht: Springer Netherlands, 2007. ISBN: 978-1-4020-6290-2. DOI: 10.1007/978-1-4020-6291-9.
- [12] V. J. Haupt and Michael Schroeder. ‘Old friends in new guise: repositioning of known drugs with structural bioinformatics’. In: *Briefings in Bioinformatics* 12.4 (July 2011), pp. 312–326. ISSN: 1467-5463. DOI: 10.1093/bib/bbr011.
- [13] Maria Laura Bolognesi et al. ‘A perspective on multi-target drug discovery and design for complex diseases’. In: *Clinical and Translational Medicine* 7.1 (2018). ISSN: 2001-1326. DOI: 10.1186/s40169-017-0181-2.
- [14] Xuan-Yu Meng et al. ‘Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery’. In: *Current Computer Aided-Drug Design* 7.2 (June 2011), pp. 146–157. ISSN: 15734099. DOI: 10.2174/157340911795677602.
- [15] Nataraj S. Pagadala, Khajamohiddin Syed and Jack Tuszynski. ‘Software for molecular docking: a review.’ In: *Biophysical reviews* 9.2 (2017), pp. 91–102. ISSN: 1867-2450. DOI: 10.1007/s12551-016-0247-1.

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

- [16] Maria Kontoyianni. 'Docking and Virtual Screening in Drug Discovery'. In: *Proteomics for Drug Discovery: Methods and Protocols*. Ed. by Iulia M. Lazar, Maria Kontoyianni and Alexandru C. Lazar. New York, NY: Springer New York, 2017, pp. 255–266. ISBN: 978-1-4939-7201-2. DOI: 10.1007/978-1-4939-7201-2_18.
- [17] Stefano Forli. 'Charting a Path to Success in Virtual Screening'. In: *Molecules* 20.10 (Oct. 2015), pp. 18732–18758. ISSN: 1420-3049. DOI: 10.3390/molecules201018732.
- [18] Qingliang Li and Salim Shah. 'Structure-Based Virtual Screening'. In: *Methods in Molecular Biology*. 2017, pp. 111–124. DOI: 10.1007/978-1-4939-6783-4_5.
- [19] Campbell McInnes. 'Virtual screening strategies in drug discovery'. In: *Current Opinion in Chemical Biology* 11.5 (2007), pp. 494–502. ISSN: 13675931. DOI: 10.1016/j.cbpa.2007.08.033.
- [20] Johannes Kirchmair et al. 'How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information'. In: *Journal of Chemical Information and Modeling* 49.3 (Mar. 2009), pp. 678–692. ISSN: 1549-9596. DOI: 10.1021/ci8004226.
- [21] Ingo Muegge and Prasenjit Mukherjee. 'An overview of molecular fingerprint similarity search in virtual screening'. In: *Expert Opinion on Drug Discovery* 11.2 (2016), pp. 137–148. ISSN: 1746-0441. DOI: 10.1517/17460441.2016.1117070.
- [22] Robert D. Brown and Yvonne C. Martin. 'The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding'. In: *Journal of Chemical Information and Computer Sciences* 37.1 (Jan. 1997), pp. 1–9. ISSN: 0095-2338. DOI: 10.1021/ci960373c.

REFERENCES

- [23] Hongmao Sun. 'Pharmacophore-Based Virtual Screening'. In: *Current Medicinal Chemistry* 15.10 (Apr. 2008), pp. 1018–1024. ISSN: 09298673. DOI: 10.2174/092986708784049630.
- [24] Bruno J. Neves et al. 'QSAR-based virtual screening: Advances and applications in drug discovery'. In: *Frontiers in Pharmacology* 9.NOV (2018), pp. 1–7. ISSN: 16639812. DOI: 10.3389/fphar.2018.01275.
- [25] James Melville, Edmund Burke and Jonathan Hirst. 'Machine Learning in Virtual Screening'. In: *Combinatorial Chemistry & High Throughput Screening* 12.4 (May 2009), pp. 332–343. ISSN: 13862073. DOI: 10.2174/138620709788167980.
- [26] Jürgen Bajorath. *Cheminformatics: Concepts, Methods, and Tools for Drug Discovery*. Ed. by Jürgen Bajorath. Vol. 275. Methods in Molecular Biology. Totowa: Humana Press, 2004. ISBN: 978-1-58829-261-2. DOI: 10.1385/1592598021.
- [27] Roberto Todeschini and Viviana Consonni. *Molecular Descriptors for Chemoinformatics*. Ed. by Roberto Todeschini and Viviana Consonni. Methods and Principles in Medicinal Chemistry. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, July 2009. ISBN: 9783527628766. DOI: 10.1002/9783527628766.
- [28] Nina Nikolova and Joanna Jaworska. 'Approaches to Measure Chemical Similarity—a Review'. In: *QSAR & Combinatorial Science* 22.910 (2003), pp. 1006–1026. ISSN: 1611-020X. DOI: 10.1002/qsar.200330831.
- [29] Alan R. Katritzky, Victor S. Lobanov and Mati Karelson. 'QSPR: the correlation and quantitative prediction of chemical and physical properties from structure'. In: *Chemical Society Reviews* 24.4 (1995), pp. 279–87. ISSN: 0306-0012. DOI: 10.1039/cs9952400279.
- [30] Alexander Golbraikh et al. 'Data set modelability by QSAR'. In: *Journal of Chemical Information and Modeling* 54.1 (2014), pp. 1–4. ISSN: 15499596. DOI: 10.1021/ci400572x.

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

- [31] Artem Cherkasov et al. 'QSAR Modeling: Where Have You Been? Where Are You Going To?' In: *Journal of Medicinal Chemistry* 57.12 (June 2014), pp. 4977–5010. ISSN: 0022-2623. DOI: 10.1021/jm4004285.
- [32] Douglas Young et al. 'Are the chemical structures in your QSAR correct?' In: *QSAR and Combinatorial Science* 27.11-12 (2008), pp. 1337–1345. ISSN: 1611020X. DOI: 10.1002/qsar.200810084.
- [33] Samina Kausar and Andre O. Falcao. 'An automated framework for QSAR model building'. In: *Journal of Cheminformatics* 10.1 (Dec. 2018), p. 1. ISSN: 1758-2946. DOI: 10.1186/s13321-017-0256-5.
- [34] D. Fourches, E. Muratov and a. Tropsha. 'Trust but verify: on the importance of chemical structure curation in chemoinformatics and QSAR modeling research'. In: *J. Chem. Inf. Model.* 50.7 (2010), pp. 1189–1204.
- [35] Florence Stahura and Jurgen Bajorath. 'Virtual Screening Methods that Complement HTS'. In: *Combinatorial Chemistry & High Throughput Screening* 7.4 (June 2004), pp. 259–269. ISSN: 13862073. DOI: 10.2174/1386207043328706.
- [36] Ralf Mueller et al. 'Discovery of 2-(2-Benzoxazolyl amino)-4-Aryl-5-Cyanopyrimidine as Negative Allosteric Modulators (NAMs) of Metabotropic Glutamate Receptor 5 (mGlu 5): From an Artificial Neural Network Virtual Screen to an In Vivo Tool Compound'. In: *ChemMedChem* 7.3 (Mar. 2012), pp. 406–414. ISSN: 18607179. DOI: 10.1002/cmdc.201100510.
- [37] Brian K. Shoichet. 'Virtual screening of chemical libraries'. In: *Nature* 432.7019 (2004), pp. 862–865. ISSN: 0028-0836. DOI: 10.1038/nature03197.
- [38] Ralf Mueller et al. 'Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening'. In: *ACS Chemical Neuroscience* 1.4 (Apr. 2010), pp. 288–305. ISSN: 1948-7193. DOI: 10.1021/cn9000389.

REFERENCES

- [39] Thompson N. Doman et al. ‘Molecular Docking and High-Throughput Screening for Novel Inhibitors of Protein Tyrosine Phosphatase-1B’. In: *Journal of Medicinal Chemistry* 45.11 (May 2002), pp. 2213–2221. ISSN: 0022-2623. DOI: 10.1021/jm010548w.
- [40] A. L. Rodriguez et al. ‘Discovery of Novel Allosteric Modulators of Metabotropic Glutamate Receptor Subtype 5 Reveals Chemical and Functional Diversity and In Vivo Activity in Rat Behavioral Models of Anxiolytic and Antipsychotic Activity’. In: *Molecular Pharmacology* 78.6 (Dec. 2010), pp. 1105–1123. ISSN: 0026-895X. DOI: 10.1124/mol.110.067207.
- [41] Mariusz Butkiewicz et al. ‘Benchmarking Ligand-Based Virtual High-Throughput Screening with the PubChem Database’. In: *Molecules* 18.1 (Jan. 2013), pp. 735–756. ISSN: 1420-3049. DOI: 10.3390/molecules18010735.
- [42] Natasha Thorne, Douglas S. Auld and James Inglese. ‘Apparent activity in high-throughput screening: origins of compound-dependent assay interference’. In: *Current Opinion in Chemical Biology* 14.3 (June 2010), pp. 315–324. ISSN: 13675931. DOI: 10.1016/j.cbpa.2010.03.020.
- [43] Andrew Anighoro, Jürgen Bajorath and Giulio Rastelli. ‘Polypharmacology: Challenges and Opportunities in Drug Discovery’. In: *Journal of Medicinal Chemistry* 57.19 (Oct. 2014), pp. 7874–7887. ISSN: 0022-2623. DOI: 10.1021/jm5006463.
- [44] Violeta I Pérez-Nuño. ‘Using quantitative systems pharmacology for novel drug discovery’. In: *Expert Opinion on Drug Discovery* 10.12 (Dec. 2015), pp. 1315–1331. ISSN: 1746-0441. DOI: 10.1517/17460441.2015.1082543.
- [45] Alan Talevi. ‘Multi-target pharmacology: possibilities and limitations of the “skeleton key approach” from a medicinal chemist perspective’. In: *Frontiers in Pharmacology* 6.SEP (Sept. 2015), pp. 1–7. ISSN: 1663-9812. DOI: 10.3389/fphar.2015.00205.

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

- [46] Eugen Lounkine et al. 'Large-scale prediction and testing of drug activity on side-effect targets'. In: *Nature* 486.7403 (June 2012), pp. 361–367. ISSN: 0028-0836. DOI: 10.1038/nature11159.
- [47] Samuel D. Kim et al. 'Parkinson disease'. In: *Handbook of Clinical Neurology* 159 (2018), pp. 173–193. ISSN: 22124152. DOI: 10.1016/B978-0-444-63916-5.00011-2.
- [48] James B. Koprach, Lorraine V. Kalia and Jonathan M. Brotchie. 'Animal models of α -synucleinopathy for Parkinson disease drug development'. In: *Nature Reviews Neuroscience* 18.9 (2017), pp. 515–529. ISSN: 14710048. DOI: 10.1038/nrn.2017.75.
- [49] Serge Przedborski. 'The two-century journey of Parkinson disease research'. In: *Nature Reviews Neuroscience* 18.4 (2017), pp. 251–259. ISSN: 14710048. DOI: 10.1038/nrn.2017.25.
- [50] Eduardo De Pablo-Fernández et al. 'Neuroendocrine abnormalities in Parkinson's disease'. In: *Journal of Neurology, Neurosurgery & Psychiatry* 88.2 (Feb. 2017), pp. 176–185. ISSN: 0022-3050. DOI: 10.1136/jnnp-2016-314601.
- [51] Seppo Kaakkola. 'Clinical pharmacology, therapeutic use and potential of COMT inhibitors in Parkinson's disease.' In: *Drugs* 59.6 (June 2000), pp. 1233–50. ISSN: 0012-6667. DOI: 10.2165/00003495-200059060-00003.
- [52] Livia Dezsí and Laszlo Vecsei. 'Monoamine Oxidase B Inhibitors in Parkinson's Disease'. In: *CNS & Neurological Disorders - Drug Targets* 16.4 (July 2017). ISSN: 18715273. DOI: 10.2174/1871527316666170124165222.
- [53] Clas Fehling. 'Treatment of Parkinson's syndrome with L-dopa. A double blind study.' In: *Acta neurologica Scandinavica* 42.3 (Jan. 1966), pp. 367–72. ISSN: 0001-6314. DOI: 10.1111/j.1600-0404.1966.tb01188.x.

REFERENCES

- [54] Peter A LeWitt and Stanley Fahn. 'Levodopa therapy for Parkinson disease: A look backward and forward.' In: *Neurology* 86.14 Suppl 1 (Apr. 2016), S3–12. ISSN: 1526-632X. DOI: 10.1212/WNL.0000000000002509.
- [55] M. Angela Cenci. 'Presynaptic Mechanisms of L-DOPA-Induced Dyskinesia: The Findings, the Debate, and the Therapeutic Implications'. In: *Frontiers in Neurology* 5 (Dec. 2014). ISSN: 1664-2295. DOI: 10.3389/fneur.2014.00242.
- [56] M. Angela Cenci and Christine Konradi. 'Maladaptive striatal plasticity in L-DOPA-induced dyskinesia'. In: *Progress in Brain Research*. 2010, pp. 209–233. DOI: 10.1016/S0079-6123(10)83011-0.
- [57] Seong-Ho Koh et al. 'Inhibition of glycogen synthase kinase-3 reduces L-DOPA-induced neurotoxicity'. In: *Toxicology* 247.2-3 (May 2008), pp. 112–118. ISSN: 0300483X. DOI: 10.1016/j.tox.2008.02.007.
- [58] Miguel Medina, Juan Jose Garrido and Francisco G. Wandosell. 'Modulation of GSK-3 as a Therapeutic Strategy on Tau Pathologies'. In: *Frontiers in Molecular Neuroscience* 4 (2011). ISSN: 1662-5099. DOI: 10.3389/fnmol.2011.00024.
- [59] M Golpich et al. 'Glycogen synthase kinase-3 beta (GSK-3 β) signaling: Implications for Parkinson's disease'. In: *Pharmacological Research* 97 (2015), pp. 16–26. ISSN: 10436618 (ISSN). DOI: 10.1016/j.phrs.2015.03.010.
- [60] Gang Chen et al. 'Glycogen synthase kinase 3beta (GSK3beta) mediates 6-hydroxydopamine-induced neuronal death.' In: *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 18.10 (July 2004), pp. 1162–4. ISSN: 1530-6860. DOI: 10.1096/fj.04-1551fje.
- [61] Hojin Choi and Seong Ho Koh. 'Understanding the role of glycogen synthase kinase-3 in L-DOPA-induced dyskinesia in Parkinson's disease'. In: *Expert Opinion on Drug Metabolism and Toxicology* 14.1 (2018), pp. 83–90. ISSN: 17447607. DOI: 10.1080/17425255.2018.1417387.

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

- [62] K. Rutherford et al. 'Crystal Structures of Human 108V and 108M Catechol O-Methyltransferase'. In: *Journal of Molecular Biology* 380.1 (June 2008), pp. 120–130. ISSN: 00222836. DOI: 10.1016/j.jmb.2008.04.040.
- [63] I. Reenilä and P.T. Männistö. 'Catecholamine metabolism in the brain by membrane-bound and soluble catechol-o-methyltransferase (COMT) estimated by enzyme kinetic values'. In: *Medical Hypotheses* 57.5 (Nov. 2001), pp. 628–632. ISSN: 03069877. DOI: 10.1054/mehy.2001.1430.
- [64] P T Männistö and S Kaakkola. 'Catechol-O-methyltransferase (COMT): biochemistry, molecular biology, pharmacology, and clinical efficacy of the new selective COMT inhibitors.' In: *Pharmacological reviews* 51.4 (Dec. 1999), pp. 593–628. ISSN: 0031-6997.
- [65] Thomas Müller. 'Catechol-O-methyltransferase inhibitors in Parkinson's disease'. In: *Drugs* 75.2 (2015), pp. 157–174. ISSN: 11791950. DOI: 10.1007/s40265-014-0343-0.
- [66] Daniel Offen et al. 'Catechol-O-Methyltransferase Decreases Levodopa Toxicity In Vitro'. In: *Clinical Neuropharmacology* 24.1 (Jan. 2001), pp. 27–30. ISSN: 0362-5664. DOI: 10.1097/00002826-200101000-00006.
- [67] Angelo Antonini et al. 'COMT inhibition with tolcapone in the treatment algorithm of patients with Parkinson's disease (PD): relevance for motor and non-motor features.' In: *Neuropsychiatric disease and treatment* 4.1 (Feb. 2008), pp. 1–9. ISSN: 1176-6328.
- [68] Seppo Kaakkola. 'Problems with the Present Inhibitors and a Relevance of New and Improved COMT Inhibitors in Parkinson's Disease'. In: *International Review of Neurobiology*. 2010, pp. 207–225. ISBN: 9780123813268. DOI: 10.1016/B978-0-12-381326-8.00009-0.

REFERENCES

- [69] D. M. Longo et al. 'Elucidating Differences in the Hepatotoxic Potential of Tolcapone and Entacapone With DILIsym(®), a Mechanistic Model of Drug-Induced Liver Injury.' In: *CPT: pharmacometrics & systems pharmacology* 5.1 (2016), pp. 31–9. ISSN: 2163-8306. DOI: 10.1002/psp4.12053.
- [70] N Embi, D B Rylatt and P Cohen. 'Glycogen synthase kinase-3 from rabbit skeletal muscle. Separation from cyclic-AMP-dependent protein kinase and phosphorylase kinase.' In: *European journal of biochemistry* 107.2 (June 1980), pp. 519–27. ISSN: 0014-2956. DOI: 10.1111/j.1432-1033.1980.tb06059.x.
- [71] Katrina MacAulay et al. 'Glycogen synthase kinase 3alpha-specific regulation of murine hepatic glycogen metabolism.' In: *Cell metabolism* 6.4 (Oct. 2007), pp. 329–37. ISSN: 1550-4131. DOI: 10.1016/j.cmet.2007.08.013.
- [72] Shaunta Guha et al. 'Glycogen synthase kinase 3 beta positively regulates Notch signaling in vascular smooth muscle cells: role in cell proliferation and survival'. In: *Basic Research in Cardiology* 106.5 (Sept. 2011), pp. 773–785. ISSN: 0300-8428. DOI: 10.1007/s00395-011-0189-5.
- [73] J. Alan Diehl et al. 'Glycogen synthase kinase-3beta regulates cyclin D1 proteolysis and subcellular localization'. In: *Genes & Development* 12.22 (Nov. 1998), pp. 3499–3511. ISSN: 0890-9369. DOI: 10.1101/gad.12.22.3499.
- [74] Thomas Force and James R. Woodgett. 'Unique and Overlapping Functions of GSK-3 Isoforms in Cell Differentiation and Proliferation and Cardiovascular Development'. In: *Journal of Biological Chemistry* 284.15 (Apr. 2009), pp. 9643–9647. ISSN: 0021-9258. DOI: 10.1074/jbc.R800077200.
- [75] S. Shin et al. 'Glycogen synthase kinase (GSK)-3 promotes p70 ribosomal protein S6 kinase (p70S6K) activity and cell proliferation'. In: *Proceedings of the National Academy of Sciences* 108.47 (Nov. 2011), E1204–E1213. ISSN: 0027-8424. DOI: 10.1073/pnas.1110195108.

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

- [76] Piyajit Watcharasit et al. 'Glycogen Synthase Kinase-3 β (GSK3 β) Binds to and Promotes the Actions of p53'. In: *Journal of Biological Chemistry* 278.49 (Dec. 2003), pp. 48872–48879. ISSN: 0021-9258. DOI: 10.1074/jbc.M305870200.
- [77] Samantha F. Moore et al. 'Dual Regulation of Glycogen Synthase Kinase 3 (GSK3) α/β by Protein Kinase C (PKC) α and Akt Promotes Thrombin-mediated Integrin α IIb β 3 Activation and Granule Secretion in Platelets'. In: *Journal of Biological Chemistry* 288.6 (Feb. 2013), pp. 3918–3928. ISSN: 0021-9258. DOI: 10.1074/jbc.M112.429936.
- [78] Jyhyun Ahn et al. 'GSK3 β , But Not GSK3 α , Inhibits the Neuronal Differentiation of Neural Progenitor Cells As a Downstream Target of Mammalian Target of Rapamycin Complex1'. In: *Stem Cells and Development* 23.10 (May 2014), pp. 1121–1133. ISSN: 1547-3287. DOI: 10.1089/scd.2013.0397.
- [79] Woo-Yang Kim et al. 'GSK-3 is a master regulator of neural progenitor homeostasis'. In: *Nature Neuroscience* 12.11 (Nov. 2009), pp. 1390–1397. ISSN: 1097-6256. DOI: 10.1038/nn.2408.
- [80] Michael J. Zigmond et al. 'Compensations after lesions of central dopaminergic neurons: some clinical and basic implications'. In: *Trends in Neurosciences* 13.7 (July 1990), pp. 290–296. ISSN: 01662236. DOI: 10.1016/0166-2236(90)90112-N.
- [81] Jean-Martin Beaulieu et al. 'An Akt/beta-arrestin 2/PP2A signaling complex mediates dopaminergic neurotransmission and behavior.' In: *Cell* 122.2 (July 2005), pp. 261–73. ISSN: 0092-8674. DOI: 10.1016/j.cell.2005.05.012.
- [82] Masaaki Aoki et al. 'Structural insight into nucleotide recognition in tau-protein kinase I/glycogen synthase kinase 3 beta.' In: *Acta crystallographica. Section D, Biological crystallography* 60.Pt 3 (Mar. 2004), pp. 439–46. ISSN: 0907-4449. DOI: 10.1107/S090744490302938X.

REFERENCES

- [83] J.A. Bertrand et al. 'Structural Characterization of the GSK-3 β Active Site Using Selective and Non-selective ATP-mimetic Inhibitors'. In: *Journal of Molecular Biology* 333.2 (Oct. 2003), pp. 393–407. ISSN: 00222836. DOI: 10.1016/j.jmb.2003.08.031.
- [84] Guanglin Luo et al. 'Discovery of Isonicotinamides as Highly Selective, Brain Penetrable, and Orally Active Glycogen Synthase Kinase-3 Inhibitors.' In: *Journal of medicinal chemistry* 59.3 (Feb. 2016), pp. 1041–51. ISSN: 1520-4804. DOI: 10.1021/acs.jmedchem.5b01550.
- [85] Corwin Hansch and Toshio Fujita. 'p - σ - π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure'. In: *Journal of the American Chemical Society* 86.8 (Apr. 1964), pp. 1616–1626. ISSN: 0002-7863. DOI: 10.1021/ja01062a035.
- [86] A Z Dudek, T Arodz and J Galvez. 'Computational methods in developing quantitative structure-activity relationships (QSAR): a review'. In: *Comb Chem High Throughput Screen* 9.3 (2006), pp. 213–228. ISSN: 13862073. DOI: 10.2174/138620706776055539.
- [87] Corwin Hansch et al. 'Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients'. In: *Nature* 194.4824 (1962), pp. 178–180. ISSN: 0028-0836. DOI: 10.1038/194178b0.
- [88] ChangKyoo Yoo and Mohsen Shahlaei. 'The applications of PCA in QSAR studies: A case study on CCR5 antagonists.' In: *Chemical biology & drug design* (2017). ISSN: 1747-0285. DOI: 10.1111/cbdd.13064.
- [89] Roberto Todeschini and Viviana Consonni. *Handbook of Molecular Descriptors*. Vol. 11. July. Wiley-VCH Verlag GmbH, Weinheim, Germany, 2008, p. 688. ISBN: 9783527613106. DOI: 10.1002/9783527613106.

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

- [90] George Lambrinidis and Anna Tsantili-Kakoulidou. 'Challenges with multi-objective QSAR in drug discovery'. In: *Expert Opinion on Drug Discovery* 13.9 (2018), pp. 851–859. ISSN: 1746045X. DOI: 10.1080/17460441.2018.1496079.
- [91] Weilin Zhang, Jianfeng Pei and Luhua Lai. 'Computational Multitarget Drug Design.' In: *Journal of chemical information and modeling* 57.3 (2017), pp. 403–412. ISSN: 1549-960X. DOI: 10.1021/acs.jcim.6b00491.
- [92] Alexander Tropsha. 'Best practices for QSAR model development, validation, and exploitation'. In: *Molecular Informatics* 29.6-7 (2010), pp. 476–488. ISSN: 18681743. DOI: 10.1002/minf.201000061.
- [93] Igor V. Tetko et al. 'Critical assessment of QSAR models of environmental toxicity against tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection'. In: *Journal of Chemical Information and Modeling* 48.9 (2008), pp. 1733–1746. ISSN: 15499596. DOI: 10.1021/ci800151m.
- [94] Asad U Khan. 'Descriptors and their selection methods in QSAR analysis : paradigm for drug design'. In: 21.8 (2016), pp. 1291–1302.
- [95] Mati Karelson. 'Molecular Descriptors in QSAR/ QSPR'. In: March (2000), p. 35168. ISSN: 1433-7851. DOI: 10.1002/1521-3773(20010316)40:6<1136::AID-ANIE1136>3.0.CO;2-M.
- [96] Ana L. Teixeira, João P. Leal and Andre O. Falcao. 'Random forests for feature selection in QSPR models - An application for predicting standard enthalpy of formation of hydrocarbons'. In: *Journal of Cheminformatics* 5.2 (2013), p. 1. ISSN: 17582946. DOI: 10.1186/1758-2946-5-9.
- [97] Peixun Liu and Wei Long. 'Current mathematical methods used in QSAR/QSPR studies'. In: *International Journal of Molecular Sciences* 10.5 (2009), pp. 1978–1998. ISSN: 14220067. DOI: 10.3390/ijms10051978.

REFERENCES

- [98] Maykel Pérez González et al. 'Variable selection methods in QSAR: an overview.' In: *Current topics in medicinal chemistry* 8.18 (2008), pp. 1606–1627. ISSN: 15680266. DOI: 10.2174/156802608786786552.
- [99] Robin Genuer, Jean-michel Poggi and Christine Tuleau-malot. 'Variable selection using Random Forests'. In: *Pattern Recognition Letters* 31.14 (2012), pp. 2225–2236. DOI: <https://doi.org/10.1016/j.patrec.2010.03.014>.
- [100] Matthias Dehmer et al. *Statistical Modelling of Molecular Descriptors in QSAR / QSPR*. February. Weinheim, Germany: Wiley-VCH Verlag GmbH, 2012, p. 32434. ISBN: 9783527324347.
- [101] Dimitar Dobchev, Girinath Pillai and Mati Karelson. 'In silico machine learning methods in drug development'. In: *Current Topics in Medicinal Chemistry* 14.16 (2014), pp. 1913–1922. ISSN: 15680266. DOI: 10.2174/1568026614666140929124203.
- [102] Angélica Nakagawa Lima et al. 'Use of machine learning approaches for novel drug discovery.' In: *Expert opinion on drug discovery* 11.3 (2016), pp. 225–239. ISSN: 1746-045X. DOI: 10.1517/17460441.2016.1146250.
- [103] Antonio Lavecchia and Carmen Cerchia. 'In silico methods to address polypharmacology: current status, applications and future perspectives'. In: *Drug Discovery Today* 21.2 (Feb. 2016), pp. 288–298. ISSN: 13596446. DOI: 10.1016/j.drudis.2015.12.007.
- [104] Steven L Dixon et al. 'AutoQSAR: an automated machine learning tool for best-practice QSAR modeling'. In: *Future medicinal chemistry* (2016).
- [105] Alexander Tropsha and Alexander Golbraikh. 'Predictive QSAR modeling workflow, model applicability domains, and virtual screening.' In: *Current pharmaceutical design* 13.34 (2007), pp. 3494–504. ISSN: 1873-4286. DOI: 10.2174/138161207782794257.

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

- [106] Andrej-Nikolai Spiess and Natalie Neumeyer. ‘An evaluation of R² as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach.’ In: *BMC pharmacology* 10 (2010), p. 6. ISSN: 1471-2210. DOI: 10.1186/1471-2210-10-6.
- [107] Andrew P. Bradley. ‘The use of the area under the ROC curve in the evaluation of machine learning algorithms’. In: *Pattern Recognition* 30.7 (July 1997), pp. 1145–1159. ISSN: 00313203. DOI: 10.1016/S0031-3203(96)00142-2.
- [108] D. M. W. Powers. ‘Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation’. In: *Journal of Machine Learning Technologies* 2.1 (2011), pp. 37–63.
- [109] Pierre Baldi and Søren Soren Brunak. ‘Bioinformatics: The Machine Learning Approach’. In: *MIT Press* (2001), IXXI, 1452. ISSN: 0269-8889. DOI: 10.1017/S0269888904220161.
- [110] Andrew M Davis, Stephen A St-Gallay and Gerard J Kleywegt. ‘Limitations and lessons in the use of X-ray structural information in drug design.’ In: *Drug discovery today* 13.19-20 (Oct. 2008), pp. 831–41. ISSN: 1359-6446. DOI: 10.1016/j.drudis.2008.06.006.
- [111] David R Cooper et al. ‘X-ray crystallography: assessment and validation of protein–small molecule complexes for drug discovery’. In: *Expert Opinion on Drug Discovery* 6.8 (Aug. 2011), pp. 771–782. ISSN: 1746-0441. DOI: 10.1517/17460441.2011.585154.
- [112] P. Debye. ‘Interferenz von Röntgenstrahlen und Wärmebewegung’. In: *Annalen der Physik* 348.1 (1913), pp. 49–92. ISSN: 00033804. DOI: 10.1002/andp.19133480105.
- [113] Arghya Barman and Rajeev Prabhakar. ‘Protonation States of the Catalytic Dyad of β -Secretase (BACE1) in the Presence of Chemically Diverse Inhibitors: A Molecu-

REFERENCES

- lar Docking Study'. In: *Journal of Chemical Information and Modeling* 52.5 (May 2012), pp. 1275–1287. ISSN: 1549-9596. DOI: 10.1021/ci200611t.
- [114] Tímea Polgár and György M. Keserü. 'Virtual Screening for β -Secretase (BACE1) Inhibitors Reveals the Importance of Protonation States at Asp32 and Asp228'. In: *Journal of Medicinal Chemistry* 48.11 (June 2005), pp. 3749–3755. ISSN: 0022-2623. DOI: 10.1021/jm049133b.
- [115] John J. Irwin and Brian K. Shoichet. 'ZINC—a free database of commercially available compounds for virtual screening.' In: *Journal of chemical information and modeling* 45.1 (2005), pp. 177–82. ISSN: 1549-9596. DOI: 10.1021/ci049714+.
- [116] Diogo Santos-Martins et al. 'AutoDock4 Zn : An Improved AutoDock Force Field for Small-Molecule Docking to Zinc Metalloproteins'. In: *Journal of Chemical Information and Modeling* 54.8 (Aug. 2014), pp. 2371–2379. ISSN: 1549-9596. DOI: 10.1021/ci500209e.
- [117] Birte Seebeck et al. 'Modeling of metal interaction geometries for protein-ligand docking.' In: *Proteins* 71.3 (May 2008), pp. 1237–54. ISSN: 1097-0134. DOI: 10.1002/prot.21818.
- [118] Huameng Li and Chenglong Li. 'Multiple ligand simultaneous docking: orchestrated dancing of ligands in binding sites of protein.' In: *Journal of computational chemistry* 31.10 (July 2010), pp. 2014–22. ISSN: 1096-987X. DOI: 10.1002/jcc.21486.
- [119] Jon A. Read et al. 'Chloroquine Binds in the Cofactor Binding Site of Plasmodium falciparum Lactate Dehydrogenase'. In: *Journal of Biological Chemistry* 274.15 (Apr. 1999), pp. 10213–10218. ISSN: 0021-9258. DOI: 10.1074/jbc.274.15.10213.

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

- [120] Nicholas M. Levinson et al. 'A Src-Like Inactive Conformation in the Abl Tyrosine Kinase Domain'. In: *PLoS Biology* 4.5 (May 2006). Ed. by Tony Pawson, e144. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0040144.
- [121] Yi Liu and Nathanael S. Gray. 'Rational design of inhibitors that bind to inactive kinase conformations'. In: *Nature Chemical Biology* 2.7 (July 2006), pp. 358–364. ISSN: 1552-4450. DOI: 10.1038/nchembio799.
- [122] R. O. Dror et al. 'Identification of two distinct inactive conformations of the 2-adrenergic receptor reconciles structural and biochemical observations'. In: *Proceedings of the National Academy of Sciences* 106.12 (Mar. 2009), pp. 4689–4694. ISSN: 0027-8424. DOI: 10.1073/pnas.0811065106.
- [123] Paul R. Gouldson et al. 'Toward the active conformations of rhodopsin and the β 2-adrenergic receptor'. In: *Proteins: Structure, Function, and Bioinformatics* 56.1 (Apr. 2004), pp. 67–84. ISSN: 08873585. DOI: 10.1002/prot.20108.
- [124] H. Frauenfelder et al. 'A unified model of protein dynamics'. In: *Proceedings of the National Academy of Sciences* 106.13 (Mar. 2009), pp. 5129–5134. ISSN: 0027-8424. DOI: 10.1073/pnas.0900336106.
- [125] Daniel E. Koshland. 'The Key–Lock Theory and the Induced Fit Theory'. In: *Angewandte Chemie International Edition in English* 33.2324 (Jan. 1995), pp. 2375–2378. ISSN: 0570-0833. DOI: 10.1002/anie.199423751.
- [126] A. P. Kornev et al. 'Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism'. In: *Proceedings of the National Academy of Sciences* 103.47 (Nov. 2006), pp. 17783–17788. ISSN: 0027-8424. DOI: 10.1073/pnas.0607656103.
- [127] Zengming Zhang et al. 'Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction'. In: *Bioinformatics* 27.15

REFERENCES

- (Aug. 2011), pp. 2083–2088. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr331.
- [128] Douglas B. Kitchen et al. ‘Docking and scoring in virtual screening for drug discovery: methods and applications’. In: *Nature Reviews Drug Discovery* 3.11 (Nov. 2004), pp. 935–949. ISSN: 1474-1776. DOI: 10.1038/nrd1549.
- [129] Renxiao Wang, Yipin Lu and Shaomeng Wang. ‘Comparative Evaluation of 11 Scoring Functions for Molecular Docking’. In: *Journal of Medicinal Chemistry* 46.12 (June 2003), pp. 2287–2303. ISSN: 0022-2623. DOI: 10.1021/jm0203783.
- [130] Sheng-You Huang, Sam Z. Grinter and Xiaoqin Zou. ‘Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions’. In: *Physical Chemistry Chemical Physics* 12.40 (2010), p. 12899. ISSN: 1463-9076. DOI: 10.1039/c0cp00151a.
- [131] Gregory A. Ross, Garrett M. Morris and Philip C. Biggin. ‘One Size Does Not Fit All: The Limits of Structure-Based Models in Drug Discovery’. In: *Journal of Chemical Theory and Computation* 9.9 (Sept. 2013), pp. 4266–4274. ISSN: 1549-9618. DOI: 10.1021/ct4004228.
- [132] T Fatima Sapundzhi et al. ‘Determination of the relationship between the docking studies and the biological activity of δ -selective enkephalin analogues’. In: *Journal of Computational Methods in Molecular Design* 5.2 (2015), pp. 98–108.
- [133] Gregory L. Warren et al. ‘A Critical Assessment of Docking Programs and Scoring Functions’. In: *Journal of Medicinal Chemistry* 49.20 (Oct. 2006), pp. 5912–5931. ISSN: 0022-2623. DOI: 10.1021/jm050362n.
- [134] ‘Pearson Correlation’. In: *A Practical Approach to Using Statistics in Health Research*. Ed. by Adam Rowe and Mackridge Philip. Hoboken, NJ, USA: John Wiley & Sons, Inc., Apr. 2018, pp. 165–172. DOI: 10.1002/9781119383628.ch17.

6. COMPARATIVE ANALYSIS OF QSAR MODELING AND MOLECULAR DOCKING: A RATIONAL APPROACH IN POLYPHARMACOLOGY

- [135] Andrew L. Hopkins, Colin R. Groom and Alexander Alex. 'Ligand efficiency: a useful metric for lead selection'. In: *Drug Discovery Today* 9.10 (May 2004), pp. 430–431. ISSN: 13596446. DOI: 10.1016/S1359-6446(04)03069-7.
- [136] Sandro Cosconati et al. 'Virtual screening with AutoDock: theory and practice'. In: *Expert Opinion on Drug Discovery* 5.6 (June 2010), pp. 597–607. ISSN: 1746-0441. DOI: 10.1517/17460441.2010.484460.
- [137] Juan Alvarez and Shoichet Brian. *Virtual Screening in Drug Discovery*. Ed. by Juan Alvarez and Brian Shoichet. 1st Editio. Vol. 1. Drug Discovery Series. CRC Press, Mar. 2005, p. 496. ISBN: 978-0-8247-5479-2. DOI: 10.1201/9781420028775.
- [138] Gareth Jones et al. 'Development and validation of a genetic algorithm for flexible docking.' In: *Journal of molecular biology* 267.3 (Apr. 1997), pp. 727–48. ISSN: 0022-2836. DOI: 10.1006/jmbi.1996.0897.
- [139] ChemicalComputingGroupInc. 'Molecular Operating Environment (MOE)'. In: (2018).
- [140] Christopher A. Lipinski. 'Drug-like properties and the causes of poor solubility and poor permeability'. In: *Journal of Pharmacological and Toxicological Methods* 44.1 (July 2000), pp. 235–249. ISSN: 10568719. DOI: 10.1016/S1056-8719(00)00107-6.
- [141] G. Richard Bickerton et al. 'Quantifying the chemical beauty of drugs'. In: *Nature Chemistry* 4.2 (Feb. 2012), pp. 90–98. ISSN: 1755-4330. DOI: 10.1038/nchem.1243.
- [142] Jonathan B. Baell and Georgina A. Holloway. 'New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays'. In: *Journal of Medicinal Chemistry* 53.7 (Apr. 2010), pp. 2719–2740. ISSN: 0022-2623. DOI: 10.1021/jm901137j.

REFERENCES

- [143] Jürgen Bajorath. ‘Activity artifacts in drug discovery and different facets of compound promiscuity’. In: *F1000Research* 3 (Oct. 2014), p. 233. ISSN: 2046-1402. DOI: 10.12688/f1000research.5426.1.
- [144] Jean-Paul Ebejer, Garrett M. Morris and Charlotte M. Deane. ‘Freely Available Conformer Generation Methods: How Good Are They?’ In: *Journal of Chemical Information and Modeling* 52.5 (May 2012), pp. 1146–1158. ISSN: 1549-9596. DOI: 10.1021/ci2004658.
- [145] Jiangming Sun et al. ‘ExCAPE-DB: An integrated large scale dataset facilitating Big Data analysis in chemogenomics’. In: *Journal of Cheminformatics* 9.1 (2017), pp. 1–9. ISSN: 17582946. DOI: 10.1186/s13321-017-0203-5.
- [146] Michael R. Berthold et al. ‘KNIME - The Konstanz Information Miner’. In: *SIGKDD Explorations* 11.1 (2009), pp. 26–31. ISSN: 19310145. DOI: 10.1145/1656274.1656280.
- [147] Ana L Teixeira and Andre O Falcao. ‘Noncontiguous atom matching structural similarity function’. In: *Journal of Chemical Information and Modeling* 53.10 (2013), pp. 2511–2524. ISSN: 15499596. DOI: 10.1021/ci400324u.
- [148] David Meyer et al. *Misc functions of the Department of Statistics (e1071)*, TU Wien. 2014. DOI: citeulike-article-id:9958545.

7

General discussion and conclusions

7.1 Contribution

Polypharmacology-based approaches are very powerful tools in which systems biology-based knowledge of the potential targets is integrated in drug discovery for the identification of the most promising drug candidate with broader activity profiles. Selection of the right multi-targeting molecules is important to further successfully elucidate drug action mechanism in systems pharmacology studies of complex diseases. Each study conducted under this project contributed to the advancement of computational approaches to virtual screening(VS), which are adopted in polypharmacology. Such efforts can enhance the performance of multi-target drug designing. The main contributions of this work are summarized in the following sections.

7. GENERAL DISCUSSION AND CONCLUSIONS

7.1.1 Automation of quantitative structure-activity relationship

A complete automated pipeline to build quantitative structure-activity relationship (QSAR) models using state-of-the-art machine learning approaches for fast mining of chemical databases was developed in the first phase of this project. The whole process of model building involves several critical tasks including data collection and processing, appropriate data representation (descriptors and fingerprints calculation), high dimensional data handling and selection of the best predictors sufficient to predict the desired biochemical property, machine learning models fitting and unbiased validation. All these most important aspects of QSAR modeling were addressed and consistently applied in the generated automated QSAR modelling framework (Chapter 3).

The advantages of automation of repetitive tasks in the drug discovery process are numerous and include increased research quality by reducing error along with significant time saving, boosted up productivity and capacity, to name a few. Indeed, an open source customizable QSAR modeling platform to automate the laborious tasks in the QSAR modelling life cycle, is an important addition to the QSAR community, especially to researchers without extensive knowledge of modeling methodology. Following are the most distinct features and advantages of the developed QSAR modeling workflow:

- **Data access:** “Fully automated” mode automatically accesses up to date data from on-line molecules database and a “Customized” mode deals with different input data sets options. However, availability of a flexible handing of input data enables automated QSAR modeling pipeline a widely used platform.
- **Data curation:** The constructed framework covers many common needs of modelers by providing a reliable and consistent data processing method for preparing a good quality QSAR modeling data sets.
- **Prior estimate of data set modelability:** To identify difficult problems the workflow

calculates a prior estimate of the feasibility to obtain robust QSAR models by using a given data set of molecules. A prior estimation of data modelability is calculated to inform users/modelers if there is any need of additional data processing and manual curation to address different data problems. However, data quality check helps to avoid time-consuming modeling trials.

- **Feature selection and validation:** A novel feature of automated QSAR modeling workflow is automated procedure for variable selection coupled with a stringent model validation methodology.
- **Interactive prediction system:** QSAR modeling framework have a high level of automation and with default options, a reliable model can easily be build. Nonetheless, it is not a black-box prediction system as it allows modellers to control all parts of the modeling process and everything is accessible to users. Thus, it is an extensible and highly customizable tool, which provides the output of all modeling task for the diverse application, reproduction of historical predictions and updating models with new molecules as they become available.
- **Workflow performance assessment:** The performance of implemented methodology of QSAR modeling was tested on thirty datasets of different CNS therapeutic targets. The analysis of the obtained results showed that the developed variable selection procedure in automated QSAR modeling workflow was able to remove 62–99% redundant data and performed consistently with high dimensional data sets. Comparison of performance of QSAR models with and without features selection revealed that the large reduction of irrelevant variables contributed in improving model predictability and in better understanding the underlying relationship between the property of interest and the relevant features.

Moreover, automated QSAR modeling framework was also evaluated by comparing its performance to the published QSAR model. The performance of QSAR model

7. GENERAL DISCUSSION AND CONCLUSIONS

generated using fully automated procedure was significantly better than the authors QSAR modeling efforts.

In conclusion automated QSAR modeling framework was able to generate robust QSAR models without any expert interventions and advanced parameterization for the customization of complex modeling algorithms and procedures.

7.1.2 Molecular structural representation

QSAR models based on the assumption that molecular structures are mainly responsible for molecular properties, therefore, molecular structural information is considered critical to accurately predict biological activity [1, 2, 3, 4, 5]. The selection of the best molecular representation for efficiently decoding information from molecular structures into computer-readable formats is still a challenging task in cheminformatics. New numeric representation of structures is used as input data matrices to model and understand quantitative relationships between structures and biological activity in QSAR modeling.

To verify how well each molecular representation type is capable to capture the more relevant structural elements, a thorough comparative analysis of the vector and metrics space representations was conducted (Chapter 4). For a fair assessment of these methods, both vector and metric space data modeling approaches were subjected to state-of-art machine learning methods that included different dimensionality reduction methods, both from feature selection and linear dimensionality reduction, as these typically produce more robust and higher quality models. Significance of each modeling approach was estimated and results showed that, in general, there is no general gain in using metric-space based approaches for modeling, as the results are similar to using individual vectorized descriptors. Secondly, molecular fragments (fingerprint based) alone produce models that are superior to the use of specialized descriptors, yet different fingerprint models are more prone to produce better models than others; Thirdly, the graph matching similarity (NAMS) most of the times

surpassed fingerprints in model quality and this finding was further tested using challenging QSAR data sets of highly diverse and sparsely distributed compounds in molecular space (remote chemical space regions).

7.1.3 Chemical space visualization

The findings of the comparative study of molecular structural representation further assisted to extend this study for the implementation and evaluation of dimension reduction methods to represent molecular high dimensional metric spaces into reduced dimensionality for visual characterization and diversity analysis of chemical activity spaces. The aim of this work was to design a methodology to capture the highest probability regions for molecules being active within reduced metric spaces.

Under this study, a novel and reliable methodology was developed that can efficiently be used to build probabilistic surfaces of molecular activity (PSMAs) for visual characterization of molecules in molecular activity spaces (Chapter 5). The basic principle of molecular space mapping approaches based on the concept of molecular structure and activity similarity, therefore in molecular activity maps, molecules are represented into reduced metric spaces (2D projected space) where small distances between molecules represent high activity similarity. Property of each molecule in 2D activity maps is added as property of each molecule as a third dimension that generates a surface on maps

Main challenges including a) Choice of a molecular space representation, b) accuracy of dimensionality reduction (DR) methods and c) performance of the interpolation algorithm were addressed for achieving descriptive and predictive accuracy of chemical space visualization tool. The proposed molecular space mapping methodology integrates the advantages of the following state-of-the-art methods in each domain of data visual analysis:

- **Molecular representation:** In the presented methodology, NAMS-based similarity

7. GENERAL DISCUSSION AND CONCLUSIONS

was used to represent understudy data sets into metrics spaces (distance matrices). NAMS, a graph-based approach has a high discriminative power for very similar molecules over other structural or graph matching approaches. NAMS was concluded the most robust molecular representation method in the comparative analysis of molecular structural representation, a study conducted under the second objective of this thesis.

- **DR methods:** Four non-linear DR methods were tested. t-SNE applications in molecular space diversity analysis is a first effort to build activity spatial classification model using this algorithm by comparing its performance with other commonly used tools.
- **PSMA:** A Non-parametric 2D KDE function was applied for mapping the most likely activity regions (activity probability maps (PSMAs)) from sparsely distributed active and inactive compounds. Integration of KDE in molecular activities space visualization approach to compute probability density function for active and inactive molecules within 2D projected space and to generate surfaces (3D map) is a totally novel approach to build a non-parametric model with predictive properties.

The obtained results showed the reliability of the proposed methodology as all the produced PSMAs from the four data sets appeared consistent and were able to characterize active and inactive molecules in clear separate groupings. Thus, PSMAs can be served as a classification model as well as a chemical space visualization tool that can significantly contribute for the understanding of structure-activity relationships (SARs), which is important for library design, chemical classification and virtual screening in drug designing projects.

7.1.4 Polypharmacology based virtual screening

Today, one of the main interest in the modern drug discovery process is the development of fast and robust approaches for VS to predict compounds with large therapeutic profiles

(multi-targeting activity) [6, 7, 8, 9]. Thus, the aim of this study was to design a rational and re-usable polypharmacology-based VS pipeline by integrating different chemical and biological information for improving the success rate in hits identification and to find novel ligands (scaffold) that are used as a promising starting point in the drug discovery process. For this purpose, the state-of-the-art QSAR modeling methodology (designed in the first phase of this project) was further complemented with molecular docking, the most widely used structure-based VS method. Parkinson's disease (PD), a multifactorial neurodegenerative disease that involves dopaminergic degenerative process was chosen as a case study (Chapter 6).

In many previous studies, dual-targeting ligands designing has been focused as a promising tool against a complex disease like cancer for efficient targeting of tumor signalling pathways affected by abnormal mutations [10]. However, existing knowledge of PD underlying neuro-pathomechanism was thoroughly studied to derive a dual-targeting drug designing model of PD for implementing and validating the designed polypharmacology-based approach. In one of the base-line PD treatments, dopamine biosynthesis and its synaptic availability is increased for improving the motor symptoms by administrating dopamine precursor L-DOPA with COMT inhibitors, a major catabolic regulator directly involved in L-DOPA metabolism [11, 12, 13, 14]. COMT inhibitors prolong L-DOPA half-life by reducing its degradation, but long exposure of L-DOPA is reported as the main cause of neurotoxicity and L-DOPA-induced dyskinesia. Other studies of neurotoxicity in PD have shown that a kinase enzyme GSK3 β plays a main regulatory role in several processes of neural development (neurogenesis, proliferation, neural differentiation, and synaptic plasticity); thus, its dysregulation (increased activity) has been suggested as a principal pathogenic cause in neurodegenerative diseases including L-DOPA-induced dyskinesia and neurotoxicity in PD. Identification of GSK3 β inhibitors may play an important role to control neurotoxicity [15].

However, a pair of targets including COMT and GSK3 β selected to build QSAR and docking models to perform NCI database screening and to predicted novel dual-targeting

7. GENERAL DISCUSSION AND CONCLUSIONS

inhibitors. The newly identified inhibitors may provide promising scaffolds to improve the motor functions of PD patients by enhancing the bioavailability of dopamine and avoiding neurotoxicity. All the top-ranked hits from both approaches were categorized into three groups: a) best-docked hits b) consensus predictions, and c) QSAR best hits. Top-ranked hits from each group were selected to perform experimental validation for the final assessment and comparison of hits rates of molecular docking and QSAR-based VS. The presented comparative analysis of QSAR modeling and molecular docking can be an important contribution for optimizing and enhancing the predictive performance of polypharmacology-based VS and can provide a rational and re-usable drug designing pipeline. Moreover, the designed pipeline can easily be further adapted for a more complex network of several targets and anti/off-targets to achieve increased efficacy and reduced toxicity in multi-factorial diseases such as CNS disorders and cancer.

7.2 Limitations and future work

The results presented in this thesis demonstrated the contribution of each objective for the advancement of different VS approaches. Nonetheless, this work can be further developed in a number of ways:

- Automated QSAR modeling framework was only developed for QSAR regression problems. Although this framework is highly customisable and easily extendable, it would be further helpful to add automated facility to build it for QSAR classification data sets for the users lacking knowledge in machine learning. Moreover, the performance of generated framework was tested only for QSAR data sets, it would be interesting to validate its performance and customize it for other modeling datasets.
- In the second objective, an extensive analysis of molecular representation methods was conducted to assess different data analysis and modeling approaches. From this

comparative study, the best performing molecular representation NAMS was further selected to develop a new chemical space visualization approach. The main limitation of NAMS is large computational cost to calculate structural similarity between millions of molecules for actual visualization of big chemical spaces and virtual screening efforts. Thus, future efforts are required to improve NAMS execution time that may be achieved by parallelizing the calculations.

Moreover, the developed molecular space mapping method was validated using single-target activity space, which represent a tiny part of known activity spaces. In future work, the applicability domain of this activity space visualization method would be vastly increased using larger data sets of multiple-targets to develop a chemical data spatial characterization tool for VS using only structural similarity of molecules.

- One further objective concerns the comparative analysis and integration of QSAR modeling and molecular docking to develop a polypharmacology approach. Only computational part of this work is presented in this document. Experimental validations of all the selected hits including target selective and dual-targeting compounds would be completed for the final assessment of the implemented polypharmacology-based VS pipeline.

References

- [1] Artem Cherkasov et al. 'QSAR Modeling: Where Have You Been? Where Are You Going To?' In: *Journal of Medicinal Chemistry* 57.12 (June 2014), pp. 4977–5010. ISSN: 0022-2623. DOI: 10.1021/jm4004285.
- [2] A Z Dudek, T Arodz and J Galvez. 'Computational methods in developing quantitative structure-activity relationships (QSAR): a review'. In: *Comb Chem High Throughput Screen* 9.3 (2006), pp. 213–228. ISSN: 13862073. DOI: 10.2174/138620706776055539.

7. GENERAL DISCUSSION AND CONCLUSIONS

- [3] Corwin Hansch et al. ‘Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients’. In: *Nature* 194.4824 (1962), pp. 178–180. ISSN: 0028-0836. DOI: 10.1038/194178b0.
- [4] ChangKyoo Yoo and Mohsen Shahlaei. ‘The applications of PCA in QSAR studies: A case study on CCR5 antagonists.’ In: *Chemical biology & drug design* (2017). ISSN: 1747-0285. DOI: 10.1111/cbdd.13064.
- [5] Roberto Todeschini and Viviana Consonni. *Handbook of Molecular Descriptors*. Vol. 11. July. Wiley-VCH Verlag GmbH, Weinheim, Germany, 2008, p. 688. ISBN: 9783527613106. DOI: 10.1002/9783527613106.
- [6] Andrew Anighoro, Jürgen Bajorath and Giulio Rastelli. ‘Polypharmacology: Challenges and Opportunities in Drug Discovery’. In: *Journal of Medicinal Chemistry* 57.19 (Oct. 2014), pp. 7874–7887. ISSN: 0022-2623. DOI: 10.1021/jm5006463.
- [7] Eugene C. Butcher, Ellen L. Berg and Eric J. Kunkel. ‘Systems biology in drug discovery’. In: *Nature Biotechnology* 22.10 (Oct. 2004), pp. 1253–1259. ISSN: 1087-0156. DOI: 10.1038/nbt1017.
- [8] Violeta I Pérez-Nuño. ‘Using quantitative systems pharmacology for novel drug discovery’. In: *Expert Opinion on Drug Discovery* 10.12 (Dec. 2015), pp. 1315–1331. ISSN: 1746-0441. DOI: 10.1517/17460441.2015.1082543.
- [9] Alan Talevi. ‘Multi-target pharmacology: possibilities and limitations of the “skeleton key approach” from a medicinal chemist perspective’. In: *Frontiers in Pharmacology* 6.SEP (Sept. 2015), pp. 1–7. ISSN: 1663-9812. DOI: 10.3389/fphar.2015.00205.
- [10] Nulgumnalli Manjunathaiah Raghavendra et al. ‘Dual or multi-targeting inhibitors: The next generation anticancer agents’. In: *European Journal of Medicinal Chemistry* 143 (Jan. 2018), pp. 1277–1300. ISSN: 02235234. DOI: 10.1016/j.ejmech.2017.10.021.

REFERENCES

- [11] Seppo Kaakkola. 'Clinical pharmacology, therapeutic use and potential of COMT inhibitors in Parkinson's disease.' In: *Drugs* 59.6 (June 2000), pp. 1233–50. ISSN: 0012-6667. DOI: 10.2165/00003495-200059060-00003.
- [12] Samuel D. Kim et al. 'Parkinson disease'. In: *Handbook of Clinical Neurology* 159 (2018), pp. 173–193. ISSN: 22124152. DOI: 10.1016/B978-0-444-63916-5.00011-2.
- [13] Thomas Müller. 'Catechol-O-methyltransferase inhibitors in Parkinson's disease'. In: *Drugs* 75.2 (2015), pp. 157–174. ISSN: 11791950. DOI: 10.1007/s40265-014-0343-0.
- [14] Daniel Offen et al. 'Catechol-O-Methyltransferase Decreases Levodopa Toxicity In Vitro'. In: *Clinical Neuropharmacology* 24.1 (Jan. 2001), pp. 27–30. ISSN: 0362-5664. DOI: 10.1097/00002826-200101000-00006.
- [15] Hojin Choi and Seong Ho Koh. 'Understanding the role of glycogen synthase kinase-3 in L-DOPA-induced dyskinesia in Parkinson's disease'. In: *Expert Opinion on Drug Metabolism and Toxicology* 14.1 (2018), pp. 83–90. ISSN: 17447607. DOI: 10.1080/17425255.2018.1417387.

