Universidade Lusófona de Humanidades

Escola de Ciências e Tecnologias da Saúde

Development of a QSAR Models for the Prediction of Plasma Protein Binding

João Pedro da Silva Gonçalves Poiares

Orientador: Professor Paulo Paixão

Lisboa

2014

Development of a QSAR Models for the Prediction of Plasma Protein Binding

Dissertação de Mestrado Integrado em Ciências Farmacêuticas apresentada na Universidade Lusófona de Humanidades e Tecnologias/ Escola de Ciências e Tecnologias da Saúde

João Pedro da Silva Gonçalves Poiares

Orientador: Professor Paulo Paixão

Lisboa

2014

Acknowledgment

A number of people have kindly contributed (although I am solely responsible for any flaws and inconsistencies) during the process of the development of this thesis and thus a million thanks are due!

First, I would like to thank all the people that were involved in my work. Namely, Professor Paulo Paixão for always being available and helping when needed, and for providing all the necessary feedback for the development of this thesis.

Since the development of this theses lasted for six months, there are other people that indirectly influenced my work, and whom I have to thank.

First of all, my girlfriend for always being by my side and giving me all the support and encouragement I needed. Second, my family for always supporting my decisions and for providing me with all the resources I needed not only for the development this thesis, but also for my development as a human being. Without them I would not be the same person.

Abstract

One of the most important factors, affecting the pharmacokinetic profile of a drug is binding to plasma protein. As such, this study aimed at the development of a quantitative structure–activity relationship model, to predict the fraction unbound in plasma (f_{ub}) for four species, using artificial neural network ensemble (ANNE).

To this end a database of 363 drugs was used, and molecular descriptors were determined. The dataset was divided in two groups, a train and an external validation, to avoid overfitting. The ANNE optimization reduced the descriptors required to determine the f_{ub} to 37, and 150 ANN were randomly selected, trained and the optimal configuration was collected. The different ANNE were built by averaging the output values of the selected ANN and the best ANNE was selected.

The model created was able to predict, with a small amount of error, the f_{ub} values (root mean square error of 0.16798 and 0.193705 for train and test dataset respectively), however, it tends to underestimate this value (mean error of -0.00291 and -0.015780 for train and test dataset respectively). The ANNE interpretation showed that the main characteristics of that affect f_{ub} were the molecule charge, size, structure and lipophilic and hydrophilic affinity.

Key-words: ANNE, ANN, plasma protein binding, QSAR, fraction unbound in plasma

Resumo

Um dos factores mais influentes na farmacocinética de um fármaco é a ligação às proteínas plasmáticas. Sendo assim, com este estudo pretendeu-se desenvolver um modelo QSAR, para prever facção do fármaco livre no plasma (f_{ub}) para quatro espécies, usando um "*ensamble*" de redes neuronais (ANNE).

Para tal, utilizou-se uma base-de-dados de 363 fármacos, e determinou-se os seus descritores moleculares. Esta base-de-dados foi dividida em dois grupos, um para treino e outro para validação externa, para evitar "*overfitting*". O ANNE foi optimizado, reduzindo o número de descritores para 37, e 150 redes foram aleatoriamente selecionadas, treinadas e a sua configuração optimizada registada. Os diversos ANNE foram obtido através da média aritmética dos valores das redes seleccionadas, e o melhor ANNE foi escolhido.

Este modelo foi capaz de prever com um erro reduzido, o valor da f_{ub} (erro quadrático médio de 0.16798 e 0.193705 para o grupo de treino e teste respectivamente), no entanto tendencialmente subestima o seu valor (erro médio de -0.00291 e -0.015780 para o grupo de treino e teste respectivamente). A interpretação do modelo permitiu observar que o tamanho da molécula, a sua estrutura, carga, lipofilia e hidrofilia são as características que mais afectam o valor da f_{ub}.

Palavras-chave: ANNE, redes neuronais, ligação a proteínas plasmáticas, QSAR, fração livre no plasma

Table of Contents

Table of Figures and Tables

Abbreviations

AAG - α1-Acid Glycoprotein

ADME - Absorption, Distribution, Metabolism and Excretion

ANN – Artificial Neural Network

ANNE - Artificial Neural Network Ensemble

ANOVA - Analysis of variance

F_{ub} - fraction unbound in plasma

PPB – Plasma Protein Binding

QSAR - Quantitative structure–activity relationship

RMSE - Root Mean Squared Error

SMILES - Simplified molecular-input line-entry system

SVM - support vector machines

Pharmacokinetics and Drug Fraction Unbound in Plasma

Pharmacokinetics is the science that provides a mathematical basis to assess the time course of drugs in the body, as such it incorporates the processes of absorption, distribution, metabolism and excretion (ADME) [1] [2] [3].

The effectiveness of a drug dosage is determined by the concentration of the drug in the body, and ideally, it should be measured at the site of action of the drug. To understand the importance of the drug substance and its formulation on absorption, and distribution of the drug to the site of action, one must first consider the sequence of events that precede its therapeutic effect, as described in Figure 1 [1] [2].

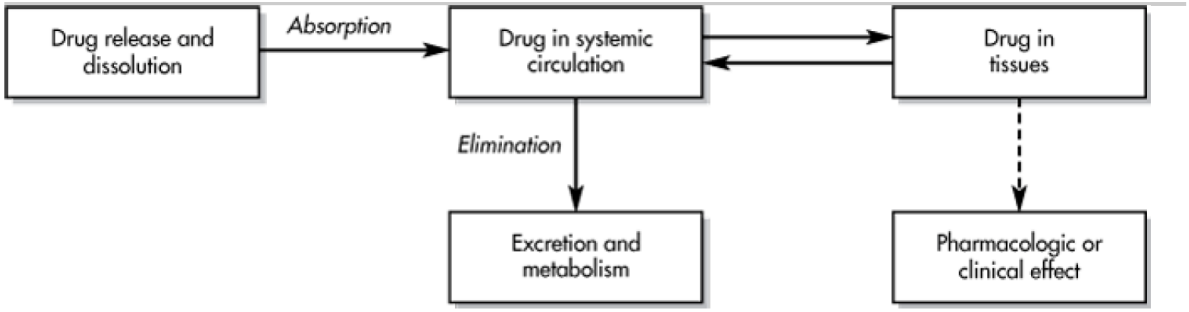

Figure 1 - Scheme demonstrating the dynamic relationship between the drug, the drug product, and the pharmacologic effect [2]

As consequence, a fundamental understanding of these parameters is required to design an appropriate drug regimen for a patient, or even to design a new drug for testing, as almost 39% of the costly late-stage failures in drug development until the late 1990s were caused by poor pharmacokinetics, and even in recent years, about 15% of failures in drug development are due to this reason [1] [2] [4] [5].

However, since measuring the concentration of a drug in its site of action is difficult, concentrations are normally measured in whole blood from which serum or plasma is generated. The simplest pharmacokinetic concept is that based on total drug in plasma, however, drug molecules may be bound to a greater or lesser extent to the proteins present within the plasma, thus free drug levels may be vastly different from those of total drug levels. Additionally, only the unbound fraction of the drug can produce the desired pharmacological effect and be metabolized [1] [3].

Plasma protein binding (PPB) of drugs is expressed as percentage of total drug that is bound to plasma protein such as albumin, α1-acid glycoprotein (AAG) and lipoproteins. A number of acidic drugs have high affinity for sites on blood proteins, particularly albumin, whereas the main binding protein for many basic drugs is AAG. Each protein have its own proprieties, their concentration in plasma may vary depending on gender, age and health state and they can contribute simultaneously to the binding of the drug [6] [7] [8] [9].

PPB is a reversible association of a drug with the proteins of the plasma due to hydrophobic and electrostatic interactions such as van der Waals and hydrogen bonding. The unbound drugs can passively diffuse through the barriers into the organs where they are

metabolized, biliary excretion or glomerular filtration in kidney, and to the sites where they interact with therapeutic targets to produce therapeutic effects [6] [7].

Since numbers of protein-binding sites are limited, competition will exist between two drugs, and the drug with higher affinity will displace the other, causing increased free drug. The protein-binding properties are related to plasma clearance, elimination half-life, apparent volume of distribution and area under curve [10].

QSAR for Fraction Unbound in Plasma

As consequence of what was written in the previous chapter, the analysis of the PPB drug capability is a vital attribute for the assessment of the drug's pharmacokinetics and pharmacological effects and also the toxicity risk that the drug may pose [6] [7] [11].

There are several *in vitro* assays that can be used to determine the extent of plasma protein binding, such as equilibrium dialysis, ultrafiltration, ultracentrifugation, chromatographic methods, fluorescence spectroscopy, ultraviolet spectroscopy, circular dichroism, nuclear magnetic resonance spectroscopy, and capillary electrophoresis [6] [12] [13].

However *in vitro* and *in vivo* ADME are relatively expensive in terms of resources, reagents and detection techniques, therefor there is a need for reliable *in silico* technique to predict PPB of virtual compounds in order to avoid the synthesis of chemicals which do not have the potentiality of being approved drugs, and also to gain insights into the chemical nature of drug–protein interactions [6] [4] [9].

One possible approach is the quantitative structure-activity relationship models (QSAR), which are computer-based models for the prediction of toxicological, biological and physico-chemical properties. They aim at establishing, if it exists, a mathematical relationship between structural-derived properties of chemicals and their experimental properties, such as toxicity [14] [15] [16].

Since a QSAR model is a methodology that starts with generated descriptors based on molecular structures and uses computational algorithms to relate the key descriptors to the dependent property value of interest, a number of QSAR studies on the analysis and prediction of PPB have been performed during the last two decades. Several molecular descriptors were used, such as constitutional, topological, electrotopological, physicochemical, quantum chemical descriptors and a variety of mathematical models were also used, such as multiple linear regression with variable selection, artificial neural networks (ANN) and support vector machines (SVM) [4] [9].

Nevertheless, prediction of binding affinity of drugs to plasma proteins is a rather complicated task, since the quality of the model depends on the quality of the dataset, which in this case can be a tricky task because like other pharmacokinetic parameters, binding affinity data vary significantly from report to report as a result of differences in methodology, experimental conditions, and mathematical approaches [4].

ANN for Developing a QSAR

Applying machine learning algorithms to QSAR models has long been done, it started with the used linear regression models, but these were quickly supplanted by Bayesian neural networks. The use of ANN for QSAR models appeared in the 90s, and has developed into a well- established scientific area with numerous ideas, theoretical approaches, and successful practical applications [17] [18] [19].

ANN are powerful non-linear models for classification, regression, or dimensionality reduction, which maps input vectors to output vectors with repeated compositions of simpler modules called layers. The internal layers re-represent the input and learn features of the input useful for the task [17] [20].

Deep neural networks, or neural networks with multiple hidden layers, have recently been highly successful in numerous applications because they are capable of learning complicated, rapidly-varying non-linear functions and are also capable of extracting a hierarchy of useful features from their input [17]

Non-Bayesian ANN have also been applied to QSAR, initially with a single small hidden layer, but more recently, a tremendous improvements in training methods for deep and wide neural networks as well as a renewed appreciation for the advantages of deeper networks as been done [17].

In practice, ANNs are used for solving so-called ill-posed problems, for which numerous alternative solutions can be suggested. These problem exactly correspond to tasks performed by NN in an absolute majority of QSAR studies, by the means of learning or training, which can be supervised, unsupervised, or reinforced [18].

The advantages of ANN over statistical estimation technique is that no *a prior* knowledge of underlying statistical nature of problem is required and no simplifying assumption need to be made for application of this technique in a sparse data environment [21].

However, it is often difficult to decode the final model to identify the changes to molecular structure needed to obtain a desired property, and this mathematical model has a tendency to ‘memorize’ rather than learn and are particularly susceptible to over-fitting, especially if the training data is noisy [20].

Objective

This study aims at the development of a QSAR model to predict plasma protein binding, for four different species, human, dog, rat and monkey, using ANN software, and using Microsoft Excel to building an ANNE.

Materials and Methods

Data Base of Fraction Unbound in Plasma

The QSAR model, was created based on a data set of intravenous pharmacokinetic data from human, rat, dog, and monkey for 363 compounds, previously developed by Franco Lombardo and his colleges [22].

The available fraction unbound in plasma (f_{ub}) information was collected for the four species. A multifactor ANOVA was preformed to both data sets, using StatPoint Statgraphics Centurion v15.2.11., where the two factors analysed were molecule and species. A statistically significant difference was observed between the different species, and as consequence, a binary input was added to the model to describe each specie (Table 1).

Table 1 - Species binary code

Human	0	0
Monkey	0	1
Rat	1	0
Dog	1	1

Calculation of the Molecular Descriptors

The following methodology was used for the calculation of the molecular descriptors: SMILES notation of each molecule was obtained using the on-line PubChem Compound database (http://www.ncbi.nlm.nih.gov), ionization descriptors, i.e. acid pKa and base pKa, were predicted by using ChemAxon (http://www.chemicalize.org), additionally, for molecules without an acid group a value of 15 was attributed to the acid pKa, and for drugs without a basic group, a value of -1 was to the base pKa. The remaining descriptors, related to size, hydrogen bonding potential, lipophilicity and others, were obtained from the on-line E-Dragon 1.0 software using CORINA to convert the SMILES notation to the 3D representation of the molecule. From all the molecules in the database, twelve were excluded, since the E-Dragon Software didn't accept a molecules with that size (the molecules were to big for the software). At the end of the creation of this database, for each one of the 351 molecules, 1670 descriptors were created.

ANNE Optimization

The Artificial Neural Network (ANN) non-linear regression was performed using the backpropagation neural modelling system Qnet for Windows v.2000 build 751 (Vesta Services inc, USA) and an in-house developed Microsoft Excel® VBA procedures for process automation.

Both molecular descriptors and binary descriptors for species, where considered as the ANN input and the f_{ub} values where considered as the ANN outputs. It's important to say at this point that, each molecule had an individual input entry for each specie f_{ub} values.

To allow the calculation of the relative relevance of the molecular descriptors, all networks designed were built using normalised variables both at the input and output, and a sigmoid transfer function was used in all connections.

In order to avoid network overfitting, input space dimension pruning was preformed, and early stopping of training was used, where 25% of the train molecules were randomly selected to act as a sub-set for the internal testing of the model, not effectively used in the regression process, and ANN train was performed until degradation on the RMSE for the internal test data was observed. Furthermore, each network architecture was trained 20 times, with random initial values and different sub-sets of the internal test cases, to avoid training convergence to local minima.

The input space dimension pruning was a two-step process. At first, molecular descriptors where eliminated based on three factors:

- Molecular descriptors classification;
- Molecular descriptors highly correlated (r>0,9);
- Molecular descriptors with highly repetitive values (90%).

The second step consisted of the optimisation of the network structure for the most relevant molecular descriptors. To this end, ANN with one hidden layer, and a ρ value between 2 and 3, where ρ is the ratio between the number of train cases and the number of connections in the network, where trained 20 times, and relative input weights and standard deviation were collected for all networks.

The 10 best networks (networks with smallest RMSE) were selected, and the molecular descriptors with relative weight smaller than the binary code weights where eliminated. This procedures was preformed until no molecular descriptor could be eliminated.

To reduce the input space further, the same procedure as described above for the input weight value was preformed, but this time focused on the standard deviation values.

In the last step of the optimization process, several ANN structures were tested, varying the number of hidden layers (up to three) and the number of hidden neurons in order to obtain a ρ above 1. To this end, 265 ANN were built, and 150 ANN were randomly selected, trained and the optimal ANN configuration for each ANN was collected.

After the optimization process, the predicted output values and input weight information, for each optimal ANN configuration was collected, and the ANN ensemble (ANNE) was built for the best 10, 20, 30, 40, 50, 100 and 150 networks based on the RMSE values.

ANN Validation

External validation was done by comparing the values predicted by the best ANNE (ANNE with smallest RMSE) to the observed f_{ub} values of the drugs in the external validation group of data, not previously used in the training and optimization process.

Results

Data Base of Fraction Unbound in Plasma

The available f_{ub} for the four species were analysed, and a multifactor ANOVA was performed using StatPoint Statgraphics Centurion v15.2.11., where the two factors analysed were molecule and species. The outcome of this analysis can be seen in Table 2, and as result a statistically significant difference was observed between the different species.

Table 2 - Analysis of Variance for Ligação_1 - Type III Sums of Squares

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
MAIN EFFECTS					
A:Espécie	0.215402	3	0.0718008	2.98	0.0314
B:Molécula	68.2189	238	0.286634	11.88	0.0000
RESIDUAL	10.8604	450	0.0241343		
TOTAL (CORRECTED)	79.9453	691			

ANNE Optimization

Regarding the reduction of the input space, the subset of E-Dragon 1.0 software descriptors chosen for this study are displayed in Table 3. Additional a binary input was added to the model to describe each specie. Highly correlated descriptors (r>0.9), and those with highly repetitive values (90%) were eliminated, resulting in the reduction of the input space from 1670 to 281 descriptors.

Table 3 - Table of molecular descriptor classes

Descriptor Class
Constitutional descriptors
Topological descriptors
Walk and path counts
Connectivity indices
Information indices
Topological charge indices
Geometrical descriptors
WHIM descriptors
GETAWAY descriptors

Functional group counts
Molecular properties

The selection of the most relevant descriptors was described under methods, and at the end of the first stage, the input space was still too large, since the number of molecular descriptors was 197. However, after the second approach, the input space size was reduced to 37 descriptors, which was considered a reasonable number of molecular descriptors to build the ANNE, and can be seen in Table 4. The evolution of the RMSE and correlation of e train and test groups during this process of elimination can be seen in Figure 2 and Figure 3.

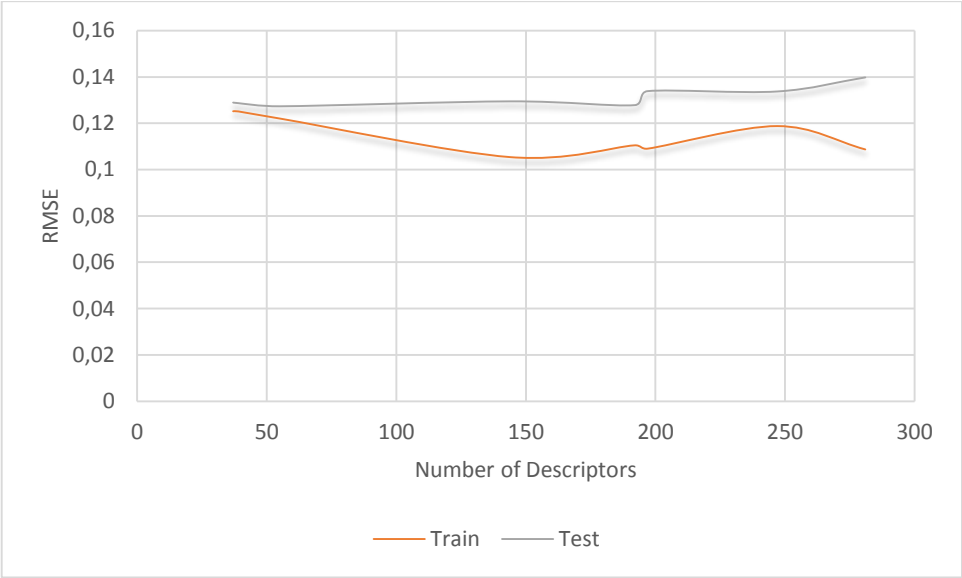

Figure 2 - Evolution of the RMSE with the reduction of molecular descriptor number

Figure 3 - Evolution of the correlation of the predicted and observed values with the reduction of the molecular descriptors

Table 4 - Molecular descriptors at the end of the ANNE optimization

Abreviation	Description	Class
Bin 1	Species binary code 1	Species
Bin 2	Species binary code 2	Species
RBN	number of rotatable bonds	constitutional descriptors
nS	number of Sulfur atoms	
nR06	number of 6-membered rings	
J	Balaban-like index from topological distance matrix	topological descriptors
D/Dr09	distance/detour ring index of order 9	
SRW05	self-returning walk count of order 5	walk and path counts
MPC06	molecular path count of order 6	
IVDE	mean information content on the vertex degree equality	connectivity indices
SIC1	Structural Information Content index	information indices
GGI3	topological charge index of order 3	topological charge indices
JGI4	mean topological charge index of order 4	
JGI5	mean topological charge index of order 5	
JGI6	mean topological charge index of order 6	
DISPe	displacement value / weighted by Sanderson electronegativity	geometrical descriptors
L2u	2nd component size directional WHIM index / unweighted	WHIM descriptors
E1v	1st component accessibility directional WHIM index / weighted by van der Waals volume	
E2v	2st component accessibility directional WHIM index / weighted by van der Waals volume	
G2p	2nd component symmetry directional WHIM index / weighted by polarizability	
Du	D total accessibility index / unweighted	
H7u	H autocorrelation of lag 7 / unweighted	GETAWAY descriptors
HATS2u	leverage-weighted autocorrelation of lag 2 / unweighted	
R3u	R autocorrelation of lag 3 / unweighted	
R1u+	R maximal autocorrelation of lag 1 / unweighted	
R6m+	R maximal autocorrelation of lag 6 / weighted by mass	
R5v+	R maximal autocorrelation of lag 5 / weighted by van der Waals volume	
R5e+	R maximal autocorrelation of lag 5 / weighted by Sanderson electronegativity	
R5p+	R maximal autocorrelation of lag 5 / weighted by polarizability	
nCs	number of total secondary C(sp3)	Functional group count
nRNH2	number of primary amines (aliphatic)	
nRNR2	number of tertiary amines (aliphatic)	
nOHs	number of secondary alcohols	
nArX	number of X on aromatic ring	
ALOGPS_logP	Gthose-Crippen octanol-water partition coeff. (logP)	molecular properties
Pka Base	Basic pKa	
pKa Acid	Acid pKa	

After the input space was optimized, and in order to optimize the structural space, 265 ANN were considered, varying the number of hidden layers (up to three) and the number of hidden neurons in order to obtain a ρ above 1. Of these, 150 ANN were randomly selected, trained and the optimal ANN configuration, based on the lower internal test RMSE for each ANN, was collected.

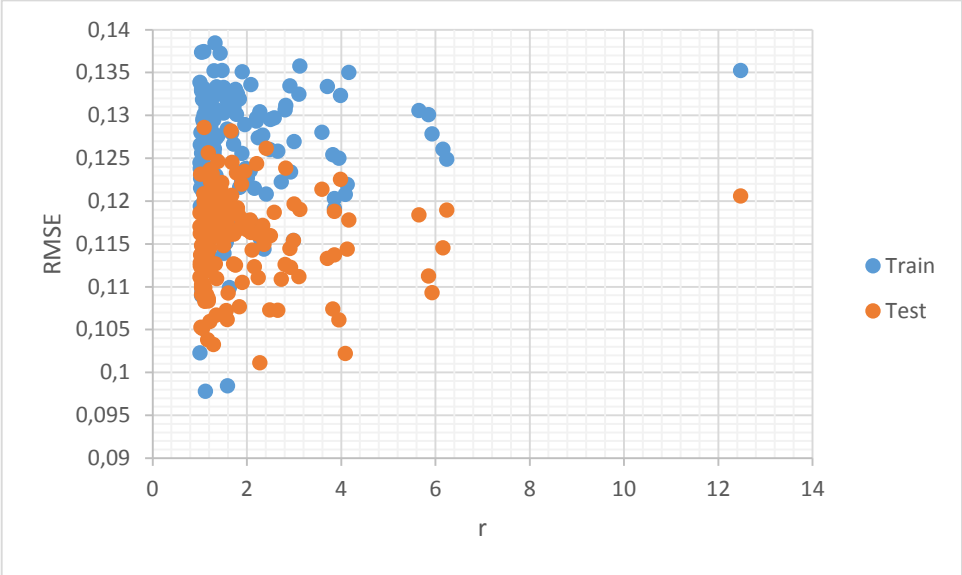

Figure 4 - Performance, measured in terms of the train and internal test RMSE, for the 150 ANN during the optimization procedure. Complexity of each individual ANN is described by the ratio between the number of patterns to the number of connections (r)

At the end of the ANN optimization an ANNE was built by averaging 10, 20, 30, 40, 50, 100 and 150 of the best ANN based on the RMSE values for the test group, and the predicted output (mean and S.D) for each ensemble was collected. The evolution of the RMSE of the mean predicted outputs can be seen in Figure 5, and the comparison of the predicted and observed f_{ub} for each ensemble can be seen in Figures 6 to 12.

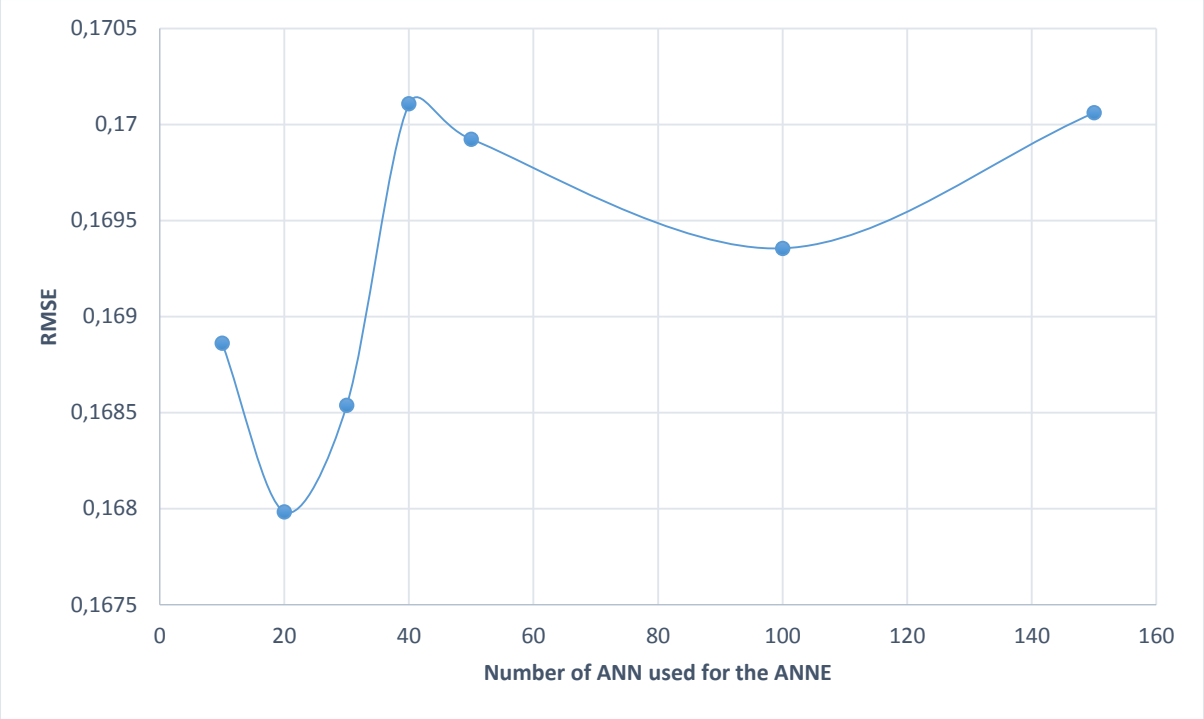

Figure 5 - Evolution of the RMSE for each ANNE

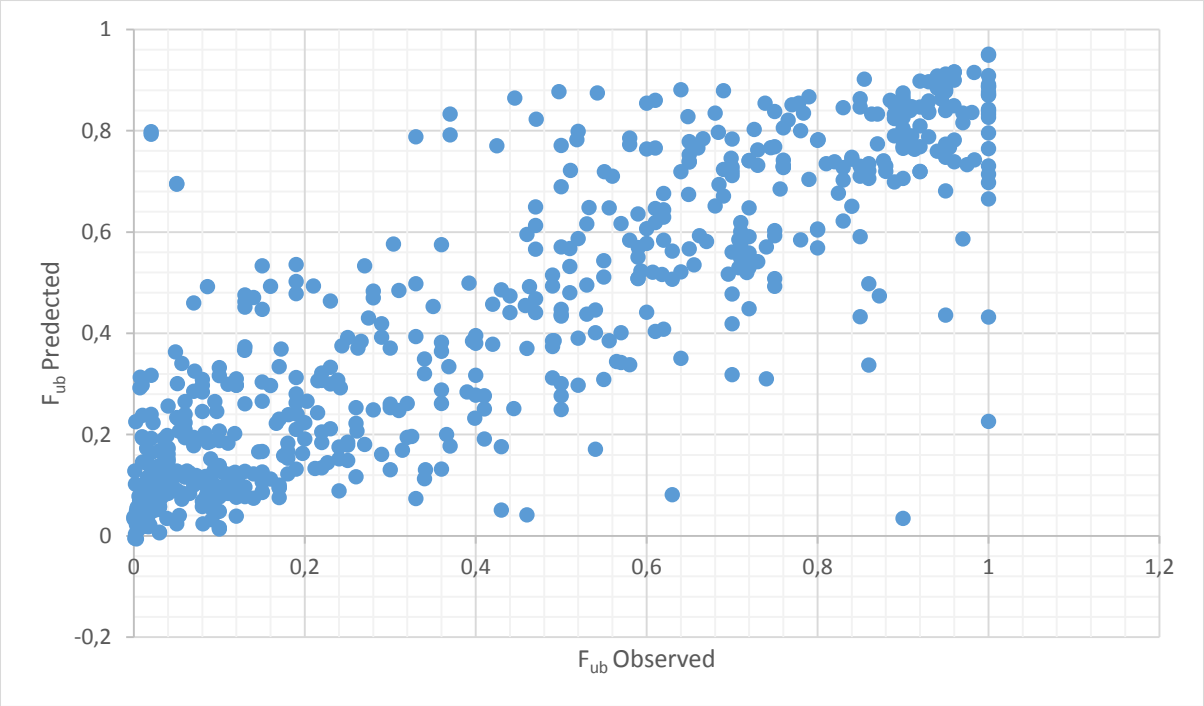

Figure 6 - Plot of the observed vs predicted f_{ub} in the ensemble with the 10 best ANNs

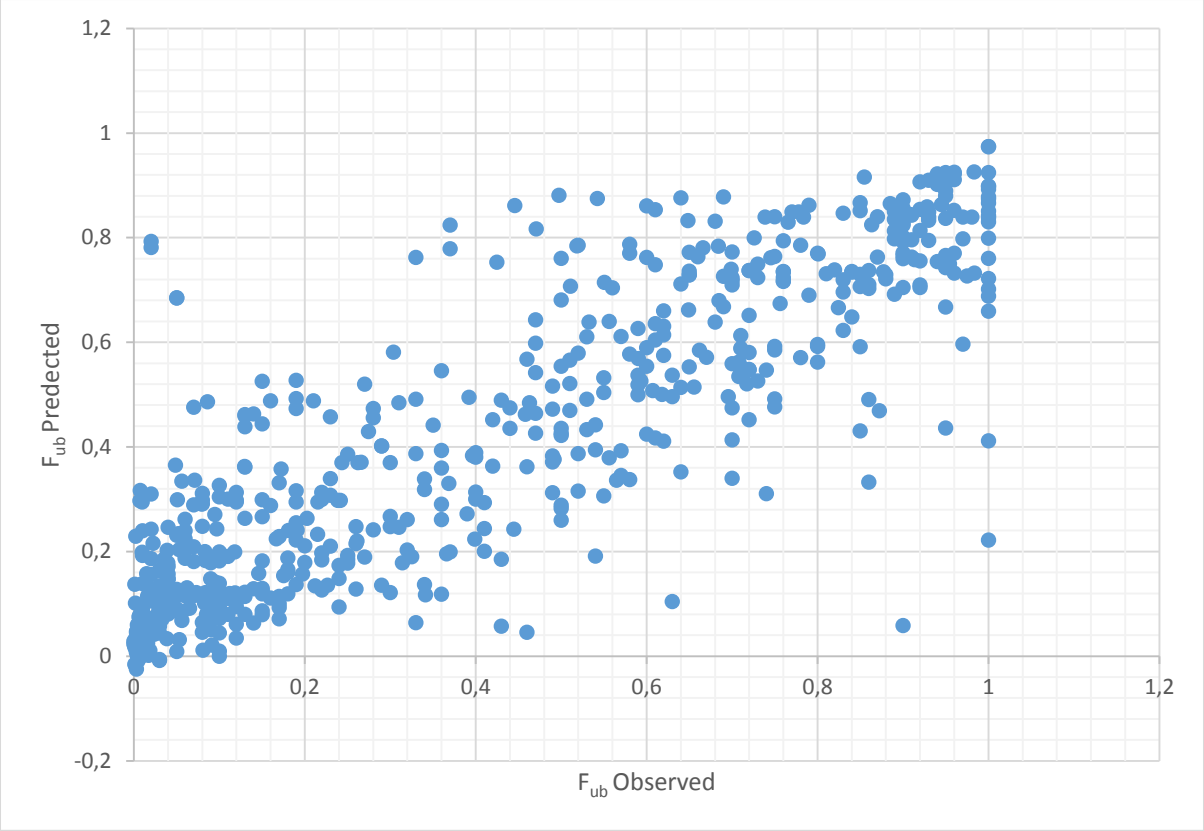

Figure 7 - Plot of the observed vs predicted f_{ub} in the ensemble with the 20 best ANNs

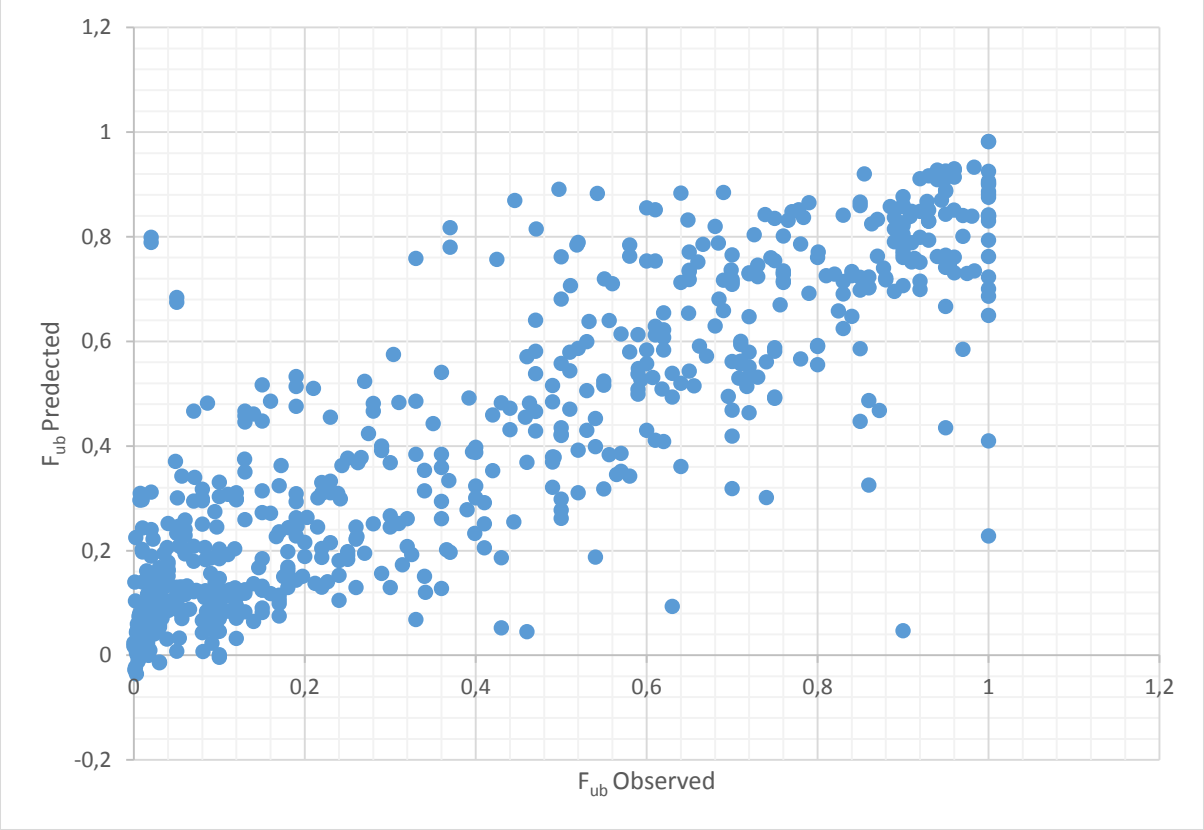

Figure 8 - Plot of the observed vs predicted f_{ub} in the ensemble with the 30 best ANNs

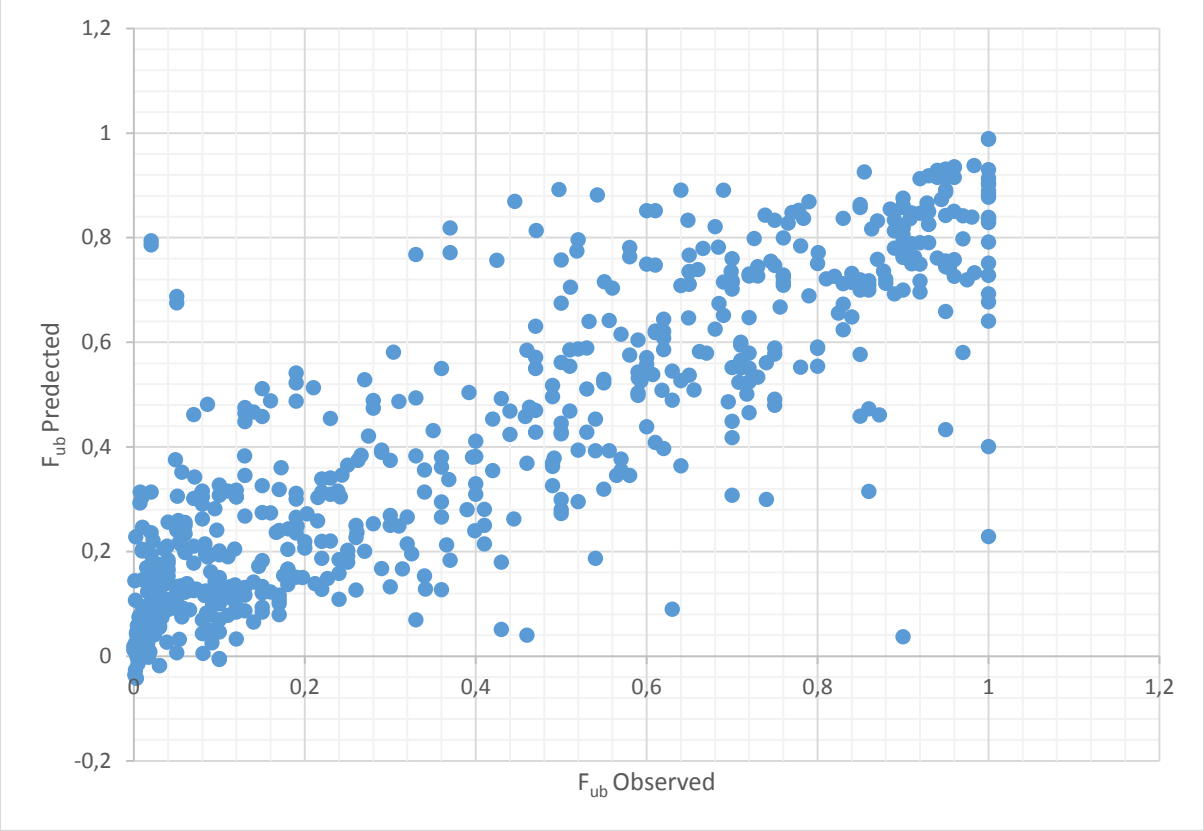

Figure 9 - Plot of the observed vs predicted f_{ub} in the ensemble with the 40 best ANNs

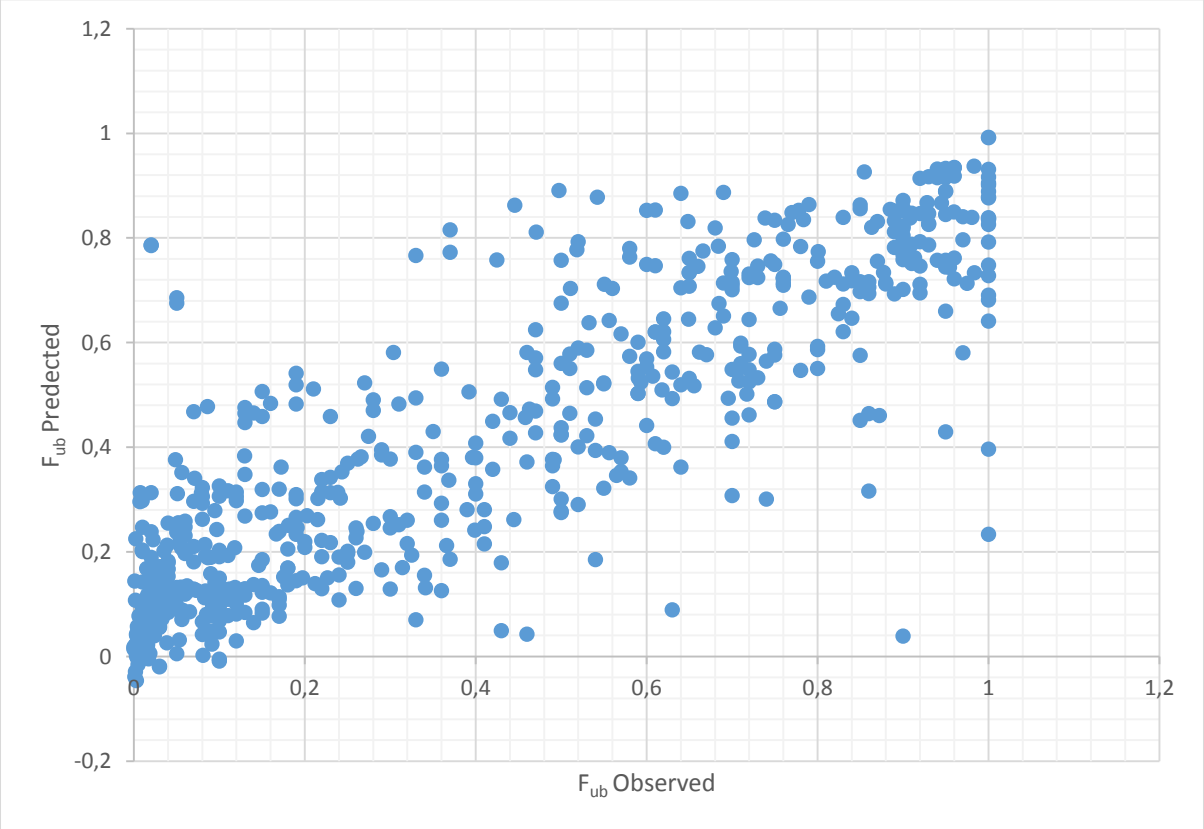

Figure 10 - Plot of the observed vs predicted f_{ub} in the ensemble with the 50 best ANNs

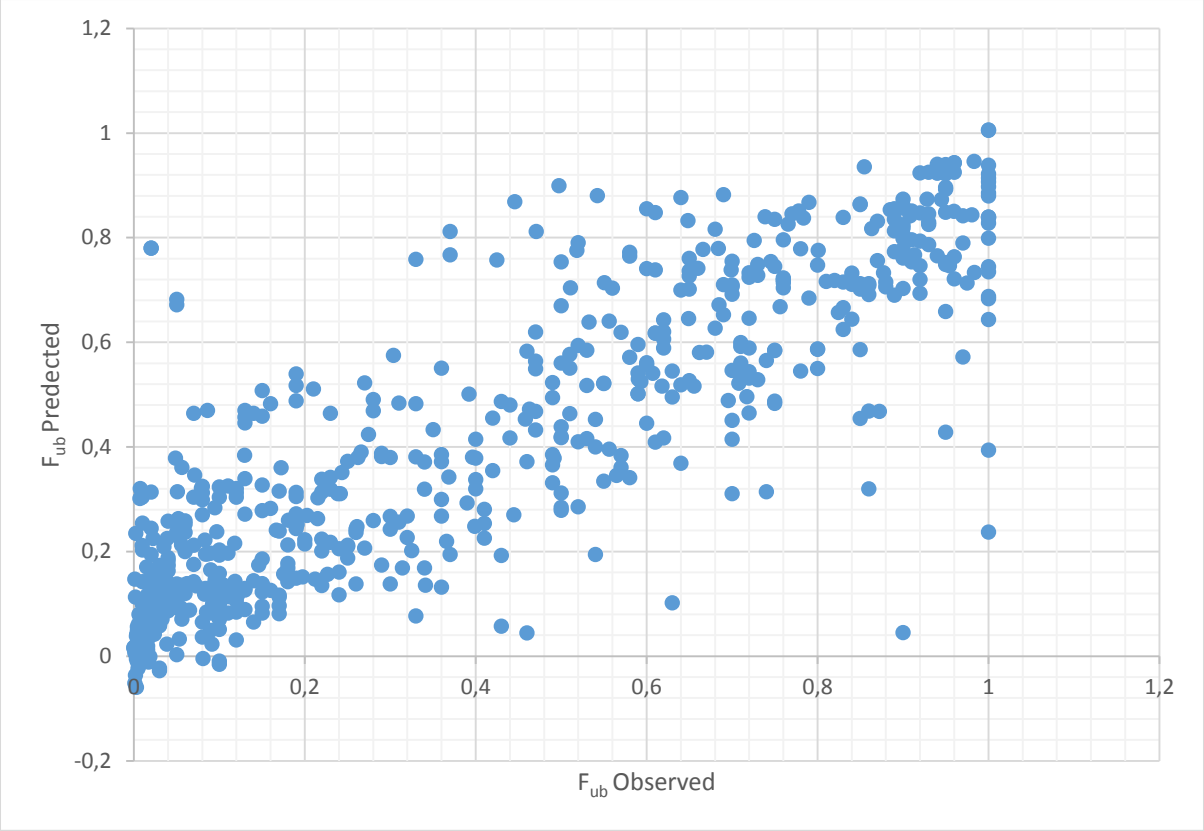

Figure 11 - Plot of the observed vs predicted f_{ub} in the ensemble with the 100 best ANNs

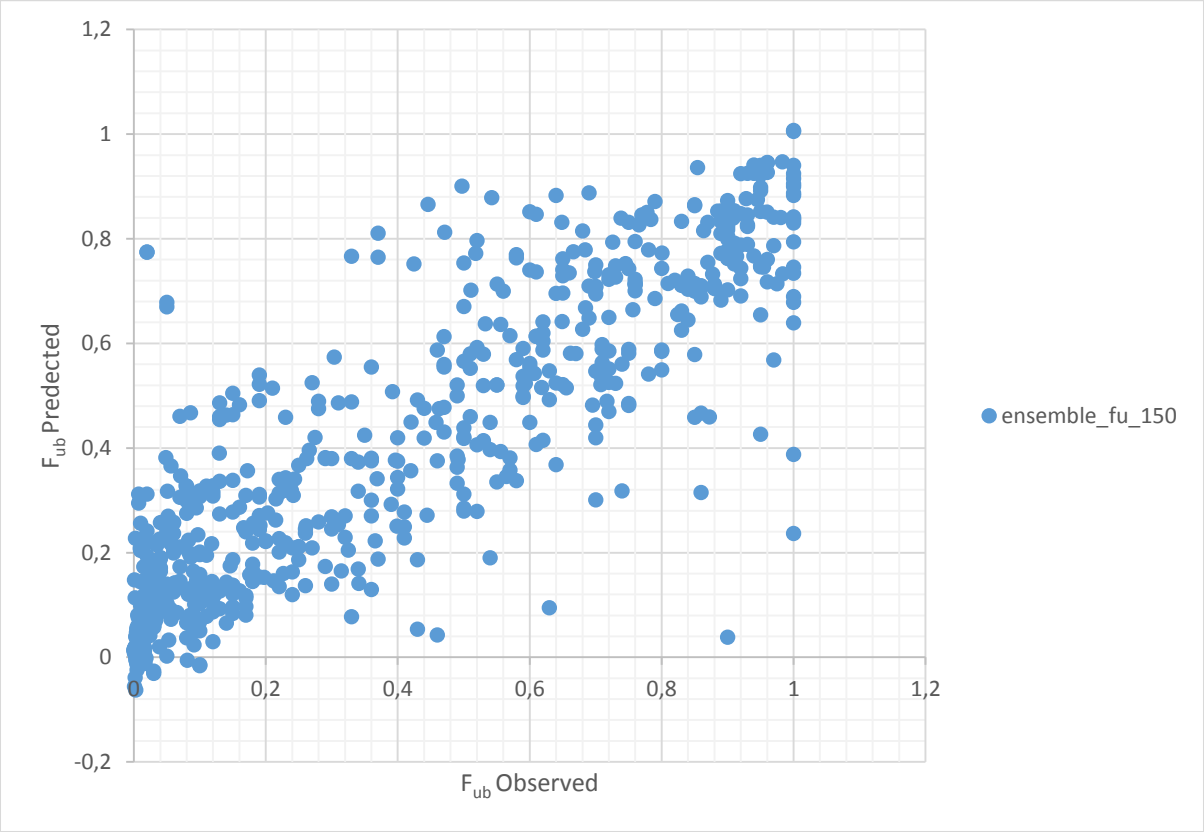

Figure 12 - Plot of the observed vs predicted f_{ub} in the ensemble with the 150 best ANNs

The best ANNE was built using the 20 best ANN, i.e. the ANN with smallest RMSE, and the statistical evaluation of this ANNE is displayed in Table 5.

Table 5 - Statistical evaluation of the best ANNE

APE	0.02822
RMSE	0.16798
Mean Error	-0.00291
Max res	0.84147
Min Res	0.00044
Spearman correlation	0.85722

APE – Averaged squared error of prediction

RMSE – Root mean square error

Max res – Maximum residue

Min res – Minimum residue

Regarding the input weight information, all data was collected, and the average and standard deviation was calculated for each input used. Additionally, the average weight of each class of molecular descriptors was also calculated.

Table 6 - Input weight information

Molecular Descriptor	Average Weight	Standard Deviation	Descriptor Class	Average Weight
Bin 1	0.29	0.138	Specie	0.69
Bin 2	0.40	0.091		
RBN	2.37	0.362	constitutional descriptors	10.94
nS	3.54	0.386		
nR06	5.04	0.345		
J	1.81	0.681	topological descriptors	3.27
D/Dr09	1.45	0.325		
SRW05	1.20	0.319	walk and path counts	4.31
MPC06	3.11	0.751		
IVDE	2.25	0.501	connectivity indices	2.25
SIC1	3.98	0.551	information indices	3.98
GGI3	2.59	0.595	topological charge indices	10.05
JGI4	3.35	0.561		
JGI5	1.42	0.515		
JGI6	2.69	0.336		
DISPe	1.74	0.375	geometrical descriptors	1.74
L2u	2.85	0.561	WHIM descriptors	16.89
E1v	2.83	0.389		
E2v	3.45	0.401		
G2p	2.77	0.505		

Du	4.98	0.374		
H7u	2.29	0.511	GETAWAY descriptors	21.72
HATS2u	3.85	0.708		
R3u	3.21	0.539		
R1u+	3.13	0.504		
R6m+	2.10	0.534		
R5v+	2.76	0.540		
R5e+	2.12	0.448		
R5p+	2.27	0.715		
nCs	3.69	0.483	functional group counts	11.48
nRNH2	2.33	0.423		
nRNR2	1.88	0.302		
nOHs	3.11	0.539		
nArX	0.47	0.261		
ALOGPS_logP	8.50	0.754	molecular properties	12.67
Pka Base	2.44	0.314		
pKa Acid	1.74	0.403		

ANNE Validation

As stated earlier, the external validation was done by comparing the values predicted by the best ANNE to the observed fub values of the drugs in the external validation group of data, using the best ensemble obtained in the optimization process (ensemble with the 20 best ANNs). The statistical information of the ANNE validation is displayed in Table 7

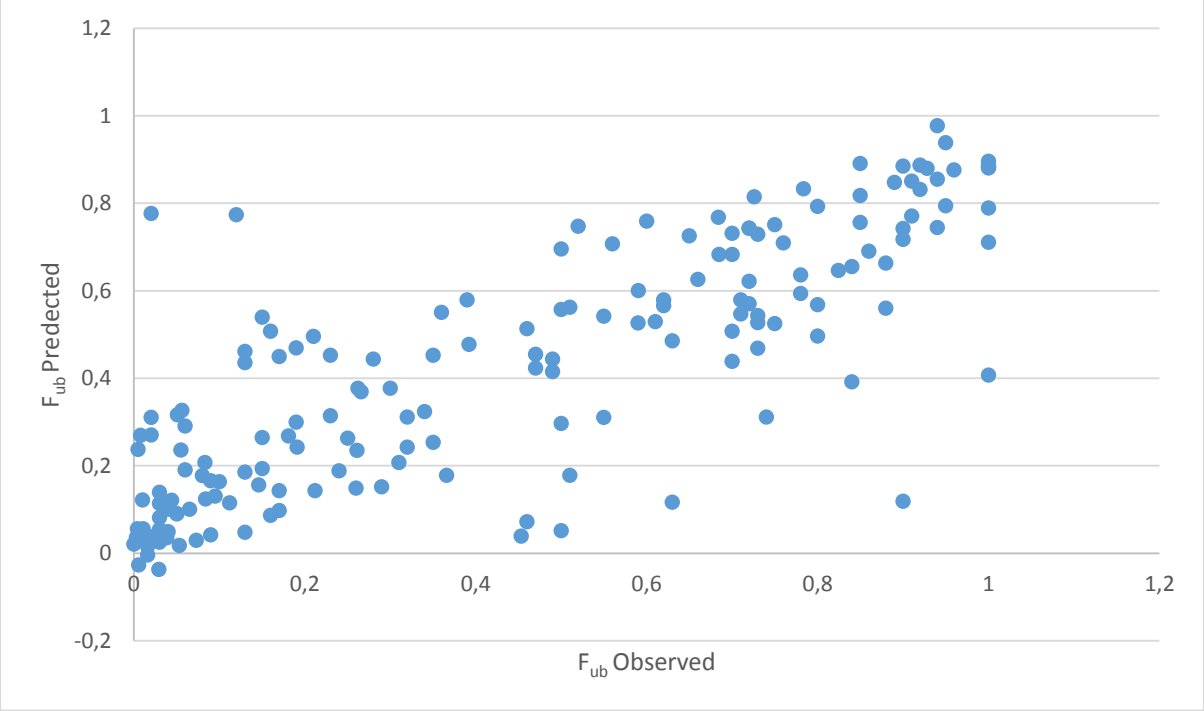

Figure 13 - Plot of the observed vs predicted f_{ub} in the ensemble with the 20 best ANNs for the test subset

Table 7 - Statistical information for the validation of the best ANNE

APE	0.037522
RMSE	0.193705
Mean Error	-0.015780
Max res	0.781048
Min Res	0.000048
Spearman correlation	0.812773

Discussion

As stated in the first chapter of this essay, the analysis of the PPB drug capability is vital for the assessment of the drug's pharmacokinetics, pharmacological and toxicological effects. Also the PPB capacity of a drug might account for numerous interactions between different drugs, because the number of binding sites is limited [6] [7] [10] [11].

Additionally, the need for reliable *in silico* technique to predict PPB of existing and new drugs is of great importance, in order to reduce development and research cost [6] [4] [9].

To this end, this study aimed at the development of a QSAR model for the prediction of drug's PPB, using an ANNE. Since the binding affinity of drugs to plasma proteins is a rather complicated task, because it depends on the quality of the dataset, a previously developed dataset of intravenous pharmacokinetic data from human, rat, dog, and monkey for approximately 400 compounds was used, since data set was carefully compiled from literature reports and expanded with some inhouse determinations for plasma protein, and also because to the authors' knowledge, it was the largest publicly available data set [22].

Several QSAR models have already been developed to predict PPB, several types of molecular descriptors and size of drug database have been used, however, they provide little mechanistic understanding of binding relationships [13] [4].

Furthermore, no evidence have been found in the bibliographic review of a QSAR model that could predict the f_{ub} for different species.

Based on what was stated previously, and given the existence of f_{ub} values for different species in the database used, it was imperative to evaluate if there was a statistically significant difference between de f_{ub} of a drug for the different species, because if no statistically significant difference was found, the model developed could predict automatically the f_{ub} for the four species.

To this end an ANOVA test was performed, and a statistically significant difference was observed between the different species for the f_{ub} at a 95.0% confidence level.

This result was expected since previous studies also showed difference in plasma protein binding in different species. This fact can be explained by difference in physiology, development and biological phenomena in the different species, and also by difference in the plasma proteins composition and binding characteristics, due to structural differences between species [23] [24] [25] [26] [27].

ANNE Optimization and Validation

Optimization of the neural networks was done as reported in methods. Validation is a critical aspect of any model construction, for neural network model validation, usually model validation is based upon some specified network performance measure of data that was not used in model construction (a “test set”) [28] [29].

The methodology mentioned above is also known as train-and-test and the proportion set aside for training of the available data has ranged, in practice, from 25% to 90%, in this case a 75/25 training testing data set ratio was used [28].

Additionally, during the model optimization procedure, and in each individual training step, the train data was again divided randomly into an actual train dataset and an internal test dataset again in a 75/25 ratio, which is a variation of the train-and-test methodology [28].

This internal train/test random division process was done in order to allow the early-stop of the optimization, by training until a degradation of the RMSE in the internal validation data was observed. This methodology is based on the principal that the test error reaches a minimum and then increases as training goes on, while the training error monotonically decreases as show in Figure 14 [30].

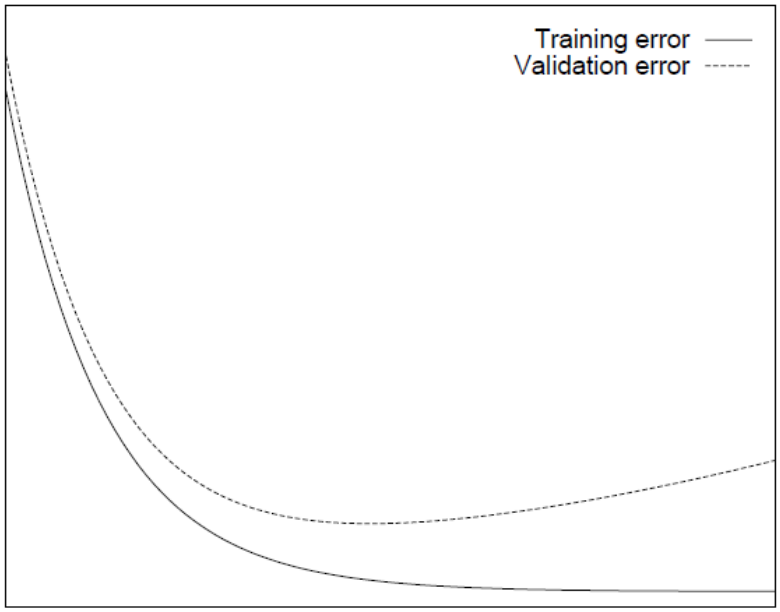

Figure 14 - Idealized behaviour of training and validation error [30]

Nevertheless, in real life, things are a lot more complex with error curves having almost always more than one local minimum. The curve showed in Figure 15, which is a generic error curve, exhibits as many as 16 local minima and of these local minima, 4 are the global minimum up to where they occur [30].

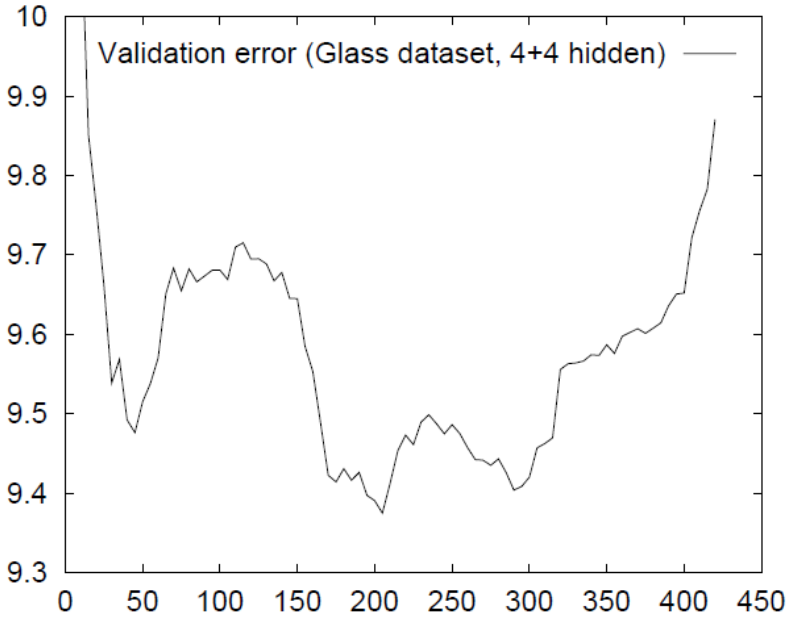

Figure 15 - Real validation error curve [30]

Due to this fact, each network was run for an excessive number of iterations, and the iteration that resulted in the lowest residual mean square error (RMSE) of the testing group was kept. Addition, each of the selected ANN structures was started 20 times with random initial values, in order to sweep the parameters space and avoid convergence to local minima.

As can be seen in Figure 4, RMSE of the train datasets, within each individual structure, after optimization varied between 0.098 and 0.138, as for the RMSE of the internal validation datasets varied between 0.101 and 0.129.

The similarity of the range of RMSE of training and testing datasets show that the early stopping method was able to avoid overtraining and memorization. Additionally, the small range of RMSE indicates that local minima was avoided.

Another problem that could affect the performance of the model, and that had to be addressed early in the development process of the ANNE model, was the number of molecular descriptors available, because with the increase in size of the ANN system, the number of neural connections also increases proportionately. This condition naturally

increases the likelihood of training algorithm convergence problems. Besides, the memory requirement and the processing time have to be addressed have as well [31].

To this end the input space dimension reduction was done in a two-step process.

The first step, was a brute force approach, where molecular descriptors where eliminated based on three factors:

- Only molecular descriptors with the classification mentioned in Table 3 were selected;
- Molecular descriptors highly correlated ($r>0.9$) were eliminated, allowing the removal of redundant descriptors, which contained information already within another descriptor;
- Molecular descriptors with highly repetitive values (90%) were eliminated, ensuring that the neural network wouldn't fit to a specific kind of molecules, based on the training molecules randomly assigned to each network, which would reduce the applicability domain of the final model.

The second step consisted on the optimisation of the network structure for the most relevant molecular descriptors, by a pruning process, as described under methods. ANN with one hidden layer, and a ρ value between 2 and 3, were used to reduce the ability of the network to memorize the data and avoid over-fitting [32]. This ANNs where trained 20 times, and relative input weights and standard deviation were collected for all networks.

The elimination process based on the RMSE and standard deviation values of the species binary code, as described under methods, because this attributes were considered *a priori* relevant descriptors, based on the ANOVA test results.

Additionally, and to validate the input space reduction process, Figure 2 and Figure 3 show that with the reduction of the number of descriptors, RMSE decreased and correlation increased in the test subset, which indicates that this process was able to remove complexity to the model, and increased the capability of prediction, without compromising the adjustment capability of the model, since the RMSE of the training subset didn't increase that much.

Furthermore, the results described previously for the correlation parameter, allows the conclusion that this reduction process was able to eliminate molecular descriptors that only adjusted to small group of molecules.

Finally, and in order to optimize the ANN architectural structures, a last step of the optimization process consisted on training 150 ANN randomly selected form a set of 256

ANN, to ensure a reasonable computational time. The ANN had varying the number of hidden layers (up to three) and the number of hidden neurons in order to obtain a ρ above 1, because as reported above, it reduces the ability of the network to memorize the data and avoids overfitting.

Subsequently to the optimization process, the ANNE was built for the best 10, 20, 30, 40, 50, 100 and 150 networks based on the RMSE values, due to the fact that they are able to improve generalization performance [33] [34].

The creation of an ANNE can be divided into two steps, creating individual members, which in these case are the 150 ANN developed, and combining the output of the ensemble members, to produce the ensemble output. In this study, this last step was obtained by averaging the output data of the ANNs selected for each ensemble [34].

As can be seen from figure 5 to 12, the best ANNE obtained was the ensemble with 20 ANN, because it show the best (smallest) RMSE. The results attained here, suggests that the evolution of the RMSE isn´t proportional to the number of ANN used in the ensemble, it rather shows that the RSME stats to fall with the increasing number of ANN used, however, after a while, the RMSE stats to increase and after that this, it tends to maintain a certain level of RMSE.

The results mentioned above, can indicate that although the ANNE provides a better generalization performance, if the number of ANN used is too big, the additional ANN added are adding noise to the model, instead of creating a better approximation model.

Analysing the statistical results of the best ANNE for the train dataset, it shows that the model has a good ability to adjust to this dataset, since the RMSE is small. Also, it shows that the model generally tends to underestimate the values of the f_{ub} since the mean error has a negative value.

Additionally for the statistical analysis of the ANNE, a Spearman's rank-order correlation was preformed, to measures the strength of association between the f_{ub} observed and predicted. Given the result (0.85722), there is a good correlation between the values predicted and observed [35] [36].

Regarding the validation of the ANNE model, the best ANNE was used to obtain the predicted output for the test dataset, not previously used in the training and optimization process, and the results were compared to the observed values, as showed in Figure 13.

Analysing the validation results and comparing them to the training results, the model has a good ability to predict f_{ub} values, since a small RMES was obtained for the

previously unused dataset, and this RMSE value although bigger it has the same magnitude of the value obtained for the training dataset. Furthermore, and based on the Spearman's rank-order correlation result, a good correlation between the values predicted and observed can be observed.

ANNE Interpretation

As mentioned in methods, to build this model an input size reduction was performed. As consequence, and due to the "black-box" nature of the ANN models, it is important to create a bridge between the mathematical model, and what occurs in nature.

To this end, it is not only relevant to analyze the molecular descriptors obtained in the optimization process and their relative weights in the final output, but also to understand how molecules bind to the plasma proteins, and how this two things relate.

Frist, let's start by understanding how drugs interact with plasma proteins, and which proteins they interact with. The most important plasma proteins in terms of drug binding are albumin and α1-acid glycoprotein, followed by lipoproteins. The serum albumin is the primary constituent in human plasma, accounting for 60% of total plasma protein [6].

As said earlier, acidic drugs have high affinity to albumin, whereas the main binding protein for many basic drugs is AAG. For drugs, there are two main binding sites and a variable number of secondary (lower affinity) sites on albumin, both main sites are elongated hydrophobic pockets, however, site I especially binds bulky heterocyclic anions (e.g. warfarin), whilst site II preferentially recognizes small aromatic carboxylic acids (i.e. ibuprofen) [6] [7] [37].

Initial studies indicate that binding of drugs to AAG appear to involve hydrophobic rather than electrostatic forces, and some acidic drugs such as warfarin can compete with basic drugs for what appears to be a single binding site, perhaps in the protein part of the glycoprotein molecule [11].

More recent studies show that both hydrophobic and electrostatic forces have an important role in interactions between AAG and drugs. Furthermore, AAG exists in a mixture of two or three genetic variants, with different number of binding sites (varying from 2 to 3 binding sites), and different affinity to drugs and their structural functional groups [38].

Lipoproteins, have been described to bind some basic drugs such as chlorpromazine and imipramine. The forces involved here are mainly hydrophobic. However the clinical relevance of these findings are unclear due to their present in small amounts in blood [4] [11] [37].

Regarding the 37 molecular descriptors selected after the ANNE optimization, the class with smallest relative weight in the output values is the binary code for the species characteristic, since in the optimization process, all molecular descriptors with weight smaller to these ones were eliminated.

After this descriptor class, those with smallest influence in the model are the geometrical descriptors, information indices, connectivity indices, topological descriptors and walk and path counts.

The geometrical descriptor selected by the model was the displacement value weighted by Sanderson electronegativity, and it represents the displacement between the geometric and the electronegativity centres of the molecule [39]. Since the interaction of the drugs and the plasma protein depend on the charge of the molecule and it´s 3D structure, it makes sense that this descriptor is relevant for the affinity of a drug to a certain protein.

The connectivity index used by the model is the mean information content on the vertex degree equality, which is a measure of the lack of structural homogeneity or the diversity of a molecule [40].

The topological descriptors selected by the model are Balaban-like index from topological distance matrix (Balaban distance connectivity index) and the distance/detour ring index of order 9. The Balaban distance connectivity index is able to differentiate isomer molecules, and as consequence, highlights the importance of the molecular structure in the affinity of a drug to the plasma protein, which as stated earlier is a known fact [41]. The distance/detour ring index of order 9 is a measure of the cyclicity of the molecule, and as such is related to the morphology of the molecule [42].

The information indices selected were the Structural Information Content index (neighborhood symmetry of 1-order), which is a structure related descriptor, and as stated before, the structure of the drug is a major factor for its affinity for the plasma proteins [43].

The walk and path counts descriptors selected in the model were the self-returning walk count of order 5 and molecular path count of order 6, which also relate to the size, complexity and structure of the molecule [44].

The descriptor classes with biggest influence in the output values are constitutional descriptors, topological charge indices, WHIM descriptors, GETAWAY descriptors, functional group counts and molecular properties.

The constitutional descriptors selected in this model were number of rotatable bonds, number of sulphur atoms and number of 6-membered rings, and as their names indicate, they represent the flexibility of the molecule, the number of certain kind of atoms or structures in the molecule, so they represent structural aspect of the drug.

The topological charge indices used in the molecule evaluate the charge transfer between pair of atoms, and therefor evaluate the capability of the molecule to form dipoles, which can be related to the charge affinity of certain molecules to specific binding sites of some plasma proteins [45].

The WHIM descriptors are 3-dimensional molecular indices that represent different sources of chemical information such as the whole 3D-molecular structure in terms of size,

shape, symmetry and atom distribution, therefore its relevance in the model can be related to the structural requirements of the drugs to bind to the plasma proteins [46].

The GETAWAY descriptors are used to describe the molecular structure of the drug, and as such, can be related to the capability of a drug to bind to a plasma protein, the same way the WHIM descriptors does [47].

The functional group counts descriptors indicate the number of certain functional groups in the molecule, and as such can be related affinity of certain functional groups and structural groups to specific binding sites of some plasma proteins.

The molecular properties define the lipophilic and hydrophilic affinity of the molecule, expressed by the logP value, and the ionization state of the molecule, expressed by de values of pKa and pKb. Since this properties influence the affinity of the molecules to the hydrophilic and lipophilic binding sites of the plasma proteins, and the charge of the molecule, it is expected that they contribute heavily to the output value.

Future Research

After this study, it is clear that the model here developed is capable of predicting the f_{ub} of drugs, with a small amount of error, if molecular descriptors required are calculated. Furthermore, the determination of the applicability domain of this model should made, and also the different capability of the model to predict the f_{ub} for the different species. Additional when presenting the final result for a desired molecule, the model should automatically correct the value, if a negative value is determined.

Bibliography

[1] S. Dhillon and A. Kostrzewski, Basic pharmacokinetics, 1st ed., London: Pharmaceutical Press, 2006.

[2] L. Shargel, S. Pong and A. Yu, Applied Biopharmaceutics and Pharmacokinetics, 6th ed., McGraw-Hill Education, 2012.

[3] D. Smith, H. Waterbeemd and D. Walker, Pharmacokinetics and Metabolism in Drug Design, Wiley-VCH Verlag GmbH, 2001.

[4] Z. Zhivkova and I. Doytchinova, "Quantitative Structure—Plasma Protein Binding Relationships of Acidic Drugs," *Journal of Pharmaceutical Sciences,* vol. 101, no. 12, pp. 4627 - 4641, 2012.

[5] M. Spedding, "New directions for drug discovery," *Dialogues in Clinical Neuroscience,* vol. 8, no. 3, pp. 295-301, 2006.

[6] T. Ghafourian and Z. Amin, "QSAR Models for the Prediction of Plasma Protein Binding," *BioImpacts,* vol. 3, no. 1, pp. 21-27, 2013.

[7] F. Yang, Y. Zhang and H. Lian, "Interactive Association of Drugs Binding to Human Serum Albumin," *Internation Journal of Molecular Science,* vol. 15, no. 3, p. 3580–3595, 2014.

[8] A. Shibukawa, M. E. R. Rosa and T. Nakagawa, "Application of High-Performance Frontal Analysis to Enantioselective Studies on Drug - Plasma Protein Binding," *Chromatography,* vol. 22, no. 1, pp. 25-31, 2001.

[9] T. Moda, C. Montanari and A. Andricopuloa, "In Silico Prediction of Human Plasma Protein Binding Using Hologram QSAR," *Letters in Drug Design & Discovery,* vol. 4, no. 7, pp. 502-509, 2007.

[10] M. Alam, M. Awal, N. Subhan and M. Mostofa, "In-vitro Relationship between Protein-binding and Free Drug Concentrations of a Water-soluble Selective Beta-adrenoreceptor Antagonist (Atenolol) and Its Interaction with Arsenic," *Journal of Health, Population and Nutrition,* vol. 27, no. 1, pp. 20-30, 2009.

[11] P. Routledge, "The plasma protein binding of basic drugs," *Britsh Pharmacological Society,* vol. 22, no. 5, p. 499–506, 1986.

[12] Y. Kawai, Y. Fujii, K. Akimoto and M. Takahashi, "Evaluation of Serum Protein Binding by Using in Vitro Pharmacological Activity for the Effective Pharmacokinetics Profiling in Drug Discovery," *Chemical and Pharmaceutical Bulletin,* vol. 58, no. 8, p. 1051—1056, 2010.

[13] K. Lexa, E. Dolghih and M. P. J. , "A Structure-Based Model for Predicting Serum Albumin Binding," *PLoS ONE,* vol. 9, no. 4, p. e93323, 2014.

[14] ORCHESTRA , “In Silico methods,” ORCHESTRA | In-silico Methods for Evaluating Chemical Toxicity, 2010. [Online]. Available: http://in-silico-methods.eu/about-in-silico-methods. [Accessed 6 July 2014].

[15] Organisation for Economic Co-operation and Development, “Guidance Document on the Validation of (Quantitive) Structure-Activity Relationship[(Q)SAR] Models,” Organisation for Economic Co-operation and Development, Paris, 2007.

[16] E. Benfenati, S. Pardoe, T. Martin, R. G. Diaza and A. Lombardo, “Using Toxicological Evidence from QSAR Models in Practice,” *ALTEX,* vol. 30, no. 1, pp. 19-40, 2013.

[17] G. Dahl and N. Jaitly, “Multi-task Neural Networks for QSAR Predictions,” *ArXiv,* pp. 1-21, 2014.

[18] I. I. Baskin, V. A. Palyulin and N. S. Zefirov, “Neural Networks in Building QSAR Models,” in *Artificial Neural Networks: Methods and Protocols*, Hatfield, Humana Press, 2008, pp. 137-158.

[19] F. Cheng and V. Sutariya, “Applications of Artificial Neural Network Modeling in Drug Discovery,” *Clinical and Experimental Pharmacology,* vol. 2, no. 3, pp. 1-2, 2012.

[20] D. Butina, M. D. Segall and K. Frankcombe, “Predicting ADME properties in silico: methods and models,” *Drug Discovery Today,* vol. 7, no. 11, pp. S83-S88, 2002.

[21] S. K. Singh, S. Saini, B. Verma and D. N. Mishra, “Quantitative Structure Pharmacokinetic Relationship Using Artificial Neural Network: A Review,” *International Journal of Pharmaceutical Sciences and Drug Research,* vol. 1, no. 3, pp. 144-153, 2009.

[22] F. Lombardo, N. J. Waters, U. Argikar, M. K. Dennehy, J. Zhan, M. Gunduz, S. P. Harryman, G. Berellini, I. L. Rajlic and R. S. Obach, “Comprehensive Assessment of Human Pharmacokinetic Prediction Based onIn Vivo Animal Pharmacokinetic Data, Part 2: Clearance,” *The Journal of Clinical Pharmacology,* vol. 53, no. 2, p. 178 –191, 2012.

[23] I. Sabolic, D. Breljak, M. Ljubojevic and H. Brzica, “Are mice, rats, and rabbits good models for physiological, pharmacological and toxicological studies in humans?,” *Periodicum Biologorum,* vol. 113, no. 1, pp. 7-16, 2011.

[24] H. Matsushima, H. Kamimura, Y. Soeishi, T. Watanabe, S. Higuchi and M. Tsunoo, “Pharmacokinetics and Plasma Protein Binding of Tamsulosin Hydrochloride in Rats, Dogs and Humans,” *Drug Metabolism and Disposition,* vol. 26, no. 3, pp. 240-245, 2011.

[25] T. Akabane, K. Tabata, K. Kadono, S. Sakuda, S. Terashita and T. Teramura, “A Comparison of Pharmacokinetics between Humans and Monkeys,” *Drug Metabolism and Disposition,* vol. 38, no. 2, pp. 308-316, 2010.

[26] J.-a. Sun, D.-z. Kong, Y.-q. Zhen, Q. Li, W. Zhang, J.-h. Zhang, Z.-w. Yin and L.-m.

Ren, “Stereoselective binding of doxazosin enantiomers to plasma proteins from rats, dogs and humans in vitro,” *Acta Pharmacologica Sinica,* vol. 34, no. 12, pp. 1568-1574, 2013.

[27] I. Motrescu, S. Oancea, A. Rapa and A. Airinei, “Spectrophotometric analysis of the blood plasma for different mammals,” *Romanian Journal of Biophysics,* vol. 16, no. 3, pp. 215-220, 2006.

[28] N. Kartam, I. Flood and J. Garrett, Artificial Neural Networks for Civil Engineers: Fundamentals and Applications, Carnegie Mellon: ASCE Press, 1996.

[29] S. L. Özesmia, U. Özesmia and C. O. Tanb, “Methodological Issues in Building, Training, and Testing Artificial Neural Networks,” *Ecological Modelling,* vol. 195, no. 1-2, p. 83–93, 2006.

[30] L. Prechelt, “Early Stopping - but when?,” *Neural Networks: Tricks of the Trade,* vol. 7700, pp. 53-67, 2012.

[31] S. Muknahallipatna and B. Chowdhury, “Input Dimension Reduction in Neural Network Training - Case Study in Transient Stability Assessment of Large Systems,” *IEEE,* pp. 50 - 54, 1996.

[32] M. Shahin and M. Jaksa, “Investigation into the Robustness of Artificial Neural Networks for a Case Study in Civil Engineering,” *Modelling & Simulation Society of Australia & New Zealand,* pp. 79-83, 2005.

[33] S. Lawrence, C. L. Giles and A. C. Tsoi, “What Size Neural Network Gives Optimal Generalization? Convergence Properties of Backpropagation,” *Technical Report UMIACS-TR-96-22 and CS-TR-3617,* 1996.

[34] C. Shu and D. Burn, “Artificial neural network ensembles and their application in pooled flood frequency analysis,” *Water Resources Research,* vol. 40, no. W09301, pp. 1-10, 2004.

[35] J. McDonald, “Spearman rank correlation,” 2014. [Online]. Available: http://www.biostathandbook.com/spearman.html. [Accessed 20 September 2014].

[36] Lund Research Ltd, “Spearman's Rank-Order Correlation,” 2013. [Online]. Available: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide-2.php. [Accessed 20 September 2014].

[37] M. Wood, “Plasma Drug Binding: Implications for Anesthesiologists,” *Anesthesia & Analgesia,* vol. 65, pp. 786-804, 1986.

[38] K. Taguchi, K. Nishi, V. T. G. Chuang, T. Maruyama and M. Otagiri, “Molecular Aspects of Human Alpha-1 Acid Glycoprotein — Structure and Function,” in *Immunology, Allergology and Rheumatology » "Acute Phase Proteins"*, Intech, 2013, pp. 139-162.

[39] J. Akbar, S. Iqbal, F. Batool, A. Karim and K. W. Chan, "Predicting Retention Times of Naturally Occurring Phenolic Compounds in Reversed-Phase Liquid Chromatography: A Quantitative Structure-Retention Relationship (QSRR) Approach," *International Journal of Molecular Science,* vol. 13, no. 11, pp. 15387-15400, 2012.

[40] S. Deshpande, V. R. Solomon, S. B. Katti and Y. Prabhakar, "Topological Descriptors in Modeling the Antimalarial Activity: N1-(7-Chloro-4-quinolyl)-1,4-bis(3-aminopropyl)piperazine as Prototype," *Journal of Enzyme Inhibition and Medicinal Chemistry,* vol. 24, no. 1, pp. 94-104, 2009.

[41] E. Estrada and Y. Gutierrez, "The Balaban J Index in the Multidimensional Space of Generalized Topological Indices. Generalizations and QSPR Improvements," *Communications in Mathematical and in Computer Chemistry,* vol. 44, pp. 155-167, 2001.

[42] H. K. Jain and R. K. Agrawal, "Topological Descriptors in Modeling Tumor Necrosis Factor alpha Inhibitory Activity of Xanthines, Pteridinediones and Related Compounds," *Internet Electronic Journal of Molecular Design,* vol. 6, no. 8, pp. 218-228, 2007.

[43] D. Bonchev, "Shannon's Information and Complexity," *Mathematical Chemistry Series,* vol. 7, pp. 155-187, 2003.

[44] I. Lukovits, A. Miliþeviü, S. Nikoliü and N. Trinajstiü, "On Walk Counts and Complexity of General Graphs," *Internet Electronic Journal of Molecular Design* , vol. 1, p. 388–400, 2002.

[45] F. Torrens, "A New Chemical Index Inspired by Biological Plastic Evolution," *Indian Journal of Chemistry,* vol. 42A, pp. 1258-1263, 2003.

[46] R. Todeschini and P. Gramatica, "The Whim Theory: New 3D Molecular Descriptors for Qsar in Environmental Modelling," *SAR and QSAR in Environmental Research,* vol. 7, no. 1-4, pp. 89-115, 1997.

[47] R. Todeschini and V. Consonni, "Descriptors from Molecular Geometry," in *Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes*, WILEY-VCH Verlag GmbH & Co. KgaA, 2008, pp. 1004-1033.