Master's degree project in Bioinformatics

Department of Biology, Lund University

Building predictive unbound brain-to-plasma concentration ratio ($K_{p,uu,brain}$) models

Srinidhi Varadharajan

August 2013-May 2014

Table of Contents

ABSTRACT

The blood-brain barrier (BBB) constitutes a dynamic membrane primarily evolved to protect the brain from exposure to harmful xenobiotics. The distribution of synthesized drugs across the blood-brain barrier (BBB) is a vital parameter to consider in drug discovery projects involving a central nervous system (CNS) target, since the molecules should be capable of crossing the major hurdle, BBB. In contrast, the peripherally acting drugs have to be designed optimally to minimize brain exposure which could possibly result in undue side effects. It is thus important to establish the BBB permeability of molecules early in the drug discovery pipeline.

Previously, most of the in-silico attempts for the prediction of brain exposure have relied on the total drug distribution between the blood plasma and the brain. However, it is now understood that the unbound brain-to-plasma concentration ratio ($K_{p,uu,brain}$) is the parameter that precisely indicates the BBB availability of compounds. $K_{p,uu,brain}$ describes the free drug concentration of the drug molecule in the brain, which, according to the free drug hypothesis, is the parameter that causes the relevant pharmacological response at the target site.

Current work involves revisiting a model built in 2011 and uploaded in an in-house server and checking for its performance on the data collected since then. This gave a satisfying result showing the stability of the model. The old dataset was then further extended with the temporal dataset in order to update the model. This is important to maintain a substantial chemical space so as to ensure a good predictability with unknown data. Using other methods and descriptors not used in the previous study, a further improvement in the model performance was achieved. Attempts were also made in order to interpret the model by identifying the most influential descriptors in the model.

1. INTRODUCTION

1.1 Background

Trends and technologies associated with Drug discovery and development have experienced an upsurge over the last few decades. Being an interdisplinary field of research, this area has seen some profound expansion in the knowledge base with the improvements in understanding of basic chemistry, genetics, and molecular biology and so on. The sequencing of human genome was one such advancement that further opened up the prospects by aiding in the discovery of new targets. These advancements have led to a lot of success stories in treating various diseases.

One of the major advancements in the field of pharmaceutical research can be pointed out as the advanced computational technologies that are now being employed in all major stages of the drug discovery process. These have been shown to considerably reduce resources and time and limit chemical synthesis of undesirable compounds.

The drug discovery and development is in general a long, tedious and expensive process coupled with a high risk of failure. Typically, it takes around 15-20 years to develop a drug from the initial stages of target identification till the introduction of the drug into the market. The cost of introducing a drug molecule into the market has been approximated to be almost $1.2 billion. [1] U.S. Food and Drug Administration (FDA) estimates that eventually only 8% of the compounds that enter the Phase 1 clinical trials can reach the market. [2]

Though the investment in the pharmaceutical research has increased through the years, associated developmental cost of drugs also seems to be accruing [3] while the number of approved drugs has reduced owing to the tighter regulatory requirements for the drug approval. A lot of efforts have been made to address the fundamental issue of reducing attrition while also speeding up the process of drug development.

Traditionally, in the early phases of drug discovery, the focus was mainly on the efficacy and the selectivity of the molecule towards the target. The pharmacokinetics and ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicology) studies were usually carried out later in the later phase. Through the years, some studies had indicated that a major portion of drug attrition is attributable to poor ADMET properties. [4] This called for a need to move towards the *'Fail fast Fail cheap'* strategy by investigating the ADME properties in

early phases of drug discovery. [5] The cost associated increases as a compound progresses through the drug development cascade. [6]

To reduce the late-stage attrition, it is critical to identify compounds that are unlikely to succeed and to terminate the development of these as early as possible. Only the compounds exhibiting a good ADMET profile should be advanced into the clinical trials thus reducing the resources spent.

Computational models are one of the attractive solutions for predicting the appropriate ADME characteristics of molecules under consideration. These approaches can be used as a cost effective filter for choosing compounds that are most likely to meet the desired needs [7], even before their actual synthesis.

1.2 Drug discovery process

Historically random experiments were designed by trial and error basis to find novel drugs [8]. With the extensive efforts in understanding molecular biology, chemistry, biochemistry, genetics and so on, in context of human body and various related functions, the process of discovering drugs has become much more streamlined and continues to improve.

Traditional linear model of drug discovery and development processes starts off with defining the disease to be investigated in the project. This is followed by identification of a target involved in the disease using various genomics and proteomics analysis and then followed by the validation of the identified target. A druggable target is usually a biological component like enzymes, receptors etc which can bind to the ligand and elicit the required response.

Once a suitable target has been identified, thousands and millions of compounds are screened for an interaction with the target to discover “hit” compounds which serve as the starting compounds for the drug development process. The hit compounds are then progressed into the lead identification, validation and optimization phase. During this phase the compounds that can interact with the chosen target are then assessed for various properties like selectivity and affinity using biochemical assays. Potency of the compounds is also determined. This phase involves multitude of in-vitro/in-vivo screening assays to determine and confirm the characteristics of the molecule. The techniques like structure activity relationships (SARs) and ADMET studies are carried out to evaluate the development potential of the lead compounds.

After the initial selection of lead compounds series, optimization work is done to improve the efficacy and safety of the chosen compounds. Structural variations of the lead series are often performed at this stage to tailor the molecule for desired qualities. These analogues are then tested using various assays to find the best compound to serve as a drug candidate.

The next step involves studying the effects of the drug candidate in animal models. In this stage the safety and efficacy of the candidate molecule is further studied and pharmacokinetic profiles of the molecule is established in the animal models. Drug toxicity profiles are extensively evaluated using in-vitro and in-vivo assays. High levels of safety have to be established before the human trials in the next phase. The above mentioned steps are referred to as pre-clinical research where the safety and efficacy of the drug molecule is established prior to its advancement towards the clinical trials.

Clinical research is the final phase which involves various phases where the drug molecule is tested in human. After a drug molecule successfully passes these phases, the approval for the new drug is sought from the regulatory agencies. Once approved, the drug can eventually reach the market. However, the follow-up clinical studies still need to be conducted in of form of post-marketing investigations.

1.3 Challenges in CNS drug discovery

Many Central nervous system (CNS) diseases do not have effective drugs in the market. Most small molecule CNS drugs in the market are focused on certain therapeutic areas like migraine, epilepsy etc., while leaving some of the other common, often devastating CNS disease with no effective cure. [9] There is a great unmet need in the area of neurodegenerative diseases like Alzheimer's disease, Parkinson's disease and so on. [10] In many instances, despite identification of some promising molecules, the complexity of CNS has kept them much away from becoming successful drugs. On the other hand, targeting CNS might also be necessary in certain non-CNS diseases.

CNS drug candidates have been observed to have a lower success rate and longer development phases, as compared to their non-CNS counterparts. [11] This difference can be largely associated with the numerous complexities involved in targeting CNS, like the intricacies of the human brain, lower predictability of animal models, CNS side effects and so on.[11] This area thus entails a careful investigational research of alternate approaches for increasing the success rate.

In drug discovery phase, it is very essential to determine the possibility of a molecule to cross the blood-brain barrier. The drugs targeted to CNS must successfully permeate the BBB to achieve an optimal distribution to the brain. The peripherally acting drugs, on the other hand, may have to be kept away from the brain to avoid unwanted toxic effects. This necessitates a better understanding of the complexities that surround BBB and the properties of molecules that can increase or decrease the permeation.

Various in-vivo and in-vitro experiments have been designed for the determination of extent of brain penetration for drugs. However, these methods are quite time-consuming and costly. Computational methods to predict molecular properties will be very useful in initial screening so as to make a good decision of which compounds can go forward to the more laborious and expensive tests involved in *in-vivo* and *in-vitro* studies. [12]

1.4 Barriers to access the brain

Existence of physiological barriers that separate CNS from the systemic blood circulation has been well established. Two vital barriers in the CNS are the blood-brain barrier (BBB) and the blood-cerebrospinal fluid barrier (BCSFB). The former acts as a barrier between the blood and brain interstitial fluid (ISF) and is composed of brain capillary endothelial cells with tight junctions. The latter is present at the choroid plexus, separating blood and ventricular cerebrospinal fluid (CSF) [13] and is formed of epithelial cells linked by tight junctions which are however more permeable than the BBB. [14] The BCSFB is found to have a relatively much smaller surface area as compared to the BBB, thus BBB is thought to play the major role in drug delivery. [15] Various metabolic enzymes and transporters are also present to shield/protect the brain from endogenous toxins and various other xenobiotics.

1.4.1 The blood brain barrier

BBB presents the major hurdle for a drug to reach a target in brain. It primarily functions to regulate the transport of compounds to and from the brain for protecting it from harmful xenobiotics and other potential neurotoxins. It has been observed that almost 98% of the small molecules do not cross the BBB. [16] BBB is thus crucial to maintain homeostasis in the CNS.

The existence of such a barrier in the brain was first realized in the 19th century through the experiments performed by Paul Ehrlich. [17] BBB has been actively under research and scientists are attempting to gain deeper understanding of the complexities.

The structural feature that is of primary importance in BBB is the tight junctions that exist between capillary endothelial cells. These effectively restrict inter-cellular transfer of solutes. The tight junctions are characterized by absence of fenestrations and usually display a low pinocytosis.[18] Efflux transporters are another defense mechanisms flaunted by the BBB. They serve to pump toxins and xenobiotics out of the brain. BBB has also been observed to show a high electrical resistance, it thus keeps polar and ionic molecules, especially the acid compounds, away from penetrating into the brain. [19]

1.4.2 Transportation of molecules across the BBB

Usually transport of compounds across the BBB occur transcellularly as the paracellular transportation is restricted by the presence of tight junctions. However, transcellular mode of transport is further affected by various efflux transports present at the BBB. There are various mechanisms that occur at BBB influence the brain permeation of compounds.

Passive diffusion

Most commonly, compounds enter the brain by passive diffusion, where the concentration gradient is the main driving force. Equilibrium is attained when the concentration of the drug compound at either side of the membrane are equal [20]. The capacity of the drug to passively permeate depends on its physicochemical properties. For example, lipophilicity has been identified to be a key factor for diffusion of drug into the brain. It is seen that an increases in lipophilicity usually corresponds to a higher BBB permeation. Other properties like molecular weight, polar surface area etc, also play a vital role.

Carrier mediated transport

Various transporters are present at the BBB evolved to function in effectively protecting the brain. Some of the compounds that are hydrophilic and cannot undergo passive diffusion use transporters to aid the permeation process.

Influx transporters are involved in transport of molecules like glucose, amino acids from blood to brain [21] to provide the nutrients required by the brain. These mainly aid the transport of small hydrophilic molecules which can otherwise not pass through the BBB by passive diffusion.

Efflux transporters are a very critical defense system evolved at the BBB. They serve the gatekeeper function by pumping out the potentially toxic compounds from the brain. The most important efflux transporters at BBB are P-glycoprotein (P-gp), breast cancer resistance protein (BCRP) and multidrug resistance protein (MRP). [22] They fall into the ABC (ATP-binding cassette) superfamily. [23]

1.4.3 Characteristics of molecules that can potentially cross BBB

Drug-like molecules ought to have a good bioavailability, ADMET properties and potency. Druglikeness has often been defined based on "Rule of Five" proposed by Lipinski[84], which is a set of rules assessing the oral absorption of the compounds. The "Rule of Five" states a rule of thumb for a compound to possess good bioavailability, a molecular weight less than 500; Number of hydrogen bond donors less than 5; number of hydrogen bond acceptors less than 10 ; a ClogP (octanol-water partition coefficient) less than 5. Compounds that do not exhibit the forementioned characteristics are most likely to suffer poor bioavailability. Similar efforts have been made by many other groups to examine the relationship between the molecular properties of pharmaceutically relevant compounds and their potential to become drugs. The purpose has been to discover trends in the physicochemical properties for the compound in a particular developmental stage or in a certain disease area and to identify the key factors for compound related attrition.

In general CNS drugs tends to be on the higher side with respect to the lipophilicity, and rigidity of the molecule while they need to be smaller (lower molecular weight), and possess lower hydrogen-bond acceptor and donors, fewer negative charges and a lower PSA, compared to the non-CNS drugs. Numerous studies have attempted to look into the molecular physiochemical properties related to BBB penetration. Similar to Lipinski's "Rule of five", some simple rules have been formulated to define CNS drug-likeness.

For example, it has been proposed that the following attributes are advantageous for a potential lead to be able to permeate the BBB. [24] A molecular weight less than about 400-450; number of hydrogen bond donors less than 3;number of hydrogen bond acceptors less than 7; a PSA of 60-70 $Å^2$; pKa of 7.5-10.5 and fewer rotatable bonds.

Lipophilicity has been known to be one of the most critical factors for the BBB permeation. Higher lipophilicity enables the compounds to permeate the lipid rich membranes. [25] Lipophilicity is often expressed in terms of logP. Though a high lipophilicity is favorable, it is important for a molecule to possess optimal values of logP, because with the increase in lipophilicity the non-specific binding of the molecules to the plasma proteins also increases.

Permeation of a compound across BBB is highly influenced by hydrogen bonding potential of the molecule. BBB permeation decreases significantly with increase in number of hydrogen bonds. [26] This necessitates that the sum of nitrogen and oxygen atoms in the molecule should preferably be kept below 5. [24]

For the passive diffusion, it has been observed that ionization of the molecule plays an important role. Weak bases and neutral compounds have much higher chances of permeating the BBB as compared to acids. On the other hand, strong acids and bases are usually not capable of penetrating the BBB. [27]

1.5 Measurement of brain exposure

1.5.1 Total brain-plasma concentration ratio (Kpb)

The total brain-plasma concentration ratio denoted by $K_{p,brain}$ or logBB has been the most widely used parameter for in-silico prediction of brain exposure. It is calculated as logarithm of the ratio of the concentration of the drug molecule in the brain to that in blood, at equilibrium (Equation 1 [28]). It basically measures the way the drug molecule partitions itself between the brain and the blood.

$$K_{p,brain} = \frac{C_{u,brainISF}}{C_{u,p}} \quad (1)$$

However, it has been argued that logBB, being based on total concentrations, is affected by the non-specific binding of the molecules to the plasma protein and brain tissue [29] and may be misleading [28,31,32] since it is only the free drug that is available for transport across BBB and for binding to the target proteins in the brain.

1.5.2 Permeability solubility product

Upon realization of the incomplete description given by the conventional analysis using LogBB, it was suggested to alternatively use logPS (logarithm of permeability solubility) as a measure of unbound molecule. The permeability solubility product measures the rate of drug transport over the BBB [33]. It is measured by in-vitro brain perfusion experiments. However, this measure does not represent the free drug concentration either as it does not consider the efflux clearance at BBB. Furthermore, PS is a measure of penetration rate and therefore is not necessarily correlated with the extent of penetration.

1. 5.3 Unbound Brain-to-plasma Concentration ratio ($K_{p,uu,brain}$)

Unbound brain-to-plasma concentration ($K_{p,uu,brain}$) is a parameter that estimates the amount of free drug in the brain ISF. It is defined as the ratio of unbound drug concentration in brain to the unbound drug in plasma, in steady state [28](Equation 2).

$$K_{p,uu,brain} = \frac{C_{u,\ brainISF}}{C_{u,plasma}} \quad (2)$$

Where $C_{u,brainISF}$ is the free drug concentration in the ISF and $C_{u,plasma}$ is the free drug concentration in the brain. The unbound drug concentration in brain, $C_{u,brainISF}$ can be directly measured through microdialysis in brain[45,46]. The method is experimentally challenging and involves a large amount of resources to carry out. The non-specific binding associated with highly lipophilic molecules pose a further challenge for this method [34] and is therefore of a limited usability in drug discovery projects. However, an alternative method of determining $K_{p,uu,brain}$ has been proposed, which can be used to circumvent the problems associated with microdialysis[28].Where two *in-vitro* experiments and one in-vivo experiment are used to determine $K_{p,uu,brain}$.

$$K_{p,uu,brain} = \frac{K_{p,brain}}{V_{u,brain}\, f_{u,p}} \quad (3)$$

Where $K_{p,brain}$ is the total brain-blood concentration ratio, $V_{u,brain}$ is the unbound volume of distribution in brain and $f_{u,p}$ is the unbound fraction of drug in plasma. $V_{u,brain}$ is commonly measured using brain-slice method[62] and $f_{u,p}$ is determined by equilibrium dialysis technique.[85]

Generally it has been concluded that when $K_{p,uu,brain}$ is close to 1, the compound is expected to be able to cross the BBB by passive diffusion and is also not a substrate for the transporters at the BBB. The compounds with a $K_{p,uu,brain}$ of greater than 1 are substrates for the influx transporters and are thus actively transported, while the compounds that have a $K_{p,uu,brain}$ less than 1 tend to be substrates for efflux transporters.[35]

According to the eq. 3, experimental determination of $K_{p,uu,brain}$ involves measurement of the total brain-to-plasma concentration ratio obtained from in-vivo animal experiments and in-vitro determination of plasma protein binding and binding to the brain tissue. Predictive in-silico models can be of great value in circumventing the necessity of performing such resource and time intensive experiments.

1.6 *In-silico* predictive models

Predictive modeling has now gained popularity in the area of drug discovery. It is based on using algorithms that can learn from the provided examples and can later be used for the prediction of unseen examples. *In-silico* analysis involves various statistical methods like multiple-linear regression, Partial least squares (PLS), machine learning methods etc. Data modeling often involves the well known concept of structure activity relationships. As the

name suggests, structure activity relationships refers to the method of correlating structural features that a molecule possesses to its biological activity.

1. 6.1 Quantitative structure activity relationships (QSAR)

The idea of relating the structure of a compound to its biological activity can be traced back to 1869, when Crum brown and Fraser proposed the concept of biological response (φ) being a function of chemical structure (C) of a compound (equation 4)[36]. Thus a change in the chemical feature will alter the biological activity exhibited by the molecule.

$$\varphi = f(C) \qquad (4)$$

Subsequent studies further supported the view of correlating the structure with the activity of molecules. The essence of these methods are in the fact that the structural features of molecule can be used to infer physical and chemical properties of the which can in turn be correlated to the biological activity of the molecule. [37] Various statistical methods are thus used to find features that make a compound active or inactive and the relationships between these features and the bioactivity. These relationships can be either qualitative (SAR) or quantitative (QSAR).

Quantitative structure activity relationship attempts to determine the quantitative relationship between the chemical features and the desired biological response. This relationship between the biological endpoint and the descriptors of the compound are modeled using statistical methods. The modern QSAR studies were initiated by Corwin Hansch around 1963. [66]

The crucial stages in building a reliable QSAR model firstly involve collection of a good dataset as model quality will wholly depend on the quality of the dataset. The subsequent step involves calculating a set of relevant descriptors that can describe the dataset well. The model generation then begins by choosing a suitable method to establish the correlation between the compounds and the calculated descriptors. Validation of the built model is then done to investigate the predictive powers of the model (Figure 1.1).

1.6.2 Molecular descriptors

A molecular descriptor is a calculated numerical representation of various properties inherent in a molecule.

Several common molecular descriptors exists that represent various attributes of a structure , ranging from descriptors that rely on simple counts [38] of important features like hydrogen bond donors acceptors , number of rotatable bonds, number of a aromatic rings systems and

so on to more complex descriptors, for example, based on quantum chemistry calculated properties.[68]

Of great importance are the descriptors for describing physicochemical properties of molecules, such as hydrophobicity measured in terms of octanol-water partition coefficient etc [38]. Other commonly used descriptors include topological indices, shape indices [38], fingerprints [69] etc. Designing of 3D descriptors for molecules has also been an attractive concept to capture the properties that 2D descriptors might fail to describe.

Molecular descriptors thus form the basis to use the properties inferred from the chemical structure in a mathematical setting.

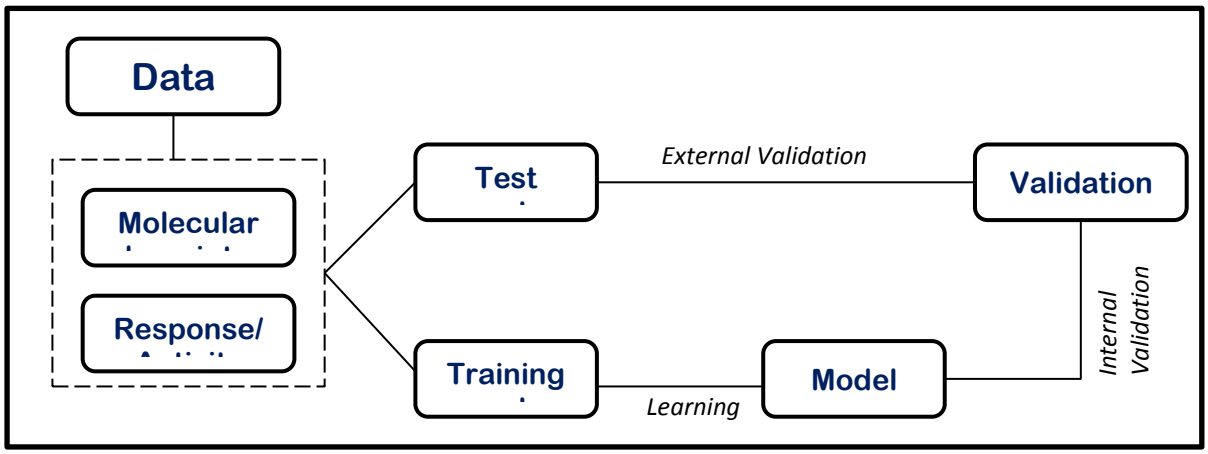


Figure 1.1 General workflow of a QSAR experiment.

1.6.3 Machine learning algorithms

Classical linear QSAR models often utilize linear statistical methods like Partial least squares (PLS) and multiple linear regression. These often suffered from the problem of over fitting [1] and nonlinear relationships cannot be addressed well using these methods. Thus non-linear machine learning methods like support vector machine(SVM), random forest (RF) provide an attractive solution to this problem by offering the advantage of handling large amounts of data more accurately.[39] Machine learning is thus one of the methods commonly used to build QSAR models.

Machine learning is a sub field under artificial intelligence that involves creating computer algorithms that can learn from data. It basically seeks to establish relationships from the data by finding sensible patterns in it, which can then be used to make predictions on new examples. Machine learning as a technique possesses some attractive qualities when

compared to direct programming. It is more accurate and can process larger data more efficiently.

Machine learning includes supervised and unsupervised learning methods. Supervised learning involves algorithms that learn complex patterns from a set of labeled examples. Labeled examples refer to those that have both features and the associated labels. The algorithm thus learns a hypothesis to fit appropriately to the dataset in consideration

Unsupervised Learning, on the other hand involves use of Unlabelled dataset thus there are no associated labels. The algorithm searches for patterns within the data to make useful inferences.

There are two major categories of machine learning problems, namely, classification and regression problems. In regression, the data consists of continuous values, and the model thus predicts a real value for the new example. Classification, on the other hand, deals with categorical data.

1.6.4 Validation of the QSAR model

A key point to consider about a QSAR model is its predictive power, which is indicative of how well the model can predict an example that it has not seen before. Determination of this is done through different validation methods. The validation methods can be primarily categorized into internal validation and external validation.

Internal validation is used to determine the model fit, which gives an idea of how an unseen example might be handled by the model. Most commonly employed methods are leave one out cross-validation [65], k-fold cross validation [66] and so on. It is measured in terms of cross validated correlation coefficient or Q^2.

External validation refers to the use of an external test set which is not included in the dataset used to build the model. The performance of the model on the test set is often measured in terms of correlation coefficient R^2 and RMSE.

1.7 In silico BBB penetration models

1.7.1 logBB Models

Previous work in this field has often been focused on logBB predictive models for describing the capacity of a compound to permeate and distribute across the BBB. Various statistical methods and machine learning algorithms have been employed for building these models. Studies have also analyzed dependencies of the logBB to various vital physico-chemical

descriptors. It has been consistently noted that logBB is mainly dependent on hydrogen bonding potential, molecular volume and lipophilicity [42].

One of the initial efforts towards QSAR modeling of logBB was taken by Young et al., where the correlation of logBB with logP was established in 20 anti-histamine molecules. Later studies by Waterbeemd et al. [44], attempted to correlate logBB to molecular volume and PSA. Futher, Abraham et al. [45], worked on logBB model based on 60 compounds and relating them to five solute descriptors The major limitation with these initial studies were the smaller size of dataset used. Thus, subsequent research saw a lot of progress in such logBB modeling based on extending the Young dataset along with using better statistical methods.

The size of the publicly available logBB data has increased gradually. The biggest public logBB set so far compiled by Lanevskij *et al* [48] constitutes about 400 compounds. Diverse molecular descriptors have been utilized, for example, 2D physicochemical descriptors [49-54] describing information about the molecular size, shape, lipophilicity etc., and 3D molecular structure [54, 55, 56]. The early logBB models typically used a smaller set of descriptors to build the model and the model building strategy was often limited to simple MLR statistics. The recent studies have attempted to build models using larger number of descriptors along with more complex algorithms that can deal with the increased number of variables [42, 57, 58]. The models utilizing non-linear algorithms [42,58] have, in general shown a higher accuracy than the linear models [57,58,59].

However, recently it has been realized that logBB is not very relevant for making inferences on BBB permeability. Thus using $K_{p,uu,brain}$ data for the purpose is a much more attractive solution.

1.7.2 Classification models

Various classification models have been developed for classifying compounds based on their ability to cross the blood brain barrier and elicit the required effect. A common strategy has been to classify compound into BBB+/BBB- based on whether they are permeate through BBB by passive diffusion. These classification models have shown an accuracy of about 75-95 %.

Ajay et al., performed a study on CNS active/inactive drug to build a compound library with potential CNS activity. They also analyzed the difference in the CNS active and inactive terms of seven important descriptors like molecular weight, number of rotatable bonds, kappa2, logP, hydrogen bond donors and acceptors and so on. The models built in this study could produce upto 80% predictability. Zhang et al., [42] performed similar studies using a

dataset of 156 compound and building QSAR classification models mainly employing kNN and SVM algorithms. These models, built on different types of descriptors like Dragon, MOE and MolConnZ, showed a good accuracy of prediction.

There have been several other attempts to build classification models for BBB penetration. Various modeling schemes have been used for building such models. Decision tree based techniques like recursive partitioning [43] has also seen to have a lot of potential for this purpose.

CNS+/CNS- Classification have also been extensively studied. This usually involves categorizing compounds based on whether they are centrally active or not. The CNS+ compounds are usually BBB+, on the contrary, the BBB+ compounds do not necessarily have to be CNS+. The BBB+ compounds can also be CNS- which implies that they permeate the BBB while not showing any activity. This makes the definitions rather simple for the analysis of complexities related to permeation of compounds through the BBB.

1.8 $K_{p,uu,brain}$ Models

The published $K_{p,uu,brain}$ modeling studies performed by Fridén et al [28], utilized a dataset of 43 compounds to build in-silico $K_{p,uu,brain}$ models using PLS. $K_{p,uu,brain}$ was assessed based on $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ using Equation(3) .

16 molecular descriptors were included in the study which comprised of standard descriptors like ClogP, molecular weight, hydrogen bond donors (HBD) and so on. Irrelevant descriptors were later excluded based on calculation of variable importance for projection score (VIP). It was observed that the significant descriptors as picked by the VIP scores were related mainly to hydrogen bonding. The model utilizing only HBA as the descriptor was found to possess comparable predictive power to the model utilizing the set of 16 descriptors.

Chen et al., in 2011, extended the Fridén's dataset to include 247 in-house compounds in total for $K_{p,uu,brain}$. The model was primarily built based on SVM and RF machine learning algorithms. Descriptor set used for the model building consists of 196 in-house descriptors. Modeling strategy included building an indirect model, which utilized the individual datasets of $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ for model building and the $K_{p,uu,brain}$ values then calculated from the individual predictions, and a direct model based on the $K_{p,uu,brain}$ data.

Studying through various single component and consensus model, it was found that the consensus model with SVM direct, RF indirect and RF direct components gave the best

prediction with an R^2 of 0.58 and RMSE of 0.46. It was thus seen that consensus models in general perform better than the single component models.

1.9 The goal of Master Thesis project

The aim of the current master thesis project was to collect the up-to-date AstraZeneca in-house $K_{p,uu,brain}$ data, examine the performance of previously published $K_{p,uu,brain}$ model [39] on the temporal test dataset and build new models by using the expanded dataset. During the model building, various QSAR modeling strategies were used and compared. It was particularly interesting to apply some new QSAR methods such as the combination of support vector machine (SVM) and molecular signature descriptors [63] and conformal prediction [64] etc. on the $K_{p,uu,brain}$ dataset. The work further involved examining the substructures within the dataset to make useful inferences about the structural features that have strong influence on the penetration and distribution of the molecule across the BBB.

2. METHODS

2.1. Dataset

A predictive model for $K_{p,uu,brain}$ was built in 2011 based on a set of in-house data and was uploaded into an in house server for routine usage. This study was based on a dataset consisting of 248 compounds that had the values for $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$. While a separate set of dataset consisting of other measured values of $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ consisted of 505, 3235 and 474 compounds respectively.

Since 2011, there has been additional data accumulated for these parameters. This dataset was collected and cleaned to remove duplicates and overlaps. The old dataset was then extended by the addition of the newly measured compounds (since 2011). This new set of data had 100 compounds for which the values of $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ were available.
The present dataset compiled for the model building and validation consists of 722, 1210 and 5756 compounds for $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ respectively while the $K_{p,uu,brain}$ dataset consists of 347 compounds in total.

2.2 Molecular descriptors

In this work, two types of molecular descriptors were employed which are described below. The first set, called AZ descriptors (AZdesc), an in-house descriptor set consisting of 196 2D and 3D descriptors describing various physico-chemical properties like molecular weight, lipophilicity, hydrogen bonding properties, electrostatics and topology. An in-house program, Clab, was used for the calculation of AZ descriptors with input of SMILES strings.
The second type of descriptor is the signature molecular descriptors. This descriptor was developed by Faulon *et al*. It is a class of atom based descriptor based on the concept of molecular graph. Such a molecular graph can be expressed as G = (VG, EG), where VG represents the atoms in the molecule being described while EG denoted the edges which represents the bonds between the atoms. A molecule is thus defined in terms of a set of canonical sub graphs which represent all the atoms that are at a predefined distance (height) from the central atom in consideration. Thus, for a molecular graph represented by G and an atom x in that molecule, the signature of height h of x I can be denoted by hσG(x).
Signature molecular descriptor thus explains the extended valence of the atoms of the molecule under consideration [71]. This way of representation gives a tree structure, where the first layer constitutes the neighbours of the atom x in the molecular graph G and the

subsequent layers consist of the neighbours of the vertices of the previous layer except the atom x[72]. Thus, each molecule under consideration is associated with a vector whose components are the frequency of occurrence of the particular signature in the structure of the molecule. The signature descriptors have been previously used successfully in various QSAR modeling strategy [71, 73].

2.3 Modeling methods

In the current study, for building $K_{p,uu,brain}$ models, two non-linear machine learning algorithms were used, namely, SVM and Random Forest .

2.3.1 Support Vector Machine

Support vector machine (SVM) is a supervised learning algorithm that was developed by Vapnik and co-workers [73]. It is largely based on the concepts of statistical learning or VC theory (developed by Vapnik and Chervonenkis) and structure risk minimization theory. It was originally proposed for classification but is now also widely applied in regression problems. Typically, the goal of SVM algorithm is to map a n-dimensional input vector into a high dimensional feature space and define a optimal hyperplane that can maximize the margin between the classes, in case of a binary classification problem. In case of Support vector regression the selected optimal hyperplane is the one from which the distance to all the data points is minimum. This mapping is done to the training examples to make it closer to a linearly separable case and is accomplished using a kernel function. Radial basis function is a commonly used type of kernel for SVM algorithm.

SVM algorithm depends on some hyper-parameters namely, C and gamma. C refers to the soft margin constant. These Parameters have to be optimized based on the nature of dataset under consideration. This is often done by k-fold cross validation where for each split of data into k subsets, the cross validation error is computed using different values of C and gamma. The values of C and gamma corresponding to the least cross validation error are then used for training an SVM model. SVM is particularly attractive as it effectively addresses risk of overfitting. It can handle high dimensional feature space and also local minimization.

2.3.2 Random forest

Decision trees methods are another category of widely used machine learning algorithms. In this algorithm, criteria are found for splitting the dataset into branches, thus forming a tree structure, hence the name. Each of these branches is referred to as a node and the terminal nodes are called leaves of the tree. The splitting of the nodes is achieved using some decision

rules which once established, are used to predict the future examples. A decision tree can be used in case of both regression and classification. Each end node of the tree denotes quantitative data in the former case and categorical data in case of classification problem.

These set of methods find a great utility owing to their ability to handle high dimensional data while ignoring irrelevant descriptors [74] and providing a better ease of interpretation. On the other hand, decision tree algorithms may be on the lower side with respect to the prediction accuracy. [74]

While numerous improvements have been made to such decision tree algorithms to improve its applicability, Random forest is one such improvement. Random forest is an ensemble method proposed by Leo Breiman[75] which aggregates results from multiple decision tree based learners. A random forest is a collection of trees constructed from the training dataset and validated internally to be capable of yielding predictions for future observations [76]. Every tree in the collection, called a base learner, is constructed from a bootstrap sample drawn with replacement from the original dataset. This random sample often comprises of approximately two third of the data while the remaining one third of the dataset is referred to as the ‘Out-of-bag' sample. The OOB sample is then run down the constructed tree for prediction and the error rate is computed. The main improvement in case of the random forest algorithm as compared to many of the previous decision tree based methods is the introduction of an additional layer of randomness. Instead of splitting using all the available variables, the RF algorithm selects only a random subset of variables to find the best split at each of the nodes. The trees are then grown to the maximum levels without pruning. Finally, the predictions from these ensembles of trees are combined using majority voting in case of classification problems and average values in case of regression problems. The number of descriptors considered at every node is often the parameter that has to be optimized depending on the dataset under consideration.

RF has been found to be very useful in cases where the number of Independent variables is much greater compared to the number of observations [76]. Some of the other attractive features of random forest predictors are high predictability and speed along with an inherent estimation of prediction accuracy and measures of descriptors importance. Such an ability of determining the measures of descriptor importance is of great utility in ranking the variables based on their capability to predict the response from the model.

2.3.3 Consensus models

Consensus model refers to a kind of data fusion which considers ways to combine predictions from various models. This approach, in general, has been shown to improve the predictive performance of a model. This improvement is probably attributable to the fact that when predictions from various component models are combined, the errors are averaged out and the methods show a greater accuracy by complimenting each other. A commonly used method for building consensus model is to take the average of predictions from various models.

2.3.4 Conformal predictors

Conformal predictors are a set of predictors that provide confidence for the prediction, based on past experience. These can be built on any traditional algorithm. Initially developed for classification problems, conformal prediction is now also being applied on regression problems. This method is mainly based on the usefulness of hedged predictions in analyzing datasets. Predictions are said to be "hedged" when they are associated with scores of how confident and accurate the predictor is in predicting the values. [77]

For a typical classification problem, conformal prediction associates the prediction with confidence and credibility values. The confidence score indicates how likely it is for the predicted label to be correct while credibility evaluates the suitability of the training data to classify the given test example. On the other hand, conformal prediction applied on a regression problem gives a range of confidence levels and outputs region predictions (intervals) which indicate a range of possible values at that confidence level. The main criteria to be fulfilled for being able to apply conformal prediction are that the data should follow the IID (independent and identically distributed). The algorithm first produces a point prediction and then non-conformity scores are generated which evaluate how different the test example is compared to the previous examples seen by the algorithm [78]. This score is then utilized in determining the region prediction. The conformal prediction is valid if the probability of true label lying outside the predicted region is not more than the error ϵ. Thus as the confidence level increases the width of the region increases giving rise to nested prediction sets. This width is a kind of measure for the efficiency. Thus there is a trade-off between the reliability denoted by the confidence and the accuracy.

2.4 Model building workflow

In the present study, two model building approaches were defined based on the dataset being used, namely, direct and indirect models.

- **Direct model** represents a model built using Kp,uu,brain data determined from experimental values of $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$.
- **Indirect model** consists of the three single component models ($K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$) built based on the respective experimental data, whose individual predictions on the test set are combined to calculate $K_{p,uu,brain}$ values (using equation (3)) .

Strategy of building the indirect model is to be able to effectively incorporate the experimental data available for each of the parameters, which is comparatively more compared to that of the $K_{p,uu,brain}$.

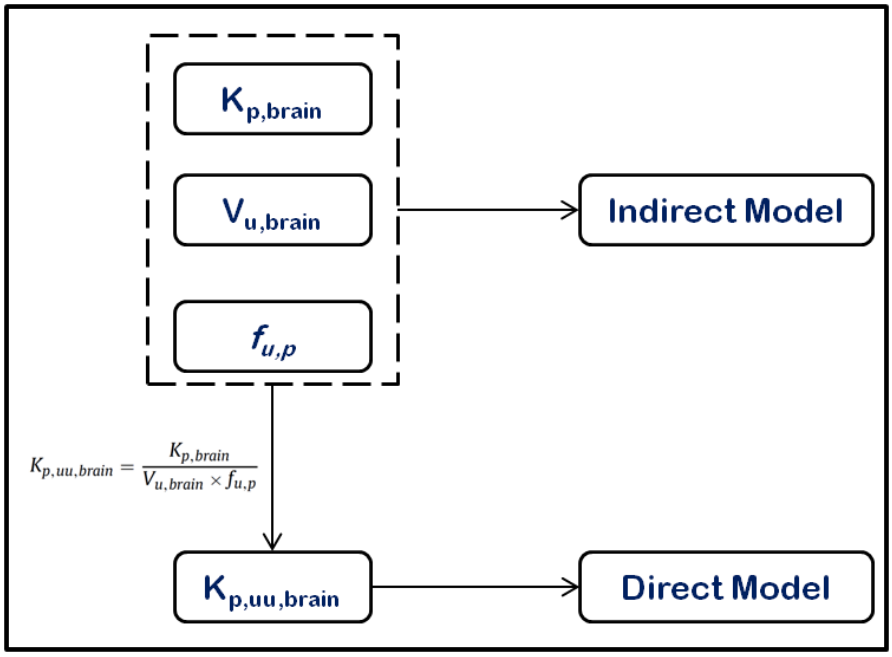


Figure 2.2: $K_{p,uu,brain}$ Model workflow

Various attempts were made to build models and improve the performance. The SVM and RF models with AZ descriptors were built using AZOrange [81], an in-house implementation of the open source package Orange [80], which is software developed for data mining. The models using Signature molecular descriptors were built using LIBSVM [79], an open source SVM library. The statistical analysis was performed using R [82] and TIBCO Spotfire. Table 2.1 explains the main models built during this study.

Table 2.1 Various $K_{p,uu,brain}$ models built

1	**I. Signature SVM model built on the old data**	O_ SVM(S,d)	Signature SVM ***Direct model*** built on the old data
2		O_ SVM(S,i)	Signature SVM ***Indirect model*** built on the old data
3	**II. Single component models**	SVM(A, d)	SVM ***Direct model*** built using AZ descriptors
4		RF(A, d)	RF ***Direct model*** built using AZ descriptors
5		SVM(S, d)	SVM ***Direct model*** built using signature descriptors
6		SVM(A, i)	SVM ***Indirect model*** built using AZ descriptors
7		RF(A, i)	RF ***Indirect model*** built using AZ descriptors
8		SVM(S,i)	SVM ***Indirect model*** built using signature descriptors
	III. Consensus models	**AZ Descriptors**	
9		SVM(A,d)_RF(A,d)	Consensus models based on the two ML algorithms, SVM and RF, and the two modeling workflows (direct and indirect), with only AZ descriptors set
10		SVM(A,i)_RF (A,i)	
11		SVM(A,d)_ SVM(A,i)	
12		RF(A,d)_ RF (A,i)	
13		SVM(A,d)_ RF (A,i)	
14		SVM(A,i)_RF(A,d)	
15		SVM(A,d)_SVM(A,i)_RF(A,d)	
16		SVM(A,d)_SVM(A,i)_RF(A,d) +RF(A,i)	
17		SVM(A,d) _RF(A,d)_RF(A,i)	
		Signature Descriptor	
18		SVM(S,d)_SVM(S,i)	Consensus model based on only signature descriptors
		AZ Descriptors & Signature Descriptor	
19		SVM(A,d)_RF(A,d)_SVM(S,d)	Consensus models based on two different descriptors, model workflows and ML algorithms
20		SVM(A,d)_RF(A,d)_SVM(S,i)	
21		SVM(A,d)_RF(A,d)_ SVM(S,d) _SVM(S,i)	
22		SVM(A,d)_RF(A,i)_RF(A,d)_SVM(S,d)	
23		SVM(A,d)_ SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)	
24		SVM(A,d)_SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)_SVM(S,i)	
25	**IV. Conformal prediction**	CP_SVM(S, i)	Conformal predictor on SVM Indirect model built based on signature descriptors

2.5 Model Validation

Validation is a vital step for building a good QSAR model as it shows the goodness of fit of the model under consideration. Two types of validations are commonly performed to assess the predictive powers of a QSAR model.

Internal validation is used to determine the model fit, which gives an idea of how an unseen example might be handled by the model. Most commonly employed methods are leave one out cross –validation Q^2, k-fold cross validation and so on. It is measured in terms of cross validated correlation coefficient or Q^2. In the present work, k-fold cross validation was used during the model validation.

External validation refers to the use of an external test set which is not included in the dataset used to build the model. The performance of the model on the test set is often measured in terms of the coefficient of determination, R^2 and RMSE.

R^2 value denotes the correlation coefficient, which describes how good the predictions from the model are. A high R^2 value is thus indicative of a good predictability of the model under consideration. In essence, R^2 basically represents the percent of data that is closest to the best fit line, thus gives a picture of how well the data under consideration is explained by the regression equation set up. The value of R^2 ranges from 0 to 1. An R^2 of zero is attributable to a case where none of the variation in the observations can be explained by the variation in the independent variables whereas a value of 1 describes an ideal case of exact explanation [38].

RMSE (Root mean square error) represents the extent by which the predicted values deviate from the true experimental values. It is calculated as the square root of mean squared error that evaluates the square of difference between the observed and the predicted values.

$$RMSE = \frac{\sqrt{(Y_{obs}-Y_{pred})^2}}{n} \quad (5)$$

Where Y_{obs} is the observed value of the dependent variable and Y_{pred} is predicted value of the dependent variable.

2.1.6 Classification model

Classification models were evaluated by categorizing the data into classes based on certain criteria. Classification criteria used in this study are as follows.

Two-class classification: The prediction results from the model were classified based on a cut off at -1, where a $\log K_{p,uu,brain}$ value greater than or equal to -1 renders the compound as BBB positive while values less than -1 implies BBB negative. BBB positive implying the

ability of the compound to permeate the BBB while BBB negative implying inability of the compound to do so.

Three-class classification: A three-class model was also built based on a cut off at two levels where a $logK_{p,uu,brain}$>=-0.52 is defined as HIGH, $logK_{p,uu,brain}$ <-1.3 as LOW and all compounds between $logK_{p,uu,brain}$ - 1.3 and -0.52 as MODERATE. Here the compounds under the HIGH category are said to have a greater chance of permeating the BBB while those in LOW class have little chance.

Throughout the analysis, the classification performance have been measured based on certain parameters. To calculate these parameters, a confusion matrix is first constructed. A confusion matrix is a matrix representing how well the prediction fits the actual values. This is analyzed based on the fraction of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) in the predictions. The primary measures used are as described in table 2.2.

- *Accuracy* describes the fraction of correctly predicted instances.
- *Sensitivity* or recall is proportion of positives that are correctly predicted as positive.
- The fraction of negatives that are correctly predicted as negatives comprise the *Specificity*.
- *Negative Precision* and *Positive precision* describe the accuracy of prediction of negative and positive class respectively.
- *F-score* provides a measure of accuracy considering the harmonic average of recall and precision.
- *Kappa* score measures the difference between the observed agreement and the agreement expected to be present just by chance.
- *Matthew's coefficient* represents the correlation between the observed and predicted binary classification.

Table 2.2 Classification performance measures

$Accuracy = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}}$	$Positive\ precision = \frac{\mathrm{TN}}{\mathrm{FN}+\mathrm{TN}}$
$Specificity = \frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}}$	$Negative\ precision = \frac{\mathrm{TP}}{\mathrm{FP}+\mathrm{TP}}$
$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$	$F-score = \frac{2 * \mathrm{Sensitivity} * \mathrm{Positive\ precision}}{\mathrm{Sensitivity} + \mathrm{Positive\ precision}}$
$Kappa = \frac{\mathrm{Accuracy}-\mathrm{Cp}}{1-\mathrm{Cp}}$ Where , $Cp = \frac{(TP+FP)*(TP+FN)*(TN+FN)*(TN+FP)}{(TP+FP+FN+TN)^2}$	
$Mathews\ coefficient = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FN} * \mathrm{FP}}{\sqrt{(\mathrm{TP}+\mathrm{FN}) * (\mathrm{TP}+\mathrm{FP}) * (\mathrm{TN}+\mathrm{FN}) * (\mathrm{TP}+\mathrm{FP})}}$	

2.6 Model Interpretation

Further, an analysis of substructures and descriptors were performed for the training sets used to build the $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ models. This gives an idea of the overall trend represented by the model. This was performed to analyze the dataset for potential indications of association of values of $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ with the substructures that they possess or the descriptors that are used to best describe them and to corroborate the data already known about the same.

2.6.1 Signature descriptor gradient

An in-house script was used to produce the SVM decision function gradient values for each of the training set predictions made by the individual models of $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$.
These were used to infer the substructure that can possibly have the most effect (positive or negative) on the values of $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ respectively.

2.6.2 AZ descriptor gradient

A similar analysis was performed with AZ descriptors using some in-house python scripts using modules from the in-house implementation of the Orange package. The gradient values were analyzed to evaluate the descriptors showing the highest positive or negative effects on the end point values.

2.6.3 VIP values (Variable importance of projection)

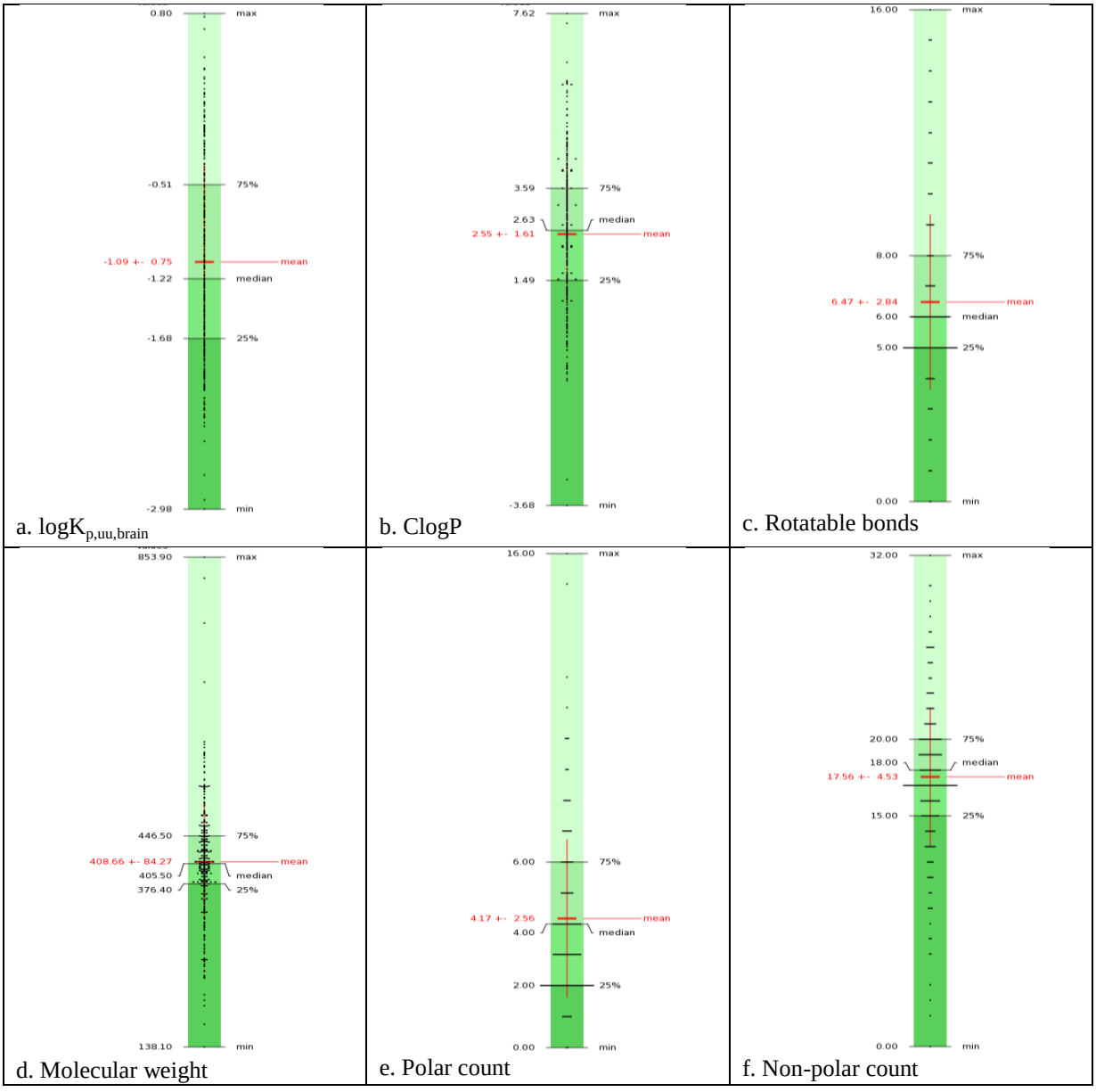
An in-house python script that calculates the VIP values based on random forest model was used to analyze the descriptor importance in each of the $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ models. The descriptors with the highest VIP values were analyzed to make specific inferences.
Based on the calculated VIP values, the important descriptors were ranked in both models using RF with AZ descriptors and RF with signature descriptors.

3. RESULTS AND DISCUSSIONS

3.1 Compounds in the dataset

As mentioned earlier, some characteristics of compounds are favorable for a good distribution across the blood-brain barrier. An overview of the trend of these properties among the compounds in the dataset used is as shown in figure below.

Figure 3.1a shows the trend of the $K_{p,uu,brain}$ values in the dataset. The average value is around -1.09± 0.75 as indicated. It can be noticed that a large majority of the data points lie in the intermediate region.



a. $logK_{p,uu,brain}$ b. ClogP c. Rotatable bonds

d. Molecular weight e. Polar count f. Non-polar count

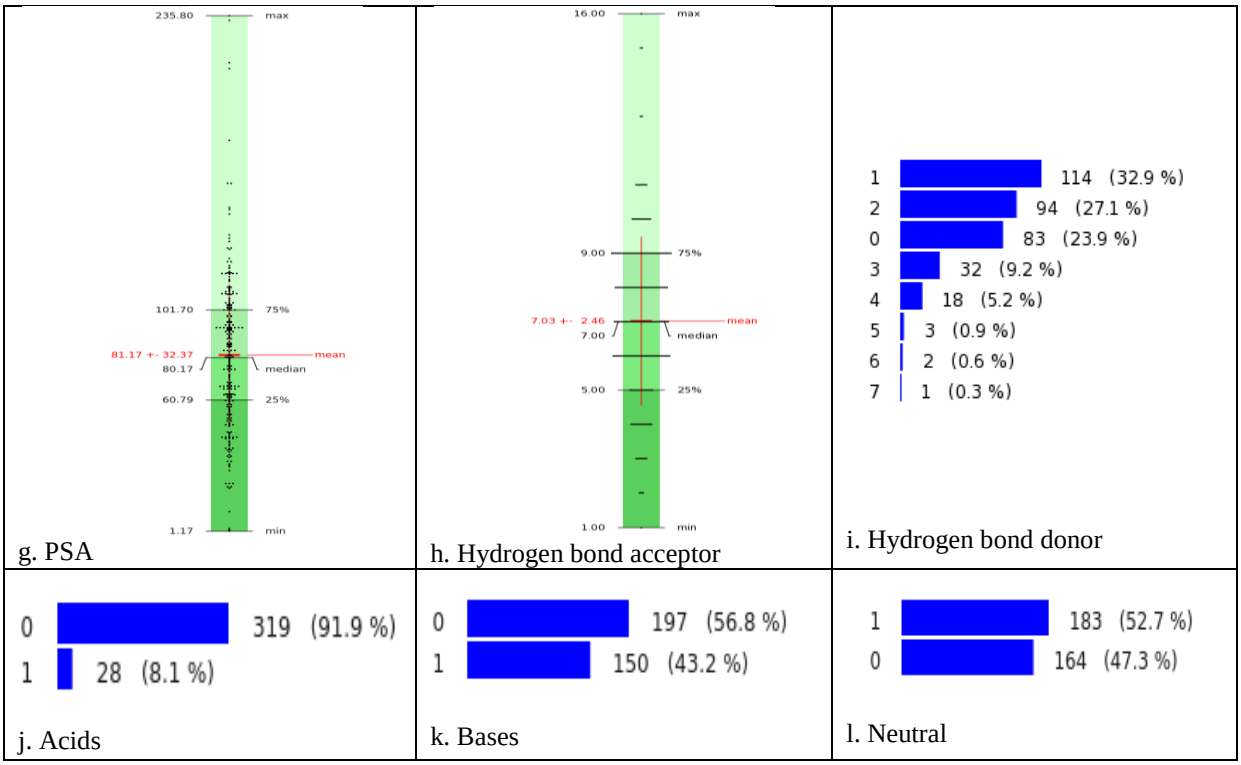


Figure 3.3: Overview of the trend of some properties across the dataset.

ClogP denotes lipophilicity of the compound, and is an important parameter for BBB penetration. Some studies have shown the mean ClogP for CNS drugs to be around 2.1[24]. The dataset used in the study shows an average ClogP value of around 2.55 ± 1.61(Figue1 (b)) which seems to be close to that value. The molecular rigidity is often defined using number of rotatable bond; a potential CNS drug is thought to have a slightly higher rigidity than Non-CNS drugs. The dataset here consists of an average rotatable bond count of around ~6.47 (Figure 1(c)).

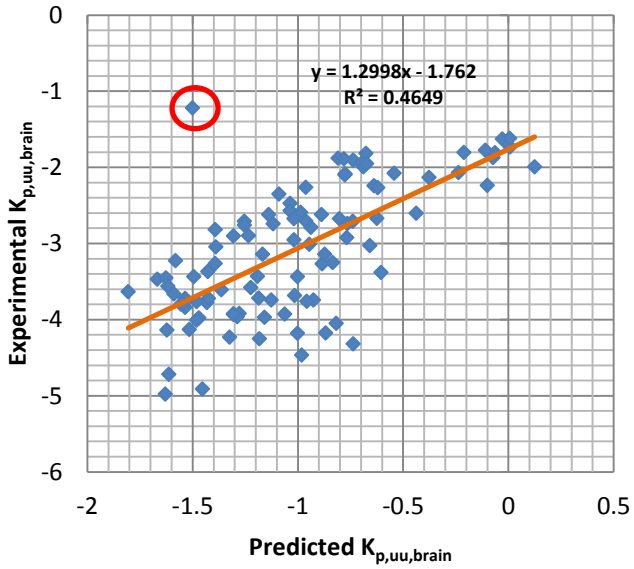
The dataset represents a set of compounds with a molecular weight average at approximately 409±84 (Figure1 (d)), which seems good as for a CNS drug the range usually suggested is around 400-450. It is also known that CNS drugs have a higher non-polar count than their non-CNS counterparts, which is reflected in the figures 1(e) and 1(f).

Hydrogen bonding properties are critical for CNS drugs, an overall picture of how HBA and HBD are distributed across the dataset is represented in Figure 1(h) and 1(i). The figures 1(j) 1(k) and 1(l) represent the details of number of acids, bases and neutral compounds in the dataset.

3.2 Validation of the Current $K_{p,uu,brain}$ model

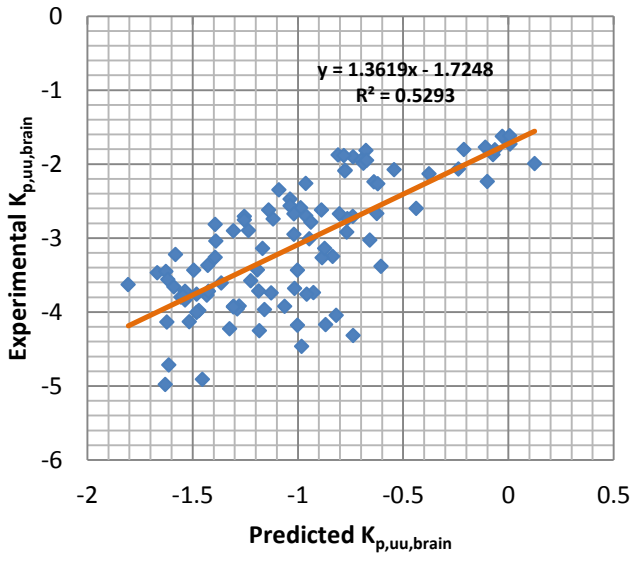
Upon the external validation of the model built in 2011 using test set consisting of compounds from the new dataset for which $K_{p,uu,brain}$ data was available (100 compounds), it was seen that the model gave a correlation coefficient, R^2 of 0.46 which increased to 0.53 on removal of a clear outlier(Figure 3.2a and b), while the RMSE decreased from 0.63 to 0.58. The original study in 2011 had seen a R^2 value of 0.58 with an external test set. This was thus indicative of the stability of the performance of the current model.

This outlier had a unusually high experimental value, which can be noticed in the distance of the point marked from the best fit line.



a.

R^2	0.46
RMSE	0.63
Observations	100



b.

R^2	0.53
RMSE	0.58
Observations	99

Figure 3.4 Prediction results on the temporal test set.

3.2.1 Classification performance:

Classification performance of the model was determined based on the two difference methods of categorization, namely, two-class classification and three-class classification.

Two class classification:

A two class classification on the prediction given by the current model on the temporal test set (with the outlier removed) is as shown (Table 3.1 and 3.2). For assessing the classification performance, firstly, a confusion matrix was constructed where the prediction was evaluated and categorized into true positives, false positives, true negatives and false negatives. Determining the performance measures showed that the prediction had a decent accuracy of around 76% with a good sensitivity and specificity of 74% and 77% respectively.

Table 3.1 Confusion Matrix for two-class classification

TP	34
FP	12
TN	41
FN	12
Total	**99**

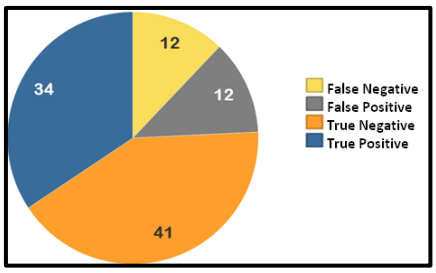


Table 3.2 Performance Measures for two-class model

Accuracy	0.76
Sensitivity	0.74
Specificity	0.77
Positive Precision	0.74
Negative Precision	0.77
F-score	0.74
Kappa	0.51
Matthews correlation coefficient	0.51

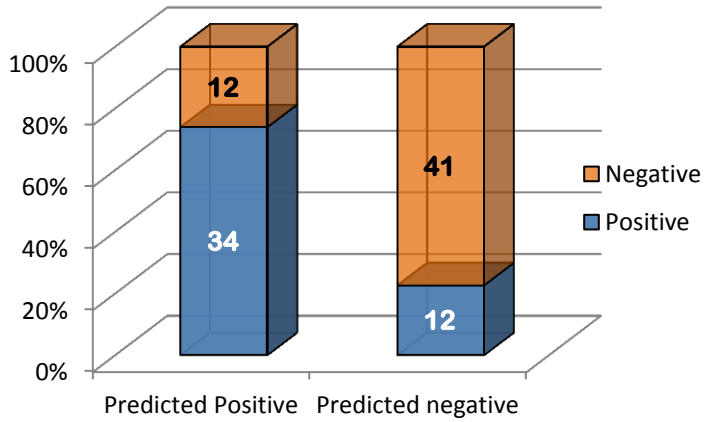


Figure 3.5 A chart representing a comparison between the experimental and predicted data of $logK_{p,uu,brain}$.

Three –class classification:

From the various performance measures calculated from the confusion matrix, a classification accuracy of 92%, 81% and 40% for HIGH, LOW and MODERATE class respectively (table 3.3), was observed. These results were again in line with what was observed in the previous study in 2011.

It is important to note that about 57% of the compounds belong to the Moderate class (table 3.3 and figure 3.4). The low prediction accuracy for the moderate class largely deteriorates the accuracy of the three class model for the whole dataset.

Table 3.3 Confusion matrix for High, Low and Moderate classes

HIGH			LOW			MODERATE	
TP	11		TP	25		TP	23
FP	1		FP	5		FP	34
Precision	0.92		Precision	0.833		Precision	0.40

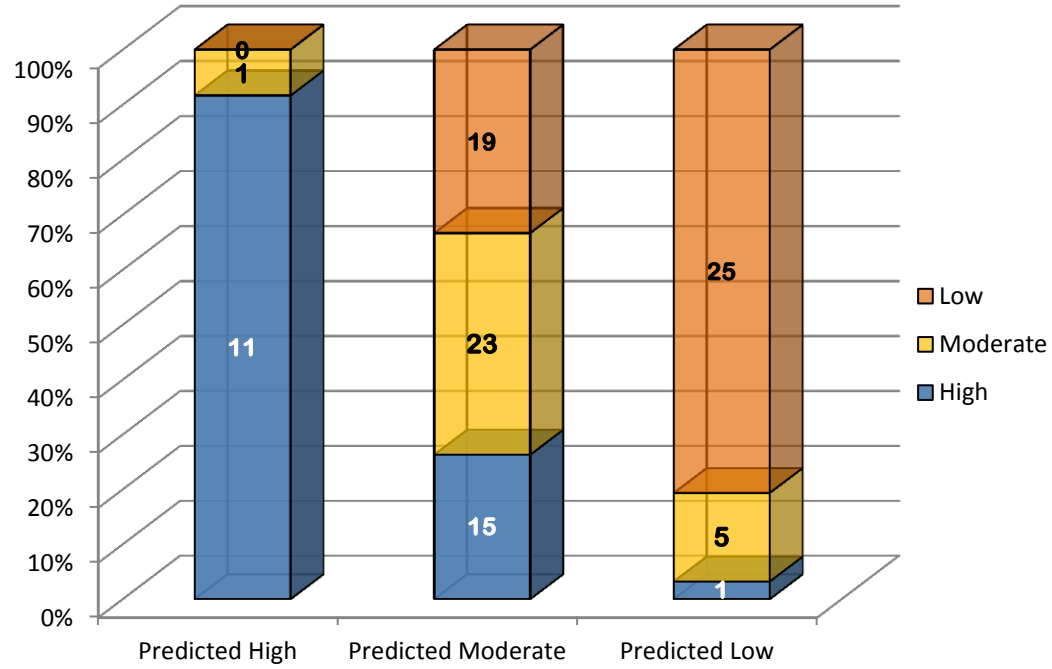


Figure 3.6 Comparison of predicted results with the experimental values using a three-class model

Overall, these observations give us a good confidence in the predictive powers of the current model. On the other hand, it is important to realize that it is necessary to update these in-silico models with temporal datasets to further extend the chemical space represented by the training set, thus further improving the chances of a good prediction of unknown compounds.

3.3 Signature SVM model built on the old dataset

After the validation of the current model, an attempt was made to check the performance of the model built using signature molecular descriptors. This model was built based on the old dataset (used in the 2011 study) to evaluate the performance using signature molecular descriptor as the study in 2011 used only the AZ descriptors set. The specifications of the model are as listed in the table below (table 3.4).

Table 3.4 Model specifications

Model type	Direct and Indirect model
Machine learning algorithm	SVM
Descriptor	Signature molecular descriptors
Size of dataset (Training set)	$K_{p,uu,brain}$: 173 (Direct model) $K_{p,brain}$: 432 $V_{u,brain}$: 399 $f_{u,p}$: 3161
Size of dataset (Test set)	74

External validation of the model gave an R^2 of 0.52 and RMSE of 0.50 for the direct model and R^2 of 0.46 and RMSE of 0.59 for indirect model (Figure 3.5 and 3.6). This was observed to be similar to the results that were obtained using the set in-house physico-chemical descriptors in the 2011 study (Direct Model R^2 of 0.53 and RMSE of 0.48 and Indirect Model R^2 of 0.42 and RMSE of 0.54)

From this study it could be inferred that signature molecular descriptor alone does not greatly improve the performance of this particular model. Further the descriptor was employed in the model building along with the AZ descriptors to determine if consensus between the AZ descriptor and the signature descriptor can further improve the model.

At this stage, different signature height ranges were checked for any improvement of the model. Since there was no significant improvement, a signature height range of 0-3 was used throughout the study.

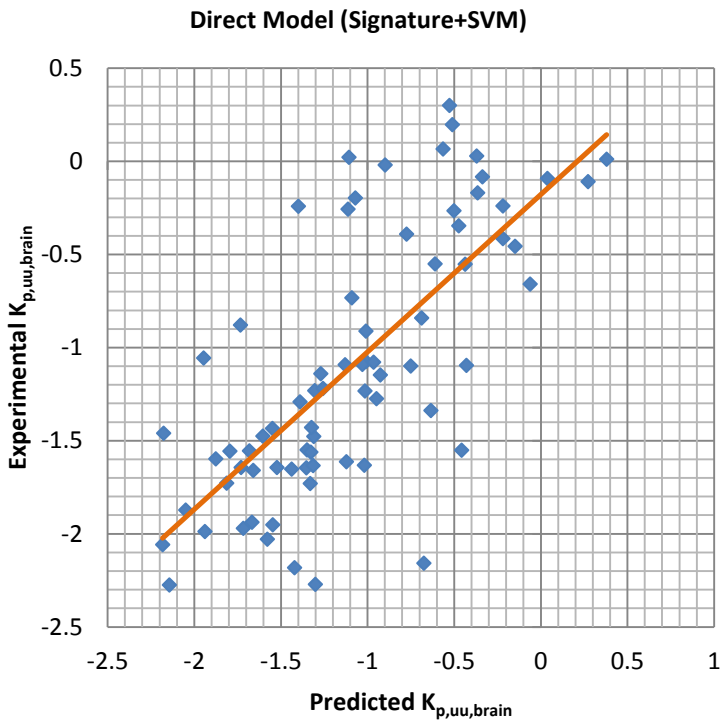


Figure 3.7 Direct Model: Predicted vs Experimental Values

Model type	Direct
R^2	0.52
RMSE	0.50

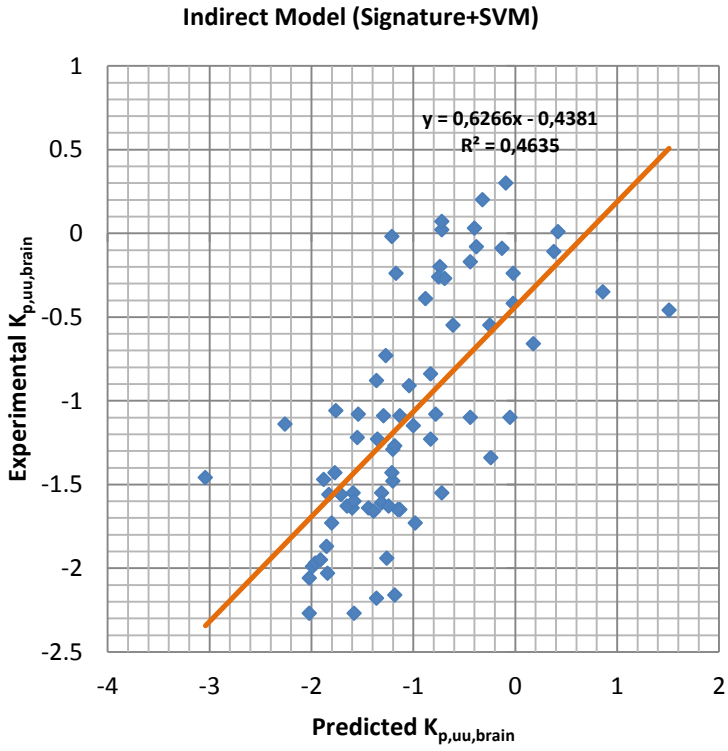


Figure 3.8 Indirect Model: Predicted vs Experimental Values

Model type	Indirect
R^2	0.46
RMSE	0.59

The subsequent studies aimed at building a Model with the inclusion of the new data in the original dataset. The outlier (represented in Figure 3.2) was removed from the dataset. It was then attempted to improve the performance of the model using a different approach to model building.

3.4 Single component models

The subsequent studies involved model building using the dataset where the new and the old data (until 2011) were combined. During the model building, the whole dataset was randomly split into training and test set with the ratio of 7:3 and random splitting was repeated 10 times and thus 10 models were built of each model type.

Table 3.6 lists and explains the notations used to represent the models, throughout the report.

Table 3.5 : Model specifications

Model type	Direct model and Indirect
Machine learning algorithm	SVM and RF
Descriptor	AZ Descriptor and Signature descriptor
Size of dataset (Training set)	$K_{p,uu,brain}$: 242 (Direct model) $K_{p,brain}$: 617 $V_{u,brain}$: 1105 $f_{u,p}$: 5651
Size of dataset (Test set)	104

Table 3.6 Notations for the single component models.

Model	**ML Algorithm**	**Model workflow**	**Descriptor**
SVM_(A, d)	SVM	Direct	AZ Descriptors
RF_(A, d)	RF	Direct	AZ Descriptors
SVM_(S, d)	SVM	Direct	Signature descriptor
SVM_(A, I)	SVM	Indirect	AZ Descriptors
RF_(A, I)	RF	Indirect	AZ Descriptors
SVM_(S, I)	SVM	Indirect	Signature descriptor

3.4.1 Internal Validation

For the models using AZ descriptor, the internal cross validation q^2 was calculated for all the datasets. The Average cross validation R^2 (q^2) for direct model was found to be 0.63 with SVM and 0.64 with RF respectively (Figure 3.7) and in case of indirect model, $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ had an average of 0.67, 0.67 and 0.80 respectively with SVM and 0.70 ,0.65 and 0.78 with RF (Figures 3.8, 3.9, 3.10). The range of R^2 values obtained with the internal validation is indicative of the good internal predictive performance of the models.

The graphs represent the trend of the variation in cross-validation R^2 (q^2) for the direct model and the 3 component models of the indirect model. The x-axis represents the models and y-axis represents the R^2 values.

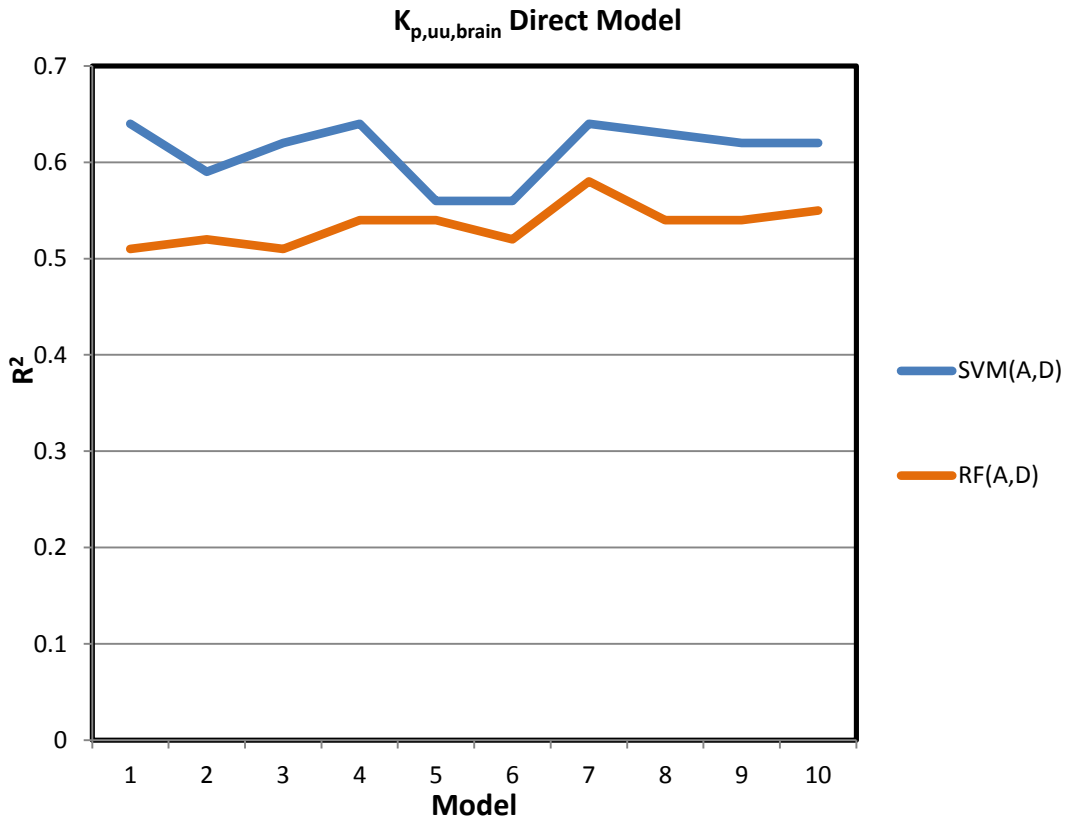


Figure 3.9 Graph representing the q2 of the 10 training sets (Direct Model).

ML Algorithm	SVM	RF
R^2	0.61	0.54
RMSE	0.46	0.50

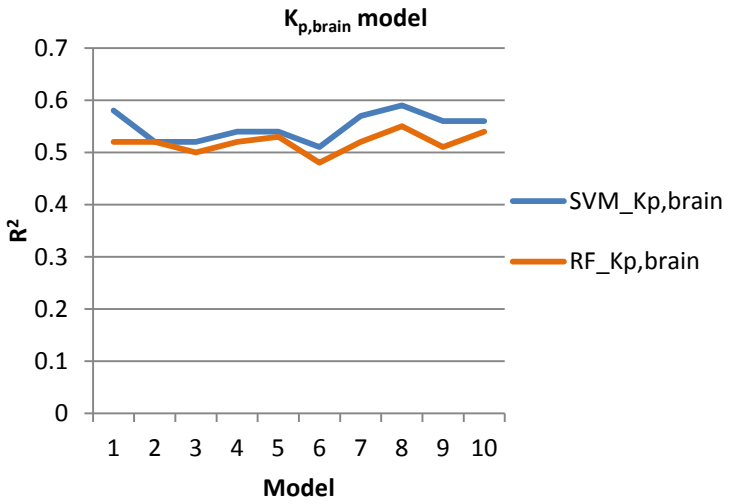


Figure 3. 10 Graph representing the q^2 values of $K_{p,brain}$ model across the 10 sets of data.

ML Algorithm	SVM	RF
R^2	0.55	0.52
RMSE	0.53	0.53

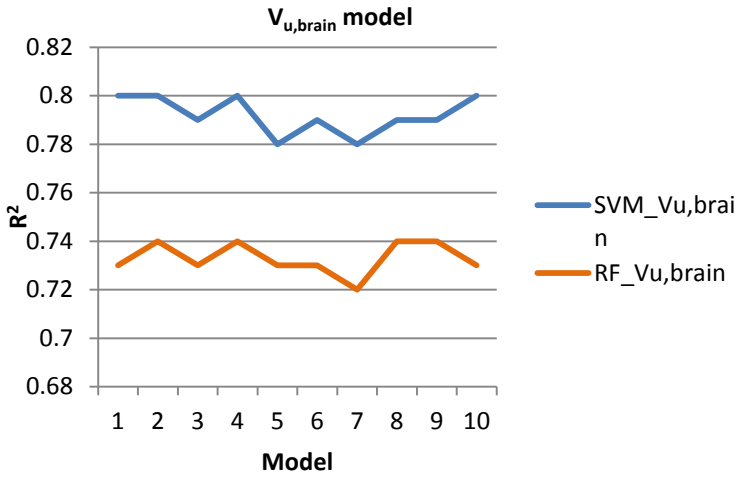


Figure 3.11 Graph representing the q^2 values of $V_{u,brain}$ model across the 10 sets of data

ML Algorithm	SVM	RF
R^2	0.8	0.73
RMSE	0.32	0.36

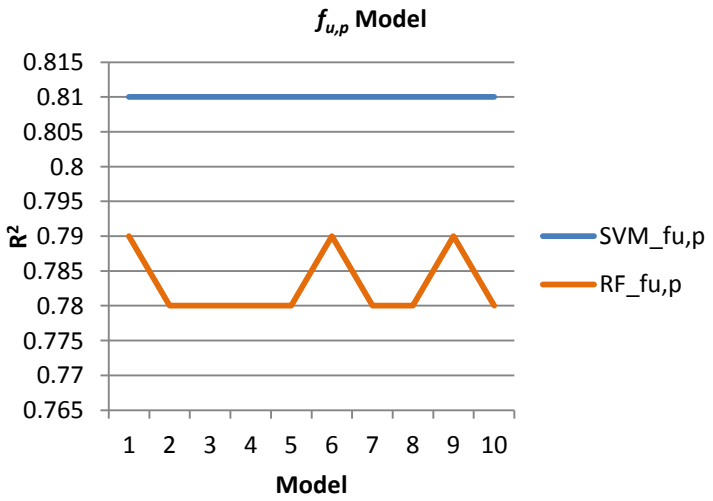


Figure 3.12 Graph representing the q^2 values of $f_{u,p}$ model across the 10 sets of data

ML Algorithm	SVM	RF
R^2	0.81	0.78
RMSE	0.38	0.4

3.4.2 External validation of the $K_{p,uu,brain}$ models

External validation involved testing the 104 compounds of the test set on the model built. The predictions on the test sets from the six single component models gave an insight into the machine learning method and the descriptors that are providing the best predictions for each of the dataset. Among the models built, it was seen that the indirect models based on SVM and RF with the in-house physicochemical descriptors gave the best predictions with an average R^2 of 0.59 and an average RMSE of 0.49 in both cases. But, the model based on RF was considered to be the best among the two, owing to the more consistent predictions as seen in the graph (Figure 3.11). This was inferred based on the average R^2 over the predictions of all the 10 models in each case (Table 3.7).

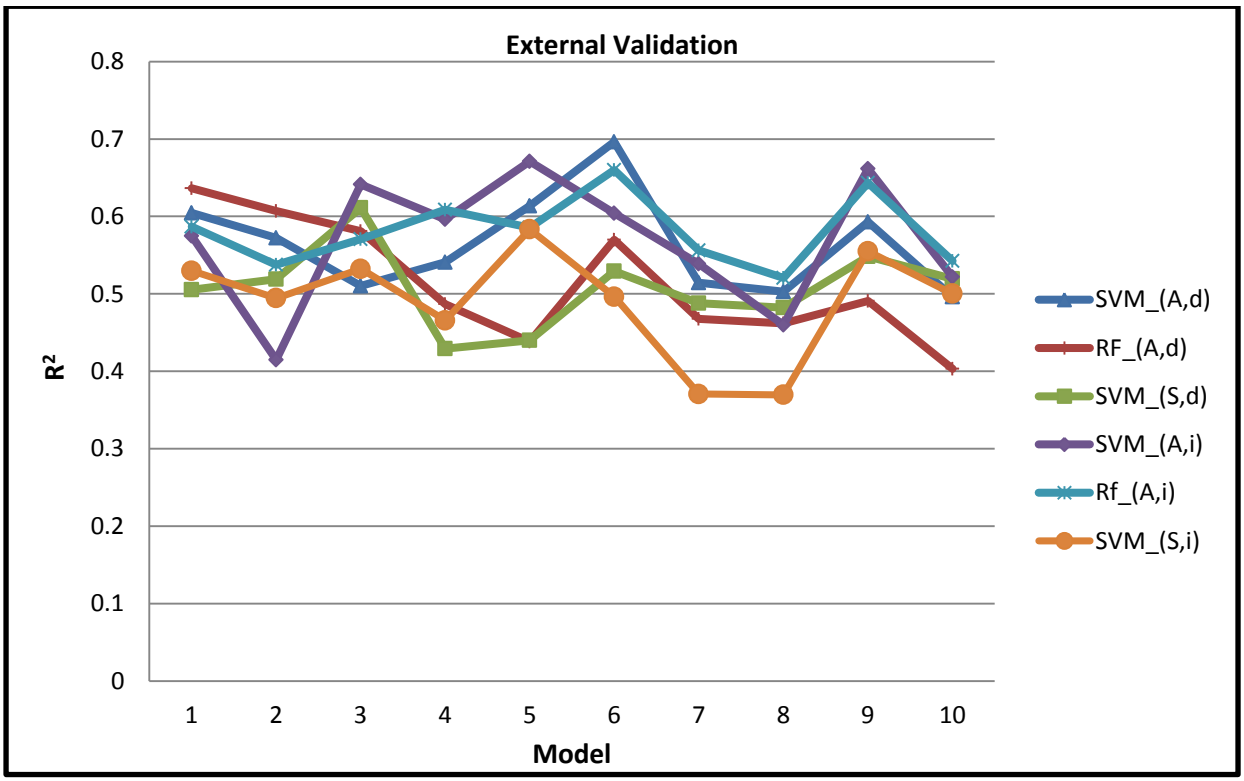


Figure 3.13 R^2 across the ten datasets for each of the models.

Table 3.7 Performance of the models (Average over 10 sets)

Model	R^2	RMSE
SVM_(A, d)	0.57	0.49
RF_(A, d)	0.54	0.52
SVM_(S, d)	0.51	0.53
SVM_(A, I)	0.59	0.49
RF_(A, I)	0.59	0.49
SVM_(S, I)	0.5	0.57

Some regression curves for the six single component models as explained above are represented with the equation and respective R^2 values (Figure 3.12). In the regression graphs, the x-axis represents the $K_{p,uu,brain}$ prediction values and y-axis represents the experimental or observed values. The line represented in the graphs is called the best fit line which basically best describes the data on the scatter plot.

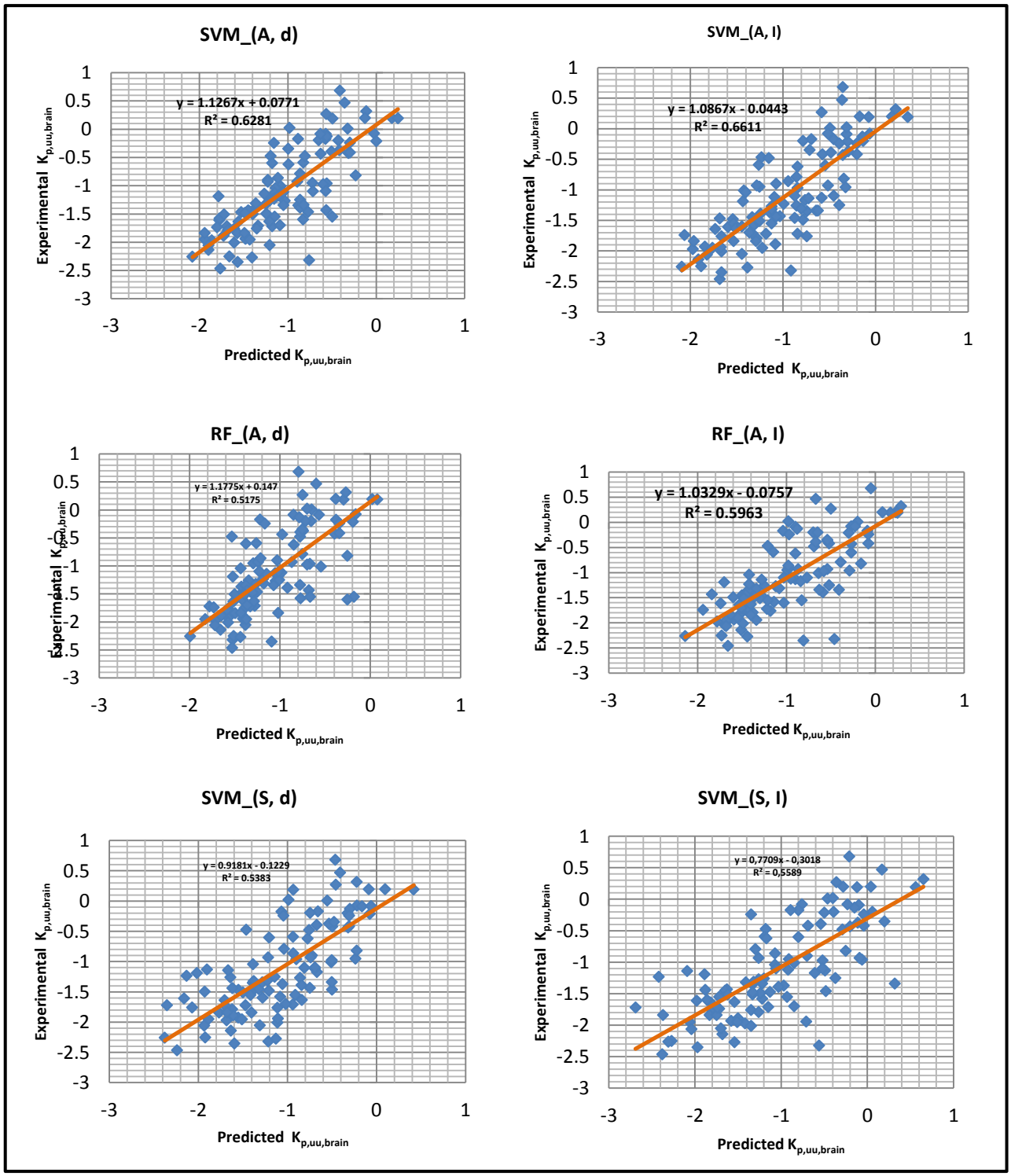


Figure 3.14 Regression curves for the six single component models of one of the datasets

3.5 Consensus Models:

Following the analysis of the single component models, consensus models were built by taking the average of the predictions from the individual components. Various combinations of the descriptors, the machine learning algorithms and modeling schemes (direct and indirect) were tried out (Table 3.8).

As it can be observed from the R^2 values, consensus model seem to perform better as compared to their single component counterparts. Most of the models gave an R^2 of above 0.6. The consensus model based on only signature descriptor seemed to perform slightly poorer compared to the ones with AZ Descriptors.

Based on the R^2 and RMSE values the best four consensus models were picked for further analysis (Figure 3.13)

Table 3.8: Consensus models

Model	**R^2**	**RMSE**
AZ Descriptor		
SVM(A,d)_RF(A,d)	0.60	0.48
SVM(A,i)_RF (A,i)	0.63	0.46
SVM(A,d)_ SVM(A,i)	0.62	0.46
RF(A,d)_ RF (A,i)	0.60	0.48
SVM(A,d)_ RF (A,i)	0.63	0.46
SVM(A,i)_RF(A,d)	0.62	0.47
SVM(A,d)_SVM(A,i)_RF(A,d)	0.63	0.46
SVM(A,d)_SVM(A,i)_RF(A,d) +RF(A,i)	0.63	0.47
SVM(A,d) + RF(A,d) + RF(A,i)	0.64	0.46
Signature Descriptor		
SVM(S,d)_SVM(S,i)	0.56	0.51
AZ Descriptors & Signature Descriptor		
SVM(A,d)_RF(A,d)_SVM(S,d)	0.61	0.48
SVM(A,d)_RF(A,d)_SVM(S,i)	0.62	0.46
SVM(A,d)_RF(A,d)_ SVM(S,d) _SVM(S,i)	0.62	0.47
SVM(A,d)_RF(A,i)_RF(A,d)_SVM(S,d)	0.63	0.46
SVM(A,d)_ SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)	0.65	0.45
SVM(A,d)_SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)_SVM(S,i)	0.65	0.45

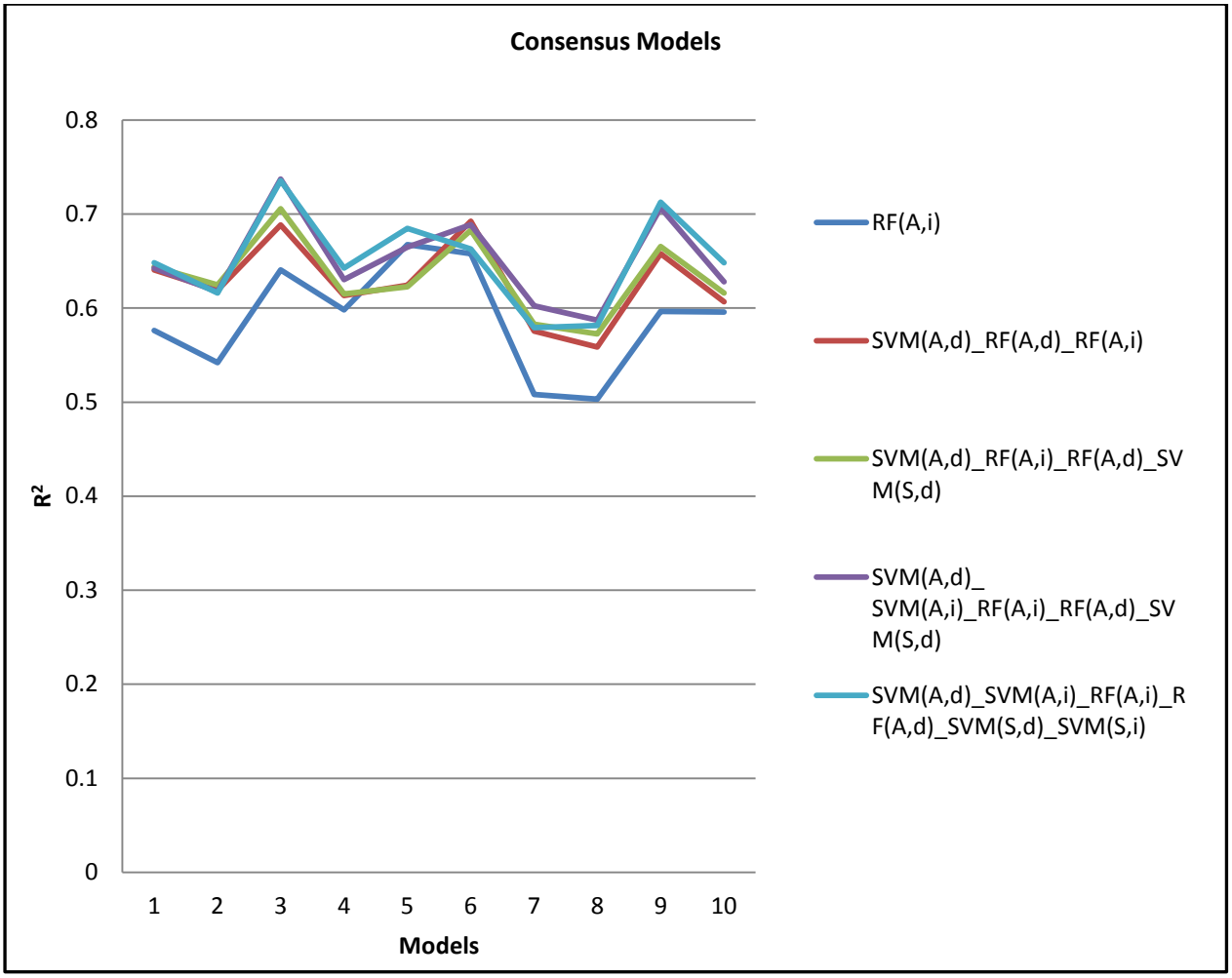


Figure 3.15 Graph showing the variation of R^2 values across the ten sets.

Further, classification performance was checked for the top four consensus models that showed the best performance.

Classification models using the best consensus models involved classifying the prediction of the model based on the 2-class ($logK_{p,uu,brain}$ >= -1 as positive and $logK_{p,uu,brain}$ < -1 as negative) or 3- class ($logK_{p,uu,brain}$ >= -0.52 (HIGH), $logK_{p,uu,brain}$ < -1.3 as (LOW), -0.52 < $logK_{p,uu,brain}$ >= -1.3 as (MODERATE)) categories. In general, different approaches exist to derive consensus models. In the current study, different approaches for calculations of consensus prediction were attempted by using Average, median, maximum value (where the maximum value among the predictions from the different model is taken as the consensus prediction) and minimum value (where the minimum value among the predictions from the different model is taken as the consensus prediction).

3.5.1 Two class classification

Table 3.9 Classification performance (Two Class) of the best consensus models using different ways of data fusion (Average, Median, Max Value and Min Value)

	Accuracy	Sensitivity	Specificity	Positive precision	Negative precision	F-score	Kappa	Matthews Coefficient
AVERAGE								
SVM(A,d) _RF(A,d)_RF(A,i)	0.836	0.802	0.858	0.799	0.861	0.8	0.66	0.659
SVM(A,d)_RF(A,i)_RF(A,d)_SVM(S,d)	0.836	0.814	0.848	0.791	0.867	0.8	0.662	0.66
SVM(A,d)_ SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)	0.843	0.823	0.852	0.8	0.874	0.809	0.677	0.675
SVM(A,d)_SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)_SVM(S,i)	0.839	0.817	0.849	0.796	0.869	0.804	0.668	0.666
MEDIAN								
SVM(A,d) _RF(A,d)_RF(A,i)	0.835	0.789	0.863	0.803	0.855	0.795	0.658	0.655
SVM(A,d)_RF(A,i)_RF(A,d)_SVM(S,d)	0.831	0.805	0.847	0.785	0.861	0.792	0.651	0.65
SVM(A,d)_ SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)	0.837	0.813	0.846	0.793	0.868	0.8	0.664	0.661
SVM(A,d)_SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)_SVM(S,i)	0.839	0.818	0.848	0.795	0.871	0.804	0.669	0.666
MAXIMUM								
SVM(A,d) _RF(A,d)_RF(A,i)	0.799	0.89	0.729	0.7	0.905	0.782	0.6	0.612
SVM(A,d)_RF(A,i)_RF(A,d)_SVM(S,d)	0.78	0.918	0.678	0.67	0.921	0.774	0.57	0.592
SVM(A,d)_ SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)	0.762	0.949	0.628	0.646	0.948	0.766	0.541	0.585
SVM(A,d)_SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)_SVM(S,i)	0.75	0.968	0.597	0.631	0.966	0.762	0.523	0.58
MINIMUM								
SVM(A,d) _RF(A,d)_RF(A,i)	0.825	0.654	0.941	0.885	0.795	0.749	0.621	0.638
SVM(A,d)_RF(A,i)_RF(A,d)_SVM(S,d)	0.816	0.607	0.96	0.917	0.775	0.729	0.6	0.624
SVM(A,d)_ SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)	0.816	0.589	0.971	0.933	0.771	0.721	0.599	0.628
SVM(A,d)_SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)_SVM(S,i)	0.806	0.565	0.972	0.935	0.76	0.702	0.574	0.609

From the 2-class classification performance measures it can be seen that consensus prediction based on minimum value has lower sensitivity and higher specificity while the one with the maximum value has a lower specificity and higher sensitivity (Table 3.9). Thus, as it can be noted, among the different methods, average calculation seemed to give a more consistent result. On an average, the models showed an accuracy of around 0.84 with a sensitivity or recall of approximately 0.81 and specificity of around 0.85. A Kappa value of around 0.67 on shows a good predictive performance of these models.

3.5.2 Three class Classification

Similar to the trend noticed with the 2-class classification results, the average method of calculating consensus prediction for 3-class model seemed to be more consistent across the models (table 3.10) . The precision fell in the range of 81-86%, 81-82% and 46-50% for the HIGH, LOW and MODERATE classes respectively.

Table 3.10 Classification performance (three-class) of the best consensus models using different ways of data fusion (Average, Median, Max Value and Min Value)

	Precision HIGH	**Precision LOW**	**Precision MODERATE**
AVERAGE			
SVM(A,d) _RF(A,d)_RF(A,i)	0.816	0.806	0.456
SVM(A,d)_RF(A,i)_RF(A,d)_SVM(S,d)	0.808	0.808	0.459
SVM(A,d)_ SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)	0.862	0.819	0.479
SVM(A,d)_SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)_SVM(S,i)	0.81	0.828	0.491
MEDIAN			
SVM(A,d) _RF(A,d)_RF(A,i)	0.815	0.806	0.454
SVM(A,d)_RF(A,i)_RF(A,d)_SVM(S,d)	0.834	0.801	0.457
SVM(A,d)_ SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)	0.85	0.817	0.479
SVM(A,d)_SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)_SVM(S,i)	0.826	0.817	0.492
MAXIMUM			
SVM(A,d) _RF(A,d)_RF(A,i)	0.674	0.874	0.438
SVM(A,d)_RF(A,i)_RF(A,d)_SVM(S,d)	0.645	0.889	0.419
SVM(A,d)_ SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)	0.653	0.885	0.397
SVM(A,d)_SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)_SVM(S,i)	0.619	0.892	0.375
MINIMUM			
SVM(A,d) _RF(A,d)_RF(A,i)	0.89	0.744	0.431
SVM(A,d)_RF(A,i)_RF(A,d)_SVM(S,d)	0.907	0.714	0.424
SVM(A,d)_ SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)	0.909	0.719	0.404
SVM(A,d)_SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d)_SVM(S,i)	0.915	0.701	0.4

From all the results described above, it can be noted that the best performance is exhibited by the 5-component model: SVM(A,d)_SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d). This model has an average R^2 of 0.65 with RMSE of 0.45, average two-class classification accuracy of

84% and an average precision of 48% for predicting Moderate class in the 3-class classification. Here we can see a clear improvement in the performance as compared to the validation result. The figures below (Figure 3.14,3.15 and 3.16) summarize sample results from one of the 10 SVM(A,d)_ SVM(A,i)_RF(A,i)_RF(A,d)_SVM(S,d) models.

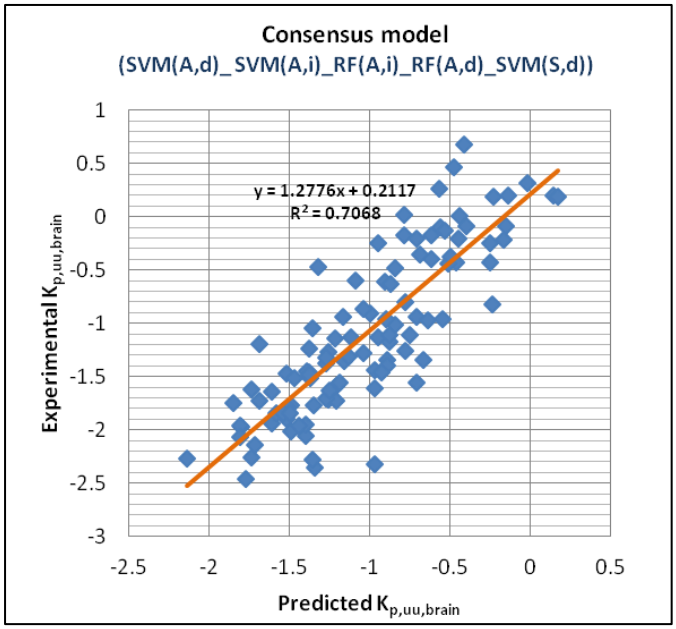


Figure 3.16 Sample regression plot for the 5 component consensus model (one of the 10 runs).

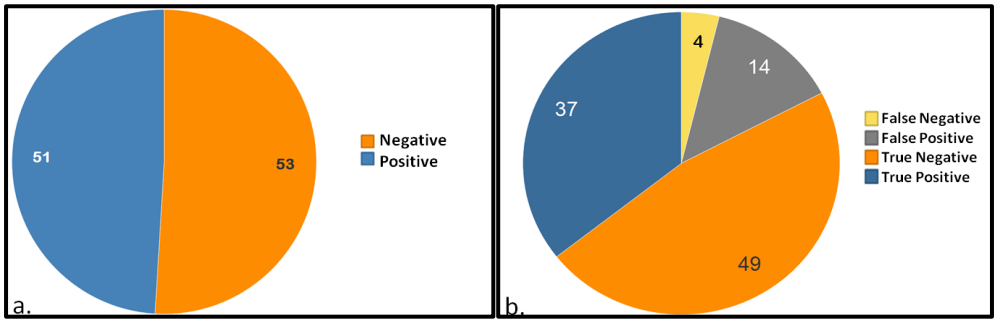


Figure 3.17 (a) Two-class classification and (b) confusion matrix for the 5-component model (one of the 10 runs)

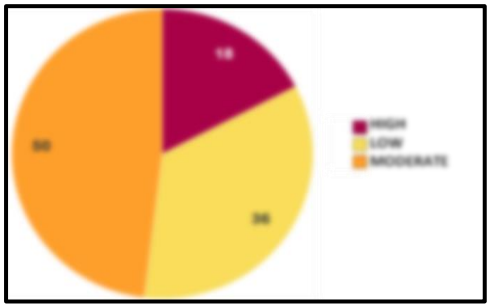


Figure 3.18: Three-class classification for the 5 component consensus model (one of the 10 runs)

3.6 Conformal Prediction

An indirect conformal prediction model was built using the dataset corresponding to the best model among the 10 runs used in section 3.5. The model utilized signature molecular descriptors. The purpose of using conformal predictors is to be able to associate the predictions from the model to a confidence score.

In this case, the error percentage and interval length at every confidence level was computed for each of $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$. $K_{p,uu,brain}$ interval was calculated by combining the minimum and maximum values of each of the components. The table below (table 3.11) shows the specifications of the model and the dataset used.

Table 3.11 Model specifications (Conformal predictor)

Model type	SVM Signature -> Conformal predictor (Indirect Model)
Machine learning algorithm	SVM
Descriptor	Signature descriptor
Size of dataset (Training set)	$K_{p,brain}$: 617 $V_{u,brain}$: 1105 $f_{u,p}$: 5651
Size of dataset (Test set)	104

To make useful inferences of the intervals obtained from the conformal predictor, experimental ranges of $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ were analyzed as summarized in table below (Table 3.12).

Range of Experimental Values:

Table 3.12 Experimental values range for the various parameter.

log$K_{p,brain}$	- 2.46 to 1.33
log$V_{u,brain}$	-0.24 to 3.52
log$f_{u,p}$	-1.98 to 2.00
log$K_{p,uu,brain}$	-2.46 - 0.68

The $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ component models (Tables 3.13, 3.14 and 3.15) were built and the predictions from these were used to calculate $K_{p,uu,brain}$. The models showed a consistency between the RMSE and the interval lengths.

Table 3.13 Prediction from $K_{p,brain}$ model.

Confidence Levels	%Error in the prediction	Average Interval Length	Average coefficient of variation
40	61.54	0.87	1.14
45	58.65	0.99	1.29
50	53.85	1.11	1.44
55	47.12	1.27	1.66
60	37.50	1.50	1.95
65	34.62	1.65	2.14
70	32.69	1.87	2.43
75	31.73	2.01	2.61
80	21.15	2.68	3.48
85	16.35	3.06	3.97
90	6.73	3.83	4.97
95	2.88	4.93	6.40

Model Prediction	
R^2	RMSE
0.57	0.53

Table 3.14 Prediction from $V_{u,brain}$ model

Confidence levels	%Error in the prediction	Average Interval Length	Average coefficient of variation
40	68.27	0.19	0.16
45	60.58	0.26	0.21
50	57.69	0.30	0.24
55	54.81	0.33	0.27
60	51.92	0.37	0.30
65	44.23	0.49	0.40
70	42.31	0.55	0.45
75	35.58	0.62	0.50
80	26.92	0.76	0.61
85	20.19	0.99	0.80
90	11.54	1.24	1.00
95	4.81	1.80	1.45

Model Prediction	
R^2	RMSE
0.75	0.36

Table 3.15 Prediction from $f_{u,p}$ model

Confidence levels	%Error in the prediction	Average Interval Length	Average coefficient of variation
40	68.27	0.19	0.17
45	64.42	0.24	0.21
50	53.85	0.32	0.29
55	44.23	0.39	0.35
60	41.35	0.46	0.42
65	34.62	0.53	0.48
70	30.77	0.58	0.52
75	23.08	0.77	0.69
80	18.27	0.88	0.79
85	14.42	1.10	0.99
90	9.62	1.52	1.37
95	7.69	1.69	1.53

Model Prediction	
R^2	RMSE
0.76	0.36

$K_{p,uu,brain}$ Calculations:

For calculating $K_{p,uu,brain}$ value range the formula the formula as given in the table (3.16) below was used. This formula is based on the Equation (3).

Table 3.16 Equation used to calculate Kp,uu,brain ranges.

Maximum $\log K_{p,uu,brain} = \text{Max}\log K_{p,brain} - \text{Min}\log V_{u,brain} - \text{Min}\log f_{u,p} + 2$
Minimum $\log K_{p,uu,brain} = \text{Min}\log K_{p,brain} - \text{Max}\log V_{u,brain} - \text{Max}\log f_{u,p} + 2$

From Table 3.17, it can be noted that the conservative validity of the conformal prediction is not satisfied as the error rates at each confidence level do not correspond to the ϵ and $1-\epsilon$ relationship (5% error at 95% confidence). This can probably be overcome with a different approach to calculation of the $K_{p,uu,brain}$ interval. The interval length increases from 1.26 (at 40% confidence) to 8.42 (at 95% confidence) these high interval lengths make the interpretation of the results somewhat complex. However, the model performance in terms of the RMSE seems to be consistent with the interval length (low RMSE).

Table 3.17 Prediction from $K_{p,uu,brain}$ indirect model

Confidence levels	%Error in the prediction	Average Interval Length	Average coefficient of variation
40	47.12	1.26	0.17
45	41.35	1.49	0.21
50	33.65	1.73	0.29
55	27.88	2.00	0.35
60	18.27	2.33	0.42
65	13.46	2.67	0.48
70	9.62	3.00	0.52
75	6.73	3.39	0.69
80	3.85	4.32	0.79
85	1.92	5.15	0.99
90	0	6.58	1.37
95	0	8.42	1.53

Model Prediction	
R2	**RMSE**
0.48	0.58

As described by the Three Sigma rule (or 68-95-99.7 rule), for a normally distributed dataset, ±1 standard deviation interval of the mean is where ~68% of the data points lie.

Thus checking for the average for experimental values of $K_{p,brain}$, $V_{u,brain}$ and *fu,p* and doubling the value we get the interval length where 60% of the data points are to lie, assuming that the data is normally distributes (Table 3.18).

To understand the interval length given by the conformal prediction, experimental interval length was determined as above and compared to the interval lengths in case of confidence levels of 60%, 65% and 70% (Table 3.19).

Table 3.18 Comparison of the standard deviation and interval length.

	Avg Standard deviation	Interval Length
$K_{p,brain}$	0.1069	0.2138
$V_{u,brain}$	0.2208	0.4416
fu,p	0.1008	0.2016

Table 3.19 Comparison of the interval lengths at the confidence levels of 60, 65 and 70%.

	Interval Length		
	60	**65**	**70**
$\mathbf{K_{p,brain}}$	1.5	1.65	1.87
$\mathbf{V_{u,brain}}$	0.37	0.49	0.55
$\boldsymbol{f_{u,p}}$	0.46	0.53	0.58

It can be noted from the above results that the interval length obtained $K_{p,brain}$ and $f_{u,p}$ are much higher than the experimental interval lengths thus making it difficult to give accurate predictions with high confidence, while the interval length in case of $V_{u,brain}$ seems to be slightly better (which is again consistent with the better model performance as shown by the R^2 and RMSE values).

Conformal prediction probably points to the inherent noise in the experimental measurement data due to which the IID (Independent and identically distributed) assumption may not be satisfied.

However, in all cases shown above, the conformal prediction results were consistent in comparison to the model performance as observed from the R^2 and RMSE values.

3.7 $K_{p,brain}$ Modeling approach

This model was an approach to understand if improving the $K_{p,brain}$ component alone can improve the $K_{p,uu,brain}$ prediction of the model while utilizing experimental data for $V_{u,brain}$ and *fu,p*.

Among the various $K_{p,brain}$ consensus models, the model with all the 3 component models (SVM+AzDesc, RF+AZDesc and SVM+Signature descriptors) gave the best performance with an R^2 of 0.72 and a RMSE of 0.47. This model was used to build a kind of indirect model where the $K_{p,brain}$ consensus model was used along with the experimental $V_{u,brain}$ and *fu,p* to calculate $K_{p,uu,brain}$. Values obtained from such a calculation are as represented in the table below (Table 3.20). Average over all the ten models gave an R^2 of 0.64 and RMSE of 0.45.

Comparing these results to the values previously obtained from the consensus models, it can be seen that this model has a performance comparable to the performance of the best consensus models.

Table 3.20 Model Performance

Dataset	R^2	RMSE
1	0.68	0.41
2	0.62	0.49
3	0.64	0.46
4	0.62	0.48
5	0.67	0.48
6	0.6	0.48
7	0.61	0.43
8	0.54	0.46
9	0.67	0.43
10	0.72	0.4
Average	**0.64**	**0.45**

For these models, classification performance was also evaluated and averaged over the 10 runs (Table 3.21). Two-class classification gave an accuracy of around 84%, while with 3 class, a precision of 81.4%, 82.3% and 51.3 % was obtained for HIGH, LOW and MEDIUM class respectively (Table 3.22). However, it can be noted that the precision for moderate class is somewhat higher than the previous models.

Table 3.21 Two class classification performance.

Two-class classification	
Accuracy	0.839
Sensitivity	0.791
Specificity	0.867
Positive Precision	0.809
Negative Precision	0.856
F-score	0.8
Kappa	0.666
Matthews correlation coefficient	0.662

Table 3.22 Three class classification performance

Three-class classification	
Precision HIGH	0.814
Precision LOW	0.823
Precision MODERATE	0.513

3.9 $K_{p,uu,brain}$ Model Interpretation

3.9.1 Random Forest VIP values

The table below summarizes the analysis of the descriptors based on RF VIP values (Table 3.23).

Among the top 15 descriptors for $K_{p,uu,brain}$ most of them were related to polarity and molecular topology. These properties have been previously shown to be important determinants of $K_{p,uu,brain}$. However, it is also noticed that lipophilicity, though a very important factor for total brain-plasma concentration ratio ($K_{p,brain}$), does not appear in the list of top influential descriptors based on RF VIP values (ClogP). This is due to the fact that higher lipophilicity will lead to higher non-specific brain tissue and plasma protein binding and in the end has little influence on the $K_{p,uu,brain}$.

Table 3.23 Top 15 descriptors based on RF VIP.

Descriptor	VIP	
MM_SAS_EP_P_SUM	0.023	Sum of positive electrostatic potentials on solvent accessible surface.
HBAsum	0.021	Sum of acceptor free energies according to Raevsky (HYBOT).
Kappa2	0.021	Topological index.
VDW_AREA	0.016	Van der Waals molecular surface area.
MM_SAS_EP_P_MEAN	0.016	Mean of positive electrostatic potentials on solvent accessible surface.
MM_VDW_EP_P_SUM	0.015	Sum of positive electrostatic potentials on Van der Waals surface.
Kappa1	0.014	Topological index.
CMR	0.014	Calculated molar refractivity. Largely a volume descriptor, highly correlated with molecular weight.
HBAmax	0.012	Highest free energy factor for H-bond acceptors according to Raevsky (HYBOT).
OVAL_NEW	0.012	TSA / the area of a sphere with the volume given by MolVol2D
HBD	0.012	Lipinski number of HB donors = number of OH+NH.
MM_VDW_EP_P_AREA	0.012	Area of Van der Waals surface with positive electrostatic potential.
AREA	0.012	Van der Waals radius surface, summed over all atoms, with a 1-3 overlap correction.
Chi3p	0.011	Sum of reciprocal square roots of valences over all 4-count linear atom paths.
VOL	0.011	Gaussian volume. A measure of molecular volume.

3.9.2 AZ descriptor gradient values

The most Influential AZ descriptors were also determined using the AZ Descriptor model with SVM. This was determined based on SVM decision function values. Based on the average decision function gradient values, a list of descriptors that potentially influence $K_{p,uu,brain}$ positively (Table 3.24) and negatively (Table 3.25) was highlighted. Based on the standard deviation within the gradient values, the descriptors that are not very influential in spite of having high average gradient values were removed from the list (Standard deviation > absolute value of the average gradient values).

This calculation performed on $K_{p,uu,brain}$ data produced results that are very complex to interpret.

Table 3.24 Top 15 positively influential descriptors based on SVM decision function gradient

Descriptor	Average Gradient value	Standard deviation	Median	Description
MM_HASA	4.168	2.554	3.737	A measure of the dispersion of the charge on hydrogen bond acceptor atoms on the surface.
PolarCountMW	1.783	1.383	1.693	Polar count divided by molecular weight
MM_HADSA	1.024	0.852	0.916	A measure of the dispersion of the charge on hydrogen bond donor and acceptor atoms on the surface.
MM_QnegVar	0.938	0.853	0.862	Variance of negative charges .
NonpolarCountMW	0.861	0.524	0.913	Nonpolar count divided by molecular weight
MaxNegChargeGM	0.644	0.582	0.560	Maximum negative charge using the Gasteiger-Marsili partial charge equilibration.
AverNegCharge_GM	0.510	0.368	0.469	Average negative charge using the Gasteiger-Marsili partial charge equilibration.
MM_QposMean	0.264	0.166	0.271	Mean of positive charges.
SPEC_SAS_NONPOL_AREA	0.202	0.131	0.192	SAS_NONPOL_AREA / SAS_TOT_AREA.
SIC	0.196	0.139	0.204	Structural information content of 0 order.
SPEC_HB_TOT	0.187	0.117	0.175	HBsum/HeavyAtomCount.
SPEC_VDW_HB_A_AREA	0.173	0.117	0.180	VDW_HB_A_AREA / VDW_AREA.
MM_HACA	0.163	0.106	0.152	A measure of the dispersion of the charge on hydrogen bond acceptor atoms on the surface.)
FractionNeutral	0.116	0.112	0.091	10 ^ (ACDlogD74 - ACDlogP)
AverNegCharge_GH	0.106	0.081	0.076	Average negative charge using the Gasteiger-Huckel partial charge equilibration.

Since $K_{p,uu,brain}$ is a parameter involving many complex processes, interpretation of these descriptor is complex.

Further work has to be carried out to be able to get a fundamental understanding of the listed influential descriptors.

Table 3.25 Top 15 Negatively influential descriptors based on SVM decision function gradient

Descriptor	Average Gradient value	Standard deviation	Median	Description
MM_HDSA	-4.572	3.677	-4.992	A measure of the dispersion of the charge on hydrogen bond donor atoms on the surface.
MM_QnegMean	-0.636	0.323	-0.659	Mean of negative charges.
SPEC_SAS_HB_D_AREA	-0.568	0.213	-0.557	SAS_HB_D_AREA / SAS_TOT_AREA.
SPEC_VDW_HB_D_AREA	-0.364	0.196	-0.373	VDW_HB_D_AREA / VDW_AREA.
SPEC_FLEX_BND	-0.276	0.110	-0.286	Defined as ratio FLEX_BND/HEAVIES.
MM_HDCA	-0.274	0.174	-0.293	A measure of the dispersion of the charge on hydrogen bond donor atoms on the surface.
SPEC_SAS_POL_AREA	-0.202	0.131	-0.192	SAS_POL_AREA / SAS_TOT_AREA.
OVAL_NEW	-0.155	0.094	-0.122	TSA / the area of a sphere with the volume given by MolVol2D
FractionIonized	-0.116	0.112	-0.091	(1 - FractionNeutral)
MinEV3	-0.109	0.068	-0.098	3rd smallest minimum eigenvalue from connectivity matrix, where diagonal has atomic weights.
SPEC_VDW_POL_AREA	-0.100	0.080	-0.094	VDW_POL_AREA / VDW_AREA
HBAmax	-0.070	0.045	-0.076	Highest free energy factor for H-bond acceptors according to Raevsky (HYBOT).
LUMO	-0.063	0.045	-0.059	Huckel molecular orbitals, Lowest unoccupied molecular orbital energy.
Balaban	-0.059	0.055	-0.034	Topological distance matrix based index related to ring structures.
MaxEV2	-0.014	0.008	-0.015	2nd largest maximum eigenvalue from connectivity matrix, where diagonal has atomic weights.

3.9.3 Signature gradient values

Decision function Gradient values were generated by a script that calculates the signature descriptors and builds an SVM model. These values somewhat represent the effect of the particular signature on the end point value. Thus this can be used to identify substructures within the molecule that could potentially exert a positive or negative effect.

Based on the average gradient values, and analysis of other statistical parameters like standard deviation a set of top signatures have been selected and represented in the tables below (Tables 3.26 and 3.27).

Complexity of the parameter $K_{p,uu,brain}$ is reflected in the absence of an exact trend in the substructures obtained with signatures as with other methods (Figures 3.17 and 3.18). This necessitates a detailed study of the substructures obtained to possibly find any solid correlations with $K_{p,uu,brain.}$

Table 3.26 Top positively influential signatures for $K_{p,uu,brain}$

Signature	Number of Occurences	Average gradient values	Standard deviation (of gradient values)	Classification (3- class)	Average Experimental Value	Standard Deviation (of Experimental values)
[C](p[C](p[C])p[C]([C]p[C]))	147	0.0641	0.0095	(L=66,H=33,M=48)	-1.12	0.75
[N](p[C]p[C])	103	0.0509	0.0093	(L=38,H=28,M=37)	-1.02	0.78
[C](p[C](p[C])p[C](p[C][F]))	33	0.0427	0.0101	(L=12,H=11,M=10)	-0.96	0.81
[C]([C]p[C]p[C])	175	0.0424	0.0071	(L=76,H=42,M=57)	-1.09	0.75
[C](p[C]([C]p[C])p[C](p[C][C]))	30	0.0411	0.0037	(L=12,H=4,M=14)	-1.20	0.55
[C]([C](p[C]p[N]))	23	0.0325	0.0034	(L=6,H=12,M=5)	-0.52	0.84
[C](p[C][N]p[N])	16	0.0323	0.0033	(L=2,H=10,M=4)	-0.45	0.77
[C](p[C](p[C])p[C](p[C])[C](p[C]p[C]))	23	0.0321	0.0038	(L=0,H=13,M=10)	-0.25	0.50
[C]([C]([C])[C]([C])[C]([C]))	13	0.0312	0.0038	(L=3,H=5,M=5)	-0.76	0.63
[F]([C](p[C](p[C])p[C](p[C])))	21	0.0309	0.0087	(L=8,H=7,M=6)	-0.99	0.81
[C](p[C](p[C])p[C](p[C])[F])	21	0.0309	0.0087	(L=8,H=7,M=6)	-0.99	0.81
[C](p[C](p[C])p[C](p[C])[C]([N]))	16	0.0294	0.0018	(L=1,H=6,M=9)	-0.56	0.44
[N](p[C]([C]([C][O])p[N](p[N,0]))p[N](p[N,0]([C])))	10	0.0293	0.0025	(L=3,H=3,M=4)	-0.93	0.54
[N](p[C]([C]([C][O])p[N](p[N,0]))p[N]([C](p[C]p[C])p[N,0]))	10	0.0293	0.0025	(L=3,H=3,M=4)	-0.93	0.54
[C]([C]([C][O])p[N](p[N])p[N](p[N]))	10	0.0293	0.0025	(L=3,H=3,M=4)	-0.93	0.54

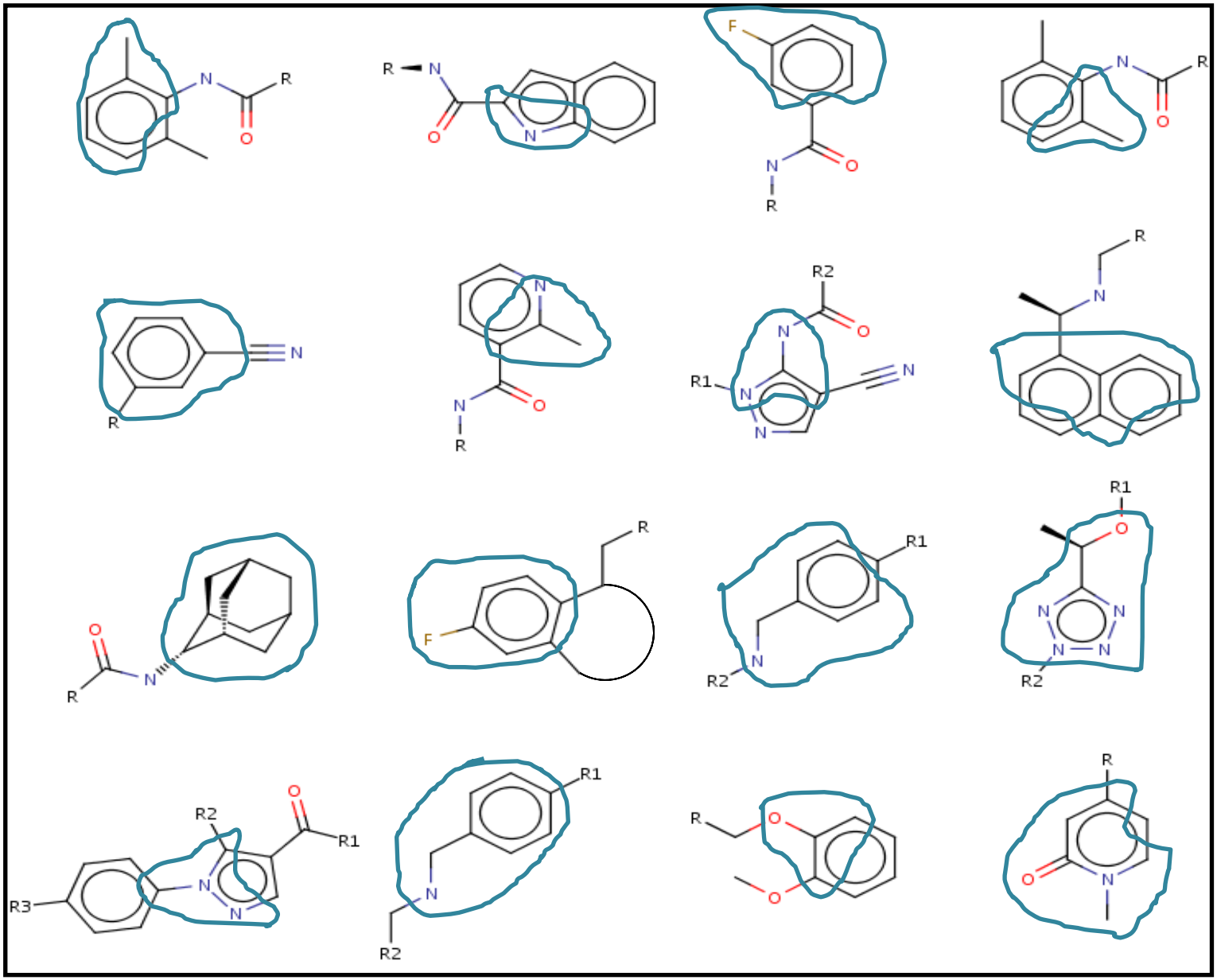


Figure 3.19 Substructures showing positive influence on $K_{p,uu,brain}$.

Table 3.27 Top Negatively influential signatures for $K_{p,uu,brain.}$

Signature	Number of Occurences	Average gradient values	Standard deviation (of gradient values)	Classification (3- class)	Average Experimental Value	Standard Deviation (of Experimental values)
[C]([C][C][N])	116	-0.0706	0.0106	(L=67,H=16,M=33)	-1.35	0.69
[C]([C](=[C])[N]([C])=[O])	12	-0.0613	0.0048	(L=8,H=0,M=4)	-1.65	0.43
[O](=[C]([C](=[C])[N]([C])))	12	-0.0613	0.0048	(L=8,H=0,M=4)	-1.65	0.43
[C](p[C](p[C])p[C](p[C][C]))	95	-0.0506	0.0096	(L=37,H=17,M=41)	-1.12	0.67
[N](p[C]([C]p[N])p[N](p[C]))	69	-0.0399	0.0021	(L=44,H=4,M=21)	-1.41	0.47
[C](p[C](p[C])p[C](p[C])[C]([N]=[O]))	24	-0.0384	0.0087	(L=15,H=2,M=7)	-1.47	0.56
[C]([C]p[N]p[N])	75	-0.0375	0.0031	(L=45,H=6,M=24)	-1.36	0.52
[C]([C]([C][C][N]))	13	-0.0364	0.0171	(L=10,H=2,M=1)	-1.56	0.79
[C]([C]([C][C]))	11	-0.0361	0.0138	(L=7,H=0,M=4)	-1.56	0.39
[N]([C]([C])[C]([C])[C]([C][C]))	23	-0.0349	0.0085	(L=14,H=1,M=8)	-1.46	0.44
[C]([C](=[C]([C,0])[C](p[N]p[N]))=[N]([N]([C,0])))	22	-0.0339	0.0029	(L=19,H=0,M=3)	-1.67	0.32
[C]([C](=[C]([C][C,0]))[N]([N](=[C,0]))=[O])	22	-0.0339	0.0029	(L=19,H=0,M=3)	-1.67	0.32
[C]([C](=[C])[N]([N])=[O])	22	-0.0339	0.0029	(L=19,H=0,M=3)	-1.67	0.32
[C]([C]([N]([N,0])=[O])=[C]([C](=[N,0])[C](p[N]p[N]))|0|	22	-0.0339	0.0029	(L=19,H=0,M=3)	-1.67	0.32
[N]([C]([C](=[C,0])=[O])[N](=[C]([C,0]))|0|	22	-0.0339	0.0029	(L=19,H=0,M=3)	-1.67	0.32

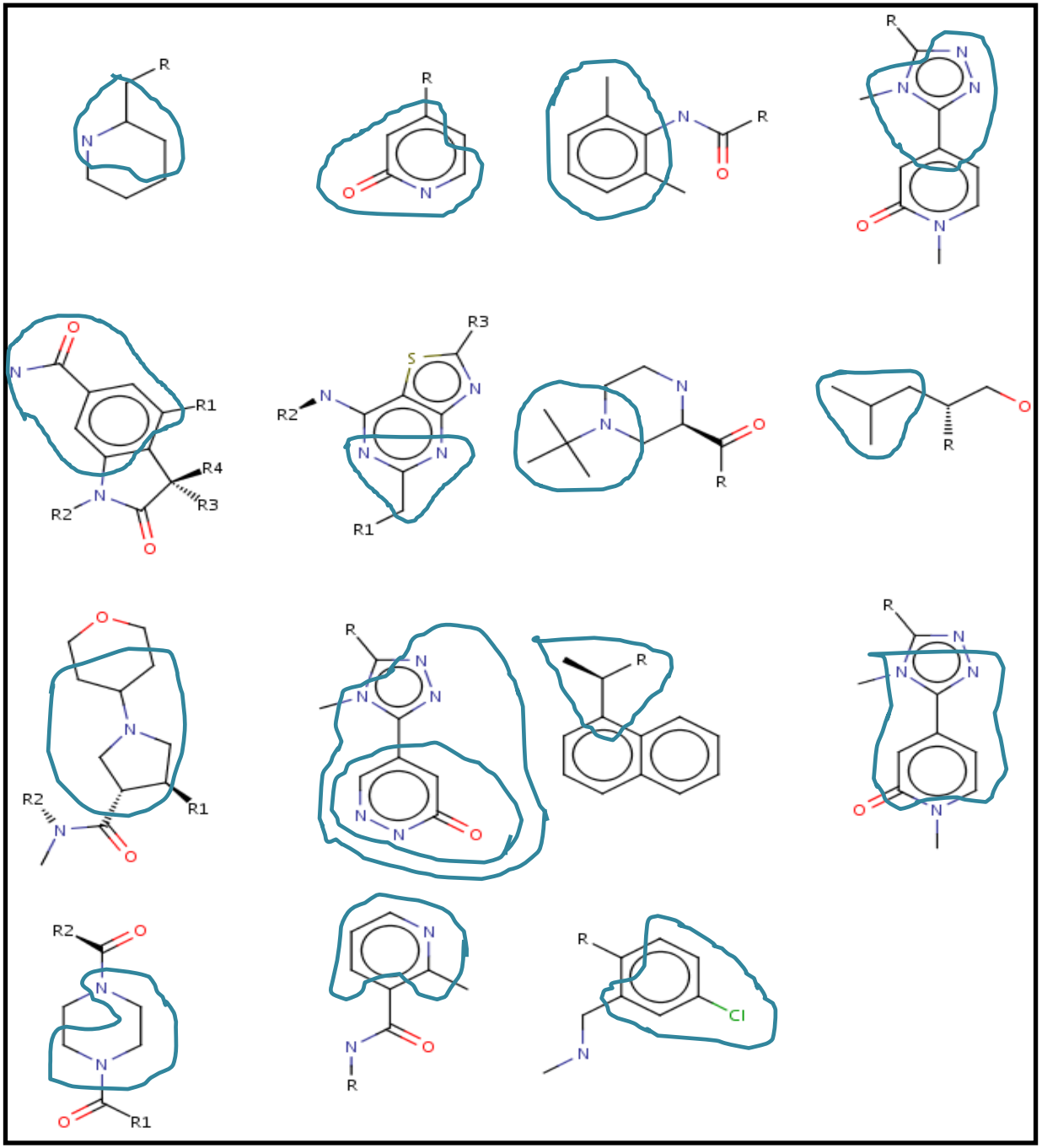


Figure 3.20 Substructures showing negative influence to $K_{p,uu,brain}$.

3.10 Descriptor analysis for the individual components ($K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$)

While the substructure analysis for $K_{p,uu,brain}$ proved to be quite complex, it was attempted to do such an analysis on the individual models of $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ using the RF VIP value analysis and the signature descriptor gradient analysis.

3.10.1 RF VIP:

Calculation of RF VIP values for the three component models $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ produced results where some clearly vital descriptors came at the top positions.

The table 3.28 shows the analysis on the $K_{p,brain}$ model. The top descriptor in the list came up to be PSA which has been established as a very important determinant for the $K_{p,brain}$. There were also other descriptors based on hydrogen bonding properties in the list.

Table 3.28 Top 15 Descriptors based on RF VIP for $K_{p,brain}$.

Descriptor	VIP	
PSA	0.013	Van der Waals radius surface, summed over all N, O and attached hydrogens, 1-3 overlap correction.
SAS_POL_AREA	0.012	Solvent accessible surface polar area.
MM_VDW_EP_P_MEAN	0.011	Mean of positive electrostatic potentials on Van der Waals surface.
HBDmax	0.011	Highest free energy factor for H-bond donors according to Raevsky (HYBOT).
MM_SAS_EP_P_SUM	0.011	Sum of positive electrostatic potentials on solvent accessible surface.
MM_VDW_EP_P_SUM	0.01	Sum of positive electrostatic potentials on Van der Waals surface.
HBD_Selma	0.01	Number of hydrogen bond donors.
SPEC_SAS_NONPOL_AREA	0.008	SAS_NONPOL_AREA / SAS_TOT_AREA.
PAT	0.008	Number of polar atoms (O, N, S, P).
MM_QO	0.008	Sum of atomic charges on O.
HBsumTotal	0.008	Sum of donor and acceptor free energies according to Raevsky (HYBOT).
MWPat	0.008	MW *Pat/AT_TOT Proportion of MW accounted for by the polar atoms (by number).
SPEC_SAS_POL_AREA	0.007	SAS_POL_AREA / SAS_TOT_AREA.
SPEC_SAS_HB_D_AREA	0.007	SAS_HB_D_AREA / SAS_TOT_AREA.
MM_SAS_EP_N_SUM	0.007	Sum of negative electrostatic potentials on solvent accessible surface.

On the other hand, for the parameter $V_{u,brain}$ is highly dependent on the lipophilicity, ClogP scored the maximum VIP value, thus coming on top of the list (Table 3.29). This table clearly shows many lipophilicity related descriptors.

Table 3.29 Top 15 descriptors based on RF VIP for $V_{u,brain}$.

Descriptor	VIP	
ClogP	0.025	ClogP is a predicted octanol/water partition coefficient from Daylight/Biobyte
ACDlogP	0.02	ACDlogP is calculated as the octanol/water partition coefficient for the neutral species.
Motoc	0.019	Topological distance matrix based index related to ring structures.
GClogP	0.019	Octanol/water partition coefficient based on Ghose/Crippen atom types
VDW_NONPOL_AREA	0.017	Van der Waals non-polar surface area.
NNlogP	0.017	Octanol/water partition coefficient using a neural network approach based on Ghose/Crippen atom types
SAS_NONPOL_AREA	0.016	Solvent accessible surface non-polar area.
HuckelResEnergy	0.015	Huckel molecular orbitals, resonance energy.
MinEV2	0.013	2nd smallest minimum eigenvalue from connectivity matrix, where diagonal has atomic weights.
M1M	0.013	Moment of inertia along the first principal axis of the molecule.
Kappa2	0.012	Topological index.
MWNPat	0.011	MW * NPat/AT_TOT Proportion of MW accounted for by the excess of non-polar atoms (by number)
FractionIonized	0.011	(1 - FractionNeutral)
VDW_AREA	0.011	Van der Waals molecular surface area.
AromCount	0.011	Number of aromatic atoms.

$f_{u,p}$, a parameter that related to plasma protein binding is highly influenced by the charges and lipophilicity of the molecules, which is clearly seen in the results (Table 3.30).

Table 3.30 Top 15 descriptors based on RF VIP for $f_{u,p}$.

Descriptor	VIP	
Base	0.014	Presence of a basic function.
POS_charges	0.0118	Number of basic groups likely to be ionised at pH 7.4.
NNlogP	0.010	Octanol/water partition coefficient using a neural network approach based on Ghose/Crippen atom types
GClogP	0.010	Octanol/water partition coefficient based on Ghose/Crippen atom types
Amine3	0.01	Number of tertiary amines.
ClogP	0.01	ClogP is a predicted octanol/water partition coefficient from Daylight/Biobyte
ACDlogP	0.008	ACDlogP is calculated as the octanol/water partition coefficient for the neutral species.
CHARGES	0.008	POS_charges + NEG_charges.
HOMO	0.008	Huckel molecular orbitals, Highest occupied molecular orbital energy.
ACDlogD74	0.007	ACDlogD74 is calculated as the octanol/water distribution coefficient at pH 7.4.
HuckelPiEnergy	0.007	Huckel molecular orbitals, pi electrons energy.
PIAT	0.007	Number of pi atoms (number of atoms linked to double bonds + number of halogen atoms).
AromCount	0.007	Number of aromatic atoms.
MaxPosChargeGH	0.006	Maximum positive charge using the Gasteiger-Huckel partial charge equilibration.
M1M	0.006	Moment of inertia along the first principal axis of the molecule.

3.10.2 Signature Descriptor

In the signature descriptor analysis of $K_{p,brain}$, it was noticed that as expected the groups that increase lipophilicity like $–CH_3$, -Cl substitutions came up in the list of top 15 signatures positively influencing $K_{p,brain}$(Table 3.31). While groups like amides and ethers were seen in the negative influence (Table 3.32) which is due to their hydrogen bonding properties. However, some signatures representing aromatic aliphatic esters came up as positively influencing $K_{p,brain}$ possibly due to their poor hydrogen bonding capacities.

$V_{u,brain}$ is mainly lipophilicity driven. Tertiary and secondary amines, lipophilic substitutions like sulphur, methyl groups were seen to influence $V_{u,brain}$ positively (Table 3.33).
As expected, ether and amides groups were frequently represented in the negative list (Table 3.34).

Since $f_{u,p}$ depends on the plasma protein binding, signatures representing groups that are vital for such interactions were observed in the analysis like the hydroxyl group, amides, ethers, long alkyl chains etc (Table 3.35 and 3.36).

An effort has been made to interpret the machine learning models for $K_{p,uu,brain}$ by calculating the decision function gradient for the descriptors. Some descriptors having large gradient values have been provided in the tables below. However, there is still a lack of understanding of existence of a clear trend in these descriptors. In the future work, further exploration of the relationship between the descriptors and the $K_{p,uu,brain}$ would be needed in order to improve the model interpretation.

Table 3.31 Top 15 positively influencing signatures for $K_{p,brain}$.

Signature	Number of occurences	Gradient Values	Average Experimental Value	Standard Deviation
[C](p[C](p[N])p[N](p[C]))	14	0.1288	-0.82	0.80
[C](p[C](p[C][C])p[C](p[C][C]))	23	0.1174	-0.15	0.95
[C]([C]([C])[C]([C])[N]([C][C]))	14	0.1099	-0.48	1.08
[O]([C][C])	236	0.0791	-0.85	0.79
[C](p[C](p[C](p[C,0]))p[C](p[C](p[C,0])[Cl]))	156	0.0745	-0.97	0.65
[C]([C]p[C](p[C])p[C](p[C]))	62	0.0722	-1.13	0.72
[C]([C](p[C](p[C])p[C](p[C])))	62	0.0722	-1.13	0.72
[C]([N])	330	0.0648	-0.96	0.83
[N](p[C](p[C])p[C](p[C][O]))	27	0.0639	-0.85	0.55
[C](p[C](p[N](p[C,0]))p[N](p[C]([C]p[C,0])))	14	0.0639	-0.82	0.80
[C]([C]([C])[C]([C][N]))	123	0.0637	-0.98	0.74
[C](p[C](p[C](p[C,0]))p[C]([C]p[C](p[C,0])))	56	0.0627	-1.14	0.74
[C](p[C](p[C]([C]p[C,0]))p[C](p[C](p[C,0][Cl])))	122	0.0624	-0.94	0.66
[O]([C][C](p[C](p[C])p[N](p[C])))	41	0.0605	-0.85	0.56
[O]([C][C](p[C]p[N]))	44	0.0593	-0.84	0.59

Table 3. 32 Top 15 negatively influencing signatures for $K_{p,brain.}$

Signature	Number of occurences	Gradient Values	Average Experimental Value	Standard Deviation
[C]([C]([C][N])[O]([C]))	21	-0.1175	-1.16	0.66
[O]([C]([C]([N,0]))[C]([C]([C][N,0])))	14	-0.1173	-1.12	0.66
[C]([C][S])	35	-0.0777	-1.16	0.59
[N]([C]([C]([C,0]))p[C]([C](p[C]p[C])p[N](p[N,1]))p[C](p[N,1][N]([C][C,0])))	23	-0.0749	-0.97	0.47
[C](p[C](p[C](p[C,0]))p[C](p[N](p[C,0]))[C](p[N](p[N])p[N]([C]p[C])))	33	-0.0744	-1.15	0.54
[C]([N]([C](=[C])[C]([C]=[O])))	31	-0.0742	-1.39	0.40
[N]([C][C](=[C])[C]([C]=[O]))	31	-0.0742	-1.39	0.40
[C](=[C][N])	58	-0.0729	-1.48	0.47
[C]([S]([C]=[O]))	15	-0.0708	-1.61	0.24
[C]([C](p[C](p[C])p[C](p[C]))p[N](p[N](p[C,0]))p[N]([C]p[C,0]([N])))	47	-0.0694	-0.89	0.63
[C](p[C](p[C])p[C](p[C])[C]([N]=[O]))	45	-0.0683	-0.66	0.95
[N]([C]([C]=[O]))	12	-0.0677	-1.13	0.79
[C]([C]p[N]p[N])	303	-0.0672	-1.12	0.63
[C]([C]([C]([C,0]))[C]([N]([C][C,0])))	21	-0.0671	-0.91	0.58
[N](p[C]([C]p[N])p[N](p[C]))	277	-0.0666	-1.14	0.60

Table 3.33 Top 15 positively influencing signatures for $V_{u,brain}$

Signature	Number of occurences	Gradient Values	Average Experimental Value	Standard Deviation
[C](p[C]p[C])	1014	0.1019	1.64	0.66
[C]([C]([C]([C,0]))[N]([C,0][C](p[C]p[C])))	22	0.0905	2.21	0.63
[S](p[C]p[C])	213	0.0815	1.86	0.75
[C](p[C]p[N][O])	43	0.0805	1.24	0.73
[S]([C][C])	38	0.0665	1.80	0.75
[C]([C]([N]([C][C,0]))[N]([C][C]([C,0])))	30	0.0653	1.85	0.66
[C](p[C](p[C])p[C](p[C])[Cl])	112	0.0652	1.79	0.73
[Cl]([C](p[C](p[C])p[C](p[C])))	112	0.0652	1.79	0.73
[N]([C][C])	572	0.0620	1.56	0.72
[C](p[C]p[C][O])	502	0.0608	1.83	0.58
[C]([C](p[C]p[C]))	161	0.0601	1.64	0.69
[C]([C]([C])[C](p[C]p[C]))	48	0.0551	2.08	0.47
[S](p[C](p[C](p[C])p[C](p[C]p[N,0]))p[C]([N]([C])p[N,0]))	118	0.0543	2.22	0.41
[C](p[C]p[C]p[N])	217	0.0528	1.99	0.59
[C](p[C]p[C][Cl])	319	0.0512	1.87	0.66

Table 3.34 Top 15 negatively influencing signatures for $V_{u,brain}$.

Signature	Number of occurences	Gradient Values	Average Experimental Value	Standard Deviation
[C](p[C](p[C])p[C](p[C]))	439	-0.0838	1.43	0.64
[C]([C]p[N]p[N])	125	-0.0786	1.13	0.50
[N]([C]([C]([O,0]))[C]([C]([O,0]))[C](p[C](p[C])p[C](p[C][O])))	38	-0.0774	2.01	0.49
[C](p[C]p[N])	365	-0.0743	1.30	0.63
[C]([C][O])	460	-0.0739	1.59	0.67
[C](p[C](p[C])p[C]([C]p[N]))	56	-0.0652	1.29	0.60
[C]([C][N]=[O])	877	-0.0647	1.59	0.68
[O](=[C]([C][N]))	877	-0.0647	1.59	0.68
[N](p[C](p[C])p[C](p[C][C]))	53	-0.0625	1.24	0.41
[C](p[C](p[C](p[C,0]))p[C](p[C](p[C,0])[C]([N]=[O]))[Cl])	46	-0.0615	1.90	0.51
[C]([C]([C]([C,0]))[N]([C]([C])[C,0]))	28	-0.0591	2.13	0.50
[C]([C]p[C]p[N])	260	-0.0590	1.37	0.60
[C](p[C][C]p[N])	216	-0.0580	1.28	0.63
[C](p[C](p[N](p[C,0]))p[C](p[C]([C]p[C,0])))	59	-0.0580	1.03	0.43

Table 3.35 Top 15 positively influencing signatures for $f_{u,p}$

Signature	Number of occurences	Gradient Values	Average Experimental Value	Standard Deviation
[C](p[C]p[C])	1014	0.1019	1.64	0.66
[C]([C]([C]([C,0]))[N]([C,0][C](p[C]p[C])))	22	0.0905	2.21	0.63
[S](p[C]p[C])	213	0.0815	1.86	0.75
[C](p[C]p[N][O])	43	0.0805	1.24	0.73
[S]([C][C])	38	0.0665	1.80	0.75
[C]([C]([N]([C][C,0]))[N]([C][C]([C,0])))	30	0.0653	1.85	0.66
[C](p[C](p[C])p[C](p[C])[Cl])	112	0.0652	1.79	0.73
[Cl]([C](p[C](p[C])p[C](p[C])))	112	0.0652	1.79	0.73
[N]([C][C])	572	0.0620	1.56	0.72
[C](p[C]p[C][O])	502	0.0608	1.83	0.58
[C]([C](p[C]p[C]))	161	0.0601	1.64	0.69
[C]([C]([C])[C](p[C]p[C]))	48	0.0551	2.08	0.47
[S](p[C](p[C](p[C])p[C](p[C]p[N,0]))p[C]([N]([C])p[N,0]))	118	0.0543	2.22	0.41
[C](p[C]p[C]p[N])	217	0.0528	1.99	0.59
[C](p[C]p[C][Cl])	319	0.0512	1.87	0.66

Table 3.36 Top 15 negatively influencing signatures for $V_{u,brain}$.

Signature	Number of occurences	Gradient Values	Average Experimental Value	Standard Deviation
[C](p[C](p[C])p[C](p[C]))	439	-0.0838	1.43	0.64
[C]([C]p[N]p[N])	125	-0.0786	1.13	0.50
[N]([C]([C]([O,0]))[C]([C]([O,0]))[C](p[C](p[C])p[C](p[C][O])))	38	-0.0774	2.01	0.49
[C](p[C]p[N])	365	-0.0743	1.30	0.63
[C]([C][O])	460	-0.0739	1.59	0.67
[C](p[C](p[C])p[C]([C]p[N]))	56	-0.0652	1.29	0.60
[C]([C][N]=[O])	877	-0.0647	1.59	0.68
[O](=[C]([C][N]))	877	-0.0647	1.59	0.68
[N](p[C](p[C])p[C](p[C][C]))	53	-0.0625	1.24	0.41
[C](p[C](p[C](p[C,0]))p[C](p[C](p[C,0])[C]([N]=[O]))[Cl])	46	-0.0615	1.90	0.51
[C]([C]([C]([C,0]))[N]([C]([C])[C,0]))	28	-0.0591	2.13	0.50
[C]([C]p[C]p[N])	260	-0.0590	1.37	0.60
[C](p[C][C]p[N])	216	-0.0580	1.28	0.63
[C](p[C](p[N](p[C,0]))p[C](p[C]([C]p[C,0])))	59	-0.0580	1.03	0.43

4. CONCLUSION AND FUTURE PERSPECTIVES

In a drug discovery project, it is very critical to determine whether or not a drug molecule will pass through the BBB. Computational prediction of such properties prove to be of great utility in reducing the time and resources spent by aiding in the early elimination of compounds possessing undesirable qualities. The work thus involved building a predictive model that can help in assessing BBB permeability properties of compound.

Revisiting a previous in-house $K_{p,uu,brain}$ model and extending the dataset along with applying newer techniques saw a further improvement in the performance, where the R^2 increased to 0.64 for the best consensus model. This model is composed of 5 different components (using different Machine learning algorithms and descriptors) and has a two-class accuracy of 84% along with the moderate class precision of 48%, in contrast to the 40% seen in the validation results. Here, we see a clear improvement in the overall predictive powers of the model, on the other hand, it can be noted that the model complexity has further increased.

Conformal prediction applied on the SVM model with signature descriptor pointed to the possible noise in the experimental data by giving results that were not clearly interpretable. However, a consistency was always noted between the model performance, in terms of the R^2 and RMSE, and the interval length output by the conformal predictor. The $K_{p,uu,brain}$ indirect model in this case gave a high interval length, also probably due to the formula used for calculating the $K_{p,uu,brain}$ range not being very appropriate. The results from the conformal prediction for the $K_{p,uu,brain}$ indirect can be further studied possibly by defining a more suitable equation for calculating the range. It is also important to note that the results from the conformal prediction might probably point to the unsuitability of its use with the dataset used in the study, as the dataset requires to agree with IID assumptions to be able to successfully apply Conformal prediction.

Model interpretation involving understanding of the factors influencing the $K_{p,uu,brain}$ values based on RF VIP values showed a consistency with previous studies by suggesting many descriptors related to topology and polarity as highly influential. The importance of these descriptors have been described previously. On the other hand, the substructure analysis using the signature gradient showed some unclear trends which still remain to be analyzed further. This reflects the fact that $K_{p,uu,brain}$ is a parameter describing a highly complex process, making it difficult to have a clear understanding of concepts like the important substructures. Work will continue in an attempt to understand these factors.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisors, Dr. Hongming chen and Dr. Ola Engkvist for their invaluable guidance and support throughout the project. I would also want to acknowledge the advice and assistance rendered by Dr. Susanne Winiwarter, Dr. Lars Carlsson, Dr. Jonna Stålring and Dr. Markus Fridén. Special thanks Ajay Anantha and other colleagues in the Computational Chemistry group.

5. REFERENCES

1. Lisa Michielan and Stefano Moro. Pharmaceutical Perspectives of Nonlinear QSAR Strategies. J. Chem. Inf. Model. 2010, 50, 961–978.
2. Peck RW. Driving earlier clinical attrition: if you want to find the needle, burn down the haystack. Considerations for biomarker development. Drug Discovery Today. 2007, 12, 289-294.
3. Michael D. Rawlins. Cutting the cost of drug development? Nature Reviews Drug Discovery. 2004, 3, 360-364.
4. T.J Hou and X.J.Xu. ADME Evaluation in Drug Discover. 3. Modelling Blood-brain barrier partitioning using simple molecular descriptors. j. Chem. Inf. Comput. Sci. 2003, 43, 2137-2152.
5. David E. Clark. In silico prediction of blood-brain barrier permeation.DDT, 2003, VOl. 8, No. 20.
6. Michael Dickson and Jean Paul Gagnon. Key factors in the rising cost of new drug discovery and development. Nature Reviews Drug Discovery, 2004, 3, 417-429.
7. Katya Tsaioun, Michel Bottlaender, Aloise Mabondz. ADDME – Avoiding Drug Development Mistakes Early: central nervous system drug discovery perspective.BMC Neurology, 2009, 9
8. Jürgen Drews. Drug Discovery: A Historical Perspective Science 2000,287, 1960.
9. Vivian I. Teichberg. From the liver to the brain across the blood–brain barrier. Proc Natl Acad Sci U S A. 2007, 104(18): 7315–7316.
10. Roberto Scatena, Giuseppe E Martorana, Patrizia Bottoni, Giorgia Botta, Paola Pastore & Bruno Giardina. An update on pharmacological approaches to neurodegenerative diseases. Informa UK Ltd , 2007, ISSN 1354-3784
11. Li Di, Edward H Kerns & Guy TCarter. Strategies to assess blood–brain barrier penetration. Opin. Drug Discov. 2008.3(6):677-687
12. Mati Karelson, Dimitar Dobchev, Tarmo Tamm, Indrek Tulp, Jaak Jänes, Kaido Tämm, Andre Lomaka, Deniss Savchenko, Gunnar Karelsona. Correlation of blood-brain penetration and human serum albumin binding with theoretical descriptors.ARKIVOC 2008 (xvi) 38-60
13. N. Joan Abbott, Lars Rönnbäck & Elisabeth Hansson. Astrocyte–endothelial interactions at the blood–brain barrier.Nature Reviews Neuroscience 7, 41-53 (January 2006)
14. Sandipan Roy. Strategic Drug Delivery Targeted to the Brain: A Review. Der Pharmacia Sinica, 2012, 3(1):76-92
15. Andreas Reichel. The role of Blood-Brain barrier studies in the pharmaceutical industry. Current Drug Metabolism, 2006,7,183-203
16. Domenico Ribatti, Beatrice Nico , Enrico Crivellato , and Marco Artico. Development of the Blood-Brain Barrier: A Historical Point of View. Anat Rec B New Anat. 2006 Jan; 289(1):3-8.
17. Jeffrey P, Summerfield SG. Challenges for blood-brain barrier (BBB) screening. Xenobiotica 2007; 37:1135–1151.
18. Brian T. Hawkins and Thomas P. Davis. The Blood-Brain Barrier/Neurovascular Unit in Health and Disease. Pharmacological Reviews. 2005 vol. 57 no. 2173-185.
19. Deeken JF, Löscher W.The Blood-Brain Barrier and Cancer: Transporters, Treatment, and Trojan Horses. Clin. Cancer Res. 2007 13; 1663.
20. Elizabeth CM de Lange. The mastermind approach to CNS drug therapy: translational prediction of human brain distribution, target site kinetics, and therapeutic effects. Fluids Barriers CNS. 2013;10(1):12.
21. Ohtsuki S, Terasaki T. Contribution of carrier-mediated transport systems to the blood-brain barrier as a supporting and protecting interface for the brain; importance for CNS drug discovery and development. Pharm Res. 2007 Sep; 24(9):1745-58.
22. David J. Begley. Delivery of therapeutic agents to the central nervous system: the problems and the possibilities. Pharmacol Ther. 2004; 104(1):29-45.
23. Wolfgang Löscher and Heidrun Potschka. Blood-Brain Barrier Active Efflux Transporters: ATP-Binding Cassette Gene Family. NeuroRx. 2005 January; 2(1): 86–98.
24. Hassan Pajouhesh and George R. Lenz. Medicinal chemical properties of successful Central Nervous System drugs. NeuroRx. 2005 October; 2(4): 541–553.

25. Roland Nau,Fritz Sörgel,Helmut Eiffert. Penetration of Drugs through the Blood-Cerebrospinal Fluid/Blood-Brain Barrier for Treatment of Central Nervous System Infections. Clin. Microbiol. Rev. 2010 vol. 23 no. 4, 858-883.
26. William M. Pardridge. The Blood-Brain Barrier: Bottleneck in Brain. Drug Development. NeuroRx. Jan 2005; 2(1): 3–14.
27. Fischer H, Gottschlich R, Seelig A. Blood-brain barrier permeation: molecular parameters governing passive diffusion. J Membr Biol. 1998 Oct 1;165(3):201-11.
28. Fridén M, Winiwarter S, Jerndal G, Bengtsson O, Wan H, Bredberg U, Hammarlund-Udenaes M, Antonsson M. Structure-brain exposure relationships in rat and human using a novel data set of unbound drug concentrations in brain interstitial and cerebrospinal fluids. J Med Chem. 2009 Oct 22; 52(20):6233-43
29. Di L, Rong H, Feng B. Demystifying brain penetration in central nervous system drug discovery. Miniperspective. J Med Chem. 2013 Jan 10; 56(1):2-12.
30. Dennis A. Smith, Li Di & Edward H. Kerns. The effect of plasma protein binding on in vivoefficacy: misconceptions in drug discovery. Nature Reviews Drug Discovery.2010. 9, 929-939.
31. Martin, I. Prediction of blood-brain barrier penetration: are we missing the point? Drug Discov. Today 2004, 9, 161–2.
32. Hammarlund-Udenaes, M. Active-site concentrations of chemicals - are they a better predictor of effect than plasma/organ/tissue concentrations? Basic Clin. Pharmacol. Toxicol. 2010, 106, 215–20.
33. Oldendorf WH. Measurement of brain uptake of radiolabeled substances using a tritiated water internal standard. Brain Res. 1970 Dec 1;24(2):372-6.
34. Xingrong Liu, Kristine Van Natta, Helen Yeo, Olga Vilenski, Paul E. Weller, Philip D. Worboys and Mario Monshouwer. Unbound Drug Concentration in Brain Homogenate and Cerebral Spinal Fluid at Steady State as a Surrogate for Unbound Concentration in Brain Interstitial Fluid.DMD, 2009 vol. 37 no. 4 787-793.
35. Mehdipour AR, Hamidi M. Brain drug targeting: a computational approach for overcoming blood-brain barrier. Drug Discov Today. 2009 Nov;14(21-22):1030-6.
36. T.Wayne Schultz,Mark T.D. Cronin, John D. Walker, Aynur O. Aptula.. Quantitative structure–activity relationships (QSARs) in toxicology: a historical perspective. Journal of Molecular Structure: THEOCHEM 2003; Volume 622, Issues 1–2, Pages 1–2
37. James D. McKinney, Ann Richard, Chris Waller, Michael C. Newman, Frank Gerberick.The Practice of Structure Activity Relationships (SAR) in Toxicology. Toxicol. Sci. (2000) 56 (1): 8-17.
38. Andrew R. Leach (2007). An Introduction to Chemoinformatics. The Netherlands: Springer.
39. R. Burbidge, M.Trotter, B. Buxton, S.Holden. Drug design by machine learning: support vector machines for pharmaceutical data analysis. Computers and Chemistry 26(2001) 5-14.
40. Hongming Chen, Susanne Winiwarter, Markus Fridén, Madeleine Antonsson, Ola Engkvist. In silico prediction of unbound brain-to-plasma concentration ratio using machine learning algorithms. J Mol Graph Model. 2011 Aug;29(8):985-95.
41. Ajay, Guy W. Bemis, and Mark A. Murcko. Designing Libraries with CNS Activity. J. Med. Chem. 1999, 42, 4942-4951.
42. Liying Zhang, Hao Zhu, Tudor I. Oprea, Alexander Golbraikh, and Alexander Tropsha. QSAR Modeling of the Blood–Brain Barrier Permeability for Diverse Organic Compounds. Pharm Res. 2008 Aug;25(8):1902-14
43. Claudia Andres and Michael C. Hutter. CNS Permeability of Drugs Predicted by a Decision Tree. QSAR & Combinatorial Science .2006; 25(4):305 - 309.
44. H. van de Waterbeemd, M. Kansy. Hydrogen-bonding capacity and brain penetration. Chimia, 46 (1992), pp. 299–303
45. M.H. Abraham et al.Hydrogen bonding. 33. Factors that influence the distribution of solutes between blood and brain. J. Pharm. Sci., 83 (1994), pp. 1257–1268.
46. De Lange, E. C.; Danhof, M.; De Boer, A. G.; Breimer, D. D. Critical factors of intracerebral microdialysis as a technique to determine the pharmacokinetics of drugs in rat brain. Brain Res. 1994, 666, 1–8.

47. Lindén, K.; Ståhle, L.; Ljungdahl-Ståhle, E.; Borg, N. Effect of probenecid and quinidine on the transport of alovudine (3'-fluorothymidine) to the rat brain studied by microdialysis. Pharmacol. Toxicol. 2003, 93, 226–32.
48. Lanevskij, K.; Dapkunas, J.; Juska, L.; Japertas, P.; Didziapetris, R. QSAR analysis of blood-brain distribution: the influence of plasma and brain tissue binding. J. Pharm. Sci. 2011, 100, 2147–60.
49. Kaliszan, R. Brain/blood distribution described by a combination of partition coefficient and molecular mass. Int. J. Pharm. 1996, 145, 9–16.
50. Salminen, T.; Pulli, A.; Taskinen, J. Relationship between immobilised artificial membrane chromatographic retention and the brain penetration of structurally diverse drugs. J. Pharm. Biomed. Anal. 1997, 15, 469–77.
51. Clark, D. E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration. J. Pharm. Sci. 1999, 88, 815–21.
52. Kelder, J.; Grootenhuis, P. D.; Bayada, D. M.; Delbressine, L. P.; Ploemen, J. P. Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. Pharm. Res. 1999, 16, 1514–9.
53. Feher, M.; Sourial, E.; Schmidt, J. M. A simple model for the prediction of blood-brain partitioning. Int. J. Pharm. 2000, 201, 239–47.
54. Osterberg, T.; Norinder, U. Prediction of polar surface area and drug transport processes using simple parameters and PLS statistics. J. Chem. Inf. Comput. Sci. 40, 1408–11.
55. Keseru, G. M.; Molnar, L. High-Throughput Prediction of Blood-Brain Partitioning: A Thermodynamic Approach. J. Chem. Inf. Model. 2001, 41, 120–128.
56. Lombardo, F.; Blake, J. F.; Curatolo, W. J. Computation of brain-blood partitioning of organic solutes via free energy calculations. J. Med. Chem. 1996, 39, 4750–5.
57. Narayanan, R.; Gunturi, S. B. In silico ADME modelling: prediction models for blood-brain barrier permeation using a systematic variable selection method. Bioorg. Med. Chem. 2005, 13, 3017–28.
58. Muehlbacher, M.; Spitzer, G. M.; Liedl, K. R.; Kornhuber, J. Qualitative prediction of blood-brain barrier permeability on a large and refined dataset. J. Comput.-Aided Mol. Des. 2011, 25, 1095–106.
59. Iyer, M.; Mishru, R.; Han, Y.; Hopfinger, A. J. Predicting blood-brain barrier partitioning of organic molecules using membrane-interaction QSAR analysis. Pharm. Res. 2002, 19, 1611–21.
60. Norinder, U.; Haeberlein, M. Computational approaches to the prediction of the blood-brain distribution. Adv. Drug Deliver. Rev. 2002, 54, 291–313.
61. Gratton, J. A.; Abraham, M. H.; Bradbury, M. W.; Chadha, H. S. Molecular factors influencing drug transfer across the blood-brain barrier. J. Pharm. Pharmacol. 1997, 49, 1211–6.
62. Fridén, M.; Gupta, A.; Antonsson, M.; Bredberg, U.; Hammarlund-Udenaes, M. In vitro methods for estimating unbound drug concentrations in the brain interstitial and intracellular fluids. Drug Metab. Dispos. 2007, 35, 1711–9.
63. Jean-Loup Faulon; Donal P.ViscoThe, Jr. and Ramdas S. Pophale. Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies.J. Chem. Inf. Comput. Sci., 2003, 43 (3), pp 707–720.
64. Glenn Shafer; Vladimir Vovk. A Tutorial on Conformal Prediction. Journal of Machine Learning Research,2008, 9, 371-421.Osten D.W. Selection of optimal regression models via cross-validation. J Chemometr 1988,2,39–48
65. Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Artificial Intelligence (IJCAI), 1995.
66. Corwin Hansch; Peyton P. Maloney; Toshio Fujita; Robert M. Muir. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. Nature, 1962, 194, 178 – 180.
67. Mati Karelson; Victor S. Lobanov. Quantum-Chemical Descriptors in QSAR/QSPR Studies. Chem. Rev. 1996, 96, 1027−1043.
68. R. Todeschini, V Consonni (2009). Molecular descriptiors for chemoinformatics. WILEY-VCH

69. Malcolm J. McGregor and Steven M. Muskal. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. J. Chem. Inf. Comput. Sci. 1999, 39, 569-574.
70. Jean-Loup Faulon. The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies J. Chem. Inf. Comput. Sci. 2003, 43, 707-720
71. Jean-Loup Faulon, Michael J. Collins, and Robert D. Carr. The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequences.J. Chem. Inf. Comput. Sci. 2004, 44, 427-436
72. Carla J. Churchwell, Mark D. Rintoul , Shawn Martin, Donald P. Visco, Jr., Archana Kotu, Richard S. Larson, Laurel O. Sillerud, David C. Brownc, Jean-Loup Faulon. The signature molecular descriptor 3. Inverse-quantitative structure–activity relationship of ICAM-1 inhibitory peptides. Journal of Molecular Graphics and Modelling 22 (2004) 263–273
73. Corinna Cortes, Vladimir Vapnik. Support-Vector Networks. Machine Learning. 1995, 20, 273-297
74. Vladimir Svetnik, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J. Chem. Inf. Comput. Sci. 2003, 43, 1947-1958.
75. Leo Breiman. Random Forests. Machine Learning. 2001, 45, 5–32.
76. Anne-Laure Boulesteix, Silke Janitza, Jochen Kruppa and Inke R. König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics.Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.2012,2, 6, 493–507.
77. Alexander Gammerman, Vladimir Vovk. Hedging Predictions in Machine Learning. The Computer Journal. 2007, 50: 2, 151–177.
78. Glenn Shafer, Vladimir Vovk. A Tutorial on Conformal Prediction. Journal of Machine Learning Research.2008, 9, 371-421.
79. Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011, 2:27:1--27:27, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
80. Demšar, J., Curk, T., & Erjavec, A. Orange: Data Mining Toolbox in Python; Journal of Machine Learning Research 14(Aug):2349−2353, 2013.
81. Jonna C Stålring, Lars A Carlsson,Pedro Almeida and Scott Boyer.AZOrange - High performance open source machine learning for QSAR modeling in a graphical programming environment.Journal of Cheminformatics 2011,3:28.
82. R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
83. Tibco Spotfire,version 3.1.1, Tibco Software inc, Palo Alto, California.
84. Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, Paul J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings .Advanced Drug Delivery Reviews.Volume 23, Issues 1–3, 15 January 1997, Pages 3–25.
85. Wan, H.; Rehngren,M. High-throughput screening of protein binding by equilibriumdialysis combined with liquid chromatography and mass spectrometry. J. Chromatogr., A 2006, 1102, 125–134.