UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Mining Cardiac Side-Effects of Known Drugs

Mestrado em Bioinformática e Biologia Computacional
Especialização em Bioinformática

Joana Barros

Dissertação orientada por:
Prof. Doutor André Osório e Cruz de Azerêdo Falcão

2015

Resumo

O canal iónico hERG é crucial para a manutenção do funcionamento do batimento cardíaco no entanto a sua suscetibilidade a uma variedade de fármacos torna-o sensível a várias de terapias comuns. Os métodos mais típicos para a verificação de inibidores do hERG baseiam-se em análises de eletrofisiologia *patch-clamp*, no entanto estas metodologias estão associadas a um elevado custo e duração. Novas abordagens para a avaliação da segurança de fármacos nas fases iniciais de desenvolvimento são cruciais. Atualmente, estudos *in silico* que permitem estabelecer relações entre moléculas e a sua resposta biológica (QSAR) são amplamente usados uma vez que permitem selecionar as propriedades moleculares cruciais para a função de um fármaco. Com esta tese propõem-se novos modelos de predição da inibição da proteína hERG construídos aplicando técnicas QSAR a um conjunto de moléculas recolhidas da literatura e da base de dados ChEMBL. De forma a complementar os modelos QSAR foi testada uma nova abordagem usando semelhança molecular calculada através da ferramenta NAMS. Vários métodos, como Florestas Aleatórias, Máquinas de Vetores de Suporte, Regressão Linear e *Least Absolute Shrinkage and Selection Operator*, para a redução de variáveis foram aplicados a diferentes coleções de descritores e *fingerprints* moleculares. O melhor método para a seleção de variáveis resultou da combinação de Regressão Linear com Máquinas de Vetores de Suporte que obteve um R^2 de 0.61. O processo de seleção de variáveis é crucial para a obtenção de bons modelos QSAR e a utilização de várias coleções de descritores e *fingerprints* moleculares permitiu selecionar o conjunto melhor adaptado ao caso em questão. O melhor modelo de predição sem a aplicação de limiares de semelhança molecular obteve um R^2 de ≈ 0.56. Os modelos QSAR complementados usando uma

abordagem baseada em semelhança estrutural para a seleção de moléculas 80% e 90% semelhantes obtiveram valores de R^2 de ≈ 0.63 e ≈ 0.37, respetivamente. Utilizando os mesmos limiares de semelhança foram adicionalmente obtidos dois novos modelos baseados na média ponderada entre o pIC50 e o valor da semelhança das moléculas selecionadas, estes modelos obtiveram R^2 de ≈ 0.52 e ≈ 0.57, respetivamente. Com esta abordagem foi possível verficar a importância da semelhança molecular para a predição do comportamento de fármacos. Modelos que utilizaram informação estrutural direta, com o uso de *fingerprints* moleculares, ou indiretamente, através seleção de moléculas estruturalmente semelhantes, obtiveram melhores resultados o que suporta a premissa de que moléculas estruturalmente semelhantes exibem características físico-químicas idênticas. O conhecimento obtido com a elaboração desta dissertação permitiu a construção de uma ferramenta *online*, hERGIP, que permite a predição de inbidores do hERG e que se encontra disponível para toda a comunidade. Esta ferramenta foi comparada com o Pred-hERG, que aplica classificação binária para determinar inibidores do hERG, e foi possível verificar que ambas apresentam um comportamento semelhante o que faz do hERGIP uma ferramenta inovadora na medida em que permite obter a predição do pIC50 para a determinação de inibidores do hERG.

Palavras Chave: Quimio-informática, human Ether-à-go-go-Related Gene, Aprendizagem Automática, Relações Estrutura-Atividade Quantitativas

Abstract

The hERG protein is crucial in maintaining the heart's normal function however, many drugs target it due to its promiscuity to a variety of structurally different molecules thus representing a therapeutic challenge. The most common approach is based on conventional patch-clamp electrophysiology which is expensive and time consuming. Viable approaches to evaluate a compounds safety are needed to be applied in the early stages of the drug development process. Studies that help establish relations between a molecule and the respective biological response (QSAR) are widely used in this type of situations. This thesis proposes new prediction methods to determine hERG inhibitors. To achieve this QSAR methodologies were applied to a collection of molecules retrieved from the literature and the ChEMBL database. A new approach using structural similarity performed with NAMS was also tested to complement the QSAR prediction models. Different variable reduction methods, such as Random Forests, Support Vector Machines an Linear Regression and Least Absolute Shrinkage Method were applied to different molecular descriptors collections and molecular fingerprints. The best variable selection method was a combination between a Linear Regression and Support Vector Machines which yield an R^2 of 0.61. The variable reduction process is a vital step for the development of good QSAR models and the application of several molecular descriptors and fingerprint collections enabled the selection of the more appropriate variables to be used. The best prediction model without using a similarity threshold achieved an R^2 of ≈ 0.56. The QSAR prediction models using a structural similarity approach to select molecules 80% and 90% similar yield R^2 values of ≈ 0.63 and ≈ 0.37. Two other prediction models using the weighted mean between molecules pIC50 an similarity value

using the same similarity thresholds returned R^2 of ≈ 0.52 for the 80% threshold and ≈ 0.57 for the 90% threshold. This research was useful to gain insight regarding the use of molecular similarity for the prediction of a drug's bioactivity. Prediction models directly or indirectly using molecular similarity for QSAR model building yield the best results supporting the assertion "similar molecules have similar features". The knowledge gained with this research enabled the construction of hERGIP, a free webtool for the prediction of hERG inhibitors. A comparison of this tool with Pred-hERG, which utilises binary classification, suggested that they performed equally, making hERGIP a novel tool since it applies regression to determine hERG inhibitors by measuring the molecule's pIC50.

Keywords: Chemoinformatics, human Ether-à-go-go-Related Gene, Machine Learning, Quantitative Structure–Activity Relationship

Resumo Alargado

A saúde da população é uma das maiores preocupações atuais, o aparecimento de novas condições médicas carentes de métodos de tratamento potenciou um crescimento exponencial na indústria farmacêutica. Apesar do crescimento da população o processo de descoberta e desenvolvimento de novos fármacos manteve-se inalterado desde 1960 e não está adaptado para as necessidades atuais. Estima-se que apenas 16% dos processos de desenvolvimento de fármacos chegam ao fim e levam à comercialização de uma nova terapia. Este processo é precedido pela descoberta de "compostos potenciais" para um determinado alvo biológico, esta é uma fase crucial que apresenta um efeito gargalo e reduz o número de compostos a serem testados na fase seguinte. A fase de desenvolvimento está associada a um elevado custo e duração devido à quantidade de testes clínicos em laboratório e no Homem portanto métodos que permitam a redução destas desvantagens são necessários. O tempo médio do processo de desenvolvimento de novos fármacos é estimado entre 9 a 13 anos, uma forma de reduzir este valor passa por restringir o número de moléculas resultantes do processo de descoberta. Métodos *in silico* podem ser aplicados a uma grande variedade de moléculas e são particularmente úteis para lidar com compostos que afetam alvos indesejados além disso permitem reduzir o número de "compostos potenciais" o que torna o processo de desenvolvimento mais eficiente. Este trabalho foca-se no canal de iónico hERG que constitui um desafio terapêutico por ser suscetível a uma grande variedade de classes de fármacos. Este canal é responsável pela condução de corrente elétrica necessária ao correto funcionamento do coração. Alterações, de natureza congénita ou adquirida, no funcionamento deste processo podem levar a arritmias cardíacas fatais. Dada a natureza desta dissertação serão apenas focadas as alterações provocadas por fármacos. O hERG apresenta características estruturais

que o tornam suscetível a uma variedade de moléculas sendo um alvo não intencional de várias terapias. Os métodos mais típicos para a verificação de inibidores do hERG baseiam-se em análises de eletrofisiologia *patch-clamp* no entanto estas metodologias estão associadas a um elevado custo e duração. A Quimio-informática é uma área na fronteira entre a Química e a Informática que utiliza técnicas computacionais para responder a questões químicas e permite estabelecer correlações que apenas são possíveis utilizando metodologias computacionais. A sua utilização na indústria farmacêutica teve um elevado impacto uma vez que engloba técnicas que facilitam os processos de descoberta de novos fármacos.

Este projeto teve como objetivo desenvolver novos modelos de predição da inbição da proteína hERG através do seu valor de pIC50. Para tal foram utilizadas técnicas que permitem estabelecer relações entre estrutura de moléculas e a sua atividade (QSAR) e foram aplicados descritores moleculares - representações matemáticas que descrevem as moléculas - calculados por diferentes ferramentas, nomeadamente CDK, RDKit e e-Dragon, assim como *fingerprints* moleculares obtidos pelo CDK e OpenBabel. Este tipo de metodologia é comum na literatura no entanto este projeto distingue-se pela aplicação de uma nova abordagem usando semelhança estrutural em conjunto com metodologias QSAR.

Como primeiro passo foram recolhidas 2719 moléculas da literatura e da base de dados CHEMBL que tinham registo do valor de inibição, obtido por técnicas laboratoriais, para o hERG, no entanto na fase inicial desta tese foram apenas utlizados 2 conjuntos com 258 e 105 moléculas de forma a selecionar o melhor método para a redução de variáveis a ser utilizado nos modelos de predição. Foram testadas Máquinas de Vetores de Suporte, Florestas Aleatórias, *Least Absolute Shrinkage and Selection Operator LASSO*) e um método com base em Regressão Linear em junção com Máquinas de Vetores

de Suporte. Testes utilizando apenas descritores moleculares permitiram concluir que Florestas Aleatórias era o melhor algoritmo para o caso das coleções de descritores do RDKit e CDK, retornando valores de R^2 de ≈ 0.56 e ≈ 0.52, respetivamente. Com a adição de *fingerprints* OpenBabel os melhores resultados foram obtidos usando o algoritmo Regressão Linear em junção com Máquinas de Vetores de Suporte e foram obtidos R^2 de ≈ 0.63 para o RDKit e *fingerprints* OpenBabel e ≈ 0.61 para o CDK e *fingerprints* OpenBabel. Os resultados da coleção de descritores e-Dragon não foram utilizados uma vez que produziram resultados bastante mais baixos do que as restantes coleções o que sugere que não são apropriados para os nossos modelos. O processo de seleção de variáveis é crucial para a obtenção de bons modelos QSAR e a utilização de várias coleções de descritores e *fingerprints* permitiu selecionar o conjunto melhor adaptado ao caso em questão. De acordo com estes resultados foram selecionados os conjuntos de descritores RDKit e CDK para a fase de validação cruzada, devido à importância dos *fingeprints* foi decidido que seriam testados os *fingerprints* OpenBabel bem como *fingerprints* CDK de modo a verificar se existiam diferenças significativas entre as duas ferramentas. Foi decido utilizar o algoritmo da Regressão Linear mas juntando-o a Florestas Aleatórias, Máquinas de Vetores de Suporte e *LASSO* de forma a verificar se existiam diferenças na magnitude dos resultados obtidos, uma vez que as Florestas Aleatórias retornaram valores de desempenho elevados quando não eram considerados *fingerprints* era esperado que ao combinar Regressão Linear e Florestas Aleatórias fossem obtidos os melhores resultados.

Para a fase de validação cruzada em *5-fold* foram utilizadas duas abordagens: *Standard*, que não tem em consideração a utilização de limiares de semelhança molecular e corresponde ao modelo FullQSAR, e a *Similarity Based*, que é dividida em dois métodos: um de acordo com a média ponderada e um segundo relacionado com a semelhança de acordo com limiares de 80% e 90% e engloba os modelos WM80, WM90, QSAR80 e QSAR90. Para a abordagem FullQSAR o melhor

resultado foi obtido com as variáveis RDKit e *fingerprints* CDK que resultou num R^2 de ≈ 0.54, para o modelo QSAR80 foi obtido um R^2 de ≈ 0.65 com os descritores RDKit e *fingerprints* OpenBabel e para o modelo QSAR90 foi obtido um R^2 de ≈ 0.67 com os descritores CDK e *fingerprints* CDK. Os modelos baseados na média ponderada retornam um R^2 de ≈ 0.54 e ≈ 0.61 para o limiar de 80% e 90%, respetivamente. Todos os resultados foram obtidos com o algoritmo Regressão Linear em junção com Florestas Aleatórias.

No caso dos modelos QSAR80 e FullQSAR e apesar do melhor modelo ter sido obtido com a coleção de descritores do RDKit foi utilizado o melhor resultado obtido com os descritores CDK que corresponde a um R^2 de ≈ 0.53 e ≈ 0.63, respetivamente, para a validação com um conjunto de moléculas externo. Esta seleção foi necessária uma vez que para a construção da ferramenta de predição online não foi possível utilizar a ferramenta RDKit. A validação com o conjunto externo retornou valores de R^2 de ≈ 0.56, ≈ 0.63 e ≈ 0.37 para os modelos FullQSAR, QSAR80 e QSAR90. Os resultados obtidos para o modelo QSAR90 foram mais baixos do que o esperado podendo estar relacionados com conjunto de moléculas usadas para treinar o modelo. Os modelos WM80 e WM90 retornaram um R^2 de ≈ 0.52 e ≈ 0.57.

Os resultados obtidos no decorrer desta tese permitiram dar ênfase às metodologias aplicadas na seleção de variáveis assim como à natureza das variáveis utilizadas. Os resultados obtidos aplicando *fingerprints* e fazendo um pré-seleção de moléculas usando limiares de semelhança molecular permitiram verificar a importância da estrutura das moléculas para o desempenho da sua função biológica. No final foi construída uma ferramenta *online* que utiliza os modelos FullQSAR, QSAR80 e WM80 e que permite identificar moléculas inibidoras do hERG através do cálculo do pIC50. O hERGIP foi comparado com uma ferramenta que aplica uma classificação qualitativa na determinação de inibidores do hERG, Pred-hERG, o que permitiu verifcar que ambas apre-

sentavam um comportamento semelhante fazendo do hERGIP uma ferramenta única que permite determinar moléculas inibidoras para o hERG através do cálculo do seu pIC50. Esta ferramenta tem o intuito de ser utilizada por toda a comunidade e potencialmente como um método de verificação da segurança de compostos para o canal hERG a ser aplicado nas fases iniciais do processo de Descoberta e Desenvolvimento de fármacos.

Acknowledgements

I would like to express my gratitude to my supervisor professor André Falcão for the useful comments, remarks and engagement through the learning process of this masters thesis. His expertise, understanding and patience added considerably to my graduate experience which enabled me to become a better researcher. I would also like to thank him for introducing me to Chemoinformatics, a topic I knew nothing about until one of our first meetings but that would quickly become my favourite research area.

Also, I like to thank João Monteiro for his time, help and patience in the development the web application. Without his willingness to work with me and to apply countless new ideas the hERGIP, this tool would not be as it is today.

I would like to thank Ana, Isa and Rafaela for their friendship and support even though they were far away and dealing with their own problems they still had time for me and for some ice cream. I would also like to acknowledge the good friends I made since I started my masters and who helped me get through two years of graduate school, Catarina, Daniela and Maria.

Finally, I would like to thank my parents and my brother. They were always supporting me and encouraging me with their best wishes.

Contents

List of Figures

List of Tables

List of Algorithms

Glossary

FullQSAR QSAR prediction model trained using the complete molecule collection. Given a molecule the model is able to predict its pIC50.

pIC50 negative logarithm for the half maximal inhibitory concentration. This value measures the effectiveness of compound inhibition towards a biological or biochemical target.

QSAR80 QSAR prediction model trained using molecules at least 80% similar. This model enables the prediction of the pIC50 of a molecule but can only be used if the molecule with the unknown pIC50 is at least 80% similar with one of the molecules utilised to build the model.

QSAR90 QSAR prediction model trained with molecules with similarity values equal or above 90%. It allows the prediction of a given compound pIC50 if it is at least 90% similar with any of the training molecules.

Similarity-based approach Two similarity threshold values, 80% and 90%, were used to select only molecules that had high similarity this enabled the reduction of the dataset to 1696 and 1323 molecules, respectively. This allowed to build more precise Train sets using the molecules structural similarity. A 5-fold Cross-Validation was applied to find the mean MSE and R^2 in order to evaluate the performance.

Standard approach The complete molecule collection was used to train the models. For each combination of descriptor collections the MSE and R^2 were determined through the mean result of each fold in a 5-fold Cross-Validation..

Tanimoto Coefficient is a statistic measure to determine the diversity of sample sets by measuring the number of chemical features that are common to both molecules compared to the number of chemical features that are in either.

WM80 model for the prediction of a molecule's pIC50 using the weighted mean of the pIC50 of molecules at least 80% similar with the similarity value as weights.

WM90 pIC50 prediction model using the weighted mean of the pIC50 of molecules at least 90% similar with the similarity value as weights.

List of Acronyms and Symbols

$\boldsymbol{R}^2$ Coefficient of Determination.

ADME Absorption, Distribution, Metabolism and Excretion.

CDK Chemistry Development Kit.

hERG human Ether-à-go-go Related Gene.

hERGIP human Ether-à-go-go Related Gene Inhibition Predictor.

IVS Independent Validation Set.

LASSO Least Absolute Shrinkage and Selection Operator.

LR Linear Regression.

MSE Mean Squared Error.

NAMS Noncontiguous Atom Matching Structural Similarity.

QSAR Quantitative Structure-Activity Relationship.

R&D Research/Discovery and Development.

RF Random Forest.

SMILES Simplified Molecular Input Line Entry System.

SVM Support Vector Machines.

Chapter 1

Introduction

With the amount of information published every day in the life sciences field we are faced with a problem regarding data retrieval and treatment. Data mining is on of the best methods to treat data collected during a period of time enabling the discovery of new information. Developments regarding the easiness to store and collect data increased the access to devices with high computational power and the development of more efficient algorithms to process information. This enabled the application of computational intensive methods to analyse data thus, increasing the predictive power of the data mining field (Mitchell, 1999). One important area of interest relates to public health which is a major concern in today's society due to the high demand for new therapies that can treat a disease or a medical condition.

1.1 Drug Research and Development

To develop a drug we need to understand and evaluate the risks and benefits of a new therapy and to achieve this a large quantity of molecules must be screened to determine their safety and effectiveness. These steps constitute the Drug Research/Discovery and Development (R&D) process (figure 1.1) which is divided into five stages that act as bottlenecks:

Drug Discovery

The discovery process is started due to the unavailability of a treatment or, the necessity of developing a new therapy to a given disease. The initial tasks include: drug target identification, target validation, Absorption, Distribution, Metabolism and Excretion (ADME) and pharmacokinects testing which enables hit/lead (i.e. promising compounds) generation. This stage in often performed in academia and gives insight to which protein or pathway, inhibition or activation will have a therapeutic effect. This stage may lead to a new target selection, since the initial one might not be available, thus, requiring further testing to find if the hit identification process should be continued. When the discovery process is complete several candidate compounds are found and the development phase is started.

Drug Development

This stage is composed of several steps that aim to develop a compound that can be used to treat the disease and that is safe for the population:

- **Phase I - Preclinical** - Evaluates the lead compound safety through *in vitro* pharmacodynamics tests with living organisms cells, animal models or with the aid of computational simulation.

- **Phase II - Clinical** - Begins with ADME and toxicity evaluation in humans and, subsequently, the approval or rejection of the candidate molecules by a health entity (Infarmed in Portugal) based on the previous test results.

- **Phase III - Drug Approval** - Drugs that pass the previous phases are submitted to an evaluation by a health entity resulting in the final approval or rejection of the drug before releasing it to the market. Since no drug is completely safe for the entire population this step takes into consideration the compound's risks and benefits.

- **Phase IV - Drug Surveillance** - After approval the pharmaceutical drug continues under monitoring, this is done to verify the general population reaction to it, thus, having a better assessment of the drug's safety.

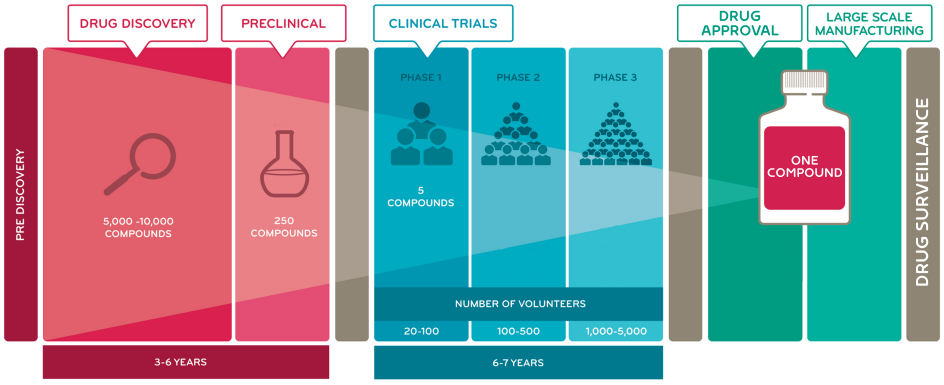

Figure 1.1: **Drug Development Process**. Each step of the process results in a reduction of suitable compounds leading or not to a selection of an appropriate therapy for the population. Based on: TRAC Services Drug Development Process

The R&D process has been roughly the same since its inception in 1960 and it is known to be long, costly and complex taking on average 9-13 years for a drug to reach the market making this process unsuitable for society's current requirements (Kaitin, 2010; Wiśniowska *et al.*, 2014). New methods that enable a faster development of safe drugs are required to speed and reduce the cost of drug R&D.

1.2 *In silico* Drug Design

The Chemical Abstract Service (CAS)[1] reports that approximately 15,000 new substances are added daily to its repository each with the potential to become a marketable pharmaceutical drug (DiMasi *et al.*, 2010; Wiśniowska *et al.*, 2014). The main problem with the R&D process is related with the high number of hit compounds generated through drug discovery. With the aid of *in silico* methods it is possible to reduce the number of compounds being tested by making the process the more cost efficient. There are several methods which can be used for *in silico* drug design, some of which include:

[1] Chemical Abstract Service, `https://www.cas.org/content/chemical-substances/faqs` (accessed March 2015)

- **Homology Modelling** - Modeling of a protein's 3D structure using, as template, an experimentally determined structure of a homologous protein. This is useful as it allows for the study of protein function and interactions.
- **Molecular Docking** - Predicts the binding between two molecules.
- **Virtual Screening** - Evaluates molecular libraries based on the potential of a molecule to bind to specific sites on target molecules.

Another important method used in computational drug design is the Quantitative Structure-Activity Relationship (QSAR) (Morris & Lim-Wilby, 2008; Wadood *et al.*, 2013), this method is applied in this thesis and it will be described in detail in Chapter 4.

These methodologies are applied to a wide variety of molecules and are particularly useful when dealing with drugs that affect unwanted targets. This work focuses on an ionic channel which is susceptible to multiple drug classes thus constituting a therapeutic challenge.

1.2.1 *Human Ether-à-go-go Related Gene*

The human Ether-à-go-go Related Gene (hERG) potassium channel is a transmembranar protein encoded by the KCNH2 gene. This channel is responsible for the conduction of the electric current necessary for the heart's normal function and the malfunction of this process may lead to cardiac arrhythmic events (Zhang *et al.*, 2014). These malfunctions can be congenital or acquired, namely through the action of pharmaceutical compounds and it is estimated that between 40% and 70% of drug-like molecules inhibit the hERG channel (Witchel, 2011).

The hERG channel is characterised by unique structural features which allow for a promiscuous binding of small molecules, therefore, the channel is of pharmacological interest due to the variety of drugs that block it with potentially fatal consequences.

Different inhibition methods for the hERG channel, namely direct blockage and channel trafficking inhibition, have been proposed. Some drugs bind to the channel and inactivate it while others inhibit the protein's trafficking to the membrane (Dempsey *et al.*, 2014). The most effective methods for prevention of

unintentional drug inhibition are through the removal of drugs that are linked to arrhythmic events from the marketplace or through usage restriction (Frolov *et al.*, 2011; Huang *et al.*, 2010). Due to this, an increase in research to find novel methods for the early assessment of hERG inhibition is crucial to a safer drug development and for a better understanding of this channel's inhibition.

1.3 Motivation

The hERG protein plays an important role in the heart's correct function and it is susceptible to drug blocking which leads to, possibly fatal, cardiac arrhythmias (Wang S. & T., 2013). Given this protein's importance it is recognized as a primary antitarget in the screening for drug candidates. However, despite of the current routinely use of hERG inhibition screens during the drug development process this ionic channel still suffers from inadvertent molecular blocking causing the removal of several drugs from the market. Methods that minimize these risks are, thus, needed for the production of safer medical treatments.

In silico approaches are a suitable alternative to apply in this situation and have been used to treat chemical data regarding molecules which inhibit the ionic channel (Braga *et al.*, 2014; Thai & Ecker, 2008; Wang *et al.*, 2013). These approaches aim to improve the Drug R&D process by reducing its cost and duration and also enabling the development of safer therapies (Bharath *et al.*, 2011).

1.4 Objectives

New methodologies must be implemented to address the assessment of the hERG channel inhibition. The main objective of this work is, therefore, to develop new prediction models to determine hERG inhibition measured as its pIC50. To achieve this goal three main objective were devised:

1. Selection of a variable reduction method;

2. Application of QSAR models using a variety of algorithms;

3. Molecular similarity incorporation in the development of QSAR prediction models. This approach was based on the assumption that a prediction model built from a dataset of structural similar molecules would help to reduce the bias present in other approaches and improve the prediction accuracy resulting in more reliable inhibition prediction result.

These methodologies are intended to be applied in the early stages of the development process, allowing the removal of hERG inhibitory molecules which would reduce the number of candidate compounds, therefore, reducing the cost and duration of the R&D process.

1.5 Contributions

The main contributions of this work are:

Contribution 1: Development of a new variable selection approach.

Contribution 2: New insight regarding the importance of molecular structural similarity for the determination of a compounds bioactivity.

Contribution 3: Collaborative of a web application for the assessment of hERG inhibition which is available at `http://hergip.lasige.di.fc.ul.pt`.

1.6 Thesis Structure

This thesis is structured into eight chapters. The first chapter introduces the difficulties with the Drug R&D process, the application of *in silico* methods for dug design and it also introduces the hERG protein, these are the foundations for this work. It also presents the objectives and motivation behind the work performed.

The second chapter states the concepts related to Chemoinformatics which are necessary for understanding this thesis.

In the third chapter it is presented the research performed by other authors which was crucial to acknowledge the methodologies already implemented and

tested to solve the hERG inhibition assessment issue. This information enabled the focus on key points which were considered in the work developed.

The fourth chapter focuses on the tools and resources applied in the approaches tested.

Chapter 5 details the methods implemented and the sixth chapter presents the results from exploratory approaches and from the application of the methodology previously stated.

Chapter 7 references the tool developed from the prediction models built with this thesis. It also provides a comparison with another tool using a small sample of molecules.

The final chapter corresponds to the conclusion which takes into consideration the knowledge discovered with this work and focuses on the problems introduced in the first chapter.

This thesis is finalised with appendices regarding the algorithm from the proposed approach for variable selection, the Python script used to retrieve molecules from ChEMBL and the list of molecules and result table from Chapter 8 tool comparison.

Chapter 2

Concepts

Chemoinformatics is a vast field which contains a specific vocabulary allowing for a unique description of methods and data representations commonly applied. This chapter will introduce the Chemoinformatics field and the concepts necessary for a correct reading and understanding of this thesis.

2.1 Chemoinformatics

Chemoinformatics was first defined by Frank Brown as "*... transforming data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization.*" (Bajorath, 2004). Since then, this relatively new discipline has been defined in a broader sense as "*... the application of informatics methods to solve chemical problems.*" (Gasteiger & Funatsu, 2006).

Chemistry produces a large and diverse amount of data and we need suitable ways of storing, accessing and retrieving that information. One way to store this chemical information is trough the use of databases, some examples are the DrugBank, ChemSpider, ChEBI, ChEMBL and CAS. However, there is a great discrepancy between the amount of data available the the number of molecules currently known. Many of the problems in chemistry are complex. The relationships between the structure of a compound and its biological activity or the influence of reaction conditions on chemical reactivity are some of them. All these problems require novel approaches using Chemoinformatics methods that allow

the management of large amounts of chemical structures and data, knowledge extraction from data and modelling of complex relationships (Gasteiger & Funatsu, 2006).

2.2 Molecular Representation

Molecules are complex structures commonly represented as images, however, this approach has limitations specially when computational methods are in use. Several methods for molecular representation have been proposed some of which became widely used due to their simplicity and implementation in several Chemoinformatics tools (Leach & Gillet, 2003). This section will focus on the methods used in this thesis to represent molecules and their properties.

2.2.1 SMILES

The Simplified Molecular Input Line Entry System (SMILES) is a line notation for entering and representing molecules and reactions. It is widely used as a general-purpose chemical nomenclature and data exchange format. However, SMILES differs in several fundamental ways from most chemical formats. SMILES represents a valence model of a molecule which uses a graph where the nodes are atoms and the edges are bonds to represent a molecule, this representation has limitations and it is not suitable for representing complex mixtures. SMILES was the main molecular representation method chosen to be used in this thesis.

2.2.1.1 SMILES Specification Rules

SMILES notation consists of a series of characters containing no spaces. Hydrogen atoms may be omitted or included. There are six generic SMILES encoding rules corresponding to specification of atoms, bonds, branches, rings, aromaticity and disconnected structures. An example for the SMILES conversion of the aspirin molecule is given in figure 2.1.

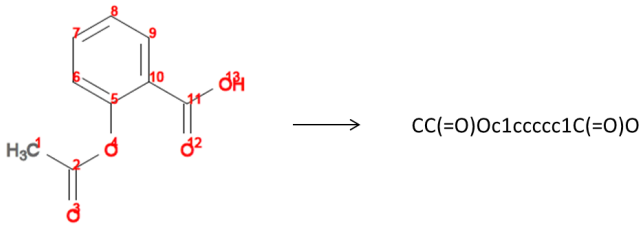

Figure 2.1: **SMILES Notation for Aspirin**. The numbers in the 2D molecular structure represent the index of each atom and the correspondent SMILES notation starts with the atom with index 1.

Atoms

- Atoms are represented by their atomic symbols, each non-hydrogen atom is specified independently by its atomic symbol enclosed in square brackets and the second letter of two-character symbols must be entered in lower case;
- Elements in the "organic subset" (i.e. B, C, N, O, P, S, F, Cl, Br, and I) can be written without brackets if the number of attached hydrogen's conforms to the lowest normal valence is consistent with explicit bonds;
- Atoms in aromatic rings are specified by lower case letters.

Bonds

- Single, double, triple, and aromatic bonds are represented by the symbols -, =, #, and : ;
- Adjacent atoms are assumed to be connected to each other by a single or aromatic bond which can be always omitted.

Branches

- Specified by enclosing them in parentheses;
- Can be nested or stacked.

Rings

The bonds are numbered in any order, designating ring opening (or ring closure) bonds by a digit immediately following the atomic symbol at each ring closure.

Aromaticity

Can be represented using alternating single and double bonds with uppercase symbols for the atoms.

Disconnected Structures

The dot '.' symbol (also called a "dot bond") is present where a bond symbol would occur but indicates that the atoms are not bonded. The most common use of the dot-bond symbol is to represent disconnected and ionic compounds.

2.2.1.2 Canonical SMILES

A canonical SMILES follows the rules previously mentioned but always writes the atoms and bonds of any particular molecule in the exact same order, regardless of the source of the molecule. This is very useful in Chemoinformatics since a given molecule will always yield the same SMILES string, allowing a chemical database system to:

- Create a unique SMILES for each molecule in the system;
- Consolidate data about one molecule from a variety of sources into a single record;
- Give a unique identifier to each molecule in a database.

Still, a canonical SMILES should not be considered a universal, global identifier, two systems that produce a canonical SMILES may use different rules in their algorithm or the same system may be improved thus changing the SMILES it produces (Daylight Chemical Information Systems, 2011).

2.2.2 Molecular Descriptors

Molecular descriptors are mathematical representations of a molecule and can be divided into two main classes: experimental (e.g. logP, molar refractivity) and theoretical molecular descriptors. These two classes differ mainly in the statistical error associated to them, theoretical descriptors are not affected by experimental errors contrary to experimental descriptors. Some theoretical descriptors, such as surface area and volume related descriptors have an experimental counterpart, showing therefore, a natural overlap with the experimental measurements. The greatest advantage of this class of descriptors is related to the cost, time and availability, in comparison to the experimental descriptors that require specialized equipment and technicians. Molecular descriptors fragment the knowledge from a molecule into several small parts. The bigger the number of the descriptors used, supposedly, the more information we can extract from that chemical substance. Also it must be taken into account that different tools may return different results for the same molecular descriptor (Todeschini *et al.*, 2008). Consequently the number of descriptors that Chemoinformatics tools can calculate is consistently growing, currently some of these can calculate as much as 4885 different descriptors (Talete SRL, 2013).

Due to their advantages, molecular descriptors have become very important variables for establishing relations between molecular structure, properties and biological activity, being, therefore, important for QSAR studies and molecular modelling (Puzyn *et al.*, 2010).

2.2.3 Molecular Fingerprints

Molecular fingerprints are bit string representations of molecules. This bit representation can be achieved using two different approaches:

- **Fragment based fingerprints** - Each bit corresponds to a structure feature (i.e. fragment with 3 rings). The substructures are selected due to their chemical relevance therefore being specific for a given task.

- **Hashing based fingerprints** - Each bit corresponds to a structure of length N. This representation is applied for similarity searching between molecules.

The bit string is composed of "1", representing the presence of a substructure in the molecule being tested, and "0" which indicates the absence of the bit, figure 2.2 shows an example.

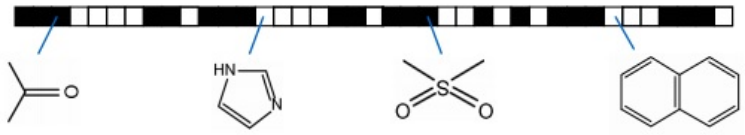

Figure 2.2: **Fingerprint Design** - Each square corresponds to a substructure of a given molecule, black squares represents fragments that are present in other molecules.

This is a simpler molecular representation that allows for a computationally efficient comparison of a molecule's structure and is also useful for calculation of simple similarity measures, such as the Tanimoto Coefficient. However, it is important to notice that different tools use different algorithms for bit generation which may cause differences when comparing different fingerprint generation tools (Andrew Dalke Scientific, 2008; Riniker & Landrum, 2013).

2.3 Molecular Similarity

Similarity is a complex and subjective concept even without the implication of chemical substances. Despite this, attempts to quantify the similarity of compounds have been a central theme in medicinal chemistry. Similarity quantifies the correspondence between properties of different compounds which is one of the most frequent tasks in chemistry and pharmaceutical research. In medicinal chemistry it allows us to understand which compounds should be considered as leads for developing a new drug, identify liabilities and determine compounds that could be used for a drug's improvement. Given its importance several measures of similarity have been proposed trough the years however, one important

aspect is that the degree of similarity depends on the method applied to measure it (Maggiora *et al.*, 2014).

Maggiora *et al.* (2014) distinguishes between two different terms: chemical similarity and molecular similarity. Chemical similarity is related to a molecule's physiochemical features such as molecular weight and solubility. Molecular similarity or structural similarity is based on the structural features such as ring systems and common substructures. Throughout this thesis when dealing with similarity I am referring to molecular/structural similarity.

With the increased availability of computational power, tools able to recognise an increasingly higher number of structural patterns than humans are being developed. This allows for a more careful similarity assessment given that more complex patterns would not be noticed by humans. Metrics that allow a quantitative representation of the similarity are important to facilitate its interpretation. The most commonly used method to calculate the similarity between two molecules is the Tanimoto Coefficient, which values range from 0 to 100% and it is represented by:

$$\frac{c}{a+b-c} \times 100 \tag{2.1}$$

where $\boldsymbol{a}$ *represents a feature only present in the first molecule,* $\boldsymbol{b}$ *a feature only present in the second molecule,* $\boldsymbol{c}$ *a feature present in both molecules and* $\boldsymbol{d}$ *a feature present in one of the molecules but not in the other.*

The more similar the molecules are, the more shared features they present and each molecule has less unique features resulting in a higher similarity value.

Since the introduction of similarity in chemistry it was stated that similar compounds should have similar properties therefore having similar biological function, because of this similarity has proven to be an important measure to be considered when dealing with chemical information (Maggiora *et al.*, 2014).

2.3.1 Noncontiguous Atom Matching Structural Similarity Method

The Noncontiguous Atom Matching Structural Similarity (NAMS) is a similarity method based on atom alignment for the analysis of structural similarity between molecules. This method is based on the comparison of the bonding profiles of atoms on comparable molecules, including features such as chirality or double bond stereoisomerism. The similarity measure is defined on the annotated molecular graph, based on an iterative directed graph similarity procedure and optimal atom alignment between atoms using a pairwise matching algorithm (Teixeira & Falcao, 2013). NAMS was the selected tool to calculate the structural similarity between molecules in this thesis and it uses the Tanimoto Coefficient to determine the similarity value.

Molecular similarity is an important measure (Bender & Glen, 2004; Nikolova & Jaworska, 2003) with a great potential to increase QSAR models performance. However, the lack of standard tools to quantitatively represent the degree of similarity between molecules has prevented its broader use. Although not applied in the final prediction models, in the course of this thesis I tested different approaches based on molecular similarity which gave insight to the importance of this feature for bioactivity prediction. The results of these approaches will be presented in section 6.1.

2.4 Quantitative Structure–Activity Relationship

The current methods for drug development are not adequate, therefore we need more efficient and robust methods which can be developed using Quantitative Structure-Activity Relationship (QSAR) models. QSAR is a widely applied technique in medicinal chemistry and it is used to correlate chemical structure with activity using computational approaches (Bharath *et al.*, 2011; Moroy *et al.*, 2012). QSAR models can predict the biological outcome and reduce the number of molecules to be tested in biological assays, pharmacokinetics tests and clinical trials. They are also useful to understand which molecular properties are determinant for a biological activity, to help optimize already existent lead drug

compounds and to predict the bioactivity of unavailable compounds with the aid of information about pre-existent molecules (Tropsha, 2010). QSAR must be implemented with caution due to their difficulty in selecting the structural features that directly influence a chemical property and determining the least amount of information needed for the representation of a molecule's structure (Teixeira & Falcao, 2014).

Given its advantages QSAR studies have been utilised in the drug R&D process potentiating its improvement.

2.4.1 Machine Learning Methods in QSAR

Classical statistical approaches such as Regression Analysis and Principal Components Analysis are commonly used in QSAR models. However, they tend to oversimplify the chemical relationship between structure and activity and, at the same time, can hinder the extraction of valuable information. Due to this, machine learning methods have been recently applied, since they offer a sophisticated approach when dealing with high dimensionality data and help produce more accurate models (Schneider & Downs, 2003). These methods include Artificial Neural Networks, k-Nearest Neighbours, Random Forest and Support Vector Machines.

Given the diversity of chemical data there is not a preferred method to apply as it varies according to the dataset size and diversity, the problem being addressed and the nature of the variables being used (Mitchell, 2014).

Chapter 3

Related Work

Health entities require hERG safety testing in the clinical trial phase, this assessment is done using conventional patch-clamp electrophysiology. This is the current gold-standard method but it is an *in vitro* technique making it expensive and time consuming. Computational approaches are a more suitable alternative for their simple and inexpensive evaluation of chemical substances. The challenge with these methods is due to the high complexity of the hERG channel, which causes reliability problems. QSAR studies help establish relations between a molecule and the respective biological response and are widely used in this type of situations, they are also useful to understand which molecular properties are determinant for a drug's function which can help to optimize already existent lead compounds.

Roche *et al.* (2002) reported one of the first applications of the QSAR methodology regarding the hERG channel inhibition. Using machine learning algorithms such as neural networks the authors were able to classify a compound as a blocker or non-blocker. Even in the early stages of these applications the importance of variable reduction was already recognized, due to the large amount of molecular descriptors used, the authors applied a Self-Organizing Map to select the best descriptor collection. The models produced were able to correctly identify 89% of the non-blocking agents and 70% of the blocking compounds and a Matthews's correlation of 0.61 was achieved. Another key aspect for a correct hERG activity assessment is the diversity of the compounds that are used to train the prediction models, the authors were aware of this despite the lack of diversity in the

molecules used. With the continuous applications of the QSAR methodology attention has been given to the selected descriptors and there has been an effort to find the smallest number of variables that can be used for compound classification. Tobita *et al.* (2005) partitioned the dataset in different sets and applied Support Vector Machines (SVM) to build prediction models. They were able to identify eight consistent molecular descriptors which were used in the models that yield the best results. These descriptors are related to hydrophobicity of molecules, surface area and polar charges, surface area and molar refractivity, molecular size, number of amines and molecular fragments that suggest a long chain and a substructure consisted of two rings connected by a bond. The best model had an accuracy of 70% for the molecule's classification. Sinha & Sen (2011) selected 12 "global descriptors" that could characterize a compound as a whole. These descriptors include: number of acceptors, number of basic nitrogen, number of carbon, acceptor-acceptor average distance, donor-donor average distance, volume, polar surface area, non-polar surface area, number of rotatable bond, donor charge, acceptor charge and dipole moment. Their prediction model achieved Coefficient of Determination ($\boldsymbol{R}^2$) values between 0.72-0.75. These authors brought attention to structural fragments and molecular features that may help explain the drug's binding characteristics to hERG.

Molecular similarity can reveal important connections between molecules especially when considering the molecular structure. Despite the lack of standardised tools to assess this parameter some researchers have been trying to use it to understand similar prediction results between different molecules. Ekins *et al.* (2006) used molecular similarity, calculated through the Tanimoto Coefficient coefficient, to verify if there was a relationship between the prediction error and the molecules similarity, they noticed that the error increased as the Tanimoto Coefficient similarity declined suggesting that with the lack of similar molecules the model performed worse, this is coherent with the statement that similar compounds have similar features. Regarding the prediction models built, the authors perform a Sammon map approach which resulted in a classification accuracy of 95% and a the Kohonen map which classified 81% of the compounds correctly. Another approach by Thai & Ecker (2009) used molecular similarity

to generate SIBAR descriptors, these descriptors are calculated through the similarity, obtained using Euclidean Distance or the Tanimoto Coefficient, between compounds from a training set and a reference set which results in a similarity matrix to be applied in the classification models. The authors also made a restricted selection of variables which resulted in the selection of 11 descriptors which include diameter related descriptors, atom counts and bond counts, partial charge descriptors, pharmacophore feature descriptors, Kier&Hall connectivity indices, physicochemical properties (SlogP), and subdivided surface areas (SlogP_VSA7,SMR_VSA5). Models built using this similarity based approach correctly classified 99% of blockers and 96% of non-blockers in the largest external test set used.

A problem which is constantly disregarded is the size of the datasets used to build the prediction models. More recently approaches using molecules contained in the ChEMBL database started to emerge. ChEMBL provides a large amount of molecules and more molecular variability than most of the datasets previously used. Czodrowski (2013) retrieved all the molecules from ChEMBL and performed a classification using a Random Forest (RF) algorithm, the best model achieved ≈80% in accuracy and used a pIC50 of 9 as a classification threshold for blockers and non-blockers. Braga *et al.* (2014) also retrieved compounds from the ChEMBL database and achieved values between 83%-84% in accuracy using a variety of machine learning algorithms such as RF and SVM with different collections of molecular fingerprints.

An increase in the diversity of the approaches as well as the addition of new methods to improve the previous is noticeable. Key aspects regarding the hERG problem are related to the molecule collection diversity, which is important to allow the discovery of biological patterns that are crucial for the channel's inhibition, and the lack of diverse approaches using molecular similarity, which can be considered a measure with great potential. All of the related works presented were taken in consideration during the development and evaluation of the methods employed in this thesis.

Chapter 4

Tools and Resources

The analysis of data is crucial for a better understanding of chemical information as well as for the extraction of important knowledge. The Chemoinformatics field utilises diverse tools and methods which enable simple and effective ways to perform chemical data analysis. These include the use of chemical tool kits which are specialised in determining mathematical representations of molecular features and statistical and machine learning methods which allow for the treatment of large quantities of data and enable a better understanding of the information given.

This chapter will introduce the tool kits and the statistical and machine learning methods applied in the course of this work.

4.1 Tools

When dealing with large amounts of data the best approach to treat and mange the information is through the use of computational tools. This section will focus on the programming languages and chemical tool kits utilised in this work.

4.1.1 Python

Python[1] is an interpreted, interactive, object-oriented programming language, it uses very clear syntax making it well suited for preforming basic programming

[1] Python programming language, `http://www.python.org/` (accessed April 2015)

tasks in Chemoinformatics. It has interfaces to many system calls and libraries some of which encompass Chemoinformatics tool kits which help with tasks such as data analysis and file parsing (O'Boyle *et al.*, 2008). Python 2.7 was utilised to: (1) retrieve molecular information from the ChEMBL database using an adapted script from Czodrowski (2013), (2) to access the RDKit functionalities enabling the calculation of molecular descriptors and (3) to use a variety of Pybel modules.

4.1.2 OpenBabel and Pybel

OpenBabel is a C++ tool kit capable of reading and writing molecular file formats as well as performing molecular data manipulation.

Open Babel includes two components, a command-line utility and a C++ library:

- **Command-line utility** - It is intended to be used to translate between various chemical file formats.
- **C++ library** - Includes all of the file-translation code as well as a wide variety of utilities. Regarding Chemoinformatics it includes standard chemistry algorithms such as determination of the smallest set of smallest rings, bond order perception, molecular fingerprints generation and calculation of some molecular descriptors (e.g. logP, polar surface area and molar refractivity).

OpenBabel was used through Pybel, a Python module that provides access to the OpenBabel C++ library (O'Boyle *et al.*, 2008, 2011). For this thesis Pybel was used for the conversion between molecular representation formats, to calculate molecular biochemical features and to generate 2D molecular representations.

4.1.3 RDKit

RDKit[1] is an open source C++ tool-kit for Cheminformatics with wrappers for Python, Java and C#. It is a multi-purpose tool which supports multiple molecular formats (e.g. SMILES), easy chemical data manipulation, substructure

[1]RDKit, `https://www.rdkit.org/` (accessed March 2015)

searching, chemical transformations, 2D molecular depiction, 2D to 3D molecule conversion, fingerprint generation, similarity measures and molecular descriptors calculation. The RDKIT tool-kit was utilised to determine molecular descriptors values for the collected molecules and it was accessed through Python.

4.1.4 R

R is a free software environment for statistical computing and graphics construction. It provides a wide variety of statistical and graphical techniques and it is highly extensible through the addition of packages (R Development Core Team, 2015), some of which specifically used for Chemoinformatics. The *randomForest* (Liaw & Wiener, 2002) and *e1071* (Meyer *et al.*, 2014) packages were utilised to perform machine learning methods, the *glmnet* (Friedman *et al.*, 2010) package was applied to perform the Least Absolute Shrinkage and Selection Operator (LASSO) and the *rcdk* (Guha, 2007) package was used to access the Chemistry Development Kit (CDK) functionalities for SMILES parsing, molecular descriptors generation and molecular fingerprints calculation. The prediction models built were tested using the *prediction* built-in function which was utilised to predict the pIC50 of a new collection of molecules therefore allowing the calculation of the Mean Squared Error (MSE) and Coefficient of Determination ($\boldsymbol{R}^2$).

4.1.5 CDK

The Chemistry Development Kit (CDK) is an open-source Java library for manipulating and processing chemical information. It provides methods for many common tasks in molecular informatics, including 2D and 3D rendering of chemical structures, SMILES parsing and generation, ring searches, isomorphism checking and structure diagram generation (Steinbeck *et al.*, 2003). For this thesis, CDK was utilised to obtain molecular descriptors and molecular fingerprints and it was accessed through the *rCDK* library in R.

4.1.6 e-Dragon

e-Dragon is an online version of the DRAGON software, which allows for the calculation of molecular descriptors (Tetko *et al.*, 2005). For the calculation of the descriptors used in this thesis the standard options were applied.

4.2 Statistical Methods

In Chemoinformatics it is common for researchers to work with massive amounts of data and variable reduction methods are crucial to remove information that is not essential as it can interfere the correct interpretation of the data and knowledge extraction. There is also a need of metrics that can translate the results obtained using different methodologies into a quantitative value enabling a better understanding of it and it is also necessary to perform a validation of those results through their application on untested data. In this section it will be presented the approaches used for variable selection, the statistical measures used to evaluate the models constructed and the validation approach performed.

4.2.1 Least Absolute Shrinkage and Selection Operator

The Least Absolute Shrinkage and Selection Operator (LASSO) is a selection and shrinkage method for linear regression, it is used to select predictors of a target variable from a large set of predictors. It does this by shrinking the regression coefficients[1], β_j, to zero based on the value of a tuning parameter λ,

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{4.1}$$

were ***RSS*** *represents the residual squared error.*

By reducing the coefficients of irrelevant variables to zero it selects the important predictors. Figure 4.1 shows the variation of number of variables considered

[1] mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant.

relevant with the variation of the maximum permissible value ($\sum_{j=1}^{p} |\beta_j|$) which is controlled by the tuning parameter.

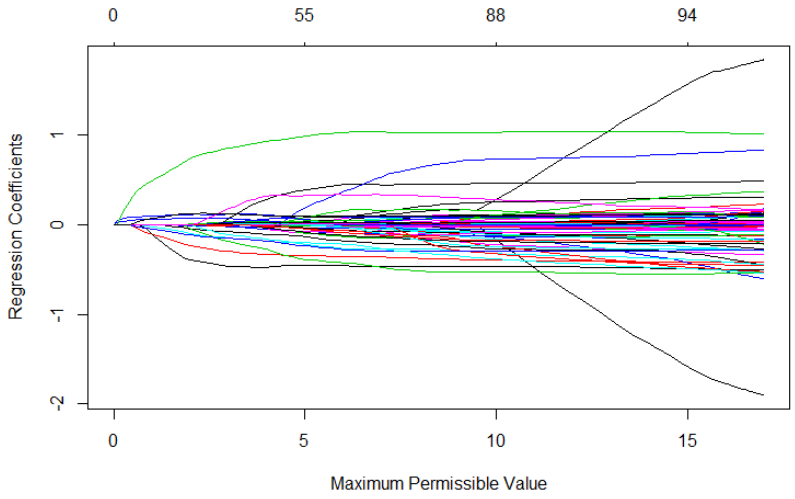

Figure 4.1: **LASSO Variable Selection**. The higher the penalty value the more permissive the model, this is shown by the increasing number of variables represented in a variety of colors.

This reduction method tries to minimise the residual squared error thus improving the prediction accuracy and facilitating the interpretation of the results by reducing the number of predictors (Tibshirani, 1994).

LASSO was applied through the *glmnet* R package and it was used with the default parameters. The predictor variables used were the molecular descriptors and fingerprints obtained from different tools.

4.2.2 Statistical Measures

Statistical measures are important to extract information from mathematical approaches. The MSE and the $\boldsymbol{R}^2$ were the two measures used to quantify the prediction model's quality.

Mean Squared Error

The Mean Squared Error (MSE) value translates the closeness of two data points, the closer they are the smaller the MSE and closer the fit is to the data. Supposing y_i is the observed value and $\hat{y}_i$ is the predicted value, the MSE is expressed by:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \tag{4.2}$$

Coefficient of Determination

The Coefficient of Determination ($\boldsymbol{R}^2$) gives the percentage of the response variable variation explained by a model. The higher the value the better the model explains the variability of the response data around its mean. The formula used to calculate the $\boldsymbol{R}^2$ is shown in equation 4.3, for the calculation consider that $\bar{y}$ corresponds to the mean of y.

$$R^2 = 1 - \frac{MSE}{Variance(y)} = 1 - \frac{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}} \tag{4.3}$$

4.2.3 Cross-Validation

Cross-validation is a method to measure the performance of a model. Data overfitting is a common problem and it may cause bias, therefore an approach that evaluates the model produced using an external dataset is necessary. Frequently the dataset used to compute the model is called Train set and the set used for validation is the Test set. Of the several procedures known, k-fold Cross-validation with k=5 was applied to calculate each models prediction accuracy. Every data point is in a test set exactly once, and gets to be in a training set k-1 times. The 5-fold Cross Validation was applied as follows:

1. The Train dataset was randomly divided into five approximately equal sets;

2. For each fold, four sets are used to train the models which were evaluated using the remaining set (Test set);

3. The mean MSE and $\boldsymbol{R}^2$ were used to determine the model's accuracy.

4.3 Machine Learning Methods

Machine Learning is related to the study, design and development of the algorithms that enable computers to learn without being explicitly programmed, as defined by Arthur Samuel, in 1959, and it can be divided into two categories: Supervised Learning and Unsupervised Learning. In the first case I have a dataset consisting of features and labels and the task is to construct an estimator which is able to predict the label of an object given a set of features. The Unsupervised Learning uses data that does not have labels and tries to find similarities between the objects (James *et al.*, 2014).

Machine Learning methods started to be applied in Chemoinformatics as an alternative to the statistical methods, they offer advantages since they use different approaches to solve the data high dimensionality issue (Schneider & Downs, 2003). In the course of this work it was applied unsupervised learning using the methods referenced in this section.

4.3.1 Support Vector Machines

Support Vector Machines (SVM) is a kernel-based method developed for classification and regression and it is commonly used in QSAR approaches to predict biochemical features. It uses decision planes (i.e. planes that separate different class objects) to define decision boundaries in a multidimensional space. Figure 4.2 represents a classification example using SVM. The chosen plane for class separation is the one that maximises the margin between the classes closest points, the points that are on the boundary line are called support vectors and the middle of the margin is called hyperplane (i.e. the optimal separating plane).

After the hyperplane calculation, points that are not on their respective side of the margin are assigned a penalty value which reduces their influence for model building. SVM are not restricted to finding linear plane separators, they also

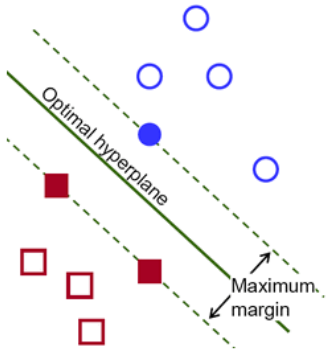

Figure 4.2: **SVM Classification Scheme**: The dots represent different classes and the hyperplane corresponds to the line that best divides the classes.

perform classification in high-dimensional feature spaces using other kernel techniques (i.e. polynomial, radial basis function and sigmoid) (Meyer, 2015).

SVM was used through the R package *e1071* with the all default parameters except the *scale* option which was redefined to "False". The SVM was trained using the molecular descriptors and fingerprints collected from the chemical tool kits mentioned before.

4.3.2 Random Forest

Random Forest (RF) is an ensemble approach that utilises a collection of decision trees. A decision tree is capable of selecting the most important variables from the given data and constructs an explicit model that describes the relationship between the selected variables and the predictions. A representation of a decision tree in shown in figure 4.3. A tree is composed of decision nodes, event nodes, terminal nodes, decision branches and event branches.

The decision nodes correspond to a point were a decision must be made and each decision branch represents all the mutually exclusive possible alternatives. The event node represents the occurrence of an event and the event branches represent all the mutuality exclusive outcomes of that event. The terminal node represents the results of a combination of decisions or events. Each decision tree is given the same information as a starting point however each tree utilises a different path to reach the terminal node. The decision trees act as a "weak

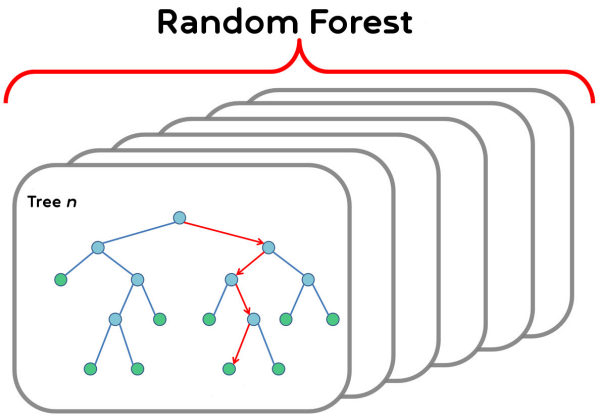

Figure 4.3: **Random Forest Representation**: A RF is an ensemble of decision trees.

predictor" and a collection of them forms a "strong predictor" corresponding to a RF. This approach allows the RF to deal with unbalanced data however, when a dataset has high noise levels it may over fit the data and it also does not allow for predictions beyond the range of the training data. The process used to obtain the best collection of molecules is similar to a "black box" in the sense that the trees selected are not explicit (Albright *et al.*, 2008).

RF have continuously been chosen for QSAR studies since they are able to have high accuracy in predictions, they have the ability to perform feature selection and they also have a method for assessing the descriptors importance (Palmer *et al.*, 2007). RF was applied using the *randomForest* R package with the default parameters. The models were trained using molecular descriptors and fingerprints.

Chapter 5

Data and Methods

This chapter will focus on the restrictions applied to the molecular dataset retrieval and it will also describe in detail the methodologies applied in this thesis.

5.1 Data

The datasets used were collected in two phases. For the first phase drug information regarding drugs that were screened for hERG activity were retrieved from Sinha & Sen (2011) and Su *et al.* (2010) resulting in a dataset with 545 molecules. A similarity analysis was performed using NAMS to verify the presence of duplicates. After this, the final dataset was randomly split into separate sets: Train with 258 molecules and Test with 105 molecules. These datasets were used to select the best variable selection method.

For the second phase I decided to retrieve drug information from ChEMBL[1] using an adapted script from Czodrowski (2013), only molecules that met the following criteria were selected:

- hERG channel as target;
- Bioactivity value expressed in IC50;
- Inhibition or Inhibitory assay type.

[1]The molecules were retrieved on June 8, 2015

The previous Train and Test sets were combined and a second similarity analysis was used to remove duplicates. The final dataset consisted of 2719 molecules which was then randomly divided into a Train1 set, with 2038 molecules, and an Independent Validation Set (IVS) with 681 molecules. The Train1 set was divided into a Train and Test sets to be used in the Variable Selection, Model Selection and Cross-Validation steps.

5.2 Methods

The workflow shown in figure 5.1 summarises this chapter illustrating each one of the steps applied in the course of this work.

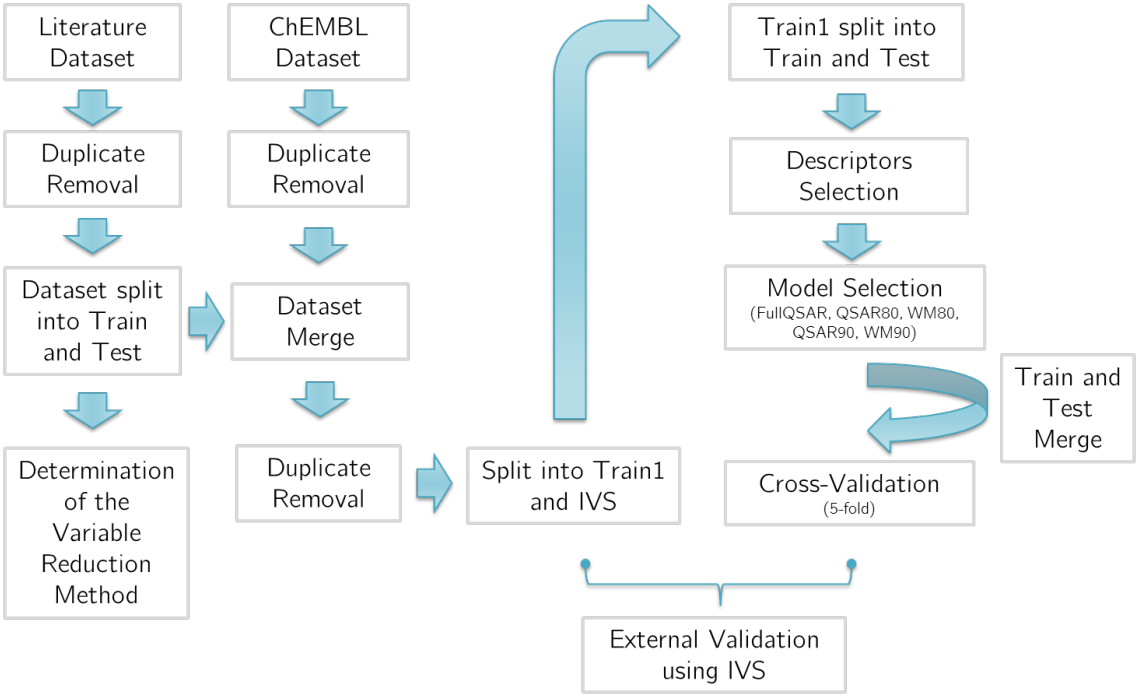

Figure 5.1: **Data and Methods Workflow** Each step of the data treatment and the methods applied in this thesis are illustrated in this workflow. The Determination of the Variable Reduction Method was performed prior to the first Dataset Merge.

5.2.1 Variable Selection

An important part of model building using chemical data relates to the choice of molecular descriptors therefore, standard methods were applied to evaluate the best algorithm for variable selection. The methods tested were LASSO, RF, Linear Regression (LR) with SVM and SVM; the molecular descriptors collections included RDKit, CDK, e-Dragon and Fingerprints from OpenBabel. The tests were performed using the first phase Train and Test sets, the algorithms tested were used through R libraries.

A new approach based on LR was tested. This method analyses each variable (i.e. molecular descriptors and fingerprints) independently by building the prediction model using only the variable being tested. The $\boldsymbol{R}^2$ calculated for each variable is used as a threshold to find the variable collection that yields the highest overall $\boldsymbol{R}^2$. If a descriptor has a higher value of $\boldsymbol{R}^2$ it is assumed that it is more important for the model that others with a lower value therefore, by determining the best threshold for the selection of variables, it only selects the necessary descriptors to achieve the best possible model. To calculate the $\boldsymbol{R}^2$ value different models can be applied, in this case only SVM was initially used. The pseudocode for this variable selection method is presented in appendix A.

5.2.2 Best model selection and 5 fold Cross-Validation

The Train1 set was randomly divided into 5 similarly sized subsets each being used once as a validation set and the rest as a training set. In the first fold, descriptor selection was performed for each model using the variable reduction method discussed before. The descriptors chosen were RDKit and CDK molecular descriptors and CDK and Open Babel Fingerprints which were used separately and combined as follows: RDKit and Open Babel Fingerprints, RDKit and CDK Fingerprints, CDK and Open Babel Fingerprints and CDK and CDK Fingerprints. The machine learning methods applied were SVM and RF; LASSO was also applied. As stated before it is assumed that similar molecules have similar features therefore a similarity based approach was also tested in addition with the standard approach. The two methodologies were implemented as follows:

- **Standard** - For each combination of descriptors the MSE and $\boldsymbol{R}^2$ of each model (SVM, RF and LASSO) were determined through the mean result of each fold. The entire molecule collection was used to train the models.

- **Similarity-based** - Two Tanimoto similarity threshold values, 80% and 90%, were used to select only molecules that had high similarity; this enabled the reduction of the dataset to 1696 and 1323 molecules, respectively. The methodology used for the Standard approach was then applied to the datasets.

Two different methods to predict the pIC50 were implemented:

- **Method 1** - Uses the weighted mean between the pIC50 of all molecules that have a similarity value equal or above the given threshold and their similarity value. Given the use of a similarity threshold to select molecules this method is only applied to the Similarity-based approach. The Cross-Validation was not applied to this method.

- **Method 2** - Includes two other descriptors to each descriptor collection: for each molecule, the similarity value and the pIC50 of the most similar molecule. This method is applied with the Standard and Similarity-based approach.

5.2.3 Validation with Independent Validation Set

The models that yield the best results with the 5-fold Cross-Validation were selected for the validation with the Independent Validation Set (IVS). Although not present in the 5-fold validation, models from the Similarity-based approach that use a weighted mean to predict a compounds pIC50 were also validated with the IVS. The selected models for the validation with the independent set were the following:

- QSAR Model with CDK descriptors and CDK Fingerprints which utilised the complete molecule collection to train the prediction model;

- QSAR Model with CDK descriptor trained using molecules with 80% similarity;
- Weighted Mean Model using molecules with 80% similarity;
- QSAR Model with CDK descriptors and CDK Fingerprints trained with molecules with 90% similarity;
- Weighted Mean Model built using molecules with 90% similarity.

Chapter 6

Results and Discussion

This chapter will focus on the results obtained through the exploratory work developed using NAMS and from the application of the methods stated in Chapter 5. It will also include an analysis of the results and the relevant points suggested by this research.

6.1 Exploratory Work

As stated before, similarity is considered an important measure to be considered when dealing with bioactivity prediction. NAMS was utilised to test two approaches with a strong application of structural similarity to predict a compounds pIC50. These approaches gave insight to the procedures that can be used when applying structural similarity and it also allowed to verify the veracity of the hypothesis that structural similarity can be useful for feature prediction.

6.1.1 Molecular Space Approach

With the aid of NAMS I built a molecular space with *kernels*, i.e. groups of molecules with high similarity and therefore, similar characteristics. The methodology to create this molecular space goes as follows:

1. Arrangement of the molecules from the most inhibitory to the least;

2. Selection of the molecule with the highest inhibition score, this is called K1;

3. Selection of a similarity threshold (50%, 60%, 70%, 80% and 85%);

4. Calculation of the similarity between the K1 and the other molecules present in the dataset;

5. Selection of the first molecule that has a similarity value with K1 bellow the selected threshold, this molecule will be named K2 and K1 will be removed from the molecule collection;

6. Calculation of the similarity between K2 and the remaining molecules in the collection;

7. Repetition until no more molecules are found bellow the selected similarity threshold.

This approach enabled the creation of *kernels* which are unique in the sense that they represent the most inhibitory molecule of a class of similar ones. Using different similarity thresholds I was able to select different numbers of *kernels* which were then used to train a SVM, table 6.1 shows the results. With this molecular space the input molecules characteristics can be predicted based on its proximity with the *kernels*.

Table 6.1: **Similarity threshold influence in the number of *kernels* and model's performance**

Similarity Threshold	Number of *kernels*	R^2
0.50	21	0.395
0.60	62	0.457
0.70	123	0.459
0.80	168	0.509
0.85	184	0.471

The results suggested the influence of the number of *kernels* in the pIC50 prediction model performance. By using different similarity thresholds the number of *kernels* being selected varied, the higher/more permissive the similarity cut-off value the more *kernels* are chosen. With the increase in the similarity threshold

it was visible an improvement in the models performance which started to decline at the 80% threshold. The best result was achieved with the 80% threshold which has 168 *kernels* and yield an $\boldsymbol{R}^2$ of $\approx 51\%$.

6.1.2 Neighbourhood Molecules Approach

An approach based solely on the molecular similarity and the pIC50 was also tested. For each molecule I selected the closest molecules ranging from the 3^{rd} to the 8^{th}, along with their respective pIC50. The pIC50 models performance results were achieved by two means:

(a) Using the closest molecules pIC50 as descriptors which were then used to train a SVM model;

(b) Using the product between the similarity value and the pIC50 of the closest molecule's as descriptors (i.e.product descriptors) used to train a SVM.

Table 6.2: **pIC50 prediction using neighbour molecules**. With a number of neighbours above 6 the models performance reduces to lower values than considering 5 or less neighbours.

Number of Molecules	Product Descriptors	$\mathbf{R^2}$
3	Yes	0.405
	No	0.404
4	Yes	0.415
	No	0.379
5	Yes	0.428
	No	0.392
6	Yes	0.396
	No	0.360
7	Yes	0.387
	No	0.342
8	Yes	0.386
	No	0.345

The approach using the product between the closest molecules pIC50 and similarity returns the overall best prediction results in comparison with the results

obtained using the pIC50 mean. The best results were obtained using the 5 closest molecules which returned an $\boldsymbol{R}^2$ of ≈ 0.43. Using the 3 and 4 closest molecules the prediction results are very similar suggesting that with this threshold the prediction has not improved as much as the other cases when using the product between pIC50 and similarity.

These results, although not as high as the final models presented in this thesis, suggest a strong relationship between the similarity and the bioactivity value.

6.2 Descriptors and Variable Selection Methods

This section presents the results obtained from the different algorithms and descriptors sets which were tested to find the best variable reduction method and the best descriptor collection to use. The results using only the RDKit descriptor set are shown on the left panel in figure 6.1. The best selection methods were the RF with an $\boldsymbol{R}^2$ of 0.557 and the LR coupled with SVM which returned an $\boldsymbol{R}^2$ of 0.553, the SVM yield a $\boldsymbol{R}^2$ of 0.514 and the LASSO algorithm returned the worse result with a $\boldsymbol{R}^2$ of 0.147. The coupling with OpenBabel Fingerprints increased the overall performance; despite the increased $\boldsymbol{R}^2$ the performance order of the algorithms roughly maintained the same with the exception of the the best algorithm which was the LR+SVM with an $\boldsymbol{R}^2$ of 0.631; the RF approach had a performance of 0.579, the SVM approach achieved an $\boldsymbol{R}^2$ of 0.501 and LASSO yield an $\boldsymbol{R}^2$ of 0.347.

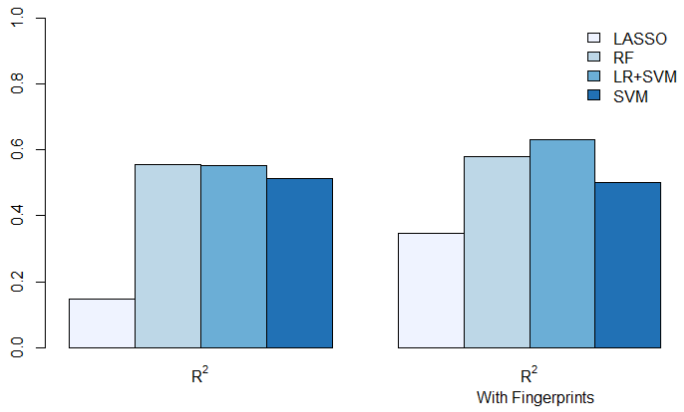

Figure 6.1: **Variable Selection - RDKit**. The addition of molecular fingerprints improves the results and enhanced the performance of the LR+SVM approach.

The second descriptor collection used was the CDK descriptor set, figure 6.2 shows the results. Considering only the CDK descriptors the pattern is similar to the one obtained with the RDKit collection, the best $\boldsymbol{R}^2$ belongs to RF with 0.515 followed by the LR coupled with SVM with 0.488, the SVM method returned a performance value of 0.443 and LASSO achieved a negative $\boldsymbol{R}^2$ of 0.177. With

the addition of fingerprints the results improved similarly to the RDKit collection coupled with OpenBabel Fingerprints and the performance order of the models is also identical. The addition of fingerprints improved greatly the results of the LR coupled with SVM yielding an $\boldsymbol{R}^2$ of 0.61, the RF approach gives a performance value of 0.552, the SVM algorithm achieved an $\boldsymbol{R}^2$ of 0.497%, LASSO still performed the worse achieving an $\boldsymbol{R}^2$ of 0.186.

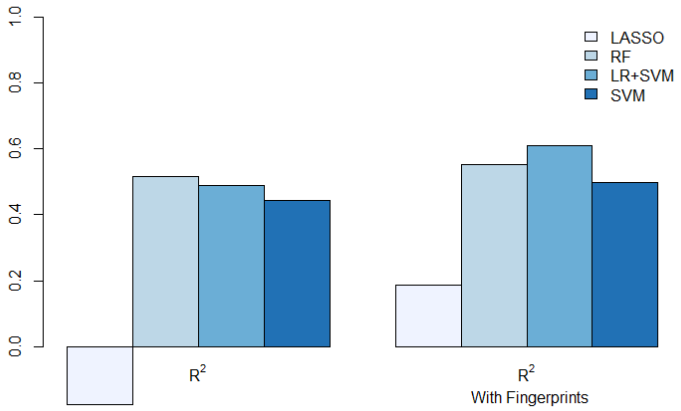

Figure 6.2: **Variable Selection - CDK**. The pattern visible with the addition of molecular fingerprints is identical to the RDKit collection.

The e-Dragon descriptor collection achieved the worst results, as seen in figure 6.3 only the LR coupled with SVM algorithm was capable of achieving a positive performance value of 5%. The SVM, LASSO and RF returned negative performance values of 0.05, 0.087 and 0.194 respectively. When adding the OpenBabel Fingerprints the results improve as seen with the RDKit and CDK descriptors collections. The best model corresponded to the LR coupled with SVM which achieved an $\boldsymbol{R}^2$ of 0.531, the second best algorithm was SVM with a performance of 0.385, LASSO was surprisingly the third best model with an $\boldsymbol{R}^2$ of 0.293 and RF performed the worse with an $\boldsymbol{R}^2$ of 0.052.

These results illustrate the importance of the descriptors used and the variable reduction method performed. There are important differences between the descriptors collections, the RDKit and CDK sets return very similar results indicating that these perform well with the reduction methods tested. However,

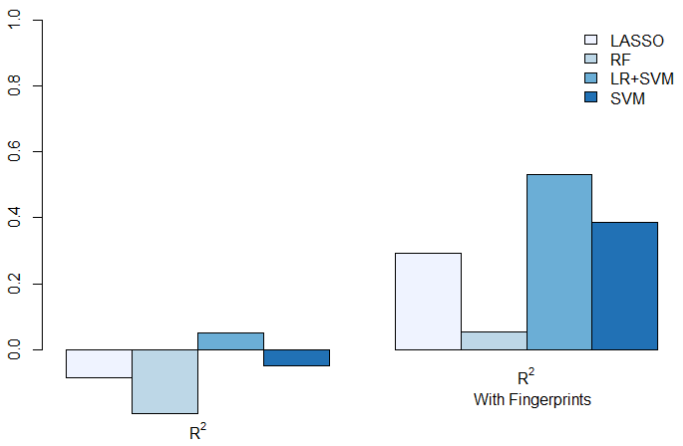

Figure 6.3: **Variable Selection - e-Dragon**. e-Dragon descriptor collection performed the worse in comparison with other collection. The best variable reduction algorithm remained the LR+SVM.

with the e-Dragon collection the algorithms tested had a very low performance suggesting that this set was not suitable for these approaches therefore, it was decided to exclude e-Dragon from the posterior methods. The addition of molecular fingerprints, namely OpenBabel Fingerprints, showed an improvement of the algorithms performance and, since the fingerprints split each molecule into structural fragments this shows the importance of the molecular structure for the prediction of biochemical features. This was a key finding and suggested the use of different sets of fingerprints in the Best Model Selection stage to find if the type of fingerprints used show as much differences as the molecular descriptor collections.

Focusing on the algorithms results for the RDKit and CDK descriptors set (figures 6.1 and 6.2) it is visible that the worst performance was achieved by the LASSO algorithm indicating that this type of shrinkage method is not well suited for the molecular descriptors tested, one hypothesis could be the lack of contrasting differences between the descriptors values which difficult the algorithm's application. The best method, without considering the addition of fingerprints is RF closely followed by the LR coupled with SVM however, with the addition of fingerprints the best algorithm is the LR coupled with SVM, which suggests that

this method can perform better than RF when molecular structure is included as a variable. Given these results, for the following steps it was decided to use the LR algorithm coupled with SVM, RF and LASSO. It was verified that some of the best results use a RF method therefore it is hypothesised that using the LR approach coupled with the RF will have the best performance, it was also decided to use LASSO to find if this algorithm improved with a previous selection of the most relevant variables using LR. The LR coupled with SVM was used as standard.

6.3 Best Model Selection and 5-Fold Cross-Validation

This section will focus on the results achieved with the 5-fold Cross-Validation performed on different descriptors collections and with the different approaches stated before: the Standard approach, which corresponds to the FullQSAR model, and the Similarity-based approach which includes the QSAR80, WM80, QSAR90 and the WM90.

The results show that the RF algorithm performed the best with the variables given and the LASSO approach performed the worst.

The FullQSAR model results are shown in table 6.3. The maximum mean $\boldsymbol{R}^2$ belongs to the model built with the RDKit descriptor set and CDK Fingerprints which has a mean MSE of 0.387 and a mean $\boldsymbol{R}^2$ of 0.538, the second best result used CDK molecular descriptors and CDK Fingerprints and has slight variations compared with the previous approach achieving a mean MSE of 0.395 and a mean $\boldsymbol{R}^2$ of 0.534, the third best model was built with RDKit descriptors and achieved a mean MSE of 0.407 and a mean $\boldsymbol{R}^2$ of 0.514. This is considered the base-line approach, it is similar to other approaches found in the literature since it only uses classic molecular descriptors and fingerprints to find patterns that enable the prediction of the pIC50. It were expected lower performance results with this model since it was used a relatively large and diverse training set, in comparison to other authors, which may account for the differences in accuracy.

With the Similarity-based approach I expected better results since the training universe is reduced to the molecules that have more structures in common with the rest. By doing this selection I am focusing more on patterns that occur in a

Table 6.3: **5-Fold Cross-Validation:** FullQSAR

Variables	Algorithm	Mean MSE	Mean R^2
Fingerprints OpenBabel	SVM	0.702	0.174
	RF	0.552	0.350
	LASSO	0.722	0.152
Fingerprints CDK	SVM	0.641	0.245
	RF	0.446	0.473
	LASSO	0.669	0.210
RDKit	SVM	0.735	0.122
	RF	0.407	0.514
	LASSO	0.627	0.251
RDKit + Fingerprints OpenBabel	SVM	0.670	0.199
	RF	0.410	0.510
	LASSO	0.568	0.322
RDKit + Fingerprints CDK	SVM	0.588	0.296
	RF	0.387	0.538
	LASSO	0.649	0.222
CDK	SVM	0.598	0.293
	RF	0.416	0.510
	LASSO	0.648	0.237
CDK + Fingerprints OpenBabel	SVM	0.529	0.377
	RF	0.420	0.505
	LASSO	0.599	0.296
CDK + Fingerprints CDK	SVM	0.540	0.365
	RF	0.395	0.534
	LASSO	0.595	0.298

smaller but coherent set of molecules. Similarity cut-off values of 80% and 90%, which enabled the selection of molecules that had a high structural similarity, were chosen. Even with the application of the similarity threshold the number of selected molecules remained relatively high, 1696 for the 80% cut-off value and 1323 for the 90% cut-off which indicates intrinsic similarities between the molecules in the Train set. With the reduction of the Train set to a more coherent sample I expected better results than with the FullQSAR model, this was in fact verified.

For the QSAR80 the results are shown in table 6.4, the best result was achieved

using RDKit descriptors and OpenBabel Fingerprints resulting in a mean MSE of 0.282 and a mean $\boldsymbol{R}^2$ of 0.654, the second best result also used RDKit descriptors but coupled with CDK Fingerprints and achieved a mean MSE of 0.296 and a mean $\boldsymbol{R}^2$ of 0.635, the third best result used CDK descriptors and had a mean MSE of 0.304 and a mean $\boldsymbol{R}^2$ of 0.629.

Table 6.4: **5-Fold Cross-Validation:** Similarity-Based QSAR80

Variables	Algorithm	Mean MSE	Mean R^2
Fingerprints OpenBabel	SVM	0.338	0.589
	RF	0.333	0.596
	LASSO	0.393	0.524
Fingerprints CDK	SVM	0.364	0.557
	RF	0.329	0.599
	LASSO	0.404	0.508
RDKit	SVM	0.353	0.564
	RF	0.302	0.628
	LASSO	0.389	0.521
RDKit + Fingerprints OpenBabel	SVM	0.366	0.548
	RF	0.282	0.654
	LASSO	0.378	0.536
RDKit + Fingerprints CDK	SVM	0.350	0.567
	RF	0.296	0.635
	LASSO	0.389	0.521
CDK	SVM	0.347	0.576
	RF	0.304	0.629
	LASSO	0.398	0.515
CDK + Fingerprints OpenBabel	SVM	0.340	0.585
	RF	0.357	0.569
	LASSO	0.589	0.289
CDK + Fingerprints CDK	SVM	0.345	0.578
	RF	0.342	0.584
	LASSO	0.587	0.287

Considering the QSAR90, I expected even better results given that with the similarity threshold of 90% the molecules used to train the model should share even more similar features. The results of this approach are shown in table 6.5, the first model used CDK descriptors and CDK Fingerprints and achieved a

mean MSE of 0.273 and a mean $\boldsymbol{R}^2$ of 0.673, the second model only used CDK descriptors and was able to achieve a mean MSE of 0.278 and a mean $\boldsymbol{R}^2$ of 0.667, the third model was built using the RDKit collection and CDK Fingerprints and achieved a mean MSE of 0.282 and a mean $\boldsymbol{R}^2$ 0.662.

Table 6.5: **5-Fold Cross-Validation:** Similarity-Based QSAR90

Variables	Algorithm	Mean MSE	Mean R^2
Fingerprints OpenBabel	SVM	0.322	0.618
	RF	0.312	0.626
	LASSO	0.609	0.257
Fingerprints CDK	SVM	0.3388	0.597
	RF	0.303	0.639
	LASSO	0.0384	0.543
RDKit	SVM	0.331	0.657
	RF	0.286	0.657
	LASSO	0.361	0.543
RDKit + Fingerprints OpenBabel	SVM	0.343	0.588
	RF	0.277	0.668
	LASSO	0.364	0.566
RDKit + Fingerprints CDK	SVM	0.352	0.577
	RF	0.282	0.662
	LASSO	0.370	0.559
CDK	SVM	0.321	0.617
	RF	0.278	0.667
	LASSO	0.365	0.455
CDK + Fingerprints OpenBabel	SVM	0.309	0.630
	RF	0.334	0.567
	LASSO	0.264	0.615
CDK + Fingerprints CDK	SVM	0.318	0.620
	RF	0.273	0.673
	LASSO	0.536	0.360

The best overall result was achieved by the QSAR90 model with a mean $\boldsymbol{R}^2$ of 0.673, this goes as according to the expected and sustains the hypothesis that structurally similar molecules have similar biochemical features. One other point worth mentioning is the selected descriptors; coupled sets of descriptors return

the best results which solidifies the importance of structure since the molecular fingerprints provide descriptors that take in account the presence of specific molecular patterns.

The pIC50 of the most similar molecule proved to be an important descriptor for model building since it was always selected for the models built. Within the best models the similarity value was only selected for the QSAR80 with the CDK descriptor set. The selected descriptors include subdivided surface areas (_VSA), Topological Polar Surface Area (TPSA), fragment counts (e.g. fr_benzene, fr_pyridine, fr_NH1), physicochemical properties (i.e. MolLogP), number of aromatic carbocycles, number of aromatic heterocycles, number of hydrogen acceptors and molecular fingerprints. Although the difference in the number, the descriptors selected are in accordance with the ones cited by the literature which evidences the ability of these descriptors to identify patterns related to the hERG inhibitors.

The Weighted Mean results are present in table 6.6 and show a mean $\boldsymbol{R}^2$ of 0.5439 and 0.610 and mean MSE of 0.380 and 0.3267 for the 80% and 90% similarity thresholds. These models apply the similarity analysis in two steps, first for the selection of the molecule's collection and second for the calculation of the pIC50 using the weighted mean of each molecule and the respective pIC50 value. The high results from the WM90 demonstrate that the similarity and the pIC50 of the molecules can be important variables to take in account for bioactivity prediction, however the results from the WM80 are not as high which may be explained by the size of the molecule collection given the more permissive similarity threshold.

Table 6.6: **5-Fold Cross-Validation**: WM80 and WM90. N corresponds to the number of molecules utilised for model training.

Approach	Mean MSE	Mean R^2	N
Similarity-Based 80%	0.380	0.544	1696
Similarity-Based 90%	0.327	0.610	1323

Although in some cases the RDKit descriptor set provides better results we were unable to successfully install the RDKit python module in the Linux server

where the human Ether-à-go-go Related Gene Inhibition Predictor (hERGIP) tool is being held therefore, and since the results with the CDK descriptor set are close to the results where the RDKit set performs better, I decided to only use the CDK molecular descriptors.

6.4 Validation with Independent Validation Set

The models selected for each approach were the following:

- **FullQSAR** - CDK descriptors collection and CDK Fingerprints with a LR variable reduction method and RF to build the prediction model.
- **QSAR80** - CDK molecular descriptors with the same methods applied in the FullQSAR approach.
- **QSAR90** - CDK descriptors collection and CDK Fingerprints using the same algorithms applied in the previous approaches.

The results for the validation with the IVS are shown in table 6.7, for each model it is present the MSE, the $\boldsymbol{R}^2$, and the number of molecules that are selected for each similarity threshold.

The IVS validation shows interesting results, the QSAR90 should have achieved the best results but the $\boldsymbol{R}^2$ is only of 0.374, the lowest yield. However the QSAR80 still shows better results than the FullQSAR model resulting in a $\boldsymbol{R}^2$ of 0.625 opposite an $\boldsymbol{R}^2$ 0.556. The WM90 model achieved an $\boldsymbol{R}^2$ of 0.5683, which was as expected better than the $\boldsymbol{R}^2$ of 0.523 of the WM80 model. The results from the WM80 were lower than the FullQSAR which was not expected based on the results from the Cross-Validation. An hypothesis for these results is related to the number of training molecules that are chosen given the similarity threshold, with no similarity limitations the number of molecules used to train the FullQSAR model was 2038 however, in the QSAR80 and WM80 the number was 1695 due to the 80% similarity threshold and for the QSAR90 and WM90 after the application of the similarity restriction the final number of molecules of the training set was only 215. The reduced number of molecules used to train the QSAR90

may be related to the low performance score since the number and diversity of the compounds might not be adequate.

Table 6.7: **IVS Validation Results**

Approach	Model	MSE	R^2	Train N	Test N
Full Train	CDK + CDK Fingerprints	0.429	0.556	2038	681
Similarity-Based 80%	CDK	0.320	0.625	1695	576
	WM	0.407	0.523	1695	576
Similarity-Based 90%	CDK + CDK Fingerprints	0.523	0.374	215	453
	WM	0.361	0.568	215	453

A representation of the relation between the predicted and the expected pIC50 regarding the results obtained for each approach is presented in figure 6.4. It is visible a good correlation between the predicted and expected pIC50 since the results follow a regression line with an y-intercept in zero. The QSAR90 plot mirrors the results obtained and clearly shows a difference in the prediction of pIC50 below 6 which is not visible in the other models. Despite the QSAR90 low results it is suggested that the QSAR90 and WM90 perform better than other models when predicting higher pIC50. A cluster of molecules is also visible in the intermediate pIC50 values which was expected since these values are more common in the collected molecular dataset.

For the hERGIP webtool only the QSAR80, WM80 and FullQSAR were used since it were the models that achieved the higher performance values. The QSAR90 returned a lower $\boldsymbol{R}^2$ than expected therefore it was not used and since the WM90 is intimately related with it was also decided to not present this model in the on-line tool.

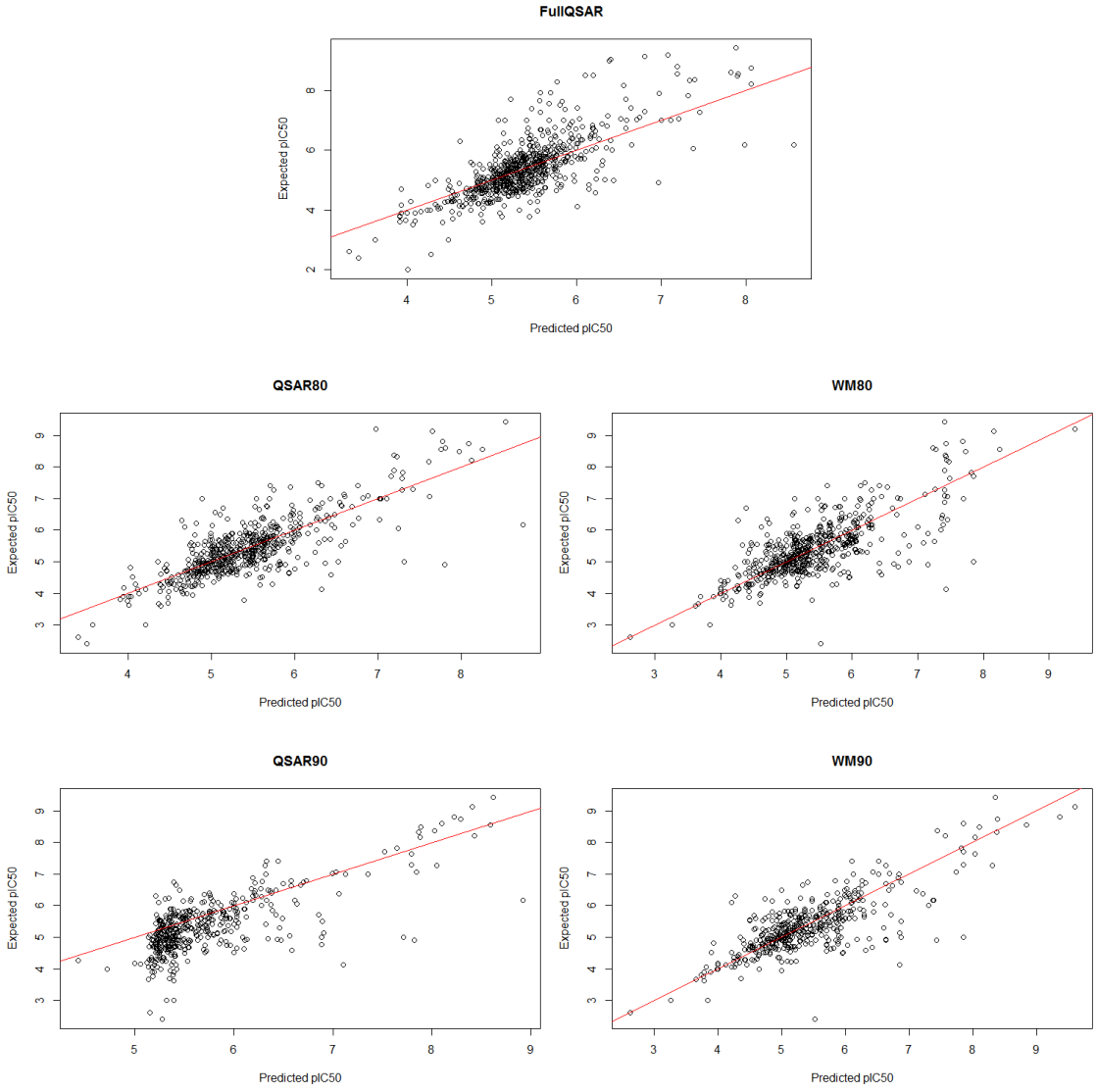

Figure 6.4: **Relationship between the predicted and expected pIC50**. The red line represents the linear function with an y-intercept in zero, the dots clustering along the line strongly suggests that the models are capable of building a direct relationship between the prediction and the real pIC50 value.

Chapter 7

hERGIP Webtool

In this chapter a detailed guide related to the hERGIP tool will be provided. As stated before this tool was developed to predict a compound's inhibition, in this case measured as the pIC50, and it was specially developed to determine hERG inhibitors. Three aspects will be focused: the Architecture, the Back-end and the User Interface. hERGIP resulted from a collaborative effort regarding the interface and the implementation of the tool's architecture.

This chapter also includes a section dedicated to a comparison of hERGIP with the Pred-hERG prediction tool.

7.1 Architecture

This tool was built in a way that enables the communication between layers, figure 7.1 shows the schematics of the layers present in hERGIP and their interactions. This is done in such a way that the user performs a query that is transmitted between the layers enabling the right information to reach the prediction models and determine the prediction result thus answering the clients query.

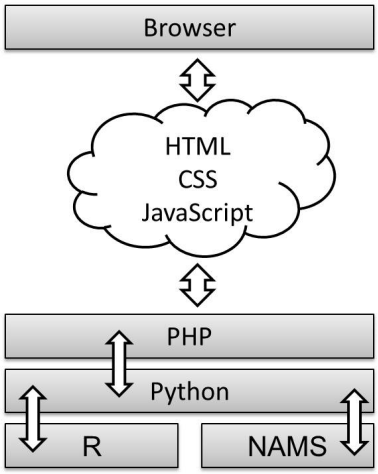

Figure 7.1: **hERGIP layered achitecture**.

7.2 Back-end

As mentioned before this tool uses a layered architecture, in the bottom are the R scripts which held the prediction models and the descriptor calculators, and NAMS which is used to perform the similarity analysis to determine which prediction model should be computed.

There are 3 R scripts that enable the prediction of the bioactivity value: ***QSAR80.R***, ***WM80.R*** and ***ModelFullQSAR.R***. The ***calcCDK.R*** is used to calculate the CDK molecular fingerprints and ***calcFingerprints.R*** is responsible for the determination of the CDK molecular fingerprints. In order to use NAMS every molecule must be converted to a nams formated file, this is done using the ***Recoder.py*** module which also uses the ***chirality.py*** and ***doubleb_e_z.py*** modules. NAMS is then accessed through Python and its results are returned to a Python script.

The tool's backbone was developed in Python and consists of several modules that are intended to better sort the results obtained from R and NAMS:

- ***Model80.py*** - Contains the functions needed to perform the QSAR80 and WM80 models;
- ***DescriptorsAndNams.py*** - Contains the functions that calculate the molecular similarity, the descriptors and fingerprints values;

- ***DescriptorsAndNamsFilter.py*** - Performs the same tasks as the previous module with the addition of molecular weight filters when calculating the molecular similarity;

- ***Tool.py*** - The main script where the results from the ***Model80.py*** and ***DescriptorsAndNams.py*** modules are returned to. It also contains functions that enable the application of the FullQSAR model.

- ***ToolFilter.py*** - Contains the results from the ***Model80.py*** and *DescriptorsAndNamsFilter.py*. It also contains the functions needed to perform the FullQSAR model;

- ***ToolSimple.py*** - Contains the functions needed to determine the prediction of the pIC50 using the FullQSAR model. This module does not perform a similarity analysis.

The ***Tool.py*** and ***ToolFilter.py*** scripts include an algorithm to decide if and which prediction model or models should be computed. The determination process goes as follows:

1. Similarity analysis

2. Assessment of the molecules presence in the database. If the molecule is present the pIC50 is returned to the browser and the tools execution ends.

3. Determination of the highest similarity value.

   (a) If the similarity is lower than 80% the FullQSAR model is calculated, only the highest similarity value found and the predicted pIC50 is returned.

   (b) If the similarity value is equal or higher than 80% the FullQSAR, the QSAR80 and the WM80 models are performed. The predicted pIC50 for each model is returned as well as the highest similarity value found and a table with information regarding the molecules with similarity $\geq$80%.

The ***ToolSimple.py*** only performs the prediction of the pIC50 using the FullQSAR model. The ***Tool.py***, ***ToolFilter.py*** and ***ToolSimple.py*** files concatenate all of the results and send them to a PHP script were an array with the output is created and sent to the browser thus returning the pIC50 value and other information to the user. To complement this tool, XLM and JSON web services were also implemented.

7.3 User Interface

The hERGIP interface was based of the B3PP webtool (Martins *et al.*, 2012) due to its simplicity and easiness to use. Alterations were performed to better suit the tools intended use, these changes were done mainly in:

- The molecular input formats, hERGIP only supports InCHI and SMILES;
- The results output. If the highest similarity found is at least 80% an output table with the FullQSAR, QSAR80 and WM80 predicted pIC50 and IC50 interval is shown.
- A second output table is also shown if the QSAR80 and WM80 are used, this table presents all the molecules in our database that have at least a similarity score of 80%.
- The button options. hERGIP allows for "Fast QSAR", "Filtered Similarity" and "Full Similarity" to predict the pIC50.

The second result table includes a 2D molecular representation, the molecule's identifier, which corresponds to the CHEMBL ID or a personal identifier in case the molecule was retrieved from Sinha & Sen (2011) and Su *et al.* (2010), the molecular SMILES, the similarity score, the molecule's pIC50 and the literature reference from which the molecule's information was retrieved. Figures 7.2, 7.3 and 7.4 show the hERGIP input and different outputs when selecting the "Fast QSAR" and "Full Similarity" buttons.

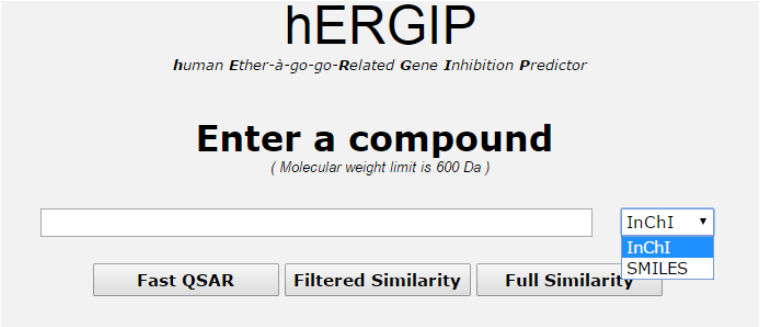

Figure 7.2: **hERGIP input**. hERGIP allows the selection of a "Fast QSAR", "Filtered Similarity" or "Full Similarity" for the pIC50 prediction.

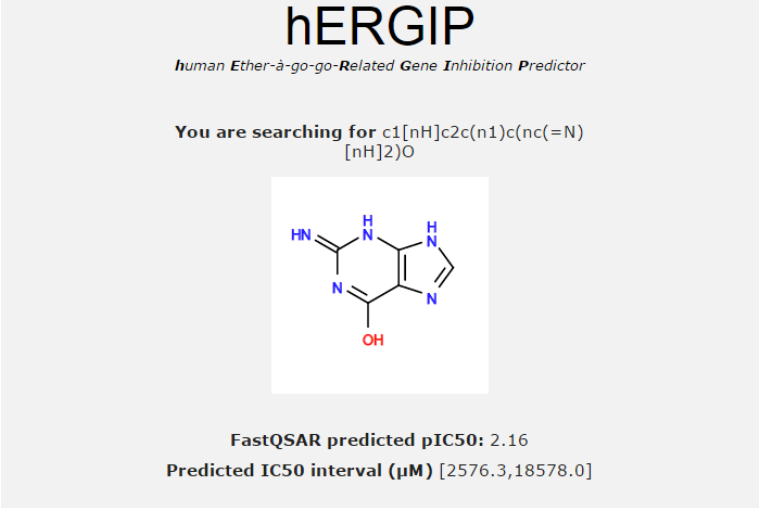

Figure 7.3: **hERGIP output - Fast QSAR**. The pIC50 and IC50 are displayed in a line notation when the input molecule is not at least 80% similar with other/others in the hERGIP database.

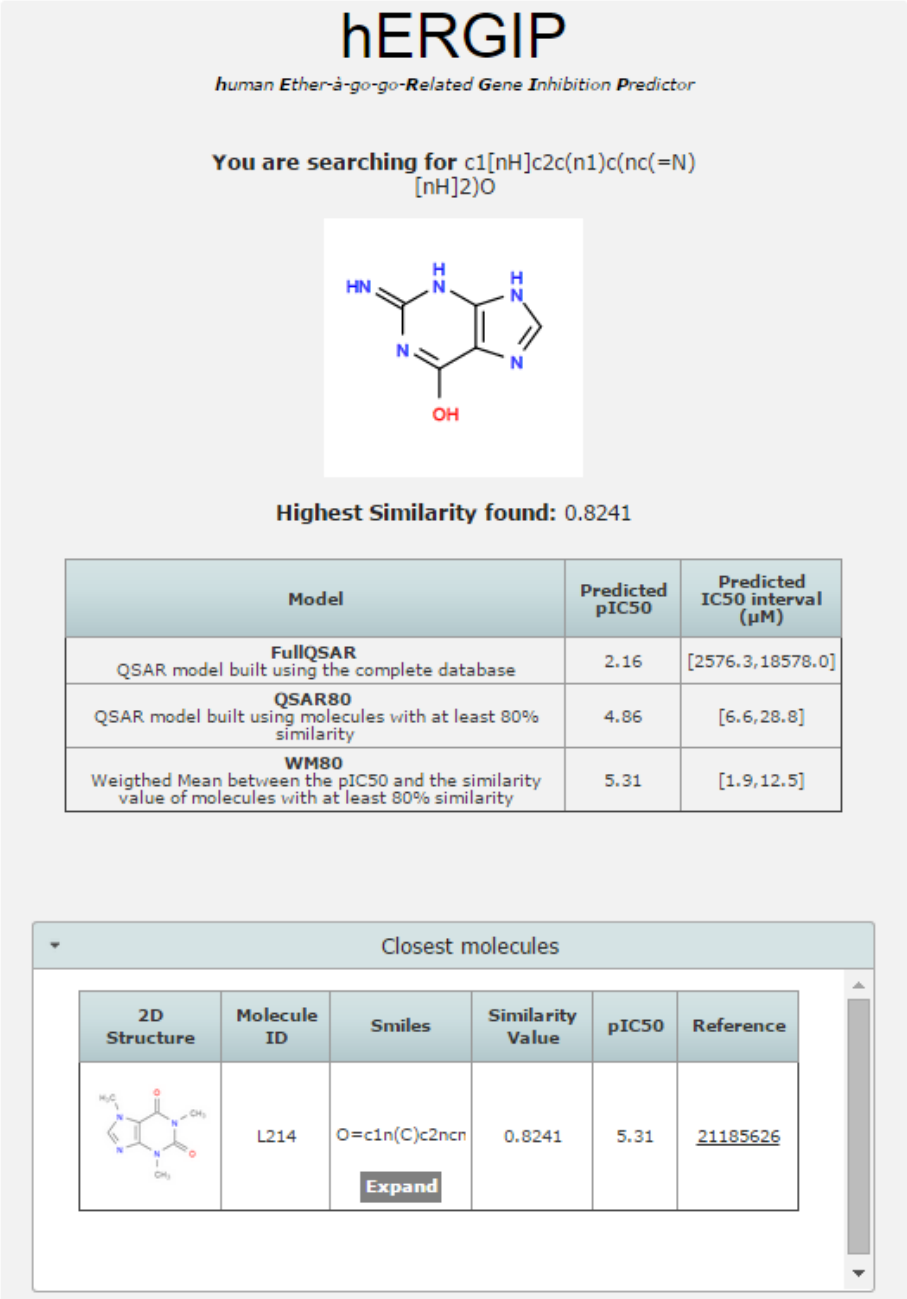

Figure 7.4: **hERGIP output - Full Similarity**. The pIC50 and IC50 prediction results are presented in a table since the highest similarity found was above 80%. The Closest Molecules section show a table with the molecules present in the data collection which have at least 80% similarity with the input molecule.

7.4 Prediction Tools Comparison

In this section the hERGIP and the Pred-hERG[1] tools will be compared. Although both applications allow to predict hERG inhibitors Pred-hERG distinguishes itself by applying a classification of Blocker or Non-Blocker instead of predicting the pIC50. I selected 30 molecules[2], with a wide range of pIC50, from the IVS to be tested in Pred-hERG.

To allow a comparison between the tools, the following thresholds were adapted from Pred-hERG (Braga *et al.*, 2015):

- Non-Blocker: pIC50 <5
- Weak/Moderate Blocker: 5 $\geqslant$ pIC50 $\geqslant$ 6
- Strong Blocker: pIC50 >6

Table 7.1 shows the classification results obtained from the hERGIP QSAR80 prediction and the Pred-hERG Multiclass prediction. A table with the full results is present in appendix C.

Table 7.1: **Comparison between hERGIP and Pred-hERG.** The QSAR80 model from hERGIP and the Multiclass model from Pred-hERG were used to achieve the results presented. The **NB**, **WB** and **SB** correspond to Non-Blocker, Weak Blocker and Strong Blocker, respectively.

	hERGIP			Pred-hERG		
Reality	**NB**	**WB**	**SB**	**NB**	**WB**	**SB**
NB	**9**	1	0	**9**	1	0
WB	0	**8**	1	0	**7**	2
SB	0	1	**10**	0	0	**11**

Pred-hERG became available in July 2015 which allowed for a comparison between hERG inhibition prediction tools which has not been done before to our knowledge. It must be taken into consideration that hERGIP uses regression prediction models and that the Pred-hERG models were trained using molecules

[1] http://labmol.farmacia.ufg.br/predherg/
[2] The complete ordered list in SMILES is present in Appendix C

collected from ChEMBL which may be a source of bias since the selected 30 molecules used for this test include compounds from the same database.

The results show that both tools misclassified only 3 out of 27 molecules. hERGIP and Pred-hERG both predict a Non-Blocker as a Weak-Blocker. However, they exhibit differences when predicting Weak Blockers and in the case of Strong Blockers Pred-hERG correctly identified all the molecules while hERGIP misclassified one compound however, hERGIP was more coherent when identifying Weak Inhibitors.

The results suggest that a regression prediction hERGIP performs at the same level as a qualitative classification tool such as Pred-hERG. This supports that the methodology implemented, in this thesis, for prediction of hERG inhibitors is a suitable and more useful approach since it can provide more information instead of only classifying molecules using a limited number of classes. This makes hERGIP a novel tool regarding the prediction of hERG inhibition.

The increasing occurrence of hERG prediction tools offers the community a choice regarding the type of inhibition prediction result, it also allows for a more efficient testing of the prediction tools which contributes to their improvement.

Chapter 8

Conclusion

Due to its importance, the drug inhibition of the hERG channel has been previously addressed in other studies and a variety of methodologies to classify a molecule as an inhibitor or a non-inhibitor have been proposed (Braga *et al.*, 2014; Thai & Ecker, 2008; Wang *et al.*, 2013). The current standard approach is based on a electrophysiological method however, *in silico* approaches are preferred since they're simple, fast and inexpensive. With this work I decided to focus on the variable reduction method and the application of molecular similarity to develop prediction models for hERG inhibition assessment.

I proposed new methods to be applied for variable selection, which used as variables different molecular descriptors collections and also tested the model's improvement when adding molecular fingerprints. The tests applied to verify each methodology performance were pivotal to identify the best descriptors and methods to achieve the best prediction models. This resulted in the selection of a new method which performs an assessment of each variable using LR and selects the best collection of variables through the application of SVM. The increased results when applying fingerprints opposite only using descriptor collections suggested the importance of structural patterns for bioactivity prediction.

It was suggested that the use of structural similarity could improve the prediction models, therefore, several methodologies based on the similarity value calculated by NAMS were tested. Two methodologies were implemented to test the application of structural similarity. Despite not being used in the final models, they were an important step for the conceptualization of the final approach.

The method which applied the combination of molecular descriptors and kernels selected with an 80% threshold performed the best, which led to the application of a pre-selection of molecules using the 80% and 90% similarity thresholds. The QSAR models built with this approach were validated and suggest that the initial hypothesis was correct, the models which reduced the molecular dataset to a coherent collection of molecules with a high similarity in structure did perform better than the Standard QSAR model. An exception to this was verified when applying the IVS to the QSAR90. This model returned the lowest results and I hypothesised that this result is due to the high difference between the number of the selected molecules in the train and test sets.

I also decided to apply an approach based on the weighted mean between the pIC50 and the similarity value. The results achieved by this approach suggested that even without the application of fingerprints and descriptors the prediction models performed well. This further supports the importance of structural similarity to predict a given molecule's biological outcome.

With this thesis I proposed solutions for the assessment of hERG inhibitors through the application of prediction models with a previous variable reduction approach. The suggested solutions included the application of molecular similarity for models selection and for the determination of new descriptors - the similarity value and the pIC50 of the closest molecule - and an approach based on the weighted mean of the pIC50 using the similarity values as weights. These methods were validated with standard validation techniques and enabled the selection of 5 prediction models (i.e. FullQSAR, QSAR80, QSAR90, WM80 and WM90).

The development of hERGIP enabled the comparison of the produced models with the Pred-hERG prediction tool which suggested that the regression models applied by hERGIP performed as well as the classification from Pred-hERG. This makes hERGIP a novel tool which applies regression to predict the pIC50 and is able to perform similarly to common qualitative prediction tools.

In this thesis I addressed the challenges of developing new methods to predict the inhibition of the hERG channel. Due to the increasing need to develop new drugs, methods which aim to facilitate the drug R&D process, such as the

ones proposed in this thesis, are becoming even more important. I was able to successfully achieve my objectives since I performed a analysis which helped to identify the best variable selection method, applied QSAR methodology to develop prediction models and I applied molecular similarity to improve QSAR models. In addition to this I also developed hERGIP using the best predication models achieved. Applying *in silico* approaches for prediction of the biological behaviour of drugs is becoming even more crucial and implementing a tool such as hERGIP in the Drug R&D process can have a positive impact and reduce the time of this, currently, slow as costly process. Therefore, I believe that the work presented is a meaningful contribution to this field an if it is applied correctly it can impact greatly the Drug R&D process.

8.1 Future Work

The work developed in this thesis gave insight to new methodologies regarding *in silico* drug inhibition assessment and it can be extended by the application of some new approaches and improvements. To improve this work I suggest the following:

- Improvement of the molecular dataset, a more diverse set of molecules can provide the ability to determine more unique differences between strong and weak inhibitors.
- Further application of molecular similarity for new approaches regarding the prediction of bioactivity. With this research I hoped to instigate the increase application of structural similarity in future drug development approaches since it is suggested to be a useful measure.
- Further improvement of the hERGIP prediction tool. Improvements would include the option to predict a set of molecules and the ability to add new compounds and their experimental pIC50 values to its molecular database.

Appendix A

Variable Selection Method

The algorithm constructed for the selection of the best variable collection using LR and SVM is presented in pseudocode and it is divided into three segments. The pseudocode represented in pseudocode A.1 allows for the computation of the $\boldsymbol{R}^2$ value of each molecular descriptor and fingerprint trough a LR model.

Algorithm A.1: R^2 calculation for each column variable

Input: Two matrices m and t with n columns and l lines
Output: A variable reduction model

```
1  v ← 0
2  i = 2
3  for element in n do
4  |  if i<n then
5  |  |  dataTrain ← m[][(1, i)]
6  |  |  dataTest ← t[][(1, i)]
7  |  |  build a linear regression model with dataTrain
8  |  |  apply predict to the model built and validate using dataTest
9  |  |  calculate R²
10 |  |  if R² ≥ 0 then
11 |  |  |  v ← v + R²
12 |  |  end
13 |  end
14 |  i ← i + 1
15 end
16 vo ← order v
```

The selection of the best $\boldsymbol{R}^2$ to be used as a threshold for the selection of the best collection of variables is presented in pseudocode A.2 and the model built using this threshold is obtained when applying pseudocode A.3.

Algorithm A.2: Calculation of best R^2 threshold for variable selection

```
1  vf ← 0
2  ncol ← 0
3  for value in vo do
4  |  i = 2
5  |  for element in n do
6  |  |  if i<n then
7  |  |  |  dataTrain ← m[][(1, i)]
8  |  |  |  dataTest ← t[][(1, i)]
9  |  |  |  build a linear regression model with dataTrain
10 |  |  |  apply predict to the model built and validate using dataTest
11 |  |  |  calculate R^2;
12 |  |  |  if R^2 ≥ value then
13 |  |  |  |  ncol ← ncol + i
14 |  |  |  end
15 |  |  |  i ← i + 1
16 |  |  end
17 |  end
18 |  dataTrain ← m[][(1, ncol)]
19 |  dataTest ← m[][(1, ncol)]
20 |  build a support vector machine model with dataTrain
21 |  apply predict to the model built and validate using dataTest
22 |  calculate R^2
23 |  vf ← vf + R^2
24 end
25 vfo ← order vf
26 best ← vfo[1]
```

Algorithm A.3: Retrieval of the best variable reduction model

```
1  i = 2
2  bestcols ← 0
3  for element in n do
4  |  if i<n then
5  |  |  dataTrain ← m[][(1, i)]
6  |  |  dataTest ← t[][(1, i)]
7  |  |  build a linear regression model with dataTrain
8  |  |  apply predict function to the model built and validate using
   |  |  dataTest
9  |  |  calculate R^2
10 |  |  if R^2 ≥ best then
11 |  |  |  bestcols ← bestcols + i
12 |  |  end
13 |  end
14 |  i ← i + 1
15 end
16 dataTrain ← m[][(1, bestcols)]
17 dataTest ← t[][(1, bestcols)]
18 build a support vector machine model with dataTrain
19 apply predict to the model built and validate using dataTest
```

Appendix B

ChEMBL Molecule Retrieval

The following Python code specifies the restrictions used for the selection of molecules from ChEMBL, this script was adapted from Czodrowski (2013). The script returns a *Chembl-Results.txt* file with the selected molecules information.

```
import urllib2
import json
import re
from sys import argv

codes = {'herg_human' : 'Q12809' }

def looks_like_number(x):
    try:
        float(x)
        return True
    except ValueError:
        return False

def QueryChembl(accession=None):
    '''
    Query chembl
    '''
    print """
```

```
20      # ============================================================
21      # 1. Use UniProt accession to get target details
22      # ============================================================
23      """
24      if not accession:
25          accession = argv[1]
26
27      if accession.find("CHEMBL") == -1:
28        target_data = urllib2.urlopen("https://www.ebi.ac.uk/chemblws/
              targets/uniprot/%s.json" % accession).read()
29        target_data = json.loads(target_data)
30
31      else:
32        target_data = {}
33        target_data['target'] = {}
34        target_data['target']['chemblId'] = accession
35
36      print """
37
38      # ============================================================
39      # 2. Get all bioactivties for target CHEMBL_ID
40      # ============================================================
41      """
42
43      bioactivity_data = json.loads(urllib2.urlopen("https://www.ebi.ac.uk/
            chemblws/targets/%s/bioactivities.json" % target_data['target']['
            chemblId']).read())
44
45      print "Bioactivity Count:        %d" % len(bioactivity_data['
            bioactivities'])
46      print "Bioactivity Count (IC50): %d" % len([record for record in
            bioactivity_data['bioactivities'] if record['bioactivity_type']
            == 'IC50'] )
47
48      print """
49
```

```
50      # ============================================================
51      # 3. Get compounds with binding affinity (IC50)
52      # ============================================================
53      """
54      ic50_skip=0
55      inhb_skip=0
56      FinalList=[]
57      dr={}
58      for bioactivity in [record for record in bioactivity_data['
            bioactivities'] if looks_like_number(record['value']) ] :
59
60        if re.search('IC50', bioactivity['bioactivity_type']):
61          if bioactivity['units'] != 'nM':
62            ic50_skip+=1
63            continue
64        elif re.search('Inhibition', bioactivity['bioactivity_type']):
65          inhb_skip+=1
66        else:
67          continue
68
69        try:
70            cmpd_data = json.loads(urllib2.urlopen("https://www.ebi.ac.uk/
                chemblws/compounds/%s.json" % bioactivity['
                ingredient_cmpd_chemblid']).read())
71            my_smiles = cmpd_data['compound']['smiles']
72            if bioactivity['bioactivity_type']=="IC50":
73                bioactivity['Smiles']=my_smiles
74                dr[count] = bioactivity
75                count+=1
76                FinalList.append([bioactivity['Smiles'],bioactivity['
                    bioactivity_type'],bioactivity["operator"],bioactivity['
                    value'],bioactivity['units'],bioactivity["assay_type"],
                    bioactivity["assay_description"],bioactivity["reference"
                    ],bioactivity["assay_chemblid"]])
77        except:
78            print " compound not found", bioactivity['
```

```
            ingredient_cmpd_chemblid']
79
80
81      print "Skipped %i IC50 values" % ic50_skip
82      print "Skipped %i Inhibition values" % inhb_skip
83
84
85      out=open("Chembl-Results.txt","wt")
86      for l in FinalList:
87          for elem in l:
88              out.write(str(elem))
89              out.write("\t")
90          out.write("\n")
91
92
93  if __name__ == '__main__':
94      for name,accession in codes.items()[:]:
95          print "Checking ", name
96          QueryChembl(accession)
```

Appendix C

Molecules for Tool Comparison

This appendix includes the ordered list of molecules used in Chapter 7 and a table with all the results from the comparison between hERGIP and Pred-hERG.

List of Molecules:

CN1[C@@H]2CC[C@H]1[C@@H]([C@H](C2)OC(=O)c1ccccc1)C(=O)O
CCN1C=C(C(=O)O)C(=O)c2cc(F)c(N3CCNC(C)C3)c(F)c12
CC(C)NC[C@H](O)COc1ccc(CC(=O)N)cc1
F[C@@H]1CN(CCN2C(=O)C=Cc3ccc(cc23)C#N)CC[C@@H]1NCc4cc5OCCOc5cn4
COc1c2c(cc(c1N1C[C@@H]3CCCN[C@@H]3C1)F)c(=O)c(cn2C1CC1)C(=O)O
Fc1c(N2CC(NCC2)C)c(OC)c2n(C3CC3)cc(c(=O)c2c1)C(=O)O
C[C@@H](CO)Oc1cc(Oc2ccc(cc2)C(=O)N3CCC3)cc(c1)C(=O)Nc4ccn(n4)C(C)C
O[C@@H](C1CC1)C(=O)N2CC(=C[C@H]2c3cccc(O)c3)c4cc(F)ccc4F
CC(C)Oc1ccc(cc1C#N)c2onc(n2)c3cnc4CN(CCc4c3C)C(CO)CO
NC(=O)c1ccc(cc1)C2=CC3(CCNCC3)Oc4ccccc24
CC(C)[C@@](CCCN(C)CCc1ccc(c(c1)OC)OC)(C#N)c1cc(c(c(c1)OC)OC)OC
C[C@@H]1CCCN1CCN2CCc3cc(ccc3C2=O)c4ccc(cc4)C(=O)N5CCCC5
CC(C)N(C(C)C)C(=O)c1ccc(cc1)C2=CC3(CCNCC3)Oc4ccccc24
S1C(=O)NN=C(c2ccc3N(CCCc3c2)/C(=N/CC)/c2ccc(OC)c(OC)c2)C1C
Oc1ccccc1[C@@H]2CC[C@@H](CC2)N3CC(C3)NC(=O)CNc4ncnc5ccc(cc45)C(F)(F)F
CCOC(=O)N1CCC(CN2CCC3(CC2)CC(=O)Nc4ncccc34)CC1
Fc1cc(OCCC2CC2C3CCN(CC3)c4ncc(Cl)cn4)ccc1C(=O)NC5CC5
Fc1ccc(n2c3c(c(c2)C(CC)CC)cc(cc3)Cl)cc1

C. MOLECULES FOR TOOL COMPARISON

[C@H]1(CNC(=O)[C@@]21CCN(CC2)C1(CCCCC1)c1ccc(cc1)F)c1ccc(cc1)F
Fc1cc(ccn1)C(NC(=O)[C@@H]2CC[C@H](C[C@H]2c3ccc(Br)cc3)N4CCOCC4)c5ccc(Cl)cc5
CN1CCCN(CC(=O)Nc2cc(nc(n2)c3oc(C)cc3)n4nc(C)cc4C)CC1
C(c1ccc(n2c3c(c(c2)C2CCN(CC2)CCN2C(=O)NCC2)cc(cc3)Cl)cc1)C(=O)OC
COc1cc(CN2CCC(CNCCCCCC(c3ccc(F)cc3)c4ccc(F)cc4)C2)cc(OC)c1OC
CCCOCCN1C(=O)C(=Nc2cnc(cc12)c3ccc(OC)nc3)N4CCN(CC4)[C@@H](C)[C@@H](C)O
CCCCCCCN(CC)CC#CCCc1ccc(Cl)cc1
S(=O)(=O)(Nc1ccc(C(O)CCCN(CCCCCCC)CC)cc1)C
c1c(ccc(c1)F)n1cc(c2c1cccc2)C1CCN(CC1)CCN1C(=O)NCC1
CC(C)(N1[C@@H]2CC[C@H]1C[C@@H](C2)Oc3cccc(c3)C(=O)N)c4ccccc4
CC(N1[C@@H]2CC[C@H]1C[C@@H](C2)Oc3cccc(c3)C(=O)N)c4ccc(C)s4
NC(=O)c1cccc(O[C@@H]2C[C@H]3CC[C@@H](C2)N3CCc4ccccc4)c1

Table C.1: **Comparison of hERGIP and Pred-hERG**

	Molecule	pIC50	FullQSAR	QSAR80	WM80	Pred-hERG Binary	Pred-hERG Multiclass
	1	2.40	3.42	3.51	5.52	Non-Blocker	Non-Blocker
	2	2.62	3.31	3.40	2.62	Non-Blocker	Non-Blocker
	3	3.00	4.48	4.21	3.84	Blocker	Weak/Moderate Blocker
	4	3.63	4.10	4.01	4.16	Non-Blocker	Non-Blocker
	5	3.89	3.94	3.94	3.70	Non-Blocker	Non-Blocker
	6	3.89	4.00	4.04	3.89	Non-Blocker	Non-Blocker
	7	4.51	4.59	4.53	4.44	Blocker	Non-Blocker
	8	4.74	4.92	4.94	5.17	Non-Blocker	Non-Blocker
	9	4.90	4.94	4.84	4.93	Non-Blocker	Non-Blocker
	10	4.94	5.60	6.14	5.35	Non-Blocker	Non-Blocker
	11	5.01	6.34	6.52	6.87	Blocker	Strong Blocker
	12	5.02	5.01	5.01	5.25	Blocker	Weak/Moderate Blocker
	13	5.17	5.72	5.58	5.12	Blocker	Weak/Moderate Blocker
	14	5.18	5.04	5.18	4.86	Blocker	Weak/Moderate Blocker
	15	5.18	5.59	5.50	5.88	Blocker	Weak/Moderate Blocker
	16	5.55	5.40	5.64	5.86	Blocker	Weak/Moderate Blocker
	17	5.59	5.73	5.98	5.92	Blocker	Weak/Moderate Blocker
	18	5.83	5.50	5.43	5.39	Blocker	Strong Blocker
	19	5.89	5.81	5.97	5.83	Blocker	Weak/Moderate Blocker
	20	6.23	6.02	6.01	6.07	Blocker	Strong Blocker
	21	6.55	6.18	6.06	6.05	Blocker	Strong Blocker
	22	6.88	6.30	6.54	7.41	Blocker	Strong Blocker
	23	7.00	7.11	7.10	6.73	Blocker	Strong Blocker
	24	7.27	5.57	5.75	6.30	Blocker	Strong Blocker
	25	7.28	7.46	7.29	7.41	Blocker	Strong Blocker
	26	7.82	7.32	7.30	7.82	Blocker	Strong Blocker
	27	8.16	6.55	7.61	7.48	Blocker	Strong Blocker
	28	8.33	7.33	7.23	7.43	Blocker	Strong Blocker
	29	8.80	7.19	7.77	7.68	Blocker	Strong Blocker
	30	9.42	7.88	8.53	7.40	Blocker	Strong Blocker
MSE	NA	NA	0.635	0.473	0.793	NA	NA

References

Albright, S., Winston, W. & Zappe, C. (2008). *Data Analysis and Decision Making with Microsoft Excel, Revised*. Cengage Learning, pages 321-323. 31

Andrew Dalke Scientific (2008). Molecular fingerprints. `http://www.dalkescientific.com/writings/diary/archive/2008/06/26/fingerprint_background.html`, accessed September 2015. 14

Bajorath, J. (2004). *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*. Biomed Protocols, Humana Press, page V. 9

Bender, A. & Glen, R.C. (2004). Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry*, **2**, 3204–3218. 16

Bharath, E.N., Manjula, S.N. & Vijaychand, a. (2011). In silico drug design-tool for overcoming the innovation deficit in the drug discovery process. *International Journal of Pharmacy and Pharmaceutical Sciences*, **3**, 8–12. 5, 16

Braga, R.C., Alves, V.M., Silva, M.F.B., Muratov, E., Fourches, D., Tropsha, A. & Andrade, C.H. (2014). Tuning hERG Out: Antitarget QSAR Models for Drug Development. *Current Topics in Medical Chemistry*, **14**, 1399–1415. 5, 21, 63

Braga, R.C., Alves, V.M., Silva, M.F.B., Muratov, E., Fourches, D., Lião, L.M., Tropsha, A. & Andrade, C.H. (2015). Pred-hERG: A Novel web-Accessible Computational Tool for Predicting Cardiac Toxicity. *Molecular Informatics*. 61

Czodrowski, P. (2013). hERG me out. *Journal of chemical information and modeling*, **53**, 2240–2251. 21, 24, 33, 71

Daylight Chemical Information Systems (2011). Daylight Theory Manual. `http://www.daylight.com/dayhtml/doc/theory/`, accessed March 2015. 12

Dempsey, C.E., Wright, D., Colenso, C.K., Sessions, R.B. & Hancox, J.C. (2014). Assessing hERG pore models as templates for drug docking using published experimental constraints: the inactivated state in the context of drug block. *Journal of chemical information and modeling*, **54**, 601–612. 4

DiMasi, J.A., Feldman, L., Seckler, A. & Wilson, A. (2010). Trends in risks associated with new drug development: Success rates for investigational drugs. *Clinical Pharmacology & Therapeutics*, **87**, 272–277. 3

Ekins, S., Balakin, K.V., Savchuk, N. & Ivanenkov, Y. (2006). Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning and kohonen and sammon mapping techniques. *Journal of Medicinal Chemistry*, **49**, 5059–5071. 20

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22. 25

Frolov, R.V., Ignatova, I.I. & Singh, S. (2011). Inhibition of HERG potassium channels by celecoxib and its mechanism. *PloS one*, **6**, 468–481. 5

Gasteiger, J. & Funatsu, K. (2006). Chemoinformatics – An Important Scientific Discipline. *Journal of Computer Chemistry, Japan*, **5**, 53–58. 9, 10

Guha, R. (2007). Chemical informatics functionality in r. *Journal of Statistical Software*, **18**. 25

Huang, X.P., Mangano, T., Hufeisen, S., Setola, V. & Roth, B.L. (2010). Identification of human Ether-à-go-go related gene modulators by three screening platforms in an academic drug-discovery setting. *Assay and drug development technologies*, **8**, 727–742. 5

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2014). *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*. Springer Texts in Statistics, Springer New York, pages 3-4. 29

Kaitin, K.I. (2010). Deconstructing the Drug Development Process: The New Face of Innovation. *Clin Pharmacol Ther*, **87**, 356–361. 3

Leach, A.R. & Gillet, V.J. (2003). *An Introduction to Chemoinformatics*. Dordrecht Kluwer Academic Publishers, pages 1-2. 10

Liaw, A. & Wiener, M. (2002). Classification and regression by randomforest. *R News*, **2**, 18–22. 25

Maggiora, G., Vogt, M., Stumpfe, D. & Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, **57**, 3186–3204. 15

Martins, I.F., Teixeira, A.L., Pinheiro, L. & Falcao, A.O. (2012). A Bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling*, **52**, 1686–1697. 58

Meyer, D. (2015). Support vector machines - the interface to libsvm in package e1071. Tech. rep., Technische Universität Wien. 30

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. (2014). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-4. 25

Mitchell, J.B.O. (2014). Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, **4**, 468–481. 17

Mitchell, T.M. (1999). Machine Learning and Data Mining Over the past. *Communications of the ACH*, **42**, 31–36. 1

Moroy, G., Martiny, V.Y., Vayer, P., Villoutreix, B.O. & Miteva, M.a. (2012). Toward in silico structure-based ADMET prediction in drug discovery. *Drug discovery today*, **17**, 44–55. 16

Morris, G. & Lim-Wilby, M. (2008). Molecular docking. In A. Kukol, ed., *Molecular Modeling of Proteins*, vol. 443 of *Methods Molecular Biology™* , 365–382, Humana Press. 4

Nikolova, N. & Jaworska, J. (2003). Approaches to Measure Chemical Similarity– a Review. *QSAR Combinatorial Science*, **22**, 1006–1026. 16

O'Boyle, N.M., Morley, C. & Hutchison, G.R. (2008). Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chemistry Central journal*, **2**, 5. 24

O'Boyle, N.M., Banck, M., James, C.a., Morley, C., Vandermeersch, T. & Hutchison, G.R. (2011). Open Babel: An Open chemical toolbox. *Journal of Cheminformatics*, **3**, 33. 24

Palmer, D.S., O'Boyle, N.M., Glen, R.C. & Mitchell, J.B.O. (2007). Random forest models to predict aqueous solubility. *Journal of chemical information and modeling*, **47**, 150–158. 31

Puzyn, T., Leszczynski, J. & Cronin, M. (2010). *Recent Advances in QSAR Studies: Methods and Applications*. Challenges and Advances in Computational Chemistry and Physics, Springer Netherlands, pages 29-35. 13

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. 25

Riniker, S. & Landrum, G.a. (2013). Similarity maps - A visualization strategy for molecular fingerprints and machine-learning methods. *Journal of Cheminformatics*, **5**, 1–7. 14

Roche, O., Trube, G., Zuegge, J., Pflimlin, P., Alanine, A. & Schneider, G. (2002). A virtual screening method for prediction of the hERG potassium channel liability of compound libraries. *ChemBioChem*, **3**, 455–459. 19

Schneider, G. & Downs, G. (2003). Editorial: Machine Learning Methods in QSAR Modelling. *QSAR & Combinatorial Science*, **22**, 485–486. 17, 29

Sinha, N. & Sen, S. (2011). Predicting hERG activities of compounds from their 3D structures: development and evaluation of a global descriptors based QSAR model. *European journal of medicinal chemistry*, **46**, 618–630. 20, 33, 58

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, **43**, 493–500. 25

Su, B.H., Shen, M.y., Esposito, E.X., Hopfinger, A.J. & Tseng, Y.J. (2010). In silico binary classification QSAR models based on 4D-fingerprints and MOE descriptors for prediction of hERG blockage. *Journal of chemical information and modeling*, **50**, 1304–1318. 33, 58

Talete SRL (2013). Dragon6. `http://www.talete.mi.it/products/dragon_description.htm`, accessed June 2015. 13

Teixeira, A.L. & Falcao, A.O. (2013). Noncontiguous atom matching structural similarity function. *Journal of chemical information and modeling*, **53**, 2511–2524. 16

Teixeira, A.L. & Falcao, A.O. (2014). Structural similarity based kriging for quantitative structure activity and property relationship modeling. *Journal of chemical information and modeling*, **54**, 1833–1849. 17

Tetko, I.V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., Palyulin, V.A., Radchenko, E.V., Zefirov, N.S., Makarenko, A.S., Tanchuk, V.Y. & Prokopenko, V.V. (2005). Virtual Computational Chemistry Laboratory – Design and Description. *Journal of Computer-Aided Molecular Design*, **19**, 453–463. 26

Thai, K.M. & Ecker, G.F. (2008). A binary QSAR model for classification of hERG potassium channel blockers. *Bioorganic & medicinal chemistry*, **16**, 4107–19. 5, 63

Thai, K.M. & Ecker, G.F. (2009). Similarity-based SIBAR descriptors for classification of chemically diverse hERG blockers. *Molecular diversity*, **13**, 321–336. 20

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288. 27

Tobita, M., Nishikawa, T. & Nagashima, R. (2005). A discriminant model constructed by the support vector machine method for HERG potassium channel inhibitors. *Bioorganic & medicinal chemistry letters*, **15**, 2886–2890. 20

Todeschini, R., Consonni, V., Mannhold, R., Kubinyi, H. & Timmerman, H. (2008). *Handbook of Molecular Descriptors*. Methods and Principles in Medicinal Chemistry, Wiley, pages 303-305. 13

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*, **29**, 476–488. 17

Wadood, A., Ahmed, N., Shah, L., Ahmad, A., Hassan, H. & Shams, S. (2013). In-silico drug design : An approach which revolutionarised the drug discovery process. *OA Drug Design & Delivery*, **1**, 1–4. 4

Wang, S., Li, Y., Xu, L., Li, D. & Hou, T. (2013). Recent developments in computational prediction of HERG blockage. *Current topics in medicinal chemistry*, **13**, 1317–1326. 5, 63

Wang S., W.J.C.L.Z.L.Y.H., Li Y. & T., H. (2013). ADMET Evaluation in Drug Discovery. 12. Development of Binary Classification Models for Prediction of hERG Potassium Channel Blockage. *Journal of Molecular Modeling*, **9**, 996–1010. 5

Wiśniowska, B., Mendyk, A., Fijorek, K. & Polak, S. (2014). Computer-based prediction of the drug proarrhythmic effect: problems, issues, known and suspected challenges. *Europace*, **16**, 724–735. 3

Witchel, H.J. (2011). Drug-induced hERG block and long QT syndrome. *Cardiovascular Therapeutics*, **29**, 251–259. 4

Zhang, K.p., Yang, B.f. & Li, B.x. (2014). Translational toxicology and rescue strategies of the hERG channel dysfunction : biochemical and molecular mechanistic aspects. *Nature Publishing Group*, **35**, 1473–1484. 4