UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

INFERENCE OF BINDING AFFINITY FROM NEURONAL RECEPTORS IN HUMANS

Mestrado em Bioinformática e Biologica Computacional
Especialização em Bionformática

Ana Sofia Grave de Jesus

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

2016

ACKNOWLEDGEMENT

I thank my teacher **Dr. André Falcão** for all the support, guidance and help with the project and thesis.

I wish to express my sincere gratitude to my family, my boyfriend and his family and my friends for all the support and encouragement.

Abstract

Only some compounds (e.g. ligands) act as neurotransmitters in the brain, binding to specific neuroreceptors. Understanding the criteria behind why a ligand binds to a particular target in the brain can help design drugs which are more effective. With the help of data-mining techniques, quantitative structure–activity/propriety relationship (QSAR/QSPR(Q (SAR)) models and machine learning methods, a supervised model can be built which can predict binding affinities for any molecule, provided sufficient experimental data is available.

Models which can predict binding affinities for specific neuroreceptors were built using three machine learning methods (Random Forests, Support Vector Machines and Least Absolute Shrinkage and Selection Operator) and two sets of molecular descriptors from different chemical toolboxes (Open Babel and CDK). Experimental data was collected to create the database and curated by removing inconsistencies and duplicates. The final dataset had 43901 binding affinity values for 53 human neuroreceptors. In the model building phase, 75% of the dataset was used for training and 25% for validation. The modelling consisted of choosing the most important variables (descriptors) for each neuroreceptor and validating using statistical measures.

Random Forests and SVM were the best methods. Random Forests was used to select the most important variables and SVM for the statistical measure. The value of root mean squared error (RMSE) was below 0.214, more than half of the receptors had the percentage of variance explained (PVE) above 50% and Pearson's correlation coefficient was above 0.50, confirming the model had a good fit. Small dataset (below 112 entries) resulted in some models having poor results. RMSE values from validation and modelling parts were similar for the best model resulting in a good therefore can predict the strength of binding between neuroreceptor and neurotransmitter. The values of RMSE for the best models were between 0.087 and 0.201 where the PVE is above 50% and correlation above 0.50.

Some molecular descriptors were selected frequently; 46 descriptors appeared in more than 20 neuroreceptors, however only 6 descriptors appeared in all neuroreceptors. The same descriptors are used to identify the same family of neuroreceptors.

Keywords: Cheminformatics, machine learning, QSAR/QSPR, neuroreceptors, binding affinity

Resumo

É importante perceber o critério que determina a ligação entre uma molécula e um recetor específico, em particular no cérebro, onde só alguns compostos atuam como neurotransmissores e ligam-se a neurorecetores especifícos. Os neurotransmissores, dependem da sua estrutura para estabelecerem uma ligação com os neurorecetores. Essa ligação pode ser medida através de valores de binding affinity. É possivel, com ajuda de técnicas de data-mining, métodos de machine learning e de relação quantitativa estrutura-propriedade/atividade (QSAR/QSPR), construir um modelo que consiga prever esses valores de binding affinity, desde que tenhamos toda a informação necessária (propriedades/estrutura da molécula e do neurorecetor e valores de binding affinity). Métodos de QSAR/QSPR foram desenvolvidos para compreender as propriedades das moléculas, prever a sua estrutura, e a relação entre os descritores moleculares da sua estrutura com as suas propriedades.

De modo a prever valores de binding affinity entre neurotransmissores e neurorecetores, neste trabalho foi criada uma base de dados , com seis dimensões referentes a espécies de animais (dimspecie), a referências bibliográficas (dimref) , a diferentes fontes de dados utilizadas para fazer a base de dados (dimsource), a recetores (dimrec) , a moléculas que vão ligar aos recetores (dimlig) e à localização do recetor (dimlocal).Os valores binding affinity foram expressos em pKi. A base de dados foi curada, os duplicados foram removidos, assim como e valores inconsistentes, como por exemplo, todos as entradas sem estrutura do composto (SMILES). A base de dados tinha 198169 valores de binding affinity.

Após a construção da base de dados, procedeu-se à escolha específica de dados para construção do modelo QSAR/QSPR, de modo a ter um bom conjunto de dados. Os critérios de escolha, foram os seguintes: os recetores tinham que estar localizados no cérebro (neurorecetores humano), e tinham que se ligar a mais de 50 ligandos. No final, o conjunto de dados tinha 43901 valores de binding affinity entre 0 e 1 para 53 neurorecetores. O conjunto de dados obtido foi dividido em 75% para o conjunto de treino e 25% para conjunto de teste, isto de forma aleatória para cada neurorecetor.

Os descritores moleculares para os compostos do conjunto de dados foram desenvolvidos com a ajuda de duas ferramentas OpenBabel e CDK que foram desenvolvidas para perceber a linguagem dos dados químicos. Essas ferramentas permitem procurar, converter, analisar e armazenar dados de modelação molecular e as caraterísticas

bioquímicas. Uma molécula pode ser codificada através de fingerprints que possibilita a determinação da similaridade entre duas moléculas. Existem mais de 5000 descritores, como por exemplo, a massa molecular, o número de átomos, entre outros.

Para a construção do modelo, foram usados três métodos combinados de machine learning (Random Forests, Support Vector Machines (SVM) e Least Absolute Shrinkage and Selection Operator (LASSO)), na escolha das variáveis mais importantes, ou seja, as que descrevem melhor a ligação entre os ligandos e os neurorecetores. Os métodos usados foram Random Forests e LASSO e depois posteriormente procedeu-se à validação com obtenção de valores de RMSE , do coeficiente de correlação de Pearson e da percentagem da variação explicada (PVE) com a ajuda do SVM e LASSO.

O método de SVM reconhece padrões e baseia-se em encontrar, nos dados , instâncias que são capazes de maximizar a separação entre dois pontos.

O método Random Forests, reduz a variância da função da predição estimada, usando para esse feito, árvores de regressão e faz média do resultado. O número de árvores usadas foram 500,enquanto LASSO é um método de regressão que envolve uma penalização do tamanho absoluto dos coeficientes de regressão, em que alguns casos serão zero.

Em relação à escolha do conjunto de dados, foi usado o método de cross-validation, em que cada combinação de métodos foram corridos cinco vezes e por cada corrida o conjunto de treino foi divido em 75%, para o conjunto de treino e 25% para o conjunto de teste de forma aleatória, para cada neurorecetor.

Os resultados obtidos demonstraram que em todos os métodos, com poucas variáveis, os valores de RMSE são elevados, mas chega a um patamar em que quantas mais variáveis são usadas, maior é o valor de RMSE. No entanto, esses valores variam consoante o recetor, pois existem recetores com um baixo valor de RMSE com 4 variáveis, no entanto, temos outros que são necessários 100 variáveis para se obter um valor baixo de RMSE. O número de variáveis mais importantes para o modelo varia entre 4 e 100.

A melhor combinação de métodos em que foram obtidos os melhores resultados para os modelos foram o Random Forests e SVM, apesar de haver três modelos que obtiveram melhores resultados com outro método (LASSO e SVM) . Para validação do modelo foi usado o conjunto de teste que tem 25% dos dados do conjunto de dados iniciais.

O RMSE é um bom indicador da qualidade do modelo, mede a distância entre os dados observados e os dados que fazem o modelo. O maior valor de RMSE para o conjunto de treino foi de 0.214.

Em geral estamos na presença de bons modelos, no entanto, alguns modelos apresentaram resultados fracos, em que os valores de RMSE são elevados, os valores de PVE e de correlação são baixos e os resultados entre os dados de treino e os dados de testes são muito diferentes, isso acontece na maior partes das vezes quando o número de dados no conjunto de dados é inferior a 112. Para ter um bom modelo, o conjunto de dados precisa de ter mais de 112 entradas, ou seja, é preciso mais de 112 valores de binding affinity para poder construir um bom modelo para esse neurorecetor de modo a prever corretamente valores de binding affinity com outros neurotransmissores .

Em relação à correlação que nos indica a força e direção da relação linear entre variáveis, o valor menor é 0, o que indica um fraca correlação, mas em média os valores da correlação são acima de 0.50, o que indica uma forte correlação.

A outra medida usada para medir a qualidade do modelo obtido foi a percentagem de variação explicada (PVE) , que em geral está acima do 50%.

Os resultados do conjunto de teste foram próximos aos obtidos com o conjunto de treino. Como por exemplo, no caso do modelo para o transportador de seratonina (5-HT transporter), em que o valor de RMSE é 0.216 e a percentagem de variação explicada de 51.1 e para a correlação 0.711, que em comparação com o conjunto de treino que foram 0.196, 57.3 e 0.759 respetivamente são próximos.

Os melhores modelos têm os valores de RMSE entre 0.087 e 0.201, em que o PVE está acima de 50% e a correlação está acima de 0.50.

Relativamente à selecão dos descritores moleculares mais importantes para a construção do modelo, verificou-se que cerca de 46 descritores moleculares foram escolhidos em pelo menos 20 recetores, isso demonstra que esses descritores são necessários para construir um bom modelo. No entanto, constatou-se que 6 descritores foram selecionados em todos os recetores, a massa molecular, a refratividade molar, o logaritmo do coeficiente partição da água/octanol, o número de ligações simples e aromáticas, demonstrando que estes descritores são os mais importantes para termos um bom modelo. Verificou-se também que os mesmos descritores servem para identificar as mesmas famílias de recetores.

Futuramente este modelo pode ser usado na fase inicial da descoberta e produção de novas drogas, pois este modelo consegue verificar a viabilidade dessa droga antes de se proceder a ensaio experimental , através da previsão de valores de binding affinity entre a droga e o seu alvo.

O desenvolvimento de uma aplicação online onde se coloca o composto e essa aplicação verifica se o composto se vai ligar a algum neurorecetor.

Palavras-chave: Cheminformatics, machine learning, modelos QSAR/QSPR, neurorecetores, binding affinity

Content

List of Tables

List of Figures

Glossary

QSAR/QSPR – quantitative structure-property/activity relationship

RMSE – Root Mean Squared Error

MSE – mean squared error

CNS – central nervous system

BBB – brain blood barrier

LASSO – Least Absolute Shrinkage and Selection Operator

SVM – Support Vector Machine

RF – Random Forests

Ki _value – Binding affinity value

PVE – percentage of variance explained

r – Pearson's correlation coefficient

1. Introduction

1.1. Overview

A drug is a chemical substance which interacts with a receptor and through this interaction, has biological effects in the organism. For the drug discovery industry, it is important to understand the criteria that determines why a molecule binds to a particular target. The efficacy and overall safety of drugs are determined by its activity profile towards many biological targets therefore it is necessary to design and predict drugs with a specific multi-target behaviour [10, 15, 30]. It is better to use computational methods to correctly predict if a ligand will bind to a receptor or has a structural affinity with it, instead of doing several experimental assays that take time and are expensive [10, 15, 30, 34, 35, 55-58]. Computational models to predict binding affinities are useful because they can rationalize a large number of experimental observations leading to saving time and costs [25].

Computer-based (in silico) methods are being developed to help select the best possible drug candidates without eliminating any of the relevant ones. The computational tools are mainly used for the conformational analyses of molecular structure to characterize the interactions between drugs and their targets and to assess and optimize the drug activity using quantitative structure-property/activity relationships (QSARs/QSPRs) methods. In drug design, the QSAR/QSPR methods are used for the estimation of physicochemical properties, biological effects and to understand the physicochemical features governing a biological response which makes QSAR a low-cost tool for the selection and optimization of drug discovery and development [10, 15, 35].

1.2. Definition of the problem

The main focus of this work is to predict the binding affinities between several molecules and receptors in the human brain. The objective is to use machine learning techniques (Random Forests, Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine (SVM)) and QSAR/QSPR methods to build the best model which can predict descriptors that are most closely related to the property of interest (in this case, binding affinity values) [10,30,34].

1.3. Work outline

To predict the binding affinity, it is necessary to build a model - this requires a robust, unbiased and sufficiently large training set. A good training set is comprised of a database with as high-quality records as possible, especially in terms of reliability, consistency and full access to the complete description of that data [10, 17].

The QSAR/QSPR methods are used to model the property or activity of a set of molecules. The modelling allows the development of a linear or non-linear model to either predict the activity of a set of molecules, or to classify a set of molecules based on a set of molecular descriptors [18, 23]. Models are only useful if they are predictive therefore they need to be validated so they are able to produce good results even when using unknown data. The process of validation is divided into two parts- the training set and the test set. The training set is used to build the model [35] and the test set is used to validate the model using unknown data.

Cross-validation is an effective approach to randomly divide the dataset. In an n-fold cross-validation, the data is distributed, either randomly or in a stratified way, into n-separate folds, with one fold being the initial test set.

Until now there has not been a single, best, machine learning algorithm that can resolve all problems. It is necessary to understand the problem and use the best methods to resolve it, sometimes it is better to combine two methods. The performance of a method depends on the size of the dataset, the nature and internal correlation of the description set available, the relevance of the non-local data, amongst several other factors. [35].

The statistical measures are used to give goodness-of-fit prediction values for the QSAR/QSPR model and to validate the model. The most common measures are correlation, the determination of coefficients (R^2) and standard error of the estimates, like root mean squared error (RMSE) [26].

1.4. Scheduling

The diagram below shows the time taken for each task necessary for this work:

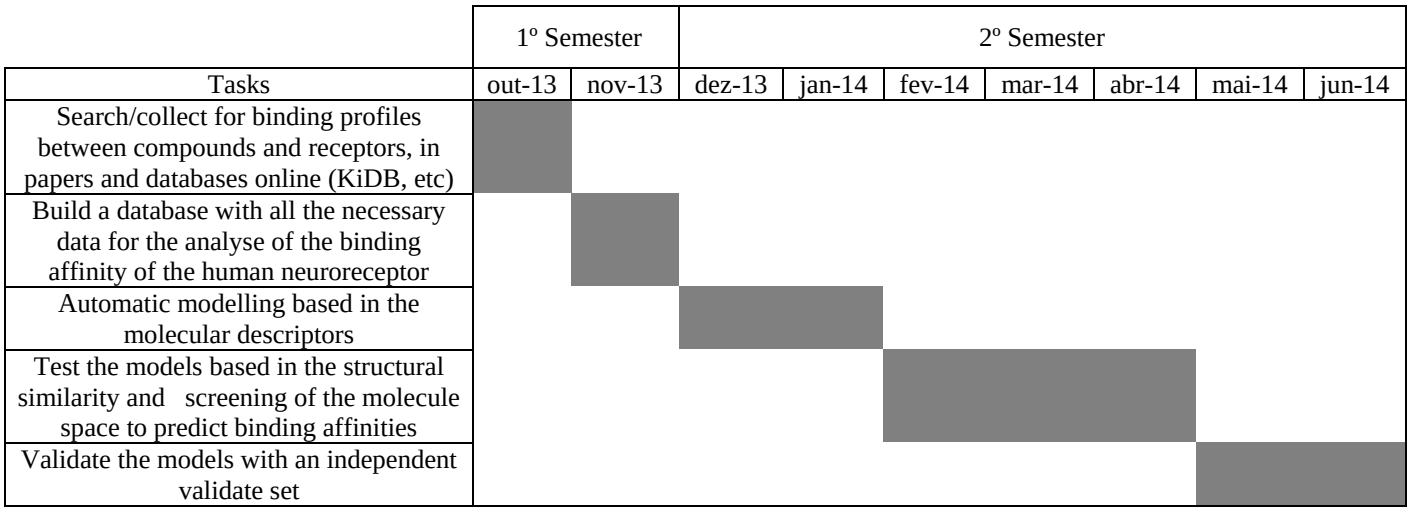

Figure 1 – Gantt chart of each task and the corresponding monthly schedule.

The tasks for this work were divided into five parts: data search, build database, data consolidation, model training and test set (Figure 1). First data was gathered for binding profiles between compounds and receptors in papers and online databases (KiDB, DrugBank, etc). After gathering the data and selecting the best sources for relevant datasets, a database was built to analyse and predict the binding affinity between the ligands and the neuroreceptors in humans using Python and MySQL.

The data was cleaned and consolidated for the automatic modelling part, where the automatic learning models were adjusted for the molecular descriptors.

The different models were tested based on structural similarity and screening of the molecular space to predict the binding affinities. This was a critical, the most complex and lengthy stage because the automatic learning models had to be adjusted to produce consistent results in order to create the best model to predict binding affinities. This stage took three months to complete (Figure 1).

Before using the whole dataset, the machine learning methods were tested with only four receptors that had the biggest datasets, in order to check if it was possible to build good models.

Adjustments needed to be made to some of the models and some took longer than others. In the end, the best model was tested and selected with an independent dataset using an Open Source computational tool.

2. Concepts and Related work

2.1. Neurotransmitters and their receptors

Neuroreceptors are glycoproteins in the nerve cellular membrane or target cell in which a transmitter compound can interact producing a biological response [1, 3]. A neurotransmitter is defined through four criteria. Firstly, a neurotransmitter needs to be present inside the nerve cell (the enzymes required for its synthesis). Secondly, a neurotransmitter needs to be present on the synaptic level for the enzymes do its inactivation. Thirdly, a neurotransmitter needs to produce the same response when it is experimentally placed on the target. Lastly, it is necessary for the neurotransmitter to be in the synaptic space for the time of spontaneous activation or the electric stimulation of the nerve cell [2, 4].

Most of these neurotransmitters are amino acids, such as glycine and glutamate, or their decarboxylation products or derivatives, including γ-Aminobutyric acid (GABA), serotonin and others. Many of these compounds are also hormonally active, but they are excluded from the brain by the blood-brain barrier. Although many neurotransmitters, such Acetylcholine are excitatory, some are inhibitory. The latter stimulate the opening of anion (Cl-) channels, thereby causing the postsynaptic membrane to become hyperpolarized so that it must be more highly depolarized to trigger an outgoing action potential. There are also some polypeptide neurotransmitters, many of which are also polypeptide hormones, that elicit complex behaviour patterns [1, 2, 5].

2.1.1. Types of neuroreceptors

Neuroreceptors are different in their structure and in their response to the bond between them and a neurotransmitter. They are classified as receptor coupled to the G-protein (GPCRs) or receptor connected with ions channels (LGICs). The GPCRs superfamily is divided into six classes and LGICs are classified in three superfamilies. Within each superfamily/class, the neuroreceptors are similar to each other. It is expected that their functions will be similar on a molecular level. However, molecules with similar structure and function can have different degrees of selectivity depending on where they are expressed (e.g. tissue or organ). Molecules expressed in the same location have a similar degree of selectivity [2, 3,16].

The degree of functional constraint at the molecular level is similar within each family, but the degree of selective constraint is distributed over a wide range, independent of molecular similarity. This happens due to the difference in the neural function in which the neuroreceptor is involved [7, 10].

The GPCRs are the central focus in pharmacological research. Below (Table I) is the type and subtype of receptors:

Receptor	**Type**	**Subtype**	**Example of agonist therapy**	**Example of antagonist therapy**
Cholinergic	Nicotinic	Nicotinic	Stimulation of the tract Gl (M1) Glaucoma	Neuromuscular blocker and muscular relaxant (N) Peptic ulcer (M)
	Muscarinic	M1-M5		
Adrenergic (adrenoreceptors)	α,α2, β	α1 A, α1 B, α2A, α1 D, α2C, β1, β2, β3	Anti-asthmatic (β2)	Blockers β(β1)
Dopamine		D1, D2, D3, D4, D5	Parkinson disease	Anti-depressing (D2,D3)
Histamine		H1-H3	Vasodilation	Treatment to allergies (H1) Anti-ulcer(H2)
Opioids		δ,μ,κ, ORL1	Analgesics(κ)	Morphine antipode (overdose)
Serotonin (5-Hidroxi-triptamine	5HT1-5HT7	5HT1A, 5HT1B, 5HT1D-1F, 5HT2A-2C, 5HT5A, 5HT5B,	Anti-migraine (5-HT1D) Stimulation of the tract Gl (5-HT4)	Anti-emetic (5HT3)

Table II - Example of receptors in terms of their type, subtype and example of agonist and antagonist therapy [3].

2.1.2. Pharmacon as a neurotransmitter

A pharmacon, meaning a biologically active substance, can be an antagonist and block the receptor of the messengers or it can be an agonist and imitate the messenger. Alternatively, agonist pharmacon can act like an inverse agonist pharmacon that binds to the same receptor as an agonist but induces a pharmacological response opposite to that agonist. [3].

There are a lot of new drugs which have different types of targets – the ion channels and other potential molecular targets [5], the neuronal gap junctions [13] and aquaporin channels [12].

The measure of the binding affinity (Ki value) is very important in terms of molecule/receptor specificity because it indicates the tendency and strength of the binding between the molecule and receptor, and characterizes the interactions [10, 30, 32]. To predict the binding affinity between small-molecule ligands and receptors, it is necessary to use knowledge, regression and first principle based methods. Knowledge-based methods are founded in experimental assays to determine the protein-ligand complexes through statistical approaches, to make rules about the interaction of geometric preferences. The regression-based methods are statistical processes of estimating the relationships among variables where many modelling and analyses techniques are applied to understand the relationship between the dependent variable and the independent variables. First principle-based methods are the foundation of the problem. They are not based on any experimental assay - it is the intuitive knowledge of the problem [5, 8].

2.2. Molecular Descriptors

To create the best model, it is necessary to have the best set of compounds, which enables the best way to understand the relationship between the activity and structure. A minimum number of compounds (between five and ten) are needed to create a QSAR/QSPR model descriptor.

Many in silico models were developed to help build good models that can predict the properties and activities of the molecules using different types of information, like physic-chemical properties, pharmacological effects and many others. Hajjo et al. [36] developed and validated binary classification QSAR/QSPR models that can predict the potential of 5-hydroxytryptamine 2B (5-HT2B) serotonin actives that can cause valvular heart disease. In another study Luo, Man et al. [39] developed binary QSAR/QSPR models of 5-hydroxytryptamine 1A (5-HT1A) serotonin in terms of the binding activity using data from PDSP Ki database. Yugandhar K., and Gromiha, M., Michael [42] developed a model that can predict the binding affinity of protein–protein complexes using machine learning approaches. They analysed several machine learning algorithms to discriminate protein-protein complexes into high and low-affinity groups based on their Kd values. In a different study Zhang L, et al [43] developed a method to design a new anti-malarial compounds through the modelling of a database defined as active or inactive towards *P.falciparum*.

Different statistical approaches help to interpret the QSAR/QSPR model. It is necessary to assure the statistical fit of the model and to pay attention to ubiquitous correlation

coefficient when using different datasets [10, 17]. To validate a QSAR/QSPR model it is better to use an external test set – a group of molecules that are not in the original dataset. One way of doing this is to divide the initial dataset into training and test sets [10, 17].

For a long time, science has tried to capture and convert, through theoretical pathways, all the information in the structure of a molecule into numbers as a way to understand the quantitative relationships between structures and properties, biological activities, or other experimental properties. A molecular description is a result of a logical and mathematical procedure where the chemical information is transformed into a symbolic representation of a molecule, useful numbers or the result of some standardized experiment [10,11].

Nowadays there are more than 5000 descriptors obtained from different ways that are computable by using dedicated software of chemical structure. The QSAR approach was divided into different parts. First, it is necessary to understand the molecular structure then define the molecular descriptors and the chemoinformatic tools [10, 17].

The molecular structure is represented through theoretical molecular descriptors and the relationships with experimental properties of molecules [10, 17].

The molecular descriptors are numerical indexes that encode information related to the structure, which can be experimental physic-chemical properties of molecules and theoretical indexes calculated by mathematical formulas or computational algorithms [10, 17]. To create a molecular descriptor (Figure 2), it is necessary to apply many principles from different theories, such as quantum-chemistry, information theory, organic chemistry and graph theory. These principles help to build the model through the different properties of a compound [10, 17]. Figure 2 (below) illustrates the relationship among molecular structure, molecular descriptors, chemoinformatics and QSAR/QSPR modelling [10, 17].

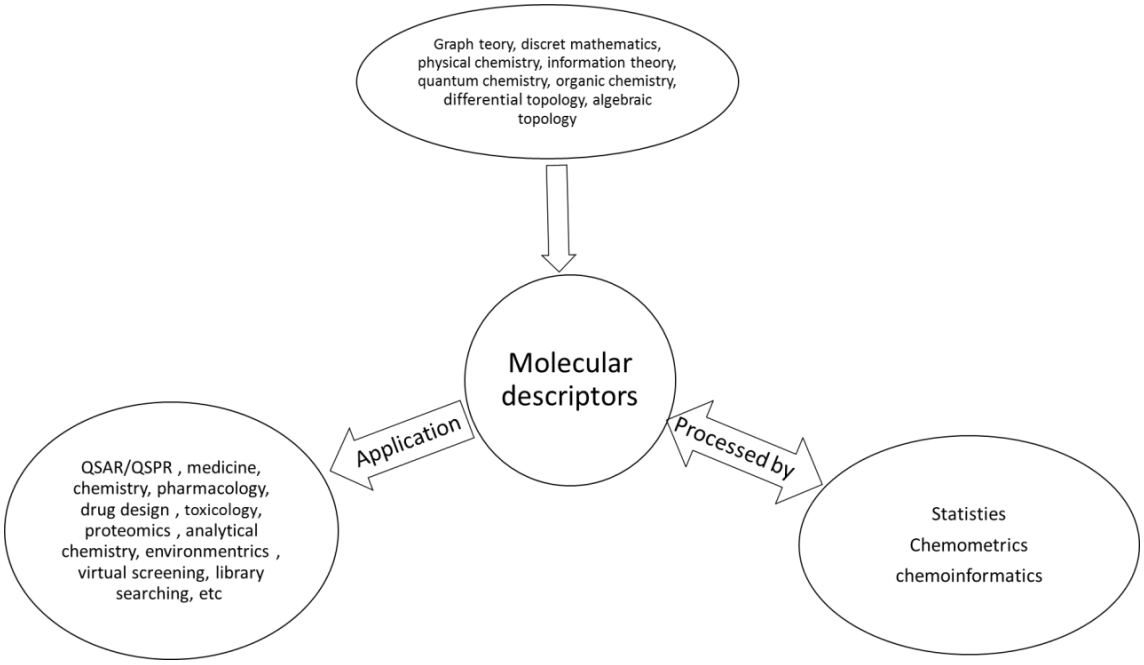

Figure 2 – Scheme showing the relationship among molecular structure, molecular descriptors, chemoinformatics and QSAR/QSPR modelling [10, 17].

The molecular descriptors are the basic tools used to transform chemical information into a numerical value capable of being used through informatics procedures. The most important variables are used for the modelling [10, 17].

The molecular descriptors are divided into experimental measures (log P, molar refractivity, dipole moment etc.) and theoretical molecular descriptors [10, 17]. There are simple molecular descriptors, like the number of different types of atoms in a molecule or the ones that are created by applying algorithms to a topological representation. These type of descriptors are called topological or 2D-descriptors [11, 14]. The descriptors that are created from the spatial (x, y, z) coordinates, are called geometrical, or 3D-descriptors. There is another type of descriptor called 4D-descriptors, which are derived from the interaction energies between the molecule, embedded into a grid and some probe [11, 14]. The last two descriptor type (3D and 4D descriptors) have more information content that the simple ones, although the "best descriptors" are those that have information content comparable with the information content of the response for which the model is sought, but in general, there isn't a best molecular descriptor which is valid for all the problems [11, 14]. The 2D-molecular descriptors that are used in the molecular structure are derived from algorithms applied to a topological representation, they have the topological indexes. One alternative to that type of two-dimensional representation molecular graph

are the linear notation systems, like the Wiswesser Line Notation (WLN) system; SMILES (Simplified Molecular Input Line Entry System) which uses line notation that represents a molecule as a single-line string of characters. InChi is another line notation, a more modern one, which resolves many of the chemical ambiguities not addressed by SMILES, in terms of the stereocenters and other model problems. SMARTS (SMiles ARbitrary Target Specification) which is similar to SMILES, but is an expression instead of simple atoms and bond, is another system [11,14,17,18].

Chemical structures that are serialized in standard formats are needed in order to enable exchange and linking of chemical information. InChi, SMILES and other systems are a standardized identifier for chemical structures [11, 17, 18]. The 3D-QSAR method should only be applied to a dataset when the analysis is expected to reveal insights of the 3D structure-activity relationships. The 3D properties of molecules govern biological activity. This method takes into account the 3D structure of ligands and additionally it is applicable to sets of structurally different compounds [11, 14]. The 3D-QSAR methods can help to predict the binding affinities if there are many compounds that bind the target protein in a similar way [11, 14, 17].

There are a number of methods to calculate the chemical formula (ID descriptors), the 2D structure (2D descriptors) and the 3D conformation (3D descriptors) for the molecular descriptors. The methods use atom types, molecular fragments and the three-dimensional structure. The descriptor with the lowest degree of information is the ID descriptors. The ID descriptors are almost never used for the QSAR/QSPR approaches [11, 19].

The topological descriptors are obtained through the molecular graph and encode molecular connectivity into numerical values called topological indexes.

The steric descriptors are involved in the size and shape of molecules. Both are critical in understanding the structural properties which modulate the biological activity of the compounds. If there are enough shape and surface complementarity between the drug and the target, efficient and specific drug target can be ensured. The volume is represented through van der Waals' force. To have the general form of a molecule, the spheroidal properties are used. The bioavailability of drugs is related to lipophilicity, polarity and hydrophobicity properties. These properties can be estimated using computational tools [11, 19].

Some websites provide free "ready to use" training sets (chemical structures associated with data activity). An example of that is www.cheminformatics.org/.

The typical features used in a descriptor calculation software are grouped as the following (Table II):

Type of descriptor	Examples
Constitutional Descriptors	Molecular weight, Simple counts e.g., number of atoms, bonds, rings, Aromaticity indices
Topological Descriptors	Balaban index, Randic indices, Wiener index Kier and Hall connectivity indices1, Kier flexibility index, Kier shape indices, Kappa shape indices, Information content indices, molecular walk counts
Atom Pairs Descriptors 2D and 3D	sum of topological distances between bonds, Presence/absence of different type of bonds like C - C at topological distance 1-7
Geometric descriptors	Gravitation index, Shadow indices
Charged Partial surface area descriptor	Charged polar surface area, total polar surface area (TPSA)
2D Descriptors	MDL keys
Electrostatic Descriptors	Maximum and minimum partial charges, Molecular polarizabilities, Dipole moments and polarity indices
Lipophilicity descriptors	Hansch substituent constant4, Log D, Log P
Quantum chemical descriptors	Charges, HOMO and LUMO energies, Orbital electron densities, Superdelocalizabilities energies

Table III – Type of descriptors and examples [19]

The ideal software to create molecular descriptor needs to be free or cheap, open source (creation of description calculation algorithms), have a GUI (easy usage), work on multiple platforms, accept multiple molecular file formats (SMILES, InChi, etc) and be able to calculate many types of descriptors. This way it can be accessed and used by anyone [11, 24]. There are many different software solutions with some of these characteristics, including CDK Descriptor Calculator GUI v.1.4.5 which is a free open source, command line and GUI, accepts MDL, SDF and SMILES file formats and can be used on multiple platforms. The most remarkable feature is that it can be used to read molecular files and calculate most of the molecular descriptors [11, 24].

2.2.1. Database building

To have a proper model, a good training set must be created. The training set must have information from many different sources and different databases. The database can be explored through chemical identifiers (registry number, chemical name), structure and

sub-structures as well as physicochemical property values. That information can be downloaded in different formats like SMILES.

To have an accurate QSAR/QSPR model, the training set, the molecular descriptors and the biological parameters must be carefully selected. The most important features for the quality of the model are the molecule diversity, the quality of the biological values (standard errors) and the range of biological activity [11, 19, 22].

QSAR/QSPR methods can help to develop drugs with absorption, distribution, metabolism, excretion and toxicity (ADMET) profiles. That means drugs which can be absorbed, distributed through the body, metabolized and excreted with no toxicity for humans. In pharmacokinetics, this is used to describe a drug that can be safely used [19]. To optimize and prioritize the drug candidates, many in silico methods for the prediction of diverse properties and activities were developed. An example of these in silico tools are the QSAR/QSPR models, the decision trees and molecular docking [20].

Chemoinformatic is a process which involves having libraries of small molecules. The molecules can be standardized by adding hydrogens or removing unconnected structures, through the calculation of molecular descriptor and visualization of chemical structures in two or three-dimensions etc. Due to this, many chemoinformatics libraries were created to deal with such tasks. These libraries are produced for chemical toolboxes like CDK and Open Babel [11, 19].

2.2.2. Similarity through fingerprints

One way of encoding the structure of a molecule is through molecular fingerprints. There are different types of fingerprints. The most common is a series of binary digits (bits) that represent the presence or absence of particular substructures in the molecule [11, 13, 24]. Through the comparison of fingerprints, it is possible to determine the similarity between two molecules, to find matches to a query substructure, etc. The different types of fingerprints are provided for OpenBabel, like the fingerprint format that is a path-based fingerprint FP2; substructure based fingerprints FP3, FP4 and Molecular ACCess System (MACCS); user-defined substructures [11, 24]. The FP2 indexes small molecule fragments based on linear segments of up to seven atoms (somewhat similar to the Daylight fingerprints) [24].

The molecular structure is analysed to identify linear fragments of length from 1-7 atoms. Single atom fragments of C, N and double bonds are ignored. When the atoms form a

ring the fragment is terminated [24]. For each of these fragments, the atoms, bonding, and whether they constitute a complete ring is recorded and saved in a set which means there is only one of each fragment type. Chemical identical versions are identified and only a single canonical fragment is retained. Each of the remaining fragments is assigned a hash number from 0 to 1020 which is used to set a bit in a 1024 bit vector [24]. The others FP3, FP4 use series of SMARTS queries and MACCS use the SMARTS patterns. The fingerprint can be created in two ways, through a vector returned by OpenBabel GetFingerprint() method, using Fingerprint (myvector) or by calling the calcfp() method of a molecule [11,24].

2.3. Machine Learning methods

When ligands dock into protein binding sites, binding affinities values can be predicted through statistical methods. An example of those methods are knowledge, regression and first-principle based methods. Several methods were developed to rank computer-generated binding modes. The best results are achieved through a combination of different scoring schemes but with the same consensus scoring approach [20, 30].
In this work, linear regression was used to predict the probability of a ligand to connect to the neuroreceptor [10, 30].

Zhu et al. [48] tried to build a good linear regression method QSAR/QSPR but failed. The correlation coefficients were less than 0.65 for self-fitting and cross validation testing. It is better to use advanced machine learning methods such as Bayesian inference, Random Forests and SVM which showed good results [30].
The knowledge-based method evaluates the increasing number of experimentally determined protein-ligand complexes. Statistical methods are used to extract rules in terms of geometric interaction. The rules are converted into pseudopotentials to be applied to score computer-generated ligand binding modes [30].
Another approach to model the protein-ligand complexes is to use a particular input potential to construct a database, then derive statistical protein-ligand potentials from it. This allows to explore and study the results when potentials are modified [10, 30].

In the beginning, it is better to use all the descriptors available than to reduce the descriptor pool to a smaller set. It is possible to choose classes like topological, electronic, geometric or a combination of those [30].

The workflow analyses one set of molecules using the necessary descriptors to load the molecular structures into the software then transform the list of molecules into a document which the software can read. A vector is created containing the names of the descriptors' classes. Different functions and machine learning techniques like Random Forests are used to choose the best set of descriptors and to build a linear model of predicted versus observed [30].

To have good performance, a combination of general purpose statistical environment and chemoinformatics is required. That means the use of different statistical approaches like LASSO and Support Vector Machines (SVM) to have the best model [10, 30]. The applications for that include exploratory analysis of the datasets, molecular selections based on a combination of statistical properties and chemical information, the development of predictive models for screening purposes and so on [30].

Nowadays, the linear regression model that assumes the regression function E (Y|X) is linear in the inputs $X_1 \ldots X_p$. This simple model gives an interpretable description of how the inputs affect the output [31].

2.3.1. Least Absolute Shrinkage and Selection Operator (LASSO)

The 'LASSO' was one of the most used methods to analyse and estimate linear models, it reduces the residual sum of squares subject to the sum of absolute value of the coefficients being less than a constant, it can produce some coefficients with the zero values. This method is good to do the subset selection and ridge regression [31].

The LASSO performs L1 shrinkage where some regression coefficients shrunk exactly to zero and few other regression coefficients with little shrinkage so that there are "corners" in the constraint, which in two dimensions corresponds to a diamond (figure 3). If the sum of squares "hits" one of these corners, then the coefficient corresponding to the axis is shrunk to zero. As p increases, the multidimensional diamond has an increasing number of corners, and so it is highly likely that some coefficients will be set equal to zero [31].

The figure below shows a geometric interpretation of LASSO:

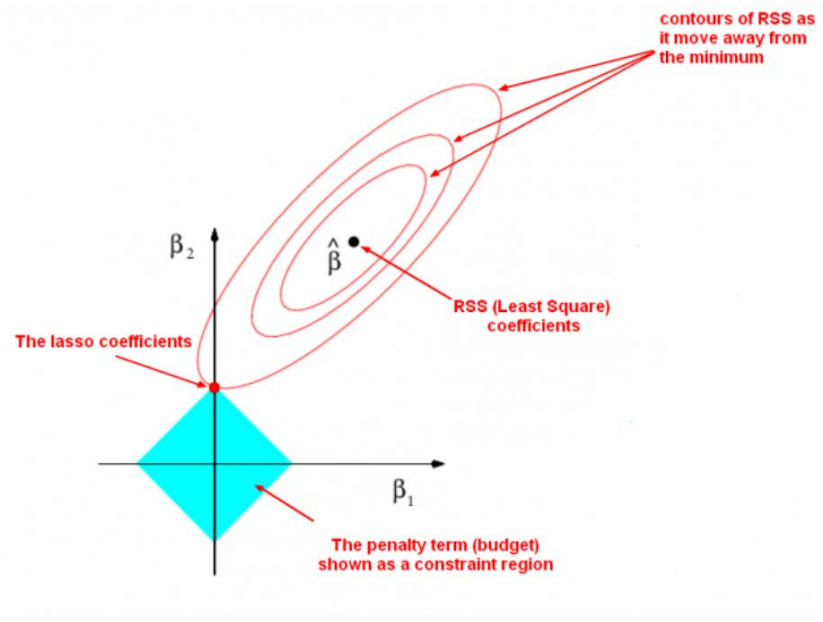

Figure 3 - Estimation picture for the LASSO, where the blue areas are constrains regions and red eclipse are the contours of least squares error function [31].

2.3.2. Random Forests

Random Forests is an ensemble learning method for classification or regression. The algorithm uses decision trees that work through a random subset of the dataset and builds a large collection of de-correlated trees and then averages them. The leaves of the decision tree represent properties/activities values and the branches the conjunctions of descriptors that lead to those properties/activities. This method is simple to train and tune, that is why this method is so popular [10, 30, 31]. The Random Forests is a good algorithm to use with QSAR/QSPR models because it allows to have fewer variables and more observations, shows a good predictive performance even when "noisy" variables are present and has only a minimal necessity to tune the default parameters to achieve a good performance. This method can be used with a mixture of categorical and continuous descriptors and gives the measures of descriptors' importance, the percentage of variation explained and shows how the set of molecular descriptors are capable of explaining the variation in the property/activity value [10, 30, 31].

2.3.3. Support Vector Machine (SVM)

The other method is SVM, which is used for classification or prediction. The idea behind this method is the construction of a decision hyperplane or set of hyperplanes in a high-

dimensional feature space that minimizes the margin using a kernel function to modify the data in way that the data is separately based on the largest distance to the nearest training data points [10]. This method provides a good predictive performance but depends on the selection of the model parameters [10].

2.4. Chemical toolboxes

Chemical toolboxes were developed to understand the many languages of chemical data. The toolbox enables search, convert, analyse or store data from molecular modelling, chemistry, solid-state materials, biochemistry, or related areas [17,18]. The OpenBabel [17] and CDK are examples of Chemical toolboxes used in this work.

Open Babel [17] can be used through the Python interface. This can be used through two options, using the OpenBabel module, that contains standard Python, or the Pybel module which provides an easier way to access OpenBabel toolkit.

The OpenBabel module provides direct access to the C++ OpenBabel library from Python (import OpenBabel) [17].

Pybel has a lot of functions and classes which allows simpler access to the Open Babel libraries, file input/output and attributes of atoms and molecules. A molecule can be created through the OBMol, using Molecule (myOBMol), or by reading from a file or a string [17].

CDK stands for Chemistry Descriptors Kit. CDK major functionality is reading and writing molecule formats, generate binary fingerprints and calculating molecules descriptors. It can be used in several applications, like R, CDK-Taverna and others. In this work CDK was used in R. The kit had several descriptors, but the ones used in this work were the chemoinformatics ones which describes the molecule's structure [18 23,29]. CDK has a large number of molecular descriptor routines located in the package cdk.qsar. The atom based and whole molecular descriptors are available.

Some examples of molecular descriptors available are atom and bond counts, topological descriptors, geometric descriptors and holistic descriptors such as Weighted Holistic Invariant Molecular descriptors (WHIM) and Burden-CAS-University of Texas eigenvalues (BCUT) descriptors [18, 29]. CDK has a variety of chemoinformatics functionality that can be used to develop a variety of applications.

The most used application is the data mining of chemical information, which combines chemoinformatics and statistical tools, like combining CDK with R [18, 23, 29].

2.5. Data sources

To have a good model it is necessary to find as high quality data as possible.

Below are different sources of database that are candidates to build the database:

Name	url	Size	Description	Chemical Structure
The IUPHAR/BPS	http://www.guidetopharmacology.org/about.jsp	485 targets (receptors), 6064 ligands, 41076 binding affinity constants, 21774 references.	Detailed, peer-reviewed pharmacological, functional and pathophysiological information on human, mouse and rat. G Protein-Coupled Receptors, Voltage- and, Ligand-Gated Ion Channels, Nuclear Hormone Receptors, Catalytic receptors, Transporters Enzymes	SMARTS
Ki Database (KiDB)	http://pdsp.med.unc.edu/pdsp.php	55524 Ki values	The Ki database serves as a data warehouse for published and internally derived Ki, or affinity, values for a large number of drugs and drug candidates at an expanding number of GPCRs, ion channels, transporters and enzymes	SMILES
AffinDB	http://pc1664.pharmazie.uni-marburg.de/affinity/	748 affinity values	An affinity database for ligand–protein complexes, including receptors, for which structures are available via the Protein Data Bank	SMILES
DrugBank	http://redpoll.pharmacy.ualberta.ca/drugbank	6811 drug entries, 4294 non-redundant protein sequences are linked to these drug entries	A useful database of approved and experimental drugs with links to potential targets.	SMILES Inchi, MOL PDB SDF
PubChem	http://pubchem.ncbi.nlm.nih.gov	119960567	A searchable database containing chemical information on a large number of compounds, enabling flexible structure-based	Canonical SMILES InChi InChiKey

			homology search. Is easily linked to other receptor-based databases	
ChEMBL DB	https://www.ebi.ac.uk/chembldb/	9356 targets and 1520172 Compound	Bioactive drug-like small molecules, it contains 2-D structures, calculated properties (e.g. logP, Molecular Weight, Lipinski Parameters, etc.) and abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data).	SMILES ChEMBL ID
The Binding Database	http://www.bindingdb.org/bind/index.jsp	1051955 binding data, 7117 protein targets and 440396 small molecules	Proteins and small molecules (nonpolyneb, organic compound, with molecular weight less than 1000 Da) binding affinities via noncovalent interactions.	SDFile, SMILES
PDBbind	http://sw16.im.med.umich.edu/databases/pdbbind/index.jsp	9,308 compounds	Protein-ligand (7121), nucleic acid-ligand (79), protein-nucleic acid (511), and protein-protein complexes (1597).	SMILES InChi
Binding MOAD mother of all databases	http://www.bindingmoad.org/	21109 Protein-Ligand Structures, 7284 Binding Data, 10156 Different Ligands	enzyme, no-enzyme	SMILES
Thomas Ray [33]"	[33]	35 drugs 67 receptors	The objective of Thomas Ray' paper was to present receptor binding profiles of the 35 drugs in way that they can be easily compare for similarity and difference, were the data for that 35 drugs and 67 receptors insert into the new database the values of binding were extract for the tables	SMILES

Table IV – Description of the databases, the name, the url, size of the database and the chemical structure representation used in databases.

3. Data and Methods

3.1. Consolidation of data and database building

The criteria used to select the sources of data to build the database were: easy and free access to the data; SMILES, InChi or SMARTS as chemical structure representation and binding affinity data from neuroreceptors in the human brain. The sources of data that followed these criteria are KiDB, Thomas Ray data [33], PDBbind and BindingDB.

Before inserting any data into the database, it was necessary to normalize the Ki data values in order to allow an easy comparison of the diverse receptor affinity profiles of different ligands [27, 33].

The Ki values are distributed in different magnitudes. It is necessary to transform the Ki values for them to be consistent in all the data. The lower Ki Values are produced for higher affinities, it is important to calculate pKi value, using $pKi=\log_{10}(Ki)$. This means higher affinities have higher pKi and each unit of pKi value corresponds to one order of magnitude of Ki value. If Ki value is bigger than 10000, pKi is equal to 4 [34].

Following criteria was used to normalize the data:

If Ki value > 10 000 the new Ki value = 0 (1)

If Ki value < 10 000 the new Ki value = min (1.0, 1.0-(log10(Ki value))/4.0)) (2)

Performing the normalization means that higher affinities will have higher values and affinities that are too low will be added as zero [27, 33].

For PDBbind database, it was necessary to convert the Ki values into nM, for example, some values were in uM, and they were converted into nM.

The relation model is in the figure below (figure 4) where the measure used was binding affinities expressed in pK_i (Ki_value):

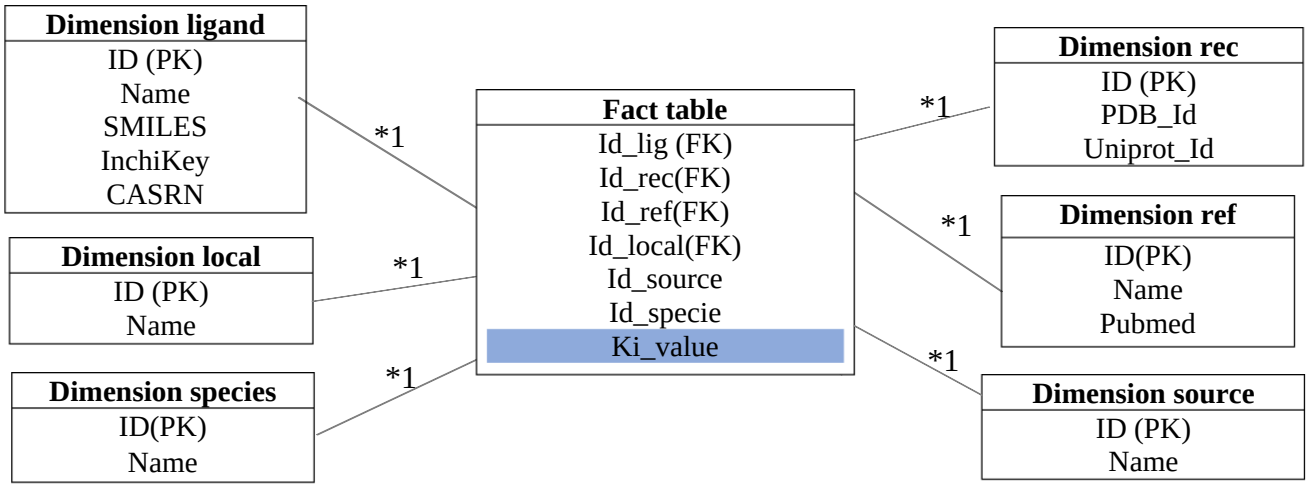

Figure 4 - Relation model with fact table with six dimensions (local, specie, rec, ligand, ref, source).

The new database has six dimensions (Table IV – VIII): the ligands (dimension lig); the receptors (dimension rec); the source organism (dimension species); the references (dimension ref); the localization/tissue (dimension local); data sources (dimension source). In total the table of facts has six attributes: ligand, rec, ref, local, species, source and one metric, the Ki value.

The following table shows the information added in the dimension local, referring to the localization of the receptor in the body:

Dimension local			
Name of the attribute	**Type**	**Description**	**Example**
ID	Numeric	Auto generated	1
Name	String	Localization/tissue	Brain

Table V - Description of the content for the dimension local, related to the localization of the receptor

The following table shows the information added in the dimension rec, referring to the receptors, the name of the receptor and ID number of the receptor in the UniProt (Universal Protein Resource) and PDB (Protein Data Bank):

Dimension rec			
Name of the attribute	**Type**	**Description**	**Example**
ID	Numeric	Auto generated	1
Name	string	The name of receptor/target	Human Glutathione S-transferase P1-1,complex with ter117
PDB_id	string	ID number of the receptor in the PDB	10gs
UniProt_ID	string	ID number of the receptor in the UniProt	42262

Table VI - Description of the content for the dimension rec, related to the receptor (name, the id reference on PDB and UniProt).

The following table shows the information added in the dimension ref, referring to the bibliographic reference of the receptor:

Dimension ref			
Name of the attribute	**Type**	**Description**	**Example**
ID	Numeric	Auto generated	1
Name	String	Reference to the authors where the data comes from	Dignam, JD; Nada, S; Chaires, JB
PubMed	String	Bibliographic reference number or link	12731874 or http://www.ncbi.nlm.nih.gov/pubmed/12731874

Table VII - Description of the content for the dimension ref, related to the bibliographic reference of the receptor.

The following table shows the information added in the dimension lig, referring to the ligand, the name and the molecular information (SMILES, InChiKey and CASRN) of the ligand:

Dimension lig			
Name of the attribute	**Type**	**Description**	**Example**
ID	Numeric	Auto generated	1
Name	String	The name of the ligand/drug(molecule)	(+)-1-(1-(2-fluorophenyl)-2-(2-(trifluoromethoxy)phenyl)ethyl)piperazine
SMILES	String	Chemical nomenclature of the ligand	C1CN(CCN1)C(CC2=CC=CC=C2OC(F)(F)F)C3=CC=CC=C3F
InchiKey	String	IUPAC International Chemical Identifier of the ligand	ZKHQWZAMYRWXGA-KQYNXXCUSA-N
CASRN	String	Unique numerical identifier of ligand	NSC664704 or 71125-38-7

Table VIII - Description of the content for the dimension ligand related to the compound that connects to the receptor (name, molecular information, like SMILES).

The following table shows the information added in the dimension source, referring to the data source of the binding affinity values between the receptor and the ligand:

Dimension source			
Name of the attribute	**Type**	**Description**	**Example**
ID	Numeric	Auto generated	1
Name	string	Data source	KIDB

Table IX - Description of the content for the dimension source, the name of the data source.

3.2. Data curation for the dataset

The selection for the receptors were based on the following criteria:

1. Receptor from the human brain;
2. Having more than 50 molecules with binding affinity data;

3. The value of binding affinity between the receptor and the ligand needs to be bigger than zero.

Note: The data where the species wasn't defined (UNDEFINED) but belongs to neuroreceptor was extracted too.

Receptors which followed the above criteria were extracted into a file. For each ligand that can bind to the receptor, a file was created with the following information regarding the Id of the ligand, the ligand's structure (SMILES), binding affinity value (Ki value) and the animal species (figure 5).

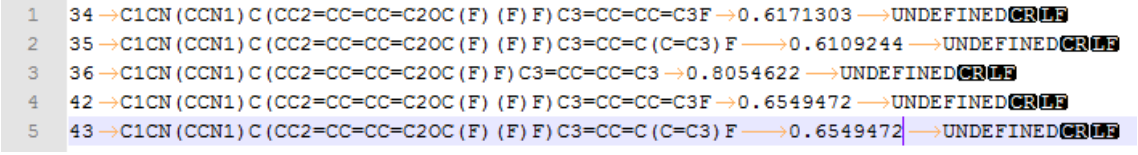

Figure 5 – Example of the content in file for 5-HT-Transporter.

Afterwards, two files were created, one with OpenBabel descriptors (fingerprints fp0001) and the other with CDK descriptors (fingerprints V1). Both have different types of descriptors from Pybel (OpenBabel) and pyCDK library (CDK).

The files were combined to create the dataset. The final result has the Id of ligand, the species, the binding affinity value (Ki value), the fingerprints and another descriptor from OpenBabel and fingerprints and another descriptor from CDK. Below is an example of one file from 5-HT-Transporter (Figure 6).

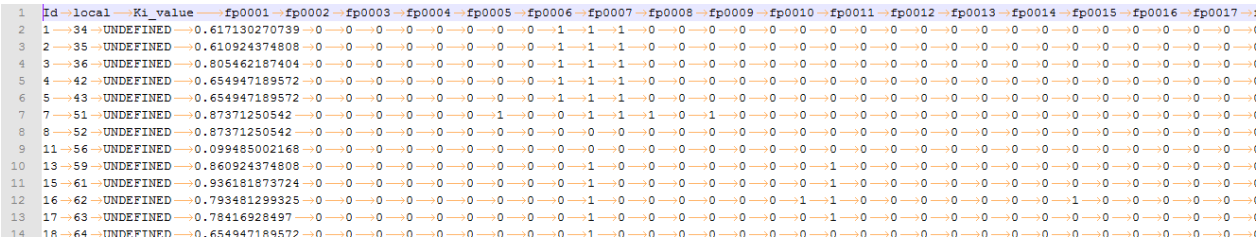

Figure 6 - Example of the content in file from 5-HT-Transporter with all the fingerprints and descriptors from OpenBabel and CDK.

The molecules were processed using the Pybel, OpenBabel libraries and pyCDK library.

3.3. Model Building

The dataset was analysed using R. It was necessary to divide the dataset into training set that is 75% of the original dataset and 25% for the test set. The test set was used to validate the model. The division was made to randomize the process.

To build the models, three machine learning methods were used: the LASSO, SVM and Random Forests. Following scheme (figure 7) represents the process:

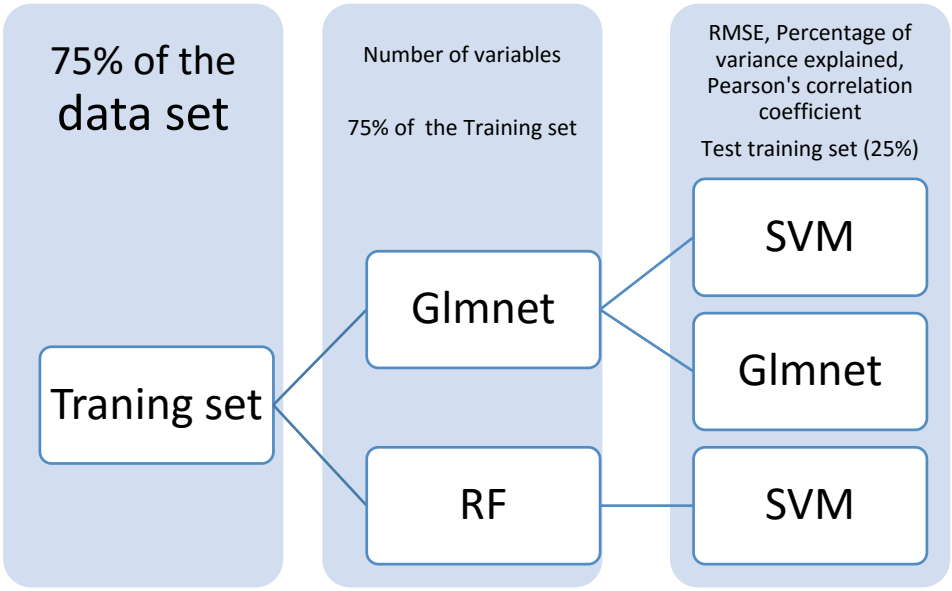

Figure 7 – Scheme showing the process to build a model.

- The Random Forests and Least Absolute Shrinkage and Selection Operator (LASSO) methods were used to select the most important variables (descriptors) which can describe a compound and predict the binding affinity;
- LASSO and SVM were used to validate the model, to verify if the selected variables were a good predict for the binding affinity values.

In each process the cross-validation was run five times, which means that the dataset (training set) was randomly divided into five equal (or almost equal) parts and divided into a training set (75% of dataset) and test set (25% of dataset). Each time a method was run a new training set and test set were created.

3.3.1. LASSO

The idea behind LASSO method is that we have data (x, y), where x is the predictor variable and y are the responses for N independent observations or we can assume that they are conditionally independent given x [31].

Then $$\sum_i \frac{x_{ij}}{N} = 0 \, , \sum_i \frac{x_{ij}^2}{N} = 1 \quad (3)$$

$$\beta' = (\beta_1' \ldots . \beta_p')$$

LASSO technique minimizes the log partial likelihood subject to the sum of the absolute values of the parameters being bounded by a constant. This will produce some coefficients that are exactly zero, because of that the result is the reduction of the estimation variance while providing an interpretable model [31].

Computing the LASSO solution is a quadratic programming problem that can be resolved through efficient algorithms that compute the entire path of solutions as λ, where making t sufficiently small will cause some of the coefficients to be exactly zero. The LASSO technique is good for variable selection, but should be used in combination with other model building tools. [31].

3.3.2. Random Forests

Random Forests are a learning machine method for classification and regression. In regression, the same regression tree is used many times to bootstrap sampled versions of the training data and average the result [10, 31]. Random Forests uses the technique of bagging to build a large collection of de-correlated trees and then averages them. The training set $X=x_1 \ldots x_n$ with responses $Y=y_1 \ldots y_n$, bagging repeatedly (B times) then selects a random sample with replacement of the training set and fits trees to those samples. Each tree generated in bagging is identically distributed with the expectation of an average of B; such trees are the same as the expectation of any one of them. This algorithm improves the variance reduction of bagging b reducing the correlation between the trees, without increasing the variance too much. This happens in the tree-growing process through random selection of the input variables.

Random Forests is an important feature of the out-of-bay samples, which means for each observation $z_i=(x_i,y_i)$, will construct its Random Forests predictor by averaging only those trees corresponding to bootstrap samples in which z_i didn't appear. The features' error estimate is almost identical to that obtained by N-fold cross-validation.

The variable importance plots can be constructed for Random Forests in the same way as they were for gradient-boosted models. For that, at each split in each tree, the improvement in the split criterion is the important measure attributed to the splitting variable and is accumulated over all the trees in the Forest separately for each variable.

The out-of-bag samples are used to construct a different variable-importance measure, to measure the prediction strength of each variable. When the bth tree is grown, the out-of-bay samples are passed down the tree and the prediction accuracy is recorded, after that the values for the j variable are randomly permuted in the out-of-bay samples, and the accuracy is again computed. When the accuracy decreases, it is a result of average permutation over all trees and is used as a measure of the importance of variable j in the Random Forests. This doesn't measure the effect on prediction were this variable isn't available, because if the model was refitted without the variable, other variables could be used as surrogates [35].

3.3.3. Support Vector Machines (SVM)

Support vector machines can be used for regression with a quantitative response, in ways that inherit some of the properties of the SVM classifier. The measures used to verify the model's validation were the root mean squared error, the percentage of variance explained and the correlation between the variables. The function used in SVM was Gaussian radial basis kernel and has two parameters, cost and gamma. Cost represents the penalty associated with larger errors. When this value increases it causes the fitting to the training data. Gamma controls the shape of the separating hyper plane, with the increase of this value, the value of the support vectors increases [10].

3.3.4. Selection of variables

The methods used for the selection of variables in order to have the minimum number of descriptors for receptors were LASSO and Random Forests.

The process of selecting variables in the case of using the Random Forests method is the following: the function 'randomForest' is used to extract the variables that are most important to the prediction. The dataset was divided to extract what we want to predict from rest of the training set. The number of trees used was 500 because the number of trees doesn't have a huge impact in the statistical results [34]. The model is trained and predicts values are calculated. The most important variables for the model are extracted in order, into a file.

In the case of using LASSO method, the function used was 'glmnet', where X is the matrix with the observations (training dataset without the values of Ki value) and Y the response variable (values of Ki value) and the type of response, in this case, a quantitative

response (code is in Appendix). The regularization path is computed for the LASSO penalty at a grid of values for the regularization parameter lambda.

3.3.5. Model Fitting

For the model fitting the SVM and LASSO methods were used. The process is below:

1. For each run (five runs in total) divide the dataset into a training set and test set. In each run, the set is randomly created.
2. The function 'svm' was used for support vector machine method and 'glmnet' for LASSO to do the model.
3. The 'predict' function is used which predicts the values based on the model trained by 'svm' or 'glmnet'.
4. The RMSE and others measures were calculated.

Each variable was ordered by importance of variables (Random Forests or LASSO). A model was predicted through SVM method and LASSO. In the SVM method, the variable entered the dataset, one by one. The LASSO method is slightly different. There is an extra argument which needs to be considered: the values of the penalty parameter lambda at which predictions are required. These values are the number of variables.

In this study, the SVM implementation used was provided by the package 'e1071, for LASSO was 'glmnet' package and for Random Forests was 'randomForest' from R.

3.3.6. Model validation

To validate the robustness and predictive ability of the models the 5-fold cross validation or out-of-bag prediction (Random Trees) was used. The measures used to determine the external predictive ability of the model [10] were root mean squared error (RMSE), percentage of variance explained and Pearson's correlation. The RMSE is used as a standard statistical metric to measure the difference between values predicted by a model and the values observed, allowing the of measure performance in different types of studies, the quality of the fit between the data and the predict model, showing the error distribution [36, 38].

The formula of RMSE is the following:

$$\text{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \qquad (5)$$

Where $\hat{y}_i$the estimator of the dependent variable y_i and N are the number of predictions. The best model will have the lowest value of RMSE [10, 41]

In regression, RMSE is used frequently but RMSE can also be used in binary classification [40, 41]. RMSE is a good measure to use when the sample is bigger than 100, as it gives reliable and robust values [37]. The formula of RMSE uses predicts values (obtain through the 'svm' function or 'glmnet'in R) and the binding affinity values.

The mean squared error (MSE) of an estimator measures the average of the squares of the "errors", which means that the difference between the estimator and what is estimated can estimate the error of the variance [37,38]

The percentage of variance explained measures the percentage to which the model accounts for the variation (dispersion) of a given dataset. For the perfect regression relationship, the percentage of variance explained is 100% when the model has 100% accuracy. When the percentage of variance explained (PVE) decreases, the estimation of the model is bad. If the value is zero, the model doesn't have any predictive value [47]. In this work the PVE was calculated using the following formula which was used in the code (a), first the mean squared error was calculated using the predicts values and the binding affinity values in the test set, and then using the function 'var' in R (b) to calculate the variance of binding affinity values in the test set.

a. mse=mean((predsvm-teste2$Ki_value)^2)
b. (1-mse/var(teste2$Ki_value))*100

The Pearson's correlation coefficient is the relation between two variables, or dataset [46, 47].

$$Corr=r_{xy}= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad (6)$$

i=1, 2,.., n, where $\bar{x}$ and $\bar{y}$ are means of x and .

Pearson's correlation coefficient measures the strength and direction of the linear relation between the variables. When the correlation is zero, it means there isn't any relationship between variables, if the correlation is one, it has perfect correlation [45]. If the correlation is higher than 0.8, there is a really strong correlation, if is lower than 0.5 the correlation is weak [**Error! Reference source not found.**, 45]. If the correlation between p

redicted and observed activity values is good, the model has higher predictive accuracy [**Error! Reference source not found.**]. The correlation between the observed values and t he predicted values was calculated using the correlation function in R, using the dataset with the values of binding affinity and the predictions numbers, calculated using SVM.

4. Results and Discussion

4.1. Data preparation

The combination of four different data sources created a dataset with 198169 binding affinity values (Ki values), 148391 compounds, 8300 references, 4524 receptors, 246 species and 197 places in the body where the receptors are located.

Only a part of the database was used to build the models because the data selected only had receptors from the brain with more than 50 entries. The final dataset had 53 receptors, 32943 compounds and 43901 binding affinity values. Most of the data is from the KIDB (55%), 40 % from BindingDB, 3% from Thomas Ray paper and 1% from PDBbind. In the next table shows the number of compounds for each receptor which exist in the final dataset.

Receptor	Dataset	Training dataset	Validation dataset
5-HT Transporter	1025	769	256
5-HT1A	1880	1410	470
5-HT1B	847	636	211
5-HT1D	880	660	220
5-HT1E	178	134	44
5-HT2A	1438	1079	359
5-HT2B	751	564	187
5-HT2C	1112	834	278
5-HT3	270	203	67
5-HT4	142	107	35
5-HT5a	208	156	52
5-HT6	1043	783	260
5-HT7	583	438	145
Adenosine A1	1680	1260	420
Adenosine A2a	1625	1219	406
Adenosine A2B	687	516	171
Adenosine A3	2194	1646	548
Adrenergic Alpha2A	252	189	63
Adrenergic Alpha2B	214	161	53
Adrenergic Alpha2C	221	166	55
Adrenergic Beta1	180	135	45
Adrenergic Beta2	217	163	54
Alpha-1a adrenergic receptor	647	486	161
Alpha-1b adrenergic receptor	668	501	167
Alpha-1d adrenergic receptor	250	188	62
Cannabinoid CB1	1246	935	311
Cannabinoid CB2	1196	897	299
Cholecystokinin B	230	173	57
Cholinergic, muscarinic M1	702	527	175
Cholinergic, muscarinic M2	866	650	216
Cholinergic, muscarinic M3	630	473	157
Cholinergic, muscarinic M4	375	282	93
Cholinergic, muscarinic M5	373	280	93
Cholinergic, Nicotinic Alpha2Beta2	346	260	86
Cholinergic, Nicotinic Alpha4Beta2	112	84	28
Delta opioid receptor	1764	1323	441
Dopamine D1	681	511	170
Dopamine D2	2860	2145	715
Dopamine D3	1884	1413	471
Dopamine D4	1811	1359	452
Dopamine D5	392	294	98

Dopamine Transporter	670	503	167
Histamine H1	797	598	199
Histamine H2	174	131	43
Histamine H3	1037	778	259
Kappa opioid receptor	2200	1650	550
Metabotropic glutamate receptor 1	136	102	34
Metabotropic glutamate receptor 5	114	86	28
Mu opioid receptor	2208	1656	552
Neurokinin NK1	328	246	82
Neurokinin NK2	393	295	98
Neurokinin NK3	235	177	58
Norepinephrine transporter	949	712	237

Table X - Shows the number of compounds for each receptor in the dataset, training dataset and validation dataset.

4.2. Model Fitting

To build the model, three methods were used together. First was the selection part, where the best number of the most important descriptors are selected. That means to select descriptors which helps to describe the compound and to predict the binding affinity between the receptor and ligand; Random Forests or LASSO were used in this part. In the validation part, the Support Vector Machine or LASSO were used.

In beginning, the value of RMSE is high and when more variables (descriptors) were added (figures 8a-8c), that value starts to decrease, levels off and then begins to increase again. For a good model, the number of descriptors is around 4 and 100, depending on the receptor and the number of compounds used for the model.

The combination of different machine learning methods for the selection part gives the following figure for the 5-HT Transport:

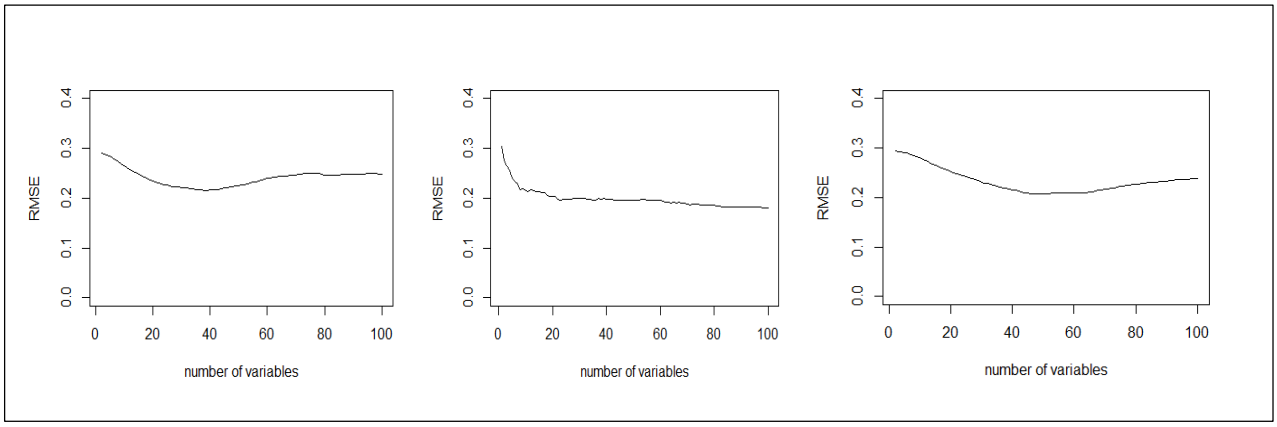

Figure 8 – (a) Plot for 5-HT Transporter that shows the value of Root Mean Squared Error (RMSE) using SVM according to the number of variables in consideration, obtained though Random Forests; (b) Plot for 5-HT Transporter that shows the value of RMSE using SVM according to the number of variables in consideration, obtained though LASSO; (c) Plot for 5-HT Transporter that shows the value of RMSE using LASSO according to the number of variables in consideration, obtained though LASSO.

The results of the training for each method are in the following table, which shows the values of RMSE, percentage of variance explained (PVE), Pearson's correlation coefficient (r) and the number of variables for each receptor.

Receptor	Training set	Random Forests and SVM				LASSO and SVM				LASSO and SVM			
		nºvars	RMSE	PVE	r	nºvars	RMSE	PVE	r	nºvars	RMSE	PVE	r
5-HT Transporter	769	98	0.196	57.3	0.759	58	0.211	49.4	0.709	40	0.227	42.2	0.655
5-HT1A	1410	99	0.189	50.9	0.714	129	0.197	46.5	0.684	29	0.220	31.4	0.563
5-HT1B	636	48	0.179	60.3	0.780	46	0.195	52.3	0.727	43	0.213	44.4	0.667
5-HT1D	660	62	0.179	60.0	0.774	45	0.199	52.3	0.725	36	0.223	39.1	0.632
5-HT1E	134	35	0.146	22.7	0.579	23	0.161	12.6	0.423	25	0.154	26.7	0.508
5-HT2A	1079	100	0.186	60.1	0.778	108	0.194	57.2	0.758	42	0.213	48.7	0.699
5-HT2B	564	90	0.140	71.7	0.847	77	0.153	68.1	0.827	34	0.183	54.8	0.743
5-HT2C	834	98	0.150	62.2	0.789	114	0.168	59.4	0.771	28	0.212	35.8	0.605
5-HT3	203	96	0.164	69.6	0.831	34	0.182	66.6	0.817	33	0.210	55.2	0.743
5-HT4	107	58	0.104	75.1	0.875	21	0.144	62.8	0.814	46	0.141	69.3	0.829
5-HT5a	156	27	0.115	51.1	0.727	20	0.124	46.5	0.720	41	0.137	35.2	0.554
5-HT6	783	100	0.182	55.9	0.747	68	0.194	46.6	0.693	40	0.200	44.3	0.668
5-HT7	438	94	0.175	61.2	0.801	57	0.188	52.3	0.724	30	0.216	38.9	0.618
Adenosine A1	1260	100	0.183	45.7	0.686	105	0.193	42.5	0.657	30	0.224	22.3	0.472
Adenosine A2a	1219	64	0.152	71.3	0.848	67	0.165	64.5	0.808	41	0.186	55.3	0.744
Adenosine A2B	516	19	0.160	61.4	0.786	42	0.173	53.0	0.730	34	0.189	43.9	0.664
Adenosine A3	1646	90	0.198	56.1	0.751	121	0.221	45.1	0.694	35	0.232	40.0	0.633
Adrenergic Alpha2A	189	17	0.159	58.2	0.764	20	0.179	51.4	0.739	32	0.183	45.1	0.682
Adrenergic Alpha2B	161	21	0.144	54.8	0.746	27	0.176	51.0	0.720	29	0.179	45.1	0.664
Adrenergic Alpha2C	166	5	0.154	67.7	0.850	39	0.145	66.6	0.813	46	0.156	61.9	0.803
Adrenergic Beta1	135	22	0.165	77.8	0.880	21	0.214	61.9	0.797	34	0.238	51.5	0.715
Adrenergic Beta2	163	35	0.165	74.7	0.860	12	0.215	60.0	0.786	38	0.252	45.7	0.691
Alpha-1a adrenergic receptor	486	25	0.187	56.5	0.755	55	0.206	55.0	0.746	28	0.235	41.7	0.649
Alpha-1b adrenergic receptor	501	79	0.182	59.6	0.780	53	0.191	54.5	0.753	34	0.211	45.9	0.675
Alpha-1d adrenergic receptor	188	36	0.099	89.6	0.947	24	0.178	62.8	0.801	45	0.201	56.0	0.751
Cannabinoid CB1	935	59	0.165	62.9	0.792	91	0.169	61.9	0.791	44	0.177	58.6	0.769
Cannabinoid CB2	897	92	0.143	76.8	0.879	50	0.159	71.4	0.845	53	0.181	63.5	0.799
Cholecystokinin B	173	4	0.214	27.4	0.556	11	0.246	30.8	0.578	28	0.254	23.9	0.471
Cholinergic, muscarinic M1	527	95	0.165	59.3	0.775	70	0.215	39.9	0.636	32	0.234	29.0	0.543
Cholinergic, muscarinic M2	650	92	0.200	55.3	0.739	63	0.210	50.8	0.718	27	0.234	39.0	0.627
Cholinergic, muscarinic M3	473	41	0.172	61.1	0.780	70	0.197	52.1	0.723	29	0.223	37.7	0.622
Cholinergic, muscarinic M4	282	77	0.193	56.6	0.751	37	0.204	52.5	0.737	32	0.227	40.8	0.648
Cholinergic, muscarinic M5	280	77	0.138	58.5	0.757	45	0.190	55.1	0.742	36	0.221	40.3	0.643
Cholinergic, Nicotinic Alpha2Beta2	260	12	0.065	64.4	0.822	10	0.121	27.7	0.558	28	0.123	24.7	0.434
Cholinergic, Nicotinic Alpha4Beta2	84	92	0.118	85.9	0.924	19	0.198	53.7	0.766	32	0.223	51.3	0.714
Delta opioid receptor	1323	95	0.200	56.1	0.751	73	0.211	49.9	0.713	38	0.228	42.1	0.651
Dopamine D1	511	97	0.181	47.3	0.686	34	0.206	34.1	0.594	28	0.215	27.9	0.519
Dopamine D2	2145	100	0.179	47.9	0.702	118	0.177	56.4	0.751	32	0.203	42.5	0.654
Dopamine D3	1413	94	0.197	58.4	0.773	55	0.196	58.2	0.764	45	0.218	47.9	0.691
Dopamine D4	1359	100	0.198	48.5	0.698	82	0.216	38.5	0.622	32	0.241	22.8	0.486
Dopamine D5	294	79	0.200	43.9	0.603	31	0.221	32.1	0.572	32	0.239	21.2	0.461
Dopamine Transporter	503	99	0.161	47.4	0.646	35	0.181	48.8	0.709	24	0.202	37.4	0.616
Histamine H1	598	94	0.187	61.4	0.787	72	0.188	63.3	0.796	43	0.209	54.0	0.737
Histamine H2	131	100	0.123	32.9	0.648	22	0.142	36.4	0.613	45	0.149	35.7	0.582
Histamine H3	778	38	0.153	60.1	0.778	53	0.177	47.9	0.696	35	0.198	35.5	0.600
Kappa opioid receptor	1650	60	0.203	58.9	0.769	126	0.229	41.5	0.657	34	0.244	35.5	0.597
Metabotropic glutamate receptor 1	102	31	0.146	29.8	0.526	1	0.153	24.6	0.559	31	0.146	36.8	0.584
Metabotropic glutamate receptor 5	86	97	0.122	70.5	0.841	17	0.161	65.3	0.810	52	0.159	61.7	0.803
Mu opioid receptor	1656	96	0.206	55.1	0.743	114	0.218	50.5	0.714	39	0.248	36.5	0.606

Neurokinin NK1	246	32	0.108	64.7	0.803	19	0.140	51.3	0.719	40	0.143	50.6	0.710
Neurokinin NK2	295	34	0.139	68.5	0.847	33	0.172	58.3	0.765	45	0.181	53.3	0.734
Neurokinin NK3	177	97	0.126	72.1	0.845	28	0.162	67.8	0.824	43	0.178	60.5	0.783
Norepinephrine transporter	712	100	0.173	51.2	0.715	86	0.184	45.9	0.682	33	0.200	37.2	0.614

Table XI - Results of the best run with all three methods, where r is the Pearson's correlation coefficient, RMSE and nºvars means number of variables (descriptors) and PVE is the percentage of variance explained.

Comparing the values for the best model in each method (table X), the Random Forests and SVM had the best results, because the values of RMSE were the lowest between the three methods in almost all the receptors; a small value of RMSE indicates better performance of the model and a small value of deviation for the model of predicts. The lowest value was 0.065 for Cholinergic, Nicotinic Alpha2Beta2 model and the highest value was 0.214 for Cholecystokinin B model, which means the values of binding affinity predicted by the models built with Random Forests and SVM methods are closer to the binding affinities values in the training set.

The values of binding affinities (in pKi) are between 0 and 1. A higher pKi value (close to 1) corresponds to lower Ki values ($pKi = 1- (\log_{10}(Ki\ value))/4$), and a higher affinity between the receptor and ligand. In cases where Ki values are bigger than 10000, the pKi is 0. If the value of pKi is 0.5, the Ki value is 100. Most of the values of RMSE from the method with the best results (Random Forests and SVM) are around 0.1 and 0.2. If the value of RMSE is around 0.1, the predicted values of pKi will be between around 0.4 and 0.6 ([0.5-0.1; 0.5+0.1) and the Ki values around 39 and 251. However, if the RMSE is around 0.2 the predicted values of pKi will be between around 0.3 and 0.7 ([0.5-0.2; 0.5+2]) and the Ki values around 15 and 631. There are models such as Adrenergic Alpha2C, Dopamine D2 and Dopamine D3 which have better results with LASSO and SVM methods meaning that those models presented lower RMSE values with those methods.

Some models were more difficult to adjust, for example, the metabotropic glutamate receptor and Histamine H2, and other models where the dataset was small (less than 200 entries). However, when the dataset is bigger, it takes longer to build the model.

Perhaps a compromise is needed when considering the time taken and the quality of the model. It isn't always necessary to a have a bigger dataset, instead with the best model the same results can be obtained with a smaller dataset. However, the dataset needs to have at least 200 entries. By choosing an adequate dataset size, valuable research time can be saved.

Most of the models have key descriptors which need to be included to build the best model (small value of RMSE). In Random Forests method, the number of key descriptors can vary from as few as 4 descriptors to a maximum of 100. Most of the models need more than 70 descriptors in order to build the best model. There are some exceptions where a small number of descriptors are needed to build the best model. In these models, if there are high number of descriptors then the values of RMSE is higher and values of percentage of variance explained are lower. For example, Cholecystokinin B only needs 4 descriptors (Table X) to have a lower value of RMSE (0.214).

4.3. Model Validation

To validate the methods (Random Forests and SVM) with the best models, a dataset with 25% of the original dataset was tested to check if the results were the same. The final results are in the following table:

Receptors	Nº descriptors	Dataset	RMSE	PVE	r
5-HT Transporter	98	1025	0.216	51.1	0.711
5-HT1A	99	1880	0.193	45.1	0.669
5-HT1B	48	847	0.190	57.1	0.756
5-HT1D	62	880	0.228	47.1	0.682
5-HT1E	35	176	0.110	41.0	0.592
5-HT2A	100	1438	0.213	50.3	0.600
5-HT2B	**90**	**750**	**0.156**	**70.5**	**0.853**
5-HT2C	98	1109	0.187	49.4	0.704
5-HT3	**96**	**270**	**0.176**	**70.7**	**0.856**
5-HT4	**58**	**142**	**0.097**	**83.8**	**0.927**
5-HT5a	27	208	0.142	45.1	0.728
5-HT6	**100**	**1042**	**0.169**	**61.9**	**0.812**
5-HT7	**94**	**583**	**0.186**	**54.4**	**0.746**
Adenosine A1	100	1680	0.162	51.7	0.728
Adenosine A2a	64	1625	0.162	59.6	0.772
Adenosine A2B	**19**	**687**	**0.151**	**69.4**	**0.839**
Adenosine A3	90	2194	0.197	57.6	0.764
Adrenergic Alpha2A	17	251	0.170	61.6	0.870
Adrenergic Alpha2B	21	214	0.182	69.4	0.640
Adrenergic Alpha2C	5	221	0.191	57.6	0.670
Adrenergic Beta1	22	180	0.202	61.6	0.731
Adrenergic Beta2	35	217	0.200	44.7	0.795
Alpha-1a adrenergic receptor	25	647	0.203	49.1	0.694
Alpha-1b adrenergic receptor	79	668	0.190	52.4	0.717
Alpha-1d adrenergic receptor	**36**	**250**	**0.083**	**94.7**	**0.979**
Cannabinoid CB1	**59**	**1246**	**0.173**	**57.5**	**0.760**
Cannabinoid CB2	92	1196	0.182	70.7	0.841
Cholecystokinin B	4	230	0.195	24.6	0.473
Cholinergic, muscarinic M1	**95**	**702**	**0.171**	**59.0**	**0.777**
Cholinergic, muscarinic M2	92	865	0.175	63.9	0.797
Cholinergic, muscarinic M3	41	630	0.212	43.9	0.626
Cholinergic, muscarinic M4	77	375	0.149	59.5	0.777
Cholinergic, muscarinic M5	77	372	0.129	62.8	0.820
Cholinergic, Nicotinic Alpha2Beta2	12	346	0.100	49.3	0.898
Cholinergic, Nicotinic Alpha4Beta2	92	112	0.118	40.0	0.877
Delta opioid receptor	95	1764	0.218	47.7	0.699
Dopamine D1	97	681	0.201	46.8	0.702
Dopamine D2	**100**	**2860**	**0.176**	**54.9**	**0.740**
Dopamine D3	94	1884	0.213	52.3	0.728
Dopamine D4	100	1811	0.203	39.5	0.627
Dopamine D5	79	392	0.180	54.5	0.830
Dopamine Transporter	99	670	0.139	50.3	0.871
Histamine H1	94	797	0.212	51.9	0.715
Histamine H2	100	174	0.114	25.5	0.507

Histamine H3	38	1035	0.167	53.5	0.735
Kappa opioid receptor	60	2198	0.201	50.2	0.712
Metabotropic glutamate receptor 1	31	136	0.145	38.2	0.862
Metabotropic glutamate receptor 5	97	114	0.099	71.2	0.913
Mu opioid receptor	**96**	**2208**	**0.201**	**55.2**	**0.742**
Neurokinin NK1	32	327	0.129	46.9	0.706
Neurokinin NK2	**34**	**393**	**0.118**	**79.8**	**0.899**
Neurokinin NK3	97	235	0.139	70.5	0.836
Norepinephrine transporter	**100**	**947**	**0.172**	**50.3**	**0.710**

Table XII - For each receptor we have the number of compounds in the training set and the test set as well as the values of the number of descriptors, RMSE, percentage of variance explained (PVE) and Pearson's correlation coefficient (r) values where the best models are highlight in bold.

The best models are highlighted in bold in table XI, where the difference between the results from the training set and validated set are similar and the RMSE value is the lowest [0.087; 0.201] and the percentage of variance explained is over 50%. Most of these models have more than 50 descriptors and a dataset with more than 200 entries, meaning perhaps those requirements are necessary to build the best model.

There is some variation of values of RMSE, PVE and correlation and the number of descriptors, the minimum value of descriptors used for the model is 4 and the maximum is 100. The RMSE is a good indicator of the quality of the model as it measures the fit between the data and the predicted model. The highest value of RMSE is 0.228 and lower is 0.083, lower values of RMSE indicate a good fit. The RMSE can be interpreted as the standard deviation of the unexplained variance and it is in the same units as the response variable. RMSE is a good measure of how accurately the model predicts the response.

In terms of correlation that gives us the strength and direction of the linear relation between the variables. The lowest value is 0.473 and the highest is 0.979, this shows a bigger variation, but the average correlation is above 0.50, which means a strong correlation.

The other measure is percentage of variance explained, which gives variance and dispersion of the values of binding affinity. The lowest value is 24.6 for Cholecystokinin B that is the descriptor with the lowest number of descriptors, but more than half of the receptors have percentage of variance explained above of 50%, meaning that the model has a good predictive value because the models show low variance and dispersion.

Comparing the results of the training models (table X) with Random Forests and SVM with the results of the validated models (table XI), some models have different values for percentage of variance explained, for example in the Cholinergic, Nicotinic Alpha4Beta2, where the values of RMSE and the correlation are similar but the values of the percentage of variance explained has a difference of more than 50%. The models where the values have the biggest difference are for the adrenergic neuroreceptors and two models for the dopamine neuroreceptors.

The final model shows, in general, a good fit because the values of RMSE, PVE and correlation of the training set are similar to the independent test set. These results show that the model can learn and produce consistent values.

The selection of most important variables shows that some variables were selected again in different receptors, meaning that the same variable (descriptor) is important in a different receptor for the prediction of the binding affinity. The descriptors which are most important in more than 20 receptors are shown in the following figure:

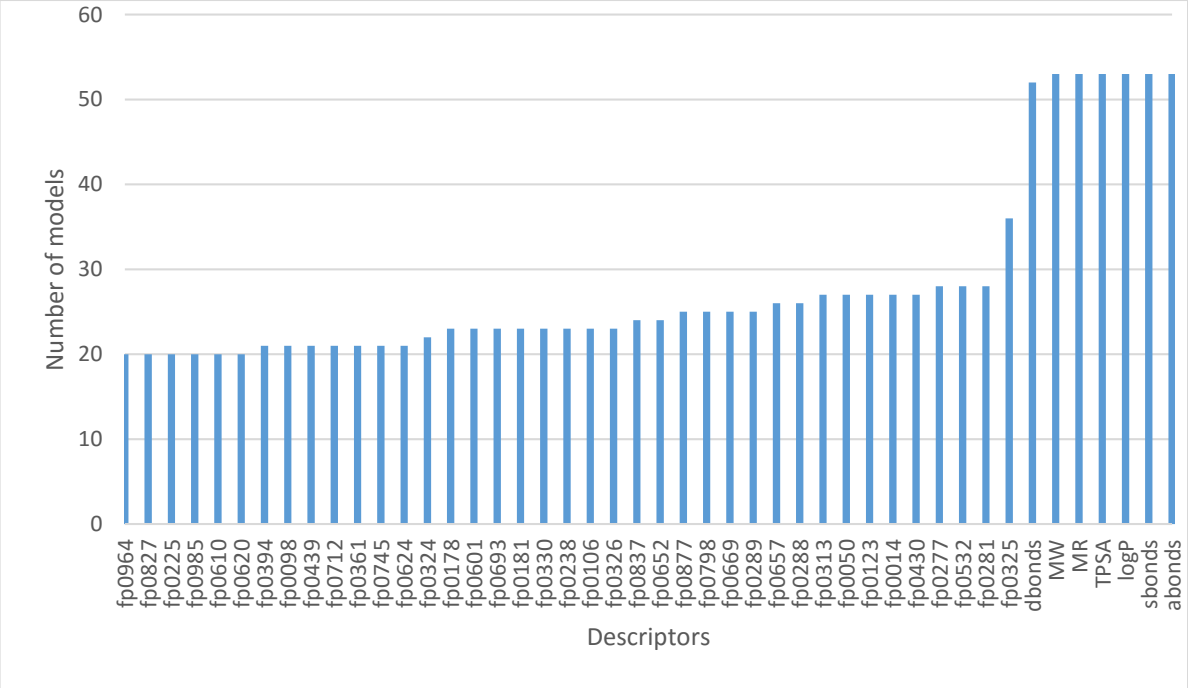

Figure 9 - The most important descriptors in more than 20 models (receptors).

From all the descriptors from OpenBabel and CDK's libraries, 46 descriptors from OpenBabel's library were selected (figure 9) in at least 20 receptors. This demonstrated that these descriptors are the most important descriptors for the construction of the model which can predict binding affinity values. Most of the descriptors are fingerprints. The descriptors

which were selected in every model were MW (Molecular Weight), MR (Molar Refractivity), TPSA (Polar Surface Area), LogP logarithm of the octanol/water partition coefficient, sbonds (number of single bonds) and abonds (the number of aromatic bonds), meaning that those are the most important descriptors to predict the binding affinity.

In order to understand if the families of receptors are related a cluster dendrogram was created (figure 10). The dendrogram shows that the same descriptors are used to identify the same family of receptors. There are some "outliers" like the histamine receptors and metabotropic glutamate receptors.

There is a relationship between dopamine and serotonin receptors because they are in the same group in the dendrogram. This interaction may be related to the fact that dopamine neural cell bodies and terminal sites are modulated by serotonin [49]. The same happens with the histamine and serotonin neuroreceptor - they are related. When the histamine increases, the 5-HT release decreases because the increased activation of histamine H3 receptor inhibits the 5-HT release [50]. The cholecystokinin B and serotonin neuroreceptor, are somehow linked because when either of these neuroreceptor interact with a ligand it is related to anxiety behaviour [51].

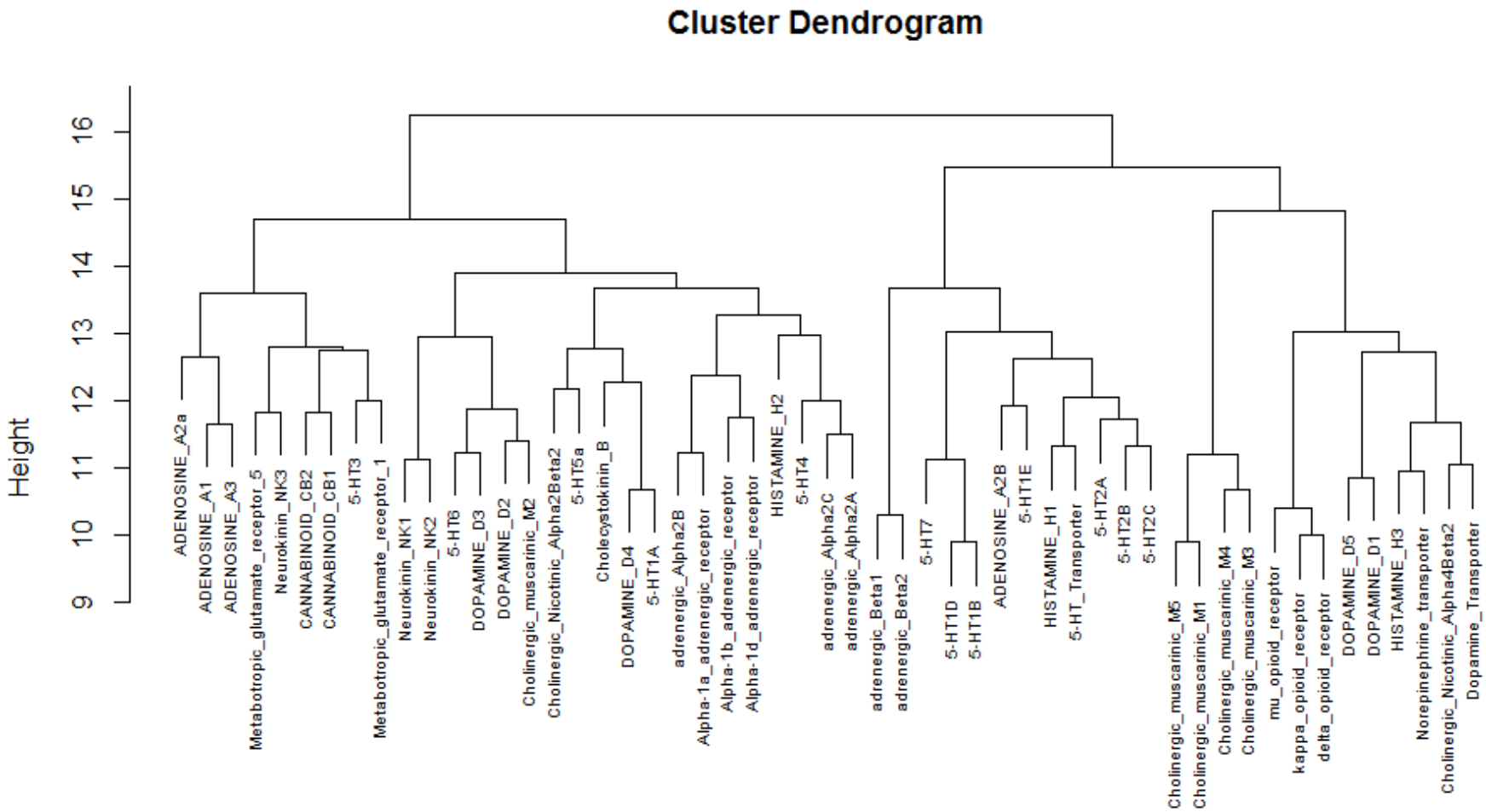

Figure 10 - Cluster dendrogram showing how the family of receptors are related in terms of descriptors

4.4. Discussion

The comparison between the results of the training set and the validated set shows a small difference of around 0.04 maximum in terms of RMSE values for the same neuroreceptor. Figure 11 shows that the values of RMSE are concentrated around the line of equality. The percentage of variance explained has the biggest difference of around 50% in one of the models. Others present some differences, but, in general, most of the models have similar values of variance and dispersion of the values of binding affinity (figure 12). The correlation values are similar in figure 13, the points are close to each other meaning that the predicted model has a strong linear relationship. Four of the models for the adrenergic neuroreceptors have the biggest difference in terms of percentage of the explained variance, RMSE and correlation values when compared to the results of the modelling part and the results of the validation part for the combined methods Random Forests and SVM. Due to the fact that the datasets used for these models are small and number of descriptors chosen are less than 25 could contribute to that difference, which means the receptor does not represent the dataset well [10]. The models for Cholinergic and Nicotinic Alpha4Beta2 showed a big difference between trained model and validation, in terms of percentage of variance explained. In this instance the problem may be related to the size of the dataset, because this model was the smallest with 112 binding affinity values [10].

Most of models with small datasets appear to have the poorest results in terms of RMSE, PVE and correlation, because when the training dataset is not large enough, the data collected may not reflect the complete property space [10]. Therefore, the model cannot be used to confidently predict the binding affinity values.

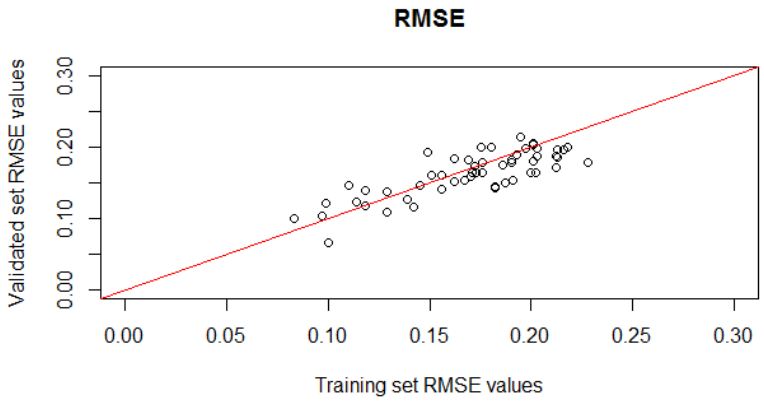

Figure 11 – Plot of the values from the training set and the validated set in terms of RMSE.

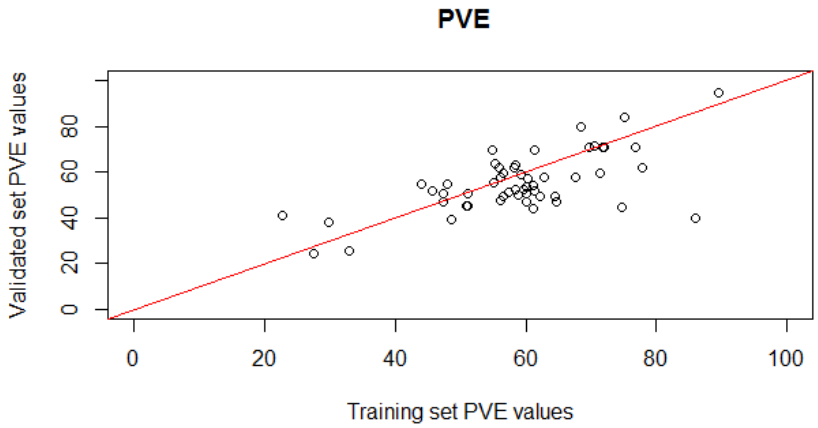

Figure 12 - Plot of the values from the training set and the validated set in terms of percentage of variance explained (PVE).

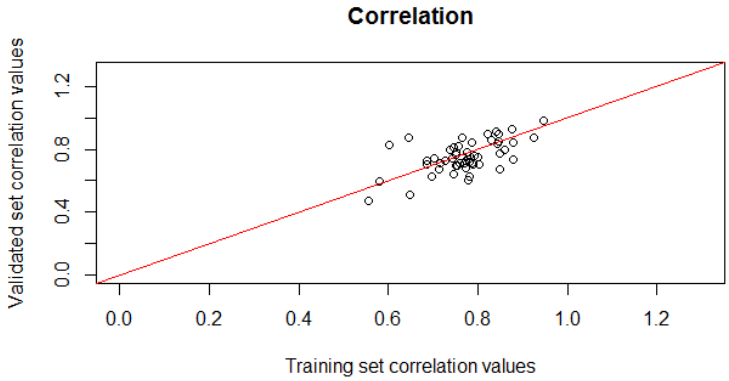

Figure 13 - Plot of the values from the training set and the validated set in terms of Pearson's correlation coefficient.

Through the method of Random Forests, the most important descriptors were selected. Molecular fingerprints (figure 9) is one of the most used molecular descriptors for the models. The fingerprints generate a pattern for each atom, the nearby atoms and bonds between them, the number produced for a particular molecule can be easily managed by the computer because the descriptors are selected in models [10, 19, 24].

There are six descriptors that were selected in every model (figure 9), three of them are constitutional descriptors (sbonds, abonds and MW). The constitutional descriptors are the most important descriptors for the model because these type of descriptors are the most common and simple, which gives an idea of the chemical composition of a compound without any information about its molecular geometry or atom connectivity [10].

The other descriptors TPSA, LogP and MR describe lipophilic, polarity and hydrophobic properties; these properties are related to bioavailability of drugs. The design and manufacturing process is partially responsible for the bioavailability; if the design of the drug is successful, the bond between the drug and the target is stronger [19, 52]. The TPSA descriptor is the sum of solvent-accessible surface areas of atoms, with an absolute value of partial charge greater than or equal to 0.2. TPSA is related with the activity of receptors [53, 54]. The LogP value of a molecule is the sum of the fragment values that are present in the molecule, which is important because the connection between the target and drug happens through those fragments [53]. Molecular refraction is related to the activity of the compounds [53]. Those descriptors are important because it gives an idea of the connections between the receptors and targets which can help to predict the binding affinity.

5. Conclusion and future work

For this work, a database was created using different types of data sources with binding affinity values. Three machine learning methods, Random Forests, support vector machines and LASSO were used to build a model with help of QSAR/QSPR models. The machine learning methods were combined in pairs. First, the best descriptors were selected in order of their importance for describing the compound, in this way it is possible to predict the binding affinity value between the neuroreceptor and the ligand. Second, the descriptors were added until the best model was found – that with the lowest RMSE. The most accurate model with lowest RMSE in the training process was Random Forests with 500 trees combined with SVM through 5-fold cross-validation. The model was validated with an independent dataset (test set). A model can be identified to have a good fit if the values of RMSE are below 0.30, more than half of the receptors have the percentage of variance explained above 50% and almost all the receptors have the correlation value above 0.50. To have a good model, the dataset needs to have more than 112 entries. The values of RMSE between the training set and the validated set were similar. The values of RMSE for the best models were between 0.087 and 0.201 where the PVE is above 50% and correlation above 0.50. Although there were some models with poor results, the dataset of those models was small.

There are 46 descriptors that were selected in at least 20 models showing that they are important to predict the binding affinity. From those descriptors, six were chosen in all the models: Molecular Weight, Molar Refractivity, TPSA Polar Surface Area, the logarithm of the octanol/water partition coefficient, number of single bonds and the number of aromatic bonds. These are the most important descriptors for the prediction of the binding affinity.

The same descriptors are used to identify a family of receptors; the result shows that all neuroreceptors are related. It was also shown that some neuroreceptors which don't belong to the same family are related because some drugs will connect to both neuroreceptors and they can be responsible for the same behaviour.

For future studies, it would be interesting to predict if a ligand connects to a receptor and to know the probability of it connecting to another neuroreceptor. This is important because the drug needs to connect to the right receptor to have the desired effect. This is difficult to test in a laboratory hence why having the above information is useful before doing feature tests in the laboratory.

With this information, we can create a Web platform which can be queried by anyone to find out the possibility of that compound connecting to the neuroreceptors and the value of binding affinity [34].

6. Bibliography

1. Donald Voet, Judith G. Voet, 2004, *Biochemistry*, Editora John Wley and Sons, 3ª edição.
2. Habib, Michel, *Bases neurológicas dos comportamentos*, Climepsi Editores, 1ªedição; Abril 2003, Lisboa.
3. Lullmann,H., et al,*Color Altas of Pharmacology*, Thieme, 3ªedição,2005, New York.
4. Hogg, C., R.; *Nicotinic acetylcholine receptors: from structure to brain function*; Rev Physiol Biochem Pharmacol; Springer-Verlag; 2003
5. Hille, Bertil; *Ion Channels of Excitable Membranes*; 3ª Edition; Sinauer Associates, Inc; 2001, Sunderland, Massachussetts
6. Porter RJ, Dhir A, Rogawski M, Macdonald R.;*AED mechanisms and principles of drug treatment*; In Stefan H, and Theodore W (Ed): Handbook of Clinical Neurology, 3rd series, Epilepsies Part 2: Treatment. Elsevier, Edinburgh, 2012
7. Lodish, Harvey , et al, *Molecular Cell Biology*,5ª edição,Freeman.
8. Holger Gohlke, Gerhard Klebe; *Statistical potentials and scoring functions applied to protein–ligand binding*; Elsevier Science Ltd, 2001.
9. Iwama Hisakazu and Gojobori Takashi, *Molecule Biology and Evolution*, 2002.
10. Teixeira, L., Ana; Leal, P., João; Falcão, O., André; *Random Forests for feature selection in QSPR Models - an application for predicting standard enthalpy of formation of hydrocarbons*; Journal of Cheminformatics;2013.
11. Puzyn, Tomasz et al, *Recent Advances in QSAR Studies methods and Applications*; Challenges and advances in computational chemistry and physics volume 8;Springer;
12. Peter Agre; *The Aquaporin Water Channels*; American Thoracic Society;2006
13. Consonni V,et al; *Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. Part 1. Theory of the novel 3D molecular descriptors*; J Chem Inf Comput Sci; 2002
14. Oprea, I.,Tudor; *On the Information Content of 2D and 3D Descriptors for QSAR*; Journal of the Brazilian Chemical Society; 2002
15. Hulme, C., Edward et al; *Ligand binding assays at equilibrium: validation and interpretation*; British Journal of Pharmacology; 2010

16. Allen, A., John el al; *Strategies to Discover Unexpected Targets for Drugs Active at G Protein-Coupled Receptor*; Annu. Rev. Pharmacol. Toxicol; Annual Reviews;2011
17. Hutchison, R. Geoffrey et al, *Open Babel Documentation Release 2.3.1,* December 05, 2011.
18. Beisken., Stephan; et al, *KNIME-CDK: Workflow-driven chemoinformatics*; BMC Bioinformatics; 2013
19. Yap et al; *Software News and Update PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints*; Wiley Periodicals, Inc.; 17 December 2010.
20. Stalring, C. , Jonna et al; AZOranfe – *High performance open source machine learning for QSAR modeling in graphical programming environment*; Journal of Chemiformatics; Chemistry Central Ltd; 2011.
21. Cox, Richard et al; *QSAR workbench: automating QSAR modeling to drive compound design*; J Comput Aided Mol Des; Springer;2013
22. Sun, Xian-qianf; *Structure based ensemble-QSAR model: a novel approach to the study of the EGFR tyrosine kinase and its inhibitors*; Acta Pharmacologica Sinica; CPS and SIMM; 2014
23. Spjuth, Ola et al; *Applications of the InChi in cheminformatics with the CDK and Bioclipse*; Journal of Chemiformatics; Chemistry Central Ltd; 2013
24. Craig, A. J.; Weininger, D.; Delany, J.; *In Daylight Theory Manual*; Daylight Chemical Information Systems, Inc., 2005. Available at http://www.daylight.com/dayhtml/doc/theory/index.html
25. Xue,X.,C., et al; *QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine*; J.Chem.Inf.Comp.Sci; American Chemical Society; 2004
26. Cherkasov, Artem; *QSAR Modelling: Where have you been? Where are you going to?;* J Med Chem. Author manuscript; June 2015
27. Butkiewicz, Mariusz et al; *Benchmarking Ligand-Based Virtual High-Throughput Screening with the PubChem Database*; Molecules.Author manuscript; 2013
28. Frimayanti, Neni et al; *Validation of Quantitative Structure-Activity Relationship (QSAR) Model for Photosensitizer Activity Prediction*; International Journal of Molecular Sciences; 2011.
29. Guha, Rajarshi; *Using the CDK as a backend to R*;CDK news, vol 2, 1 March 2005

30. Gohike, Holger and Klebe, Gerhard; *Statistical potentials and scoring functions applied to protein-ligand biding*; Elsevier Science Ltd. ;2001
31. Hastie,Trevor et al;*The Elements of Statistical Learning, Data Mining, Inference, and Prediction*;2ª edition; Springer Series In Statistics;2008.
32. Besnard, Jérémy et al; *Automated design of ligands to polypharmacological profiles*; Macmillan Publishers Limited; 2012.
33. Ray, S., Thomas; *Psychedelics and the Human receptorome*; PloS ONE; 2010.
34. Martins, F. Inês; Texeira, L., Ana; Pinheiro, Luis; Falcão, O., André; *A Bayesian Approach to in Silico Blood- Brain Barrier Penetration Modeling*; Journal of Chemical Information and Modeling; 2012.
35. Mitchell, B., O., John; *Machine learning methods in chemoinformatics*; John Willey & Sons,Ltd; 2014.
36. Hajjo R, et al; *Development,Validation, and Use of Quantitative Structure-Activity Relationship Models of 5- Hydroxytryptamine (2B) Receptor Ligands to Identify Novel Receptor Binders and Putative Valvulopathic Compounds Among Common Drugs*; J. Med. Chem; 2010
37. Chai, T. et al; *Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature*; Geosci. Model Dev.; 2014
38. https://en.wikipedia.org/wiki/Root-mean-square_deviation
39. Luo, Man et al; *Application of Quantitative Structure – Activity Relationship Models of 5-HT$_{1A}$ Receptor Binding to Virtual Screening Identifies Novel and Potent 5-HT$_{1A}$ Ligands*; Journal of Chemical Information and Modeling; 2014
40. Procopio, Michael J.; *An Experimental Analysis of Classifier Ensembles for Learning Drifting;* ProQuest Information and Learning Company; 2009.
41. https://en.wikipedia.org/wiki/Mean_squared_error
42. Yugandhar K., and Gromiha, M., Michael; Feature selection and classification of protein–protein complexes based on their binding affinities using machine learning approaches; Proteins; WILEY PERIODICALS, INC.;2014
43. Zhang L., et al; A. *Discovery of Novel Antimalarial Compounds Enabled by QSAR-Based Virtual Screening*. J. Chem. Inf. Model. 2013.
44. Benigni, Romualdo; *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*; Taylor & Francis Group; 2003.
45. http://mathbits.com/MathBits/TISection/Statistics2/correlation.htm

46. Cassotti, M ; Grisoni, F; *Variable selection methods:an introduction; Molecular Descriptors*; http://www.moleculardescriptors.eu/
47. Benfenati, Emilio; *Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes*;Elsevier;2007
48. Zhu, Ruixin et al; *Investigations on Inhibitors of Hedgehog Signal Pathway: A Quantative Estructure – Activity relationship Study*; International Journal of Molecular Sciences; 2011
49. Kelland MD et al; *The modulation of dopaminergic neurotransmission by other neurotransmitters*. New York: CRC Press; 1996.
50. Parastoo Hashemi et al; *In Vivo Electrochemical Evidence for Simultaneous 5-HT and Histamine Release in the Rat Substantia Nigra pars Reticulata Following Medial Forebrain Bundle Stimulation*; Journal of Neurochemistry; PCM,2012.
51. Rex, André et al; *Cortical 5-HT-CCK interactions and anxiety-related behaviour of guinea-pigs: a microdialysis study*; Neuroscience letters; Elsevier Science Ireland Ltd; 1997
52. http://www.merckmanuals.com/professional/clinicalpharmacology/pharmacokinetics/drug-bioavailability
53. Todeschini, Roberto and Consonni, Viviana; *Handbook of Molecular Descriptors: Methods and Principals in Medical Chemistry*; Wiley-VCH Verlag GMbH; Weinheim (Federal Republic of Germany); 2000.
54. S. Prasanna and R. J. Doerkse; T*opological Polar Surface Area: A Useful Descriptor in 2D-QSAR*; Bentham Science Publishers Ltd; 2009.
55. Kaur, H., et al; *Computational Analysis and In silico Predictive Modeling for Inhibitors of PhoP Regulon in S. typhi on High-Throughput Screening Bioassay Data set*; Interdisciplinary Sciences: Computational Life Sciences; Springer; August 2015.
56. Temerinac-Ott, M., et al; *Deciding when to stop: efficient experimentation to learn to predict drug-target interactions*; BMC Bioinformatics; 2015
57. Naga Nagisetty et al; *Building a knowledge base to assist clinical decision-making using the Pediatric Research Database (PRD) and machine learning: a case study on pediatric asthma patients*; BMC Bioinformatics; 2014
58. Sonam Gaba et al; *Cheminformatics Models for Inhibitors of Schistosoma mansoni Thioredoxin Glutathione Reductase;* The Scientific World Journal; Hindawi Publishing Corporation; 2014

7. Appendix

Code used in the LASSO method:

```
cv<-5
  K<-1
  N<-nrow(pro)
  sq<-as.integer(seq(1,cv+1,length.out=N+1))[1:N]
  cross<-sample(sq, nrow(data),replace=T)


  for(K in 1:cv){
    #remove qualitative variables and transform the data
frame in matrix
    treino<-pro[cross!=K,]
    teste<-pro[cross==K,]

    Y<-as.matrix(treino$Ki_value)
    #remove the select variable in teste
    rv<-treino$Ki_value
    X<-as.matrix(treino[,-rv])


    #analyse
    mdlrec<-glmnet(X,Y,family="gaussian")

    #penalty
    spot<-which.min(mdlrec$dev.ratio)

#prediction
       for(spot in 1:length(mdlrec$df)) {
      preds<-predict(mdlrec,Xtest,s=mdlrec$lambda[spot])
      rmse<-sqrt(mean((preds-Ytest)^2))
    }
```