UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Combining metric and vector space data mining methods for screening CFTR rescuers in cystic fibrosis

Inês Cristina Ferreira Coelho

MESTRADO EM BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL
ESPECIALIZAÇÃO EM BIOINFORMÁTICA

Dissertação orientada por:

Prof. Doutor André Osório e Cruz de Azerêdo Falcão

2017

Resumo

O objetivo principal desta dissertação é desenvolver modelos para propor moléculas de interesse que podem se tornar em princípios ativos no tratamento da fibrose quística. Neste projeto apresenta-se uma abordagem *in silico* para a seleção de moléculas que possivelmente têm capacidades terapêuticas em relação à fibrose quística. Este processo é efetuado computacionalmente com a utilização de ferramentas de prospeção de dados e os objetivos primordiais deste processo são a identificação e seleção de moléculas que podem ajudar no combate à doença para posterior teste em laboratório. Para este efeito foram desenvolvidos previsões para a capacidade terapêutica das moléculas em dois espaços: i) espaço vetorial; ii) espaço métrico. No espaço vetorial as previsões foram realizadas tendo em conta os descritores moleculares das moléculas, com recurso ao método estatístico e computacional *random forest*. As moléculas foram também representadas num espaço métrico construído com a dissemelhança molecular entre as mesmas, onde ocorreu uma redução de dimensões tornando possível a representação das instâncias num plano bidimensional - este espaço métrico foi subsequentemente analisado por uma ferramenta estatística denominada kriging.

Para comprovar os métodos escolhidos, usaram-se dois conjuntos de dados; potenciadores e ativadores de anoctaminas (moléculas de possível interesse para o tratamento da fibrose quística) e corretores da proteína causadora da fibrose quística (CFTR - *Cystic fibrosis transmembrane conductance regulator*). No âmbito desta dissertação, foram identificadas 10 moléculas provenientes do estudo com potenciadores de anoctaminas e 18 moléculas provenientes do estudo com corretores da CFTR, para serem testadas em laboratório.

Adicionalmente, foram recolhidos dados de repositórios de informação biológica para validar os métodos utilizados. Este passo adicional permite concluir que algumas das moléculas escolhidas têm ligações diretas e indiretas à fibrose quística, dando credibilidade ao método desenvolvido.

É importante referir que a forma como este projeto foi desenvolvido permite a utilização de diferentes conjuntos de dados de ligandos para proteínas alvo, o que torna este método flexível e adaptável à doença que seja objeto de estudo.

Palavras Chave: Chemoinformatics, Aprendizagem Automática, Fibrose Quística, Design de Medicamentos In Silico

Abstract

The main goal of this dissertation is to develop models in order to identify and propose lead chemical molecules that can possibly become principal actives against the cystic fibrosis disease.

In this project it is presented an *in silico* approach to perform a molecular screening on possible therapeutic candidates for the cystic fibrosis disease. This process is done computationally with data mining tools and the main objectives are the identification and selection of molecules to further testing in the laboratory. To achieve this goal the data mining exercise was developed on two spaces: i) vectorial space, ii) metric space. On the vectorial space, molecular descriptors were selected to implement a random forest algorithm (a supervised machine learning method) in order to realize forecasts on the molecule ability to treat the disease. The studied molecules were also represented in a metric space that was developed using molecular dissimilarity between all molecules. This dissimilarity values were modelled to fit in a 2 dimensional representation - In this metric space the statistical tool chosen was kriging.

To prove the chosen methodology, two main datasets were used: A dataset with Anoctamin activators or potentiators (molecules of interest to treat the cystic fibrosis disease) and a dataset with correctors of the protein which causes cystic fibrosis (CFTR - Cystic fibrosis transmembrane conductance regulator). Based on these datasets, 10 and 18 molecules were selected respectively to be further tested in a lab environment.

To conclude the work and validate the workflow results, an additional analysis was performed using selected information repositories. This additional step has confirmed that some of the chosen molecules are

directly and indirectly related to cystic fibrosis, giving some credibility to the proposed method.

Finally, the way this project was developed enables the use of different datasets with ligands of the target proteins as input, making the method flexible and adaptable to any disease in study.

Keywords: Chemoinformatics, Machine Learning, Cystic Fibrosis, In Silico Drug Design

Resumo Alargado

Atualmente existem diversas doenças para as quais se conhecem os seus mecanismos biológicos, mas para as quais ainda não existe qualquer tipo de medicamento, ou os disponíveis não são suficientemente eficazes. Nos últimos anos, o número de novas moléculas a chegar ao mercado tem sido cada vez menor, como tal é necessário novas abordagens para a descoberta e desenvolvimento de novos princípios ativos. Além disso, todo o processo para criar um novo medicamento é extremamente caro e demorado e a grande maioria dos compostos desenvolvidos nunca chegam a ser comercializados. Como tal, qualquer hipótese de reduzir o tempo e o preço da investigação bem como de aumentar as probabilidades de sucesso é importante para estudo. O desenvolvimento computacional de medicamentos apresenta-se como um método capaz de contribuir para a resolução destes problemas. Comparativamente aos tradicionais métodos de descoberta e desenvolvimento de medicamentos, o desenvolvimento computacional utiliza algoritmos, conhecimento biológico prévio e grandes quantidades de informação para uma seleção rápida e precisa de moléculas para posterior análise em laboratório. O desenvolvimento pode ser baseado na estrutura do alvo ou no ligando.

O foco deste projeto foi a seleção de moléculas de interesse com a utilização de uma abordagem baseada na estrutura do ligando, mais especificamente, um modelo QSAR (da relação entre estrutura e atividade de forma quantitativa,) com uma pesquisa de moléculas utilizando uma base de dados ZINC, que é constituída por todos os compostos comercialmente disponíveis.

O caso de estudo escolhido para desenvolver a metodologia deste projeto foi a fibrose quística, uma doença genética recessiva das mais comuns. Embora atualmente já existam alguns medicamentos para esta

doença, estes não são considerados eficazes nem podem ser administrados em todos os pacientes. A CFTR (*cystic fibrosis transmebrane condutance regulator*) é uma proteína que quando tem alguma mutação e subsequente malformação, leva à manifestação desta doença. A mutação mais frequente é a F508del, uma deleção no resíduo 508, uma fenilalanina. É uma proteína canal cuja função é transportar cloro, mais especificamente Cl^- .

Para este método selecionou-se moléculas provenientes de dois conjuntos de dados: i)moléculas potenciadoras ou ativadoras de anoctaminas e ii)moléculas corretoras da CFTR com a mutação F508del. As anoctaminas são um grupo de proteínas também canal que transportam Cl^- ativadas por Ca^{2+} e que estão bem distribuídas em vários tipos celulares. Estas características tornam as anoctaminas em alvos interessantes para combater indiretamente a fibrose quística. Neste grupo de dados estão presentes 10 moléculas corretores ou ativadores de anoctaminas. No conjunto de dados com corretores da CFTR estão presentes 109 moléculas que foram testadas em laboratório com recurso à técnica de Western Blot.

O método de desenvolvimento de medicamentos computacional foi construído numa plataforma chamada KNIME, que permite alojar pequenos programas (em diferentes linguagens de programação) de forma interligada, bem como já contém diversos nodos de interesse estatístico e/ou biológico entre outros. Tendo em conta o objetivo final deste projeto - a seleção de moléculas de interesse para serem posteriormente testadas em laboratório - este método foi assente em dois ramos de trabalho, um que analisa os dados num espaço vetorial e o segundo num espaço métrico. Os resultados de ambos foram conjugados para selecionar as moléculas a serem testadas posteriormente. Mais detalhadamente, existiu uma seleção na base de dados ZINC por moléculas semelhantes às iniciais (tanimoto > 0.75) sendo que estas foram posteriormente alvo de previsões relativamente à sua capacidade terapêutica em relação à fibrose quística. Cada um dos ramos

recebeu as moléculas iniciais, os seus valores testados em laboratório em relação à doença, bem como a estrutura das moléculas semelhantes retiradas do ZINC. No ramo vetorial, as moléculas iniciais passaram por um nodo do RDKIT que calculou os seus descritores moleculares. Esta informação juntamente com os valores previamente obtidos em laboratório, foram utilizados para treinar a *random forest* - método estatístico e estocástico baseado na construção de uma floresta de árvores de decisão. O modelo foi assim criado e utilizado para prever os valores das moléculas selecionadas do ZINC. No ramo métrico, as moléculas foram comparadas utilizando o NAMS, um programa que utiliza a estrutura das moléculas, nomeadamente os átomos e as ligações entre átomos, para comparação da semelhança molecular. Neste programa, seleciona-se uma das cinco matrizes de substituição de átomos disponíveis, permitindo a escolha de uma matriz mais ou menos estrita a substituições. Com a semelhança molecular foi calculada a distância entre compostos com -log(semelhança). A partir da matriz de distâncias entre moléculas confirmou-se com a análise de componentes principais que era significativo representar as moléculas em duas dimensões. Foi também efetuada uma análise das coordenadas principais para calcular as coordenadas das moléculas nesse espaço. De forma a ser possível representar as moléculas semelhantes retiradas do ZINC no mesmo plano, foi calculada uma matriz de transformação com recurso à matriz de distâncias iniciais e à matriz das coordenadas das moléculas iniciais. Foi então efetuado uma previsão dos valores de melhoria da atividade das anoctaminas ou da CFTR (dependendo do conjunto de dados) de cada molécula semelhante, com recurso a uma técnica de geo-estatística, conhecida como kriging. Esta técnica de geo-estatística baseia-se na interpolação espacial utilizando pontos espaciais com valores já determinados para prever valores de outras localizações.

Com as previsões resultantes de ambos os ramos de trabalho, foi construído um gráfico dos valores previstos, para auxiliar na seleção de moléculas com mais interesse para posterior análise laboratorial.

No que diz respeito aos dados dos potenciadores e ativadores das anoctaminas, a comparação molecular com recurso ao NAMS indicou que a correlação só existia nas matrizes mais estritas, o que significa que substituições, mesmo que de átomos semelhantes, afetam muito a capacidade terapêutica da molécula. Neste conjunto de dados foram selecionadas 10 moléculas : 8 com valores elevados nos dois ramos de trabalho e 2 adicionais que apresentam um valor previsto elevado num ramo e baixo no outro. Estas 2 moléculas com valores discordantes, foram principalmente selecionadas para posteriormente se avaliar em laboratório o ramo com a melhor previsão e identificar possíveis melhorias. Após a escolha destas moléculas foram recolhidos dados de repositórios de informação biológica para validar os métodos utilizados. Com base nesta recolha de informação, foi possível concluir que as moléculas selecionadas apresentam ligações possíveis às anoctaminas e a uma proteína canal com algumas características em comum com as anoctaminas. A recolha de informação permitiu também saber que uma das moléculas selecionadas foi já testada como molécula que poderia atuar como chaperona, indicando que poderia corrigir malformações em proteínas.

Não se efetuou validação do método utilizando os dados das anoctaminas umas vez que o seu número é demasiado reduzido. Quando se efetuou a validação com o conjunto de dados de corretores da CFTR encontrou-se valores incoerentes com a previsão adequada, o que após revisão de todas as técnicas e algoritmos levou a uma análise detalhada dos dados. Estes revelaram-se como erros de etiquetagem das moléculas, o que foi posteriormente corrigido com a utilização de *support vector machine*. Os dados corrigidos foram testados com recurso a um estudo de atividade molecular sobre a CFTR, (retirado de um repositório publico - Pubchem) e posteriormente utilizados então para o processo de virtual screening.

Em relação aos dados de corretores da CFTR, foi possível observar que as matrizes de substituição apresentaram correlação nas duas ma-

trizes mais estritas, bem como numa mais equilibrada. Com os resultados obtidos após a combinação das previsões foi possível escolher 18 moléculas para posterior análise: 6 com valores elevados nos dois ramos de trabalho e nas 3 matrizes, 4 com valores elevados nos dois ramos e em 2 matrizes, 3 com um valor elevado no espaço vetorial mas baixo no métrico, 3 com um valor elevado no espaço métrico e baixo no vetorial e 2 com valores discordantes entre matrizes. Também neste caso foram recolhidos dados de repositórios de informação biológica confirmando que algumas moléculas tinham ligações à fibrose quística, especialmente as anoctaminas e histonas deacetilases que já foram referidas em artigos científicos como tendo influência na doença.

As informações encontradas nos repositórios parecem dar credibilidade a todo o processo apresentado nesta dissertação, no entanto testes em laboratório são necessários para um maior entendimento e até possível melhoria do método. Sugere-se que após obtidos os resultados em laboratório, se faça outra iteração deste processo, de forma a aumentar a sua capacidade de previsão. Para além disso, como foram observadas diversas zonas de interesse no espaço métrico onde não existiam moléculas, seria interessante desenhar moléculas com as características necessárias para se localizarem nesses pontos, e testar as mesmas quer no espaço vetorial quer em laboratório.

É importante referir que a forma como este projeto foi desenvolvido permite a utilização de diferentes conjuntos de dados de ligandos para proteínas alvo, o que torna este método flexível e adaptável à doença que seja objeto de estudo.

Acknowledgements

Foremost, I would like to thank the support received from my advisor Prof. André Osório Falcão, it was a journey a bit more complicated then we anticipated however was a very engaging and enjoyable project to realize.

I would also like to thank Prof. Carlos Farinha, Prof. Margarida Amaral and Rainer Schreiber for trusting me with their lab results and Sara Canato for re-analysing the lab membranes and helping me understand the lab files.

To all the amazing people in Lasige that kept me sane, since I was able to talk about my work and at the same time realize that all of you also had some kind of issues in your projects, a big thanks and all the best wishes.

Obviously, I need to thank my mother for everything but specially for all the encouragement during this process.

Last but not the least, Pedro, now is my turn to mention you in my dissertation. This is the end of a cycle for me in so many ways, and I would like to thank you for being there during all the way, TVB.

This work was possible by MIMED project - Mining the Molecular Metric Space for Drug Design. PTDC/EEI-ESS/4923/2014

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

"The nitrogen in our DNA, the calcium in our teeth, the iron in our blood, the carbon in our apple pies were made in the interiors of collapsing stars. We are made of starstuff."

Carl Sagan, Cosmos

Today it is already possible to understand how a big number of diseases work in a biologically way, however, the information is scattered across different databases while several work groups keep trying to solve the same issues around the same diseases. Plus, the financial landscape in the investigation sphere is not always aligned with the current spending needs for each problem.

One may try to tackle some of these problems by using previous information about biological targets stored across the aforementioned databases and a computational methodology. The traditional drug design approach requires a tremendous amount of money, time, as well as human resources to identify and bring into the market a completely new drug (figure 1.1). This in combination with the ever-increasing demand for new drugs, has called in recent times for more efficient and cost-effective computational and automated drug design approaches. Automated programs being fed by the vast amount of biological data currently available predict how substances would react and require fewer investigators and

less utilization of consumables at the lab (reagents or otherwise), resulting in fewer resources (human and financial) needed to reach the desired outputs.

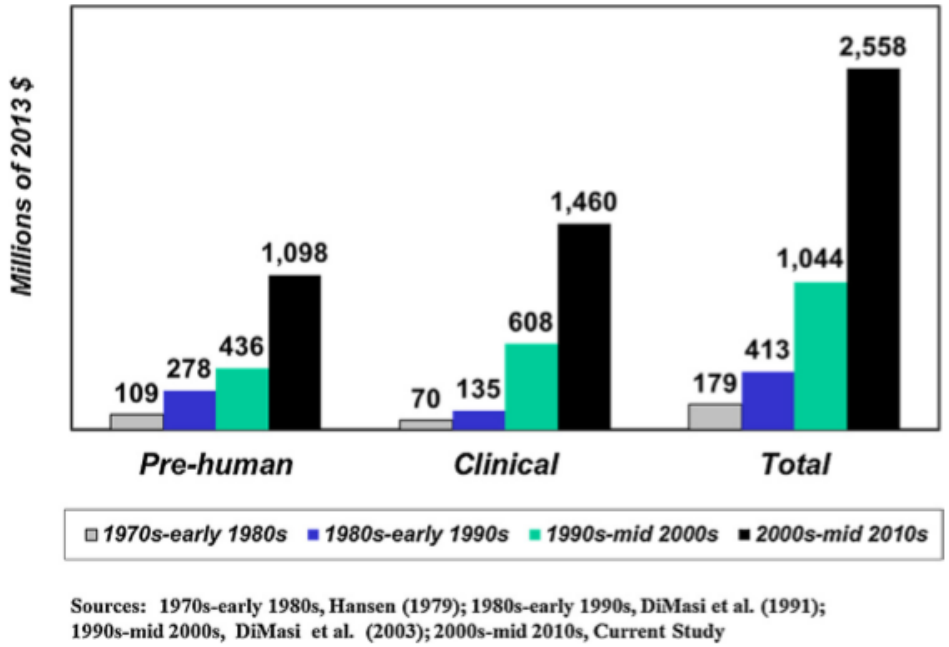


Figure 1.1: Cost of different stages in drug investigation from 1970 to 2010. From (DiMasi et al., 2016)

For the purposes of this dissertation, cystic fibrosis (CF) was chosen as the target disease to be further studied and analysed. This disease causes a thick mucus in the respiratory and digestive system, with several deadly complications, mainly, difficulty in breathing and bacterial infection in the lungs. It is also common to have digestive problems like difficulties in absorbing nutrients and constipation (Boucher, 2007). The average life expectancy for CF patients is about 37 years old, with lung or cardiorespiratory complications being the main causes of death. Cystic fibrosis is one of the most common recessive genetic diseases, affecting 1 in every 2500 newborns in Caucasian populations, (Zielenski and Tsui, 1995) with different prevalences across countries and within human populations. It is characterized by a misfolding of a channel protein called CFTR (cystic fibrosis transmembrane conductance regulator,) which is a ABC (ATP-binding cassette) transporter of chloride. The deletion of a phenylalanine at

residue 508 (F508del) is the most common cause of CFTR misfolding (Wang and Li, 2014b). More than 90% of cystic fibrosis patients carry at least one copy of the CFTR gene with the F508del mutation, causing impaired exocytic trafficking (Cheng et al., 1990), and the retention of the malformed protein in the endoplasmic reticulum, leading to its consequent degradation.

The drugs available by the end of 2016 are very expensive[1], scarce, not very effective and were only designed to work in specific patients groups (Cholon et al., 2015). As such it becomes truly important to find drugs that cure and/or counteract the CF effects. To treat this disease there are two possible courses, the correction of the protein responsible for the illness,(cystic fibrosis transmembrane conductance regulator,) or potentiating another channel protein that transport the same ions (like the anoctamins family) and therefore, counteract the poorly functioning CFTR. It is very clear that allying the need for a new CF drug and the feasibility of predicting chemical compounds reactions in an automated way is compelling and most important, a necessary evolution that many patients expect.

1.2 Objectives

The main goal of this dissertation is to propose lead chemical molecules that can possibly become principal actives against cystic fibrosis. Although all the following work was developed for this disease, the methods used in this dissertation can be applied for other diseases, depending only on the input dataset.

Hypothesis: Using datasets from cystic fibrosis drug response in combination with machine learning methods, it is possible to predict the therapeutic capacity of chemical compounds present in databases.

[1]DRG blog, While Kalydeco Sailed Through, Vertex's Orkambi Faces Strong Headwinds in Europe, last accessed in 19 July 2016

1.3 Contributions

The main contributions of this work will be:

Contribution 1: Create a new computational model to predict therapeutic capability of each compound using metric and vectorial distances.

Contribution 2: Identify molecular features of interest.

Contribution 3: Select lead compounds to be tested *in vitro*.

In this dissertation, the drug design process will use different methods to assert the capability of each compound, (retrieved from a database,) to become a possible drug for cystic fibrosis. This method can support the decision process around the selection of which molecules should be tested, reducing the human and financial resources, as well as the time needed for the drug design process.

1.4 Drug Discovery and Development

Drug discovery is considered the process from the start of the investigation of a specific disease until the designed drugs are ready for *in vitro* and/or *in vivo* testing. The development phase includes pre-clinical, clinical, approval and surveillance steps.

The process of drug research has changed dramatically over the years. Initially, drugs were developed using traditional remedies from natural medicine that were analysed using old-fashioned methods, in order to identify the respective active component, and subsequently transform it into a usable drug. Along the years, techniques evolved and became more complex and capable, such as, using cells to perform tests on molecules and natural extracts. Nowadays, drug design techniques are far more complex involving several scientific fields such as medicine, biotechnology, pharmacology, structural biology, computational chemistry and much more.

The entire drug discovery and development process - that starts with understanding the disease in study, and ends when a drug reaches the market - includes several refinements of the selected compounds, where the objective is to

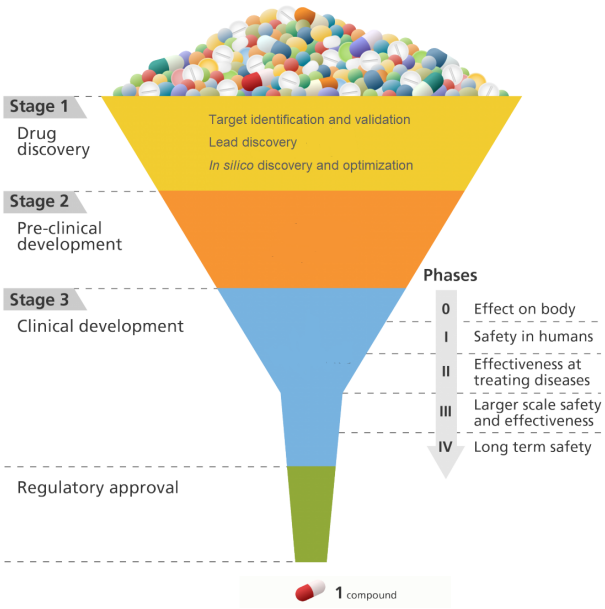


Figure 1.2: Representation of Drug research stages, incuding the continuated selection of compounds. Adapted from Genome Research Limited

find the cure and at the same time be relatively inclusive to not lose any important compound in one of this refinements (fig 1.2). The complete process can be described in 8 steps: First, it is necessary to understand the disease, this means find patient symptoms and possible targets, and proceed to isolate and test this targets for their functions and relations with the disease in study (Anderson, 2003). After this, the target needs to be validated, with several tests being performed to ensure the connection with the disease, their specific function as well as their mechanism of biological control (for example binding partner molecules) (Chen and Chen, 2008). Once the targets are found and validated, it is necessary to find compounds with the ability to interact with the drug target in a positive way - lead identification (Kalyaanamoorthy and Chen, 2011). It is very important to select several lead compounds to ensure a broader probability to identifying a clinical candidate. After the identification, an analysis of the selected compounds should be performed to find important characteristics, evaluate their potency and improve it. Some pharmacological or biological properties are required to be evaluated as well and can be done with the aid of computational biology. After this

process, the selected compounds are synthesized into a drug and the preclinical stages begin where *in vitro* and/or *in vivo* tests are performed on the synthesized drug (and potential reformulations) to evaluate the toxicity and the potency of the compounds. If the preclinical tests show positive and promising results, the selected compounds move on to the clinical trials stage, where for the first time, humans are used to evaluate the effectiveness and safety of the drug. A clinical trial is carried out to ascertain safety, side-effects, dosage, and efficacy and it is divided into several phases as described in figure 1.2. The compound candidates that passed the 2 previous phases are subjected to a final evaluation from the regional regulatory entity. This is the last barrier to get the required approval to start marketing the drug. The regulatory entities who evaluate the drugs differ from country to country and are very restrictive and demanding on this process to ensure the safety of the population, as well as the effectiveness and the economic impact of the drug. After a drug is approved, it continues under evaluation during its lifetime. Even with the clinical trials, it is possible that the entire or a portion of the population reacts differently than predicted.

1.5 Computational Drug Design

Computational Assisted Drug Design (CADD) is the use of computational methods to facilitate the design and discovery of new therapeutic solutions and its main objective is to help identify and select the most promising candidate drugs. It is during the drug discovery process, where a significant number of molecules can be identified as candidate compounds, that CADD can intervene the most. Using *in silico* testing, CADD can help on the selection of the molecules more prone to have better results. Currently, there are two ways to design a drug using the CADD process:

Structure-based drug design Using the 3D information about the biological target, one can create molecules from scratch (*de novo* approach) or use molecular databases for screening - **virtual screening**.

Ligand-based drug design When the 3D information about the target protein is not available, it is possible to use a set of ligands to identify some properties. This identification uses one of the following approaches:

- **QSARs** Quantitative structure-activity relationships models
- **Pharmacophore based models** that use functional groups and enforces small changes in the molecules but with similar structures.

In this thesis, the approach chosen was a QSAR model because the provided dataset was based on molecular information and structure of ligands to the target.

1.6 QSAR - Quantitative structure–activity relationship

With the ever growing number of chemical compounds available in databases, it is necessary to find robust and efficient ways to select interesting molecules with the ability to become drugs (Tropsha, 2010). Quantative structure-activity relationship model is a ligand-based drug design, which creates a model from tested chemical compounds, to perform virtual screening in order to predict characteristics of new compounds.

This approach relies on a simple similarity principle: compounds with similar structures are expected to have similar biological activities.

QSAR can be divided into metric and vector approaches: the metric approach usually relies on some type of molecular comparison and a subsequently statistic analysis like k-nearest neighbour (Paredes and Navarro, 2006); the vector approach uses data mining techniques such as support vector machines (Hasegawa and Funatsu, 2010) over the calculated molecular descriptors.

1.7 Current Cystic fibrosis available treatments

Currently, the few drugs available (ivacaftor and lumacaftor) are not very effective, as they only target certain CFTR mutations (Cholon et al., 2015). Sometimes it is still necessary to perform a lung transplant, with the complications

associated with that surgery being very considerable. Also, to avoid the proliferation of bacteria from one lung to another it is necessary the transplantation of both lungs.

Ivacaftor is a potentiator and its main goal is to help the channel to transport the chloride, targets the G551D substitution mutation - the CFTR can reach the membrane but is not very competent on transporting chloride (only 4-5% of cystic fibrosis patients have it).

Lumacaftor, as a corrector, doesn't have a medical use by itself since it only corrects some of the protein misfolding ("chaperone-like" effect). When used in combination with ivacaftor, it can help in the treatment of F508del mutation. However, lumacaftor has a restrictive use and can only be applied in homozygous patients with more than 12 years old. Orkambi (lumacaftor 200 mg and ivacaftor 125 mg) was rejected by the U.K.(in March 2016) and by Ireland (in June 2016) national healthcare system due to lack of cost effectiveness [1].

1.7.1 The Anoctamins hypothesis

Anoctamins are a family of channel proteins that produce Ca^{2+} activated Cl^- currents. They are present in several tissues like endothelium, visceral smooth muscle Cajal cells, heart, airways, alveoli and colon. This family of proteins does not show any obvious homology to other ion channels(Kunzelmann et al., 2011). Since they transport the same molecule and are present in several important tissues, these channels are a possible way to counteract the effects of a dysfunctional CFTR. Ano 1 is the most promising target candidate since it exists in different cell types spread throughout the human body (Kunzelmann et al., 2011).

[1] DRG blog, While Kalydeco Sailed Through, Vertex's Orkambi Faces Strong Headwinds in Europe, last accessed in 19 July 2016

1.8 Overview

The rest of this dissertation is organized as:

2nd Chapter: State of the art with more detailed information about the tools that will be used and also some contextualization to *in silico* drug design.

3rd Chapter: Description and explanation of the datasets origin.

4th Chapter: Overview of the entire workflow.

5th Chapter: Presentation of the results.

6th Chapter: Discussion of the results and selection of molecules.

7th Chapter: Conclusions.

The appendices contain:

A - List of molecular descriptors used.

B - Workflow images.

C - Table with selected molecules SMILES.

1.9 Timeline

This project consisted in 3 main areas of activities across 21 months: Research of state of art and available tools (4 months); workflow development, testing, improvement, and validation (14 months); and report (9 months).

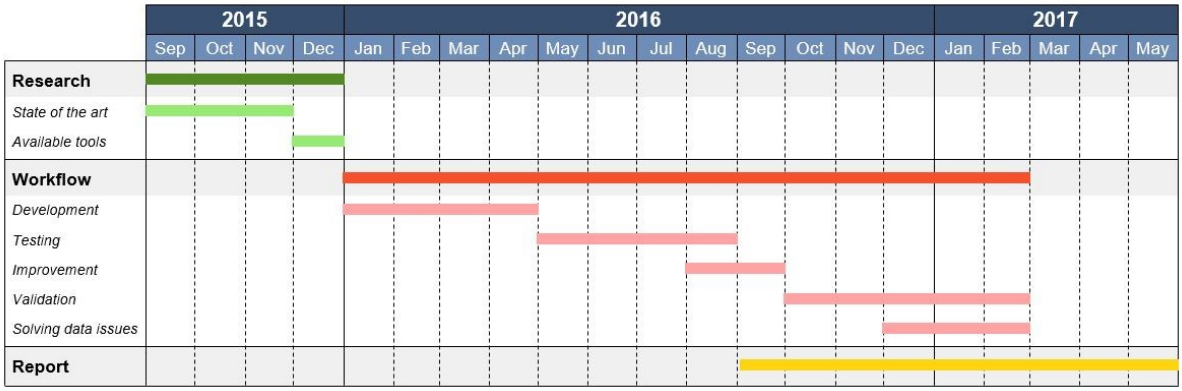


Figure 1.3: Gantt diagram representation of the project.

Chapter 2

Background

"Knowledge has to be improved, challenged, and increased constantly, or it vanishes."

Peter Drucker

Given the current cost to bring a new drug to the market (around US$2.5 billion (2013 dollars) as well as the 10-15 years needed to go through the entire process (DiMasi et al., 2016) (Clark, 2008) it is necessary to develop tools that ensure the investment is applied to the most promising molecules. The number of new molecular entities (NMEs) is growing substantially each year but the number of molecules that reach the market continues to diminish (Paul et al., 2010). The adoption of CADD increases the probability of commercializing a drug, since, compared with other methods (like high-throughput screening (HTS)) this rational drug design approach is based on knowledge instead of testing and aims to understand the disease giving information about the ligands affinity and their mechanism (Lionta et al., 2014).

2.1 Computer assisted Drug Design

CADD is the development of drug design with the aid of computational methods. As in (Martins et al., 2012), computational models can be a viable alternative to high throughput screening, as computer power becomes more affordable and more data available, complex and accurate models may be produced.

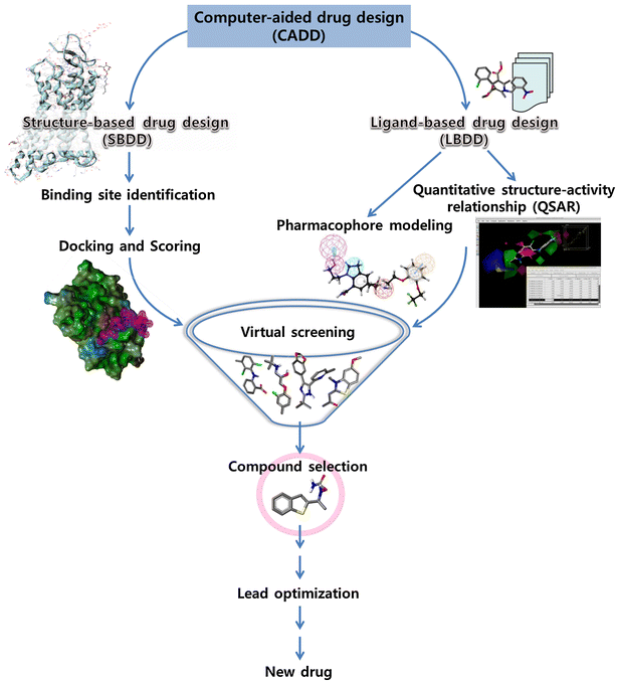


Figure 2.1: Workflow of LBDD and SBDD in computer-aided drug design. From (Joy et al., 2015)

The amount of molecules with probability to become a drug is extensive, making it impractical to test them all *in vitro*. *In silico* testing can help solve this situation, as well to aid in decreasing the costs when compared to HTS. This method can reduce the time needed for the initial steps of drug discovery diminishing the time-to-market (Martins et al., 2012).

As previously described, computational drug design refers to the discovery and optimization of drug candidates and can be divided into two categories (figure 2.1): structure based (SBDD) and ligand based (LBDD) drug design.The previous biological knowledge of either the target structure or ligands with bioactivities is the determining factor in choosing which approach should be used.

2.1.1 LBDD - Ligand based drug design

When the 3D structure of the target is not known it is possible to use several active ligands of the target and make predictions about the structure of the target (or of the ideal candidate molecule). Also, with the molecular structure of the ligands, it is possible to make predictions of their chemical/molecular properties. This type of design is based on the premise that similar structural compounds will interact with the same target. As this approach relies on the similarity between compounds, choosing the correct algorithm/formula to compare them is a determinant step. LBDD techniques can be used even when the information about the target and the ligand are scarce or nonexistent, by using the genetic sequence of the target and assuming that similar receptors interact with similar ligands (Klabunde et al., 2009). The most common methods are QSAR/QSPR and Pharmacophore modelling.

2.1.1.1 QSAR/QSPR - Quantitative structure-activity/property relationship

This method relies on the *principle of similarity* were molecules with similar structure will be more prone to have similar bioactivities and the variation in bioactivities is related with structural and/or molecular variations. Using a set of ligands, statistical tools, and a machine learning method, the goal of this project is to create an algorithm that can predict characteristics of new entities, using only information about their structure. (Tropsha, 2010) has defined a generic representation of a QSAR model as:

$$P_i = k(D_1, D_2, ..., D_n) \tag{2.1.1.1}$$

Where;

$\mathbf{P}_i$ - Characteristic of interest to predict;

k - Transformations to the data executed by the model;

$\mathbf{D}_1$, $\mathbf{D}_2$ **and** $\mathbf{D}_n$ - Molecules;

QSAR modelling can be divided into two broader categories: i) regression, where the variable to predict is continuous and ii) classification, where the algorithm will predict in which class does one element belong.

There are several commercial and free software available, some with a black-box approach. CoMFA (Comparative Molecular Field Analysis) uses a superimposed strategy to compare the molecular structure and make inferences about their biological activities (Cramer et al., 1988). GOLPE, (Generating Optimal Linear PLS Estimations) program obtains partial least squares estimations with the best possible prediction (Baroni et al., 1993). ADAPT (Automated Data-Analysis and Pattern Recognition Toolkit)[1] is a commercial program that predicts values using a training set and has several subsystems to receive the graphical input of molecules, stored them in connection tables and create and evaluate the substantial structure molecular descriptors available(Stuper and Jurs, 1976). Regardless all the programs currently available, other statistical or data mining software (not created specifically for drug design) can be used, such as: R[2], MatLab[3], SPSS[4], KNIME[5].

2.1.1.2 Pharmacophore modeling

Pharmacophore modeling uses compounds with a similar 3D structure on the principal functional groups, and a wide range of structure and different atoms in the other groups not considered as principal (Vuorinen and Schuster, 2015) therefore, selects the common chemical features that represent the ability of that set of ligands to interact with the target. This type of modeling has two main steps: i)represent the conformational flexibility of the ligands and ii) find the minimum required common features, to develop the pharmacophore model (Yang, 2010). There are several programs for this technique: the Ligand Scout is a very specific

[1]ADAPT http://research.chem.psu.edu/pcjgroup/adapt.html, last accessed on 14 August 2016

[2]R https://cran.r-project.org/, last accessed on 14 August 2016

[3]MatLab http://www.mathworks.com/products/matlab, last accessed on 14 August 2016

[4]SPSS http://www.ibm.com/analytics/us/en/technology/spss/, last accessed on 14 August 2016

[5]KNIME https://www.knime.org/

program that uses only a set of six types of chemical features and volume constraints when searching in databases(Wolber and Langer, 2005); the Discovery studio software suite[1] has several tools available including pharmacophore approach with validation and a virtual screening; the MOE (Molecular Operating Environment)[2] has an automatic pharmacophore generation.

These programs have different algorithms for conformational flexibility of the ligands and molecular comparison. Regarding flexibility issues, there are two main approaches, either every possible conformation is calculated before or at the same time the pharmacophore model is created. The molecular similarity can be done in a superimposed way, (posing a challenge if the ligands are very different and a similar origin point for comparison is difficult to determine) or, using molecular field descriptors.

2.1.2 SBDD - Structure based drug design

Structure based drug design approach uses the structure of the receptor to find molecules of interest and can be achieved in two steps:

1 - Identification or Prediction of the binding site.

2 - Docking and Scoring.

A concave region with several chemical functionalities is the ideal binding site (Anderson, 2003). Docking a molecular combination requires a prediction using the structure of the binding site and calculates their affinity with ligands (Cheng et al., 2012). Scoring a docking prediction is not an easy task as one protein cannot be seen as a static entity; each protein can have a significant number of conformations, so scoring the ability of proteins to be docked by molecules becomes a key and difficult step to achieve. This is the main reason why different scoring functions exist, mainly based on force field, empirical and knowledge-based (Huang et al., 2010). Consensus scoring (using more than one type of scoring in a combination) has been described as a more reliable way to find the

[1]Discovery studio http://accelrys.com/products/collaborative-science/biovia-discovery-studio/, last accessed 15 August 2016

[2]MOE http://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm, last accessed 15 August 2016

correct ranking of lead compounds (Houston and Walkinshaw, 2013). There are several software tools available for the binding site, docking and scoring prediction (table 2.1).

Table 2.1: Binding site and Docking prediction softwares based on table from (Joy et al., 2015)

Predicts	**Program/server**	**Description**
Binding site	CASTp	Uses weighted Delaunay triangulation and the alpha complex for shape measurements
	Cavitator	Pocket prediction using a grid based geometric analysis
	ConCavity	Based on combining evolutionary sequence and 3D structures
	eFindingSite	Common ligand binding site prediction using set of evolutionary related proteins
	SiteComp	Binding site comparison based on molecular interaction fields
Docking	AutoDock	Flexible side chains (genetic algorithm)
	GOLD	Complete solution for docking small molecules into protein binding sites
	BP-Dock	Flexible docking
	Glide	Ligand exhaustive search and flexible side chains
	idock	Flexible ligand docking
	SwissDock	Rigid ligand with less than 10 rotational bonds

2.2 Molecular descriptors

In computational drug design, there is the need to represent compounds in such a comprehensive way where it is possible to quantify their characteristics in a useful manner. Molecular descriptors are numerical values which express properties of a molecule and can be theoretical or a result of an experimental standardized test (like dipole moment or polarizability). In theoretical descriptors, it is possible to find 5 classes (molecular descriptors 0D to 4D) each one more complex than

the last. The first class (0D) encompass descriptors directly obtained from the chemical formula, like molecular weight, in the second class (1D) the structure is represented in structural fragments (e.g. funcional groups) , the third class (2D) have topological representations of the molecule, fourth class (3D) is a rigid geometrical representation and the last class (4D) adds flexibility to the 3D descriptors giving the possibility to evaluate, for example, molecular interaction. This values can bring insights between molecular structure and their associated properties, as well as their biological activities. This makes them very useful as input for the statistical methods in QSAR/QSPR and virtual screening studies (Karelson, 2000). Such as the blood-brain barrier penetration model (Martins et al., 2012) and the prediction of standard enthalpy of formation of hydrocarbons (Teixeira et al., 2013).

2.3 Supervised machine learning in QSAR

Supervised machine learning has a very important role in CADD (computer aided drug design) since it is the way to refer to several methods of inferring results after receiving a training data set with examples. This data set should contain the object (it is usually a vector with a set of variables, e.g. molecular descriptors) and the expected output. The training data set is analysed and a learning algorithm builds a model. Afterwards, the machine receives a test set, and using the model constructed in the last phase, assigns each new data to a category, returning a predicted class or value. The output of the supervised machine learning depends on the characteristic to predict, either is in form of classes (classification), or in form of a value/ number (regression problem).

2.3.1 Linear models

2.3.1.1 Multiple Linear Regression

MLR(multiple linear regression) is the simplest regression model where a linear relation between molecular descriptors and relevant biological activities is assumed. The generic representation of this linear relation with specific weights for each molecular descriptor can be represented as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \quad 2.2.1.1$$

Where;

Y - Characteristic of interest to predict;

$\mathbf{X}_{0,...,k}$ - Molecular descriptors;

$\beta_{0,...,k}$ - Weight of each molecular descriptor;

This linear approach can give us information of which molecule descriptors have positive or negative effects in the characteristic of interest since it is possible to see the weight of each descriptor. However, the descriptors chosen to be used with this method should be non-linear from each other, since the weights can be influenced by linear related descriptors. The relation between the number of molecules and the number of molecular descriptors should be around five to one (Topliss and Edwards, 1979). There are several papers using multiple linear regression, in (Yang et al., 2016) was used to predict the anti-androgenic activity of bisphenols, in (Min et al., 2016) aid in the prediction of inhibitors of hepatitis B virus replication and (Grzonkowska et al., 2016) predicting the toxicity of ionic liquids to Vibrio fischeri gram-negative bacteria.

2.3.2 Non-linear models

2.3.2.1 Neural networks

Neural network is an algorithm based on a neuron where inputs are received and transformed in outputs using a mathematical function. The idea behind a neural

network is a series of nodes with different mathematical functions and weights, which pass their outputs from one node layer to the next. In this type of approach, several weak characteristics are used to obtain a good classification. The current approaches are based in a perceptron, an algorithm created with the objective to recognize patterns using only addition and subtraction functions (Rosenblatt, 1958). In supervised learning, the neural network is trained, and the weight of each node is adjusted to achieve the best prediction possible (Terfloth and Gasteiger, 2001). Some examples of studies where neural networks algorithm was applied are the aqueous solubility prediction of drugs (Huuskonen et al., 1998), prediction of hepatic drug clearance in humans (Schneide et al., 1999) and QSAR study of anti-HIV activity for a large group of HEPT (protein family) derivatives (Jalali-Heravi, 2000).

2.3.2.2 Support vector machines

Support vector machines (SVM) proposed by (Vapnik, 1998) , can address classification and regression problems. In this supervised machine learning method the training data is mapped in a very high dimension feature space or hyperplane, which allows the representation of all instances. Given the classes or values in each training instance and their location in the hyperplane, the algorithm creates areas that will be used to determine predictions for the new data. The high-dimension space is established using a Kernel function (linear, polynomial, radial basis and sigmoid) (Yao et al., 2004). SVMs are broadly used, for example in SAR/QSAR study of phenethylamines (Niu et al., 2007), prediction of mutagenicity of compounds (Ferrari et al., 2009), QSAR models for predicting anti-HIV-1 activity of TIBO (molecule) derivatives (Darnag et al., 2010).

2.3.2.3 Ensemble methods

Ensemble methods are a combination of several supervised machine learning algorithms, that give a final result usually by weighted or unweighted voting of all the results from individual models. This combination of methods are often much more accurate than the individual models(Dietterich, 2000) and can be used for classification or regression.

Bootstrap aggregating (or bagging) is an ensemble method where a subset/sample with the same number of instances from the original training set with replacement is created, therefore some instances will be repeated and some will not be present in the sample. For each bootstrap, a classifier is created and then aggregated to form the final model. This method encompasses several small models, therefore, can be a very robust tool reducing the over-fitting problem.

Random forest algorithm (RF) develops decision trees using a different bootstrap sample of the data apart from the past tree already constructed, the variables are arbitrarily selected at each node and chosen by their ability to divide the sample. RF can be used for classification and regression, this forest of decision trees, performs a classification depending on the class predicted by the majority of trees, and the regression by the average of each tree prediction. Given its construction is insusceptible to noisy variables and can operate with more variables than examples. Random forest is widely used from QSAR prediction of compounds aquatic toxicity (Polishchuk et al., 2009), to prediction of standard enthalpy of formation of hydrocarbons (Teixeira et al., 2013), or even QSAR based model for discriminating EGFR inhibitors and non-inhibitors (Singh et al., 2015).

Support vector machines and random forest provide good results despite a large number of variables, making them the most used algorithms in QSAR and virtual screening studies (Martins et al., 2012) (Teixeira et al., 2013). Given their characteristics, this type of statistic methods needs to be performed in a vectorial space environment.

2.4 Molecular similarity

Automated prediction (like QSAR) relies heavily in similarity tools, as such a very small difference between molecules can have critical consequences in their chemical properties and activities. It is then important to use methods with a good ability to differentiate similar compounds (Martins et al., 2012). When comparing small molecules, there are several problems that arise: first of all

the representation of this type of data is difficult, since molecules have very specific structures and their size varies, becoming a true challenge to store them in structured databases. Therefore when there is an attempt to define a molecule it is necessary to know that some information will be lost and that there are several other ways to represent the same molecule. Even if representation strategies exist, for example, SMILES (Simplified Molecular Input Line Entry System, where a string represents the three-dimensional location of atoms and the type of bond they share), there is always the challenge of choosing the right methodology for comparison. Currently there are some methods for this comparison, however, none seems to work universally (Teixeira and Falcao, 2013). All these methods can be classified in three groups: i) structural descriptors (as in molecular descriptors previously referenced), ii) molecular fragments (E.g. molecular fingerprints), iii) graph theory.

2.4.1 Molecular Fingerprints

The fingerprints technique is a way of encoding the structure of a molecule and is one of the most used techniques for molecular similarity. There are two main types: a)hashed and b)circular.

Hashed fingerprints are a bit string representation of the molecule were "0" represents the absence of a certain characteristic and "1" the presence of the same characteristic. Depending on the type of fingerprint these characteristics can come from a previous list (called keyed fingerprints) or can be a spatial representation of each molecular fragment of a specific size, for example, of each N atoms and N-1 bonds in each linear substructure.

Circular fingerprints (e.g. Morgan) evaluate the neighbours of each atom with a defined radius. The radius parameter defines the distance between one atom to their neighbour; when radius=1 the atoms evaluated are all the atoms bonded to the atom in study, if the radius=2 it is possible to evaluate the atoms who have a bond to an atom connected with the atom in study, as well as the atoms with radius=1. Usually, the radius is set between 2 to 3 bonds. It is attributed to each atom an unique identifier that is updated according to the evaluations of the neighbour atoms, after each iteration, there is a duplicate structural removal

step. When the last evaluated neighbour reached the defined radius, all the repeated identifiers are removed again and then converted to a bit string.(Rogers and Hahn, 2010)

To examine the similarity, these string bits have to be evaluated and a score from zero to one is calculated. The most used metric is Tanimoto similarity or Jaccard similarity coefficient that is calculated using all presented characteristics common (shared) to both molecules divided by the total number of characteristics present in both molecules (shared and unshared). The formula to find the Jaccard similarity is :

$$jaccard_{x,y} = \frac{X \bigcap Y}{X \bigcup Y}$$

2.3.1.1

Where;

X - characteristics present in X's fingerprint.

Y - characteristics present in Y's fingerprint.

Which can be also represented as:

$$jaccard_{x,y} = \frac{c}{a + b - c}$$

2.3.1.2

Where;

a - number of ones(or characteristics) in X's fingerprint.

b - number of ones(or characteristics) in Y's fingerprint.

c - number of ones(or characteristics) in both X and Y fingerprints.

2.4.2 Non-contiguous Atom Matching Structural Similarity

Non-contiguous Atom Matching Structural Similarity (NAMS) is a tool to compare the structural similarity between small molecules(Teixeira and Falcao, 2013). For a given pair of molecules, NAMS aims to find a pair of atoms (one from each molecule) that are the most equivalent between each other, representing each molecule in a graph with atoms and their respective bonds. All the pair combinations identified by NAMS are evaluated with the best match being selected for the molecular comparison. However, one atom of a given molecule can only be matched to one atom of a second molecule and vice-versa. The non-contiguous part of this method relies on finding the best matching score, using all the possible combinations of atoms. Such matching process does not require contiguous atoms or bond profiles. After this step, NAMS calculates the Jaccard similarity coefficient (as mentioned in 2.3.1.2). In this situation, the variable "c" from Jaccard formula is the common substructure in both molecules, where variables "a" and "b" represent the self-similarity of each molecule. The main differentiation of NAMS is the ability to discriminate molecules by chirality or double bound stereoisomerism that are rarely present in other similarity approaches.

NAMS also allows to choose between 5 scoring matrices that weight in different ways the comparison of the replaced atoms:

- **Atom distance matrix 0:** The most strict matrix, where each atom has only 100% similarity with itself and 0 from the remaining atoms.
- **Atom distance matrix 1:** Almost as strict as matrix 0, but instead of 100% similarity the value given is 90%.
- **Atom distance matrix 2:** Similarity matrix based on literature.
- **Atom distance matrix 3:** Empirical matrix constructed with known information regarding the atoms.
- **Atom distance matrix 4:** This matrix is the least strict matrix possible were all the atoms as 0% similarity with all possible substitutions including itself, which gives all the attention to the structure of the molecules.

Fingerprints method and NAMS were already tested by (Teixeira and Falcao, 2013) using three datasets containing molecules with very similar characteristics (therefore more difficult to differentiate). In the end, only NAMS was able to recognize the differences and separate the molecules. Given this result, NAMS was chosen as the main method for the comparison of molecules to the metric branch of this dissertation, even though it has a higher computational cost.

2.4.3 Vector and metric spaces

In a vector space, one instance is represented as one vector, with several characteristics and therefore dimensions. When one instance is represented in a metric space with 2 dimensions, the distance between all the other occurrences is the only characteristic that defines the instance, for example, dissimilarity between instances. This dissimilarity cannot be represented directly in a 2-dimensional metric space because the number of instances is the same as the number of dimensions and this number is always greater or equal than 2, for example, a dataset with 5 instances has 5 dimensions, but only 1 characteristic. To reduce to a 2-dimensional representation it is possible to do a principal component analysis. The principal component analysis can study all the linear correlations between the data and transform them in non-linear components, these components can be a combination of linear correlations, therefore the number of components will be equal or less than the number of variables (in this case the total of instances). Each component is constructed in a non-linear correlation with the previous established components and they are build up in a sequence from the most explicative component to the less one. This statistic procedure returns a summary of the information of the variance that is explained in each component, so it is possible to verify if the datasets are properly expressed in two dimensions - this procedure can be done in the R environment using prcomp function. To organize these instances in a 2-dimensional graph, it is necessary to perform a principal coordinate analysis or classic metric multidimensional scaling, which arranges the instances in a plot with the minimum distortion possible - for which

there is a function available in R denominated cmdscale[1].

2.4.4 Inference in metric spaces

To infer in a metric space it is necessary to search the plotting points most similar to the location where the inference is being performed since instances more close to each other should be more related and affect the inference of the point to be predicted. The proximity search between plotting points is calculated with a function that satisfies the triangle inequality (where the sum of any two sides of the triangle, must be greater or equal to the length of the remaining side) (Chavez et al., 2001).

After this search, it is necessary to perform the interpolation of the data. The geostatistical tool denominated kriging is a possible algorithm to execute this process, interpolating the data in a geospatial space. This interpolation can also be called Gaussian process regression since the interpolations are modelled by a Gaussian curve. This geostatistical method aims to minimize the variance of errors and to reduce the mean residual of error to zero (Teixeira and Falcao, 2014) and can use the variogram of the initial data to fit the model - Ordinary kriging is available in Gstat[2] a R library. This technique was already used in QSAR/QSPR approaches ((Fang et al., 2004), (Yin et al., 2007),(Teixeira and Falcao, 2014)). Also, considered as a widely flexible technique, it has been used to perform interpolations in several different fields: soil quality (Smith et al., 1992), atmospheric temperatures (Hudson and Wackernagel, 1994) and epidemiologic map (Carrat and Valleron, 1991). When applied to a non-spatial problem it is required to transform the data into spatial data, in order to represent the problem in a metric space. To the best of our knowledge only one study used structural similarity and a kriging algorithm (Teixeira and Falcao, 2014), all the other researches used molecular descriptors sometimes previously selected.

[1] cmdscale https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html, last accessed 17 August 2016

[2] Gstat https://cran.r-project.org/web/packages/gstat/index.html, last accessed 17 August 2016

2.5 Cystic fibrosis

The most common mutation that causes cystic fibrosis is F508del-CFTR which originates a misfold and subsequently the retention in the endoplasmic reticulum. CFTR is present throughout several cells in the body, more specifically in the membrane of epithelial tissues (eg: airways, intestine, pancreas, sweat ducts) controlling the secretion of Cl^- and fluid, which when CFTR is defective increases the viscosity (Wang and Li, 2014a).

CFTR is a ABC (ATP-binding cassette) protein. Most of these proteins are active transporters (were the transportation of a substrate is done against the electrochemical gradient at some ATP expense). Usually, ATP binds to a nucleotide-binding domain which can alter the conformation of transmembrane domains and, therefore, enforce the substrate transport. CFTR, however, does not function as an active transporter, since the flux of Cl^- is done in favour of the electrochemical gradient (Chiaw et al., 2011) .

In (Lewis et al., 2005) X-ray crystallography studies from F508del-CFTR mutated protein suggests that the deletion of this phenylalanine does not block the first nucleotide-binding domain(NBD1) folding, but should affect domain-domain assembly, and therefore causes a global conformation change resulting in loss of channel function.

2.6 *In silico* screening

Several treatments have already been proposed for cystic fibrosis, however, most of them failed to pass the drug development phases. Since this condition is caused by mutations in CFTR (almost 2000 individual mutations identified so far (Brodlie et al., 2015)), it is logical to consider the wild-type CFTR protein as a treatment itself. Therefore it was developed a non-viral liposomal vector to transport wild-type CFTR gene directly to the airway epithelial cells. The introduction of this gene seems to work for several different mutations, however, it only enhances the lung function and not the rest of epithelial cells affected in the rest of the body (Alton et al., 2015). Another approach is the use of

small molecules that interact with the defective CFTR, these molecules can be separated into three categories:

Potentiators - molecules that enhance the channel function of proteins already in the membrane;

Correctors - molecules with the aim to correct CFTR making it possible for this protein to reach the membrane;

Read-through agents - which affects the ribosome to ensure that a mutation in the CFTR gene does not result in a premature interruption of the protein production resulting in a shortened non-functional CFTR.

Ivacaftor as already described before, is a potentiator that targets the Gly551Asp CFTR mutation and can be applied to patients with at least one allele with this mutation (Accurso et al., 2010).

Regarding the correctors, there are 2 molecules of interest: Lumacaftor (VX-809) and VX-661. However, in this type of mutations the problem relies upon the defective protein as it has difficulties to achieve the membrane and even when it reaches it, their gating function does not work properly. For this reason, the correctors by themselves are not proposed as the solution, but a combination of correctors and potentiators could help to treat this type of mutations(Cholon et al., 2015) (Pilewski et al., 2015).

Ataluren (PTC124) is a read-through agent which targets non-sense mutations that produce a premature stop signal, representing around 10% of cystic fibrosis patients. This signal will stop ribosomal production of the protein before achieving the right stop codon resulting in a loss of function. Ataluren enables the ribosome to skip the premature stop signal originating a functional protein.

Some other efforts are worth mentioning: a study to use DHPs (used in a drug to treat hypertension) as CFTR potentiators (Vietin et al., 2012) and a study that targets CFTR to treat cystic fibrosis and secretory diarrheas (Verkman et al., 2006).

Chapter 3

Data

"Data! Data! Data!" he cried impatiently. "I can't make bricks without clay!"
Sherlock Holmes,The Adventure in the Copper Beaches

The main objective of this chapter is to describe each dataset, as well as their origins and the pre-processing methods if applied, and also describe the selected database.

Can we use datasets from Anoctamins potentiators & activators and CFTR correctors in order to predict these capabilities in new molecules, and furthermore proceed to a selection of molecules of interest? It is important to refer these two datasets: i)Anoctamins potentiators and activators, ii)CFTR correctors, were never tested *in silico*.

For the purpose of this project, the ZINC database was selected; ZINC is a free database with a wide range of commercially available compounds organized into subsets. To evaluate the final results it was selected a published bioactivity assay in CFTR.

3.1 Anoctamins potentiators and activators dataset

The molecules available in this dataset are the outputs of a similarity search that uses a previous discovered anoctamin 1 activator. The lab design involved a cell-based functional screening where mutant and non-mutant cells were used for

high-throughput screening using fluorescence to measure the I^- influx. A protein immunoblot was also performed to confirm the protein interaction.

3.1.1 Description

In this dataset, it is possible to encounter potentiators and activators. While these molecules have the same function - reduction on $[Ca^{2+}]$ needed to activate Cl^- currents - there is a difference in the amount of Ca^{2+} needed.

An activator strongly increases Cl^- current at 0 $[Ca^{2+}]$, whereas a potentiator is not active at a 0 $[Ca^{2+}]$ but reduces the EC_{50} - concentration where the effect is half of the maximum - for Ca^{2+} dependent activation (Namkung et al., 2011).

There are 10 compounds with their respective mean value and SMILES. The data is available for different compound concentration $0.1\mu M, 1\mu M, 10\mu M, 50\mu M$ and in several cases, $100\mu M$ - all the considered concentrations of compound were tested with $0.5\mu M$ [ATP] and without ATP. The concentration chosen for *in silico* testing was 10 μM from the test without ATP to ensure that the results were not affected by a purinergic receptor (ATP or ADP receptor).

3.2 Cystic fibrosis transmembrane conductance regulator correctors dataset

This dataset includes several groups of molecules (secondary and tertiary group) which originated from a similarity search using 4 primary CFTR correctors.

The CFTR correctors dataset comes from a western blot lab design. This technique also called protein immunoblot is used to find specific proteins in samples, and it considers the following three steps:

1 - Gel electrophoresis: Where the proteins are separated by a gradient of size (if they are polypeptides) or by type of structure. This gel has a variable number of wells where to deposit the samples.

2 - Transfer to a membrane : Transfer of molecules to a membrane that has affinity to proteins and ensures afterward that all the proteins are locked in place and that the membrane becomes not reactive to new bondings.

3 - Marking proteins: Marking the target proteins with antibodies. An antibody can react to more than one protein but the location on the molecule (that depends on the gel electrophoresis) and the marking make it possible to discard false positives.

Once these steps are finished, it is possible to identify the marked proteins. This detection can range from fluorescent to colorimetric or even radioactive. To quantify the amount of proteins in each band there are some tools that count the number of pixels giving the ability to compare measurable results.

3.2.1 Description

The files available in this dataset contain the values obtained in the western blot lab design for 4 primary molecules, 41 molecules from the secondary library and 64 from the tertiary library. One of those compounds was double counted resulting in a total number of 108 compounds. The CFTR folder presents the results of 95 membranes for 108 molecules and for all of the controls: DMSO (Dimethyl sulfoxide), F508del-cftr, Wild Type and C4a (CFTR corrector 4a). For *in silico* testing, all the compounds were selected plus the C4a - a positive control known to correct CFTR (figure 3.1).

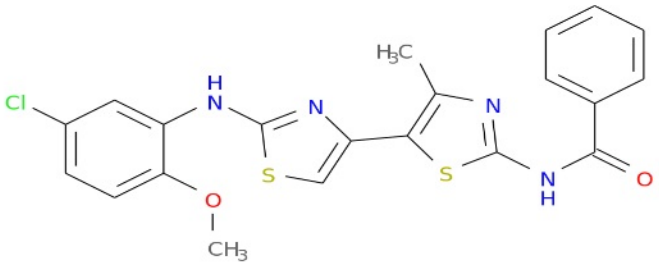


Figure 3.1:
2D representation of CFTR corrector 4a.

3.2.2 Pre-Processing

Given that each membrane has a different exposure time, a pre-processing of the data was required to normalize the lab results used as inputs. To cover this, the following formula was applied on each membrane results for each compound:

$$enhance = \frac{V_c - V_{F508del}}{V_{WT} - V_{F508del}} \tag{3.2.3.1}$$

Where;

$\mathbf{V}_c$ - compound lab result

$\mathbf{V}_{WT}$ - positive control value, for each membrane

$\mathbf{V}_{F508del}$ - negative control value, for each membrane

When the △F508 lab results for one membrane are not available, the median of all the △F508 data is used. The positive control represents the highest value in that membrane and the negative value should be the lowest.

If the result from this formula is negative, meaning a compound enhance value lower than the △F508 value, then the enhance value of that compound will be set to zero.

In the figure 3.2 it is possible to visualize the difference after the data was processed. After the treatment the enhance values of the majority become zero or close to zero; in this data, more than 50% of the molecules have an enhance value equal or below the value of the CFTR $\triangle F508del$.

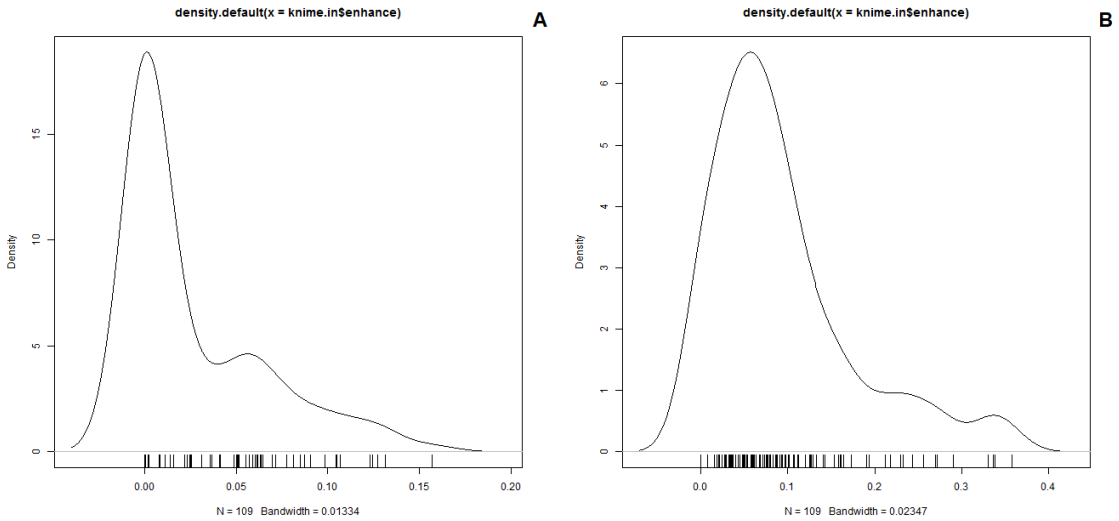


Figure 3.2: Density representation of the enhance values in the initial data. A - Before pre-processing. B - After pre processing.

3.3 ZINC - Database for virtual screening

The virtual screening step used in this project requires a compound database as a dataset to infer the medical viability of several molecules. For this purpose, the ZINC database was chosen.

As mentioned before, ZINC is a free database of commercially available compounds that contains over 35 million molecules organized into subsets (by physical properties and purchasable availability). It is possible to create small subsets for personal use and there are several queries available: by name, structure, biological activity, physical properties or even CAS number (unique identifier assigned by Chemical Abstracts Service to every chemical substance published from 1957 [1]) (Irwin et al., 2012).

This database is widely used in virtual screening approaches: structure-based discovery to identify novel Smoothened (a Hedgehog group of proteins related to cancer) ligands with different chemotypes in an effort to combat treatment resistance (Lacroix et al., 2016); virtual screening approach to find new leads

[1] CAS - A division of the American Chemical Society, CAS REGISTRY and CAS Registry Number FAQs , website last accessed 15 August 2016

against bovine brucellosis (Li et al., 2016); search for non-streroidal CYP17A1 inhibitors to treat prostate cancer (Bonomo et al., 2016); ligand-based drug design for beta-1,3-glucan synthase inhibitors to conceive less toxic antifungal molecules (Meetei et al., 2016).

It is important to mention the ability to download the entire database or subsets in different formats, making it easy to run virtual screening locally. For this project, the "all now" subset was chosen as it contains all the molecules available in stock.

3.4 CFTR bioactivity assay

The assay selected is an already published bioactivity study from The Broad Institute of MIT and Harvard [1]. This assay uses halide-sensitive yellow fluorescent protein (therefore, sensitive to Cl^-) to determine the active concentration of a drug in a cystic fibrosis bronchial epithelial cell line. In this assay, 1170 compounds were tested in the $\triangle F508$ mutation in Human bronchial epithelial cells. Out of these 1170 compounds, 605 were considered active, 559 inactive and 6 inconclusive. To facilitate the analysis these 6 inconclusive molecules were removed from the set.

[1] the study was published in 2014 with the PubChem AID 743267, external ID - 7017-01_Other_Dose_CherryPick_Activity

Chapter 4

Methods

"Truth has nothing to do with the conclusion, and everything to do with the methodology."
Stefan Molyneux

4.1 Software

R[1] is an open source programming language and environment which is designed to perform an ample range of statistical methods. There is a large number of libraries available that can aid in the construction of a program. This language was chosen to do several statistical evaluations as well to create plots and graphical representations of the data.

KNIME®[2] Analytics Platform was the chosen software to support the construction of the project workflow. It is an open source data analytics platform that allows the construction of workflows using a graphical interface. In KNIME there is a wide range of nodes available, including those with the ability to read different input files, column sorting, machine learning, or even database connections. Also, this program supports scripting in several coding languages like Python[3], R and Java, and allows the utilization of external tools in the computers command line.

[1] R - https://www.r-project.org, last accessed 18 August 2016
[2] KNIME - https://www.knime.org/, last accessed 17 August 2016
[3] Python - https://www.python.org/, last accessed 17 August 2016

RDKIT [1] is an Open-Source Cheminformatics Software that can be used to calculate substructure searching, calculate different molecular fingerprints and be used to filter functional groups. This software has several nodes available in KNIME and can be used in Python and C++ coding languages.

4.2 Workflow visualization

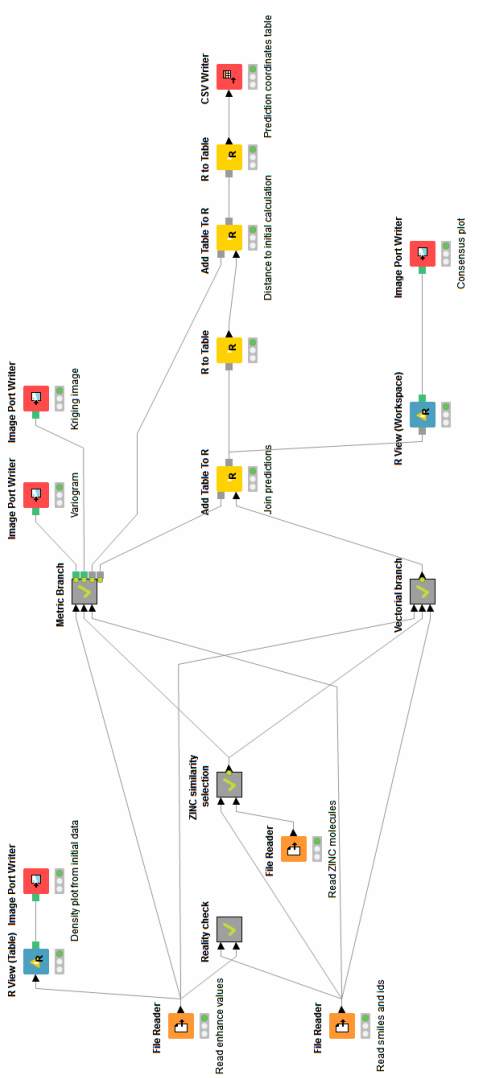


Figure 4.1: Workflow main view screenshot.

[1] RDKIT - http://www.rdkit.org/, last accessed 17 August 2016

The main view of the workflow is available in image 4.1 where there are several meta nodes (grey squares) which represent a group of nodes. Each part of this workflow will be explained in the next sections.

4.3 Workflow input

In the project workflow, the input information will be divided into three files: i).smi file with SMILES and identification numbers of the dataset in use; ii).txt file with the enhance values as well as their identification; iii).smi file from ZINC with "all now" subset which has all the molecules available in stock.

4.4 Virtual screening candidates

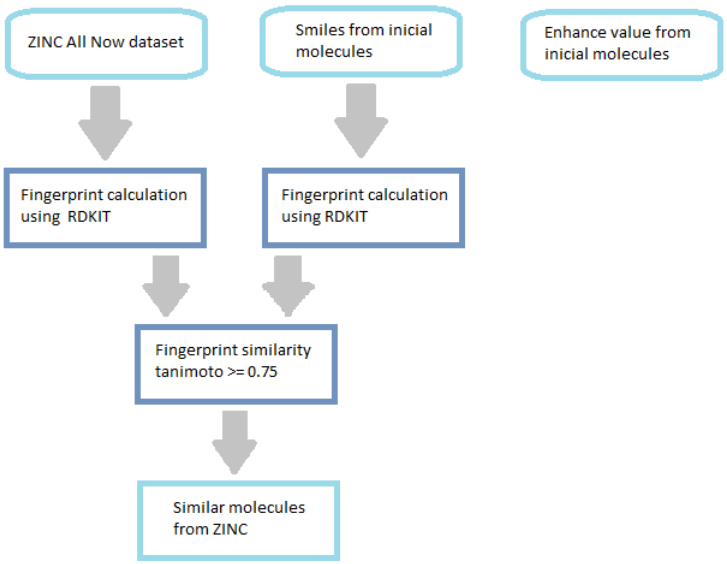


Figure 4.2: Workflow ZINC similarity representation.

The first step of this workflow is to select molecules from the zinc database that are similar to the ones of the dataset in study (4.2 and B.1). With that objective in mind, Morgan binary fingerprints will be calculated using RDKIT node present

in KNIME. The parameters to be chosen are a radius of 2 atoms and a 1024 bits length fingerprint (default values). To compare the fingerprints the similarity metric selected will be the Tanimoto coefficient, and the ZINC compounds picked for further analysis will have 0.75 Tanimoto similarity or more, as they are more accurately to predict. Otherwise, the whole process will be time-consuming and imprecise.

4.5 Vector space mining

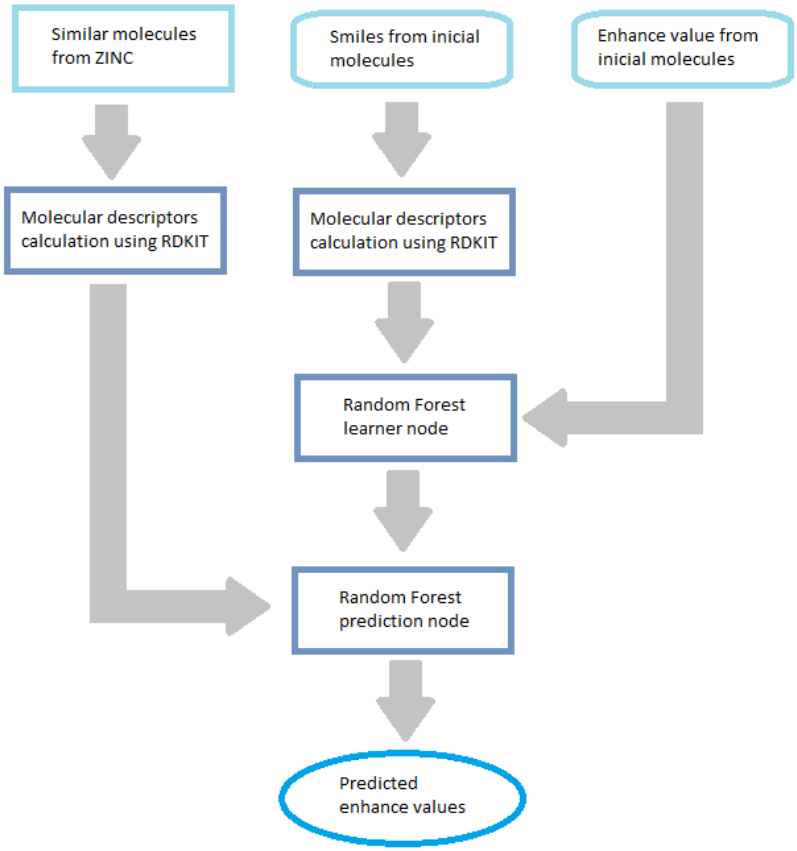


Figure 4.3: Representation of vector space workflow branch.

In the vector space mining branch (figure 4.3 and B.4) both molecular data (i)the dataset being studied, ii)previous selected molecules from ZINC) will have their

descriptors calculated via an RDKIT node (molecular descriptors list available at Appendix A). The descriptors of the initial molecules and their enhance values will be used to train the random forest learner node.

Since random forest is insusceptible to noisy variables and can operate with more variables than instances, a decision was made to keep all the molecular descriptors without a previous selection.

The model created in the random forest learner node, as well as the molecular descriptors of the ZINC molecules, will be used in the random forest node to perform a regression analysis. Here the enhance values will be predicted using an average of each prediction from individual trees.

4.6 Metric space mining

In the metric space mining part of this workflow several steps were taken:

1 - The first step of the metric space mining workflow branch (figure 4.4 and B.2) will be to calculate similarities between all the initial molecules. The selected comparison method will be *NAMS*.

2 - The similarity values will be transformed in distance using the following formula:

$$D = -log(s) \tag{4.6.1}$$

Where;

D - compound distance to other compound

s - compound similarity to other compound

3 - A principal component analysis will be then performed to evaluate the diversity to be accounted into two dimensions using *prcomp* in R.

4 - A principal coordinate analysis or classic metric multidimensional scaling will be performed to represent the compounds in a two-dimensional graph with *cmdscale* in R.

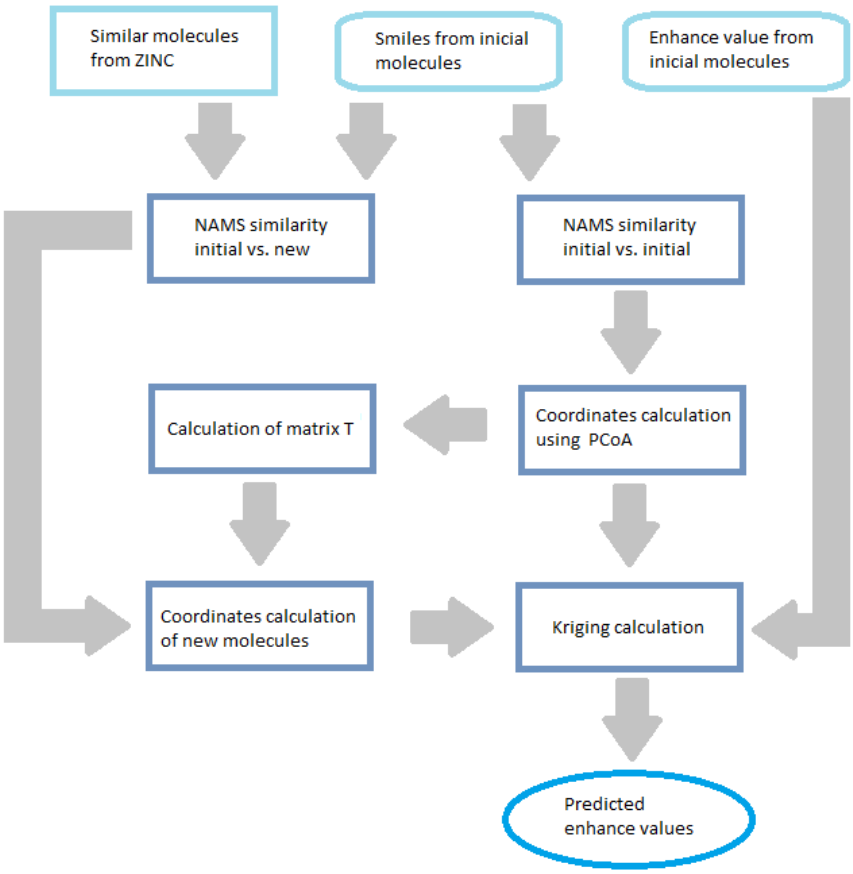


Figure 4.4: Representation of metric space workflow branch.

5 - Perform interpolation using kriging (B.3). This prediction will be executed in the R environment using ordinary kriging available in Gstat library.

In order to *cmdscale* perform a principal coordinate analysis, it needs the distances from each point to all the other points. This distance is the dissimilarity or distance calculated using NAMS results in formula 4.6.1. Since an evaluation of all the initial and new molecules would be time-consuming (due to the computational cost of NAMS), it will be necessary to construct a transformation matrix that when multiplied by the distances between the initial molecules and the new ones gives the coordinates for the new molecules. Thus, a similarity search will be performed using NAMS to return the similarity values of each new molecule to

the initial dataset and subsequently convert to distance as already described. The development of the transformation matrix depends on the following principles:

$$A * T = B \tag{4.6.2}$$

Where A is the distance matrix between all the initial molecules, T is the transformation performed by the cmdscale function and B is the matrix with the final coordinates. Since the objective is to have the new molecules represented in the same two-dimensional plan, the transformation matrix is required to calculate the coordinates of the new compounds using their distances to the initial molecules.

In order to obtain the T matrix the following formula was used:

$$T = A^{-1} * B \tag{4.6.3}$$

The T matrix will be calculated and then used to obtain the new coordinates using the new distance matrix as in the formula 4.4.2. After these steps, all the new and initial compounds can be represented in the same metric space and the prediction can be performed.

4.7 Selection of molecules of interest

The selection of lead compounds will be done by choosing the most promising molecules present in the upper right quadrant of the consensus plot. Only the top 10 molecules from all the matrices (or the majority of matrices) selected for that specific dataset will be identified for further analysis. Molecules that have a prediction higher than the midpoint value in one method and a lower than midpoint value on the other method may be considered as outliers. From this group of molecules, the selection for further testing relies on the largest difference from predictions.

Chapter 5

Results

"The devil is in the details."

Anonymous

5.1 Anoctamins

In this dataset, the random forest algorithm can be less accurate as there are only 10 molecules in the dataset, which is considered a low number to train this statistical method. Also, a proper cross validation is difficult to achieve, because each molecule contains a large amount of different information to be considered in the model; for this reason, a validation was not performed.

Considering a tanimoto score of 0.75 or above, 71 molecules were selected from ZINC database.

5.1.1 Variograms from metric space

The variogram plot is a good measurement to check if the data is spatially correlated allowing the choice of an atom distance matrix for further analysis. Considering this, the atom distance matrices that show some correlation are matrix 0, matrix 1 (figure 5.1). As matrix 2, matrix 3 and matrix 4 do not show enough correlation they were not chosen for further analysis.

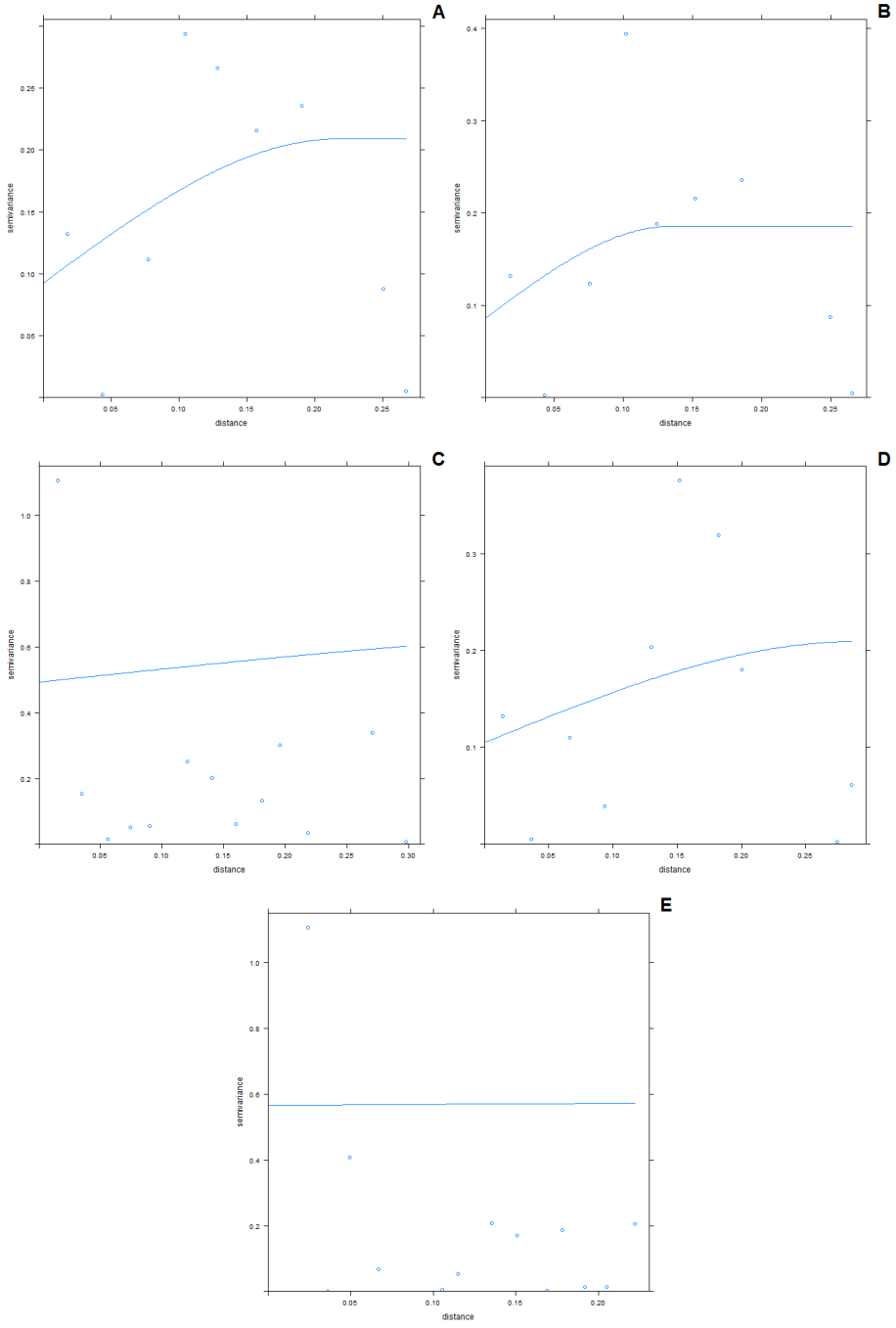


Figure 5.1: Variograms from anoctamins data with different matrices. A - matrix 0. B - matrix 1. C-matrix 2. D - matrix 3. E - matrix 4.

5.1.2 Kriging prediction

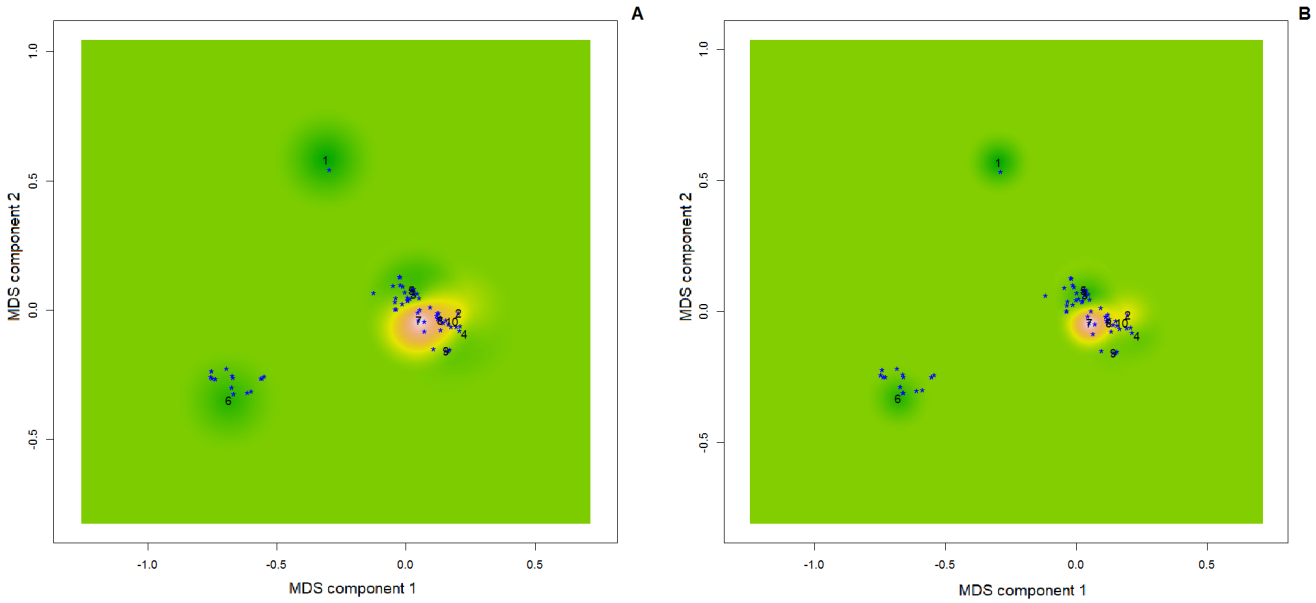


Figure 5.2: A - Matrix 0 B - Matrix 1. Kriging image interpolations, the numbers correspond to the initial molecules ids and the blue asterisks represent the locations of the new molecules from ZINC. The color palette expresses the predicted values in that metric space where green is the lower prediction and white the highest.

Giving the previously presented variograms the rest of the kriging prediction was performed for matrices 0 and 1 with the respective kriging interpolation plots (figures 5.2). The cumulative proportion of two components in the principal component analysis for matrix 0 was 0.8084 and for matrix 1 was 0.8115, demonstrating that two components are enough to substantially represent this data. In the kriging images represented in the figure 5.2, it is possible to observe the location of the hotspots as well as the new molecules location in the metric space. Both images show a low number of molecules even before the selection for further analysis. Also, it is possible to observe that between the clusters and even between molecules from the same cluster there is a lot of metric space to explore, because there were no molecules in ZINC for this locations. This can be solved with a *de novo* approach.

5.1.3 Consensus plot graph

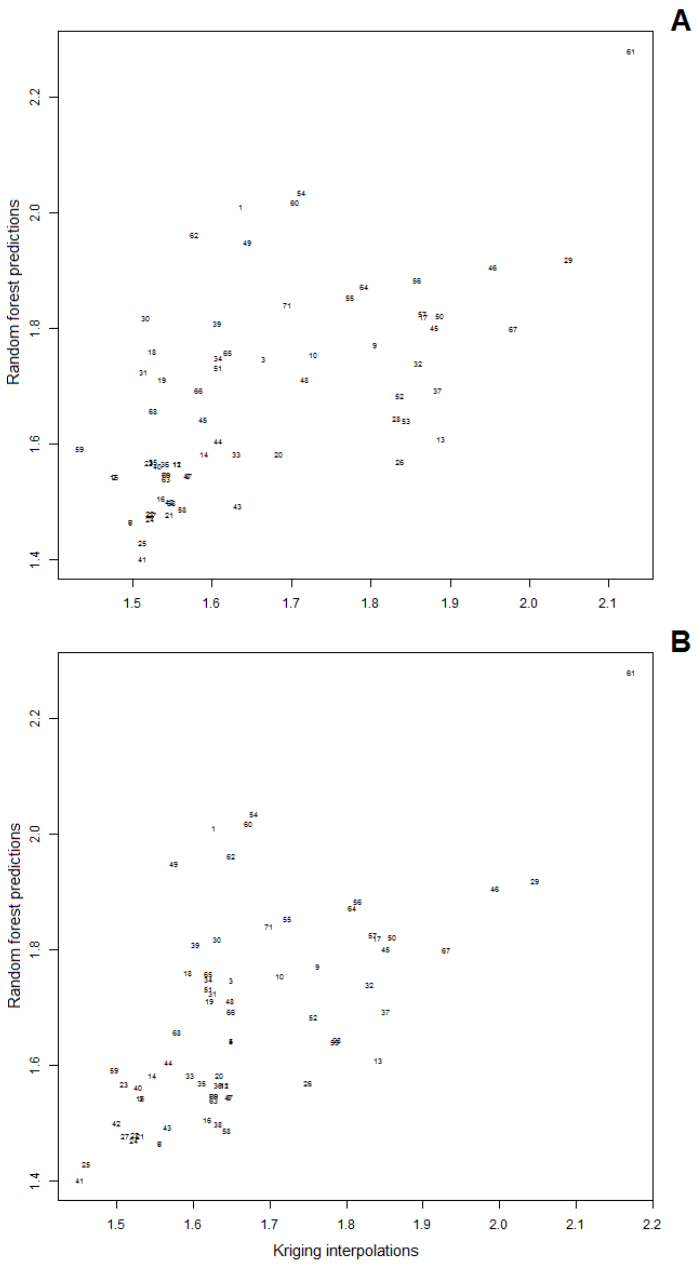


Figure 5.3: Random forest vs kriging prediction values with matrix 0 (A) and matrix 1 (B) , from the new molecules from ZINC, the numbers represent this project given ids.

When using two methods for the same prediction it is possible that some molecules will have a good prediction in one method and a bad one in the second method. Therefore a consensus plot was developed to represent the differences and the concordances in the prediction values. This plot will show the predictions for each molecule, where the X axis are enhance values predicted by kriging and the Y axis are enhance values predicted by random forest.

The consensus plot graph (figure 5.3) clearly shows, that molecule 61 as the best prediction values in both methods while molecule 62 and molecule 13 are outliers to be considered. For the remaining molecules further analysis is needed, using a supporting table where the coordinates for each molecule can be analysed.

5.2 CFTR

Using the CFTR dataset, 301 molecules were selected from ZINC to be evaluated by random forest and kriging methods. The molecules selected had a tanimoto similarity score greater or equal to 0.75.

5.2.1 Initial modelling

5.2.1.1 Variograms of the data

The methods used in the metric space depend on the correlation of the data, and for this purpose, a variogram graph can indicate the reliability and feasibility of such prediction for each individual matrix (figure 5.4). In this case, the validation was performed using matrices that revealed some correlation. It was not possible to perform this method with matrix 3 and 4 since the molecules were too similar and several were categorized as equal.

5.2.1.2 Validation

When performing a 5-fold cross validation of the random forest method using CFTR dataset (which have a standard deviation of 0.0387) the RMSE obtained was 0.0382 and a explained variance of 0.0227. In the out of the bag validation, the results were 0.0387 for RMSE and a negative explained variance of -0.0004.

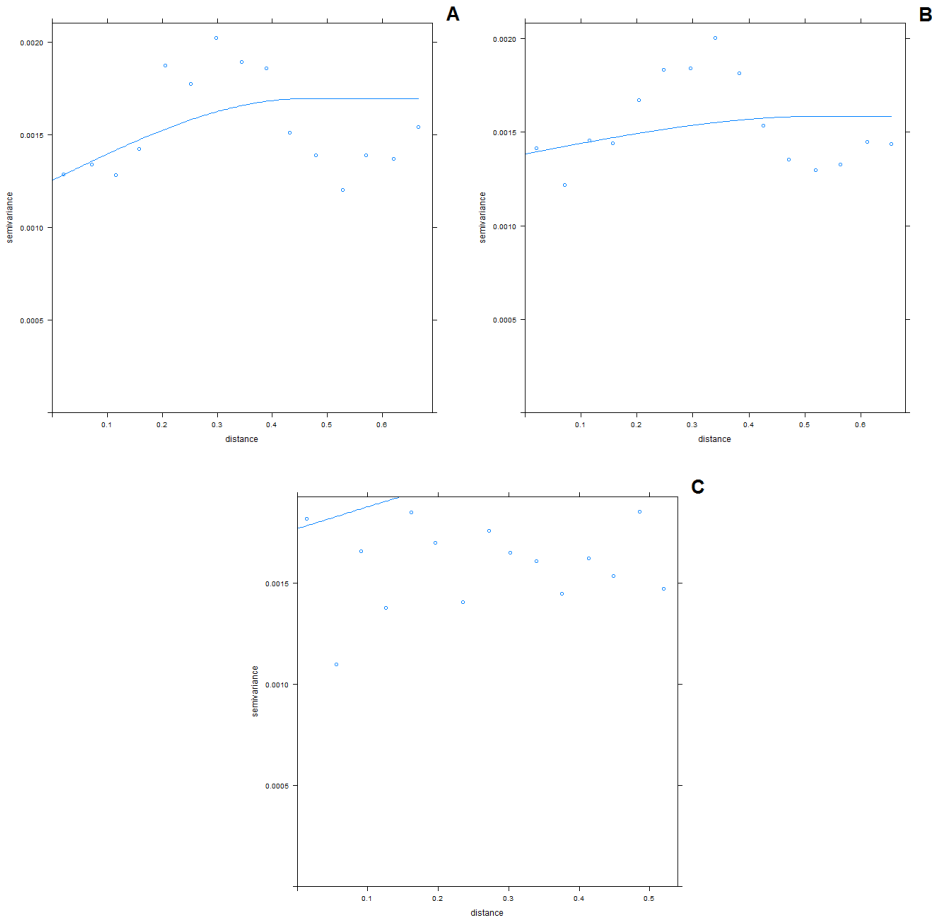


Figure 5.4: Variogram of CFTR data with matrix 0 (A) , matrix 1 (B) and matrix 2 (C).

These results show that the random forest model that receives the molecular descriptors (the standard and widely used approach) is not able to explain the data and therefore to predict values in new molecules.

The validation of the kriging algorithm was performed using a 5-fold cross validation for the matrices 0 and 1. The RMSE for matrix 0 was 0.0413 and for matrix 1 was 0.0403. As for the explained variance, the result for matrix 0 was -0.1419 and for matrix 1 was -0.0860. These results are aligned with the validation results observed in the vector space branch, where random forest was not able to predict the characteristics of new molecules accurately.

5.3 Confirmatory model with raw data

Based on this validation, all the processes were analysed for errors where no issue was found. Looking at the negative values of the explained variance in the vector space method (already a widely tested process), a decision was made to perform several cross validations using the random forest method; additionally, changes to the initial input data were also performed. When all the instances for each molecule and without the preprocessing treatment were tested, the explained variance obtained was 0.69. When the information about the membrane and well of each molecule were removed, the explained variance dropped to negative values (-0.1053). These new tests indicate that the decisive characteristics for the prediction was the molecule location (membrane and well) and the molecule compound id but not the molecular descriptors, therefore the only plausible option is the mislabeling of our initial data. In order to be possible to perform the molecular virtual screening is necessary to correct the mislabeling first.

5.4 Identifying experimental mislabeling data

5.4.1 Mislabelling identification

Mislabelling of the data is difficult to correct since the experimental part was performed by another party. Repeating the entire experiment was impossible given time and money constraints, therefore the suitable solution was to create a machine learning algorithm with the only information considered to be less prone to have experimental errors. This machine learning algorithm iterates in all the remaining data and one by one exchanging the compounds enhance values to try to understand which changes affect the explained variance and subsequently the prediction capability. This machine learning algorithm was iterating in molecular similarity comparison and performed permutations until the explained variance could not be enhanced any further. In this case, the data chosen as correct was the primary molecules as well as the C4a corrector. Using the data from this 5 compounds a support vector machine was created.

5.4.2 Reality check

Given the problems with the initial dataset and to ascertain if the model created with this transformed data can predict the activity of molecules from within the dataset and the activity of new compounds foreign to the initial dataset. To test the model created with this corrected dataset the already mentioned bioassay available at Pubmed was used.

A classification using random forest was performed to validate the model and predict the activity or inactivity of the compounds as tested by the study. A confusion matrix was created with the results of this classification. A confusion matrix is a 2x2 table were in the columns the variables are predicted positives and predicted negatives and in the rows, the variables are real positives and real negatives. Here, the values considered as real are the positive and negative classifications from the published assay. From the same classification, precision and recall values were also obtained. The precision values can help realize the proportion of true positives in all the predicted positives from the algorithm. This precision or positive predictive value are obtained following this formula:

$$Precision = \frac{TP}{FP + TP} \tag{6.2.1}$$

TP - True positives - predicted positives that are real positives;

FP - False positives - predicted positives that are real negatives.

The recall value shows the rate of true positives in all the actual positives predicted or not. Recall value or true positive rate can be calculated using:

$$Recall = \frac{TP}{FN + TP} \tag{6.2.2}$$

TP - True positives - predicted positives that are real positives;

FN - False negatives - predicted negatives that are real positives.

Given that the virtual screening selection relies on the true positives predicted, both these values (precision and recall) are important to understand the rate of true positives in all the predicted positives, as well as to realize from all the real positives in the data which are predicted as positives.

The molecules available in the published dataset have a different range of similarity when comparing to our initial dataset, as the values of similarity score obtained with NAMS (using matrix 0) can range from 0.0269 to 0.9321. When performing a prediction is necessary to have a degree of similarity, therefore in order to achieve proper validation results, the top 10 and 20% similar molecules were selected to be tested and the respective precision and recall values were calculated. With the top 10% of molecules the precision obtained was 0.60 and with the top 20% the precision obtained was 0.58. For both subsets, the recall obtained was 0.12.

Regarding the metric branch, several different approaches were performed because the enhance value is calculated using different processes, making it difficult to ascertain the validation directly. First, the values of enhance for the top 10%and 20% similar molecules were predicted using kriging, then these values were categorized as positives and negatives using the average values of the ΔF508 CFTR mutation as a threshold. With this categorization was possible to create another confusion matrix and evaluate the already mentioned metrics. The results were 0.48 for the precision value and 0.46 for the recall value. Considering that this classification was performed after the prediction there is the possibility of error associated with the prediction and associated with the classification (primarily the choice of the correct threshold), therefore a method for evaluation using only NAMS was required.

For each primary molecule a plot with the predicted density of positives and negatives throughout the similarity score was generated (graph 5.5). In these density plots, is possible to observe that the actives have higher density values than the negative compounds. The difference sometimes is more visible than others but suggests that similar compounds to the active ones are themselves active as well.

Using the similarity calculated with NAMS (matrix zero) between the primaries from our dataset and the molecules from the pubchem study, as well as

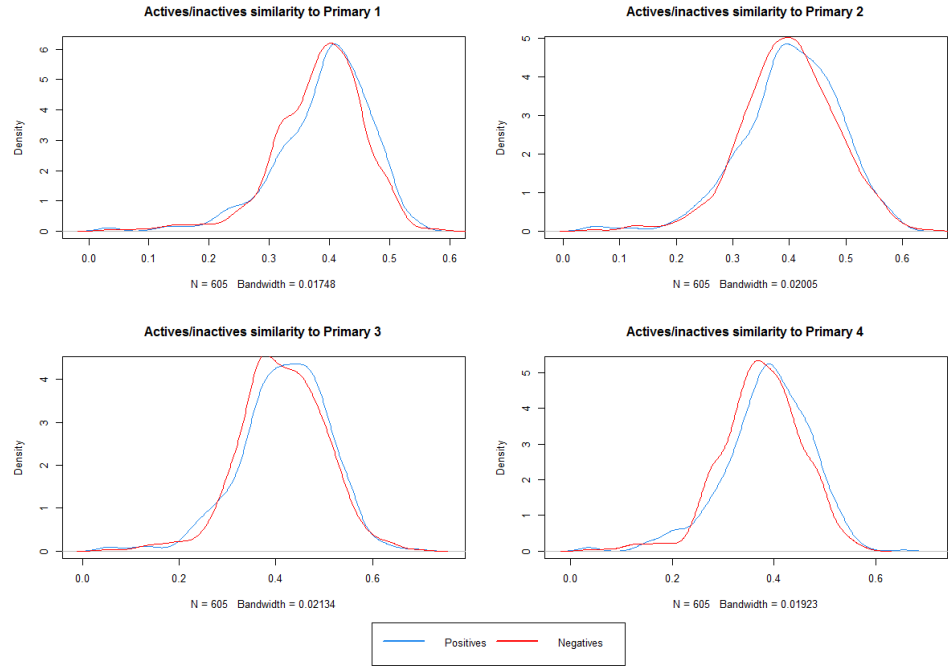


Figure 5.5: Fraction of actives and inactives per similarity and per primary compound.

their activity, a spline graph was created. A spline graph has a line which is defined piece by piece using polynomial functions, in order to be approximated to the real numbers and at the same time, construct a smooth line. This graph was build considering the molecules with more than 0.3 similarity score, given that a lower similarity score would not give pertinent information for this matter. When analysing graph 5.6, the primary molecules 1 and 4 show an increasing fraction of positives when advancing into more similar molecules. The primary molecule 1 is the best of all primaries, thus it is very important that this molecule has a good result, considering the positive fractions in increasing similarity. The primary molecules 2 and 3 have similar results (as expected as they are very similar). In both cases, the fraction of actives follows the increasing similarity and starting from ~ 0.5 similarity the fraction of actives decreased, mainly due to several negative molecules closely similar to these two primaries.

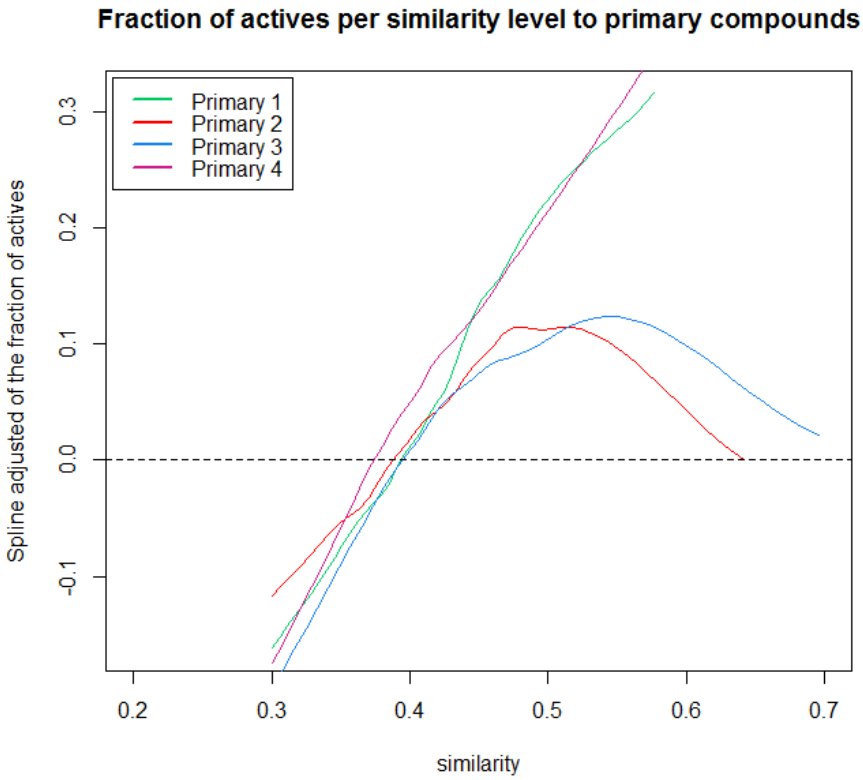


Figure 5.6: Fraction of actives per similarity and per primary compound considering only compounds with more than 0.3 similarity score

5.5 Final modelling

As previously mentioned 301 molecules were selected from ZINC to be subjected to the virtual screening approach.

5.5.1 Vector Space

5.5.1.1 Cross-validation

The cross-validation of the corrected data was executed using a 5-fold and out-of-the bag validation. This time the explained variance was 0.37 and 0.36 respectively. The MSE was 0.004 with a standard deviation of 0.083. The mean error of 100 iterations is available in a density graph 5.7. A confusion matrix was also

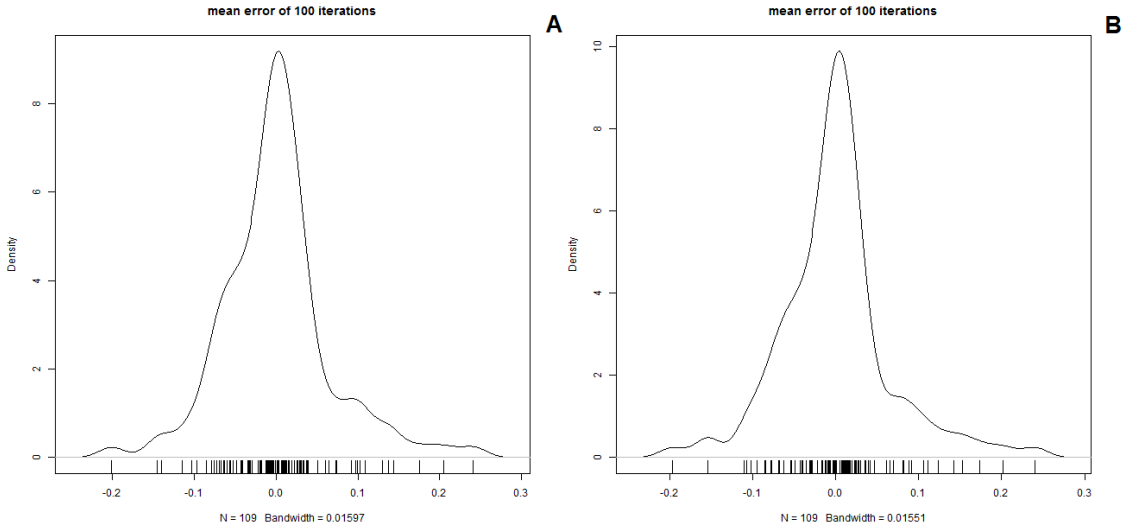


Figure 5.7: Density representation of the mean error in 100 iterations of the transformed data. A - 5-fold cross-validation. B - Out-of-the-bag cross-validation

created to compare with the results obtained in the reality check section. The results of this confusion matrix were 0.61 for the precision value and 0.12 for the recall value. These results are aligned with the ones observed in the previous section: 0.60 for the precision value and 0.12 for the recall value.

5.5.2 Metric Space

The metric space was created using matrices 0, 1 and 2. Matrices 3 and 4 described several molecules as equal given that some molecules are very close to each other.

5.5.2.1 Variograms of the data

New variograms had to be calculated given the modifications performed on the data. As it is possible to observe in graph 5.8 all the three matrices show correlation, thus the cross-validation will be performed on all of them.

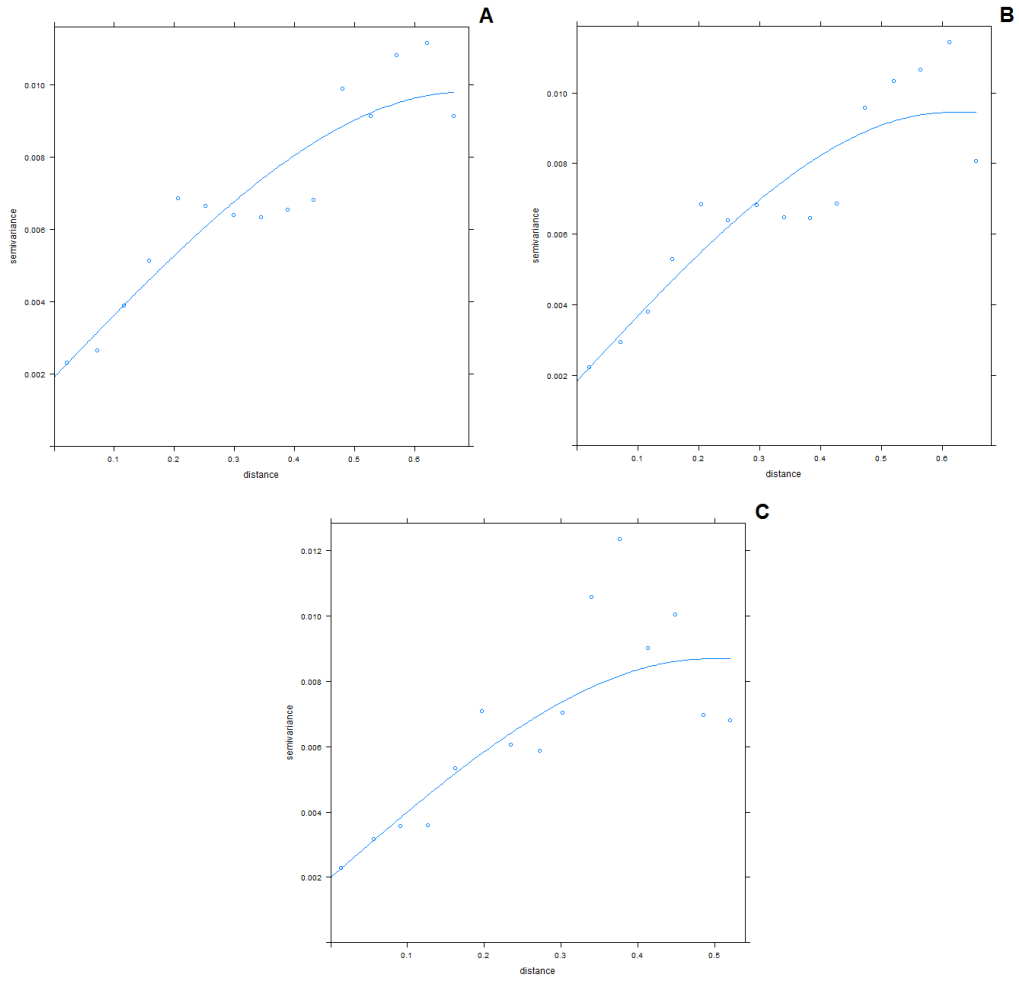


Figure 5.8: Variogram of CFTR corrected data with matrix 0 (A) , matrix 1 (B) and matrix 2 (C).

5.5.2.2 Cross-validation

The 5-fold cross-validation was performed for matrices 0, 1 and 2. The mean error of 100 iterations is represented in graph 5.9. The RMSE was 0.071 for matrix 0, 0.72 for matrix 1 and 0.070 for matrix 2. The MSE value was 0,0051 , 0.0052 and 0.0049 respectively. Matrix 2 had the best explained variance with 0.29, followed by matrix 0 with 0.27 and matrix 1 with 0.25.

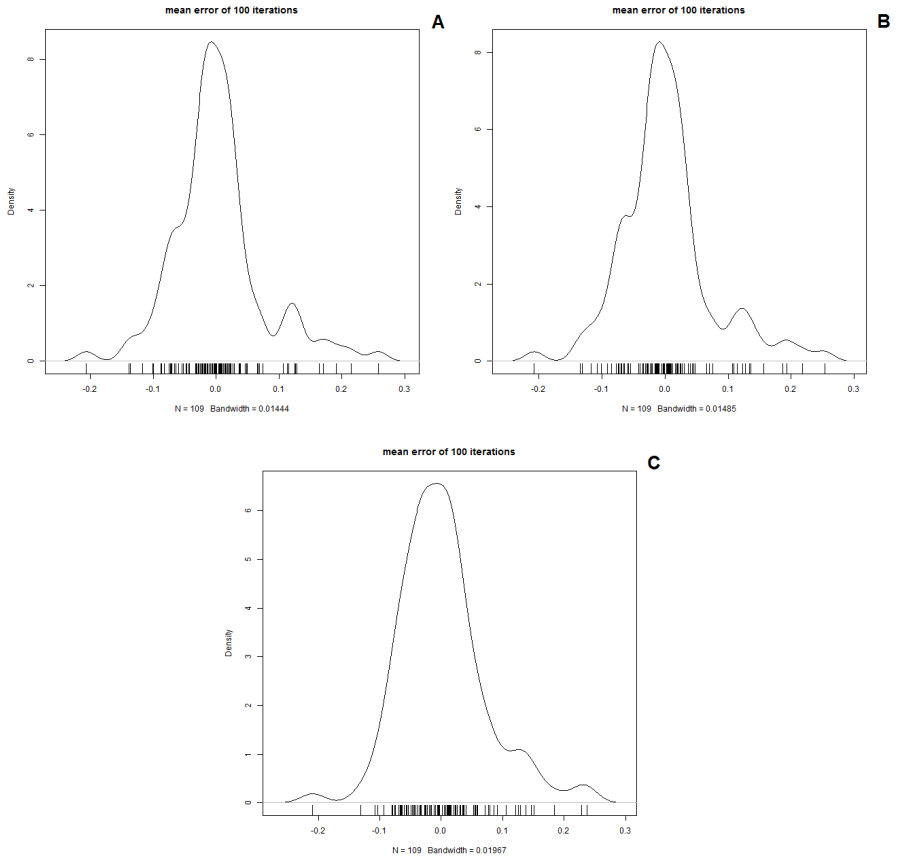


Figure 5.9: Density representation of mean error in the metric branch in 100 iterations. A - matrix 0. B - matrix 1. C - matrix 2.

5.5.2.3 Kriging - metric space prediction

The metric space prediction constructed with a principal coordinate analysis algorithm is illustrated in graph 5.10, where it is possible to see the metric space and the molecules positioning and distribution. It is also possible to see the difference between different areas concerning the amount of molecules.

The cumulative proportion of the two components ("X" and "Y" axis) range from 0.79 in matrix 2 and 0.83 for matrix 0, therefore with this values is possible to consider this metric space a good representation of the molecules in a two-dimensional space.

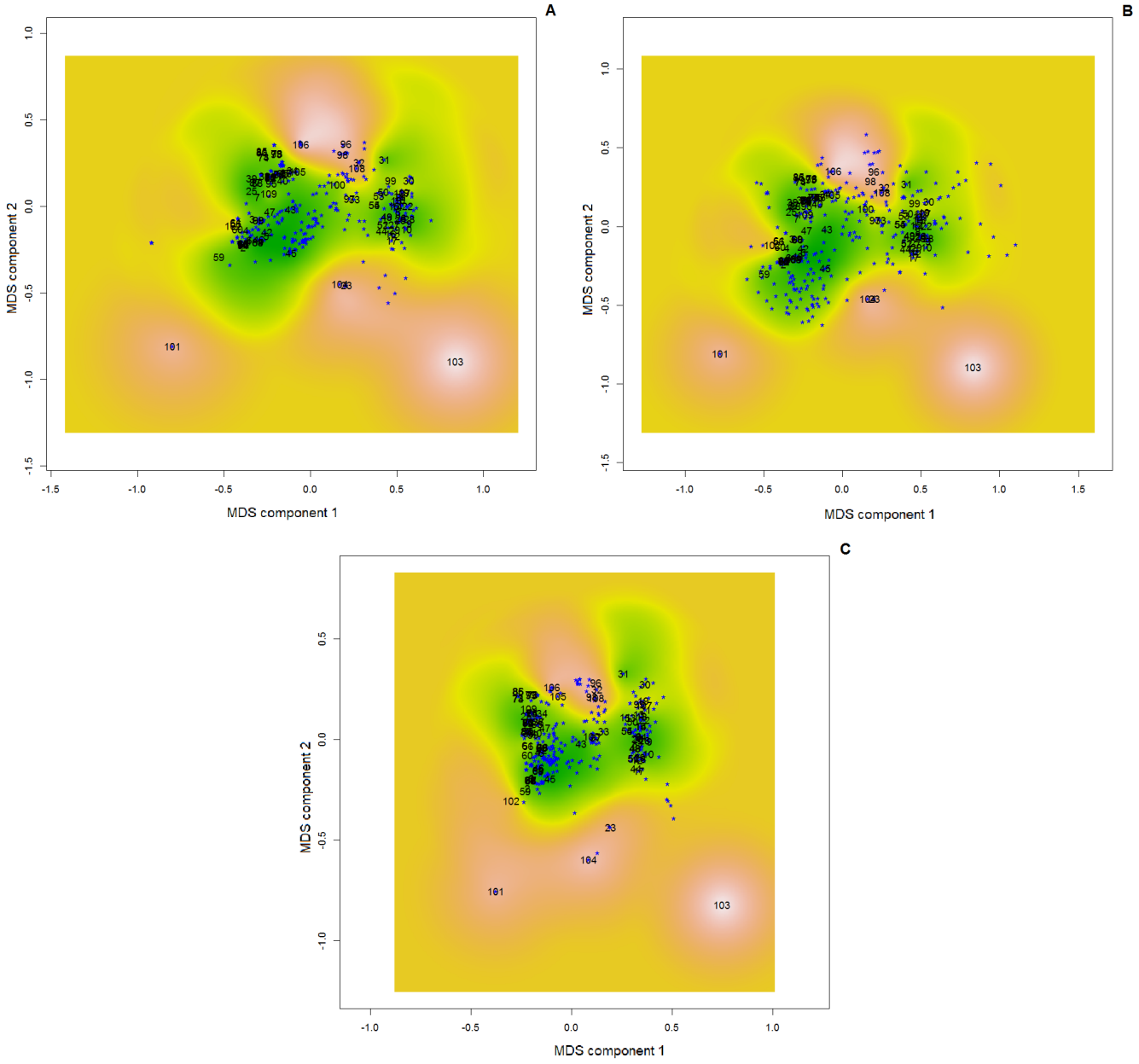


Figure 5.10: Kriging prediction image with initial and new molecules. A - Matrix 0. B - Matrix 1. C - Matrix 2. The initial molecules are represented by their id, and the new molecules are represented by blue asterisks. The color palette chosen was terrain, where green represents a low value and white a high value.

5.5.3 Consensus plot graph

The consensus plot graph (figures 5.11) represent the molecules predicted values in two axes: The X axis represents the kriging method, while the Y axis represents the random forest method. In the plots is possible to see the top molecules in both techniques, and also the molecules with low prediction values across the two.

Together with the supporting table, these plots will be used to select the top molecules and possible outliers in all the matrices. Even considering the overlap of several molecules, some molecules can be considered to be selected: molecule 9 and 208 which are top molecules across the three matrices. Molecule 49 can also be considered as it has a very low prediction from kriging and a considerable value from the vector space (49 is overlapped by molecule 50 and in the consensus plot in matrix 2). It is interesting to see a trend in the results of both spaces, which are in accordance with our theoretical assumptions.

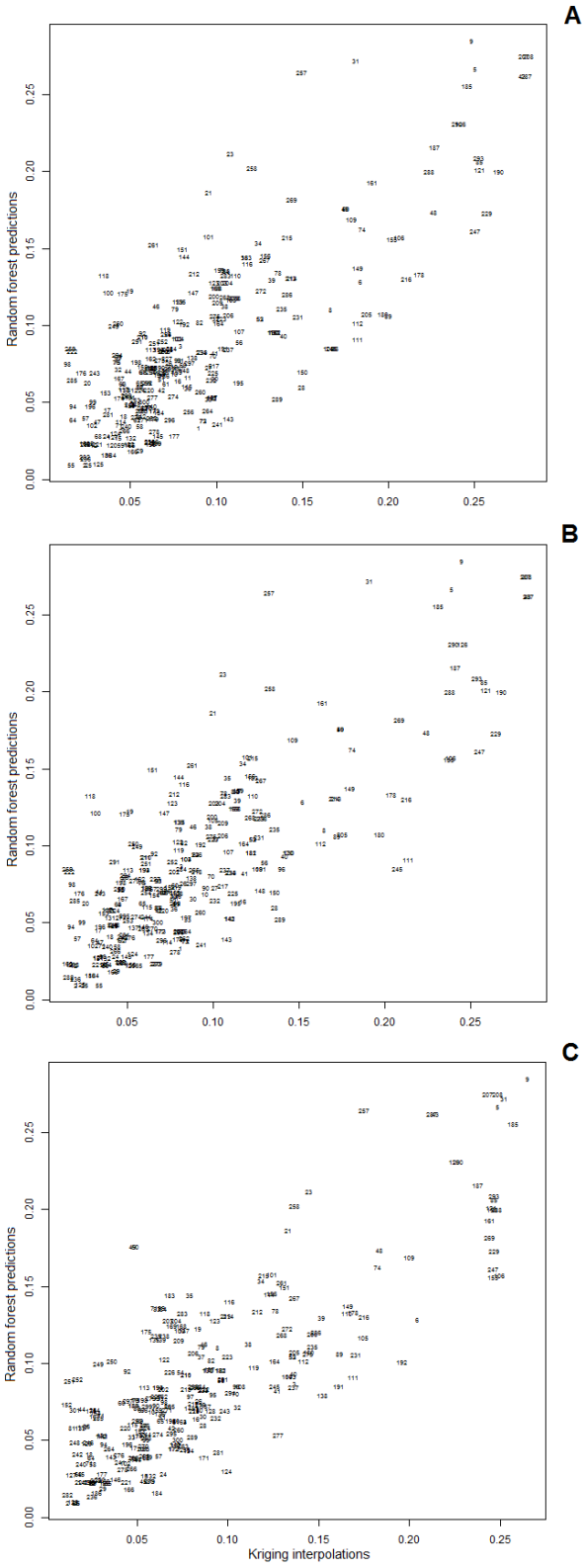


Figure 5.11: Consensus plot with the predicted values from metric branch on the x axis and the predicted values from the vector branch on the y axis. A - Matrix 0. B - Matrix 1. C - Matrix 2.

Chapter 6

Discussion

"Tell me and I forget, teach me and I may remember, involve me and I learn."
Xun Kuang

6.1 Anoctamins results

6.1.1 Most relevant descriptors in Anoctamins modelling

To evaluate the importance of each descriptor, every model created during the cross-validation had the number of trees which use the attribute as split on level 0 and 1 (root and the next immediately split) summed up and then ordered from the most used in this principal splits.

The most important variable seem to be SMR which calculates the molecular refractivity assuming the washed structure, followed by Slogp the logarithm of the coefficient octanol/water for washed structures (in the correct protonation state). The 3rd most important variable is LabuteASA (Labute's aproximate surface area) gives the surface of each atom that is not inside or coincide with other atom surface, giving the exposed surface of the molecule. The 4th and 5th most relevant descriptors are the average and exact molecular weight. There are other variables that can be also considered as relevant descriptors such as the TPSA (topological polar surface area - which acts as a representation of the molecular permeability), the number of hydrogen bond acceptors, the connectivity descriptors Chi1v and

Chi2v [1], the number of heavy atoms and finally the total number of atoms.

With these descriptors in mind, it is safe to assume the importance of the solubility characteristics of the molecule as well as the connectivity between atoms (especially the heavy) are very important characteristics to determine the potentiation or activation of anoctamins.

6.1.2 Chosen matrices

The matrices that show some correlation in variograms 5.1 presented in the last chapter were 0 and 1, the most strict matrices which can led to the assumption that a substitution of one single atom can severely affect the molecule ability to potentiate anoctamins action.

6.1.3 Selected molecules

As already referred in the methods section, it is important to select the best combined predictions from both spaces, as well to choose some outliers of each method (if there is some,) as a way to ascertain the error of this predictions when compared to *in vitro* testing.

Based on the methods previously described there are 8 top molecules that have the best results in both methodologies. These molecules are represented in image 6.1. Two outliers were also selected and represented in figures 6.2 and 6.3.

[1] the atomic valence connectivity index and can be calculated by the sum of $1/\mathrm{sqrt}(v_i j_i)$ where i<j for all the bonds enclosed by heavy atoms i and j

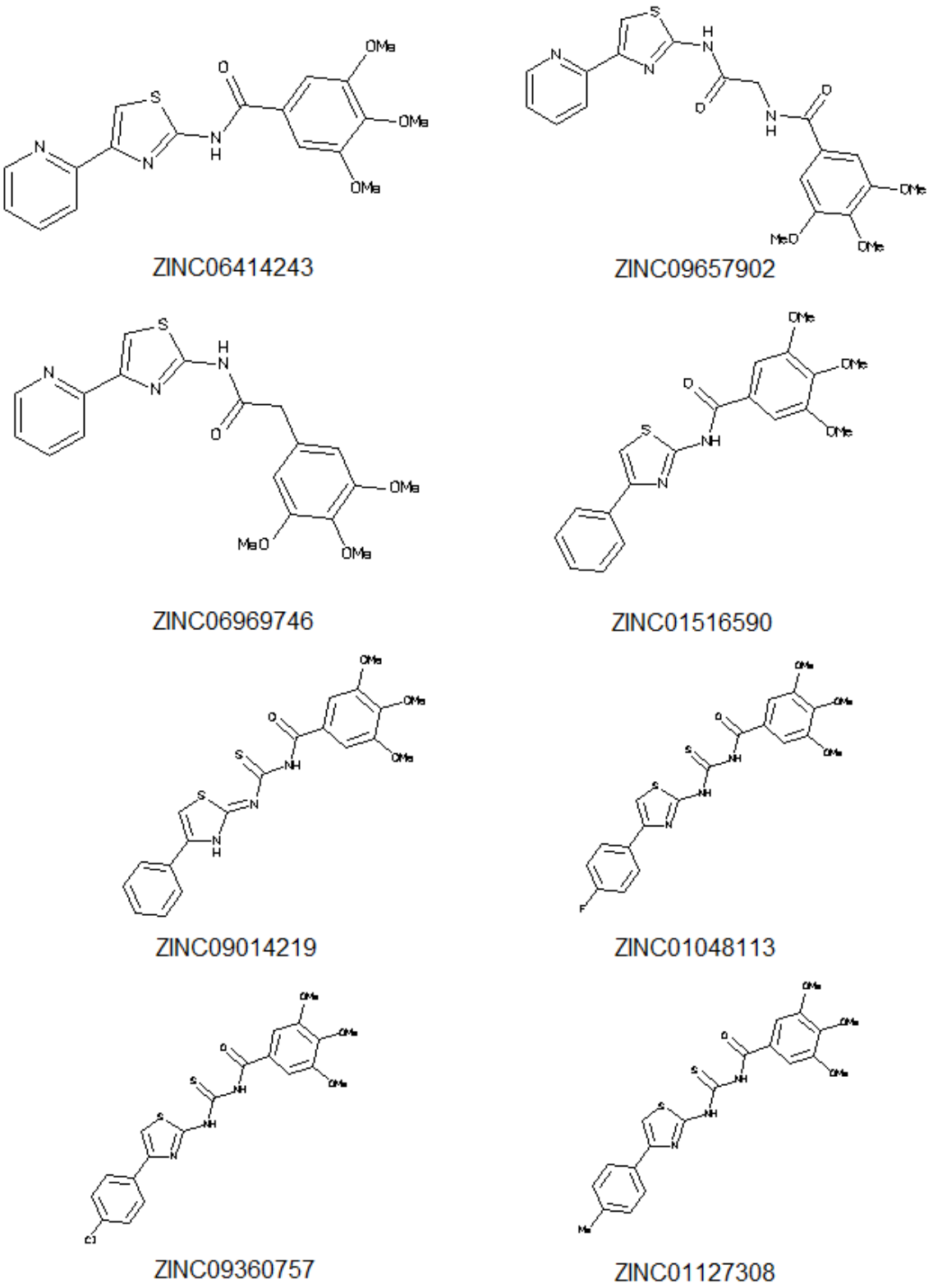


Figure 6.1: Final chosen molecules with good score from both methods. Molecular representations retrieved from ZINC.

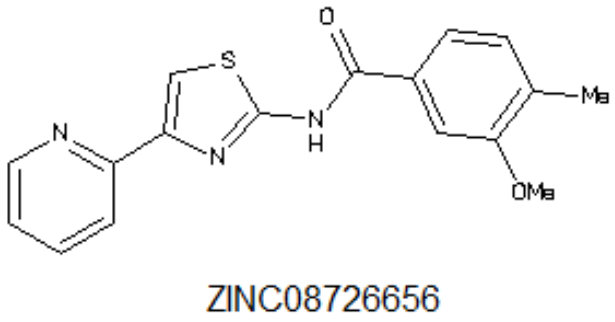


Figure 6.2: Final chosen molecule with high score in random forest and low in kriging.
Molecular representations retrieved from ZINC.

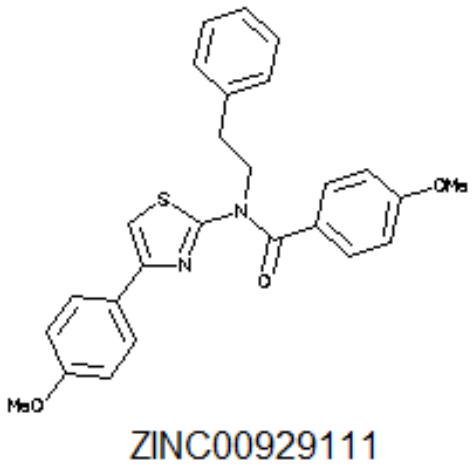


Figure 6.3: Final chosen molecule with high score in kriging and low in random forest.
Molecular representations retrieved from ZINC.

6.1.4 Confirmatory analysis

There are some evidences that support these results. It was performed a search in different information repositories and some data was collected.

In Pubchem [1] ZINC09657902 can possibly target a calcium-dependent protein kinase in Plasmodium Falciparum (malaria causing protozoan parasite) which can be interesting because anoctamins are calcium-activated channel. ZINC06969746 can possibly target anoctamin 1 in Mus Musculus (commonly called as house rat). ZINC08726656 was tested as potential chaperone treatment of Gaucher Disease.

6.2 CFTR results

6.2.1 Most relevant descriptors in CFTR modelling

The selection of important variables were performed in the same way as for the anoctamins dataset. The most important descriptor seems to be SMR, followed by LabuteASA. Then, two descriptors for the molecular weight (both average and exact) seem to have almost the same importance. In 5th place, a connectivity descriptor appears, Chi1v. There are other variables that can be also considered as relevant descriptors such as Slogp, the number of atoms, TPSA, PSA (polar surface area - defined as the sum of the surface of all the polar atoms), two connectivity descriptors (Chi2v and Chi2n - simple molecular connectivity index for path), and finally the number of heavy atoms.

These important variables are in accordance with the important characteristics in the anoctamins data namely the solubility and the number of atoms (especially heavy atoms).

6.2.2 Chosen matrices

Given the variograms presented in the last chapter, the top molecules were chosen from the matrices that have a good variance/distance relation. These matrices are matrix 0, matrix 1 and matrix 2 (figure 5.8). Given matrices 0 and 1 strictness

[1] last accessed in 15th of March of 2017

it is possible to assume that the enhance value is correlated to strict configurations changes and that a substitution of one single atom is really important for the molecule ability to correct CFTR. Nevertheless, this does not diminish the structure importance of the molecule but gives a remarkable penalization of atom substitution.

6.2.3 Selected molecules

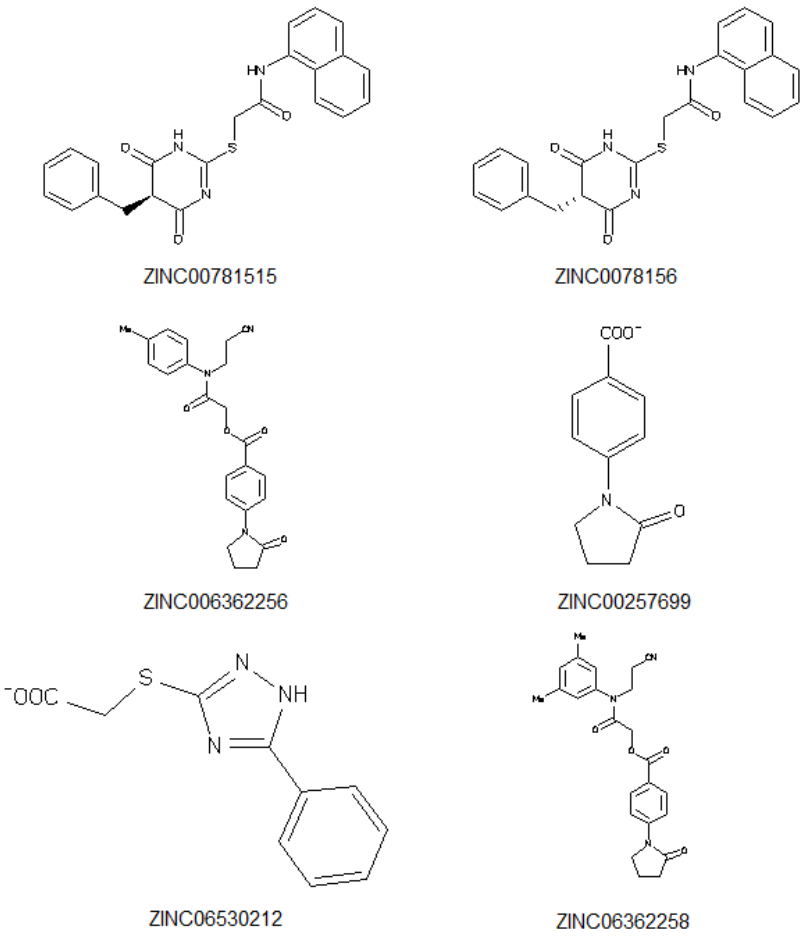


Figure 6.4: Final chosen molecules from virtual screening with top results for both techniques in the three matrices.
Molecular representations retrieved from ZINC.

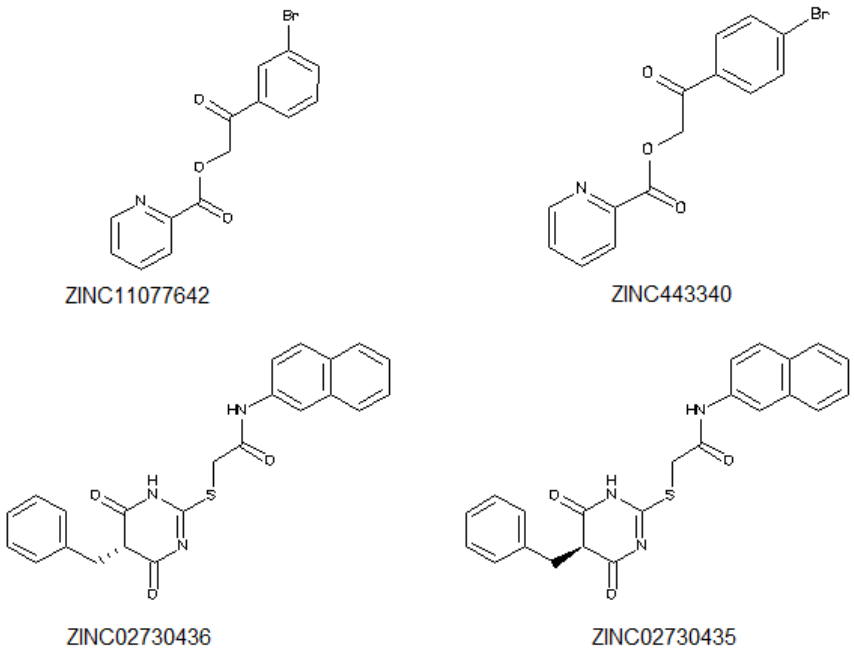


Figure 6.5: Final chosen molecules from virtual screening with top results for both techniques in two matrices.
Molecular representations retrieved from ZINC.

Analyzing the X and Y coordinates allowed the selection of molecules that had top predicted values for both methods in three matrices (6 molecules, figure 6.4), and two matrices(4 molecules, figure 6.5). A selection of 6 outliers was also performed with high random forest value and low kriging value or vice-versa (figure 6.6 and 6.7). Also, for further analysis, 2 more molecules were selected (figure 6.8). These molecules presented a top or good result for matrices 0 and 2 but a bellow midpoint prediction using matrix 1, these molecules should be study to understand which matrix is more prone to accuracy in this dataset for the continuation of this project.

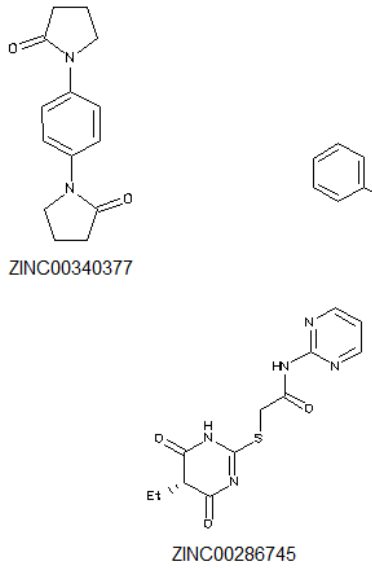


Figure 6.6: Final chosen molecules with high score in random forest and low in kriging.
Molecular representations retrieved from ZINC.

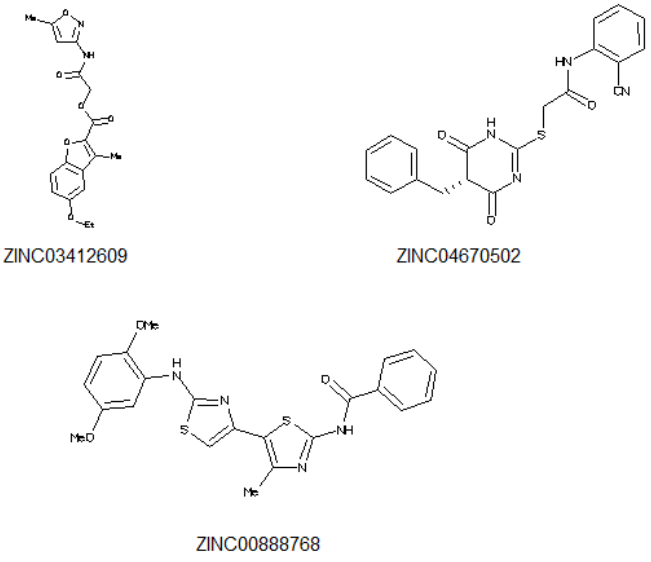
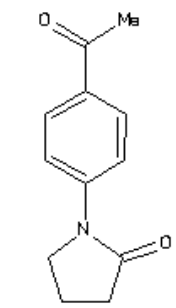
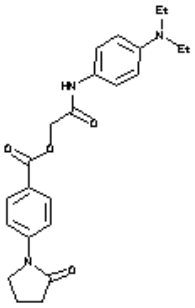


Figure 6.7: Final chosen molecules with high score in kriging and low in random forest.
Molecular representations retrieved from ZINC.



ZINC04207443



ZINC32945329

Figure 6.8: Two molecules considered outliers in matrix 1 but with good results in matrices 0 and 2.
Molecular representations retrieved from ZINC.

6.2.4 Confirmatory analysis

As in the anoctamins dataset, further information about each of the selected molecules was searched in the information repositories.

In Pubchem[1], ZINC00257699, ZINC11077642 and ZINC00443340 can possibly target anoctamin-1 in Mus Musculus (house rat). The molecule ZINC00257699 has a WIPO IPC ontology (World intellectual property organization International Patent Classification) of "Drugs for disorders of the respiratory system".

According to Surechembl [2] ZINC04207443 is an inhibitor of histone deacetylase. This inhibition capability is interesting because there are reports showing that reducing histone deacetylase 7 activity can restore function to misfolded CFTR (Hutt et al., 2010).

[1]Pubchem https://pubchem.ncbi.nlm.nih.gov/, last accessed 13th of March of 2017
[2]https://www.surechembl.org, last accessed 13 March

Chapter 7

Conclusions

In this dissertation, a new methodology combining a widely used QSAR technique and a new way to use machine learning in the molecular research environment to predict therapeutic capabilities of several molecules was presented. Cystic fibrosis is a disease were CFTR protein can have a large number of different mutations that can be treated and/or controlled by correcting the CFTR malformations. Another possible treatment course is to potentiate anoctamins action, as they are present in several different cell lines and produces Cl^- currents as CFTR.

Using an anoctamins potentiators and activators dataset as a starting point, a query was performed to find similar molecules in ZINC database. As a result of this query 71 molecules were selected for further prediction. In order to infer the therapeutic potentiator capability, two methods were applied. The first method is widely used in QSAR studies where each molecule is characterized using molecular descriptors and a machine learning method is used (in this case random forest) to learn from the initial dataset where the values to predict are known and produce forecasts using the molecular descriptors of new molecules. In this method, the molecular descriptors that seem to be more relevant to this particular potentiation are related to the molecule's solubility, molecular weight, and connectivity between atoms with a higher regard for heavy atoms. The second method uses kriging as a data mining method; this method requires data correlation and uses a topological representation of the data to infer values. In order to predict the therapeutic potentiator capability kriging receives the similarity between molecules calculated by NAMS program and represented them into

a two dimensional-plot. In this method, the selection of the atom substitution matrices from NAMS can give us some information about the molecules. In the anoctamins dataset the matrices chosen was number 0 and 1, the two more strict matrices available and therefore suggesting that a simple and similar change in one atom is sufficient to affect their ability to enhance the anoctamins function.

The outputs of both methods were then combined in a consensus plot to select the most promising molecules for further testing. In total 8 top molecules plus 2 outliers were selected based on the methods previously described. Once the molecules were selected a search in biological repositories was performed to review information about the molecules known targets or published bioassays. Several interesting findings arose in this search like a molecule that can target a calcium channel protein like anoctamins, another that can target Anoctamin-1 or even a molecule that appears to have some interaction like a chaperone.

Regarding the CFTR correctors dataset the same methodology from anoctamins could no be applied directly as the dataset had a mislabeling problem. Knowing that it would not be possible to perform new laboratory tests to the CFTR correctors, machine learning techniques were used to correct the mislabeling error. To guaranty that this corrected dataset could be used to infer characteristics of the molecules not present in the dataset, an evaluation of the prediction was performed using a pubchem bioassay with F508del-CFTR (this bioassay determines the activity or inactivity of molecules). After this correction, it was possible to apply the same methodology from the anoctamins dataset. In the CFTR dataset the matrices chosen were 0, 1 and 2, the two more strict matrices, and one more equilibrated matrix, which allow us to infer that in this case, very small differences in the atoms can affect the molecules capacity to rescue CFTR; but also that the matrix based on literature (matrix 2) have some effect on this capability, suggesting that some molecular behaviour is explained with this type of matrix and that these molecules are able to receive similar atoms and maintain the relation of enhancing CFTR function.

Based on this, 301 molecules were selected from ZINC in order to be evaluated by both methodologies. The important descriptors were also evaluated and the important characteristics seem to be similar to the anoctamins dataset

with solubility, weight, connectivity, and number of atoms as the most relevant characteristics.

Similarly to the anoctamins potentiators dataset several molecules were selected for further study: 10 top molecules with high predictions in both kriging and random forest methods, 3 outliers with high prediction in molecular descriptors and low in kriging and 3 other molecules with high prediction in kriging and low from molecular descriptors. Also, 2 extra molecules were also selected given that their predictions were good ou even top molecules in two matrices but below midpoint result in another matrix - these should be further tested to understand the differences in predicting using different matrices.

A confirmatory analysis considering the literature available was also performed showing a molecule that can possibly target Anoctamin-1 and another molecule that is an inhibitor of histone deacetylase (which can possibly restore function to misfolded CFTR).

In the CFTR dataset a 5-fold cross-validation was performed for each method: in the vector space the explained variance was 0.37, a precision of 0.61 and a recall of 0.12; in the metric branch the explained variance was 0.25 to 0.29 (depending on the matrix used).

Given the cross-validation and confirmatory analysis results, the tested methodology seems to have some predictive power, however it is necessary to test the molecules in the laboratory and improve the methodology with this results.

7.1 Future work

The immediate task is to test *in vitro* the selected molecules with a western blot lab design to evaluate the methods used in this project. It is also possible to perform another run in this workflow using this new lab results to get more precise predictions.

It is important to mention the several areas without molecules present in the metric space. Although some areas were interesting (giving the kriging predictions) they were not studied in this project because in the ZINC database was not possible to find molecules for those areas. One way to cover this issue is to design

specific molecules *in silico* to perform computational tests, and subsequently, order tailored molecules to test them. This specific molecular design can be done either in pharmacophore based models or *de novo* approach.

It can also be interesting to create a similar double space approach with a different methodology in the vector space, maintaining this direct Kriging approach since at the moment it is poorly explored in in-silico drug design methodology.

Appendix A

Listing of molecular descriptor in RDKIT

Table A.1: Table of molecular descriptors available in RDKit

SlogP	slogp_VSA1	MQN4
SMR	slogp_VSA2	MQN5
LabuteASA	slogp_VSA3	MQN6
TPSA	slogp_VSA4	MQN7
AMW	slogp_VSA5	MQN8
ExactMW	slogp_VSA6	MQN9
NumLipinskiHBA	slogp_VSA7	MQN10
NumLipinskiHBD	slogp_VSA8	MQN11
NumRotatableBonds	slogp_VSA9	MQN12
NumHBD	slogp_VSA10	MQN13
NumHBA	slogp_VSA11	MQN14
NumAmideBonds	slogp_VSA12	MQN15
NumHeteroAtoms	smr_VSA1	MQN16
NumHeavyAtoms	smr_VSA2	MQN17
NumAtoms	smr_VSA3	MQN18
NumRings	smr_VSA4	MQN19
NumAromaticRings	smr_VSA5	MQN20
NumSaturatedRings	smr_VSA6	MQN21
NumAliphaticRings	smr_VSA7	MQN22
NumAromaticHeterocycles	smr_VSA8	MQN23
NumSaturatedHeterocycles	smr_VSA9	MQN24
NumAliphaticHeterocycles	smr_VSA10	MQN25
NumAromaticCarbocycles	peoe_VSA1	MQN26
NumSaturatedCarbocycles	peoe_VSA2	MQN27
NumAliphaticCarbocycles	peoe_VSA3	MQN28
FractionCSP3	peoe_VSA4	MQN29
Chi0v	peoe_VSA5	MQN30
Chi1v	peoe_VSA6	MQN31
Chi2v	peoe_VSA7	MQN32
Chi3v	peoe_VSA8	MQN33
Chi4v	peoe_VSA9	MQN34
Chi1n	peoe_VSA10	MQN35
Chi2n	peoe_VSA11	MQN36
Chi3n	peoe_VSA12	MQN37
Chi4n	peoe_VSA13	MQN38
HallKierAlpha	peoe_VSA14	MQN39
kappa1	MQN1	MQN40
kappa2	MQN2	MQN41
kappa3	MQN3	MQN42

Appendix B

Workflow images

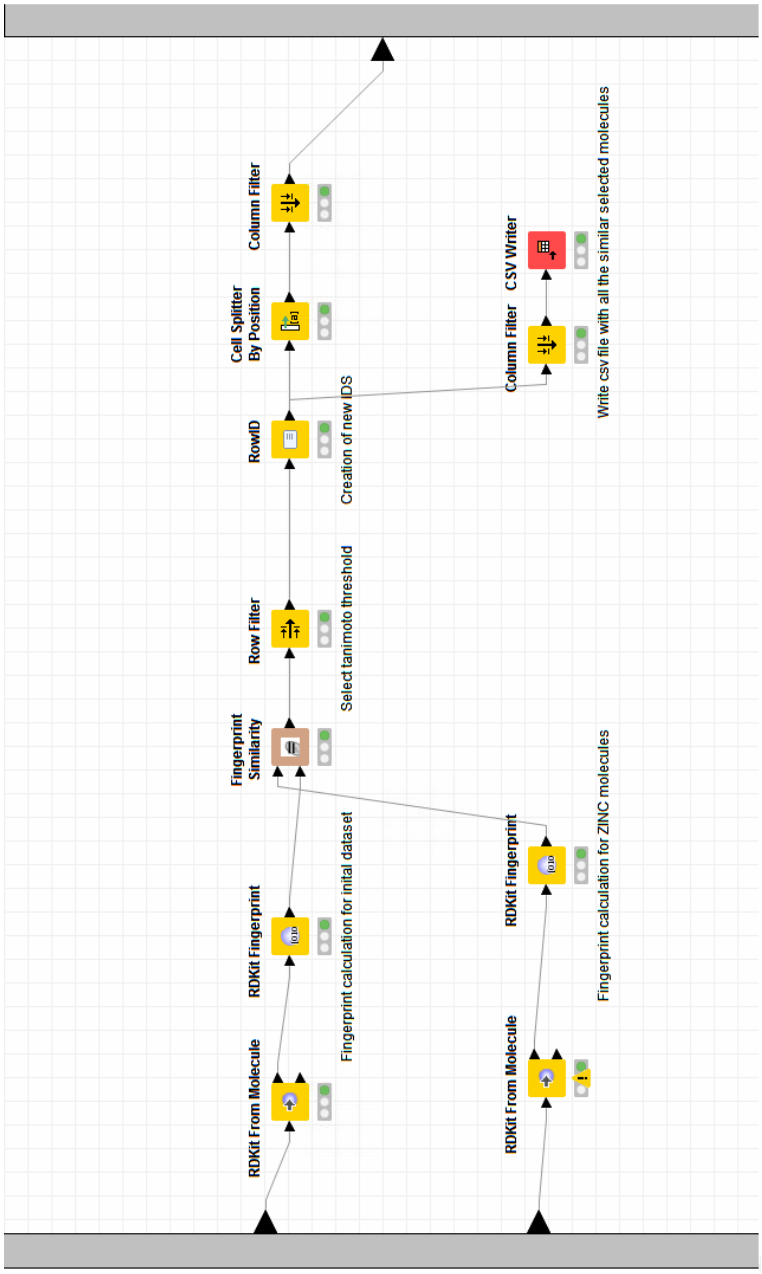


Figure B.1: Screenshot of the selection of ZINC molecules. Were it is possible to see the fingerprint calculation and further selection. RDKit from molecule node shows a yellow triangle given certain errors as duplication already mentioned about ZINC database.

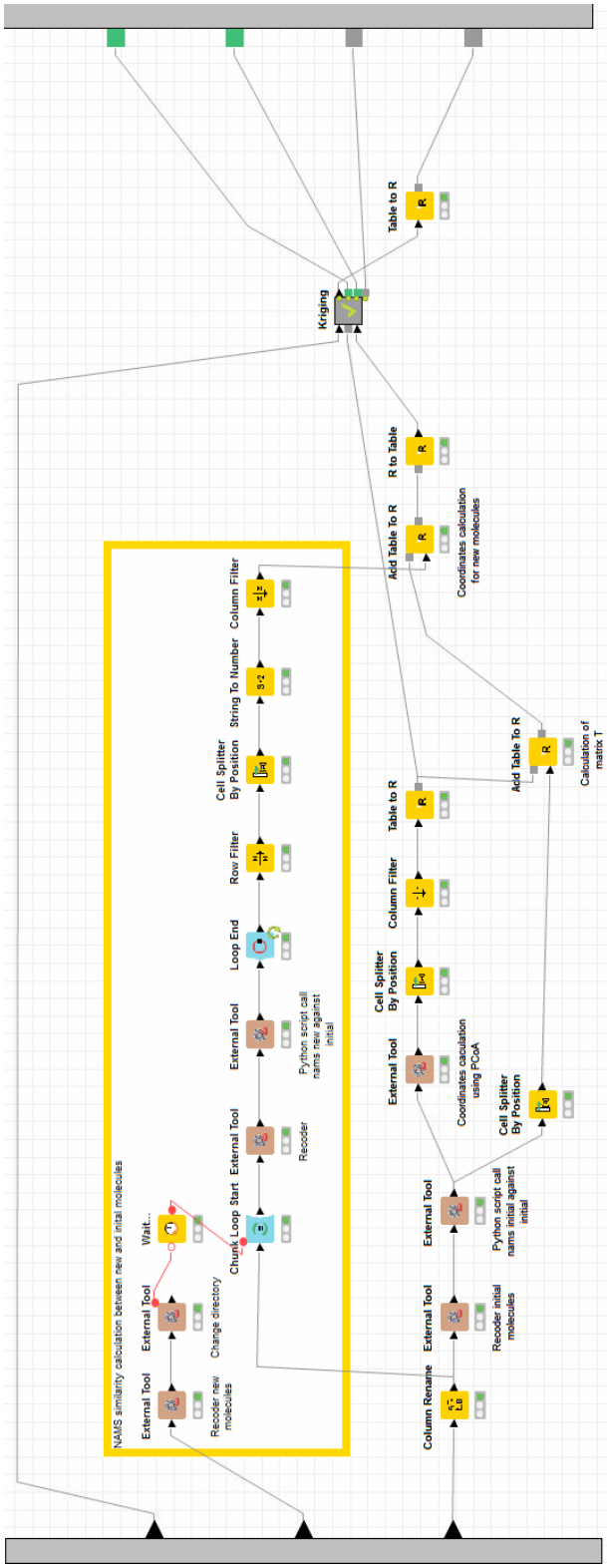


Figure B.2: Screenshot of the metric branch were all the steps can be seen, since NAMS similarity and coordinates calculation to Kriging prediction.

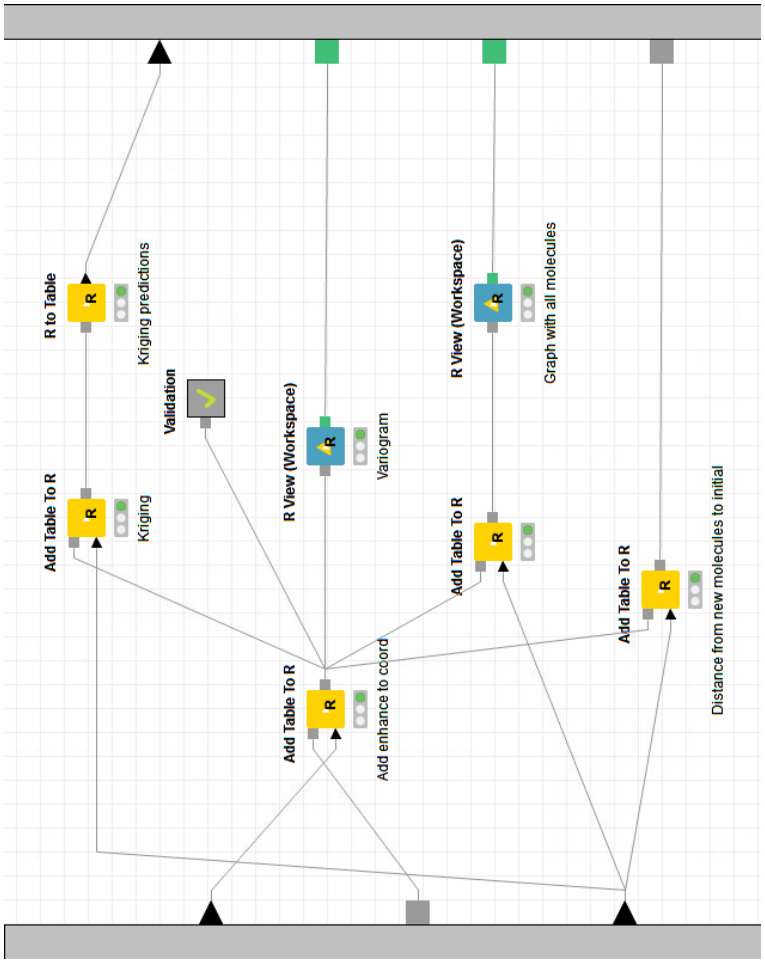


Figure B.3: Kriging metanode were it is possible to see all the calculation and graphs from this statistical method.

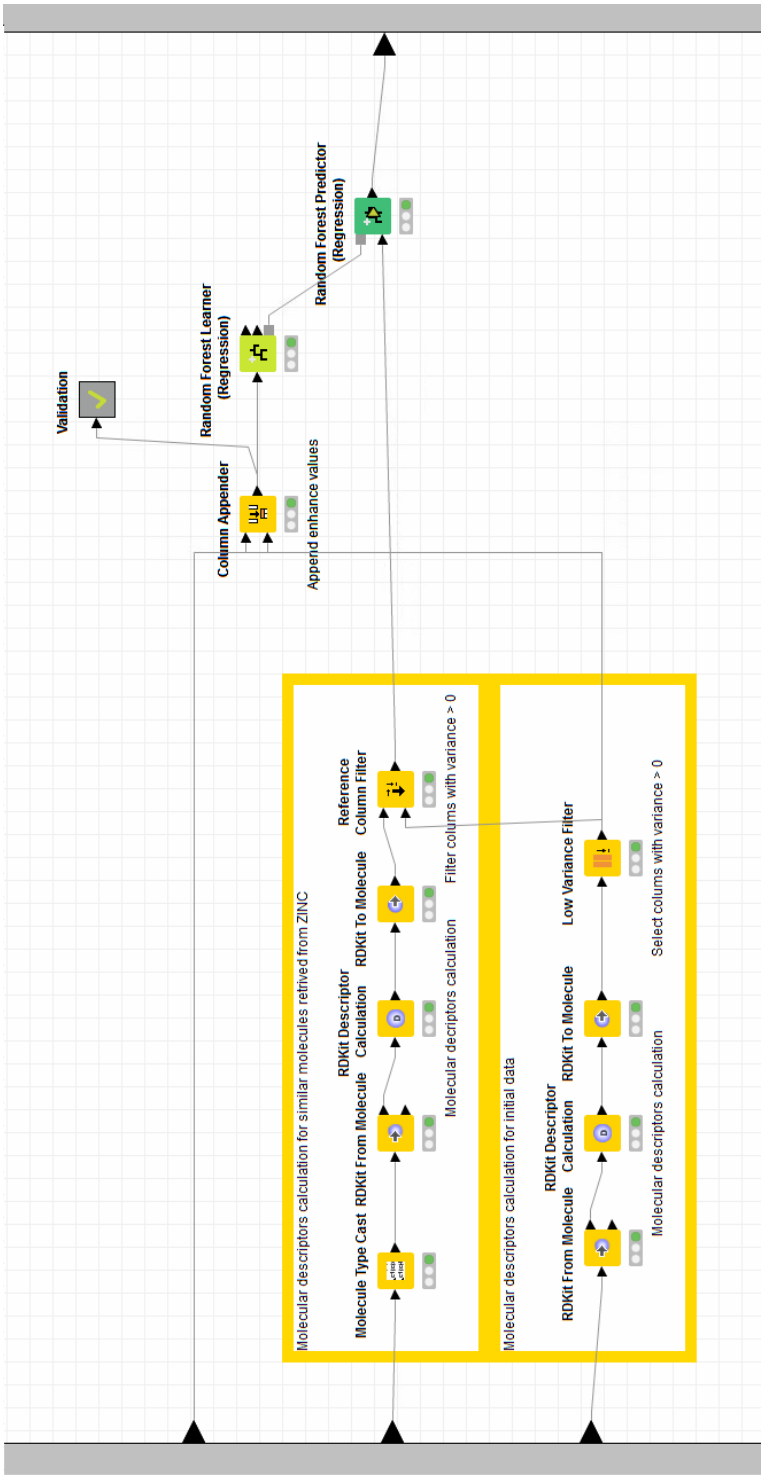


Figure B.4: Vectorial branch screenshot with the calculation of RDKit descriptors for both sets of molecules, Random forest learner and predictor node, as well as the validation metanode.

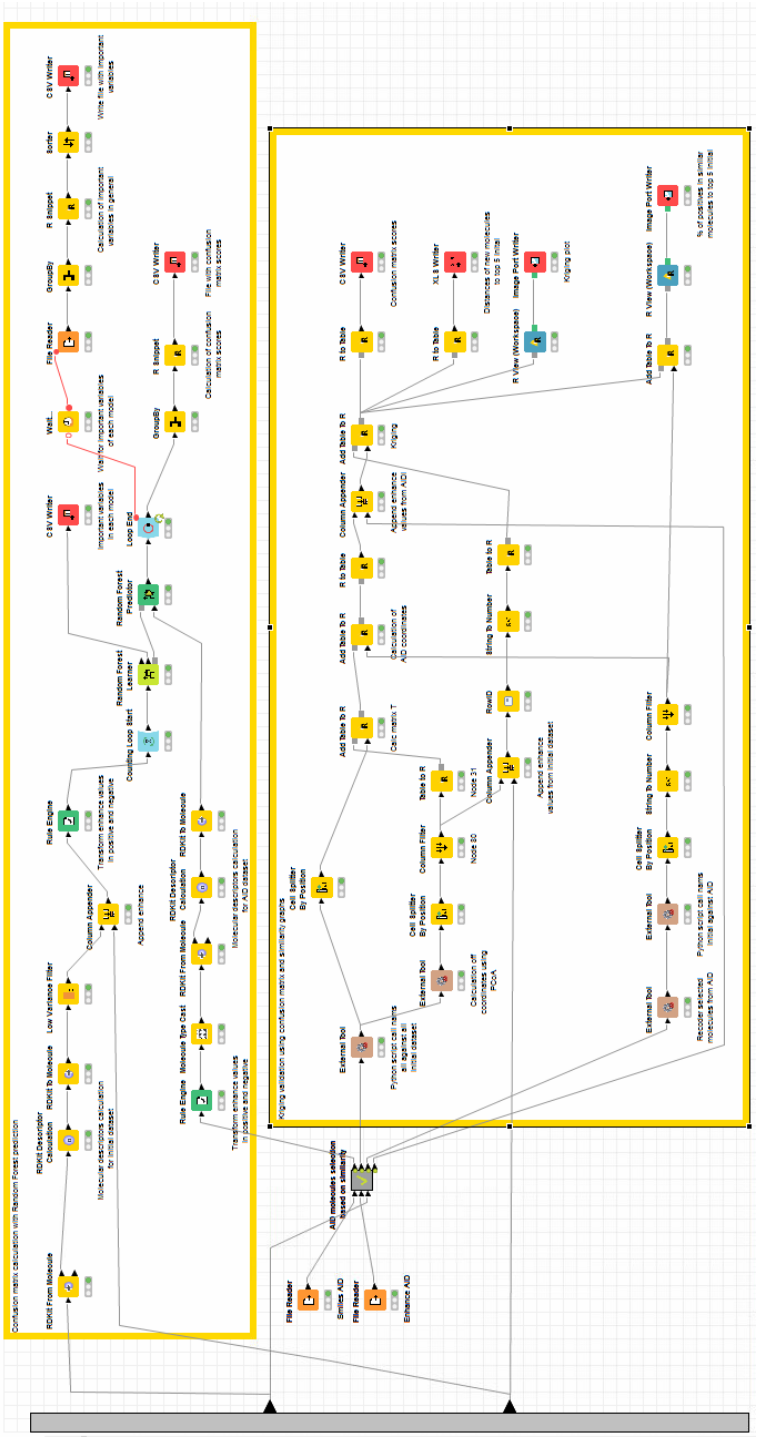


Figure B.5: Screenshot of reality check node were all the calculation using Random Forest and Kriging can be observed.

Appendix C

Selected molecules SMILES

C.1 Molecules SMILES from Anoctamins

Table C.1: SMILES of selected molecules from Anoctamins potentiators and activators.

Project ID	ZINC ID	SMILES
Molecules with high score in both techniques		
61	ZINC06414243	COc1cc(cc(c1OC)OC)C(=O)Nc1nc(cs1)c1ccccn1
29	ZINC09657902	COc1cc(cc(c1OC)OC)C(=O)NCC(=O)Nc1nc(cs1)c1ccccn1
46	ZINC06969746	COc1cc(cc(c1OC)OC)CC(=O)Nc1nc(cs1)c1ccccn1
56	ZINC01516590	COc1cc(cc(c1OC)OC)C(=O)Nc1nc(cs1)c1ccccc1
50	ZINC09014219	COc1cc(cc(c1OC)OC)C(=O)NC(=S)Nc1nc(cs1)c1ccccc1
57	ZINC01048113	COc1cc(cc(c1OC)OC)C(=O)NC(=S)Nc1nc(cs1)c1ccc(cc1)F
17	ZINC09360757	COc1cc(cc(c1OC)OC)C(=O)NC(=S)Nc1nc(cs1)c1ccc(cc1)Cl
45	ZINC01127308	Cc1ccc(cc1)c1csc(n1)NC(=S)NC(=O)c1cc(c(c(c1)OC)OC)OC
High value in random forest and low prediction from kriging		
62	ZINC08726656	Cc1ccc(cc1OC)C(=O)Nc1nc(cs1)c1ccccn1
High value in kriging and low prediction from random forest		
13	ZINC00929111	COc1ccc(cc1)c1csc(n1)N(CCc1ccccc1)C(=O)c1ccc(cc1)OC

C.2 Molecules SMILES from CFTR

Table C.2: SMILES of selected molecules from CFTR rescuers.

Project ID	ZINC ID	SMILES
Molecules with high score in both techniques in the majority of matrices		
208	ZINC00781515	c1ccc(cc1)C[C@H]1C(=O)NC(=NC1=O)SCC(=O)Nc1cccc2c1cccc2
207	ZINC00781516	c1ccc(cc1)C[C@@H]1C(=O)NC(=NC1=O)SCC(=O)Nc1cccc2c1cccc2
185	ZINC006362256	Cc1ccc(cc1)N(CCC#N)C(=O)COC(=O)c1ccc(cc1)N1CCCC1=O
9	ZINC00257699	c1cc(ccc1C(=O)[O-])N1CCCC1=O
5	ZINC06530212	c1ccc(cc1)c1[nH]nc(n1)SCC(=O)[O-]
187	ZINC06362258	Cc1cc(cc(c1)N(CCC#N)C(=O)COC(=O)c1ccc(cc1)N1CCCC1=O)C
290	ZINC11077642	c1ccnc(c1)C(=O)OCC(=O)c1cccc(c1)Br
126	ZINC00443340	c1ccnc(c1)C(=O)OCC(=O)c1ccc(cc1)Br
287	ZINC02730436	c1ccc(cc1)C[C@@H]1C(=O)NC(=NC1=O)SCC(=O)Nc1ccc2ccccc2c1
43	ZINC02730435	c1ccc(cc1)C[C@H]1C(=O)NC(=NC1=O)SCC(=O)Nc1ccc2ccccc2c1
High value in random forest and low prediction from kriging		
23	ZINC00340377	c1cc(ccc1N1CCCC1=O)N1CCCC1=O
287	ZINC02730436	c1ccc(cc1)C[C@@H]1C(=O)NC(=NC1=O)SCC(=O)Nc1ccc2ccccc2c1
49	ZINC00286745	CC[C@@H]1C(=O)NC(=NC1=O)SCC(=O)Nc1ncccn1
High value in kriging and low prediction from random forest		
28	ZINC03412609	CCOc1ccc2c(c1)c(c(o2)C(=O)OCC(=O)Nc1cc(on1)C)C
103	ZINC04670502	c1ccc(cc1)C[C@@H]1C(=O)NC(=NC1=O)SCC(=O)Nc1ccccc1C#N
192	ZINC00888768	Cc1c(sc(n1)NC(=O)c1ccccc1)c1csc(n1)Nc1cc(ccc1OC)OC
Interesting molecules with severe differences in results between matrices		
31	ZINC04207443	CC(=O)c1ccc(cc1)N1CCCC1=O
288	ZINC32945329	CCN(CC)c1ccc(cc1)NC(=O)COC(=O)c1ccc(cc1)N1CCCC1=O

References

Accurso, F. J., Rowe, S. M., Clancy, J. P., Boyle, M. P., Dunitz, J., Durie, P., Sagel, S., Hornick, D., Konstan, M., Donaldson, S., Moss, R., Pilewski, J., Rubenstein, R., Uluer, A., Aitken, M., Freedman, S., Rose, L., Mayer-Hamblett, N., Dong, Q., Zha, J., Stone, A., Olson, E., Ordoñez, C., Campbell, P., Ashlock, M., and Ramsey, B. (2010). Effect of vx-770 in persons with cystic fibrosis and the g551d-cftr mutation. *The New England journal of medicine*. 27

Alton, E. W., Armstrong, D. K., Ashby, D., Bayfield, K. J., Bilton, D., and Bloomfield, E. V. (2015). Repeated nebulisation of non-viral cftr gene therapy in patients with cystic fibrosis: a randomised, double-blind, placebo-controlled, phase 2b trial. *The Lancet Respiratory Medicine*. 26

Anderson, A. (2003). The process of structure-based drug design. *Chemistry and Biology*. 5, 15

Baroni, M., Costantino, G., Cruciani, G., Riganell, D., Valigi, R., and Clementi, S. (1993). Generating optimal linear pls estimations (golpe): An advanced chemometric tool for handling 3d-qsar problems. *Molecular informatics models - molecules - systems*. 14

Bonomo, S., Hansen, C. H., Petrunak, E. M., Scott, E. E., Styrishave, B., Jørgensen, F. S., and Olsena, L. (2016). Promising tools in prostate cancer research: Selective non-steroidal cytochrome p450 17a1 inhibitors. *Scientific Reports*. 34

Boucher, R. (2007). Cystic fibrosis: a disease of vulnerability to airway surface dehydration. *Trends in molecular medicine.* 2

Brodlie, M., Haq, I. J., Roberts, K., and Elborn, J. S. (2015). Target therapies to improve cftr function in cystic fibrosis. *Genome Medicine.* 26

Carrat, F. and Valleron, A.-J. (1991). Epidemiologic mapping using the "kriging" method: Application to an influenza-like epidemic in france. *American Journal of Epidemiology.* 25

Chavez, E., Navarro, G., Baeza-Yates, R., and Marroquín, J. L. (2001). Searching in metric spaces. *ACM Computing Surveys.* 25

Chen, Y.-P. P. and Chen, F. (2008). Identifying targets for drug discovery using bioinformatics. *Expert Opinion on Therapeutic Targets.* 5

Cheng, S. H., Gregory, R. J., Marshal, J., Paul, S., Souza, D. W., White, G. A., O'Riordan, C. R., and Smith, A. E. (1990). Defective intracellular transport and processing of cftr is the molecular basis of most cystic fibrosis. *Cell.* 3

Cheng, T., Li, Q., Zhou, Z., Wang, Y., and Bryant, S. H. (2012). Structure-based virtual screening for drug discovery: a problem-centric review. *American Association of Pharmaceutical Scientists Journal.* 15

Chiaw, P. K., Eckford, P., and Bear, C. (2011). Insights into the mechanisms underlying cftr channel activity, the molecular basis for cystic fibrosis and strategies for therapy. *Essays in biochemistry.* 26

Cholon, D. M., Quinney, N. L., Fulche, M. L., Esther Jr, C. R., Das, J., Dokholyan, N. V., Randell, S. H., Boucher, R. C., and Gentzsch, M. (2015). Potentiator ivacaftor abrogates pharmacological correction of f508 cftr in cystic fibrosis. *Science Translational Medicinel.* 3, 7, 27

Clark, D. (2008). What has virtual screening ever done for drug discovery? *Expert Opinion on Drug Discovery.* 11

Cramer, R. D., Patterson, D. E., and Bunce, J. D. (1988). Comparative molecular field analysis (comfa). 1. effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*. 14

Darnag, R., Mazouz, E. M., Schmitzer, A., Villemin, D., Jarid, A., and Cherqaoui, D. (2010). Support vector machines: Development of qsar models for predicting anti-hiv-1 activity of tibo derivatives. *European Journal of Medicinal Chemistry*. 19

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg*. 19

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of r&d. *Journal of Health Economics*. 2, 11

Fang, K.-T., Yin, H., and Liang, Y.-Z. (2004). New approach by kriging models to problems in qsar. *Journal of Chemical Information and Computer Sciences*. 25

Ferrari, T., Gini, G., and Benfenati, E. (2009). Support vector machines in the prediction of mutagenicity of chemical compounds. *Proceedings of The 28th North American Fuzzy Information Processing Society Annual Conference (NAFIPS2009)*. 19

Grzonkowska, M., Sosnowska, A., Barycki, M., Rybinska, A., and Puzyn, T. (2016). How the structure of ionic liquid affects its toxicity to vibrio fischeri? *Chemosphere*. 18

Hasegawa, K. and Funatsu, K. (2010). Non-linear modeling and chemical interpretation with aid of support vector machine and regression. *Current Computer-Aided Drug Design*. 7

Houston, D. and Walkinshaw, M. (2013). Consensus docking: improving the reliability of docking in a virtual screening context. *Journal of Chemical Information and Modeling*. 16

Huang, S.-Y., Grintera, S. Z., and Zou, X. (2010). Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Physical Chemistry Chemical Physics*. 15

Hudson, G. and Wackernagel, H. (1994). Mapping temperature using kriging with external drift: Theory and an example from scotland. *International Journal of Climatology*. 25

Hutt, D. M., Herman, D., Rodrigues, A. P. C., Noel, S., Pilewski, J. M., Matteson, J., Hoch, B., Kellner, W., Kelly, J. W., Schmidt, A., Thomas, P. J., Matsumura, Y., R Skach, W., Gentzsch, M., Riordan, J. R., Sorscher, E. J., Okiyoneda, T., Yates III, J. R., Lukacs, G. L., Frizzell, R. A., Manning, G., Gottesfeld, J. M., and Balch, W. E. (2010). Reduced histone deacetylase 7 activity restores function to misfolded cftr in cystic fibrosis. *Nature Chemical Biology*. 69

Huuskonen, J., Salo, M., and Taskinen, J. (1998). Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *Journal of Chemical Information and Computer Sciences*. 19

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. (2012). Zinc:a free toll to discover chemistry for biology. *Journal of Chemical Information and Modeling*. 33

Jalali-Heravi, M.and Parastar, F. (2000). Use of artificial neural networks in a qsar study of anti-hiv activity for a large group of hept derivatives. *Journal of Chemical Information and Modeling*. 19

Joy, S., Macalino, Y., and Gosu, V. (2015). Role of computer-aided drug design in modern drug discovery. *Archives of Pharmacal Research*. 12, 16

Kalyaanamoorthy, S. and Chen, Y.-P. P. (2011). Structure-based drug design to augment hit discovery. *Drug Discovery Today*. 5

Karelson, M. (2000). Molecular descriptors in qsar/qspr . *Jonh Wiley & Sons*. 17

Klabunde, T., Giegerich, C., and Evers, A. (2009). Sequence-derived three-dimensional pharmacophore models for g-protein-coupled receptors and their application in virtual screening. *Journal of Medicinal Chemistry*. 13

Kunzelmann, K., Tian, Y., Martins, J. R., Faria, D., Kongsuphol, P., Ousigsawat, J., Thevenod, F., Roussa, E., Rock, J., and Schreiber, R. (2011). Anoctamins. *Pflugers Archiv European Journal of Physiology*. 8

Lacroix, C., Fish, I., Torosyan, H., Parathaman, P., Irwin, J. J., Shoichet, B. K., and Angers, S. (2016). Identification of novel smoothened ligands using structure-based docking. *PLOS one*. 33

Lewis, H., Zhao, X., Wang, C., Sauder, J., Rooney, I., Noland, B., Lorimer, D., Kearins, M.C. and, C. K., Condon, B., Maloney, P., Guggino, W., Hunt, J., and Emtage, S. (2005). Impact of the deltaf508 mutation in first nucleotide-binding domain of human cystic fibrosis transmembrane conductance regulator on domain folding and structure. *The Journal of biological chemistry*. 26

Li, M., Wen, F., Zhao, S., Wang, P., Li, S., Zhang, Y., Zheng, N., and Wang, J. (2016). Exploring the molecular basis for binding of inhibitors by threonyl-trna synthetase from brucella abortus: A virtual screening study. *International Journal of Molecular Sciences*. 34

Lionta, E., Spyrou, G., Demetrios K., V., and Cournia, Z. (2014). Structure-based virtual screening for drug discovery: Principles, applications and recent advances. *Current Topics in Medicinal Chemistry*. 11

Martins, I. F., Teixeira, A. L., Pinheiro, L., and Falcao, A. O. (2012). A bayesian approach to in silico blood-brain barrier penetration. *Journal of Chemical Information and Modeling*. 11, 12, 17, 20

Meetei, P. A., Rathore, R. S., Prabhu, N. P., and Vindal, V. (2016). In silico screening for identification of novel beta-1,3-glucan synthase inhibitors using pharmacophore and 3d-qsar methodologies. *Springer Plus*. 34

Min, M., Xingjun, J., Xueding, W., Hao, Z., Weiqing, Y., Yuanyuan, Z., Changrong, P., Zicheng, L., Jing, Y., Quan, D., and Menglin, M. (2016). Synthesis and quantitative structure-activity relationships study for arylpropenamide derivatives as inhibitors of hepatitis b virus replication. *Chemical biology and drug design*. 18

Namkung, W., Yao, Z., Finkbeiner, W. E., and Verkman, A. S. (2011). Small-molecule activators of tmem16a, a calcium-activated chloride channel, stimulate ephitelial chloride secretion and intestinal contraction. *FASEB Journal*. 30

Niu, B., Lu, W.-c., Yang, S.-s., Cai, Y.-d., and Li, G.-z. (2007). Support vector machine for sar/qsar of phenethyl-amines. *Acta Pharmacologica Sinica*. 19

Paredes, R. and Navarro, G. (2006). Practical construction of k nearest neighbor graphs in metric spaces. *Proc. Fifth Workshop Efficient and Experimental Algorithms*. 7

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., and Schacht, A. L. (2010). How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*. 11

Pilewski, J., Cooke, J., Lekstrom-Himes, J., and Donaldson, S. (2015). Ws01.4 vx-661 in combination with ivacaftor in patients with cystic fibrosis and the f508del-cftr mutation. *Journal of Cystic Fibrosis*. 27

Polishchuk, P. G., Muratov, E. N., Artemenko, A. G., Kolumbin, O. G., Muratov, N. N., and Kuzmin, V. E. (2009). Application of random forest approach to qsar prediction of aquatic toxicity. *Journal of Chemical Information and Modeling*. 20

Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*. 22

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*. 19

Schneide, J., Coassolo, P., and Lavé, T. (1999). Combining in vitro and in vivo pharmacokinetic data for prediction of hepatic drug clearance in humans by artificial neural networks and multivariate statistical techniques. *Journal of Medicinal Chemistry*. 19

Singh, H., Singh, S., Singla, D., Agarwal, S. M., and Raghava, G. P. S. (2015). Qsar based model for discriminating egfr inhibitors and non-inhibitors using random forest. *Biology Direct*. 20

Smith, J. L., Halvorson, J. J., and Papendick, R. I. (1992). Using multiple-variable indicator kriging for evaluating soil quality. *Soil Science Society of America Journal*. 25

Stuper, A. J. and Jurs, P. C. (1976). Adapt: A computer system for automated data analysis using pattern recognition techniques. *Journal of Chemical Information and Modeling*. 14

Teixeira, A. L. and Falcao, A. O. (2013). Noncontiguous atom matching structural similarity function. *Journal of Chemical Information and Modelingl*. 21, 23, 24

Teixeira, A. L. and Falcao, A. O. (2014). Structural similarity based kriging for quantitative structure activity and property relationship modeling. *Journal of Chemical Information and Modeling*. 25

Teixeira, A. L., Leal, J. P., and Falcao, A. O. (2013). Random forests for feature selection in qspr models - an application for predicting standard enthalpy of formation of hydrocarbons. *Journal of Cheminformatics*. 17, 20

Terfloth, L. and Gasteiger, J. (2001). Neural networks and genetic algorithms in drug design. *Drug Discovery Today*. 19

Topliss, J. G. and Edwards, R. P. (1979). Chance factors in studies of quantitative structure-activity relationships. *Journal of Medicinal Chemistry*. 18

Tropsha, A. (2010). Best practices for qsar model development, validation, and exploitation. *Molecular Informatics*. 7, 13

Vapnik, V. (1998). Statistical learning theory. *Jonh Wiley & Sons*. 19

Verkman, A. S., Lukacs, G. L., and Galietta, L. J. (2006). Cftr chloride channel drug discovery - inhibitors as antidiarrheals and activators for therapy of cystic fibrosis. *Current Pharmaceutical Design*. 27

Vietin, S., Ermondi, G., Medana, C., Pedemonte, N., Galietta, L., and Caron, G. (2012). Ligand-based design, in silico adme-tox filtering, synthesis and biological evaluation to discover new soluble 1,4-dhp-based cftr activators. *European Journal of Medicinal Chemistry*. 27

Vuorinen, A. and Schuster, D. (2015). Methods for generating and applying pharmacophore models as virtual screening filters and for bioactivity profiling. *Methods*. 14

Wang, X. and Li, C. (2014a). Decoding f508del misfolding in cystic fibrosis. *Biomolecules*. 26

Wang, X. R. and Li, C. (2014b). Decoding f508del misfolding in cistyc fibrosis. *biomolecules*. 3

Wolber, G. and Langer, T. (2005). Ligandscout: 3-d pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *Journal of Chemical Information and Modeling*. 15

Yang, S.-Y. (2010). Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discovery Today*. 14

Yang, X., Liu, H., Yang, Q., Liu, J., Chen, J., and Shi, L. (2016). Predicting anti-androgenic activity of bisphenols using molecular docking and quantitative structure-activity relationships. *Chemosphere*. 18

Yao, X. J., Panaye, A., Doucet, J. P., Zhang, R. S., Chen, H. F., Liu, M. C., Hu, Z. D., and Fan, B. T. (2004). Comparative study of qsar/qspr correlations using support vector machines, radial basis function neural networks, and multiple linear regression. *Journal of Chemical Information and Computer Sciences*. 19

Yin, H., Li, R., Fang, K.-T., and Liang, Y.-Z. (2007). Empirical kriging models and their applicationsto qsar. *Journal of Chemometrics*. 25

Zielenski, J. and Tsui, L. (1995). Cystic fibrosis: Genotypic and phenotypic variations. *Annual Review of Genetics*. 2