Development and evaluation of ADME models using proprietary and opensource data

By

Maria-Anna Trapotsi

July 2017

A thesis submitted to the University of Hertfordshire in partial fulfilment of the requirements for the degree of Master of Science by Research

Abstract

Absorption, Distribution, Metabolism and Elimination (ADME) properties are important factors in the drug discovery pipeline. Literature ADME data are often collected in large chemical databases like ChEMBL, which might be an asset to improve the prediction of ADME properties. Pharmaceutical companies build ADME Quantitative Structure Property Relationships (QSPR) models using proprietary data and thus the inclusion of literature data might be a valuable source for the development of predictive models. The aim of this study was to investigate whether merging literature and proprietary data could improve the predictive activity of proprietary models and enlarge their applicability domain (AD).

ADME predictive models for Caco-2 (A to B) permeability and $LogD_{7.4}$ were built with data extracted from Evotec and ChEMBL database. Predictive models were developed for each property and three different training sets were used based on: proprietary compounds (Evotec models), literature compounds (ChEMBL models) and a merged set of proprietary and literature compounds (Evotec+ChEMBL models). The Random Forest (RF), Partial Least Squares (PLS) and Support Vector Regression (SVR) were used to develop the models. The performance of the models was evaluated by using two types of test sets: a diverse test set (20 % compounds of available data randomly selected) and a temporal test set (data published after the models were built). The descriptors that used were the physiochemical descriptors, the structural Molecular Access System (MACCS) descriptors and the Partial equalisation of orbital electronegativity – van der Walls surface areas (Peoe-VSA) descriptors. The AD of the models was evaluated with four distance to model metrics, which were the: kNN with Euclidean distance, kNN with Manhattan distance, Leverage and Mahalanobis distance.

The ability of an existing Evotec Caco-2 permeability model to assess literature compounds (extracted from ChEMBL) was evaluated. The literature test set was predicted with a higher RMSE compared to the RMSE in prediction for internal compounds. Additionally, a number of literature compounds was found to be outside the AD of the Evotec model, thus highlighting an area of improvement for proprietary Evotec models. Furthermore, the effect of the inclusion of literature data in the existing Caco-2 permeability and $LogD_{7.4}$ Evotec proprietary models was evaluated. The RF algorithm was the highest performing method for the development of Caco-2 permeability models and the SVR for the $LogD_{7.4}$ models. In addition, the leverage method proved to be the most appropriate for the evaluation of the models' AD. The permeability model built merging literature and proprietary data (Evotec+ChEMBL model) predicted a literature temporal test set with an RMSE of 0.68 while the Evotec model showed an RMSE of 0.74. Even in the case of the Evotec temporal test set, the two models performed similarly and the AD of the mixed models (incorporating both literature and proprietary data) was enlarged. The 86.15% of the compounds in the proprietary temporal test set were within the AD of the Evotec+ChEMBL model, while 76.50% of the compounds of the same test set appeared to be within the AD of the Evotec model. Similarly, the $LogD_{7.4}$ Evotec+ChEMBL model predicted a literature temporal test set with an RMSE of 0.77 while the Evotec model showed an RMSE of 0.83. Even in the case of the Evotec temporal test set, the two models performed similarly but the AD of the mixed models (incorporating both literature and

proprietary data) was enlarged. The 94.86% of the compounds in the proprietary temporal test set were within the AD of the Evotec+ChEMBL model, while 88.49% of the compounds of the same test set appeared to be within the AD of the Evotec model.

This study demonstrated that the inclusion of public ADME data into proprietary models improved the performance of proprietary models and enlarged at the same time their AD. The methodology presented herein will be applied by Evotec computational scientists to re-build the Caco-2 and $LogD_{7.4}$ Evotec proprietary models considering literature data as discussed in this thesis.

Acknowledgments

I would like to thank my academic supervisor Professor Mire Zloh from the University of Hertfordshire and my industrial placement supervisors, Dr Mirco Meniconi and Dr Mike Bodkin from Evotec for their help, advice and support throughout this project. I would also like to thank Dr Patrick Barton form the Evotec DMPK department for the help and support throughout this project.

I would also like to thank the Research Informatics group at Evotec that made me feel welcome and part of the group.

Finally, I would like to thank my family for the encouragement and support throughout my studies.

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION

1.1 ADME properties in drug development process

The pharmaceutical drug design and development process is time consuming, complex and characterised by high risk and cost (Wang & Urban, 2004). It has been estimated that the probability of success in Phase II clinical trials is only 34 % (Cumming, Davis, Muresan, Haeberlein, & Chen, 2013). The efficacy and ADME (Absorption, Distribution, Metabolism, Elimination) properties play a significant role in the drug mechanism (Thompson, 2000) and are considered as an integral part of the drug design process (Di & Kerns, 2016).

A molecule should be able to exhibit both a pharmacological effect and also to have the appropriate ADME properties to reach the market as a drug. Or in other words, a drug should not only be efficacious for the target disease but also with an acceptable pharmacokinetic and safety profile (Davies et al., 2015). Some of these parameters include the lipophilicity, ionisation, solubility and molecular mass (Livingstone & Davis, 2012). For example, a highly lipophilic drug can be more permeable (i.e. greater absorption) (Riley, Parker, Trigg, & Manners, 2001), can undergo greater metabolic clearance (Patrick, 2013) and it can be better absorbed in the GI tract (Avdeef & Tam, 2010). In addition, lipophilicity can affect the ability of a drug to cross the Blood Brain Barrier (BBB) and the volume of distribution (Poulin & Theil, 2002) because of the drug ability to bind to serum albumin (Patrick, 2013). As a result, parameters such as lipophilicity should be taken into account from the early stages of drug design in order to exclude compounds with unwanted properties.

The total loss rate due to poor ADME properties was near 50% in 2004 (Khanna, 2012). Although the failure rate was reduced to 14% (Tsaioun, 2007) due to the preclinical testing, there is a potential to improve cost-effectiveness of the drug discovery and development by using predictive ADME predictive models. Therefore, it is of major significance for pharmaceutical industries to improve the productivity of the drug design process (Paul et al., 2010) and reduce failure due to poor ADME properties. Computational chemistry can be a great asset in drug discovery process (Liao, Sitzmann, Pugliese, & Nicklaus, 2011), as its application can reduce the risk and cost of the drug design process (Tan et al., 2010). A useful tool of the computational medicinal chemistry is the *in-silico* predictive ADME models. The great advantage of these models is the prediction of a molecule's ADME properties (Zhang, Luo, Ding, & Lu, 2012) prior to chemical synthesis and *in-vitro* or *in-vivo* testing, which will save time and money (Zhang & Surapaneni, 2012) in preclinical testing (figure 1). Therefore, the number of compounds that have to be synthesised to obtain the required biochemical and physicochemical profile is reduced (Moroy, Martiny, Vayer, Villoutreix, & Miteva, 2012).

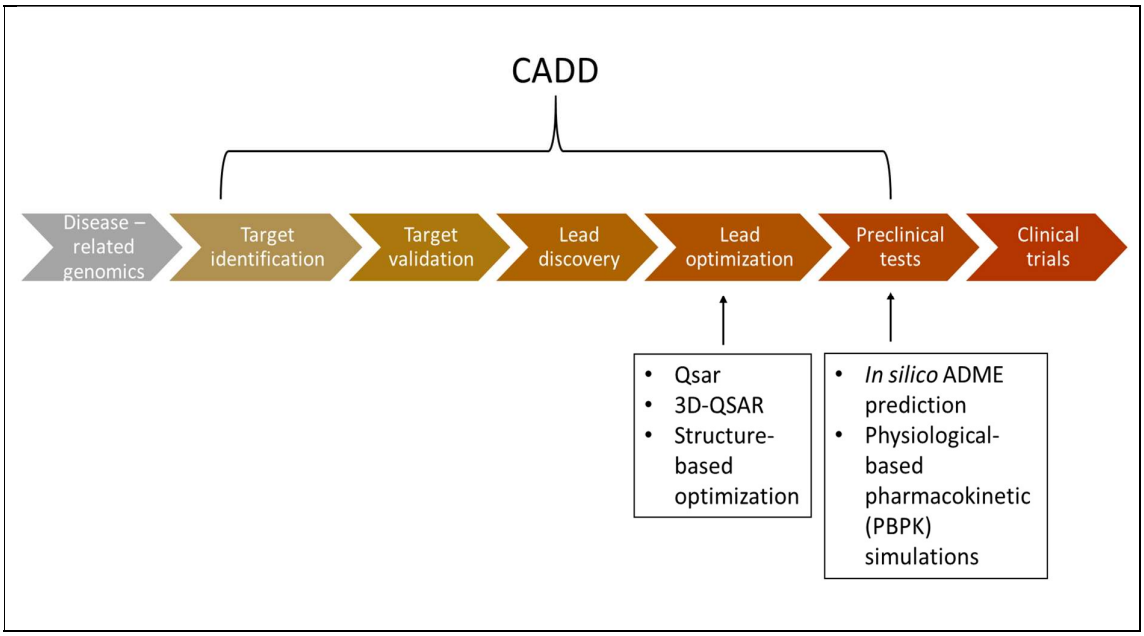

Figure 1: Computer Aided Drug Design (CADD) in drug design and development process (adapted from Kore, Mutha, Antre, Oswal, & Kshirsagar, 2012).

1.2 QSAR and QSPR modelling

Quantitative Structure Activity Relationship (QSAR) and Quantitative Structure Property Relationship (QSPR) modelling are major and commonly employed computational tools in medicinal chemistry to help the lead optimization process in drug discovery (Cramer, 2012; Kore et al., 2012). QSAR is widely used to provide optimisation of the pharmacological activity, and QSPR can provide information about pharmacokinetic or ADME properties (Puzyn, Leszczynski, & Cronin, 2010). QSPR models are mathematical models, which relate the chemical structure of the compound to a physiochemical property and this relation can be used to predict ADME properties (Yee & Wei, 2012). QSPR modelling can provide exploration and exploitation of the relationship between the chemical structure of the compounds and their ADME properties (Tropsha, 2010) prior to the synthesis of a compound (Park et al., 2014). The introduction of QSAR/QSPR models, has raised concerns for the predictability and applicability of these models (Jaworska, Nikolova-Jeliazkova, & Aldenberg, 2005). Therefore, five principles have been established for QSAR/QSPR model validation*:* 1. a defined endpoint, 2. an unambiguous algorithm, 3. a defined domain of applicability, 4. appropriate measures of goodness-of-fit, robustness and predictivity and 5. a mechanistic interpretation, if possible (Sahigara *et al.*, 2012). One of the most important principles is the applicability domain, which will be further discussed.

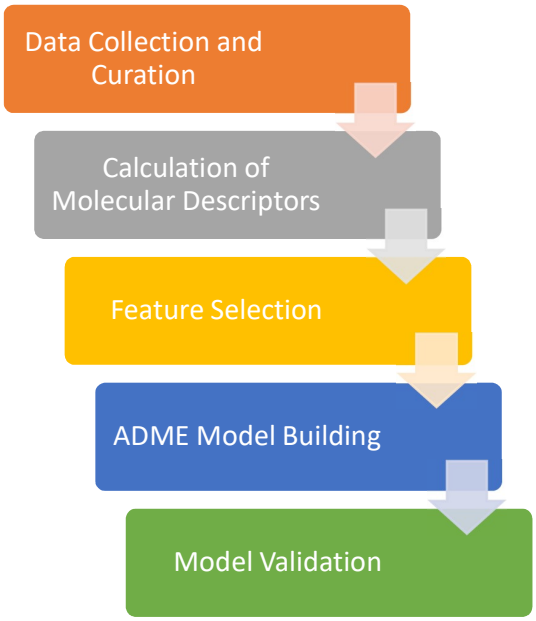

Figure 2: The steps of the QSPR development process (adapted from Cherkasov et al., 2014).

1.3 Data collection and curation

Figure 2 is schematically depicting the process of building a QSPR. The first step of that process involves the data collection and curation. This is a significant part of the QSPR development because the performance of the model depends on the quality of the training set (Yee & Wei, 2012). Literature data and databases can be considered as an increasingly important source for collection of compounds and these data have been used for the development of QSPR models (Wang, Cao, Zhu, & Yun, 2015; Wang et al., 2016).

1.3.1 Literature data and databases for ADME data collection for QSPR modelling

Literature data are published in journal articles (peer-reviewed or scientific) and thus it is usually difficult to manually search and extract information. For example, literature chemical structures are usually depicted as images and that is making the extraction and use of literature data in QSPR development difficult (Gaulton et al., 2012). Therefore, in the recent years a variety of publicly available databases have been developed due to the high demand for easy, free and open access to the literature information. As a result, the construction of QSPR models is greatly assisted by the development of large publicly available compound databases like PubChem BioAssay (Li, Cheng, Wang, & Bryant, 2010; Y. Wang et al., 2010), ChemBank (Seiler et al., 2008), ZINC (Irwin, Sterling, Mysinger, Bolstad, & Coleman, 2012), ochem.eu (online chemical database with modelling environment) (Wang *et al.*, 2016) and ChEMBL (Bento et al., 2014; Gaulton et al., 2012).

The three databases that store information for ADME assays are the PubChem BioAssay, ochem.eu and ChEMBL. The other databases like ZINC is used mainly for ligand discovery (Irwin et al., 2012) and ChemBank has been developed to guide chemists in the synthesis of

novel compounds and biologists to search for small molecules that catalyse a specific process (Seiler et al., 2008). ChEMBL is the database, which is considered as a key representative of the current plethora of publicly available data (which also include the majority of the information available in PubChem BioAssay and ochem.eu) (Papadatos & Overington, 2014; Wang et al., 2009) and has dramatically changed the way that the drug discovery community shares and deposits experimental data. Moreover, CHEMBL extracts the information from the medicinal chemistry literature (Papadatos, Gaulton, Hersey, & Overington, 2015), mainly from 12 prominent chemistry journals (Bender, 2010). Moreover, companies like AstraZeneca deposited compounds into ChEMBL (Clark et al., 2015).

ChEMBL contains information obtained by various assays, which are divided into four categories: 1. Binding (B), Functional (F), Toxicity (T) and ADME (A) and additionally include annotations related to the relevant assays. This is a great advantage of ChEMBL, which other databases lack. These supplementary annotations are useful and help the data curation process but the level of detail of annotations is not always sufficient to truly identify the protocols of the ADME assays (Papadatos et al., 2015). Even when the assay conditions seem to be the same, a significant variability is observed between measurements by different laboratories (Kalliokoski et al., 2013). Therefore, ChEMBL team has set future plans to improve the quality and consistency of the data by including more detailed description of the assays' parameters (Papadatos et al., 2015). One of the main disadvantages in ChEMBL is the quality and reliability of the literature sources. For example, an error in chemical structure might result into an erroneous descriptor calculation (Tropsha, 2010), which will ultimately affect the predictability of the model. Manual curation of the data downloaded from public databases can substantially improve the accuracy of prediction (Young, Martin, Venkatapathy, & Harten, 2008). The error in commercial or public databases ranges from 0.1% - 3.4% (Fourches, Muratov, & Tropsha, 2010) and another example is that of WOMBAT (world of molecular bioactivity) database with an overall error rate of 8% (Young et al., 2008). Therefore, it is important to curate the data extracted from large chemical databases before the development of the models.

1.4 Calculation of molecular descriptors

After the first step in QSPR process, which is the data collection and curation (figure 2), the next step is the calculation of molecular descriptors. Molecular descriptors are a basic tool for cheminformatics, which is used to transform chemical information (like physiochemical properties) into a numerical data and they can be theoretically (derived from symbolic molecule representation) or experimentally derived (Puzyn, Leszczynski and Cronin, 2010).

Topological descriptors are widely used for QSPR modelling and they refer to 2D molecular descriptors (Rajkhowa & Deka, 2014), which are based on the distances between atoms calculated by the number of intervening bonds (Puzyn et al., 2010) and thus considering the internal arrangement of compound's atoms (Pillai, 2015). Therefore, topological descriptors can give numerical information about molecular size, presence of heteroatoms, multiple bonds (Gozalbes & Doucet, 2002) and enable for the identification of the individual atoms and the

bonded connections between them (Roy, Kar, & Das, 2015). The Molecular Access System or "MACCS keys" is considered as the best known and the prototype of key-based fingerprints (Chackalamannil, Rotella, & Ward, 2017). MACCS are structural descriptors and are based on pattern matching of the chemical structure of a compound to a pre-defined set of structural fragments, (166 MACCS keys) (Wale, Watson, & Karypis, 2008). Another set of descriptors that can be used are the partial equalization of orbital electronegativity - van der Walls surface areas (Peoe-VSA) descriptors, which capture the direct electrostatic interactions (Bajorath, 2004). For example, electrostatic interactions play a role in the metabolism and protein binding, because these interactions can affect the binding of a compound to the active site of the metabolic enzyme and the plasma proteins (Cyprotex, 2015).

Other important 2D descriptors that can be used for ADME models are Polar Surface Area (PSA), number of hydrogen bond acceptors/donors, LogP, LogD at various pH (which can be either experimentally or theoretically calculated) and pKa (Hou, Li, Zhang, & Wang, 2009). For example, the H bonding behaviour of a compound can be useful for the description of drug permeability because as the number of hydrogen bonds increases, the polarity of the compound increases too and the lipophilicity becomes weaker. As a result the compound is less able to cross the cell membrane by passive diffusion (Wang *et al.*, 2016) because hydrogen bonds are formed with the outer phase of the membrane. In addition, PSA is one of the most significant molecular descriptors in QSPR studies and is a measure of polarity of the compound, which indicates the presence of a dipole moment (Caron & Ermondi, 2016). PSA is an area of Van der Waals surface, which results from oxygen, nitrogen or hydrogen atoms bound to polar areas (Danielle, 2014). As a result, PSA is related to the hydrophobicity and polarity of a molecule and is useful in estimating the compound's absorption, BBB permeability and other ADME characteristics (Kubinyi, Folkers, & Mannhold, 2008). For example, PSA should be low (60-70$Å^2$) for BBB penetration and no more than 140 $Å^2$ for cell membrane permeation (Pajouhesh & Lenz, 2005) and generally PSA gives excellent correlation with drug permeability in Caco-2 monolayers (Artursson, Palm, & Luthman, 2012).

In addition to the 2D QSAR descriptors, there also the 3D descriptors for QSAR modelling (3D QSAR) like randic molecular profiles, geometrical descriptors etc. One of the most widely used 3D QSAR method is the Comparative Molecular Field Analysis (CoMFA), which concerns mainly the electrostatic field and steric relationships between the ligand and biological target (Cherkasov et al., 2014). However, it is considered as a computationally intense process (Goodarzi & Dejaegher, 2012) and one example might be the conformational analysis to find the best conformer.

1.5 Feature Selection

A feature selection or variable selection is usually performed to choose the descriptors with goal to reduce the dimensionality and the redundancy of the descriptors that are chosen. Feature selection usually depends on two parameters. The first is the correlation and variance of descriptors and the second is the algorithm that is applied to the training set. Correlated descriptors are those which are different views of the same molecular aspect (Puzyn et al.,

2010). Therefore, some algorithms like MLR cannot produce meaningful results with correlated descriptors, whereas other methods like PLS and SVR can handle sets that contain correlated descriptors. In addition, zero or very low variance descriptors can be removed. They do not carry any information because they are constant for all the chemical compounds. There are various methods to perform feature/variable selection and they are categorised into three groups: filters, wrappers and embedded methods. The filter methods use a metric or score for each feature, based on a statistical measure and based on their score are excluded or included (Brownlee, 2016). An example of a filter method is the ReliefF, which randomly picks dataset points and finds their nearest neighbours. Then it assigns weight to the features/descriptors based on how good they can discriminate the observations from their neighbours (Eklund, Norinder, Boyer, & Carlsson, 2014). The wrapper methods use a learning algorithm and identify descriptors subsets. Models are developed and assess which descriptor combinations can result in a good model accuracy (Brownlee, 2016). The embedded methods incorporate the feature selection during the application of learning algorithm (Eklund et al., 2014). However, it was shown that the use of different feature selection methods did not improve the prediction accuracy of models developed with "state-of-the-art" algorithms (RF, ANN, SVM) (Eklund et al., 2014). The reason is that these algorithms can handle correlated descriptors.

1.6 Model Building and Machine learning in QSPR model development

One of the most significant factor in QSPR building process is the selection of an appropriate method. QSPR models have evolved significantly since scientists decided to utilise approaches from recent developments in other fields like data mining, pattern recognition, machine learning and artificial intelligence (Dudek, Arodz, & Gálvez, 2006; Geppert, Vogt, & Bajorath, 2010). Various algorithms are used to identify patterns and correlations within a dataset/training set, and through data mining process, a model is derived (Lavecchia, 2015). Each compound is considered as a vector and each molecular descriptor corresponds to 1 dimension/variable. The resulting model relates a set of descriptors with biologically relevant properties like lipophilicity and other parameters, which can affect ADME properties. There are various types of models and they are usually divided into two broad categories: continuous (regression) and classification (categorical) (Dudek, Arodz, & Gálvez, 2006) (figure 3).

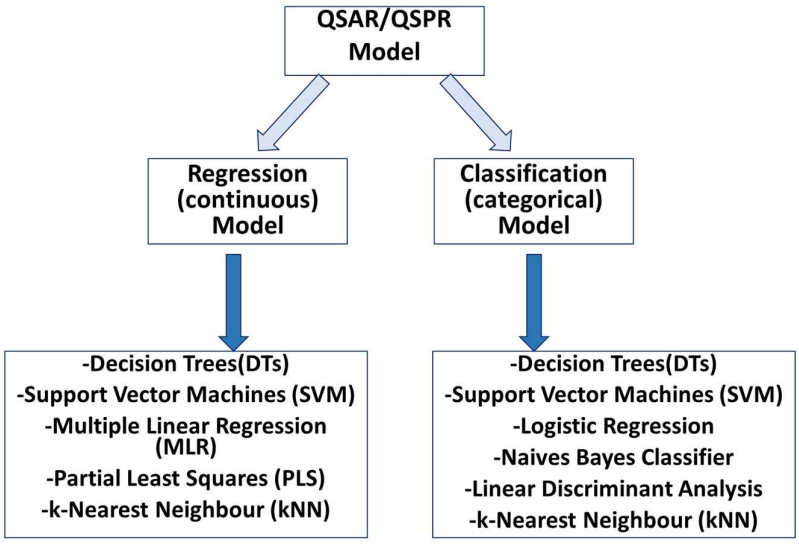

Figure 3: Summary of the QSAR or QSPR building methods (adapted from Dudek, Arodz and Gálvez, 2006; Danielle, 2014).

1.6.1 Multiple Linear Regression (MLR)

MLR is a supervised machine learning method that is able to establish a linear mathematical relationship between a property of the training compounds and a set of descriptors that encode the chemical information (Ventura, Latino, & Martins, 2013). MLR is a commonly used method for constructing QSPR models (Liu & Long, 2009) and the prediction is derived as a linear function of all descriptors (Sethi, 2012). The following equation gives the linear relationship between the target value and the compounds' features/descriptors:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad \text{(Equation 1)},$$

where n is the number of descriptors, x_1, x_2, …, x_n are the molecular descriptors, β_1, β_2, ..., β_n are descriptors' coefficients and β_0 is the model constant

Equation 1 represents a hyperplane in a space of n-dimensions. In addition, the coefficients of that equation are calculated with methods like the least-squares method, which minimizes the sum of squared residuals (Dehmer, Varmuza, & Bonchev, 2012). However, there are disadvantages related to MLR. For example correlated descriptors and a large descriptors to compounds number ratio are two factors that MLR cannot handle and result in unstable predictions (Dudek, Arodz & Gálvez, 2006). The underlying reason is that descriptors influence the calculation of the coefficients and therefore correlated descriptors could result in erroneous estimation. Additionally, the number of compounds should be at least five times the number of descriptors to reduce the possibility of erroneous coefficient calculation.

1.6.2 Partial Least Squares (PLS)

PLS is a more popular method compared to MLR because it overcomes the disadvantages of MLR mentioned above. PLS uses similar principles with Principal Component Analysis (PCA) and it is suitable to overcome the issues related to the multicollinearity and the high ratio of the number of descriptors over the number of compounds (Dudek, Arodz & Gálvez, 2006). PLS is able to project the original variables (i.e. descriptors) into latent variables (LVs) and thus reducing the dimensionality (Xing et al., 2014). LVs do not only explain the variation in the x variables (descriptors) as the PCA does. They also take into account how the variation in the x variables corresponds to the variation of the dependent variable y (target value) (Brown, 2015). The following equations correspond to the latent variables (LV_i), which are linear combinations of the variables/ descriptors (x_i).

$$y = a_1 LV_1 + a_2 LV_2 + \cdots + a_n LV_n \quad \text{(Equation 2),}$$

where y is the target value, α is the regression coefficient and LV are the latent variables in a chemical space with n descriptors/dimensions

$$LV_1 = b_{1.1}x_1 + b_{1.2}x_2 + \cdots + b_{1.n}x_n$$
$$LV_2 = b_{2.1}x_1 + b_{2.2}x_2 + \cdots + b_{2.n}x_n$$
$$\vdots$$
$$LV_i = b_{i.1}x_1 + b_{i.2}x_2 + \cdots + b_{i.n}x_n \quad \text{(Equation 3),}$$

where LV are the latent variables, i the number of the LVs, b are the variable coefficients, x are the molecular descriptors and n the number of descriptors.

Each LV (equation 3) is a linear combination of the x values and also their corresponding coefficient (b), which gives an approximation to the variation of the target value (y) (Leach & Gillet, 2007). This method decomposes the input matrix of descriptors into loadings and LVs and the later are orthogonal and are capturing the descriptor information (Sethi, 2012).

1.6.3 Decision Trees (DTs) and Random Forest (RF) in machine learning

Decision Trees (DTs) are algorithms that are used for both regression and classification models and thus they are usually referred as Classification And Regression Trees (CART) (Brownlee, 2016). The DTs are predictive models that map observations to target values (Lodhi, 2010). DTs as every machine learning algorithm has an input and output. In ADME predictive modelling, the aim is to develop a model that can predict the value of a target (e.g. permeability, lipophilicity, protein binding etc.) based on a set of input variables (descriptors) (Tsaioun & Kates, 2011). The input data are in the form of $(\mathrm{x}, \mathrm{y}) = [(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n), \mathrm{y}]$, where n is the number of descriptors and y represents the target value. In a DT, there are three types of nodes: a root node, internal nodes, and leaf nodes. Leaf nodes are also known as terminal nodes. An example of how a DT works is shown in figure 4. It is an example of a classification problem and thus the DT classifies the compounds on target property y_1 or y_2. The

classification of the test compounds is based on the leaf/terminal node that they reach after going through a series of questions (Yee & Wei, 2012). For example, according to the DT shown in figure 4, a test compound will be assigned with the y_1 if it displays a certain condition for molecular descriptor A. If it does not fulfil that condition, then the molecular Descriptor B is examined. If the molecular descriptor B has a value less than 1, then the test compound will be assigned with the target property y_1 or if it has a value greater or equal to 1, then the test compound will be assigned with the target property y_2.

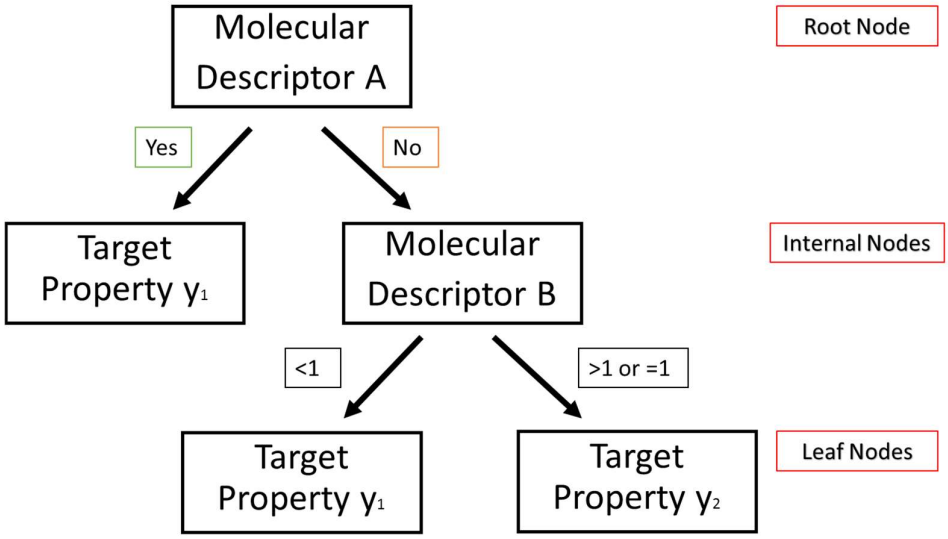

Figure 4: Schematic representation of a decision tree (adapted from Dehmer et al, 2012).

As the example above shows, a DT works by systematically subdividing the information within a training data (in the root and internal nodes) based on rules and there are various algorithms to define these rules (Dehmer et al., 2012). One of them is the recursive binary splitting (Brownlee, 2016). According to that algorithm, different split points are tried and evaluated with a cost function. The cost function that is used for regression models is expressing the sum squared residuals and is the following:

$$\sum_{i=1}^{n}(y_i - prediction_{i)}{}^2 \quad \text{(Equation 4)},$$

where i is the number of compounds and y the experimental value

The output of this algorithm represents the assignment of y value of each leaf for the test set compounds. However, this procedure provides a greedy approach because at each step a split point is defined, which might be good for that specific step but not for the overall of the DT. This limitation of the DTs can be overcome with the use of ensembles DTs like Random Forest (RF).

RF is based on an ensemble of DTs (Mitchell, 2014; K. Roy et al., 2015), which are built by training data of multiple features. Ensemble is the procedure that combines the results/predictions from multiple predictive algorithms in order to make a more accurate prediction compared to each individual prediction (Brownlee, 2016), as it benefits from the "wisdom of crowds" effect (Mitchell, 2014). RF is an improvement of the DTs because the learning algorithm is limited to a random sample compared to DTs, which are searching all the data to identify the ideal split point based on the minimisation of the sum of the squared residuals. The data are partitioned into progressively increasing homogeneous group through the tree. As a result, each terminal node of the DTs is comprised by molecules, which exhibit a similar value of the ADME property evaluated (Mitchell, 2014). RF is generally a unique combination of prediction accuracy, model interpretability and it is able to handle missing values and a variety of variables (binary, continuous, categorical) (Qi, 2012). RF can be used to perform both classification and regression models (Bajorath et al., 2012; Oprea, 2006) and the choice depends on the property that is predicted. Therefore, it is increasingly used in the field of biological computational sciences (Yang, Yang, Zhou & Zomaya, 2010). RF is an algorithm used in the literature to develop ADME predictive models like lipophilicity (Rodgers, Davis, Tomkinson, & van de Waterbeemd, 2011; Schroeter et al, 2007; Wang et al., 2015), permeability (Fredlund, Winiwarter, & Hilgendorf, 2017) and solubility (Palmer, O'Boyle, Glen, & Mitchell, 2006). The main disadvantage of RF method is that its performance can be influenced by a small sample size and also by the number of trees selected (Dehmer et al., 2012). The selection of optimal parameters can be achieved through cross validation (Statnikov, Wang, & Aliferis, 2008).

1.6.4 Support Vector Machines (SVM)

The SVM is an algorithm developed by Vapnik and co-workers and it is a widely used algorithm in the field of data mining in cheminformatics. It can be used for both classification and regression problems and when it is used for continuous/regression models can be referred as Support Vector Regression (SVR). It is an algorithm extensively used to predict properties like hERG blockade (Doddareddy, Klaasse, & IJzerman, 2010; Li, Jørgensen, Oprea, & Brunak, 2008), toxicity related properties (Mitchell, 2014), protein inhibition (Dong et al., 2009) etc. It has also been used to predict physiochemical properties like $LogD_{7.4}$ (Schroeter et al., 2007; Wang et al., 2015), melting point (Hughes, Palmer, & Nigsch, 2008) and pK_a (Harding & Wedge, 2009). For example, for a two class classifier in a 2D space, where the data are linearly separable, the SVM algorithm aims to find the maximum margin hyperplane that divides the data in a way that all the data with target value +1 lie on the opposite site from those with target value -1 (Basak, Pal, & Patranabis, 2007). This hyperplane is also referred as separate hyperplane and the margin is the distance between the separating hyperplane and data samples that are closest to that hyperplane and are called support vectors (Raschka, 2015) (figure 5). Therefore, the SVM for a classification problem aims to identify the optimal hyperplane for which the margin of separation between the chemical compounds is maximised (Khan, 2012). If w is a normal vector to the hyperplane then the hyperplane equation can be written as:

$\overrightarrow{w}\vec{x} - b = 0$ (Equation 5)

and the equations of the two parallel hyperplanes can be written as:

$\overrightarrow{w}\overrightarrow{x_+} - b = 1$ (Equation 6),

$\overrightarrow{w}\overrightarrow{x_-} - b = -1$ (Equation 7).

As the w vector is perpendicular to the hyperplane it is also perpendicular to the parallel hyperplanes and therefore the vector from the x(-) to x(+) is scalar multiple (r) of the vector w and the following equation can be written:

$\overrightarrow{x_+} = \overrightarrow{x_-} + \mathrm{r}\overrightarrow{w}$ (Equation 8).

By using equation 6 and substitute equation 8 to the x(+), the equation 9 is obtained:

$$\text{(Eq.6)} \overset{(Eq.8)}{\Longrightarrow} \overrightarrow{w}(\overrightarrow{x_-} + r\vec{w}) - b = 1 \Rightarrow$$

$$\Rightarrow \overrightarrow{w}\overrightarrow{x_-} + r\left|\left|w\right|\right|^2 - b = 1 \Rightarrow$$

$$\Rightarrow \overrightarrow{w}\overrightarrow{x_-} - b + r\left|\left|w\right|\right|^2 = 1 \Rightarrow$$

$$\Rightarrow -1 + r\left|\left|w\right|\right|^2 = 1 \Rightarrow$$

$$\Rightarrow r\left|\left|w\right|\right|^2 = 2 \Rightarrow$$

$$\Rightarrow r = 2/\left|\left|w\right|\right|^2 \text{ (Equation 9)}$$

The Margin (M) is the half of the distance between x(-) and x(+). Therefore:

$$2M = \left|\left|x_+ - x_-\right|\right| = \left|\left|rw\right|\right| \overset{(Eq.5)}{\Longrightarrow}$$

$$\Rightarrow \left|\left|rw\right|\right| = \frac{2}{\left|\left|w\right|\right|^2}\left|\left|w\right|\right| \Rightarrow$$

$$\Rightarrow 2\mathrm{M} = \frac{2}{\left|\left|w\right|\right|}(Equation\ 10)$$

Therefore, this margin distance should be maximised to identify the optimal hyperplane.

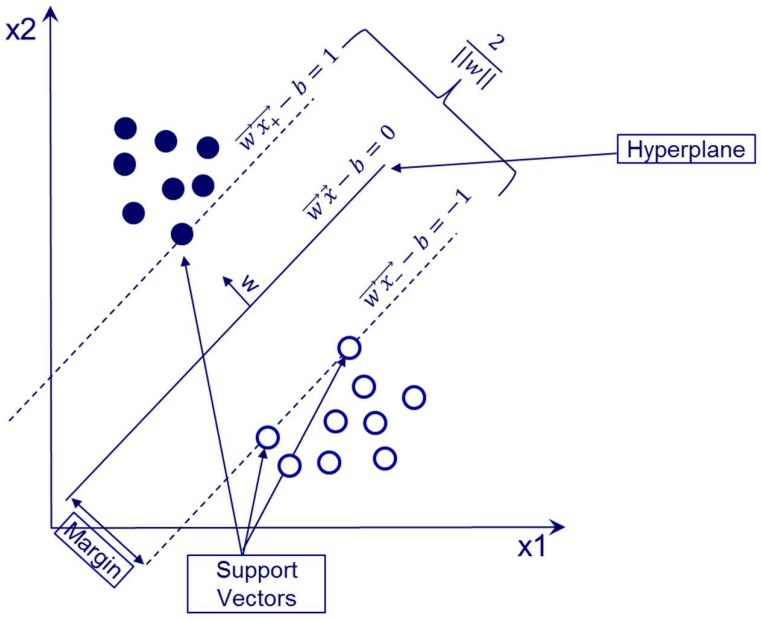

Figure 5: Schematic representation of two data classes in a 2D space by the SVM algorithm.

The case outlined above is the simplest case, where the data are linearly separable in a 2D space and can be easily schematically represented. In more complicated cases, where the data i) are not linearly separable, ii) exist in a higher dimensional space and iii) the aim is the development of a regression model, there are additional strategies to follow. In the non-linearly separable cases, the data are projected in a higher dimension space with the aim to be able to linearly separate them. The kernel trick is used to map the training set data into a higher dimensional space with a mapping function (Φ) (Khan, 2012). There are various kernels that could be used and one of the most widely used is the radial basis function (rbf) kernel ($K(x, x_i)$) for two samples/vectors x, x_i of the input space (Raschka, 2015). The rbf kernel can be expressed as the inner product of the projected x, x_i and uses the following equation to map the data in a higher dimension:

$$K(x, x_i) = e^{-\gamma x \sum((x-x_i)^2)} \quad \text{(Equation 11),}$$

where x, x_i are two vectors of the input space and γ is a hyperparameter.

To train the data with the SVM algorithm and the rbf as a kernel, three hyperparameters (ε, γ and C) should be optimised. The ε parameter is affecting the number of support vectors and it can have a value in the range of 0-1. The larger the ε value is, the lower is the number of support vectors (Khan, 2012). The γ parameter is also taking values in the range of 0-1 and the usual default value is 0.1. If the γ increases, the influence of each data sample is also increased (Raschka, 2015). The C parameter is one of the most important parameter because it can affect both the trained and predicted data (Wang et al., 2015). The C value represents a

balance between the margin maximisation and the training error minimisation (Khan, 2012). If the C is too large then the SVM algorithm will produce an overfitted model (Brownlee, 2016) and if it is too small, insufficient stress is introduced on fitting the training data (Khan, 2012; Wang et al., 2015). A grid search is usually used to find an optimal combination for the three hyperparameters described above.

1.6.5 Konstanz Information Miner (KNIME) in QSPR model building

Literature databases have significantly increased the availability and accessibility of data (Schadt, Linderman, Sorenson, Lee, & Nolan, 2010) and as a result there is a high demand of data mining tools that respond to these needs. KNIME is a data mining workflow framework, which has significantly evolved to meet the new demands of automating predictions and machine learning. It is a pipeline package, which provides a user friendly workspace (Mazanetz, Marmon, Reisser, & Morao, 2012). It uses nodes for data input and various nodes are interconnected to create a pipeline, where information is flowing through them (a process known as "visual programming"). This software offers the advantage of preparing workflows, which can be quickly customised to manage data and information in order to automatize tasks (Mazanetz *et al.*, 2012). KNIME is used in both academia and industry and special nodes have been designed for the KNIME software, which can be used in chemistry, biology and in drug design process. For example, two cheminformatics node packages that are widely used for the development of ADME models are the: ChemAxon/Infocom Marvin package and the Weka (Waikato Environment for Knowledge Analysis). Other examples of package nodes, which are developed from the KNIME community contributions are: Enalos (Melagraki, Afantitis, Sarimveis, & Koutentis, 2009; Melagraki & Afantitis, 2013) and RD-kit, Chemical Development Kit (CDK) (Mazanetz *et al.*, 2012). The KNIME software is coded in Java based on an Eclipse environment (Warr, 2012) and thus it is an extensible programme through plug-ins, which offers additional functionality (Berthold et al., 2009). KNIME also offers nodes, which are serving as interfaces for statistic/mathematic programmes (Matlab, R), programme languages (Python) and database readers (Jagla, Wiswedel, & Coppée, 2011). Finally, KNIME can be used in the development of ADME models because data mining and specialised KNIME nodes can be used for the development of predictive models.

1.7 Model Validation

Model validation is a very important process that should be performed after the model training. Model validation can be internal or external (Chackalamannil et al., 2017). An example of internal validation is the k-fold cross-validation, which partitions the initial dataset in k samples. Then a subsample is excluded and a model is built with the k-1 subsamples as training set. This procedure is repeated for k times and every subsample has been used once as the validation test set (Alpaydin, 2014). Moreover, an external validation set should also be used because it investigates the generalisability of the model to predict new chemicals (Puzyn et al., 2010). There are also measures that estimate the goodness-of-fit of the model. Two of the most commonly used are the Root Mean Square Error (RMSE) in prediction and the Pearson Correlation coefficient or the coefficient of variation in the fit to training set (R^2)

(Chackalamannil et al., 2017). The RMSE is a useful measure as it has the same units as the units in the QSPR experiment and provides indication of the likely error associated with the model's predictions. The RMSE (equation 12) is generally used as a statistical metric to establish model performance (Chai & Draxler, 2014) and the lower the values of RMSE the higher the accuracy of the model. The R^2 (equation 13) is often used to measure model quality (Wermuth, 2008). According to Wermuth (2008) the R^2 can be misleading because it depends heavily on the variation, whereas RMSE relates directly to the experimental variability but it is meaningful to report both values (Alexander & Tropsha, 2015).

$$RMSE = \sqrt{\frac{(Observed - Predicted)^2}{N}} \quad \text{(Equation 12),}$$

where N is the number of compounds

$$R^2 = 1 - \frac{\sum(Observed - Predicted)^2}{\sum(Observed - Observed\ mean)^2} \quad \text{(Equation 13)}$$

Another important way to validate and assess the QSPR models is the evaluation of their Applicability Domain (AD). A focus on methods for the AD evaluation is given in this thesis.

1.7.1 Applicability domain (AD)

Applicability domain is considered as one of the most important problems in the QSPR analysis (Tropsha, 2010). AD can establish the scope and limitations of QSPR models (Netzeva et al., 2005) and it can estimate the range of chemical compounds whose properties can be reliably predicted (Jaworska et al., 2005). AD is actually estimating the confidence in predictions or in other words it is predicting the predictability (Dragos, Gilles, & Alexandre, 2009) and it is considered as a tool to avoid predictions with a large error probability. Moreover, it is generally accepted that the compounds that are "close" to the model's chemical space (based on the training set) have higher chances to have their properties more accurately predicted than compounds that are "far" (Cumming et al., 2013). Therefore, the chemical space of the model must be defined and then asses if the compounds in the test set fit into that space. The AD is dependent on the descriptors that are used for the model. The descriptors are numerical representations of the chemical space (Todeschini & Consonni, 2009) and thus by changing the descriptors, the chemical space is also altered (Mathea, Klingspohn, & Baumann, 2016). Moreover, there is also a possibility of presence of compounds that are "far" from the model's chemical space and they are called prediction outliers. These can be present in both train and test sets (Furusjö, Svenson, Rahmberg, & Andersson, 2006).

1.7.1.1 Distance to model metrics

There are various ways to establish the AD of a model and one of them is the distance to model metrics. These approaches calculate the distance of the test compounds from a defined

point within the chemical space of the training compounds (Sahigara et al., 2012). The distances are compared between this defined point and compared to a user-pre-defined threshold. Some of the most commonly used methods are the following: Euclidean, Manhattan and Mahalanobis distance and Leverage test.

1.7.1.2 Mahalanobis distance

Mahalanobis distance (MD) is measuring the distance of a given compound (i.e. a test compound) from the distribution of the training set compounds (equation 14). MD takes into account the correlation in the data since it uses the inverse of the covariance matrix of descriptors (Netzeva et al., 2005). Other methods like Euclidean distance and Manhattan distance cannot do that automatically and other pre-treatments like PCA are necessary (Gadaleta et al., 2016). Moreover, MD is a method that can be used to detect potential multivariate outliers, which are actually compounds really far from the compounds' distribution and also squared MD approximately follows a chi-square distribution (Varmuza & Filzmoser, 2016). These features can be used to set a threshold and distinguish between compounds that are within an acceptable distance from the model.

$$Mahalanobis\ Distance\ (MD) = \sqrt{(x-\mu)S^{-1}(x-\mu)^T} \quad \text{(Equation 14),}$$

where MD is the distance of an observation x from a set of descriptors with mean μ and S (covariance matrix) and T is the transpose of the matrix.

1.7.1.3 Leverage

Another distance to model metric to estimate the AD is the Leverage method, which is based on the concept of the extent of extrapolation (Melagraki et al., 2009). The model space is comprised by a k-Dimensional space of the n chemicals (rows) and k variables (columns) and this is the X = k x n, the descriptor matrix. The leverage method measures the distance of each compound from the centroid of X matrix (Netzeva et al., 2005), by manipulating the Hat matrix (H), which is the following:

$$H = X(X^T x\, X)^{-1}X^T \quad \text{(Equation 15),}$$

where X is the descriptor matrix and X^T is the transpose matrix of X.

The next step involves the calculation of the leverages (h_i), which are the diagonal elements of the H matrix and are calculated with the following equation:

$$h_i = X_i^T (X^T x\, X)^{-1} X_i$$ (Equation 16),

where X_i is the descriptor row vector of the query compound and X is the descriptor matrix.

The final step of the leverage method involves the estimation of the threshold, which is fixed at 3p/n, where p is the number of variables/descriptors plus one and n is the number of compounds in the training set (Gadaleta et al, 2016; Puzyn et al., 2010; Sahigara et al., 2012).

1.7.1.4 Other Distances

There are also other distances that are used for the estimation of AD like the Euclidean distance (ED) and the Manhattan distance (ManD). ED is the square root of the squared differences between the corresponding elements in the descriptor matrix of two compound A and B (equation 17). ManD, between two compounds A and B, is the sum of the absolute differences of their coordinates in the n-variable/descriptor space (equation 18).

$$Eucliean\ Distance(ED) = \sqrt{(x_{A1} - x_{B1})^2 + (x_{A2} - x_{B2})^2 + \ \ldots + (x_{An} - x_{Bn})^2}$$ (Equation 17),

where ED is the distance of 2 compounds A and B with n descriptors.

$$Manhattan\ Distance\ (ManD) = \sum_{n=1}^{n} |An\ - Bn|$$ (Equation 18),

where A and B are two compounds and n is the number of descriptors.

1.7.2 k-Nearest Neighbour (kNN)

This method is establishing the distance of a test/query compound from its nearest k compounds in the training set (Sahigara et al., 2012). However, this method is not a pure distance to model metric method because it also takes into account the structural or chemical similarity of the compounds (Sahigara, Ballabio, Todeschini, & Consonni, 2013). The similarity of the test compounds to the training compounds can be assessed by using: a) descriptors, b) Principal Components (PCs) and c) Extended Connectivity Fingerprints (ECFP4). The distance between the compounds can be computed using different distance functions. The ED and the ManD can be used to calculate distance between compounds with the descriptors and Tanimoto and Dice coefficients can be used to calculate similarity with the ECFP4 fingerprints.

1.7.3 Fingerprints and Similarity measures used with kNN

Fingerprints are a popular method to evaluate chemical similarity due to their ability to translate the chemical complexity into a numeric string (Gadaleta et al., 2016). ECFP have been developed as a modified Morgan algorithm methodology (Leach & Gillet, 2007) to represent molecular characteristics, which are associated to their molecular activity (Rogers & Hahn, 2010) and they can also be used for other purposes like chemical similarity. In addition, they exhibit several advantages like that they are rapidly calculated, they can represent a great number of different molecular features and they are able to reflect both the absence and the presence of a chemical functionality (Kovacs, 2016).

Tanimoto (equation 19) and Dice (equation 20) coefficients are similarity measures, which take into account the overlapping of chemical fingerprints to quantify molecular similarity (Jasial, Hu, Vogt, & Bajorath, 2016). The difference between Dice and Tanimoto is that Dice gives twice the weight to the positive common bits and as a result emphasises more on the positive matches (Al-Shamri, 2014), whereas Tanimoto is really popular because in includes a degree of size normalisation with the denominator term (Leach & Gillet, 2007). Both give a range of 0-1, where 0 means no similarity and 1 means highest similarity.

Tanimoto	Dice
• $\boldsymbol{T} = \frac{\boldsymbol{Nc}}{\boldsymbol{Na+Nb-Nc}}$ (Equation 19), where ➢ Na the number of bits set to "1" in molecule A, ➢ Nb the number of bits set to "1" in molecule B and ➢ Nc the number of bits in both A and B.	• $\boldsymbol{D} = \frac{\boldsymbol{2xNc}}{\boldsymbol{Na+Nb}}$ (Equation 20), where ➢ Na the number of bits set to "1" in molecule A, ➢ Nb the number of bits set to "1" in molecule B and ➢ Nc the number of bits in both A and B.

1.8 Principal Component Analysis (PCA)

PCA is a method used in multivariate data analysis, in which the observations are described by inter-correlated quantitative dependent variables (Abdi and Williams, 2010). PCA could be used as part of the model validation process to establish if the compounds in the test set occupy a similar chemical space as the training compounds. The aim of this method is to reduce the dimensionality of the data, to extract the important information from the data table and express this information as a set of new variables (Abdi & Williams, 2010). As a result, the data are represented with a smaller number of variables, which are the result of the reduction of dimensionality and are called principal components (PCs) (Ringnér & Ringner, 2008; Yousefinejad, Bagheri, & Moosavi-Movahedi, 2015). The concept behind the PCA is to find PCs (e.g. PC1, PC2, …, PCn), which are linear combinations of the original variables (Varmuza & Filzmoser, 2016), which in this case are the QSPR descriptors. In addition, the PCs are chosen in a way that the first principal component (PC1) accounts for the most of the variance

in the data, the PC2 for the next largest variance etc. (Miller & Miller, 2010) The PCs are orthogonal linear combination transforms of the original descriptors (Hemmateenejad, Miri, & Elyasi, 2012).

To calculate the PCs of a matrix, which is composed by x compounds and n-descriptors (i.e. n-dimensional space), four simple steps should be followed. Firstly, the mean of each dimension is calculated and the mean is subtracted from each dimension, producing a data set whose mean is zero (Smith, 2002). The second step is the calculation of the covariance matrix, which is formed by measuring covariance values (Equation 21) between all the dimensions (Fukunaga, 2013). The covariance matrix is a square matrix, from which the eigenvalues and eigenvectors are calculated, which can reveal information for the data (Tran, Vu, & Wang, 2013). The eigenvalues and eigenvectors are special features of a matrix. An eigenvector x ($x \epsilon R^n$) is a non-zero vector of a matrix A, when Ax is a scalar multiple (λ) of x: $Ax = \lambda x$ (Anton, 2010). The scalar multiple λ is called eigenvalue of matrix A and corresponds to x eigenvector.

$$cov(Z,Y) = \frac{\sum_{i=1}^{n}(Z_i - Z_{mean})(Y_i - Y_{mean})}{(x-1)} \quad \text{(Equation 21),}$$

where z and y are 2 dimensions of the n-dimensional space and x is the number of compounds (i.e. sample size).

An eigenvalue decomposition of the matrix is performed to obtain the eigenvalues, which represent the total variance explained by the corresponding eigenvector, which indicates the direction of the new axes (Smith, 2002). At the beginning, the compounds' dataset is described with values, which cannot relate to the rest of the data, whereas the new data points (scores, equation 22) of PCs show how the points are related to the rest of the data.

$$Scores = orginal\ data\ x\ eigenvector \quad \text{(Equation 22)}$$

1.9 Permeability

Permeability is considered as a valuable parameter during the drug discovery process because it can significantly affect the ADME properties and it correlates to the velocity of a compound passage through a biological membrane barrier (Di & Kerns, 2016). Permeability extendedly affects the absorption and thus the bioavailability because a low permeable compound is not able to cross the cell membranes and ultimately interact with the biological target. Permeability also affects Distribution because it relates to the ability of the drugs to penetrate BBB and cell membranes. The ability or inability of a drug to permeate a biological membrane barrier (usually the intestinal membrane barrier) impacts on the drug's efficacy. Therefore, a low permeability value results in a reduced bioavailability, which ultimately prevents the formulation

of orally administered drugs. This is a major limitation because the oral route is the most desired and is associated with high patient compliance (Wang & Hou, 2015).

1.9.1 Structure of the cell membrane and Drug Transport

To appreciate the drug permeability through the biological cell membranes, it is significant to consider the morphology and structure of the cell membrane. The main structural component for all biological membranes is the lipid bilayer, which is consisted of amphipathic phospholipids (figure 6) (Yeagle, 2011). The chemical structure of phospholipids is characterised by a head group and long saturated lipophilic side chains. The variations in head group result in different types of phospholipids like Phosphatidylcholine (PC), Phosphatidic acid (PA), Phosphatidylglycerol (PG), Phosphatidylethanolamine (PE) and others. Additionally, cell membranes also contain membrane transporters like ion channels and uptake or efflux transporters (Goñi, 2014).

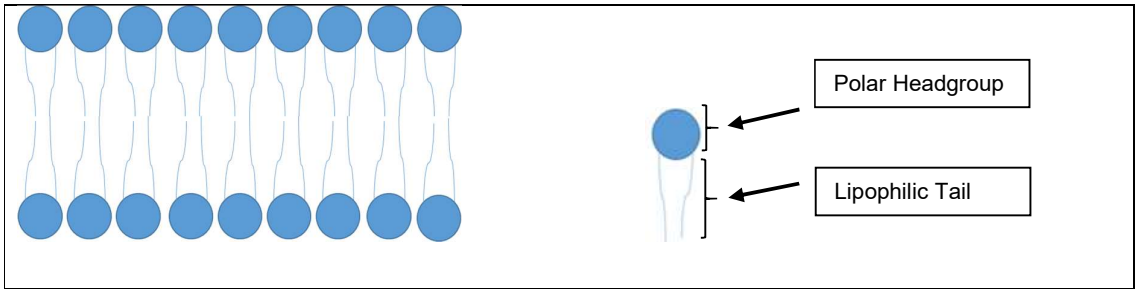

Figure 6: Illustration of the lipid bilayer and the structural unit of the lipid bilayer, the phospholipids.

The main methods that a drug can overcome biological barriers are the passive transport and active transport (figure 7). The passive transport mainly refers to either passive diffusion or paracellular permeability because no energy is required. In transcellular passive diffusion, the drug cross the cell membrane driven by Brownian motion (Di & Kerns, 2016) due to the concentration gradient (Tsaioun & Kates, 2011). The drug moves from the aqueous phase, where the drug is in high concentration, into the cell. Drugs that undergo paracellular permeability move between the tight junction of epithelial cells (figure 7). However, this route is only observed for a low percentage of drugs in the intestines (less than 5%) (Di& Kerns, 2016). Passive diffusion is believed to be the main mechanism for intestinal absorption and it has been reported that about 95% of the commercial drugs undergo passive diffusion (Mandagere & Thompson, 2002). Active transport is mediated by protein transporters present in the cell membrane (Kell, Dobson, & Oliver, 2011; Kell & Oliver, 2014).

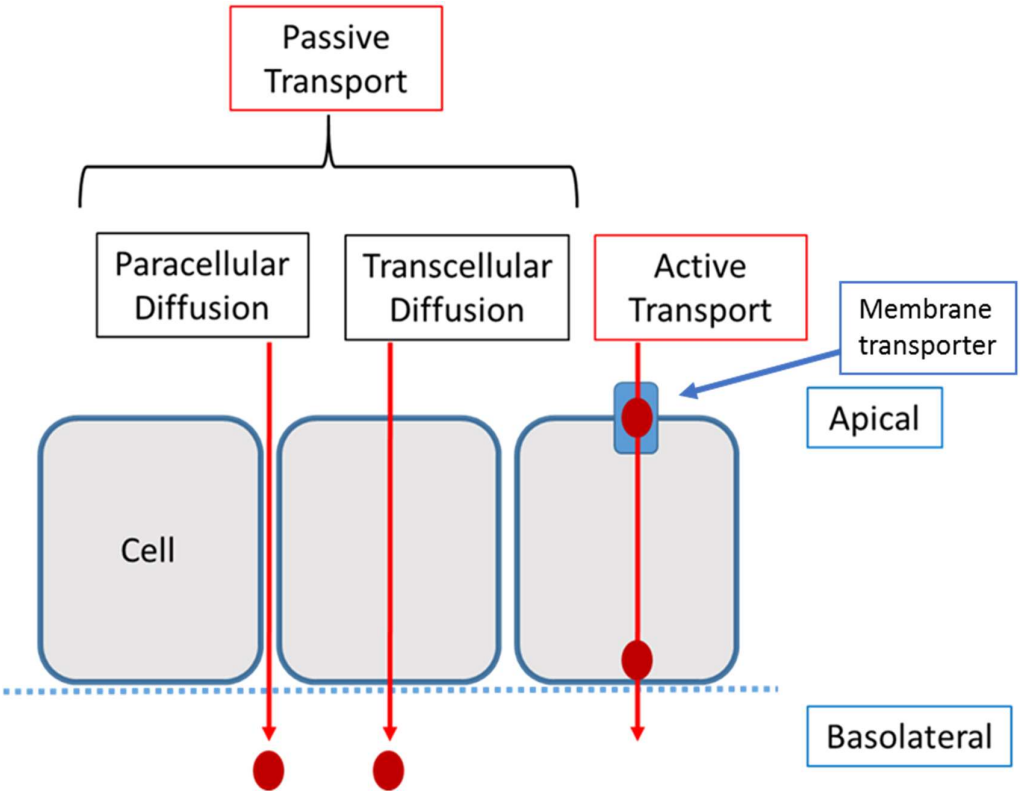

Figure 7: A simplified view of the two main permeability mechanisms.

1.9.2 *In-vitro* models of cell permeability

Permeability can be measured by various assays and is correlated to drug absorption. Monolayers of Caco-2 cells are used as an *in-vitro* model to estimate intestinal absorption of compound. These cells are human epithelial colorectal adenocarcinoma cancer cell lines (van Breemen & Li, 2005) and are extensively used (Cyprotex, 2015; Li, Volpe, Wang, Zhang, & Bode, 2011). The Caco-2 cells have the ability to mimic the morphology and functionality of the human enterocytes (Press, 2011) and since the cells are derived from colon adenocarcinoma, they exhibit both colonocytic and enterocytic characteristics (Volpe, 2008). The assay is based on the ability of Caco-2 cells to undergo spontaneous enterocyte differentiation in cell culture and replicate into confluent monolayers (Ehrhardt & Kim, 2008). When they are at the confluent state on a semi-porous membrane they start to polarise and form tight junctions resulting in a polarised apical (side A) and a basolateral (side B) membranes, which are creating an environment similar to human enterocytes (figure 9). As a result, they can be used as a surrogate (Thomas, Brightman, Gill, Lee, & Pufong, 2008) to predict permeability and transport of drugs.

This assay provides information related to *in-vivo* absorption of the small amounts of tested compound, as well as the rate of absorption, which consequently controls bioavailability (Pham-The et al., 2016). The Caco-2 assay is considered to be a 'gold standard' in calculating *in-vitro* permeability (Caldwell, Yan, Tang, Dasgupta, & Hasting, 2009) and it is also accepted

by the Food Drug Administration (FDA) as the assay to aid classification of compounds according the Biopharmaceutics Classification System (BCS) (Ku, 2008) (figure 8).

	High Solubility	**Low Solubility**
High Permeability	Class 1 High Solubility High Permeability Rapid Dissolution	Class 2 Low Solubility High Permeability
Low Permeability	Class 3 High Solubility Low Permeability	Class 4 Low Solubility Low Permeability

Figure 8: Biopharmaceutics Classification System (BCS)(adapted from Benet, 2013)

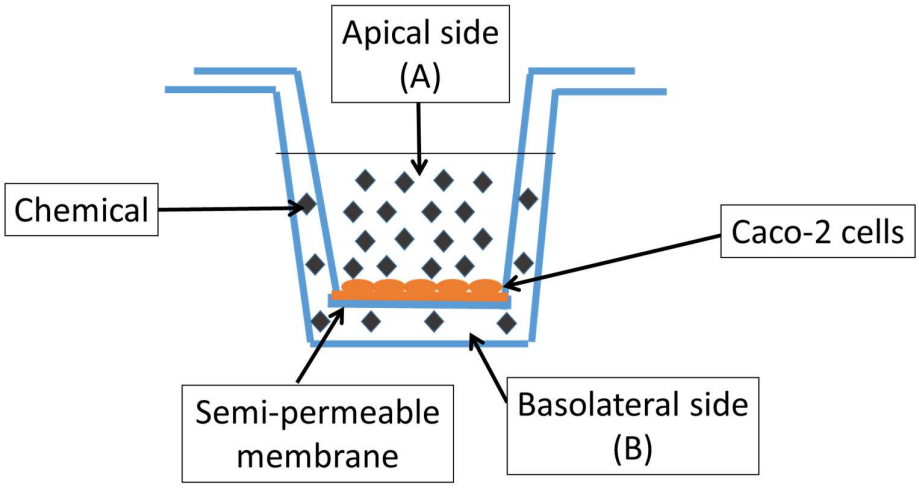

Figure 9: Schematic representation of the Caco-2 permeability assay (adapted from Li, 2001).

This assay is usually conducted in 96 well plates (Sampson et al., 2014) and the compound of interest is introduced to apical side (side A) and then the flux of the compound through the monolayer is measured. After incubation time, the amount of compound, which crossed the Caco-2 cells into the basolateral side (side B) is established usually with LC/MS (Liquid Chromatography/Mass Spectroscopy) or MS (Cyprotex, 2015). Finally, the P_{app} (apparent permeability) is calculated (equation 23) with the following formula (Volpe, 2008):

$$P_{app} = \frac{dM/dt}{C_0 \times S}$$ (Equation 23),

where C_0 is the initial concentration of the compound tested, S is the surface area of Caco-2 cell membrane and dM/dt is the rate of the amount of the compound transported to side B.

There are also other methods used to measure *in-vitro* permeability like the Parallel Artificial Membrane Permeability (PAMPA) and Mardin-Darby Canine Kidney (MDCK) assays. The PAMPA assay uses a an artificial lipid membrane to assess the likelihood of a drug to undergo passive diffusion (Avdeef & Tsinman, 2006). The MDCK assay is performed on cell lines derived from canine/dog kidney and this assay shows a good correlation with the Caco-2 assay (Cyprotex, 2015). However, the PAMPA assay can only predict passive diffusion and the cell lines used in MDCK assay exhibit differences in morphology between the canine and human cells. In addition, although Caco-2 assay is the best, it has some disadvantages. Caco-2 cells require a long culture period (~3 weeks) (Wang *et al.*, 2016) and also under or overpredict the permeability of drugs that undergo active transport (Fredlund et al., 2017) due to the different expression levels of transporters in Caco-2 cell line. Therefore, pharmaceutical industries should develop efficient models for the *in-silico* prediction of permeability, which is less time consuming and expensive compared to the *in-vitro* Caco-2 assay.

1.9.3 *In-silico* regression permeability models developed with Caco-2 data

There are various regression permeability models based on Caco-2 permeability data. The models were developed using different training sets, descriptors and algorithms like MLR, PLS, ANN, Boosting, SVR and others (table 1).

Table 1: Regression permeability models developed with Caco-2 data reported in the literature during 1997-2010.

Reference	Method	Number of Molecules	Number of Descriptors	AD Estimation? (Yes/No)
(Norinder, Österberg, & Artursson, 1997)	PLS	17	9	No
(Kulkarni, Han, & Hopfinger, 2002)	GFA (genetic function approximation)	38	6	No
(Fujiwara, Yamashita, & Hashida, 2002)	ANN	87	5	No
(Yamashita & Wanchana, 2002)	GA-PLS	73	24	No
(Nordqvist, Nilsson, & Lindmark, 2004)	PLS	51	70	No
(Hou, Zhang, Xia, & Qiao, 2004)	MLR	100	4	No
(Guangli & Yiyu, 2006)	MLR SVR	100	4	No
(Jung, Choi, Um, Kim, & Choo, 2006)	MLR	20	4	No
(Fenza, Alagona, Ghio, & Leonardi, 2007)	GA-ANN	41	5	No
(Karelson, Karelson, Tamm, & Tulp, 2009)	MLR ANN	81	6	No
(Paixão, Gouveia, & Morais, 2010)	ANN	296	12	No

Although many models reported in table 1 performed well, there are some limitations related to these studies, like the small size of the training set (table 1). This is an important factor in

predictive modelling, as a small training set (e.g. less than 100 compounds) could possibly lead to models which are not robust. For example, there is a high chance of producing overfitted models and additionally the presence of outliers could significantly affect the predictive activity (Tropsha, 2010). Furthermore, a model developed with a small training dataset would exhibit a restricted AD and thus an evaluation of the AD would be of high significance. As table 1 shows, the AD of the models (i.e. how close are the training compounds to the training set) was not established. This is a clear disadvantage because the AD can be a measure that explains a good or bad prediction based on the percentage of test compounds that fall within the AD. In addition, a small number of descriptors was used. This is a potential limitation because there is a variety of chemical features, which could be associated with ADME properties like permeability (Tao et al., 2015). Therefore, by using only a small number of descriptors, this variability might not be considered. Finally, most of the models reported (table 1) used training sets with existing drugs except the model developed by Fenza et al (2007), which was developed with proprietary only compounds. Furthermore, a model reported in 2010 (Paixão et al., 2010) used a reasonable training set, larger than the previous models, with 12 descriptors and achieved an RMSE of 0.60. An improved ANN methodology used was based on a pruning procedure and early stop approach that prevented overfitting by the model, which was a major issue noticed in the previous studies that used the ANN algorithm. However, the AD of the models was not established.

The recently reported Caco-2 regression models overcame some of the limitations mentioned above and offered the advantage of a large training set which included data extracted from ChEMBL database (Wang *et al.*, 2016) and both proprietary and ChEMBL data (Fredlund et al., 2017) (table 2). In addition, both models showed a good predictive ability and used a variety of algorithms to develop models. However, Wang et al (2016) used only one method to evaluate models' AD and Fredlund et al (2017) did not evaluate or report the models' AD. Furthermore, there are not models in the literature that investigate the effect of literature data in proprietary models in comparison to only proprietary models. The models developed by Fredlund et al (2017) merged proprietary with literature data and showed good results but the effect of merging proprietary with literature data is not reported.

Table 2: Regression permeability models developed with Caco-2 data reported in the literature during 2016-2017.

Reference	Method	Number of Molecules	Number of Descriptors	AD Estimation? (Yes/No)
(Wang *et al.*, 2016)	MLR PLS SVR Boosting	1272	193	Yes: Leverage
(Fredlund et al., 2017)	PLS SVR RF	2558	PLS, SVR: Standard AZ descriptor set RF: signature descriptors	No

1.10 Lipophilicity

Lipophilicity is a property that majorly affects the ADME of a drug and correlates to other properties like permeability, solubility and protein binding. Lipophilicity is the ability of a compound to partition into a nonpolar lipid matrix against an aqueous (Di & Kerns, 2016) and there are two different ways of expressing and calculating lipophilicity: the partition coefficient (LogP) and the distribution coefficient (LogD) (Low, Blasco, & Vachaspati, 2016). LogP is the partition coefficient of a compound between octanol (organic layer) and buffer (aqueous layer) (equation 24), whereas the LogD is the distribution coefficient of a compound between octanol (organic layer) and buffer (aqueous layer) at a specified pH (equation 25) (Caron & Ermondi, 2008).

$$LogP = log_{10}\left(\frac{C_{organic}}{C_{aqueous\ pH-all\ molecules\ neutral}}\right) \quad \text{(Equation 24)}$$

$$LogD_{pH(x)} = log_{10}\left(\frac{C_{organic}}{C_{aqueous\ pH(x)}}\right) \quad \text{(Equation 25),}$$

where C is the concentration of the compound.

The LogP refers to the partitioning of the neutral form only, whereas the LogD takes into account any acidic or basic groups that ultimately affect the distribution in octanol/water, which becomes pH dependent (Tetko & Bruneau, 2004). Moreover, a LogD value at pH 7.4 ($LogD_{7.4}$) represents the LogD at physiological pH and a value of about 1-3 is the optimal for orally available drugs (Hartmann & Schmitt, 2004). This range is optimal as it results in an intestinal absorption of a drug due to a good balance between solubility and transcellular passive diffusion (Di & Kerns, 2016). Generally, a drug should exhibit a balance between lipophilicity and hydrophilicity in order to be able to dissolve and permeate cell membrane barriers (Wang et al., 2015). A $LogD_{7.4}$ value of 3-5 shows a good permeability but the main disadvantage is that the intestinal absorption and bioavailability is reduced. A $LogD_{7.4}$ value greater than 5 results in a low solubility absorption and thus bioavailability. In addition, distribution is affected because the compound is too lipophilic and thus it gets trapped in biological tissues (Di & Kerns, 2016).

Lipophilicity can influence the possibility of a drug to be considered as drug candidate. Lipophilicity along with the Topological Polar Surface Area (TPSA) have an impact on toxicological properties of a drug (Hughes et al., 2008) and a calculated LogP<3 and TPSA>75 increase the risk of toxicity (Lu, Jessen, Strock, & Will, 2012). The toxicity increases because a calculated LogP<3 and a TPSA>75 increase the likelihood of promiscuous binding to off target pharmacology (Hughes et al., 2008). In addition, lipophilicity can affect the non-specific binding to albumin and phospholipids (Valko et al., 2012), which results in reduction of the *in-vivo* available concentration (Tarcsay, Nyíri, & Keserű, 2012). Furthermore, the pKa value is another factor that should be considered along with the lipophilicity. The pKa represents the pH that the drug is 50% ionised (Heshmati et al., 2013) and both lipophilicity and pKa determine the pharmacokinetic, pharmacological and toxicological properties of a compound (Di & Kerns, 2016).

The assay that is used to determine the LogD of compounds is the shake flask method, which is considered as the gold standard of determining the lipophilicity (Baka, Comer, & Takács-Novák, 2008). This method measures the compound's concentration in octanol (organic phase) and the aqueous phase after equilibration on both phases (Andrés et al., 2015). However, there are several limitations related to that method. For example, the use of octanol as the organic phase has several limitations (Cyprotex, 2015) because it contains a relatively high amount of aqueous content of about (4%) (Allerton, Smith, Kalgutkar,Amit, van de Waterbeemd, & Walker, 2012). As a result, octanol supports hydrogen bonding (Will, McDuffie, Olaharski, & Jeffy, 2016), which creates a different environment from that of the inner hydrocarbon core of the cell membranes. Thus octanol can overestimate the lipophilicity of compounds that are able to form hydrogen bonds (Allerton et al., 2012). However, octanol remains the most popular organic solvent for these studies in industry (Cyprotex, 2015).

1.10.1 Theoretical lipophilicity prediction and the importance *in-silico* lipophilicity models

In the past, it was extensively reported the challenge of the theoretical prediction of $LogD_{7.4}$ (Tetko & Poda, 2004). $LogD_{7.4}$ was usually measured by calculating the LogP and pK_a values

with the following equation: $(pH) = LogP - \log(1 + 10^{(pH-pKa)\Delta_i})$, where Δ_i is equal to 1 or -1 for acids and bases respectively. However, this approach can be problematic and inaccurate due to the accumulation of errors from LogP and pK_a calculations. The next evolutionary step in the theoretical prediction of LogP and $LogD_{7.4}$ was the development of software. Therefore, pharmaceutical companies tried to evaluate the possible advantage of the available software like ACD labs, Pallas PrologD and ALOGPS in the theoretical calculation of LogD. In a study (conducted in Pfizer), the $LogD_{7.4}$ for two proprietary sets of compounds was predicted with the ACD labs and the Pallas PrologD software, which resulted in very low accuracy in predictions. Pfizer (Tetko & Poda, 2004) and AstraZeneca (Tetko & Bruneau, 2004) utilized "in house" LogD data to evaluate prediction of $LogD_{7.4}$ based on a software, which predicts LogP, the ALOGPS. This software used the Associative Neural Network (ASNN) method, which allowed the user to include new data without retraining the neural network (LIBRARY mode). By incorporating $LogD_{7.4}$ data with the LIBRARY mode, the ALOGPS proved to be similar or superior compared to other software. However, the improvement was observed only for local predictions and it was difficult to produce accurate predictions for compounds with structural features not covered in the training set. Therefore, the importance of developing models for lipophilicity prediction was evident.

One of the initial attempts to develop lipophilicity models based on $logD_{7.4}$ data and the BRNN (Bayesian Regularised Neural Networks) algorithm, was conducted in AstraZeneca (Bruneau & McElroy, 2006). A set of 8200 "in house" compounds was clustered in 5000 clusters based on hierarchical clustering process and one compound from each cluster was selected to form the training set and the rest of the compounds used as "ex-cluster validation test set". In addition, a global validation test set comprised by 16325 compounds was obtained from the AstraZeneca database for "global validation". The advantage was that the model was developed with a consistent and large proprietary dataset. Model seemed to perform well for both test sets with an RMSE in prediction of 0.54 and 0.63. In addition, the AD of the models was established by using the Mahalanobis distance and the compounds were binned in 4 bins by increasing distance. The results showed a trend of increasing RMSE as the MD was increasing. However, the models were developed with only proprietary compounds.

Moreover, there are $LogD_{7.4}$ models (table 3) developed by AstraZeneca (Rodgers et al., 2011) and Bayer Schering Pharma AG (Schroeter et al., 2007) with proprietary compounds and other models developed with data extracted from the literature (Wang et al., 2015). The models could derive accurate predictions with various machine learning algorithms. However, none of these studies focus on the inclusion of literature data in the proprietary models to investigate the use of literature and opensource data in the realistic ADME evaluation in the drug discovery pipeline.

Table 3: Regression lipophilicity models developed with $logD_{7.4}$ data reported in the literature.

Reference	Method	Number of Molecules	Number of Descriptors	AD Estimation? (Yes/No)
(Bruneau & McElroy, 2006)	1. BRNN	5000	122	Yes Mahalanodis Distance
(Schroeter et al., 2007)	1. Gaussian Process 2. Linear ridge regression 3. SVR 4. RF	14556	Dragon descriptors (1664)	Yes Mahalanodis Distance
(Rodgers et al., 2011)	1. RF 2. PLS	Number of molecules varied as the models were updated over a period of 3 years	In-house descriptor set (topological, geometrical and electronic)	Yes Mahalanodis Distance
(Wang et al, 2015)	1. SVR 2. PLS	1130	121	Yes Leverage

1.11 Research Hypothesis and Aims

Pharmaceutical companies develop their ADME predictive models based on proprietary data. Therefore, these compounds are often novel and are not present in the literature. Thus, it is expected that literature data will introduce chemical diversity to the Evotec training space. This assumption is also based on a work conducted in AstraZeneca and Bayer Pharma AG, which concluded that data extracted form ChEMBL can introduce chemical diversity in proprietary databases (Kogej, Blomberg, Greasley, & Mundt, 2013). The results indicated a low molecular similarity between compounds extracted from ChEMBL database and two proprietary screening collections. In addition, various permeability and $LogD_{7.4}$ models are described in the introduction in sections 1.9.3 and 1.10.1, and were developed with either proprietary or literature data. However, none of these studies focussed on the inclusion of literature data in the proprietary models. Only a permeability model was developed in AstraZeneca including both proprietary and public available data (Fredlund et al, 2017) but the effect of the public data on the models performance and applicability domain was not reported.

Therefore, the aim of the present work is to evaluate the effects of the introduction of public data into the training set. In other words, it will be investigated whether literature compounds can be merged with proprietary data and consequently improve the ADME predictions of proprietary models. The objective is to build *in-silico* predictive ADME models by using internal and public data (i.e. literature data) and establish if the literature data can improve the performance of proprietary model and enlarge their AD. The objectives of that work are addressed in the three following parts:

1. The first objective is to evaluate the ability of the existing Evotec permeability model to predict the permeability of literature compounds. In addition, the applicability domain of the existing Evotec permeability models is evaluated with four distance to model metrics, by calculating the distance of the test compounds (compounds downloaded from ChEMBL) from the Evotec training set.
2. The second objective is to evaluate performance of Caco-2 A-to-B permeability models developed using three different algorithms and three different training sets: literature data, proprietary data and merged proprietary and literature data. Additionally, the AD of the models is evaluated to establish if the literature data could enlarge the AD of the models developed with proprietary data.
3. The third objective is to evaluate performance and AD of $LogD_{7.4}$ models using the same approach applied in the second objective.

2 MATERIALS AND METHODS

2.1 Software Framework

The data mining tool, Konstanz Information Miner (KMIME), an opensource data analysis platform, was used in automation of model development (KNIME, 2016). All the work was carried out within the KNIME 3.0 using proprietary and freely available KNIME nodes. Statistical work was carried out using the R statistical language through the R-Snippet node interface in KNIME and the software application R studio (www.rstudio.com), which is an opensource Integrated Development Environment (IDE) for R. The ChemAxon/Infocom (www.chemaxon.com) RDKit and Analytics nodes were also used. The ChemAxon KNIME nodes were used during the descriptor calculation for conversion of the Simplified Molecular Input Line Entry System (SMILES) into 2D structures, their standardisation and calculation of molecular descriptors. The RDkit was used to calculate the Peoe-VSA descriptors. The Waikato Environment for Knowledge Analysis (WEKA) data mining package nodes for KNIME, developed by the University of Waikato, was used for the implementation of the Support Vector Regression algorithm through the LibSVM node. Tibco Spotfire (www.spotfire.tibco.com) and Microsoft Excel was used for additional analysis and to generate plots and graphs.

2.2 Methods used for the evaluation of existing Evotec Caco-2 A to B permeability model with literature data

The existing Evotec Caco-2 A to B permeability model was evaluated with a test set. The test set included Caco-2 A to B permeability data extracted from the ChEMBL database. The predicted values of the ChEMBL test set (obtained with the existing Evotec model) were compared with the experimental values and the quality indicators (RMSE and R^2) were calculated. Four different distance to model metrics were applied to assess how close are the literature compounds to the proprietary training set. Figure 10 shows a schematic summary of the methodology.

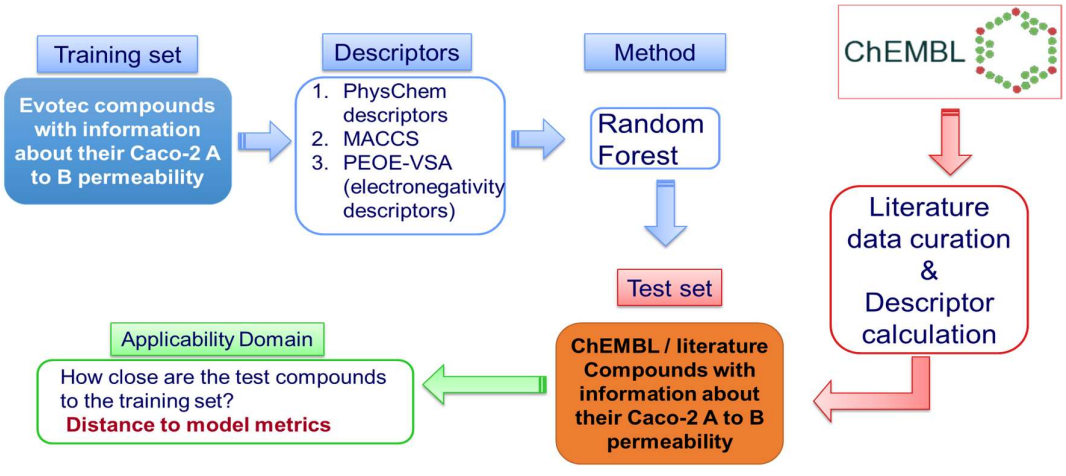

Figure 10: Schematic summary of the work and the methods used for the evaluation of the existing Evotec Caco-2 A to B permeability model.

2.2.1 Literature data curation

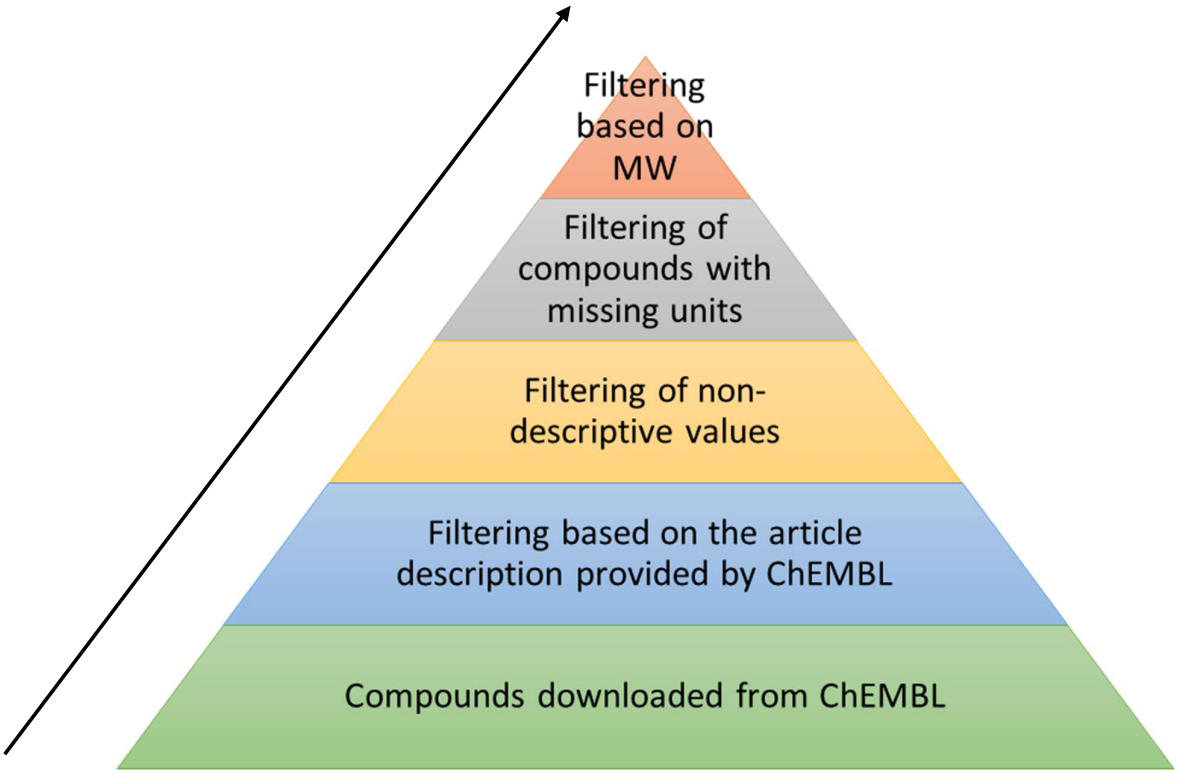

Figure 11: Schematic representation of the literature data filtering process for the compounds downloaded from ChEMBL. The arrow indicates the flow of the process.

Public compounds with Caco-2 A to B permeability data obtained from ChEMBL database. A set of 9473 compounds with Caco-2 assay information have been downloaded from ChEMBL (https://www.ebi.ac.uk/chembl/) v.21. A filtering process was applied to improve the quality and reliability of data (figure 11). A similar filtering process has been described in the literature for the development of Caco-2 QSPR models with compounds downloaded from ChEMBL (Wang *et al.*, 2016).The first step of the filtering process was to keep compounds with information only related to Caco-2 A to B permeability and compounds with MW lower than 750, as the aim is to perform the work described on similar compounds to the Evotec data set. As a result, 2670 compounds were retained. Moreover, only compounds with exact values were included and those molecules with non-descriptive or missing values were removed. Compounds with missing units have been removed. It is worth mentioning that measurement units are not consistent in ChEMBL and thus all experimental data were manually converted to a reference unit of 10^{-6} cm/s and then the Log_{10} of that value was calculated and used. Finally, where there were two or more entries of the same molecules, the permeability value mean and the standard deviation were calculated. When the standard deviation was more than one the compounds were excluded to minimise the error that could arise from a chance selection of one of the values.

2.2.2 Standardisation and Molecular descriptors calculation

After the selection and curation of the chemical compounds, the molecular descriptors were calculated. The Advanced MolConverter from Chemaxon/Infocom node was used to convert the SMILES structures into chemical structures in “Marvin document” (mrv) file format and a “Standardizer” node was applied to convert the molecules representation into a standard form (figure 13). The standardisation is essential because chemical compounds might appear in different forms depending on the source that they are obtained form. The presence of tautomers or resonance might be a potential problem in the representations and thus standardisation is required. For example, the amino group might be represented in two different forms: the charged (NH_3^+) or the neutral (NH_2) form (figure 12). Standardisation process ensures consistency in the way that chemical structures are represented prior to the descriptor calculation. The “Standardizer” KNIME node was configured by selecting the four following actions: “strip salts”, “neutralize”, “tautomerize” and “aromatize” (ChemAxon, 2016b). The “strip salt” removes predefined fragments from multi-fragment molecules (regarded as salts). The “neutralize” action neutralises the compounds with hydrogen manipulation on ionisable groups. The “tautomerize” action creates a canonical tautomer form of the molecule and the “aromatize” performs aromatisation based on the general Daylight type aromatisation.

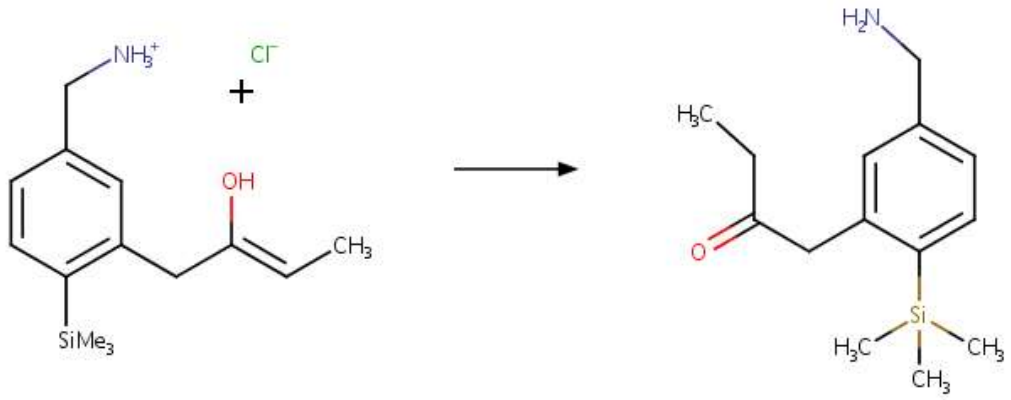

Figure 12: An example of different forms that a chemical can be represented 2016a)

The dominant protonation state of a molecule at pH equal to 7.4 was predicted using the “Major microspecies” ChemAxon node in KNIME. Three sets of descriptors have been used on this work: general physiochemical descriptors, MACCS keys and Peoe-VSA (figure 13). The general physiochemical descriptors have been calculated within KNIME using Infocom/Chemaxon KNIME nodes. The physiochemical descriptors are the following: chiral centre count, rotatable bond count, stereo double bond count, aliphatic/aromatic ring count, fsp3 (fraction of sp3 hybridized carbons), H bond Donor/Acceptor, PSA, LogP, LogD, molecular weight (MW), pKa, electric state, heavy atom count, formal charge, negative/positive ionisable groups and not carbon Heavy Atom (HA) count. The “Topology Analysis” ChemAxon node was used to calculate the: chiral centre count, rotatable bond count, stereo double bond count, aliphatic ring count, aromatic ring count and fsp3. The “Chemical Terms” ChemAxon node was used to calculate the heavy atom count, the formal charge, the not carbon HA count,

the negative and positive ionisable groups. The “Java Snippet (Simple)” node was used to determine the electric state (acid, base or zwitterion) based on the formal charge and the positive/negative ionisable groups. The code used for the configuration of the “Java Snippet (Simple)” node can be found in Appendix, table S1. The “Elemental Analysis” ChemAxon node was used to calculate the MW. The rest of the physiochemical descriptors were calculated by the homonymous ChemAxon nodes as shown in figure 13. The “MACCS keys” node by CCG (Chemical Computing Group) was used to calculate 166 substructure compound descriptors, which account for the frequency of occurrence of 166 chemical features. Finally, the Peoe-VSA descriptors have been calculated using “RD-kit descriptor Calculation” KNIME node.

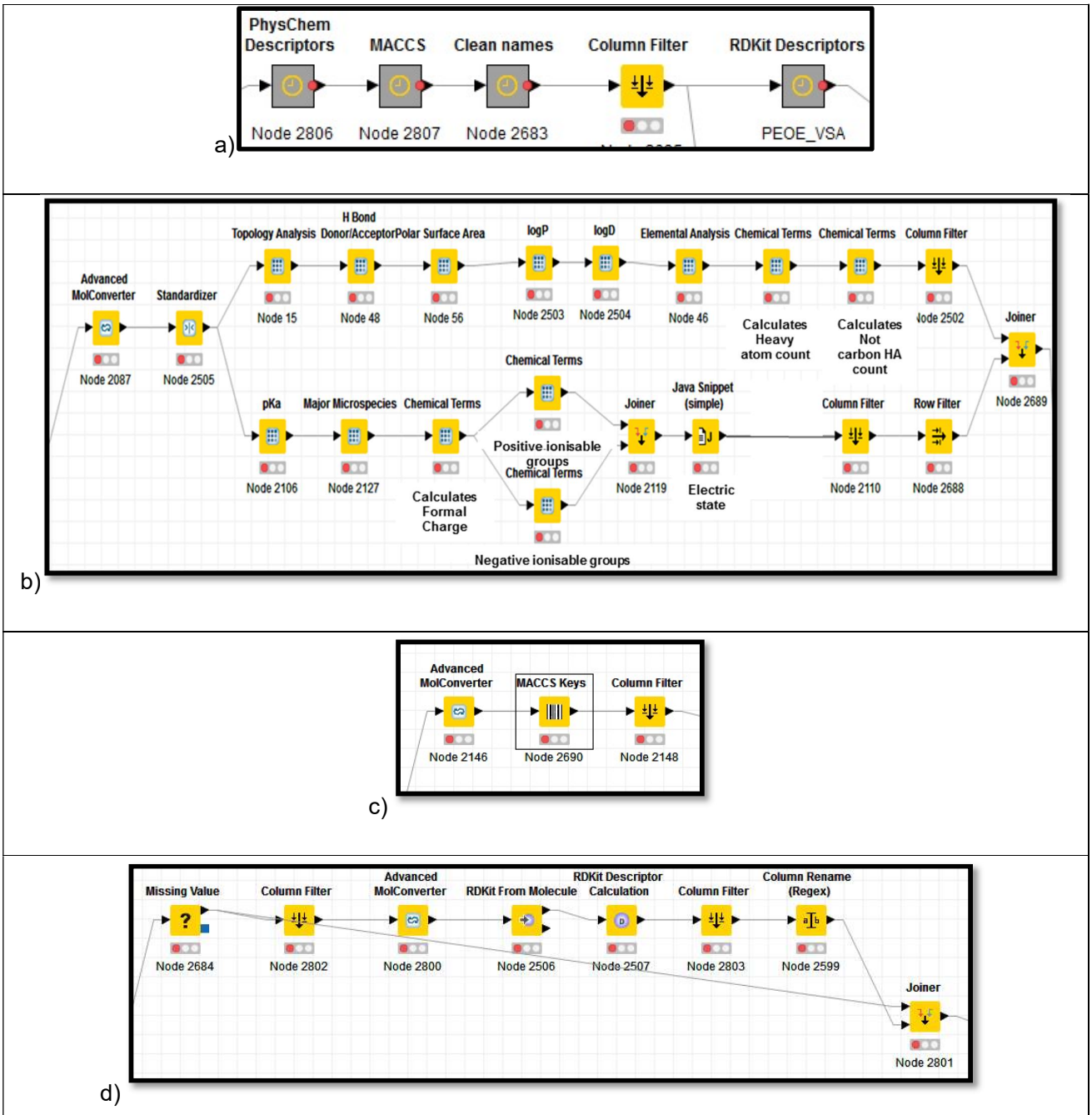

Figure 13: KNIME workflow for the calculation of descriptors: a) overall descriptor calculation workflow, b) physiochemical descriptors, c) MACCS keys and d) Peoe-VSA.

2.2.3 Prediction of Caco-2 permeability of compounds downloaded from ChEMBL by Evotec existing model

The existing Evotec permeability model was written in an R model node and this node was used along with an R predictor node. The R predictor node gave the predicted value of permeability for the ChEMBL compounds. The existing Evotec Caco-2 model uses continuous Random Forest as QSPR algorithm (Appendix, table S2) with 500 trees. In addition, the apparent Caco-2 permeability values were modelled in their logarithmic form (LogPapp).

2.2.4 Model Performance

For performance statistics, the Pearson Correlation coefficient (R^2) and the Root Mean Square Error (RMSE) were reported. The RMSE was calculated with the "RMSE calculator" node (KNIME community node) and the R^2 was calculated by the "2D/3D Scatterplot" node (KNIME community nodes).

2.2.5 Metrics to establish the Applicability Domain

The evaluation of how close are the compounds in the descriptor and chemical space and also what percentage of test compounds are within the AD of the model was carried out using PCA and distance to model metrics. For the evaluation of AD two terms were used to refer to the space that the distance to model metrics were applied. The first term was the descriptor space and referred to the calculation of kNN with Euclidean distance, kNN with Manhattan distance, Mahalanobis distance and leverage with the descriptors and PCs. The second term was the chemical space and referred to the calculation of kNN with Tanimoto and kNN with Dice with the ECFP4 fingerprints. Moreover, the PCA was used to identify if the proprietary training set compounds occupied a similar space as the literature test set compounds. The distance to model metrics were used to estimate the amount of test set compounds within the AD of the proprietary model.

2.2.5.1 Principal Component Analysis and Stopping Rule

The PCA was conducted with the KNIME and the R Snippet, which allows execution of an R script from within KNIME. The data were auto-scaled. The "princomp" function from the R "stats" package was used (Appendix, table S2). The descriptors with zero variance were excluded prior the PC calculation because the "princomp" function in R cannot handle constant (i.e. zero variance) descriptors. The PCs for the Evotec compounds were first established and they were loaded into the R learner node in order to use the same loading matrix for the calculation of the PCs for the CHEMBL compounds. Thus, an R predictor node was used to calculate the PCs for CHEMBL compounds. The results were concatenated, visualised and further analysed with the Scatter Plot (JFree Chart) node in kNIME.

An essential task, within a PCA analysis, is the identification of the significant number of principal components. This would minimize the dimensionality of the dataset and maximize the information retained. The average random method (Avg-R) (Peres-Neto, Jackson, & Somers,

2005) was used. Among a set of 20 tested methods, the Avg-R was particularly efficient in dealing with correlated descriptors. This method for identifying the number of significant principal components consisted of the following steps: 1. randomising the values within variables in the data matrix with R snippet (Appendix, table S2), 2. conducting a PCA in the reshuffled data matrix, 3. calculating the eigenvalues 4. repeating the steps1-3 for 1000 times and 5. Calculating the average eigenvalues. If an observed eigenvalue of a PC is greater than the average eigenvalue, that PC is considered as significant (non-trivial).

2.2.5.2 Evaluation of AD with Distance to model metrics

The distances between the test compounds (literature compounds downloaded from ChEMBL) and the training compounds (Evotec compounds) were considered on two different spaces: 1. descriptors (figure 14) space and 2. chemical space (figure 15). In the descriptor space the distances that were used were the Mahalanobis distance, Leverage and kNN with Euclidean and Manhattan. The descriptors were standardised with the “Normalizer” node by choosing the “Z-score normalisation” setting. Moreover, the distance measurements were performed with the PCs in the descriptor space.

In the descriptor space, Mahalanobis distance was calculated with the R Snippet node in KNIME. The “Mahalanobis distance” function in R from the “stats” package was used (Appendix, table S2). This function returns/calculates the squared Mahalanobis distance. That distance was used for the evaluation of AD and will be simply referred as Mahalanobis distance. The function of Mahalanobis distance cannot handle highly correlated descriptors and descriptors with zero variance because it requires the matrix of the descriptors and compounds to be inverted. As a result, descriptors with correlation greater than 0.85 and descriptors with zero variance were filtered out. A correlation filter KNIME node, which worked in iterations, was used to filter out correlated descriptors. In the first iteration, it identified the descriptor with the most correlations and it kept that descriptor and filtered out the correlated descriptors. Then it continued the iterations until there were no correlated descriptors. The leverage method was conducted using the “Domain Leverage” KNIME node developed by the Novamechanics (Melagraki et al., 2009; Melagraki & Afantitis, 2013) and in this case the zero variance descriptors were also excluded. The kNN with Euclidean and Manhattan distance functions were calculated. The “Similarity search” node by Analytics was used in KNIME for the calculation of the k Nearest Neighbours. For the calculation of the kNN only the descriptors with zero variance were filtered out. In the chemical space, the ECFP4 topological fingerprints (256 bits) were generated from the chemical structure (SMILES) of the compounds with the “ECFP/FCFP” ChemAxon/Infocom KNIME node. Then, the kNN method was used with Tanimoto and Dice coefficients. The “Similarity search” node by Analytics was used in KNIME for the calculation of the k Nearest Neighbours. Different values for the number of Nearest Neighbours (k) were evaluated (k = 1, 3, 5, 10, 20 and 30) and the average distance of each k was calculated. The average distances were compared by calculating correlation coefficient for pairs of average distances for different k values. The values of correlation coefficients and the computational time needed to obtain the Nearest Neighbour (NNs) were used to select the k for further AD evaluations.

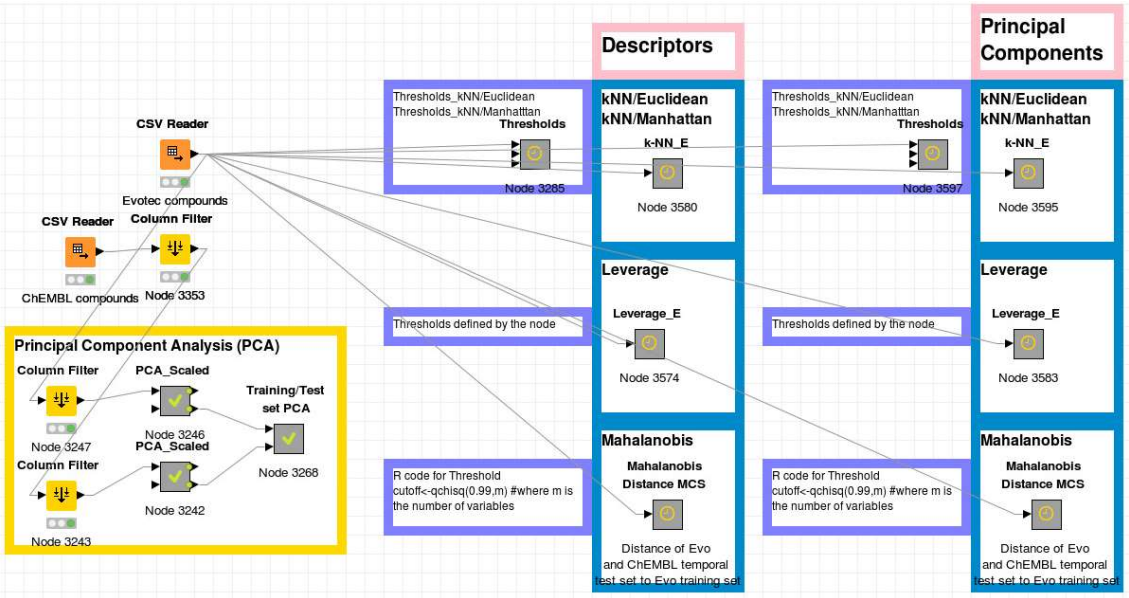

Figure 14: Screenshot of the workflow that was created for the PCA and the estimation of the AD with the four different distance to model metrics in the descriptor space.

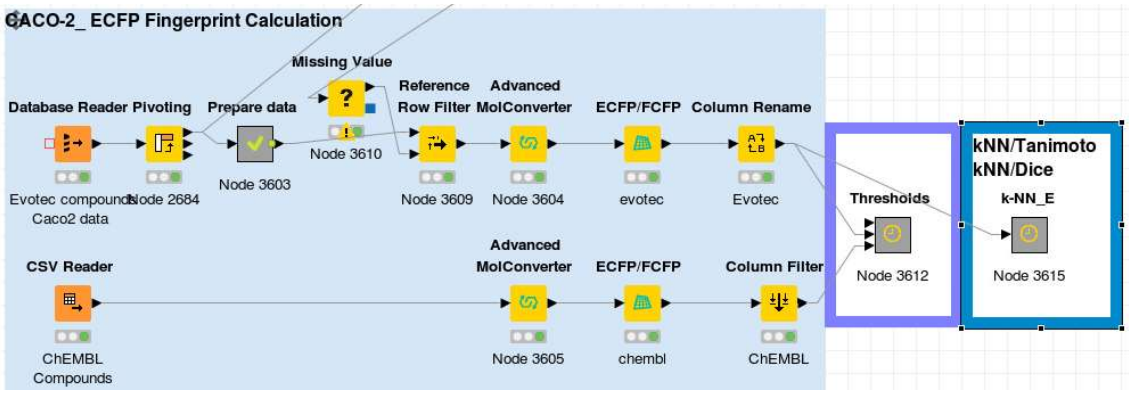

Figure 15: Overview of the workflow that was created for the PCA and the estimation of the AD with the four different distance to model metrics in the chemical space.

2.2.5.3 Distance to model metrics and thresholds

Distance thresholds were applied in the training sets to identify the test set compounds that are within and outside the AD of the Evotec proprietary models. The threshold that was applied for the leverage method is described in section 1.7.1.3. The AD threshold for the k-NN with Euclidean/Manhattan distance was calculated based on the compounds comprising the training set. The kNN distance for each of the training compound was calculated and the 5% of the most remote compounds of the training set were considered to be far from the model and false positives (Mathea et al., 2016). Therefore, the compounds were considered by increasing kNN distance and the threshold was applied at the compound representing the 95%

of the compounds with the smallest distance. The Mahalanobis distance method is explained in 1.7.1.2. However, there isn't a specific literature threshold for the Mahalanobis distance and therefore the following steps were conducted to establish the threshold: 1. estimation of the squared MD of the training set (Evotec compounds), 2. the threshold was set at the 99th quantile based on the training set squared Mahalanobis distance, 3. each ChEMBL compound was added one by one in the Evotec training set and the squared MD was established. The ChEMBL compounds that showed a value greater than the threshold, were considered not to be within the AD of the model. The 99th quantile was calculated in R with the quantile function "qchisq" and the degree of freedom was equal to the number of descriptors or PCs used. Therefore, the 99th quantile was set as the threshold.

2.2.6 Statistical Analysis

The statistical tests Mann-Whitney and Kruskall-Wallis were conducted with the homonymous nodes in KNIME.

2.3 Overview of methods used for the Development of *in-silico* predictive models

This methodology part refers to the development of *in-silico* Caco-2 A to B permeability and $LogD_{7.4}$ predictive models. The objective was to establish if the literature data can be incorporated into the Evotec proprietary models and answer the two following questions: 1. Can literature data improve the performance of proprietary models? and 2. Can literature data enlarge their AD? To provide answer to these two questions a procedure was followed and it is shown in figure 16.

Three types of training sets were used for model development. The first one was the ChEMBL training set (C) with public ADME data extracted from the literature, the second was the Evotec set (E) developed with Evotec proprietary compounds and the third one was the merging of the two previous training sets (E+C). The descriptors were calculated as mentioned in the method section 2.2.2 and then three different machine learning algorithms were applied to each training set: Random Forest (RF), Partial Least Squares (PLS) and Support Vector Regression (SVR) with a radial basis function (rbf) kernel. In this study, the term SVR was used instead of SVM because all the models developed were regression/continuous models. For each algorithm, an optimisation process was performed. In addition, a model assessment was performed with temporal and diverse test sets. The RMSE in prediction and the R^2 were also calculated to measure the model performance. Finally, the AD of the models was established with four different methods, which were outlined in sections 2.2.5.2, 2.2.5.3: a) MD, b) Leverage, c) k-NN/ED and d) k-NN/ManhD.

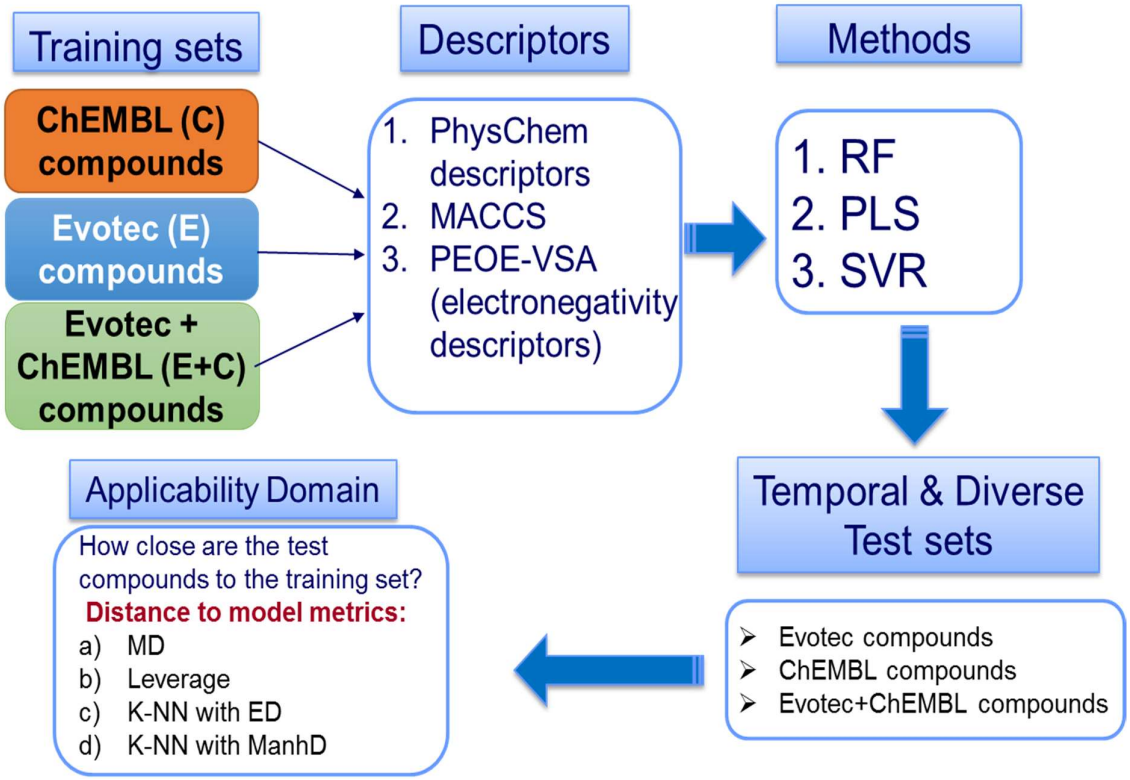

Figure 16: Overview of the methodology process followed for the development of *in-silico* Caco-2 A to B permeability and $LogD_{7.4}$ predictive models.

2.3.1 Literature data curation for the development of *in-silico* Caco-2 permeability and $LogD_{7.4}$ models

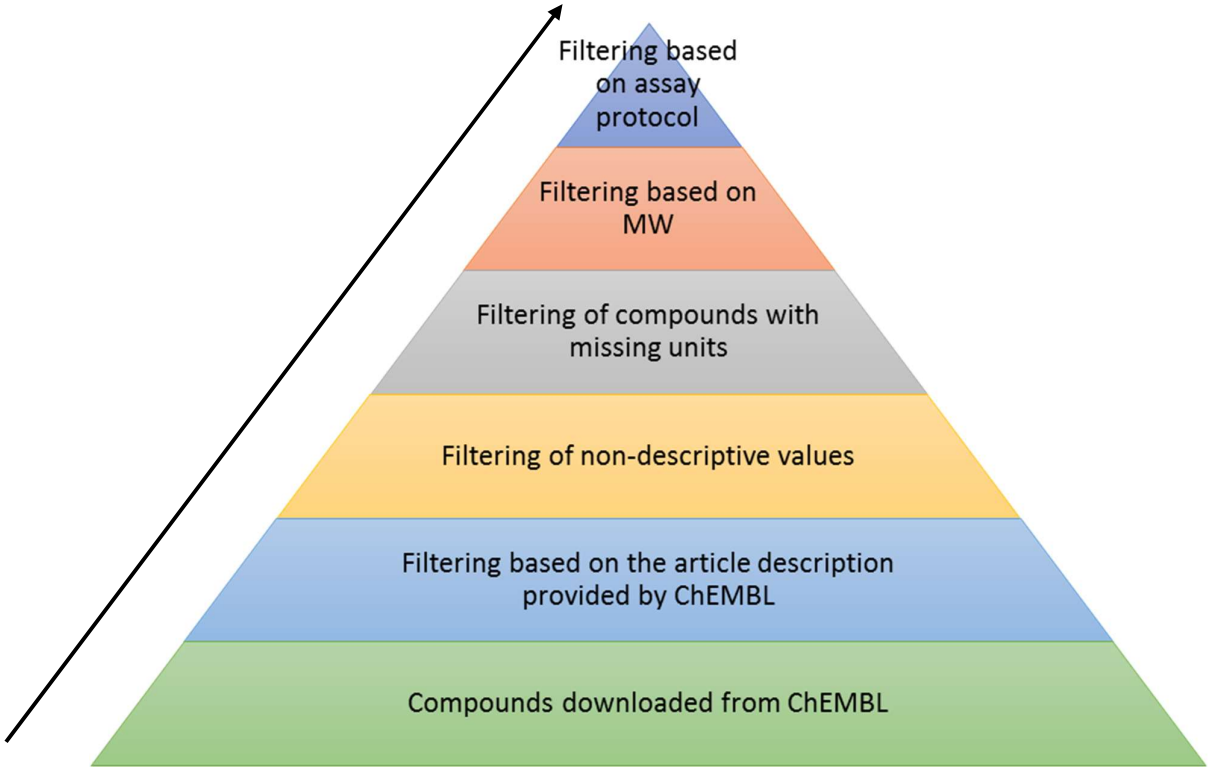

Figure 17: Schematic representation of the literature data filtering process for the compounds downloaded from ChEMBL for the development of *in-silico* Caco-2 permeability and $LogD_{7.4}$ models. The arrow indicates the flow of the process.

The same literature data curation was applied as the literature filtering process detailed in section 2.2.1 with only one difference. The difference was that for the Caco-2 permeability models, the analytical method (used in the experimental procedure) was considered. In Evotec, the Liquid Chromatography/Liquid Chromatography-Mass Spectrometry (LC/LC-MS) is used to analyse the compounds' Caco-2 permeability, whereas in the literature various methods have been reported like: HPLC, UV etc. The majority of the literature sources did not mention the analytical method used. Therefore, all the papers were manually inspected and the data obtained with a different analytical method were excluded. Different analytical methods could give different results and thus, a cross validation of analytical method is necessary to ensure the optimal conditions to accurately reproduce an analytical measurement in different laboratories (Chau, Rixe, McLeod, & Figg, 2008).

For the development of the $LogD_{7.4}$ models, 4083 compounds with $LogD_{7.4}$ information were downloaded from ChEMBL v. 22. Data were first curated based on the article description provided by ChEMBL. The compounds measured in a pH other than 7.4 and with a solvent other than octanol were excluded. The rest of the filtering process was identical to the process described above and a general workflow of filtering is shown in figure 17. The literature data

that were used as temporal tests set for the model assessment were also downloaded from ChEMBL and the same procedure was applied.

2.3.2 Selection of training and test sets

Three different types of training sets were used to develop models and more details about the source and number of compounds are shown in tables 4-7. The first type of training set was developed with only literature compounds downloaded from the ChEMBL database ("ChEMBL training set") and the models developed with that training set are the "ChEMBL models". The second type was developed with only Evotec proprietary compounds extracted from the Evotec database ("Evotec training set") and the models developed with that training set are the "Evotec models". Finally, the third type was developed with merged proprietary Evotec and literature compounds ("Evotec+ChEMBL training set") and the models developed with that training set are the "Evotec+ChEMBL models".

In addition, a temporal and a diverse test set were used to evaluate the goodness of the models. The temporal test set included compounds added to Evotec and ChEMBL databases after the initial training sets were created. The diverse test set was formed by randomly selecting 20% of compounds from the merged initial and temporal datasets. The rest 80% of the merged compounds were used as the training set to build the models.

Table 4: Training and Temporal test sets used in development of the *in-silico* permeability models.

Training set	Details	Number of compounds
Evotec	Compounds until the 31/08/2016	2075
ChEMBL	Compounds downloaded from ChEMBL v21 (8/08/2016)	1628
Evotec+ChEMBL	Compounds were merged	3703
Test set	**Details**	**Number of compounds**
Evotec temporal	Compounds from 1/09/2016 until 18/01/2017	166
ChEMBL temporal	Compounds downloaded from ChEMBL v22 (18/01/2017)	92
Evotec+ChEMBL temporal	Evotec temporal and ChEMBL temporal test sets were merged	258

Table 5: Training and Diverse test sets used in development of the *in-silico* permeability models.

Training set	Details	Number of compounds
Evotec	80% of the Evotec training set (randomly selected) until 18/01/2017	1660
ChEMBL	80% of the ChEMBL training set (randomly selected) from ChEMBL v22 (18/01/2017)	1302
Evotec+ChEMBL	Compounds were merged	2962
Test set	**Details**	**Number of compounds**
Evotec diverse	20% of the Evotec training set (randomly selected) until 18/01/2017	415
ChEMBL diverse	20% of the ChEMBL training set (randomly selected) from ChEMBL v22 (18/01/2017)	326
Evotec+ChEMBL temporal	Evotec temporal and ChEMBL temporal test sets were merged	741

Table 6: Training and Temporal test sets used in development of the *in-silico* $LogD_{7.4}$ models.

Training set	Details	Number of compounds
Evotec	Compounds until the 31/12/2016	8400
ChEMBL	Compounds downloaded from ChEMBL v22 (2/05/2017)	1209
Evotec+ChEMBL	Compounds were merged	9609
Test set	**Details**	**Number of compounds**
Evotec temporal	Compounds from 1/01/2017 until 2/05/2017	895
ChEMBL temporal	Compounds downloaded from ChEMBL v23 (19/05/2017)	86
Evotec+ChEMBL temporal	Compounds were merged	981

Table 7: Training and Diverse test sets used in development of the *in-silico* $LogD_{7.4}$ models.

Training set	Details	Number of compounds
Evotec	80% of the Evotec training set (randomly selected) until 2/05/2017	7436
ChEMBL	80% of the ChEMBL training set (randomly selected) from ChEMBL v23 (19/05/2017)	1036
Evotec+ChEMBL	Compounds were merged	8472
Test set	**Details**	**Number of compounds**
Evotec diverse	20% of the Evotec training set (randomly selected) until 2/05/2017	1859
ChEMBL diverse	20% of the ChEMBL training set (randomly selected) from ChEMBL v23 (19/05/2017)	259
Evotec+ChEMBL diverse	Compounds were merged	2118

2.3.2.1 Subsequent model assessment for Caco-2 permeability models

In a subsequent model assessment of the Caco-2 permeability models, the permeability data in the temporal test sets were merged with the training test sets and used, all together, to develop an updated model (M2). Two new temporal test sets were generated including the latest proprietary permeability data (Evotec compounds synthesised four months after the compounds in the training set) and the freshly published public permeability data from ChEMBL version 23. These new temporal test sets were referred as “New Evotec temporal test set” and “New ChEMBL temporal test sets” and represented the compounds published in the literature and synthesised in Evotec four months after the initial models (M1). The new temporal test sets were used to assess both the initial models (M1) and the new models (M2).

Table 8: Training and temporal test sets new temporal test sets used in the subsequent models assessment for the *in-silico* permeability models.

Training set (M1)	**Details**	**Number of compounds**
Evotec	Compounds until the 31/08/2016	2075
ChEMBL	Compounds downloaded from ChEMBL v21 (8/08/2016)	1628
Evotec+ChEMBL	Compounds were merged	3703
Training set (M2)	**Details**	**Number of compounds**
Evotec	Compounds until the 31/12/2016	2241
ChEMBL	Compounds downloaded from ChEMBL v22 (18/01/2017)	1720
Evotec+ChEMBL	Compounds were merged	3961
New temporal Test sets	**Details**	**Number of compounds**
Evotec temporal	Compounds from 19/01/2017 until 20/05/2017	245
ChEMBL temporal	Compounds downloaded from ChEMBL v23 (19/05/2017)	115

2.3.3 Standardisation of Molecular descriptors

A molecular descriptor standardisation process was applied as described in method section 2.2.2, using the "Normalizer" and the "Normalizer (Apply)" node in KNIME. The "Normalizer" node was used for the training set and the "Normalizer (Apply)" node for the test set compounds in order to "Standardise" them in the same range as the training set. In this way, each descriptor's values had a mean of 0 and standard deviation of 1. For example, for a dataset with m-rows and if each row contains n- different descriptors/variables, the x row for the i^{th} descriptor will be standardised with the following equation:

$$Standardised(x_i) = \frac{x_i - x_{mean}}{std(x)} \quad \text{(Equation 26),}$$

where x_i is the value of the i^{th} descriptor in row x, the x_{mean} is the mean of the x row values and std(x) is the standard deviation of the values in x row.

The std (x) and the x_{mean} were calculated with the following equations:

$$std(x) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - x_{mean})^2} \quad \text{(Equation 27),}$$

where std(x) is the standard deviation of the values in x row, the n is the number of the descriptors, x_i is the value of the i^{th} descriptor in row x and the x_{mean} is the mean of the x row values.

and

$$x_{mean} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(Equation 28),}$$

where the x_i is the value of the i^{th} descriptor in row x, x_{mean} is the mean of the x row values and n is the number of the descriptors.

2.3.4 Algorithms and their parameter optimisation for model building

In this study, three algorithms were used in model building: Random Forest (RF), Partial Least Squares (PLS) and Support Vector Regression (SVR).

2.3.4.1 Random Forest (RF) parameter selection

Random forest is based on an ensemble of decision trees (Mitchell, 2014; K. Roy et al., 2015), which are built by training data of multiple feature. The Caco-2 permeability and $LogD_{7.4}$ predictions, which were continuous variables, were provided as the average of the predictions of all the trees. Therefore, the key parameter was the number of trees (ntree). A series of 5-

fold cross-validations was performed with different number of trees (10, 20, 50, 100, 500 & 1000). It was found that 500 provided an optimal setting with a good balance between computational time and the error in the prediction. The work was performed with the "RandomForest" R package in the "R Learner" and "R predictor" node in KNIME and the "randomForest" function used to develop the model. The 200 descriptors described in section 2.2.2 were used because RF can handle both correlated and low variance descriptors.

2.3.4.2 Partial Least Squares (PLS) parameter selection

PLS is another algorithm that was used for the model development and is able to project the original variables (i.e. descriptors) into latent variables and thus reducing the dimensionality (Xing et al., 2014). This method decomposes the input matrix of descriptors into loadings and scores, and the latter are orthogonal and are capturing the descriptor information (Sethi, 2012). In this case, it was essential to choose the appropriate number of components. The dataset was shuffled 100 times and 100 PLS models were developed and assessed with a 5-fold cross validation and with maximum of 40 components. Then the mean RMSE was calculated for each component and the highest performing model was the one with the lowest mean RMSE. Then the fewest number of components that were still less than one standard error away with 95% confidence from the overall best model were chosen for the model building. The work was performed in the R learner and R predictor node with the PLS package. Descriptors were standardised with the "Normalizer" node for training set and the "Normalizer (Apply)" node for the test set. The option "Z normalisation" was applied.

2.3.4.3 Support Vector Regression (SVR) parameter selection

SVR originates from the Vapnik's structural Risk Minimisation principle for statistical theory. In this case, the radial basis function (rbf) was used as a kernel and there were three parameters to optimise: 1. epsilon (ε), 2. cost (C) and 3. gamma (γ). The goals was to tune these parameters so that the model could accurately predict the new data. The optimisation was performed with an exhaustive grid search and a 5-fold cross validation, using the tune function in the e1071 in R, to identify the optimal area and then perform a narrower grid search in that area. The grid search looks at different parameters' values and returns the best parameters to train the dataset (Chang & Lin, 2011). In addition, a 10-fold cross validation is usually used but in this case a 5-fold was selected due to the computation time. A training set of about 2000 compounds and 170 descriptors needed about 4-5 days to train, whereas a set of about 10,000 compounds more than 2 weeks. The search was carried out in the following ranges: 1. ε values from 0 to 1, 2. C values from 1- 1500 and 3. γ values from 0 to 1. After the calculation of those three parameters with the tune function from the e1071 R package, the LibSVM weka node was used in KNIME to train the models (Chang & Lin, 2011).

2.3.5 Estimation of the AD of the *in-silico* Caco-2 permeability and $LogD_{7.4}$ models with distance to model metrics

Mahalanobis Distance, Leverage, kNN with Euclidean and kNN with Manhattan were used in the descriptor space to calculate how close are the "ChEMBL temporal test set compounds" and "Evotec temporal test set compounds" from the training set of "ChEMBL models", "Evotec models" and Evotec+ChEMBL models". Therefore, the same methodology and thresholds were also applied in this part of the work as described in method sections 2.2.5.2, 2.2.5.3

The only difference is that in this part of the work the distance of 2 different test sets from 3 different training sets was calculated (figure 18) and for that reason the KNIME workflow was amended to perform these calculations. A screenshot of the workflow is shown in the Appendix (figure S1-S4).

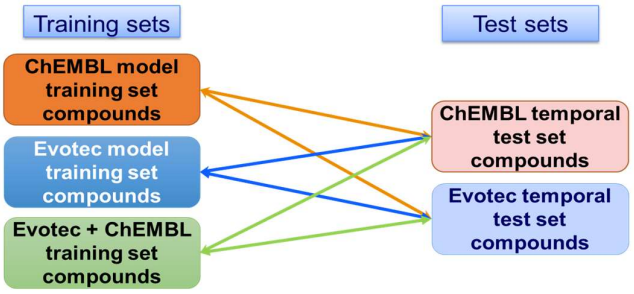

Figure 18: Schematic representation of the distances of the test set compounds from the training sets. The arrows indicate the distances that were calculated.

3 RESULTS AND DISCUSSION

3.1 Evaluation of existing Evotec Caco-2 A to B permeability model with opensource data

3.1.1 Model Assessment

The goal of this part of the work was to evaluate the performance of the existing Evotec proprietary Caco-2 permeability model on a dataset of 1770 literature compounds. This gives an indication of how well the proprietary model can predict literature compounds. The existing Caco-2 permeability model that was evaluated has been developed with Evotec proprietary compounds as the training set. The opensource test compounds (with experimental Caco2 permeability measurements) were extracted from the ChEMBL database. The ChEMBL compounds were first curated as described in method section 2.3.1. The Evotec permeability model estimates the apparent A to B Caco-2 permeability expressed as 10^{-6} cm/s; however, for the statistical computation a Log_{10} of that permeability was used. The R^2 and RMSE in prediction of the ChEMBL compounds were equal to 0.22 and 0.704 respectively. In addition, the RMSE of the Evotec training set was equal to 0.2 and the RMSE of a temporal Evotec set (assessed by the same model in the company in July 2016) was 0.42. As a result, the existing Evotec Caco-2 permeability model is not accurately predicting the A to B permeability of the ChEMBL compounds (figure 19) as accurately as it can predict the proprietary compounds.

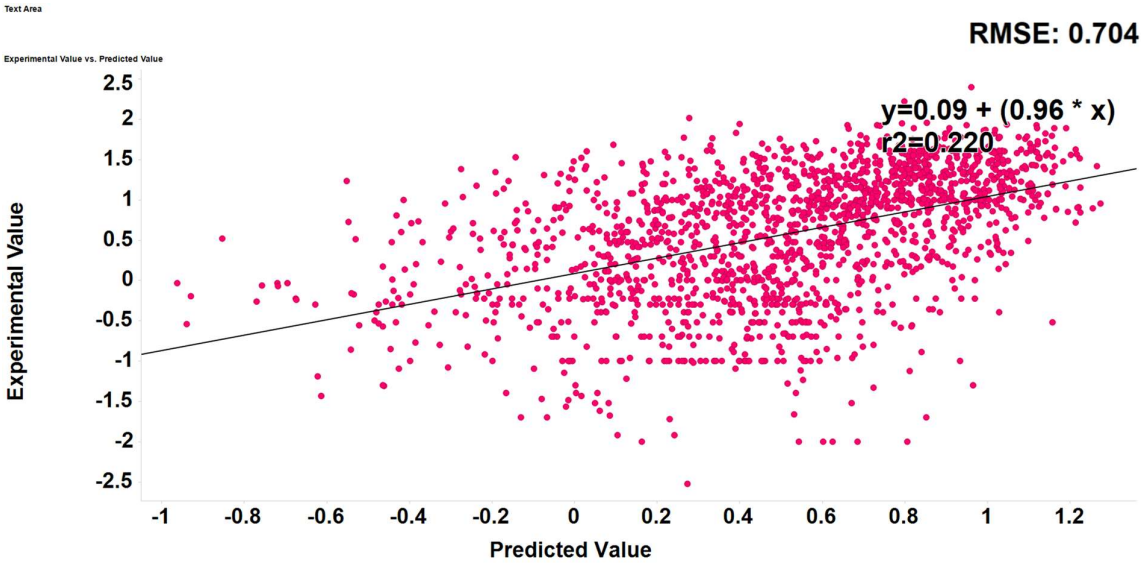

Figure 19: Experimental values for Caco-2 permeability of ChEMBL compounds vs the predicted Caco-2 permeability obtained with Evotec Caco-2 model.

A similar research was conducted by AstraZeneca (Bruneau, 2001), where they evaluated the existing proprietary solubility model with a temporal AstraZeneca solubility test set and a literature test set. The results of that study also suggested that the literature test set was not predicted as accurately as the proprietary temporal test set. The literature test set RMSE was 1.88 and the RMSE of the proprietary temporal test set was 0.78. However, these results cannot be directly compared to the Evotec results as they refer to a different model, different ADME property, different compounds in training and test sets. In another study, a literature test set and a proprietary test set were used to evaluate the model performance of a $LogD_{7.4}$ proprietary model of Bayer Shering Pharma AG (Schroeter et al., 2007). Results indicated that the model was better in predicting proprietary test set (RMSE=0.41) compared to the literature test set (RMSE= 0.66). The results from these two studies based on a solubility and a $LogD_{7.4}$ predictive model gave the same overall conclusion about the less accurate prediction of literature compounds compared to proprietary temporal test sets from proprietary models. A possible reason can be that the chemical space of the literature test set may be different from the proprietary chemical space of the training set. Therefore, a PCA analysis was conducted to compare the descriptor space covered by the public compounds from ChEMBL with that covered by the proprietary Evotec compounds used in the model building and training.

3.1.2 Principal Component Analysis

PCA has been previously used to identify the overlap of the molecular data in the descriptor space (Gavaghan, Arnby, & Blomberg, 2007), and has also been used to investigate if the distribution of training and test set are balanced and representative of the chemical domain (Roy, Kovarich, & Gramatica, 2011). Therefore, 2-dimensional PCA (2DPCA) was used to project the ChEMBL compounds into the molecular descriptors space of training compounds (Evotec compounds) in order to establish their similarity.

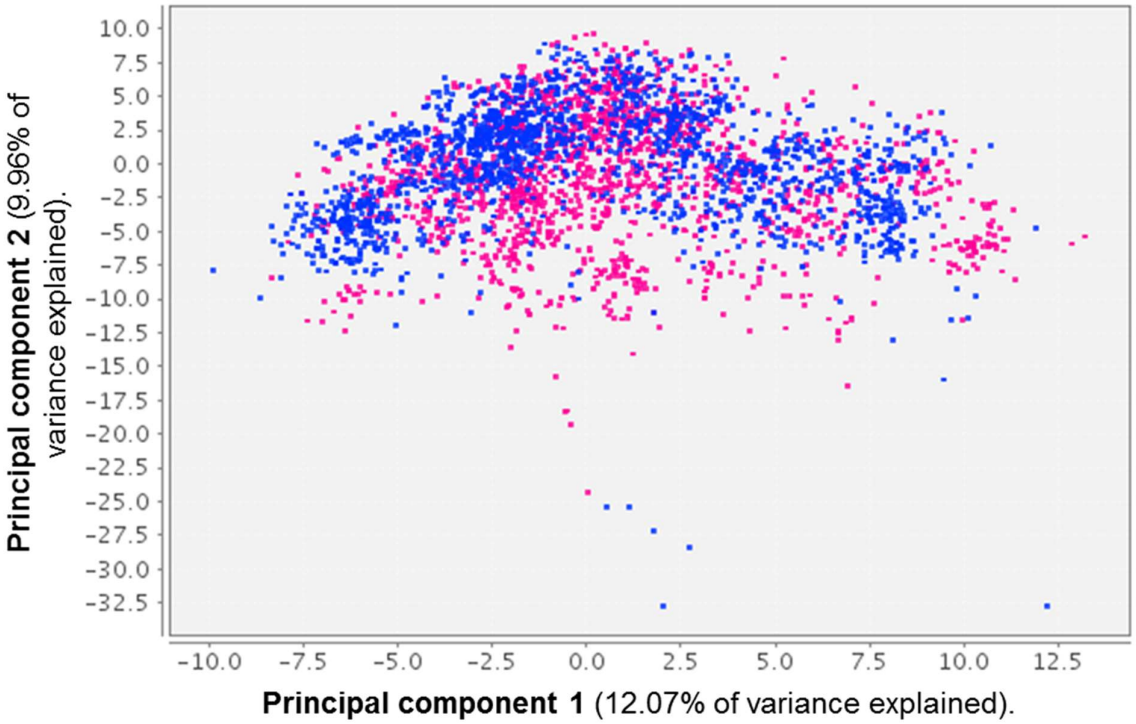

Figure 20: Principal component plot of Principal Component 1 vs Principal Component 2 for Evotec (blue) and ChEMBL (red) compounds. Figures in brackets indicate the percentage of variance explained by the corresponding PC.

The PC score plot of two sets of compounds appear to be closely related in the descriptor space (figure 20), because a significant number of CHEMBL compounds is projected on to the space covered by the Evotec compounds. However, the first two PCs visualised in this plot only account for 22.03% of the variance. Therefore, a 2D PCA analysis would be insufficient to answer the question about similarity of the two sets of compounds. Thus, the evaluation of different distance to model methods were used to estimate the “distance” of each compound in the test set from the training set. The PCA alone was not able to give specific results for each test compound but it gave an overall picture of the distribution of the two sets. Thus, PCA has been used in the literature along with other methods to establish the AD. Some examples are the use of PCA with kNN (Kaneko & Funatsu, 2014), Hotelling-T test (Venkatapathy & Wang, 2013) and Mahalanobis distance (De Maesschalck, Jouan-Rimbaud, & Massart, 2000). In more detail, a number of PCs were used instead of the descriptors.

PCA is also combined with methods (like Mahalanobis distance) that require the matrix to be inverted and cannot handle correlated descriptors or descriptors with zero variance. The multicollinearity and zero variance issues can be overcome by using PCs instead of descriptors (De Maesschalck et al., 2000; Jaworska et al., 2005). However, the problem is how many PCs are significant and should be used. In that case, a stopping rule should be applied, which will reduce the information loss (underestimation) and the noise inclusion (overestimation). It was identified that one of the most efficient stopping rule was the average random (Avg-R),

particularly efficient in dealing with correlated descriptors (Peres-Neto et al., 2005). It was found that for the Evotec compounds the first Avg eigenvalue higher than the eigenvalue is the 28th (figure 21). Therefore, the first 27 PCs were retained for subsequent analysis.

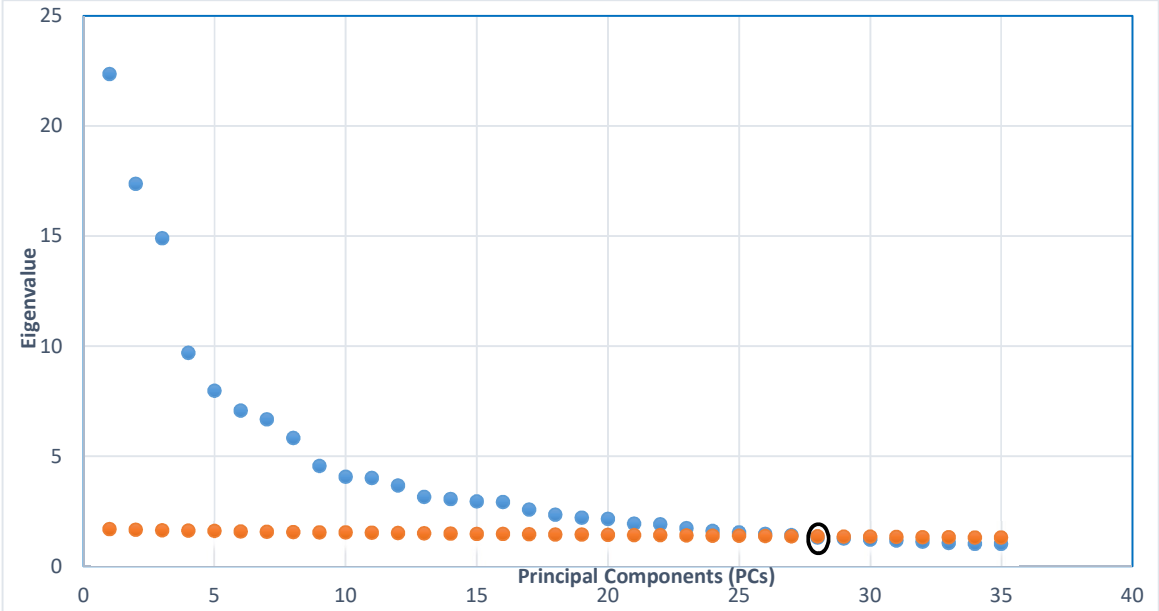

Figure 21: Scree plot of the eigenvalues from the Evotec compounds PCA (blue) and the eigenvalues obtained from the Avg-R on Evotec compounds PCA (orange).

3.1.3 Evaluation of distance to model metrics

PCA as a single method was not an efficient method to estimate the AD and other methods applied. Different distance to model metrics were evaluated for their ability to correlate the distance of the test compounds from the training set with the accuracy of the prediction. A distance to model metric, which shows such correlation, could in theory be used as a method to provide a confidence interval for a prediction. The distances between the test compounds (ChEMBL compounds) and the training compounds (Evotec compounds) were considered on two different spaces: 1. descriptors space and 2. chemical space. In the descriptor space, the distance to model metrics that were used are the kNN with Euclidean distance, kNN with Manhattan distance, Mahalanobis distance and Leverage. The distance measurements were performed with the standardised descriptors and the first 27 PCs. In the chemical space, the kNN method was used with Tanimoto and Dice coefficients. In that case, the ECFP4 fingerprints were used to calculate the Dice and Tanimoto coefficients. Moreover, for the kNN method, for both descriptor and chemical space different values for number of nearest neighbours (k) were evaluated (1, 3, 5, 10, 20 and 30) and the average distance for each k was calculated. By altering the k when computing the average distance in descriptor and chemical space has only a minimal effect on the overall value and this is shown in correlation tables that can be found in appendix (tables S3-S8). Therefore, the k=5 was used as the

number of nearest neighbours to consider, as it provided a good compromise between execution time and robustness. The same k selection process was conducted in another study (Weaver & Gleeson, 2008), where the k=5 was a good compromise. Table 9 summarises the distance to model metrics used. For all the methods shown in table 9, the compounds were: a) binned by distance and b) binned by squared residuals for further data analysis.

Table 9: Summary of the Distance to model metrics.

Distance to model metrics								
Descriptors space	**Mahalanobis Distance**		**Leverage**		**kNN**			
					Euclidean		**Manhattan**	
	Descriptors	PCs	Descriptors	PCs	Descriptors	PCs	Descriptors	PCs
Chemical space (ECFP4 fingerprints)	**kNN**							
	Tanimoto				Dice			

3.1.3.1 Bin compounds by distance

The distance to model metrics were used to evaluate the presence of a relationship between the error in the predictions (RMSE) and the calculated distance to model. Compounds in the test set were binned in 5 equally populated bins with increasing distance, and for every bin the average error in the prediction was calculated as RMSE. It was observed that there was a trend between the RMSE and the distance, especially for the first 3-4 bins (figures 22, 23). This trend indicated that as the distance of the test compounds from the training set increases, the RMSE in prediction increases too. Interestingly, this trend was observed for all the combinations of metrics/methods used (figures 22, 23). If this scenario is genuine all the distance metrics investigated could possibly be used to estimate the distance of a compound to the model (or better to the training compounds) and consequently an estimation of the expected error in the prediction could be argued. To better understand this a statistical analysis has been carried out.

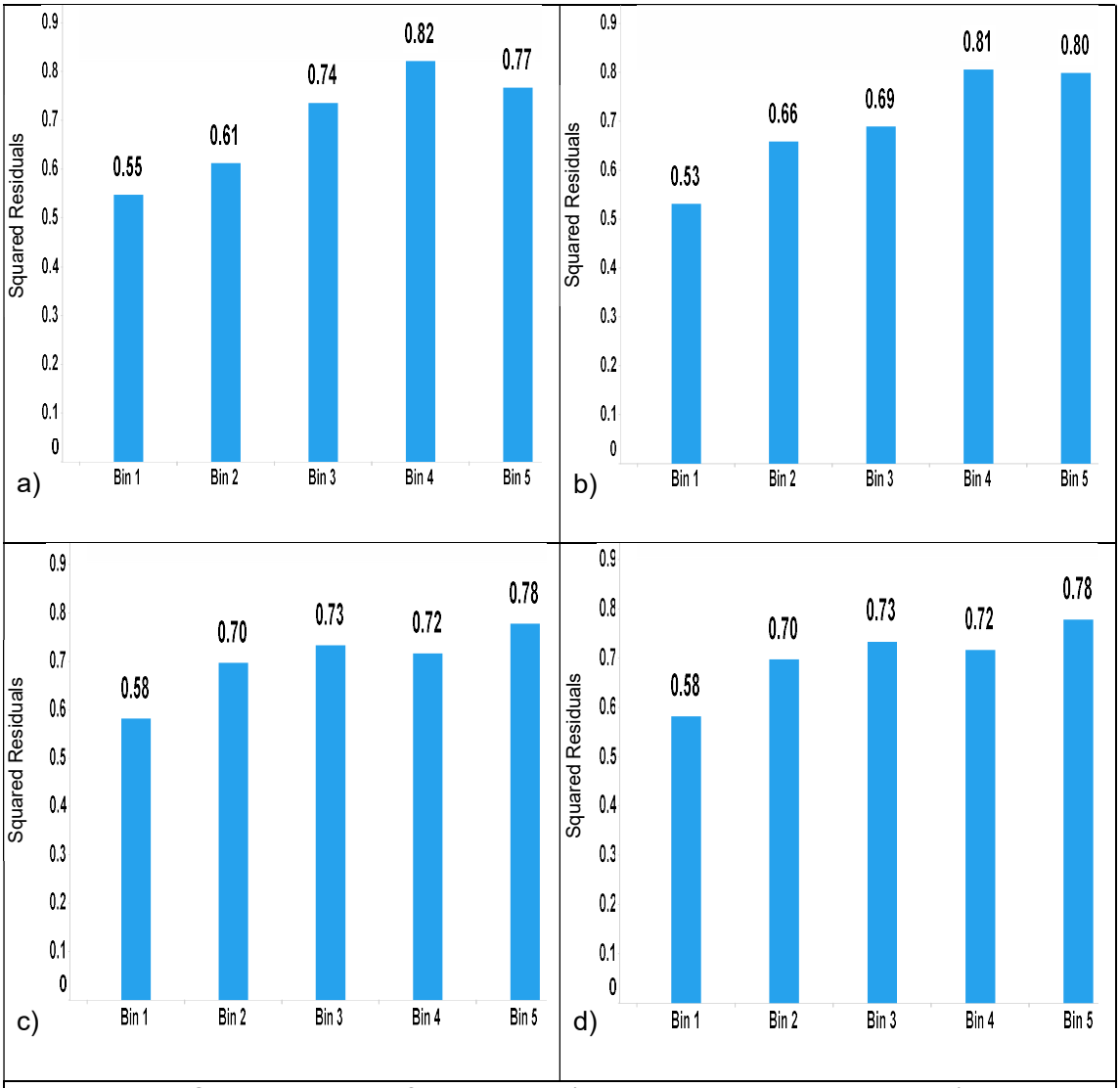

Figure 22: RMSE in prediction of the binned a) Euclidean distance to 5NNs, b) Manhattan distance to 5NNs, c) Leverages and d) Mahalanobis Distance for CHEMBL compounds calculated with the descriptors.

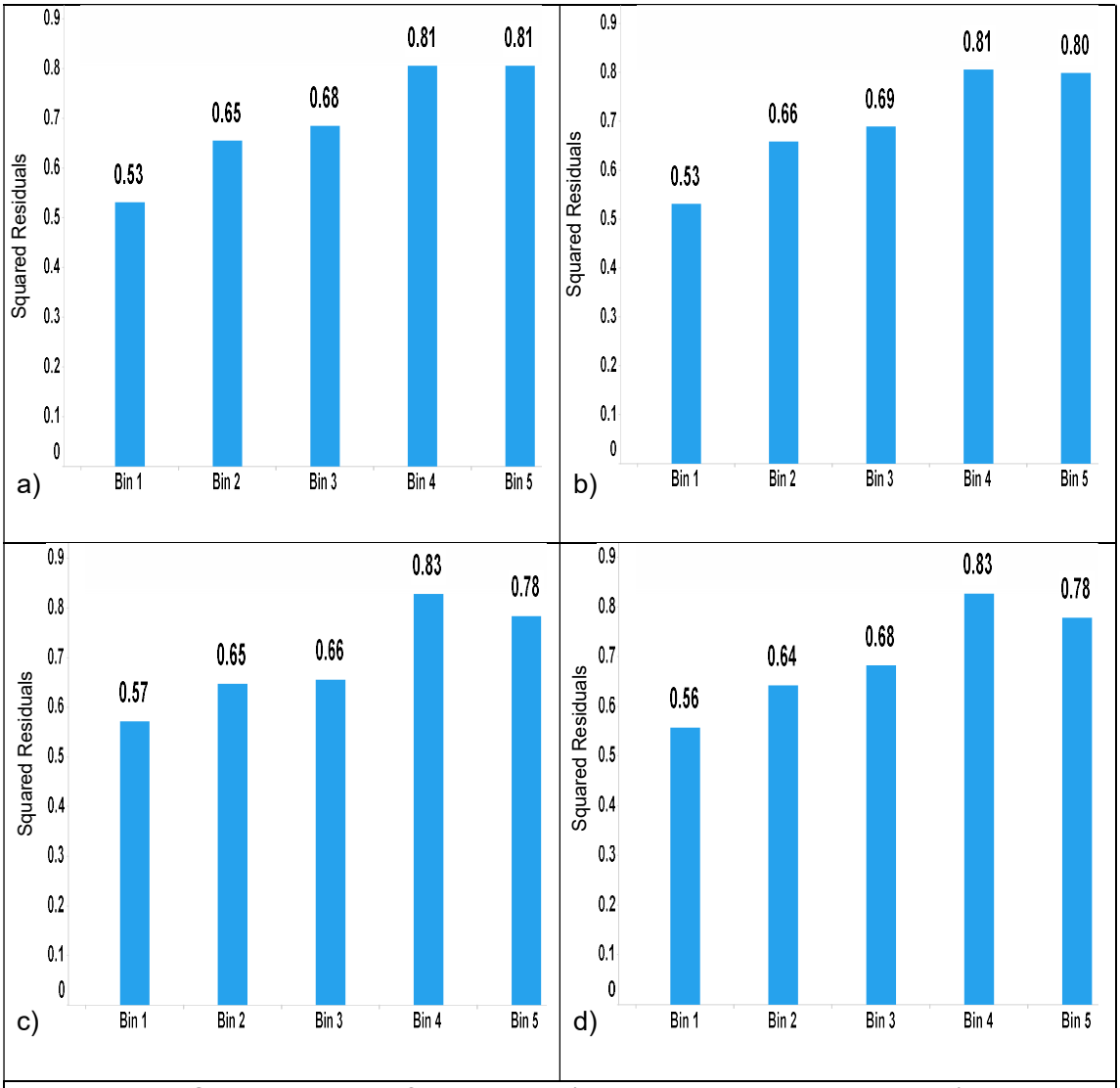

Figure 23: RMSE in prediction of the binned a) Euclidean distance to 5NNs, b) Manhattan distance to 5NNs, c) Leverages and d) Mahalanobis Distance for CHEMBL compounds calculated with the first 27 PCs.

Two different statistical tests, the Mann-Whitney test and the Kruskal-Wallis test, which is an extension of Mann-Whitney when more than 2 means are compared, were employed to determine if there is a statistically significant difference between the equally populated bins. The Kruskal-Wallis test has also been used for a similar statistical analysis by Weaver and Gleeson (2008), where they were comparing the binned distance of 5 equally populated bins with their RMSE. The authors of that work established a statistically significant overall difference between the 5 bins. Similarly, the results of Kruskal-Wallis showed that there is a statistically significant difference, whereas Mann-Whitney test did not show statistically significant difference between all the bins (table 10). The possible reason why Mann Whitney test did not always show a statistically significant difference is that compounds that are assigned in two subsequent bins might be in a similar distance from the model. However, it is not necessary that there should always be a significant difference because that depends on the compounds in the test set. For example, if the compounds in the test set are close, there will not be a significant trend between the RMSE and the distance and *vice versa*. As a result, the presence of a trend in the data depends on the compounds used as a test set. In that case, there is a weak (qualitative) trend. According to Davis and Ward (2014) the distance to model measures usually produce a weak relationship to error in prediction and as a result this is limiting the confidence that can be extracted from the statistic. In addition, a similar trend was observed in the AD investigation of $LogD_{7.4}$ models (Schroeter et al., 2007), where the error in prediction increased as the distance of the compounds in the equally populated bins increased. However, a statistical analysis has not been conducted to identify if the difference in the error in prediction within bins is statistically significant. Furthermore, in another study the AD of lipophilicity ($LogD_{7.4}$) models was evaluated (Bruneau & McElroy, 2006). A trend was observed between the error in prediction and the Mahalanobis distance of the test compounds from the training set. The trend indicated that as the distance of the test compounds increases, the error in prediction increases as well. Therefore, the findings form this study and the literature indicate that there is always a relation between the distance of the compounds and the error in the prediction.

Table 10 showed that there was not much difference in RMSE between bin 4 and 5 for most of the methods/distances as the last bin, in most of the cases, showed a lower RMSE than the previous bin (figures 22, 23). This unexpected behaviour could be explained considering that compounds in bin 4 and 5 are largely far from the model thus making the prediction unreliable. This translates into random fluctuation of the prediction error thus clearing the trend observed at smaller distances. Another possible reason is that the model could possibly extrapolate correctly outside the domain (Jaworska et al., 2005) and thus a smaller RMSE could be observed in bins 4 and 5.

Table 10: Statistical analysis of the RMSE of the bins (data are binned by distance).

	Mann-Whitney Test				Kruskal-Wallis Test
Method	Bin1_Bin2	Bin2_Bin3	Bin3_Bin4	Bin4_Bin5	Bin 1 – Bin 5
Euclidean/ Descriptors	p<0.05	p>0.05	p<0.05	p>0.05	p<0.05
Euclidean/ PCs	p<0.05	p<0.05	p<0.05	p>0.05	p<0.05
Manhattan/ Descriptors	p<0.05	p>0.05	p<0.05	p>0.05	p<0.05
Manhattan/ PCs	p>0.05	p<0.05	p<0.05	p>0.05	p<0.05
Leverage/ Descriptors	p>0.05	p<0.05	p>0.05	p>0.05	p<0.05
Leverage/ PCs	p<0.05	p>0.05	p>0.05	p>0.05	p<0.05
Mahalanobis/ Descriptors	p<0.05	p>0.05	p>0.05	p>0.05	p<0.05
Mahalanobis PCs	p>0.05	p>0.05	p<0.05	p>0.05	p<0.05

3.1.3.2 Bin compounds by squared residuals

To further asses the AD, the test compounds were binned in 5 equally populated bins by increasing squared residuals and the average distance of each bin was calculated and is shown in figures 24, 25. The bar charts did not show any trend between the distance and the prediction error. In addition, as it has been conducted previously a Mann Whitney test was applied to evaluate if there is a statistically significant difference between the bins. The figures 24, 25 indicated the absence of a trend between bins 1-4 and this was confirmed by the Mann-Whitney test (table 11). The only significant difference in the average distance was observed and confirmed by the Mann Whitney test between bin 4 and bin 5. This is an indication that compounds with the larger RMSE, which are allocated in bin 5, seem to have the greatest distance from the model. This trend between bin 4 and bin 5 is something that it was expected. However, the absence of trend between bin 1 and bin 4 can have two possible interpretations.

One possible explanation could be that a molecule might be permeable due to a property that the model cannot consider and consequently the model cannot produce accurate predictions for these compounds. The second reason is that the compounds, which are far from the chemical space of the model might not be assigned with reliable predictions.

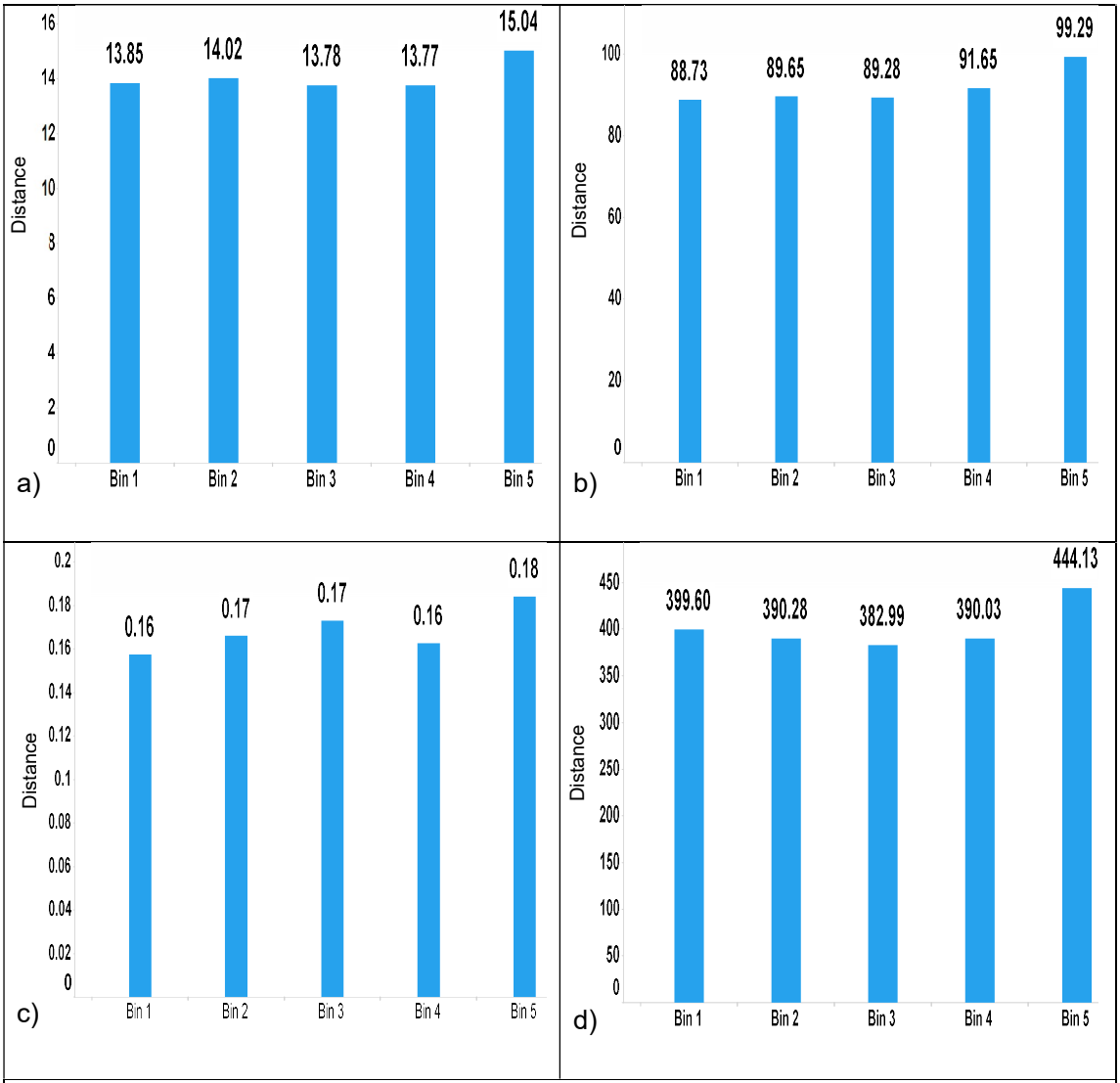

Figure 24: Average a) Euclidean distance to 5NNs, b) Manhattan distance to 5NNs, Leverages and d) Mahalanobis Distance of the binned squared residuals for CHEMBL compounds calculated with the descriptors.

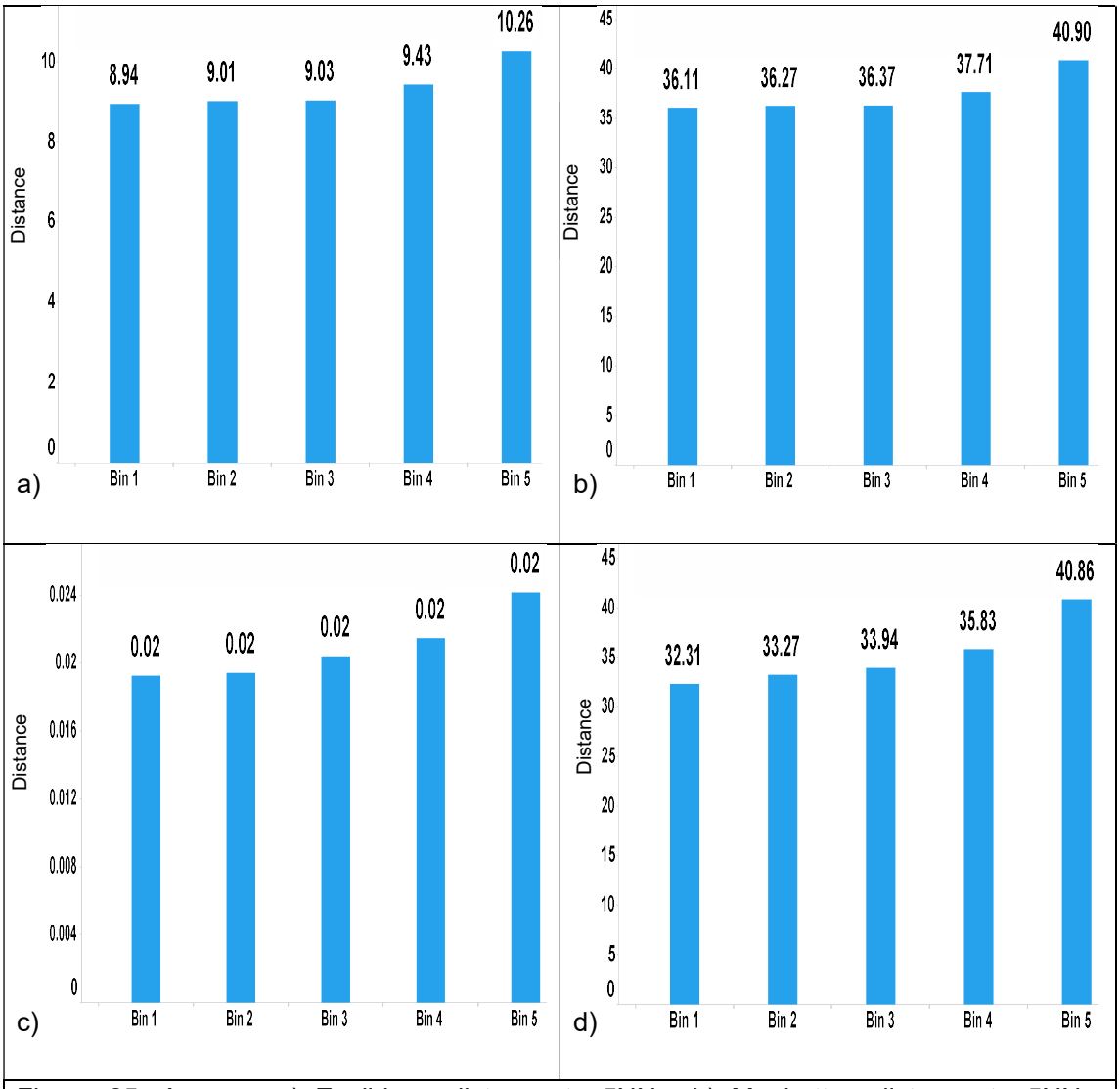

Figure 25: Average a) Euclidean distance to 5NNs, b) Manhattan distance to 5NNs, Leverages and d) Mahalanobis Distance of the binned squared residuals for CHEMBL compounds calculated with the 27 first PCs.

Table 11: Statistical analysis of the average distance of the bins (data are binned by squared residuals).

Mann-Whitney Test				
Method	Bin1_Bin2	Bin2_Bin3	Bin3_Bin4	Bin4_Bin5
kNN/ Euclidean	p = 0.838	p = 0.750	p = 0.924	p <0.001
kNN/ Euclidean PCs	p = 0.481	p = 0.387	p = 0.387	p <0.001
kNN/ Manhattan	p = 0.422	p = 0.700	p = 0.043	p <0.001
kNN/ Manhattan PCs	p = 0.843	p = 0.414	p = 0.094	p <0.001
Leverage	p = 0.838	p = 0.750	p = 0.924	p <0.001
Leverage PCs	p = 0.992	p = 0.353	p = 0.277	p <0.001
Mahalanobis	p = 0.869	p = 0.738	p = 0.841	p <0.001
Mahalanobis PCs	p = 0.992	p = 0.353	p = 0.277	p <0.001

3.1.3.3 Group compounds based on distance threshold

In addition, thresholds were applied on the distance of test set compounds from the training compounds. All the methods showed to categorise the compounds in two groups (table 12). The one group included the compounds that were within the AD and the other group the compounds that were outside. These two groups illustrated a trend indicating that the RMSE for chemicals outside the AD is larger than that for chemicals within the AD. These two groups also showed a statistically significant different RMSE between the compounds inside and outside the AD. This is in line with a study, where a trend was observed between the compounds inside and outside the AD when different distance to model metrics used with both descriptors and PCs (Jaworska et al., 2005). This is an indication that these methods can distinguish between well predicted and less accurately predicted compounds.

Moreover, the k-NN algorithm was used with two different distance functions, the Euclidean and the Manhattan in both descriptor and PCs space. The results obtained from these methods seemed to allocate a different number of compounds within the AD. A possible explanation for that will be the different way that Euclidean and Manhattan distance weight differences. The Manhattan deals with the small and large differences of each variable alike, whereas the Euclidean distance penalises those differences by squaring them (Mathea et al., 2016). In addition, the advantage of the kNN method and leverage over Mahalanobis distance was that

they calculated distances based on 174 descriptors compared to the 126 and as a result the information loss due to correlated descriptors was minimised.

Furthermore, from table 12, it is evident that each method produced different results regarding the percentage of the compounds within and outside the AD. This is something that was also observed by other studies, where different distance to model metrics were used. These studies also suggested that the results derived with different AD approaches might vary even for the same set of compounds (Jaworska et al., 2005; Sahigara et al., 2012). As a consequence, as Sahigara *et al.* (2012) concluded, none of these methods can be used on its own and it is preferable to use all the possible strategies/methods to evaluate the AD.

Table 12: The table depicts the percentage of the compounds and the RMSE for the compounds inside and outside of the AD. The Mann Whitney results and the number of descriptors or PCs used are also shown.

Method	Within AD		Outside AD		ΔRMSE	Mann Whitney test.	Number of descriptors /PCs
	%	RMSE	%	RMSE			
kNN/ Euclidean	32.60	0.56	67.40	0.76	0.20	p<0.05	174 descriptors
kNN/ Euclidean PCs	38.64	0.59	61.36	0.77	0.18	p<0.05	27 PCs
kNN/ Manhattan	19.04	0.53	80.96	0.74	0.21	p<0.05	174 descriptors
kNN/ Manhattan PCs	35.71	0.59	64.29	0.76	0.17	p<0.05	27 PCs
Leverage	47.40	0.66	52.60	0.74	0.08	p<0.05	126 descriptors
Leverage PCs	91.30	0.70	8.70	0.77	0.07	p<0.05	27 PCs
Mahalanobis	7.23	0.52	92.77	0.72	0.20	p<0.05	126 descriptors
Mahalanobis PCs	77.51	0.68	22.49	0.78	0.10	p<0.05	27 PCs

Overall, the results produced with distance to model metrics in the descriptor space indicated that there will always be an uncertainty associated with any methods for assessing the AD of QSPR models (Netzeva et al., 2005). Therefore, there were compounds with high RMSE which had a lower distance to model and the *vice versa*. There are 2 possible explanations about that: The first one is the "unexpected deviation from the model". This is happening when a prediction is considered within the AD of the model but it is still unreliable because the compound might have an additional property not accounted by the model. The second reason is that the set of ChEMBL compounds is a heterogeneous set derived from more than 300 articles and thus the experimental Caco-2 protocols may vary from the proprietary protocol and ultimately the RMSE value is affected.

3.1.3.4 kNN with Tanimoto and Dice

After the evaluation of the distance of test set compounds from the model in descriptor space, the distance in the fingerprint space was also evaluated. The idea of estimating the AD in the fingerprint space has been also considered as important in similar studies, which are trying to establish the AD of QSPR models (Gadaleta et al., 2016; Weaver & Gleeson, 2008). In this part of the study, the test compounds were binned in 5 equally populated bins by increasing similarity. The RMSE in prediction was reported for each bin as shown in figure 26.

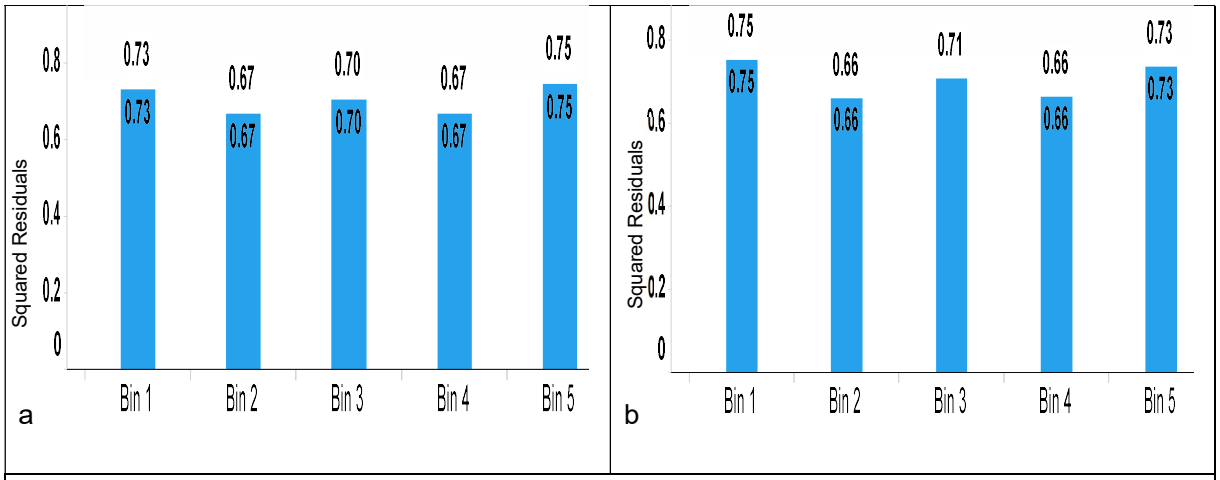

Figure 26: RMSE in prediction of the binned similarity to 5NNs for CHEMBL compounds calculated with: a) Tanimoto and b) Dice coefficients in ECFP4 fingerprint space.

The kNN method in ECFP4 fingerprint space by using the Tanimoto and Dice similarity show no correlation between the chemical similarity and the RMSE in the prediction. A possible explanation about this result is that Caco-2 permeability is a property greatly influenced by the physiochemical properties of the compounds (Artursson et al., 2012). Different functional groups, which are dissimilar, might show similar physiochemical properties. This is an interesting finding, which can be compared with other ADME models, which depend on the chemical structure of the compounds. There are ADME models, which depend on chemical

structure because examine protein interactions like the CYP-mediated metabolism or active transport (Testa & Turski, 2006) and there are models like membrane permeability which focus on the physiochemical characteristics. For example, in a study conducted by Weaver and Gleeson (2008) the kNN along with Tanimoto coefficient were used to establish the AD in the fingerprint space for a CYP450 3A4 inhibition model. The results suggested that there is a relationship with the prediction error. As a result, it would be interesting to use that distance to model metric in the ECFP4 fingerprint space in models, which aim to predict properties like metabolism or induction/inhibition of metabolic enzymes.

3.1.4 Conclusion

The PCA and AD results indicated that there were compounds which were within the chemical space of the Evotec compounds and there were some compounds, which were dissimilar. Compounds that have been predicted within the AD is not an indication that the prediction is correct but that the specific model was correctly applied for those compounds. The same implies for the compounds outside the AD and it means that there is an increased uncertainty with the prediction and model might or might not extrapolate a correct prediction. The AD estimation showed a weak trend between the error in the predictions (RMSE) and the calculated distance to model. The RMSE in predictions increased as the distance increased and this trend was observed, when the compounds were binned by increasing distance but not when binned by squared residuals. However, the results obtained from the application of a threshold, showed that a different percentage of compounds is considered within the AD based on the method used. Therefore, more than one of distance to model metrics should be considered in the estimation of AD. The distance to model metrics were able to give an indication of how far or close are the compounds from the training set but also other factors like the model ability to predict and the reliability of the compounds' source should also be taken into consideration.

For the next parts of this work, which focus on the development of *in-silico* $LogD_{7.4}$ and permeability models, the AD of the models will only be evaluated in descriptor space since $LogD_{7.4}$ and permeability are two properties dependent mainly on the physiochemical properties. In addition, the descriptors were preferred over the PCs and the reason was that for the Leverage, kNN with Euclidean and kNN with Manhattan only the descriptors with zero variance were excluded and therefore there was no information loss since zero variance descriptors were constant for all the chemical compounds. With the calculation of PCs and the selection of a number of them, it was definite that a percentage of the information was lost. The reason that all the distance to model metrics will be used is that none of them proved to be better than the other and there are suggestions in the literature to always use more than one distance to model metrics.

3.2 Evaluation of Caco-2 *in-silico* permeability models

The objective of this part was the development of QSPR models to predict Caco-2 A to B apparent permeability. Three types of models were built with different training sets, which included: i. literature, ii. proprietary and iii. merged proprietary and literature data. By comparing the performance and AD of the models, it was investigated if the merged models (Evotec+ChEMBL) could outperform the models developed with proprietary compounds (Evotec). Additionally, four distance to model metrics were applied to estimate the AD of the models and establish if the addition of literature data in proprietary models could enlarge the AD of proprietary models.

3.2.1 Models developed with literature data (ChEMBL models)

The first models reported herein were developed using only public data extracted from the ChEMBL database. These models are referred as “ChEMBL models” and were based on a set of 1628 compounds with Caco-2 permeability data extracted from ChEMBL and processed as describe in the methods section 2.3.1. Three different modelling algorithms were applied to build the QSPR models: random forest (RF), partial least square (PLS) and support vector regression (SVR).

Two different strategies were used to define and evaluate the goodness of a model. In one case, all the 1628 compounds were used to build the QSPR models and a “temporal” test set was derived subsequently, including new Caco-2 permeability data made available in a new version of ChEMBL. The temporal test set included 92 compounds. In the second case, the 1628 compounds were merged with the 92 compounds of the temporal test set and the diverse test set was built including 20% of the total number of compounds randomly selected, while the remaining 80% of the compounds have been used to build and train the model. The first testing strategy, also known as temporal test set may be more challenging and may be a better representation of a real drug discovery situation, when the Caco-2 permeability of new compounds will have to be predicted with an existing model. The RMSE of the predictions and the R^2 of the predicted versus experimental values were calculated for the test sets and used to evaluate the goodness of the model. Based on these metrics a better model will show a higher R^2 and a lower value of the RMSE for the prediction of compounds in the test set.

Table 13: RMSE in prediction and R^2 of ChEMBL diverse test set and ChEMBL temporal test set obtained with the ChEMBL model by using three different machine learning methods (RF, PLS &SVR). The red colour indicates the model that produced the lower RMSE in each testing strategy.

Model/Training set		**ChEMBL diverse test set**			**ChEMBL temporal test set**		
		RF	**PLS**	**SVR**	**RF**	**PLS**	**SVR**
ChEMBL	RMSE	0.54	0.64	**0.53**	0.69	0.78	**0.63**
	R^2	0.60	0.40	0.58	0.46	0.15	0.43

The results of the model assessment indicated that in both test sets (temporal or diverse), the nonlinear machine learning algorithms, RF and SVR, provided better performing predictive QSPR models than PLS (table 13). A possible explanation lies is the fact that there may be nonlinear relationships between Caco-2 permeability and the descriptors used. For example, MLR and SVR were compared for their ability to develop Caco-2 permeability models and the SVR performed better than MLR, due to the possible existence of non-linear relationships between Caco-2 permeability and descriptors (Karelson et al., 2009). Moreover, SVR and Boosting algorithms were able to provide more predictive Caco-2 permeability models (constructed with CHEMBL data) compared to MLR and PLS (Wang *et al.*, 2016). Cao and co-workers concluded that permeability, and in general ADME properties, are complex chemical systems not treatable or possible to be explained by mean of linear methods like PLS and MLR (Cao, Liang, Xu, Hu, & Zhang, 2011). The performance of RF and SVR was similar when assessed with the ChEMBL diverse test set. However, in the case of the ChEMBL temporal test set, the SVR algorithm performed slightly better. The SVR showed an RMSE of 0.63, whereas the RF showed an RMSE of 0.69. Both test sets were predicted with a high error in prediction and therefore the reliability of prediction by ChEMBL models is questionable and the results of prediction should be used with caution.

In general, RF and SVR are two popular methods and are probably considered as two of the best performing and more frequently used algorithms in cheminformatics (Mitchell, 2014). Moreover, there are many factors that can affect the performance of an algorithm like i) the size and distribution of compounds in chemical space, ii) the possible linearity of the chemical problem examined and iii) the nature of descriptors (Mitchell, 2014). The drawback of the SVR compared to the RF was that SVR was very time consuming due to the procedure needed to optimise the hyperparameters (C, ε and γ). On the other hand, RF required a minimal optimisation time. These algorithms also performed similarly in the development of regression models for the prediction of melting point and additionally outperformed other algorithms like kNN and PLS (Hughes et al., 2008).

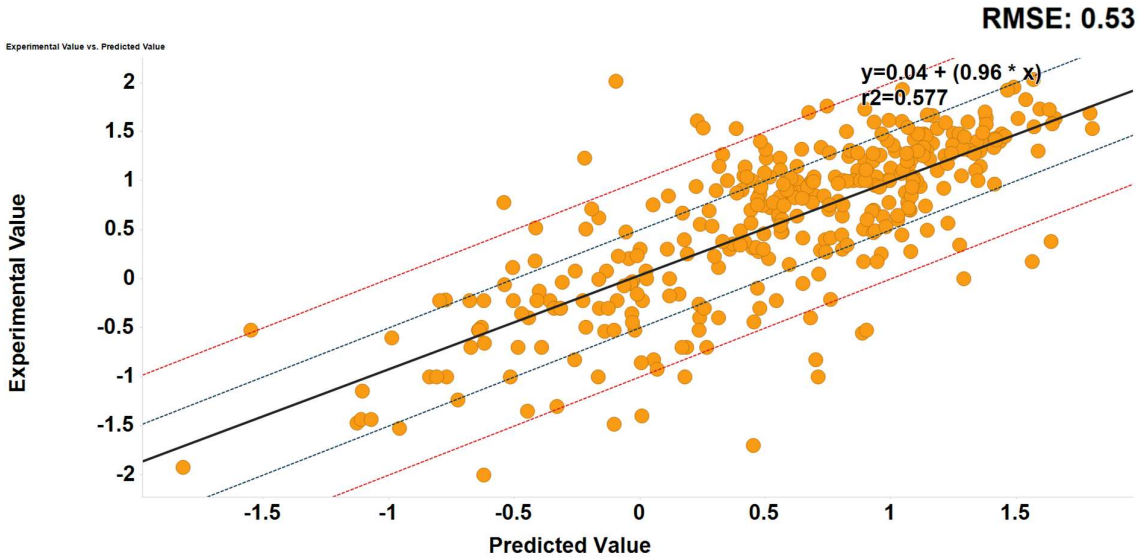

Figure 27: Experimental versus predicted Caco-2 permeability of compounds in the ChEMBL diverse test set obtained with the ChEMBL model developed with the SVR algorithm. Caco-2 permeability is reported as Log_{10} (A->B Papp[10^{-6} cm/s]). The black solid line represents the line of best fit in the form of y=b+ax. The red and dark blue dashed lines represent the y=x±1 and the y=x±0.5 respectively.

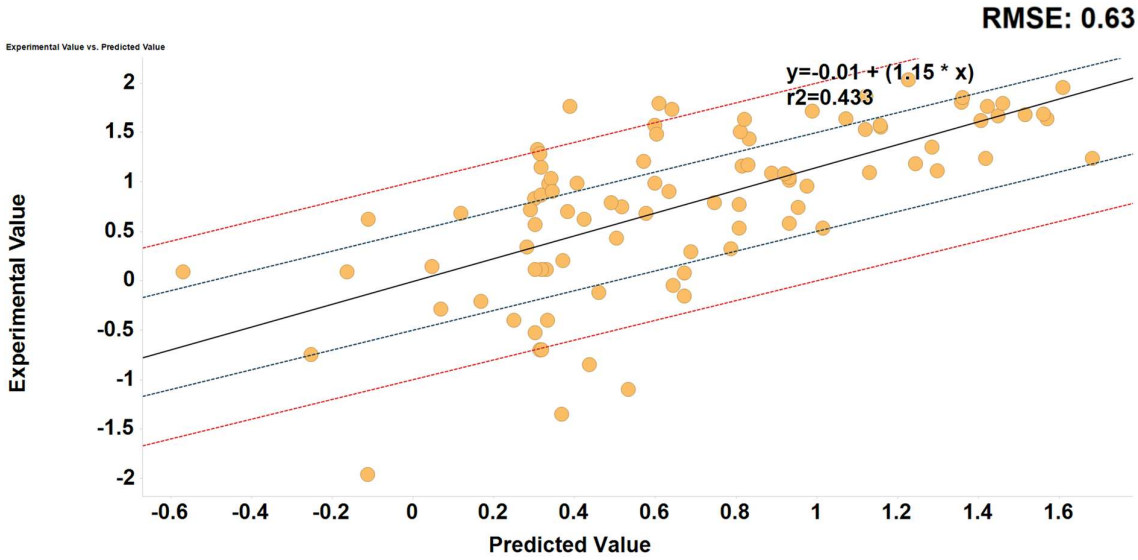

Figure 28: Experimental versus predicted Caco-2 permeability of compounds in the ChEMBL temporal test set obtained with the ChEMBL model developed with SVR algorithm. Caco-2 permeability is reported as Log_{10} (A->B Papp[10^{-6} cm/s]). The black solid line represents the

line of best fit in the form of y=b+ax. The red and dark blue dashed lines represent the y=x±1 and the y=x±0.5 respectively.

When the SVR ChEMBL model was applied on the literature diverse and temporal test sets the R^2 was equal to 0.58 and 0.43 respectively and the RMSE equal to 0.53 and 0.63 respectively (figures 27, 28). Therefore, the compounds in the diverse test set showed a better correlation between the experimental and predicted values and a lower RMSE in prediction compared to the temporal test set. However, R^2 should be considered cautiously because its value may be increased by addition of data in a narrow range of values. The diverse test set included a greater number of compounds compared to the temporal test set and thus the higher R^2 value might not indicating better model performance. The red and dark blue dashed lines enclose the compounds with predicted Caco-2 permeability within ±1 and ±0.5 log units respectively from the experimental values. These lines were used as a barrier to identify compounds with high and too high predicted values compared to experimental values (Schroeter et al., 2007). For the diverse test set, the 93.70% and the 84% of the predicted Caco-2 permeability values were within ±1 and ±0.5 log units respectively from the experimental values. For the temporal test set, the 89.13% and the 77.09% of the predicted Caco-2 permeability values were within ±1 and ±0.5 log units from the experimental values. Therefore, a smaller percentage of temporal test set compounds had prediction values from the experimental values within ±1 and ±0.5 log units. The reason might be that the compounds in temporal test set were novel or far from the model's chemical space and the model produced predictions with a higher error in prediction. Therefore, the compounds in the temporal test sets might not have been represented with compounds in the training set as it might have happened with the compounds in the diverse test set, which were randomly selected from the initial dataset.

3.2.2 Models developed with proprietary data (Evotec models)

The second models reported herein were developed using only proprietary data extracted from the Evotec database. These models are referred as "Evotec models" and were based on a set of 2075 compounds with Caco-2 permeability data. Three different modelling algorithms were applied to build the QSPR models: random forest (RF), partial least square (PLS) and supporting vector regression (SVR).

Two different strategies were used to define and evaluate the goodness of a model. In one case, all the 2075 compounds were used to build the QSPR model and a temporal test set was derived subsequently, when new compounds were added in the Evotec database with Caco2 permeability data. The temporal test set included 166 compounds. In the second case, the 2075 compounds were merged with the 166 compounds of the temporal test set and the diverse test set was built including 20% of the total number of compounds randomly selected, while the remaining 80% of the compounds have been part of the training set used to build the model.

Table 14: RMSE in prediction and R^2 of Evotec diverse test set and Evotec temporal test set obtained with the Evotec model by using three different machine learning methods (RF, PLS &SVR). The red colour indicates the model that produced the lower RMSE in each testing strategy.

Model/Training set		**Evotec diverse test set**			**Evotec temporal test set**		
		RF	**PLS**	**SVR**	**RF**	**PLS**	**SVR**
Evotec	RMSE	**0.36**	0.43	0.37	**0.57**	0.60	**0.57**
	R^2	0.75	0.64	0.73	0.44	0.45	0.49

The results of the Evotec model assessment indicated that in both test sets (temporal or diverse), the RF and SVR algorithms provided better performing predictive QSPR models than PLS (table 14) as observed with the model assessment results of the ChEMBL models (section 3.2.1). In the case of the diverse test set, the RF and SVR models provided similar RMSE values equal to 0.36 and 0.37 respectively. In the case of the temporal test set, the performance of RF and SVR was identical (RMSE = 0.57). The Evotec model predicted the diverse test set with a low error in prediction but in the case of the temporal test set the reliability of prediction by Evotec models is questionable and the results of prediction should be used with caution.

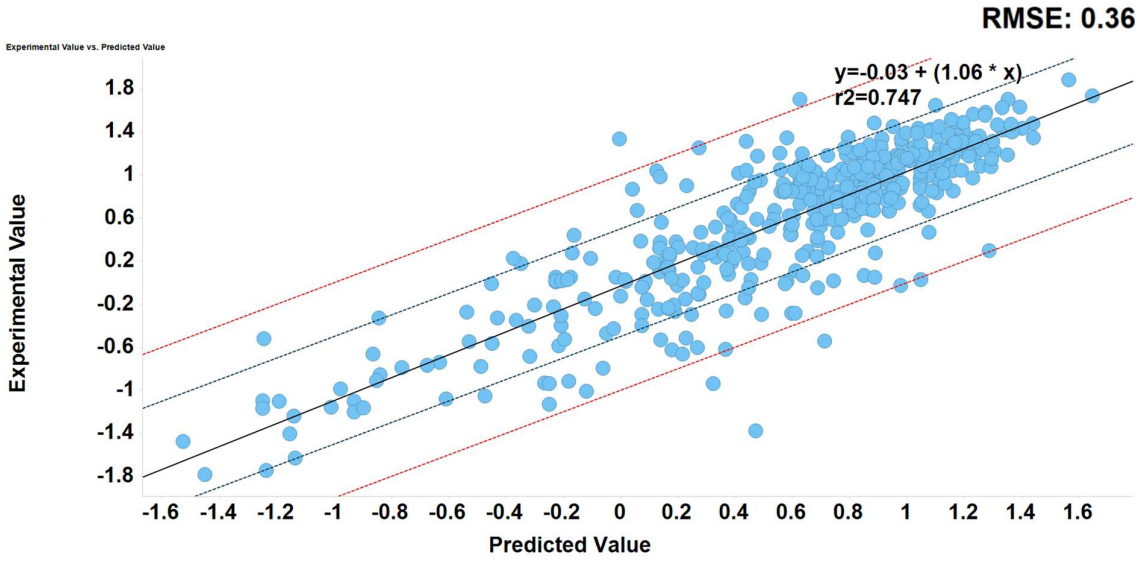

Figure 29: Experimental versus predicted Caco-2 permeability of compounds in the Evotec diverse test set obtained with the Evotec model developed with RF algorithm. Caco-2 permeability is reported as Log_{10} (A->B Papp[10^{-6} cm/s]). The black solid line represents the

line of best fit in the form of y=b+ax. The red and dark blue dashed lines represent the y=x±1 and the y=x±0.5 respectively.

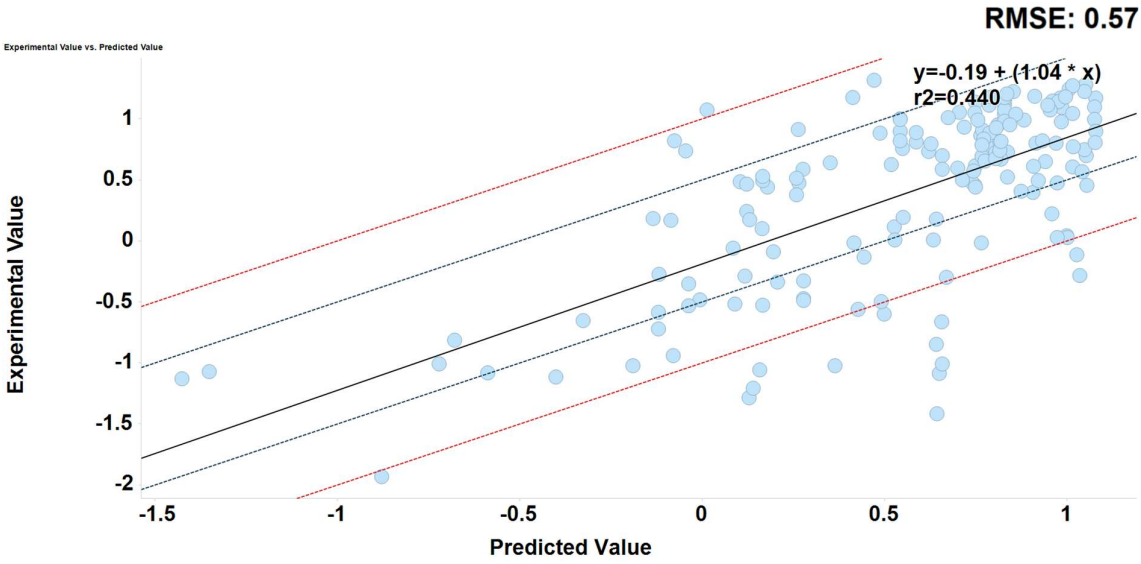

Figure 30: Experimental versus predicted Caco-2 permeability of compounds in the Evotec temporal test set obtained with the Evotec model developed with RF algorithm. Caco-2 permeability is reported as Log_{10} (A->B Papp[10^{-6} cm/s]). The black solid line represents the line of best fit in the form of y=b+ax. The red and dark blue dashed lines represent the y=x±1 and the y=x±0.5 respectively.

When the RF ChEMBL model was applied on the proprietary diverse and temporal test sets the R^2 was equal to 0.75 and 0.44 and the RMSE equal to 0.36 and 0.57 respectively (figures 29, 30). Therefore, the compounds in the diverse test set showed a better correlation between the experimental and predicted values and a lower RMSE in prediction compared to the temporal test set. However, R^2 should be considered cautiously because its value may be increased by addition of data in a narrow range of values. The diverse test set included a greater number of compounds compared to the temporal test set and thus the higher R^2 value might not indicating better model performance. The red and dark blue dashed lines enclosed the compounds with predicted Caco-2 permeability within ±1 and ±0.5 log units respectively from the experimental values. For the diverse test set, the 98.44% and the 93.10% of the predicted Caco-2 permeability values were within ±1 and ±0.5 log units respectively from the experimental values. For the temporal test set, the 91.57% and the 81.33% of the predicted Caco-2 permeability values were within ±1 and ±0.5 log units from the experimental values. Therefore, a smaller percentage of temporal test set compounds had prediction values from the experimental values within ±1 and ±0.5 log units. The reason might be that the compounds in temporal test set were novel or far from the model's chemical space and the model produced predictions with a higher error in prediction. Therefore, the compounds in the temporal test sets might not have been represented with compounds in the training set as it might have happened with the compounds in the diverse test set, which were randomly selected from the initial dataset.

3.2.3 Models developed with merged proprietary and literature data (Evotec+ChEMBL models)

The third group of models reported herein have been developed using both Evotec proprietary and literature data extracted from the ChEMBL database. These models are referred as "Evotec+ChEMBL models". The training sets from the two previous models were merged, thus resulting in 3703 compounds. Three different modelling algorithms were applied to build the QSPR models: random forest (RF), partial least square (PLS) and support vector regression (SVR).

Two different strategies were used to define and evaluate the goodness of a model. In one case, all the 3703 compounds were used to build the QSPR model and the test set (temporal) has been derived by merging the two previous (Evotec and ChEMBL) temporal test sets resulting in 258 compounds. In the second case, the 3703 compounds were merged with the 258 compounds of the temporal test set and the diverse test set was built including 20% of the total number of compounds randomly selected, while the remaining 80% of the compounds have been part of the training set used to build the model.

Table 15: RMSE in prediction and R^2 of Evotec+ChEMBL diverse test set and Evotec+ChEMBL temporal test set obtained with the Evotec+ChEMBL model by using three different machine learning methods (RF, PLS &SVR). The red colour indicates the model that produced the lower RMSE in each testing strategy.

Model/Training set		**Evotec+ChEMBL diverse test set**			**Evotec+ChEMBL temporal test set**		
		RF	**PLS**	**SVR**	**RF**	**PLS**	**SVR**
Evotec+ ChEMBL	RMSE	0.45	0.65	**0.44**	0.63	0.85	**0.62**
	R^2	0.66	0.33	0.66	0.36	0.07	0.39

The results of the model assessment (table 15) indicated that in both test sets (temporal or diverse), the RF and SVR algorithms provided more predictive QSPR models than PLS and this was also observed with the ChEMBL and Evotec models (sections 3.2.1, 3.2.2). Both SVR and RF algorithms performed similarly on the diverse and temporal test sets. For the diverse test set, the SVR produced an RMSE of 0.44 against an RMSE of 0.45 obtained with the RF model. For the temporal test set the SVR produced an RMSE of 0.62 against an RMSE of 0.63 obtained with the RF model. The Evotec+ChEMBL model predicted the diverse test set with a low error in prediction but in the case of the temporal test set the reliability of prediction by Evotec+ChEMBL models is questionable and the results of prediction should be used with caution.

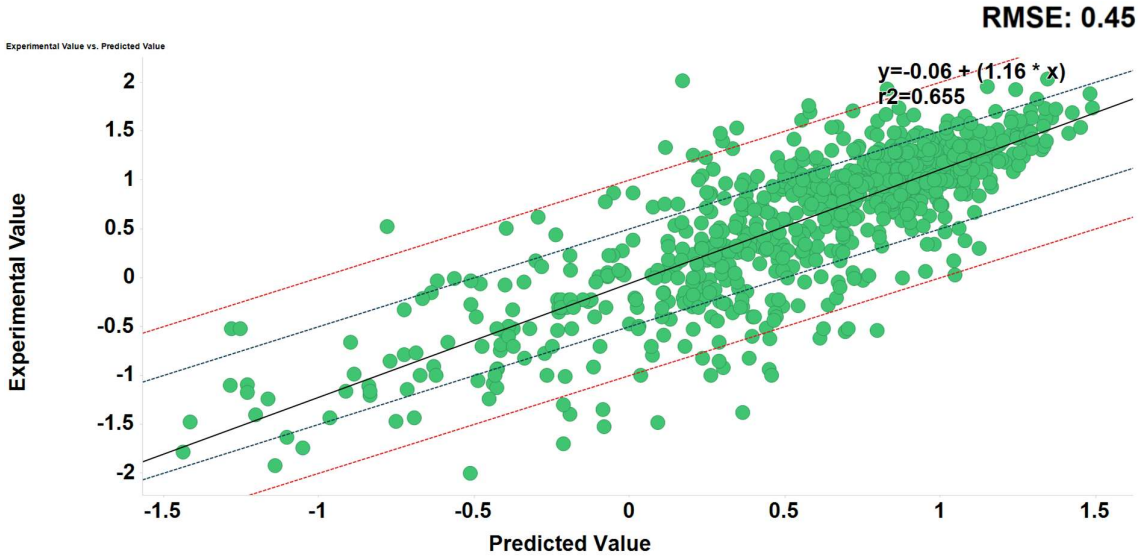

Figure 31: Experimental versus predicted Caco-2 permeability of compounds in the Evotec+ChEMBL diverse test set obtained with the Evotec+ChEMBL model developed with RF algorithm. Caco-2 permeability is reported as Log_{10} (A->B Papp[10^{-6} cm/s]). The black solid line represents the line of best fit in the form of y=b+ax. The red and dark blue dashed lines represent the y=x±1 and the y=x±0.5 respectively.

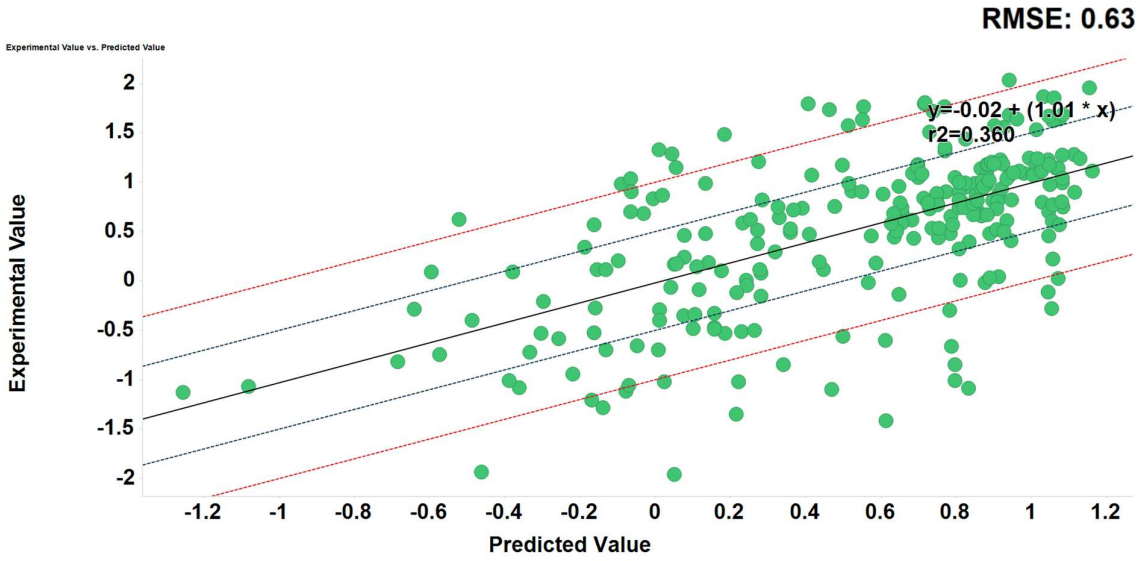

Figure 32: Experimental versus predicted Caco-2 permeability of compounds in the Evotec+ChEMBL temporal test set obtained with the Evotec+ChEMBL model developed with RF algorithm. Caco-2 permeability is reported as Log_{10} (A->B Papp[10^{-6} cm/s]). The black solid line represents the line of best fit in the form of y=b+ax. The red and dark blue dashed lines represent the y=x±1 and the y=x±0.5 respectively.

When the RF ChEMBL model was applied on the diverse and temporal test sets the R^2 was equal to 0.66 and 0.36 and the RMSE equal to 0.45 and 0.63 respectively (figures 31, 32). Therefore, the compounds in the diverse test set showed a better correlation between the experimental and predicted values and a lower RMSE in prediction compared to the temporal test set. However, R^2 should be considered cautiously because its value may be increased by addition of data in a narrow range of values. The diverse test set included a greater number of compounds compared to the temporal test set and thus the higher R^2 value might not indicating better model performance. The red and dark blue dashed lines enclosed the compounds with predicted Caco-2 permeability within ±1 and ±0.5 log units respectively from the experimental values. For the diverse test set, the 95.71% and the 88.40% of the predicted Caco-2 permeability values were within ±1 and ±0.5 log units respectively from the experimental values. For the temporal test set, the 86.05% and the 75.58% of the predicted Caco-2 permeability values were within ±1 and ±0.5 log units from the experimental values. Therefore, a smaller percentage of temporal test set compounds had prediction values from the experimental values within ±1 and ±0.5 log units. The reason might be that the compounds in temporal test set were novel or far from the model's chemical space and the model produced predictions with a higher error in prediction. Therefore, the compounds in the temporal test sets might not have been represented with compounds in the training set as it might have happened with the compounds in the diverse test set, which were randomly selected from the initial dataset.

3.2.4 Comparison of Caco-2 permeability models with models reported in the literature

The goal of this part of the work was to compare the models developed in the present study in sections 3.2.1-3.2.3 (i.e. ChEMBL, Evotec and Evotec+ChEMBL models) with models reported in the literature. Several regression permeability models have been reported and various limitations have been discussed in the introduction (section 1.9.3) regarding the training set size, the type of algorithms (linear vs nonlinear) and type of test sets (temporal vs diverse) used.

The permeability models reported in sections 3.2.1-3.2.3, were compared with the two most recent regression permeability models, published by Wang et al (2016) and Fredlund et al (2017), and reported in table 16. These two models were chosen as they exhibited two main similarities with the models developed in the present study. The first similarity is that they used a larger training set compared to other models in the literature outlined in section 1.9.3. The second is that the models' training sets incorporated literature data (Wang et al, 2016) and both literature and proprietary data (Fredlund et al, 2017) similarly to the models reported in the present study.

Table 16: The two most recent regression permeability models developed with caco-2 data.

Reference	Method	Number of Molecules	Number of Descriptors	Model Performance	AD Estimation?
(Wang et al, 2016)	MLR PLS SVR Boosting	1272	193	RMSE=0.31	Yes: Leverage
(Fredlund et al, 2017)	PLS SVR RF	2558	PLS, SVR:AZ descriptor set RF: signature descriptors	RMSE=0.45	No

The Caco-2 models developed by Wang et al (2016) showed several similarities with the ChEMBL models developed in the present study. Wang et al (2016) used Caco-2 permeability data from ChEMBL and a very similar filtering process was applied to ensure less experimental variability. The main difference in the filtering process was that in the present study the analytical method used during the Caco-2 assay had been taken into consideration. Therefore, compounds that during the Caco-2 assay, were analysed with a method different than LC/LC-MS were excluded. Different analytical methods could give different results and thus a cross validation of analytical method is necessary to ensure the optimal conditions to accurately reproduce an analytical measurement in different laboratories (Chau et al., 2008).

A training set of 1272 literature compounds was partitioned in a training set of 1017 compounds (80%) and a diverse test set of 255 compounds (20%) based on the joint x – y distances (SPXY) method to ensure that the test sets could map the measured region of the input variable space. However, this splitting might not represent a realistic drug design process. In pharmaceutical companies, ADME models are used to predict a variety of compounds, which might be chemically novel or physiochemically different from the training set compounds. By comparing the RMSE in prediction on the test sets for the models developed in the present study (ChEMBL, Evotec and Evotec+ChEMBL) and the model reported by Wang et al (2016), it seemed that their model performed better by showing a lower RMSE in prediction and a higher R^2 (Table 17). However, a direct comparison of the RMSE and R^2 would not be accurate, as different training and test sets were used.

Table 17: RMSE in prediction and R^2 of: literature ChEMBL model by Wang et al (2016), ChEMBL model, Evotec models and Evotec+ChEMBL model on their diverse test sets. The red colour indicates the highest performing modelling algorithm for each model.

	Method	RMSE	R^2
Literature ChEMBL model by Wang et al (2016)	MLR	0.36	0.75
	PLS	0.36	0.75
	SVR	0.32	0.80
	Boosting	**0.31**	**0.81**
ChEMBL model (developed in the present study)	PLS	0.64	0.45
	RF	0.54	0.42
	SVR	**0.53**	**0.59**
Evotec model (developed in the present study)	PLS	0.43	0.64
	RF	**0.36**	**0.75**
	SVR	0.37	0.73
Evotec+ChEMBL model (developed in present study)	PLS	0.65	0.40
	RF	0.45	0.74
	SVR	**0.44**	**0.73**

To make a direct comparison, the methodology developed in the present study, was applied on Wang et al (2016) training set and then the models derived, were assessed with the Wang et al (2016) diverse test set and the two external validation test sets. Therefore, the training compounds of Wang et al (2016) were trained by using the set of descriptors and algorithms used in the present study. The two external validation test sets included 298 compounds with Caco-2 permeability data and 220 compounds with MDCK permeability data obtained from ChEMBL and other literature sources. The two best performing algorithms were applied: 1. RF and 2. SVR and the results of the RMSE in prediction are shown in the table 18.

Table 18: RMSE in prediction of Boosting model developed by Wang et al (2016) and of the new model, developed with Wang et al (2016) training and test sets and the present study's methodology. The red colour indicates the highest performing model.

	Method	RMSE		
		Literature ChEMBL diverse test set by Wang et al (2016)	Caco-2 external Validation test set by Wang et al (2016)	MDCK external validation test set by Wang et al (2016)
Literature ChEMBL by Wang et al (2016)	Boosting	**0.31**	0.36	**0.38**
Literature ChEMBL model by Wang et al (2016) **developed with present study's methodology**	RF	0.34	**0.33**	0.41
	SVR	0.37	0.39	0.44

Table 18 shows that the methodology (algorithms) developed in the present study, when applied to the literature training and test sets provided comparable results. In more detail, RF results were very similar to the Boosting method as they are two similar algorithms that work by creating an ensemble of decision trees. However, the model developed with Wang et al (2016) training set and the present study's methodology performed better when evaluated external Caco-2 data. Therefore, a possible reason that the ChEMBL, Evotec and Evotec+ChEMBL models showed a higher RMSE in prediction (table 17) was due to the different partitioning of compounds in training and diverse test sets. In theory, a test set, which is representative of the training set can give a good model performance but at the same time could be unrealistic or very optimistic (Cherkasov et al, 2014). Therefore, in the present study temporal test sets and diverse test sets based on the random partition of the initial dataset were used. On the other hand, Wang et al (2016) used the joint x – y distances (SPXY) method to ensure that the test sets could map the measured region of the input variable space completely. Both random partitioning and the joint x – y distances (SPXY) partitioning offers advantages and disadvantages. The advantage of the random partitioning is that the compounds are "unknown" to the model (Martin et al., 2012). As a result, a random selection of a diverse test set gives an indication of the "realistic predictive power" of an ADME model. From another perspective, one could argue that the test should be reasonably similar and representative to the compounds of the training set. However, this approach could yield an "optimistic estimate" of the model performance (Cherkasov et al., 2014). In addition, it is important to take into account other parameters like the setting under which an ADME predictive model is used. For example, the use of a representative test set with optimistic model

assessment results might not be appropriate for a drug discovery project in a pharmaceutical company. The reason is that the newly synthesised proprietary compounds might or might not be similar to the model's training set. Therefore, a randomly selected diverse test set mimics that situation and potentially the results are more realistic.

In addition, Wang et al (2016) used only one distance to model metric to evaluate the AD of the models, whereas in the present study four different distance to model metrics compared. This comes in agreement with findings in the literature, which suggest to always use more than one distance to model metric for the AD evaluation (Sahigara et al, 2012).

The second model reported in table 16 is a regression permeability model by AstraZeneca (Fredlund et al., 2017). AstraZeneca model was developed with both proprietary and literature data and the training set of the model included 2558 compounds. The model performance was assessed with a diverse test set. The compounds in the diverse test set were randomly selected from the initial dataset and the model predicted the compounds with an RMSE equal to 0.45. This RMSE was comparable with the RMSE in prediction of the Evotec+ChEMBL model (RMSE=0.44). Both the AstraZeneca and the Evotec+ChEMBL models combined proprietary and opensource data in their training set and a reasonable error in prediction was observed. However, the AstraZeneca models had not been compared with models developed only with proprietary compounds. This comparison could indicate whether the literature data have a positive or negative impact on the proprietary models. This is an important point because there is a debate about the reliability of data in chemical databases and therefore it is interesting to investigate if their effect in proprietary models could possibly balance their experimental uncertainty. This is something investigated in the present study in section 3.2.5. Furthermore, the reliability of a model's prediction depends on two important factors. The first one is the methodology (algorithm and descriptors) and the second is the AD evaluation, which is something not reported for the AstraZeneca model. Fredlund et al (2017) monitored the performance of the model over a period of two years and the model was improved. Therefore, it would have been interesting to evaluate if the improvement in model performance relates with the possible enlargement of the AD and also to establish the effect of literature compounds on the models' AD. Therefore, the present study investigated these points regarding the AD in section 3.2.7.

3.2.5 The effect of merging proprietary and literature data in the development of Caco-2 permeability models

The three models ("ChEMBL", "Evotec" and "Evotec+ChEMBL" models), which were developed and described in sections 3.2.1-3.2.3, were used to evaluate the effect of the introduction of literature compounds in the proprietary models by testing both Evotec and Evotec+ChEMBL models on the same test sets. The test sets that were used were the Evotec and ChEMBL temporal and diverse test sets, which were also outlined in sections 3.2.1-3.2.3. The three models were trained with three different algorithms (RF, PLS and SVR) and were tested on the same diverse and temporal test sets and the results are outlined in tables 19 and 20 respectively.

The Evotec and Evotec+ChEMBL models predicted the permeability of Evotec diverse test set (table 19) with a low RMSE equal to 0.36 and 0.37 respectively (based on the RF predictions). However, in all the other cases, the models predicted the diverse and temporal test sets with a larger error in prediction. Therefore, the reliability of predictions is questionable and the results of prediction should be used with caution.

There are various reasons, which can negatively affect the performance of the models. One reason could be the problems related to the data heterogeneity (Cherkasov et al., 2014). The data extracted from ChEMBL database were used to build models (ChEMBL models) and were also merged with Evotec data to build models (Evote+ChEMBL models). The ChEMBL data are *in-vitro* ADME data, which are obtained from different sources of the medicinal chemistry literature. Therefore, inter-laboratory and/or protocol variability might have affected the models' performance. In addition, another challenge during the QSPR development is the introduction of errors during the descriptors' calculation (Cherkasov et al., 2014). There are some descriptors that can be accurately calculated (MW, atom count etc.) and some other which cannot. For example, LogP and LogD are usually calculated with a software (ChemAxon, ACD labs etc.) and thus errors might be introduced. Moreover, another factor that might have affected the model performance is the width of the AD. In general, if the test compounds are "far" from the models' training space, the models' predictions might exhibit a larger error in prediction. For example, the Evotec temporal test set compounds were extracted from the Evotec database. These compounds are novel and are synthesised for various drug discovery projects. Therefore, these compounds might have been novel and thus far from the AD of the models. Another reason that the models showed a high RMSE is the process of model validation, which exhibited both advantages and disadvantages. The advantage was that both internal (diverse) and external (temporal) test sets were used to assess the models' performance. However, the compounds in the internal/diverse validation test set were randomly selected from the initial dataset. Therefore, the test compounds were not strategically selected to be representative of the training set. Finally, another aspect that can negatively affect the model performance is the failure of the model to encounter for properties that might be important for the property investigated. For example, an important factor that could affect ADME properties is the presence of enantiomers. Appropriate descriptors should be used in the model building in order to make the model able to encounter for the effect of the enantiomers on the target value. In this study, only the number of chiral centres was considered. Therefore, there are aspects related to enantiomers, which were not considered. For example, other important features are the position of the enantiomers on the molecule and the handedness of a molecule's chiral centres (Bajorath, 2004). Thus, 3D descriptors could have been calculated to reflect these feature, which might be important.

Table 19: Table shows the model performance of “ChEMBL”, “Evotec” and “Evotec+ChEMBL” models. The RMSE in prediction and R^2 of Evotec and ChEMBL diverse test sets are reported. Results obtained by applying the RF, PLS and SVR algorithms. The red colour indicates the highest performing model between the Evotec and Evotec+ChEMBL models.

Number of compounds	Model/Training set		Evotec diverse test set			ChEMBL diverse test set		
			RF	PLS	SVR	RF	PLS	SVR
1376	ChEMBL	RMSE	0.58	0.65	0.61	0.54	0.64	0.53
		R^2	0.34	0.23	0.35	0.60	0.40	0.58
1792	Evotec	RMSE	**0.36**	**0.43**	0.37	0.72	1.13	0.75
		R^2	0.75	0.64	0.73	0.24	0.10	0.22
3168	Evotec+ ChEMBL	RMSE	0.37	0.57	**0.36**	**0.55**	**0.66**	**0.52**
		R^2	0.74	0.40	0.74	0.57	0.37	0.59

Table 20: Table shows the model performance of “ChEMBL”, “Evotec” and “Evotec+ChEMBL” models. The RMSE in prediction and R^2 of Evotec and ChEMBL temporal test sets are reported Results obtained by applying the RF, PLS and SVR algorithms. The red colour indicates the highest performing model between the Evotec and Evotec+ChEMBL models.

Number of compounds	Model/Training set		Evotec temporal test set			ChEMBL temporal test set		
			RF	PLS	SVR	RF	PLS	SVR
1628	ChEMBL	RMSE	0.60	0.72	0.71	0.69	0.78	0.63
		R^2	0.45	0.21	0.23	0.46	0.15	0.43
2075	Evotec	RMSE	0.57	**0.60**	**0.57**	0.85	1.74	0.74
		R^2	0.44	0.45	0.45	0.31	0.19	0.37
3703	Evotec+ ChEMBL	RMSE	**0.55**	0.76	0.58	**0.74**	**0.84**	**0.68**
		R^2	0.52	0.19	0.36	0.40	0.20	0.46

According to tables 19 and 20, the two best performing algorithms were the RF and SVR. The better performance of these methods over linear methods has also been observed in the literature by studies investigating permeability (Wang et al, 2016), lipophilicity (Wang et al, 2015; Rodgers et al, 2011) and plasma protein binding (Rodgers et al, 2011).

When the proprietary diverse test set used to assess the models, the Evotec+ChEMBL model (trained with the SVR algorithm) showed a similar error in prediction (RMSE = 0.36) with respect to the Evotec model (RMSE=0.37) (table 19). The Evotec+ChEMBL model also showed a similar relationship between the experimental and predicted values (R^2=0.74) compared to the Evotec model (R^2=0.73). However, the same result was not observed with the SVR and PLS algorithm. When the literature test set used to assess the models, the Evotec+ChEMBL model could better predict the Caco-2 permeability compared to the Evotec model. The Evotec+ChEMBL models provided a large improvement in the prediction of literature compounds by showing a lower RMSE in predictions and a higher R^2 (i.e. improved relationship between the experimental and predicted values) for all the algorithm tested.

Similar results and observations obtained with the temporal test sets (table 20). Evotec temporal test set was predicted similarly (i.e. with a similar error in prediction) by the Evotec and Evotec+CHEMBL models. However, when the ChEMBL temporal test set used to assess the models, the Evotec+ChEMBL model could better predict the Caco-2 permeability compared to the Evotec model. The Evotec+ChEMBL models showed a lower RMSE in predictions and a higher R^2 (i.e. improved relationship between the experimental and predicted values) for all the algorithm tested. These results indicated that the literature data could improve the prediction of newly synthesised compounds especially when the compounds are chemically novel. Temporal compounds extracted from literature (ChEMBL temporal test set) could theoretically mimic novel proprietary chemotypes and series from completely new projects or novel chemical matter in existing projects. Therefore, compounds extracted from literature can enhance the predictive ability of the proprietary models when they assess newly synthesised and chemically diverse compounds.

Although ChEMBL data might be considered as less experimentally reliable compared to proprietary data, they improved the performance of proprietary models. ChEMBL data can possibly introduce chemical diversity to the proprietary databases. AstraZeneca and Bayer Pharma AG conducted a study, which concluded that data extracted from ChEMBL can introduce chemical diversity in proprietary databases. Firstly, the two companies compared the chemical similarity of their screening collections and secondly, they compared the similarity of their screening collections with the ChEMBL database. The similarity of the two proprietary screening databases was calculated by using 2D molecular fingerprints in combination with a Nearest Neighbour (NN) approach and Tanimoto index (as a measure of molecular similarity) (Kogej et al, 2013). The outcome of that analysis was that there is a low overlap between the compound collections of these pharmaceutical companies in terms of molecular similarity. In addition, they identified the molecular similarity of the ChEMBL compounds with these 2 databases, which in total included about 3.7 million compounds. The number of ChEMBL compounds at the time was only 600K. The current ChEMBL version 23 contains about 2

million of compounds, which shows a great improvement in data reposition in that chemical database. More than the 80% of the compounds in ChEMBL database had their NN with a Tanimoto index less than 0.7. This result indicated that even in big proprietary screening databases, there is an unexplored chemical space. Therefore, ChEMBL compounds could be an asset in industry and academia to expand the chemical space and diversity of screening databases and subsequently proprietary ADME models. Therefore, this gives a possible explanation why the Evotec+ChEMBL models and ChEMBL models were better in predicting literature temporal compounds compared to proprietary Evotec models. It has also been reported that AstraZeneca develops permeability Caco-2 models, which incorporate both in house and literature data extracted from CHEMBL (Fredlund et al, 2017).

Furthermore, it is also evident from the results (table 19, 20), that the Evotec models were better in predicting Evotec temporal or Evotec diverse test set compounds compared to ChEMBL models. The same applies for the ChEMBL models; they can better predict ChEMBL temporal or diverse test set compounds compared to Evotec models. This was expected especially for Evotec models and Evotec test sets because compounds that are part of the training set and test sets might have been synthesised for the same project.

However, it is interesting to notice how the Evotec model predicted the ChEMBL compounds and how the ChEMBL model predicted the Evotec compounds. The RMSE in prediction, that obtained from both temporal and diverse test sets, indicates that the ChEMBL models can predict the Evotec test compounds more accurately than the Evotec model can predict the ChEMBL test compounds. This is an interesting point because it was expected that Evotec models could possibly be better in predicting the ChEMBL compounds due to the more experimentally reliable Caco-2 measurements. In contrast, ChEMBL were mainly compounds extracted from the literature and it was difficult to ensure the same and accurate experimental conditions due to the inter-laboratory and assay variability. Therefore, an explanation might be given with the investigation of the AD of the models. ChEMBL models might: 1. exhibit a greater chemical diversity, 2. cover a larger chemical space and consequently 3. exhibit a larger applicability domain compared to Evotec models. This was investigated in section 3.2.7.

In conclusion, the merging of the compounds from different sources (proprietary and literature) was beneficial despite the debate regarding the merging of biological data from different sources and especially from large chemical databases. The mixing of data from different sources could be dangerous as the data are generated with different experimental protocols. Therefore, there is an increasing risk to introduce errors and noise in the training set. For example, the training sets should include data, which ideally are measured based on a single protocol and by the same laboratory. (Cronin & Schultz, 2003). In addition, data with high experimental uncertainty like literature compounds could negatively influence the model performance (Wenlock and Carlsson, 2014). However, the results and the performance of the merged models (Evotec+ChEMBL) seemed to balance the experimental uncertainty of the data. The Evotec+ChEMBL models exhibited a similar performance with the Evotec models in the prediction of proprietary test sets and showed a significant improvement for the prediction of literature test sets.

3.2.6 Subsequent model assessment of the Caco-2 permeability models

In a subsequent model assessment, the permeability data in the temporal test sets were merged with the training test sets and used, all together, to develop an updated model. Two new temporal test sets were generated including the latest proprietary permeability data (Evotec compounds synthesised four months after the compounds in the training set) and the freshly published public permeability data from ChEMBL version 23. These new temporal test sets are referred as “New Evotec temporal test set” and “New ChEMBL temporal test sets”. The new temporal test sets had been used to assess both the initial models (M1) reported in the sections 3.2.1-3.2.3 and the new models introduced in this section (M2). In this analysis, only the RF algorithm was applied to build the QSPR models due to its performance (as discussed in the previous sections) and to its computational inexpensiveness.

Table 21: Table shows the model performance of the “initial” (M1) and “new” (M2) “ChEMBL”, “Evotec” and “Evotec+ChEMBL” models. The RMSE in prediction of the “new” Evotec and ChEMBL temporal test sets is reported. Results obtained by applying the RF algorithm and the red colour indicates the highest performing model between Evotec and Evotec+ChEMBL models.

Number of compounds	Model/Training set	New Evotec temporal test set		New ChEMBL temporal test set	
		M1	M2	M1	M2
M1:1628 M2: 1720	ChEMBL	0.67	0.66	0.47	0.48
M1:2075 M2: 2241	Evotec	0.47	0.42	0.67	0.68
M1:3703 M2:3961	Evotec+ChEMBL	**0.45**	**0.40**	**0.66**	**0.63**

The new temporal test sets were predicted with a lower RMSE in prediction (i.e. better predicted) with the Evotec + ChEMBL model compared to Evotec model. In addition, the RMSE in prediction obtained with the new model (M2) was lower than that of M1. The new temporal test sets were predicted with higher accuracy by the merged (Evotec+ChEMBL model) compared to the Evotec model. This is an indication of the robustness of the method and that the addition of the ChEMBL compounds in the proprietary Evotec models is beneficial.

3.2.7 Applicability Domain estimation of the *in-silico* Caco-2 permeability models

Determining the AD for a QSPR model is important to estimate the reliability of a prediction of an external compound. If the compound lies within the AD of the QSPR model used to predict a property, this prediction can be taken, otherwise the prediction should be either discarded or given a low reliability flag.

The AD of the models was estimated with the four distance to model metrics: 1. k-NN with Euclidean distance, 2. k-NN with Manhattan distance, 3. Leverage and 4. Mahalanobis distance. The distance to model metrics calculated the distance of the test compounds from the training set in the descriptors' space (i.e. the multi-dimensional space defined by the descriptors of the compounds used to train the model) and a threshold was applied. Above that threshold, compounds were considered to be outside the AD.

The goal of this section was twofold. Firstly, the AD of the Evotec+ChEMBL model was compared with the AD of the Evotec model. Secondly, once an AD distance threshold had been determined, the goal was to check whether test set compounds within the AD were predicted more accurately than compounds outside the AD. To do that, compounds in the test set were partitioned in two groups; within the AD, and outside the AD. If compounds within the AD show an RMSE in the prediction smaller than compounds outside the AD, the particular distance metric is able to clearly define an AD for the model. In addition, a Mann Whitney test was used to establish the presence of a statistically significant difference in the RMSE of the compounds inside and outside of the AD. To achieve both objectives, for every model ("ChEMBL model", "Evotec model" and "Evotec+ChEMBL"), the portion of compounds within and outside the AD was calculated by using all the four distance metrics mentioned above. The distance of compounds in the two temporal test sets ("Evotec temporal test set" and "ChEMBL temporal test set") was calculated from the training compounds of the three different models ("ChEMBL model", "Evotec model" and "Evotec+ChEMBL" model) and the percentage of the test compounds within the AD of the models was calculated. Results are reported in table 22.

Table 22: Results obtained with the kNN with Euclidean distance, kNN with Manhattan distance, Leverage and Mahalanobis distance for the three different Caco-2 permeability models and the two different temporal test sets. The table summarises: the percentage of compounds inside the AD of the models, the RMSE in prediction of compounds inside and outside the AD and the assessment of the statistical significance with the Mann Whitney (MW) test. The red colour indicates the presence of a statistically significant difference in the RMSE of the compounds inside and outside of the AD.

		kNN/ Euclidean				kNN/ Manhattan				Leverage				Mahalanobis			
		In		Out	MW	In		Out	MW	In		Out	MW	In		Out	MW
		%	RMSE	RMSE	test	%	RMSE	RMSE	test	%	RMSE	RMSE	test	%	RMSE	RMSE	test
Model	**Evotec+ChEMBL**																
Evotec temporal test set		78.31	0.56	0.52	p>0.05	80.72	0.55	0.55	p>0.05	86.15	0.50	0.81	**p<0.05**	64.46	0.48	0.66	**p<0.05**
ChEMBL temporal test set		43.48	0.64	0.80	p>0.05	33.70	0.75	0.73	p>0.05	58.70	0.66	0.84	**p<0.05**	32.21	0.60	0.80	**p<0.05**
Model	**Evotec**																
Evotec temporal test set		74.70	0.55	0.61	**p<0.05**	74.09	0.55	0.60	**p<0.05**	76.50	0.55	0.62	**p<0.05**	47.60	0.50	0.62	**p<0.05**
ChEMBL temporal test set		15.22	0.77	0.86	p>0.05	5.43	0.39	0.87	**p<0.05**	50.00	0.66	1.00	**p<0.05**	3.26	0.60	0.86	p>0.05
Model	**ChEMBL**																
Evotec temporal test set		66.27	0.57	0.66	p>0.05	67.50	0.57	0.65	p>0.05	82.33	0.51	0.89	**p<0.05**	39.15	0.66	0.56	p>0.05
ChEMBL temporal test set		53.26	0.64	0.73	p>0.05	50.00	0.63	0.74	p>0.05	61.96	0.62	0.78	**p<0.05**	42.39	0.60	0.74	p>0.05

From table 22, it can be observed that, in most cases, the RMSE in prediction of the compounds within the AD was lower than the RMSE of the compounds outside the AD. This evidence provided confidence for using these distance metrics and threshold determination method as a reliable protocol for defining the AD of the models. Similar studies are in agreement with these findings (Jaworska et al., 2005; Sahigara et al., 2012) as they have employed, the same distance metrics and reached similar conclusions.

In addition, table 22 indicated that a greater percentage of Evotec temporal test compounds and ChEMBL temporal test compounds were within the Evotec+Chembl model's AD compared to the Evotec proprietary model. Therefore, all the four distance to model metrics demonstrated that the AD of the Evotec models was enlarged with the inclusion of compounds extracted from the literature. This indicated that the literature compounds can introduce chemical diversity and cover unexplored areas of the chemical space (Kogej et al, 2013). In addition, based on the work performed in present study, the 50-90% (depending the distance to model metric considered) of the compounds extracted from ChEMBL were outside the AD of the existing Evotec model. Therefore, by merging the proprietary Evotec compounds with the compounds extracted from ChEMBL database, it was expected to introduce chemical diversity in the training set. In addition, a greater percentage of Evotec temporal compounds was within the AD of the ChEMBL models than the percentage of ChEMBL temporal compounds within the AD of the Evotec model. This observation explained the results obtained from the evaluation of model performance in section 3.2.5, where the ChEMBL model was better at predicting Evotec temporal compounds compared to Evotec model in predicting ChEMBL temporal compounds. Therefore, literature compounds offer chemical diversity and this is a clear positive impact for the proprietary chemical space.

There were also cases in which the compounds outside the AD showed an RMSE lower than the compounds inside. This may be an indication that being within the AD, although important to consider a prediction reliable, may not be sufficient for a lower RMSE in the prediction (Gadaleta et al., 2016). One reason might be the presence of activity cliffs within the chemical space. There might be chemical areas, where the permeability of the compounds change due to the presence of a particular functional group. Therefore, that means that a compound exhibits a property not encountered by the model (Netzeva et al., 2005). For example, passive transcellular permeability is considered to be the major permeability route for the compounds but compounds can also be transferred through carrier mediated transport. Therefore, if a compound has a chemical structure, which enables the binding with a membrane transporter could theoretically be permeable. Another possible explanation might be the presence of areas with a lack of chemical coverage (i.e. due to data scarcity) (Aniceto, Freitas, Bender, & Ghafourian, 2016). As a result, there is a possibility a model not to be able to make accurate predictions due to the inadequate chemical space coverage. Moreover, another possible explanation for a lower error in prediction in compounds outside the AD is that the model could possibly extrapolate correctly outside its domain (Jaworska et al., 2005). In a study, where permeability models have been developed with Caco-2 data extracted from the literature, the leverage method was used as the metric to establish AD (Wang et al, 2016). In that case, there were compounds with low leverage and a larger error in prediction and *vice versa*. The reason

of that phenomenon is at the way that the threshold is set. A threshold has a value, which simply excludes training compounds in the extremities. That means that the AD evaluation only considers the interpolation and not the possible extrapolation. Therefore, there might be compounds in the test sets that are close to that outliers and thus they can predicted with a lower error in prediction than expected. However, there is not a definite answer and all the arguments mentioned above could provide a possible explanation.

Furthermore, each distance to model metric produced different results (Jaworska et al., 2005; Gadaleta et al., 2016; Sahigara et al., 2012). This depends on: the different way that each method measures distance, the threshold definition and the descriptors used in each method (e.g. Mahalanobis distance cannot handle correlated descriptors). In addition, another difference is that the Euclidean distance assumes a normal distribution compared to other (Jaworska et al., 2005; Gadaleta et al., 2016) and that might be a potential disadvantage. Therefore, the results obtained from the evaluation of AD with the distance to model metrics should be carefully examined. The fact that each method gave a different result regarding the percentage of the compounds inside and outside of the AD and also regarding the error in prediction for the compounds inside and outside of the AD, results in a confusion about which method is the most appropriate to use. For that issue, there are reports, which suggest to always using more than one distance to model metric for the assessment of AD (Sahigara et al., 2012). The reason is that none of the existing methods can be considered as the universally the most appropriate because each method has its own advantages and disadvantages.

Therefore, a statistical analysis used to understand which method is more reliable in each case. In the literature, there are no statistical comparisons, which can distinguish between a statistically significant difference between the RMSE in prediction for compounds in and out of AD. For that reason, the non-parametric Mann Whitney test was applied. The results showed that only the Leverage method could produce a statistically significant difference in the RMSE for the compounds inside and outside of the AD. For the other three methods, there was not always a statistically significant difference. Therefore, the Leverage method could be considered as the most effective method for these models. Leverage is a method used in the literature for the estimation of the AD of permeability models (Wang et al, 2016), $LogD_{7.4}$ models (Wang et al, 2015).

In conclusion, in this case, leverage was considered as the best performing method for two reasons. The first reason was that it was able to categorise the test compounds inside and outside of the AD with a lower and higher RMSE in prediction respectively. The second reason was the presence of a statistically significant difference in all the measurements. However, if a compound is outside the AD, it is not definite that the prediction is erroneous but it provides an AD “warning” (Gupta, Adams, & Berry, 2016).

3.3 Evaluation of *in-silico* $LogD_{7.4}$ models.

The objective of this part was the development of QSPR models to predict $logD_{7.4}$. Three types of models were built with different training sets, which included: i) literature, ii) proprietary and iii) merged proprietary and literature data. By comparing the performance and AD of the models, it was investigated if the merged models (Evotec+ChEMBL) could outperform the models developed with proprietary compounds (Evotec). Additionally, four distance to model metrics were applied to estimate the AD of the models and establish if the addition of literature data in proprietary models could enlarge the AD of proprietary models.

3.3.1 Models developed with literature data (ChEMBL models)

The first models reported herein were developed using only public data extracted from the ChEMBL database. These models are referred as "ChEMBL models" and were based on a set of 1209 compounds with distribution coefficient at pH=7.4 ($LogD_{7.4}$) data extracted from ChEMBL and processed as described in the methods section 2.3.1. Three different modelling algorithms have been applied to build the QSPR models: random forest (RF), partial least square (PLS) and support vector regression (SVR).

Two different strategies were used to define and evaluate the goodness of a model. In one case, all the 1209 compounds were used to build the QSPR models and a "temporal" test set was derived subsequently, including new $LogD_{7.4}$ data made available in a new version of ChEMBL. The temporal test set included 86 compounds. In the second case, the 1209 compounds were merged with the 86 compounds of the temporal test set and the diverse test set was built including 20% of the total number of compounds randomly selected, while the remaining 80% of the compounds were used to build and train the model. The first testing strategy, also known as temporal test set, may be more challenging and may be a better representation of a real drug discovery situation. It provides an estimation of the model performance in a "real-life" situation, when lipophilicity of new compounds will have to be predicted with an existing model. The RMSE of the predictions and the R^2 of the predicted versus experimental values were calculated for the test sets and used to evaluate the goodness of the model. Based on that metric a better model will show a higher R^2 and a lower value of the RMSE for the prediction of compounds in the test set.

Table 23: RMSE in prediction and R^2 of ChEMBL diverse test set and ChEMBL temporal test set obtained with the ChEMBL model by using three different machine learning methods (RF, PLS &SVR). The red colour indicates the model that produced the lower RMSE in each testing strategy.

Model/Training set		ChEMBL diverse test set			ChEMBL temporal test set		
		RF	PLS	SVR	RF	PLS	SVR
ChEMBL	RMSE	1.01	1.34	**0.94**	0.84	1.09	**0.77**
	R^2	0.79	0.65	0.82	0.72	0.58	0.71

The results of the model assessment indicated that in both test sets (temporal or diverse), the RF and SVR algorithms provided better performing predictive $LogD_{7.4}$ models than PLS (table 23). This is an observation also reported in the literature. For example, $LogD_{7.4}$ predictive models have been developed with data extracted from ChEMBL database and the $LogD_{7.4}$ of the diverse test set was better predicted with the SVR algorithm compared to the PLS (Wang et al, 2015). There should be a nonlinear relationship between the target value ($LogD_{7.4}$) and the descriptors. In addition, proprietary $LogD_{7.4}$ models were developed with various algorithms (linear ridge regression, Gaussian process, SVR and RF) in Bayer Shering Pharma AG. Among the methods used, the linear ridge regression, which is a linear machine learning method as PLS, performed the worst (Schroeter et al., 2007) and the SVR algorithm performed better than RF. In addition, Schroeter and co-workers (2007) used literature temporal test sets to assess the performance of the proprietary $LogD_{7.4}$ models and the SVR algorithm produced the best results. Finally, $LogD_{7.4}$ models developed with RF and PLS algorithms (Rodgers et al., 2011) in AstraZeneca and the RF was more predictive than PLS.

The SVR algorithm performed better compared to the RF algorithm in both diverse and temporal test sets. In the case of the ChEMBL diverse test set, the SVR algorithm performed slightly better, by showing an RMSE of 0.94 compared to the RF which showed an RMSE equal to 1.01. In the case of the ChEMBL temporal test set, the SVR algorithm performed better, by showing an RMSE of 0.77 compared to RF that showed an RMSE of 0.84. In addition, the SVR predicted values with a higher correlation (R^2) with the experimental values for both test sets compared to RF. Both test sets were predicted with a high error in prediction and therefore the reliability of prediction by ChEMBL models is questionable and the results of prediction should be used with caution.

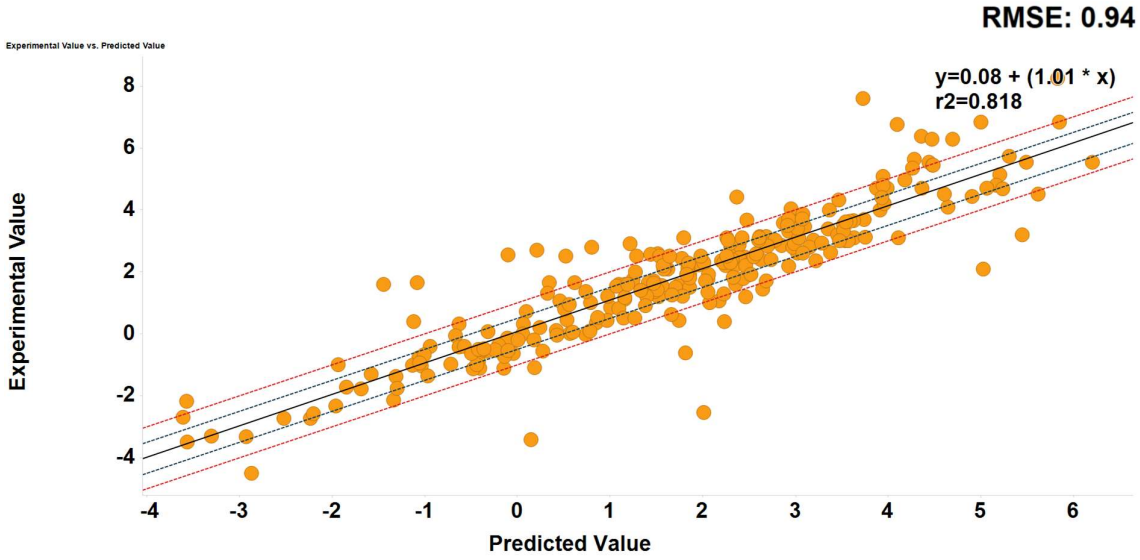

Figure 33: Experimental versus predicted $logD_{7.4}$ values of compounds in the ChEMBL diverse test set obtained with the ChEMBL model developed with the SVR algorithm. $LogD_{7.4}$ lipophilicity is reported as Log_{10} D. The black solid line represents the line of best fit in the form of y=b+ax. The red and dark blue dashed lines represent the y=x±1 and the y=x±0.5 respectively.

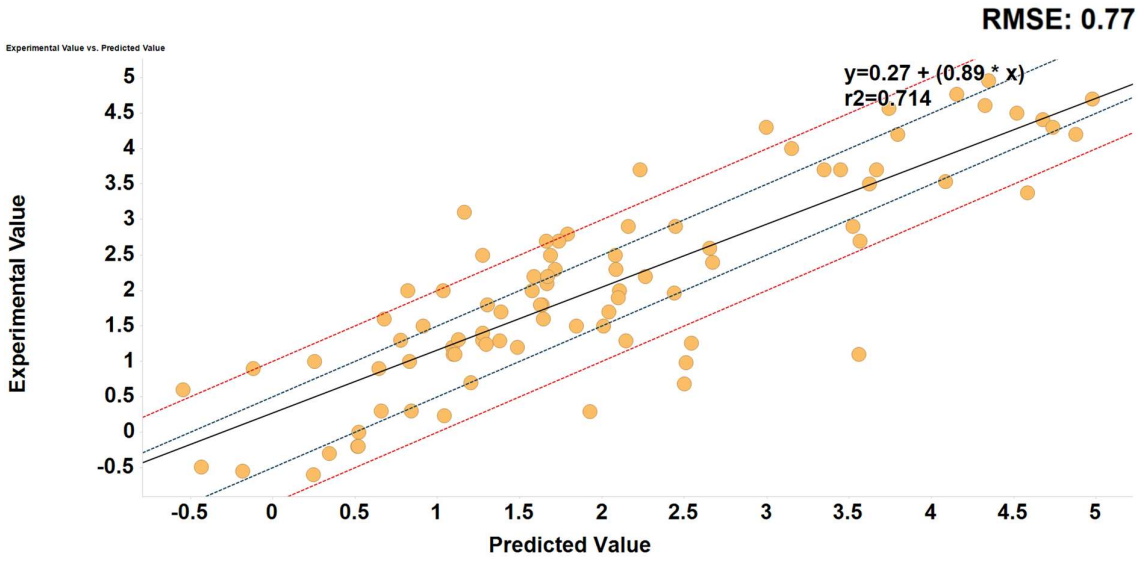

Figure 34: Experimental versus predicted $logD_{7.4}$ values of compounds in the ChEMBL temporal test set obtained with the ChEMBL model developed with the SVR algorithm. $LogD_{7.4}$ lipophilicity is reported as Log_{10} D. The black solid line represents the line of best fit in the form of y=b+ax. The red and dark blue dashed lines represent the y=x±1 and the y=x±0.5 respectively.

When the SVR ChEMBL model was applied on the literature diverse and temporal test sets the R^2 was equal to 0.82 and 0.71 respectively and the RMSE equal to 0.94 and 0.77 respectively (figures 33, 34). Thus, the compounds in the temporal test set showed a better correlation between the experimental and predicted values and a lower RMSE in prediction compared to the diverse test set. However, R^2 should be considered cautiously because its value increases with the addition of data the wider range of the data present in the test sets. The diverse test set contained a greater number of compounds with a wider value range and thus the R^2 value could have been affected. The red and dark blue dashed lines enclosed the compounds with predicted $LogD_{7.4}$ values within ±1 and ±0.5 log units respectively from the experimental values. For the diverse test set, the 82.63% and the 69.88% of the predicted $LogD_{7.4}$ values were within ±1 and ±0.5 log units from the experimental values respectively. For the temporal test set, the 80.23% and the 69.77% of the predicted $LogD_{7.4}$ values were within ±1 and ±0.5 log units from the experimental values. Therefore, approximately the same percentage of predicted values is found to be within ±1 and ±0.5 log units from the experimental values for both temporal and diverse test sets.

The RMSE in prediction for the temporal test set was lower than the RMSE of the diverse test set. The temporal test set was expected to be more challenging to predict compared to the diverse. The diverse set instead was part of the initial dataset and there was a possibility that the diverse test sets contained data similar to those present in the training set. A possible reason that the diverse test set was predicted with a higher RMSE is that the compounds that were randomly selected were not representative of the training set. There are rational division methods, like the Kennard-Stone, that can be used to partition the initial dataset into a training and a representative test set. A well representative test set can give a better result because the test compounds are represented in the training set. However, there are advantages and disadvantages in both approaches. The advantage of randomly splitting the compounds in training and test set is that the compounds might or might not be representative of the training set. As a result, a random selection of a diverse test set gives an indication of the “realistic predictive power” of an ADME model. From another perspective, one could argue that the test should be reasonably similar and representative to the compounds of the training set. However, this approach could yield an “optimistic estimate” of the model performance (Cherkasov et al., 2014). This optimistic estimate was observed by a study in which different rational methods and the random method were evaluated (Martin et al., 2012). The results of this study indicated that the rational division methods can result in better statistical results but there are cases where the predictive power of both rational and random division are comparable. In addition, in a pharmaceutical company’s drug design process, the use of a representative test set with optimistic model assessment results might not be appropriate. The reason is that the newly synthesised proprietary compounds might or might not be similar to the model’s training set. Therefore, a randomly selected diverse test set mimics that situation and potentially the results are more realistic.

3.3.2 Models developed with proprietary data (Evotec models)

The second models reported herein were developed using only proprietary data extracted from the Evotec database. These models are referred as “Evotec models” in this and the following section and are based on a set of 8400 compounds with distribution coefficient at pH=7.4 (logD7.4) data. Three different modelling algorithms were applied to build the QSPR models: random forest (RF), partial least square (PLS) and support vector regression (SVR).

Two different strategies were used to define and evaluate the goodness of a model. In one case, all the 8400 compounds were used to build the QSPR models and the test set (temporal) were derived subsequently, when new compounds were added in the Evotec database with $logD_{7.4}$ lipophilicity data. The temporal test set included 895 compounds. In the second case, the 8400 compounds were merged with the 895 compounds of the temporal test set and the diverse test set was built including 20% of the total number of compounds randomly selected, while the remaining 80% of the compounds have been used to build and train the model.

Table 24: RMSE in prediction and R^2 of Evotec diverse test set and Evotec temporal test set obtained with the Evotec model by using three different machine learning methods (RF, PLS &SVR). The red colour indicates the model that produced the lower RMSE in each testing strategy.

Model/Training set		**Evotec diverse test set**			**Evotec temporal test set**		
		RF	**PLS**	**SVR**	**RF**	**PLS**	**SVR**
Evotec	RMSE	0.60	0.63	**0.49**	0.62	0.68	**0.53**
	R^2	0.77	0.72	0.84	0.66	0.55	0.72

The results of the model assessment indicated that in both test sets (temporal or diverse), the RF and SVR algorithms provided better performing predictive $LogD_{7.4}$ models than PLS (table 24). This is an observation also reported in the literature and has discussed in section 3.3.1. The SVR algorithm performed better compared to the RF algorithm in both diverse and temporal test sets with an RMSE of 0.53 and 0.49 respectively. In addition, the correlation between experimental and predicted $LogD_{7.4}$ values was higher when the SVR applied for both diverse (R^2=0.84) and temporal (R^2=0.72) test sets compared to RF. Both test sets were predicted with a high error in prediction and therefore the reliability of prediction by Evotec models is questionable and the results of prediction should be used with caution.

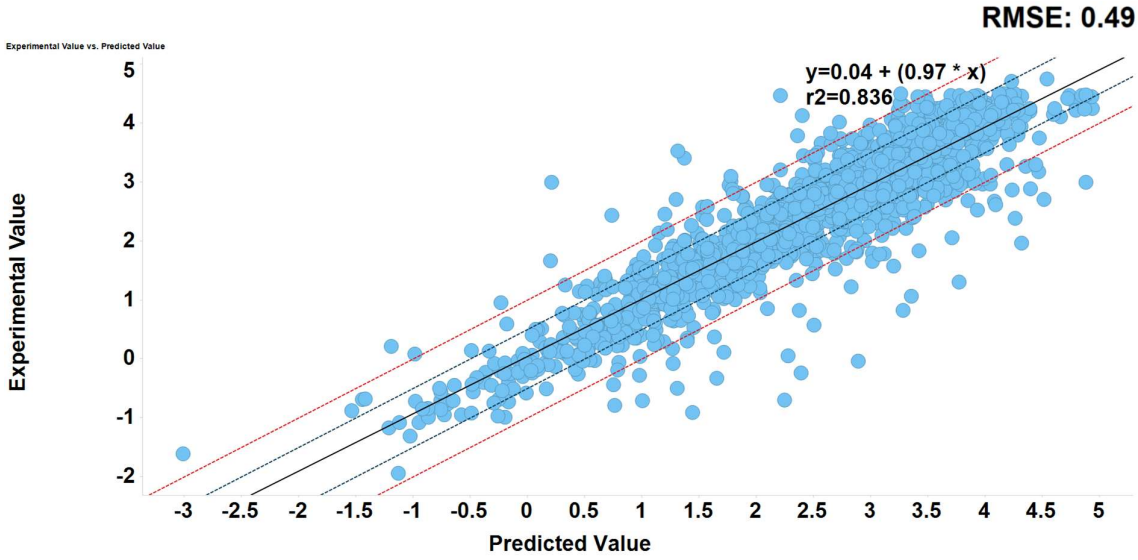

Figure 35: Experimental versus predicted $logD_{7.4}$ values of compounds in the Evotec diverse test set obtained with the Evotec model developed with the SVR algorithm. $LogD_{7.4}$ lipophilicity is reported as Log_{10} D. The black solid line represents the line of best fit in the form of y=b+ax. The red and dark blue dashed lines represent the y=x±1 and the y=x±0.5 respectively.

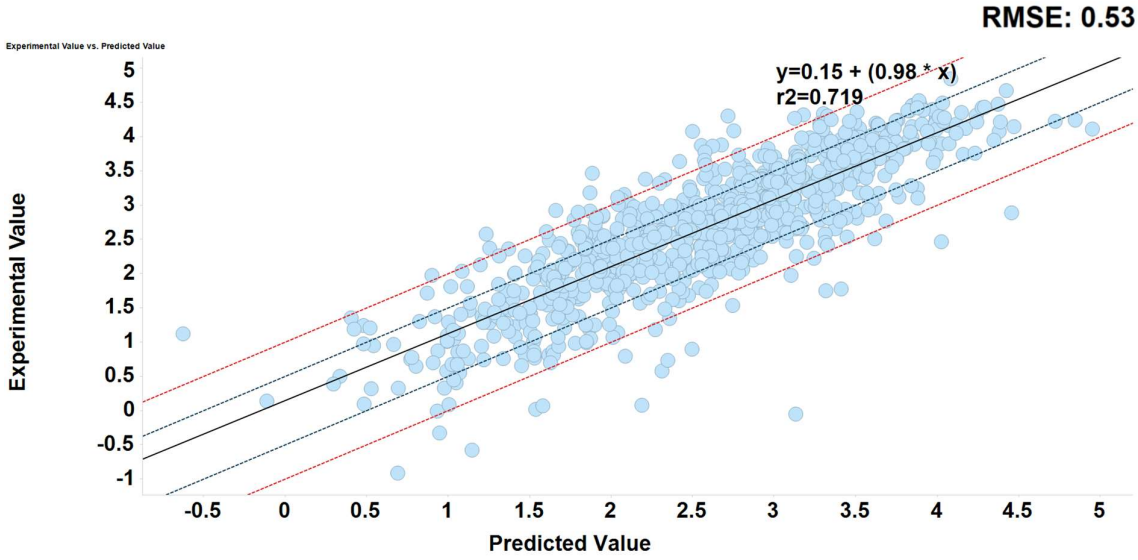

Figure 36: Experimental versus predicted $logD_{7.4}$ values of compounds in the Evotec temporal test set obtained with the Evotec model developed with the SVR algorithm. $LogD_{7.4}$ lipophilicity is reported as Log_{10} D. The black solid line represents the line of best fit in the form of y=b+ax. The red and dark blue dashed lines represent the y=x±1 and the y=x±0.5 respectively.

When the SVR Evotec model was applied on the literature diverse and temporal test sets the R^2 was equal to 0.84 and 0.72 and the RMSE equal to 0.49 and 0.53 respectively (figures 35, 36). The compounds in the diverse test set (figure 35) showed a better correlation between the experimental and predicted $LogD_{7.4}$ values (R^2=0.84) and a lower error in prediction (RMSE=0.49) compared to the temporal test set (R^2=0.72, RMSE=0.53). However, R^2 should be considered cautiously because its value may be increased by addition of data in a narrow range of values. The diverse test set included a greater number of compounds compared to the temporal test set and thus the higher R^2 value might not indicating better model performance. The red and dark blue dashed lines enclosed the compounds with predicted $LogD_{7.4}$ values within ±1 and ±0.5 log units respectively from the experimental values. For the diverse test set, the 95.37% and the 89.13% of the predicted $LogD_{7.4}$ values were within ±1 and ±0.5 log units from the experimental values. For the temporal test set, the 94.30% and the 83.35% of the predicted $LogD_{7.4}$ values were within ±1 and ±0.5 log units from the experimental values. Therefore, a smaller percentage of temporal test set compounds had prediction values within 1 and 0.5 log units from the experimental values compared to the diverse test set. The reason might be that the compounds in temporal test set were novel or far from the model's chemical space and the model produced predictions with a higher error in prediction. Therefore, the compounds in the temporal test sets might not have been represented with compounds in the training set as it might have happened with the compounds in the diverse test set, which were randomly selected from the initial dataset.

3.3.3 Models developed with proprietary and literature data (Evotec+ChEMBL models)

The third models reported herein were developed using both Evotec proprietary and literature data extracted from the ChEMBL database. These models are referred as "Evotec+ChEMBL models" and were based on the previous training sets that were used in the ChEMBL and Evotec models previously. The two previous training sets were merged resulting in 9609 compounds in total. Three different modelling algorithms were applied to build the QSPR models: random forest (RF), partial least square (PLS) and supporting vector regression (SVR).

Two different strategies were used to define and evaluate the goodness of a model. In one case, all the 9609 compounds have been used to build the QSPR models and the test set (temporal) was derived by merging the two previous (Evotec and ChEMBL) temporal test sets resulting in 981 compounds. In the second case, the 9609 compounds were merged with the 981 compounds of the temporal test set and the diverse test set was built including 20% of the total number of compounds randomly selected, while the remaining 80% of the compounds have been part of the training set used to build the model.

Table 25: RMSE in prediction and R^2 of Evotec+ChEMBL diverse test set and Evotec+ChEMBL temporal test set using different machine learning methods (RF, PLS &SVR). The red colour indicates the model that produced the lower RMSE in each testing strategy.

Model/Training set		**Evotec+ChEMBL diverse test set**			**Evotec+ChEMBL temporal test set**		
		RF	**PLS**	**SVR**	**RF**	**PLS**	**SVR**
Evotec+ ChEMBL	RMSE	0.62	0.78	**0.55**	**0.58**	0.69	**0.58**
	R^2	0.81	0.70	0.84	0.71	0.59	0.71

The results of the model assessment indicated that for both test sets (temporal or diverse), the RF and SVR algorithms provided better performing predictive $LogD_{7.4}$ models than PLS (table 25). This is an observation also reported in the literature and has discussed in section 3.3.1 and 3.3.2. In the case of diverse test set, the SVR algorithm performed better (RMSE=0.55) compared to the RF (RMSE=0.62). In addition, the correlation between experimental and predicted $LogD_{7.4}$ values of the diverse test set, was higher when the SVR (R^2=0.84) applied, compared to RF (R^2=0.84). In the case of the temporal test set, the two algorithms (RF and SVR) performed identically (RMSE=0.58, R^2=0.71). Both test sets were predicted with a high error in prediction and therefore the reliability of prediction by Evotec+ChEMBL models is questionable and the results of prediction should be used with caution.

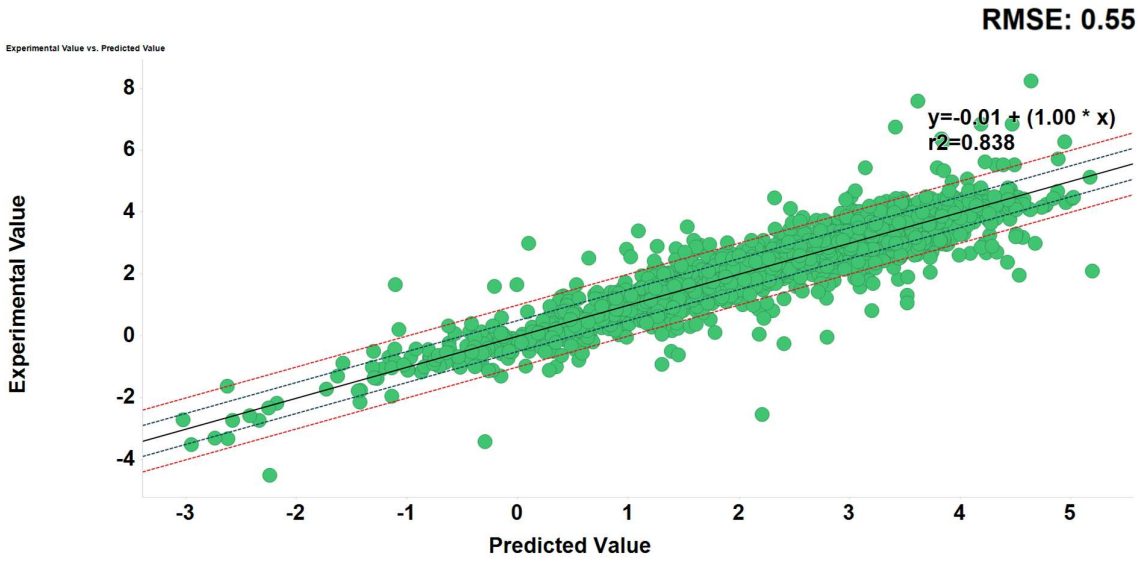

Figure 37: Experimental versus predicted $logD_{7.4}$ lipophilicity of compounds in the Evotec+ChEMBL diverse test set obtained with the Evotec+ChEMBL model developed with the SVR algorithm. $LogD_{7.4}$ lipophilicity is reported as Log_{10} D. The black solid line represents the line of best fit in the form of y=b+ax. The red and dark blue dashed lines represent the y=x±1 and the y=x±0.5 respectively.

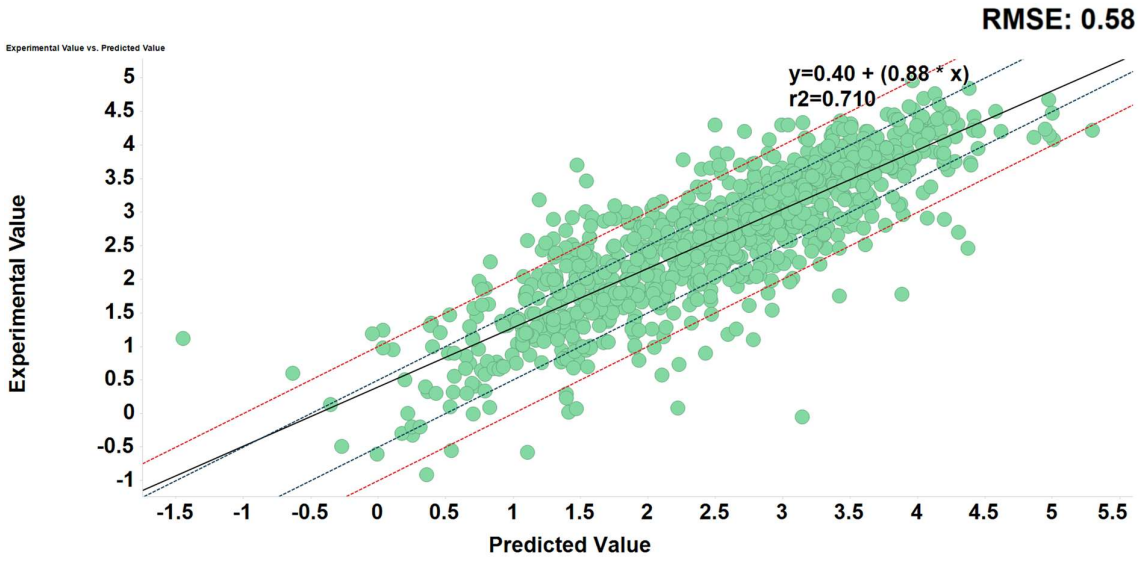

Figure 38: Experimental versus predicted $LogD_{7.4}$ lipophilicity of compounds in the Evotec+ChEMBL temporal test set obtained with the Evotec+ChEMBL model developed with the SVR algorithm. $LogD_{7.4}$ lipophilicity is reported as Log_{10} D. The black solid line represents the line of best fit in the form of y=b+ax. The red and dark blue dashed lines represent the y=x±1 and the y=x±0.5 respectively.

Although RF and SVR performed identically for the temporal test set, the performance of models for the two different test sets was compared to the overall best performing algorithm,

which is SVR. The SVR was the best performing algorithm for the diverse test set and one of the two identically best performing algorithms for the temporal test set. When the SVR Evotec+ChEMBL model was applied on the diverse and temporal test sets the R^2 was equal to 0.84 and 0.71 and the RMSE equal to 0.55 and 0.58 respectively (figures 37, 38). Therefore, the compounds in the diverse test set showed a better correlation between the experimental and predicted values and a lower RMSE in prediction compared to the temporal test set. However, R^2 should be considered cautiously because its value may be increased by addition of data in a narrow range of values. The diverse test set included a greater number of compounds compared to the temporal test set and thus the higher R^2 value might not indicating better model performance. The red and dark blue dashed lines represent the compounds with predicted $LogD_{7.4}$ within ±1 and ±0.5 log units respectively from the experimental values. For the diverse test set, the 93.96% and the 87.77% of the predicted $LogD_{7.4}$ values were within ±1 and ±0.5 log units respectively from the experimental values. For the temporal test set, the 91.44% and the 80.73% of the predicted $LogD_{7.4}$ values were within ±1 and ±0.5 log units from the experimental values. Therefore, a smaller percentage of temporal test set compounds had prediction values from the experimental values within ±1 and ±0.5 log units. The reason might be that the compounds in temporal test set were novel or far from the model's chemical space and the model produced predictions with a higher error in prediction. Therefore, the compounds in the temporal test sets might not have been represented with compounds in the training set as it might have happened with the compounds in the diverse test set, which were randomly selected from the initial dataset.

3.3.4 Comparison of $LogD_{7.4}$ models with models reported in the literature

The goal of this part of the work was to compare the models developed in the present study in sections 3.3.1-3.3.3 (i.e. ChEMBL, Evotec and Evotec+ChEMBL models) with models reported in the literature. Being $LogD_{7.4}$ a widely used parameter for pre-evaluation of compounds in drug discovery, predictive $LogD_{7.4}$ models have been published in the literature and are reported in table 26. These models developed with proprietary data in pharmaceutical companies (Bruneau & McElroy, 2006; Rodgers et al., 2011; Schroeter et al., 2007) and with data extracted from the literature (Wang et al., 2015).

Table 26: Regression lipophilicity models developed with $logD_{7.4}$ data reported in the literature.

Reference	Method	Number of Molecules	Number of Descriptors	AD Estimation? (Yes/No)
(Bruneau & McElroy, 2006)	1. BRNN	5000	122	Yes Mahalanodis Distance
(Schroeter et al., 2007)	1. Gaussian Process 2. Linear ridge regression 3. SVR 4. RF	14556	Dragon descriptors (1664)	Yes Mahalanodis Distance
(Rodgers et al., 2011)	1. SVR 2. PLS	Number of molecules varied as the models were updated over a period of 3 years	In-house descriptor set (topological, geometrical and electronic)	Yes Mahalanodis Distance
(Wang et al, 2015)	3. RF 4. PLS	1130	121	Yes Leverage

One of the initial attempts to develop lipophilicity models based on $logD_{7.4}$ data and the BRNN (Bayesian Regularised Neural Networks) algorithm, was conducted in AstraZeneca (Bruneau & McElroy, 2006). In this study a set of 8200 of AstraZeneca "in house" compounds was clustered based on hierarchical clustering process. After clustering, 5000 clusters were generated and one compound from each cluster was selected to form the training set; the rest of the compounds had been used as "ex-cluster validation" test set. In addition, a "global validation" test set, comprised by 16325 compounds, was obtained from the AstraZeneca database. The advantage was that the model was developed with a consistent and large proprietary dataset. Model seemed to perform well for both test sets with an RMSE in prediction of 0.54 ("ex-cluster validation") and 0.63 ("global validation"). The proprietary Evotec model developed in this study performed slightly better in predicting temporal test set compounds (RMSE=0.49). The Evotec+ChEMBL performed similarly with AstraZeneca models by predicting temporal test set compounds with an RMSE equal to 0.58. However, the AstraZeneca models were developed only with proprietary compounds and it will be interesting

to investigate how data extracted from the literature could affect the performance and AD of the proprietary models.

In Bayer Schering Pharma AG (Schroeter et al., 2007), $LogD_{7.4}$ models were developed using four different algorithms: Gaussian Process (GA), 2. Linear Ridge Regression (LRR), 3.SVR and 4. RF. The training set, identical for all the four models, included 14556 proprietary compounds. The models were then assessed with a literature temporal test set and proprietary temporal test set. The literature temporal test set was better predicted by the GA and SVR models, with an RMSE of 0.66 and 0.71 respectively. The Evotec model predicted the literature temporal test set with an RMSE equal to 0.83 (result obtained with SVR algorithm). On the other hand, when Schroeter et al (2007) assessed the proprietary temporal test set, they obtained an RMSE equal to 0.81 and 0.82 for the SVR and GA method respectively. The Evotec and Evotec+ChEMBL models were better in predicting proprietary temporal test set compounds with an RMSE of 0.53 and 0.58 respectively (result obtained with SVR algorithm). Moreover, the ChEMBL model obtained an RMSE equal to 0.77 which is very similar to that obtained by Schroeter et al (2007).

Furthermore, two $LogD_{7.4}$ models developed with RF and PLS algorithms (Rodgers et al., 2011) in AstraZeneca. These models were constantly updated and assessed, on a monthly basis, when new data became available over a period of 3 years. The initial model contained AstraZeneca proprietary compounds in the training set and this model, when trained with the RF algorithm, predicted the 1st temporal test set and the last temporal test set with an RMSE of 0.53 and 0.67. A final model was obtained after the 3 years of constant update and evaluation. This final model was able to predict the final temporal test set with an RMSE of 0.57. The Evotec model was able to predict the Evotec temporal test set with an RMSE of 0.53, thus indicating that the predictive activity of the proprietary model in the present study (Rodgers et al., 2011) is very similar and slightly better.

Finally, Wang et al (2015) developed $LogD_{7.4}$ models using data obtained from ChEMBL and ochem.eu database, resulting in a training set of 1130 compounds. These compounds were partitioned into training set (80% of compounds) and diverse test set (20% of compounds) with the Kennard-Stone method. This algorithm works by selecting the compounds so that they are divided evenly throughout the descriptor space of the original set of compounds (Martin et al., 2012). This rational method of splitting the compounds ensured that the test set was representative of the training set. The algorithms that used in that study were the RF and SVR. The models trained with SVR and PLS resulted in an RMSE in prediction of 0.56 and 0.69 respectively. The Evotec and Evotec+ChEMBL models were better in predicting proprietary compounds with an RMSE equal to 0.49 and 0.55 respectively. However, the models developed by Wang et al (2015) were better performing than the ChEMBL models reported in the present study. A direct comparison of the RMSE and R^2 would be not accurate, as different training and test sets were used.

In order to make a direct comparison, the methodology developed in the present study, was applied on Wang et al (2015) training set and then the models derived, were assessed with

the Wang et al (2015) diverse test set. Therefore, the training compounds of Wang et al (2015) were trained by using the set of descriptors and algorithms used in the present study. The two best performing algorithms were applied: RF and SVR, and the results of the RMSE in prediction are shown in table 27.

Table 27: Model assessment results of SVR models developed by Wang et al (2015) and of the new models developed with Wang et al (2015) training and test set and the present study's methodology.

	Method	Literature ChEMBL diverse test set by Wang et al (2015)	
		RMSE	R^2
Literature ChEMBL by Wang et al (2015)	SVR	0.56	0.89
Literature ChEMBL model by Wang et al (2015) **developed with present study's methodology**	SVR	0.55	0.90
	RF	0.62	0.87

Table 27 shows that the methodology (algorithms) developed in the present study, when applied to the literature training and test sets of Wang et al (2015) provided comparable results. The RMSE and R^2 of the two models (table 27) developed with the SVR algorithm were almost identical. The RMSE in prediction of the models reported in table 27 was smaller than the RMSE in prediction of the ChEMBL model reported in section 3.3.1. A possible explanation is that Wang et al (2015) used a representative test set, which was obtained from the initial compound dataset with the Kennard-Stone algorithm. A test set, which is representative of the training set can give a good model performance but at the same time could be unrealistic.

In addition, there are various available commercial software that can be used for the theoretical calculation of $LogD_{7.4}$ like the ChemAxon software. ChemAxon was used to calculate the $LogD_{7.4}$ of the Evotec and ChEMBL diverse and temporal test sets, which were used in the present study. The root mean square error (RMSE) of the predictions and the R^2 of the predicted versus experimental have been calculated for the test sets and are used to evaluate the goodness of the ChemAxon software (table 28). In addition, the $LogD_{7.4}$ predictions of the ChEMBL and Evotec temporal and diverse test sets obtained by the Evotec, ChEMBL and Evotec+ChEMBL, are reported in table 28.

Table 28: RMSE in prediction and R^2 of ChEMBL and Evotec diverse test sets and ChEMBL and Evotec temporal test set. Results obtained by using the ChemAxon software, and the ChEMBL, Evotec and Evotec+Chembl models developed with the SVR algorithm.

Test sets	ChEMBL diverse test set	Evotec diverse test set	ChEMBL temporal test set	Evotec temporal test set
Model	**ChemAxon**			
RMSE	1.42	1.14	1.01	0.86
R^2	0.61	0.43	0.66	0.52
Model	**ChEMBL model**			
RMSE	0.94	1.14	0.77	0.89
R^2	0.82	0.36	0.71	0.51
Model	**Evotec model**			
RMSE	1.64	0.49	0.83	0.53
R^2	0.63	0.84	0.66	0.72
Model	**Evotec+ChEMBL model**			
RMSE	0.95	0.47	0.77	0.56
R^2	0.82	0.85	0.72	0.70

The diverse and temporal test sets were better predicted by the models developed in the present study compared to the ChemAxon. The RMSE in prediction for the ChEMBL diverse test set and the ChEMBL temporal test set was 1.42 and 1.01 respectively when the ChemAxon software used. When the ChEMBL model used, the RMSE in prediction for the ChEMBL diverse test set and the ChEMBL temporal test set decreased to 0.94 and 0.77 respectively. Similarly, the correlation between the experimental and predicted values was higher for both ChEMBL diverse and temporal test sets, when the ChEMBL model used. The RMSE in prediction for the Evotec diverse test set and Evotec temporal test set was 1.14 and 0.86 respectively when the ChemAxon software used. When the Evotec+ChEMBl model used the RMSE for the Evotec diverse test set and Evotec temporal test set was 0.47 and 0.56 respectively. These results indicated that the Chemaxon software cannot predict $LogD_{7.4}$ as

accurately as the models developed in the present study. Moreover, another disadvantage of the ChemAxon is that it does not provide a measure to estimate the AD of the training set and thus the distance of the test set compounds from it. In the present study, the AD of the models was investigated (section 3.3.6).

Moreover, Chemaxon is a software that predicts the $LogD_{7.4}$ with fragment based methods and thus the predictions rely on the quality of the fragments in a large extend (Wang et al, 2015). In the study conducted by Wang et al (2015), the performance of the $LogD_{7.4}$ models developed with ChEMBL data was compared with the Chemaxon and Discovery studio software. The results indicated that the model developed with the SVR algorithm was better predicting the $LogD_{7.4}$ of the compounds in the test set. Two possible explanations, as outlined by Wang et al (2015), are: the advantage of chemical diversity that a literature test set offers and the fact that the calculation of LogD by ChemAxon is based on the pka and LogP values.

However, none of the studies outlined above investigated the effect of literature data in the proprietary models. Therefore, the effect of the literature data in the model performance of proprietary models and in the applicability domain is outlined in sections 3.3.5 and 3.3.6 respectively.

3.3.5 The effect of merging proprietary and literature data in the development of $LogD_{7.4}$ models

The three models ("ChEMBL", "Evotec" and "Evotec+ChEMBL" models), which were developed and described in sections 3.3.1-3.3.3, were used to evaluate the effect of the introduction of literature compounds in the proprietary models by testing both Evotec and Evotec+ChEMBL models on the same test sets. The test sets that were used are the Evotec and ChEMBL temporal and diverse test sets, which have been also outlined in sections 3.3.1-3.3.3 The three models were trained with three different algorithms (RF, PLS and SVR) and were tested on the same diverse and temporal test sets and the results are outlined in tables 29 and 30 respectively.

The models predicted the diverse and temporal test sets with a large error in prediction (tables 29, 30) and as a result the reliability of predictions is questionable and the results of prediction should be used with caution. There are various reasons, which can negatively affect the performance of the models and have been outlined in section 3.2.5.

Table 29: Table shows the model performance of “ChEMBL”, “Evotec” and “Evotec+ChEMBL” models. The RMSE in prediction and R^2 of Evotec and ChEMBL diverse test sets are reported. Results obtained by applying the RF, PLS and SVR algorithms. The red colour indicates the highest performing model between the Evotec and Evotec+ChEMBL models.

Number of compounds	Model/Training set		Evotec diverse test set			ChEMBL diverse test set		
			RF	PLS	SVR	RF	PLS	SVR
1036	ChEMBL	RMSE	1.09	1.54	1.14	1.01	1.34	0.94
		R^2	0.44	0.22	0.36	0.79	0.65	0.82
7436	Evotec	RMSE	0.60	**0.63**	0.49	1.56	1.43	1.34
		R^2	0.77	0.72	0.84	0.59	0.57	0.63
8472	Evotec+ ChEMBL	RMSE	**0.53**	0.69	**0.47**	**1.15**	**1.22**	**0.95**
		R^2	0.81	0.68	0.85	0.76	0.70	0.82

Table 30: Table shows the model performance of “ChEMBL”, “Evotec” and “Evotec+ChEMBL” models. The RMSE in prediction and R^2 of Evotec and ChEMBL temporal test sets are reported. Results obtained by applying the RF, PLS and SVR algorithms. The red colour indicates the highest performing model between the Evotec and Evotec+ChEMBL models.

Number of compounds	Model/Training set		Evotec temporal test set			ChEMBL temporal test set		
			RF	PLS	SVR	RF	PLS	SVR
1209	ChEMBL	RMSE	0.84	1.30	0.89	0.84	1.09	0.77
		R^2	0.58	0.34	0.51	0.72	0.58	0.71
8400	Evotec	RMSE	0.62	**0.68**	**0.53**	0.97	0.92	0.83
		R^2	0.66	0.55	0.72	0.54	0.58	0.66
9609	Evotec+ ChEMBL	RMSE	**0.58**	**0.68**	0.56	**0.86**	**0.78**	**0.77**
		R^2	0.68	0.56	0.70	0.64	0.69	0.72

When the proprietary diverse test set used to assess the models (table 29), the Evotec+ChEMBL model (trained with the RF and SVR algorithm) showed an improvement in the predictions with respect to the Evotec model. The Evotec+ChEMBL model predicted better the proprietary diverse test set with an RMSE equal to 0.53 and 0.47 (when the RF and SVR algorithm used) compared to the Evotec model, which showed and RMSE equal to 0.60 and 0.49. In addition, the Evotec+ChEMBL models showed a better correlation of experimental with predicted values compared to the Evotec model. However, the same result was not observed with the PLS algorithm. When the literature diverse test set used to assess the models, the Evotec+ChEMBL model could better predict the $LogD_{7.4}$ compared to the Evotec model. The Evotec+ChEMBL models showed a lower RMSE in predictions and a higher R^2 (i.e. improved relationship between the experimental and predicted values) for all the algorithm tested. This indicated that the Evotec+ChEMBL model can provide a large improvement in the predictions for the literature diverse test set; this evidence has been found for all the algorithms tested.

Similar results and observations obtained with the temporal test sets (table 30). The Evotec+ChEMBL model predicted better the proprietary temporal test set with an RMSE equal to 0.58 when the RF used compared to the Evotec model which showed an RMSE equal to 0.62. However, the same result was not observed with the SVR and PLS algorithm. When the ChEMBL temporal test set used to assess the models, the Evotec+ChEMBL model could better predict the $LodD_{7.4}$ compared to the Evotec model. The Evotec+ChEMBL models showed a lower RMSE in predictions and a higher R^2 (i.e. improved relationship between the experimental and predicted values) for all the algorithm tested.

Temporal compounds extracted from literature (ChEMBL temporal test set) could theoretically mimic novel proprietary chemotypes and series from completely new projects or novel chemical matter in existing projects. These results provided an indication that the compounds extracted from literature can enhance the predictive ability of the proprietary models when they assess newly synthesised and chemically diverse compounds.

The merging of the compounds from different sources (proprietary and literature) seemed to be beneficial. The results and the performance of the merged model (Evotec+ChEMBL models) balanced the experimental uncertainty of the data. The Evotec +ChEMBL models exhibited a similar performance with the Evotec models in the prediction of proprietary temporal test sets and showed a significant improvement for the prediction of literature test compounds. Similar results observed with the development and evaluation of Caco-2 permeability models. This provides a confidence in the conclusion that the literature data can have a positive effect on the performance of proprietary ADME models.

3.3.6 Applicability Domain estimation of the *in-silico* $LogD_{7.4}$ models

Determining the AD for a QSPR model is important to estimate the reliability of a prediction of an external compound. If the compound lies within the AD of the QSPR model used to predict a property, this prediction can be taken, otherwise the prediction should be either discarded or given a low reliability flag.

The AD of the models was estimated with four distance to model metrics, which were the: 1. k-NN with Euclidean, 2. k-NN with Manhattan, 3. Leverage and 4. Mahalanobis distance. The distance to model metrics calculated the distance of the test compounds from the training set in the descriptors' space (i.e. the multi-dimensional space defined by the descriptors of the compounds used to train the model) and a threshold was applied. Above that threshold, compounds were considered to be outside the AD.

The goal of this section was twofold. Firstly, the AD of the Evotec+ChEMBL model was compared with the AD of the Evotec model. Secondly, once an AD distance threshold had been determined, the goal was to check whether test set compounds within the AD are actually predicted more accurately than compounds outside the AD. To do that, compounds in the test set were partitioned in two groups; within the AD, and outside the AD. If compounds within the AD show an RMSE in the prediction smaller than compounds outside the AD, the particular distance metric is able to clearly define an AD for the model. In addition, a Mann Whitney test was used to establish the presence of a statistically significant difference in the RMSE of the compounds inside and outside of the AD. To achieve both objectives, for every model ("ChEMBL model", "Evotec model" and "Evotec+ChEMBL") the portion of compounds within and outside the AD was calculated by using all the four distance metrics mentioned above. The distance of compounds in the two temporal test sets ("Evotec temporal test set" and "ChEMBL temporal test set") was calculated from the training compounds of the three different models ("ChEMBL model", "Evotec model" and "Evotec+ChEMBL" model) and the percentage of the test compounds within the AD of the models was calculated. Results are reported in table 31.

Table 31: Results obtained with the kNN with Euclidean distance, kNN with Manhattan distance, Leverage and Mahalanobis distance for the three different $LogD_{7.4}$ models and the two different temporal test sets. The table summarises: the percentage of compounds inside the AD of the models, the RMSE in prediction of compounds inside and outside the AD and the assessment of the statistical significance with the Mann Whitney (MW) test. The red colour indicates the presence of a statistically significant difference in the RMSE of the compounds inside and outside of the AD.

		kNN/ Euclidean				kNN/ Manhattan				Leverage				Mahalanobis			
		In		Out	MW	In		Out	MW	In		Out	MW	In		Out	MW
		%	RMSE	RMSE	test	%	RMSE	RMSE	test	%	RMSE	RMSE	test	%	RMSE	RMSE	test
Model	**Evotec+ChEMBL**																
Evotec temporal test set		82.91	0.505	0.688	**p<0.05**	80.67	0.503	0.677	**p<0.05**	94.86	0.534	0.659	**p<0.05**	80.22	0.507	0.660	**p<0.05**
ChEMBL temporal test set		39.53	0.934	0.747	p>0.05	38.37	0.882	0.789	p>0.05	93.02	0.801	1.102	**p<0.05**	45.35	0.841	0.813	p>0.05
Model	**Evotec**																
Evotec temporal test set		76.98	0.499	0.630	**p<0.05**	79.30	0.505	0.598	**p<0.05**	88.49	0.520	0.613	**p<0.05**	74.08	0.497	0.621	**p<0.05**
ChEMBL temporal test set		37.21	0.912	0.785	p>0.05	32.56	0.831	0.836	p>0.05	91.86	0.805	1.108	**p<0.05**	33.72	0.871	0.815	p>0.05
Model	**ChEMBL**																
Evotec temporal test set		82.57	0.857	1.025	**p<0.05**	74.30	0.839	1.017	**p<0.05**	84.58	0.828	1.164	**p<0.05**	59.89	0.818	0.984	**p<0.05**
ChEMBL temporal test set		81.40	0.746	0.842	p>0.05	81.40	0.752	0.819	p>0.05	86.05	0.737	0.919	**p<0.05**	47.67	0.741	0.786	p>0.05

From table 31, it can be observed that, in most cases, the RMSE in prediction of the compounds within the AD was lower than the RMSE of the compounds outside the AD. This evidence, provided confidence for using these distance metrics and threshold determination method as a reliable protocol for defining the AD of the models. Similar studies are in agreement with these findings; (Jaworska et al., 2005; Sahigara et al., 2012) as they employed in their work, the same distance metrics and reached the same conclusions. Moreover, this phenomenon was observed in studies, where the AD of lipophilicity models was investigated. In more detail, in two different studies, one conducted with AstraZeneca proprietary compounds (Rodgers et al., 2011) and another conducted with Bayer Schering Pharma data (Schroeter et al., 2007), the distance of test compounds from the training set was evaluated with the Mahalanobis Distance. The test compounds with low Mahalanobis distance were closer to the model space and had a lower RMSE in prediction.

In addition, table 31 indicated that a greater percentage of Evotec temporal test compounds and ChEMBL temporal test compounds were within the Evotec+Chembl model's AD compared to the Evotec proprietary model. Therefore, all the four distance to model metrics demonstrated that the AD of the Evotec models was enlarged with the inclusion of compounds extracted from the literature. This indicated that the literature compounds can introduce chemical diversity and cover unexplored areas of the chemical space (Kogej et al., 2013).

However, there are cases, where the compounds outside the AD showed an RMSE lower than the compounds inside the AD. This phenomenon has been discussed in the present study in section 3.2.7, where various possible explanations were outlined. The same phenomenon was also observed in a study, where $LogD_{7.4}$ models have been developed with data extracted from the literature and the leverage method was used as the only metric to establish AD. In that case, there were compounds with low leverage that had a larger error in prediction and *vice versa* (Wang et al., 2016). The main reason of that phenomenon lied at the way that the threshold was defined. Generally, a threshold has a value, which simply excludes training compounds in the extremities. That means that the AD evaluation only considers the interpolation and not the possible extrapolation. Therefore, there might have been compounds in the test sets that were close to that outliers and thus they were predicted with a lower error in prediction than expected.

In addition, the AD of a $LogD_{7.4}$ model was evaluated over a period of 3 years (Rodgers et al., 2011). The $LogD_{7.4}$ models developed with RF and PLS and the training set was updated monthly. The Mahalanobis distance has been reported as the average Mahalanobis Distance of each test compounds to its 3-nearest neighbour in the training set. The final temporal test set was examined with the initial model and with the final model (updated over a period of three years). The test compounds predicted by the initial model with an RMSE of 0.67 and the average Mahalanobis Distance of the test compounds was 3.35. On the other hand, the final model produced an RMSE of 0.57 and the test compounds showed an average Mahalanobis Distance of 2.43. As a result, the inclusion of new proprietary compounds improved the model performance and enlarged its AD. This was also observed in the present study with the

inclusion of the literature data into the proprietary models. Therefore, the addition of new compounds could be beneficial for the AD and the performance of the models.

Another interesting finding regarding the distance to model metrics is that each distance to model metric produce different results. This was observed in the evaluation of the AD of the permeability models in section 3.2.7. The reason is possibly based on the fact that each distance to model metric calculated the distance of the test compounds differently.

A statistical analysis used to understand which method is more reliable and the non-parametric Mann Whitney test was applied. The Leverage method could produce a statistically significant difference in the RMSE for the compounds inside and outside of the AD. For the other three methods, there was not always a statistically significant difference. Therefore, the Leverage method could be considered as the most effective method for these models. Leverage is a method extensively used in the literature for the estimation of the AD of permeability models (Wang et al, 2016), $LogD_{7.4}$ models (Wang et al, 2015).

In conclusion, in this case, leverage was considered as the best performing method for two reasons. The first reason was that it was able to categorise the test compounds in and out of the domain with a lower and higher RMSE in prediction. The second reason was the presence of a statistically significant difference in all the measurements. However, if a compound is outside the AD, it is not definite that the prediction is erroneous but it provides an AD “warning” (Gupta et al., 2016).

4 CONCLUSION AND FUTURE WORK

The aim of this section is to summarise the overall conclusions of this study and to provide an overview of future work ideas that could be implemented in the development of ADME predictive models.

4.1 Conclusions

To address the effect of the inclusion of literature data in proprietary ADME QSPR models, new ADME models were built to include public data in their training set. Current models in Evotec make use of only proprietary data in the development of their ADME predictive models. There are also other companies like AstraZeneca, which have incorporated public data in their Caco-2 permeability models but the effect of the public data on the models was not investigated. Therefore, the aim of this study was to investigate whether the merging of literature and proprietary data could improve the predictive activity of proprietary models and enlarge their applicability domain. In order to achieve this aim, three specific objectives were devised to perform this study.

The first objective was to evaluate the ability of the existing Evotec Caco2 permeability model to predict the permeability of literature compounds and to investigate different distance to model metrics for the evaluation of the AD. A large dataset of Caco-2 permeability data was downloaded from the ChEMBL database to perform this task. The initial results showed that the literature/ChEMBL test set was predicted with a higher RMSE compared to the RMSE in prediction for the internal compounds. In addition, the AD of the existing Evotec permeability models was evaluated with four distance to model metrics: kNN with Euclidean distance, kNN with Manhattan distance, Leverage and Mahalanobis distance. The distance of the test compounds (compounds downloaded from ChEMBL) from the Evotec training set was calculated in both descriptor and chemical (fingerprint) space. The test compounds were binned in five equally populated bins, by increasing distance, and the RMSE of each bin was calculated. A weak trend was observed between distance and the RMSE of the predictions; the RMSE was increasing as the distance was increasing. The same trend was not observed when the distance between test and training compounds was evaluated in the chemical space. A possible explanation for that could rely in the fact that caco-2 permeability is a property greatly influenced by the physiochemical properties of the compounds rather than form its structure; different functional groups might show similar physiochemical properties thus translating in a similar contribution to the overall Caco2 permeability of the molecule. Since the first approach produced a weak trend between the distance of the compounds and the RMSE, a second approach was used. The goal of the second approach was to understand whether a distance threshold could be applied to distinguish between compounds within and outside the AD with a difference in their RMSE. The Mann Whitney test was also applied to identify if the difference in the RMSE of the compounds within and outside the AD is statistically significant. The compounds within the AD had a lower RMSE than the compounds outside the AD and that difference was statistically significant. A significant amount of literature compounds has been found to be outside the AD of the Evotec model, thus highlighting an area for the

improvement of proprietary Evotec models and providing a rationale for an effort aimed at incorporating public data into the proprietary Evotec model to produced improved ADME models.

The second objective was to develop new Caco-2 permeability models (referred as Evotec+ChEMBL models), which incorporate both proprietary and literature data and to evaluate their performance and AD compared to proprietary only Evotec models. In total three models were built based on three different training sets: Evotec proprietary compounds, literature compounds (downloaded from ChEMBL) and a merged set of Evotec proprietary and literature (Evotec+ChEMBL) compounds. Three different methods were used for developing QSPR models: Partial Least Squares (PLS), Random Forest (RF) and Support Vector Regression (SVR) with a radial basis function (rbf) kernel. The performance of the models was evaluated by using two types of test sets: a diverse test set (20 % compounds of available data randomly selected) and a temporal test set (data published after the models were build). In addition, four distance to model metrics used to assess the AD of the all built models by estimating the distance of test set compounds from the training set using: k-NN (k-Nearest Neighbour) with Euclidean distance (ED), k-NN with Manhattan distance (ManhD), Leverage and Mahalanobis distance (MD). The results suggested that the RF was the method of choice for developing permeability models for two reasons. The first reason was that RF is easy-implemented and is very time effective and the second was that it was able to provide a low error in the prediction of the test compounds.

A comparison of the Evotec+ChEMBL model with the existing Evotec Caco-2 showed that the inclusion of public data could be highly beneficial and could improve both the model performance and enlarged its applicability domain. The permeability model built merging literature and proprietary data predicted a temporal literature test set with an RMSE of 0.68 while the Evotec model showed an RMSE of 0.74. Similarly, the same model predicted an Evotec proprietary temporal test set with an RMSE of 0.55 while the Evotec model shower and RMSE of 0.57. Even in the case of the Evotec temporal test set, the two models performed similarly but the AD of the mixed models (incorporating both literature and proprietary data) was enlarged. The 86.15% of the compounds in the test set were within the AD of the mixed model, while 76.50% of the compounds of the same test set appeared to be within the AD of the Evotec model.

Subsequently the same protocol was applied for the third objective, which was to develop new $LogD_{7.4}$ models (referred as Evotec+ChEMBL models), which incorporated both proprietary and literature data and to evaluate their model performance and AD compared to proprietary only Evotec models. The same algorithms, distance metrics and test set strategy, used in the case of the Caco-2 permeability model, were applied. A comparison of the Evotec+ChEMBL model with the existing Evotec $LogD_{7.4}$ showed that the inclusion of public data could be beneficial and could improve both the model performance and enlarge its applicability domain. The SVR was the best performing algorithm for the lipophilicity models by providing the lowest error in prediction for the most of the cases. The SVR $LogD_{7.4}$ model built merging literature and proprietary data predicted a temporal literature test set with an RMSE of 0.77 while the

Evotec model showed an RMSE of 0.83. However, the new model predicted an Evotec proprietary temporal test set with an RMSE of 0.56 while the Evotec model shower and RMSE of 0.53. In that case, the RMSE in prediction is very similar. Even in the case of the Evotec temporal test set, the two models performed similarly but the AD of the mixed models (incorporating both literature and proprietary data) was enlarged. The 94.86% of the compounds in the test set fell within the applicability domain of the mixed model, while 88.49% of the compounds of the same test set appeared to be within the applicability domain of the Evotec model.

In conclusion, the aim of this study, which focused on investigating the effect of the introduction of public data into the proprietary ADME models, has been achieved. The inclusion of public data into proprietary data improved the performance of proprietary models and enlarged, at the same time, their AD. These observations underline the importance of the inclusion of public data in the proprietary ADME models and thus the methodology presented herein will be applied by Evotec computational scientists to re-build the proprietary Caco-2 and $LogD_{7.4}$ Evotec models. Additionally, in this study, three modelling algorithms (RF, PLS and SVR) have been used for model building and each method gave different results. The RF algorithm was the highest performing algorithm for the development of Caco-2 permeability. However, the SVR algorithm provided the best $LogD_{7.4}$ models. In Evotec, the existing method to train proprietary $LogD_{7.4}$ models, is the RF but the SVR lowered the RMSE in predictions up to 0.1 log units lower than the RF model, in the prediction of Evotec temporal test set compounds. This difference in the RMSE is not negligible and highlights the fact the different algorithms need to be assessed to find an optimal modelling approach for a particular data-set. Moreover, four distance to model metrics were assessed in the evaluation of the AD. In most of the cases the metrics used, could identify compounds inside and outside the AD with a smaller RMSE for the compounds inside. However, there were some cases were the RMSE for the compounds inside the AD was higher than the compounds outside and several reasons could be responsible for that and have been outlined. All the methods performed differently and only the Leverage method was able to distinguish between compounds inside and outside the AD with a statistically significant difference in the RMSE of the predictions. Therefore, in this study, the Leverage proved to be the most appropriate method but it is always good to use more than one metric to evaluate the AD of the model.

In conclusion, this study demonstrates that the inclusion of public data into proprietary data can improve the performance of proprietary models and enlarge at the same time their applicability domain. The work and methodology presented herein is of great value also for Evotec. Concepts and methods of this work will be implemented at Evotec by computational scientists for future ADME model building. These are: i) the inclusion of ChEMBL compounds in the proprietary training sets, ii) the addition of the SVR as an algorithm to build ADME models and iii) the implementation of a procedure for the evaluation of the AD. In the light of this actions, the recommendations for future work are outlined in the following section.

4.2 Future work

The inclusion of public data in the Evotec Caco-2 permeability and $LogD_{7.4}$ models was beneficial in terms of models performance and AD and thus it will be interesting to include public data in other ADME models. Two important ADME models during drug discovery pipeline are the microsomal stability and plasma protein binding. The plasma protein binding influences the distribution of a compound into the body's tissues. A drug with a high plasma protein binding value exhibits a decreased amount of free compound available to reach the biological target and also a slower metabolism. Therefore, plasma protein binding is a property that affects all the ADME properties and a good predictive model can benefit pharmaceutical companies. In the same aspect, the microsomal stability is equally important because can be used for the evaluation of the hepatic metabolism. Metabolism is the primary cause of failure or success of a compound (Ulenberg, Belka, Król, & Herold, 2015) as it affects their clearance (CL), half-life ($t_{1/2}$) and oral bioavailability (Di, Keefer, Scott, Strelevitz, & Chang, 2012). These parameters in turn, influence the concentration of the drug within the plasma and tissues of the body and consequently affect the efficacy and the toxicology of the drug (Cyprotex, 2015). In addition, metabolism is a difficult parameter to predict as it is influenced by the binding of the compounds with the metabolic enzyme. Therefore, public data could potentially improve the performance and AD of two important and challenging to develop ADME models.

Additionally, in this study, three modelling algorithms (RF, PLS and SVR) have been used for model building and each method gave different results. Therefore, it is important to assess different number of algorithms to find the most appropriate for each case. For example, the SVR proved to be better for the development of $LogD_{7.4}$ models compared to the RF. By taking into consideration that there are several good modelling algorithms, other could also be applied for the model building. Two interesting suggestions are the Bayesian Regularised Neural Networks (BRNN) and Boosting algorithm. The BRNN has been used for the development of $LogD_{7.4}$ models in AstraZeneca (Rodgers et al, 2011) and the boosting for caco-2 permeability models (Wang et al, 2015). In both cases, these algorithms were producing accurate predictions and it will be interesting to investigate their use with the model developed in Evotec.

Moreover, at Evotec, models are updated on a monthly basis. In this study, only a subsequent analysis was performed for the Caco- 2 permeability model with a proprietary temporal test (including compounds tested four months after the compounds in the training set) and a literature test set containing compounds from the new ChEMBL version. Results showed that the updated Evotec+ChEMBL model was better in predicting new temporal and literature temporal test sets compared to Evotec model and also was better than the initial Evotec+ChEMBL model. Therefore, it will be beneficial to regularly update the models with public data (when a new ChEMBL version is available) and proprietary data (monthly) and assess their performance and applicability domain. In addition, in terms of incorporating new public compounds from ChEMBL into the Evotec models an efficient and time effective method should be developed as the data curation of the compounds is very time consuming. For that reason, the data curation process was performed with a KNIME workflow and this workflow

can be used in the future, to filter the compounds extracted from ChEMBL database for the caco-2 permeability and $LogD_{7.4}$ lipophilicity models.

In the present study, temporal test sets and diverse test sets by randomly selecting the compounds were used. It will be interesting to add another strategy of selecting the compounds based on a rational method of splitting the initial dataset into training and test set. Some examples of rational portioning in training and test set are the Kennard Stone algorithm and the sphere exclusion method. These methods can ensure that the compounds in the test set are representative of the compounds in the training set. However, this approach might not be representative of a realistic drug discovery case scenario, but it is interesting to evaluate how the model can perform with a test set, which is representative of the training set. Finally, it will be also interesting to evaluate the presence of outliers and how these outliers affect the models' predictions.

5 REFERENCES

Abdi, H. & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459.

Al-Shamri, M. Y. H. (2014). Power coefficient as a similarity measure for memory-based collaborative recommender systems. *Expert Systems with Applications*, *41*(13), 5680–5688.

Alexander, D. & Tropsha, A. (2015). Beware of R 2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *Journal of Chemical Information and Modelling*, *55*(7), 1316–1322.

Allerton, C., Smith, D. A., Kalgutkar, Amit, S., van de Waterbeemd, H. & Walker, D. K. (2012). *Pharmacokinetics and Metabolism in Drug Design*. Weinheim: John Wiley & Sons.

Alpaydin, E. (2014). *Introduction to machine learning*. (3rd ed.). Cambridge, Massachussets: MIT Press

Andrés, A., Rosés, M., Ràfols, C., Bosch, E., Espinosa, S., Segarra, V. & Huerta, J. M. (2015). Setup and validation of shake-flask procedures for the determination of partition coefficients (logD) from low drug amounts. *European Journal of Pharmaceutical Sciences*, *76*, 181–191.

Aniceto, N., Freitas, A., Bender, A. & Ghafourian, T. (2016). A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: reliability-density neighbourhood. *Journal of cheminformatics*, *8*(1), 69.

Anton, H. (2010). *Elementary linear algebra.* (10th ed.). Hoboken: John Wiley & Sons.

Artursson, P., Palm, K. & Luthman, K. (2012). Caco-2 monolayers in experimental and theoretical predictions of drug transport. *Advanced Drug Delivery Reviews*, *64*, 280–289.

Avdeef, A. & Tam, K. Y. (2010). How Well Can the Caco-2/Madin−Darby Canine Kidney Models Predict Effective Human Jejunal Permeability? *Journal of Medicinal Chemistry*, *53*(9), 3566–3584.

Avdeef, A. & Tsinman, O. (2006). PAMPA—A drug absorption *in vitro* model. *European Journal of Pharmaceutical Sciences*, *1*(28), 43–50.

Bajorath, J. (Ed.) (2004). *Chemoinformatics: concepts, methods, and tools for drug discovery (Vol. 275).* New Jersey: Humana Press.

Bajorath, J., Vogt, M., Duffy, B. C., Zhu, L., Decornez, H. & Kitchen, D. B. (2012). Chemoinformatics in Drug Discovery. *Bioorganic & Medicinal Chemistry*, *20*(18), 5305–5315.

Baka, E., Comer, J. E. A., & Takács-Novák, K. (2008). Study of equilibrium solubility measurement by saturation shake-flask method using hydrochlorothiazide as model compound. *Journal of Pharmaceutical and Biomedical Analysis*, *46*(2), 335–341.

Basak, D., Pal, S., & Patranabis, D. (2007). Support vector regression. *Neural Information Processing- Letters and Reviews, 11*(10), 203–224.

Bender, A. (2010). Databases: Compound bioactivities go public. *Nature Chemical Biology*, *6*(5), 309–309.

Benet, L. Z. (2013). The Role of BCS (Biopharmaceutics Classification System) and BDDCS (Biopharmaceutics Drug Disposition Classification System) in Drug Development. *Journal of Pharmaceutical Sciences*, *102*(1), 34–42.

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Kruger, F.A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos,R. & Overington, J. P. (2014). The ChEMBL bioactivity database: an update. *Nucleic Acids Research*, *42*(Database issue), D1083-90.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kotter, T., Meinl, T., Ohl, P., Thiel, K & Wiswedel, B. (2009). KNIME -The Konstanz Information Miner. *AcM SIGKDD Explorations Newsletter*, *11*(1), 26–31.

Brown, N. (2015). *In Silico Medicinal Chemistry: Computational Methods to Support Drug Design.* Cambridge: Royal Society of Chemistry

Brownlee, J. (2016). *Master Machine Learning Algorithms: Discover how They Work and Implement Them from Scratch*. Retrieved March, 27, 2017, from https://bookshelflab.com/uploads/?filename=%2Fvar%2Fwww%2Fthebookshelf%2Fthebookshelf%2Fuploads%2FMaster_Machine_Learning_Algorithms.pdf.

Bruneau, P. (2001). Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets. *Journal of chemical information and computer sciences*, *41*(6), 1605–1616.

Bruneau, P. & McElroy, N. (2006). logD 7.4 modeling using Bayesian regularized neural networks. Assessment and correction of the errors of prediction. *Journal of Chemical Information and Modeling*, *46*(3), 1379–1387.

Caldwell, G., Yan, Z., Tang, W., Dasgupta, M. & Hasting, B. (2009). ADME Optimization and Toxicity Assessment in Early- and Late-Phase Drug Discovery. *Current Topics in Medicinal Chemistry*, *9*(11), 965–980.

Cao, D., Liang, Y., Xu, Q., Hu, Q. & Zhang, L. (2011). Exploring nonlinear relationships in chemical data using kernel-based methods. *Chemometrics and Intelligent Laboratory Systems*, *107*(1), 106–115.

Caron, G. & Ermondi, G. (2008). Lipophilicity: Chemical Nature and Biological Relevance. In R. Manhold (Eds.), *Molecular Drug Properties: Measurement and Prediction* (pp. 313–329). Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA.

Caron, G. & Ermondi, G. (2016). Molecular descriptors for polarity: the need for going beyond polar surface area. *Future Medicinal Chemistry*, *8*(17), 2013–2016.

Chackalamannil, S., Rotella, D. & Ward, S. (Eds.) (2017). *Comprehensive Medicinal Chemistry III* (3rd ed.). Oxford: Elsevier.

Chai, T. & Draxler, R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247–1250.

Chang, C. & Lin, C. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and technology (TIST), 2*(3), 1–27.

Chau, C., Rixe, O., McLeod, H., & Figg, W. (2008). Validation of analytic methods for biomarkers used in drug development. *Clinical Cancer Research*, *14*(19), 5967–5976.

ChemAxon. (2016a). *Standardizer – chemical business rules processing « ChemAxon – Software for Chemistry and Biology.* Retrieved July, 2, 2017, from https://www.chemaxon.com/products/standardizer/.

ChemAxon. (2016b). *Workflow tools « ChemAxon – Software for Chemistry and Biology.* Retrieved January, 4, 2017, from https://www.chemaxon.com/products/workflow-tools/.

ChEMBL. (2017). *ChEMBL*. Retrieved January, 4, 2017, from https://www.ebi.ac.uk/chembl/.

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C., Todescini, R., Consonni, V., Kuz'mins V.E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A & Tropsha, A. (2014). QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry*, *57*(12), 4977–5010.

Clark, A. M., Dole, K., Coulon-Spektor, A., McNutt, A., Grass, G., Freundlich, J. S., Reynolds,R.C. & Ekins, S. (2015). Open Source Bayesian Models. 1. Application to ADME/Tox and Drug Discovery Datasets. *Journal of Chemical Information and Modeling*, *55*(6), 1231–1245.

Cramer, R. D. (2012). The inevitable QSAR renaissance. *Journal of Computer-aided molecular design*, *26*(1), 35–38.

Cronin, M. T. D. & Schultz, T. W. (2003). Pitfalls in QSAR. *Journal of Molecular Structure: THEOCHEM*, *622*(1–2), 39–51.

Cumming, J. G., Davis, A. M., Muresan, S., Haeberlein, M. & Chen, H. (2013). Chemical predictive modelling to improve compound quality. *Nature Reviews Drug Discovery*, *12*(12), 948–962.

Cyprotex. (2015). *Everything you need to know about ADME* (2nd ed.). Macclesfield: Cyprotex.

Danielle, A. N. (2014). *Data Mining Methods For the Prediction of Intestinal Absorption Using QSAR*. Doctoral dissertation, University of Kent and University of Greenwich, Kent.

Davies, M., Dedman, N., Hersey, A., Papadatos, G., Hall, M. D., Cucurull-Sanchez, L., Jeffrey, P., Hasan, S., Eddershaw, P.J. & Overington, J. P. (2015). ADME SARfari: comparative genomics of drug metabolizing systems. *Bioinformatics (Oxford, England)*, *31*(10),1695–7.

Davis, A., & Ward, S. E. (Eds.). (2014). *The handbook of medicinal chemistry: principles and practice*. Cambridge: Royal Society of Chemistry.

De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*(1), 1–18.

Dehmer, M., Varmuza, K. & Bonchev, D. (Eds.) (2012). *Statistical modelling of molecular descriptors in QSAR/QSPR*. Weinheim: Wiley-Blackwell.

Di, L., Keefer, C., Scott, D., Strelevitz, T. & Chang, G. (2012). Mechanistic insights from comparing intrinsic clearance values between human liver microsomes and hepatocytes to guide drug design. *European Journal of medicinal chemistry*, *57*, 441–448.

Di, L. & Kerns, E. H. (2016). *Drug-like properties: concepts, structure design and methods from ADME to toxicity optimization*. (2nd ed.). London: Academic Press.

Doddareddy, M. R., Klaasse, E. C., IJzerman, A. P. & Bender, A. (2010). Prospective validation of a comprehensive *in silico* hERG model and its applications to commercial compound and drug databases. *ChemMedChem*, *5*(5), 716–729.

Dong, X., Jiang, C., Hu, H., Yan, J., Chen, J. & Hu, Y. (2009). QSAR study of Akt/protein kinase B (PKB) inhibitors using support vector machine. *European Journal of medicinal chemistry*, *44*(10), 4090–4097.

Dragos, H., Gilles, M. & Alexandre, V. (2009). Predicting the predictability: A unified approach to the applicability domain problem of qsar models. *Journal of Chemical Information and Modeling*, *49*(7), 1762–1776.

Dudek, A. Z., Arodz, T. & Gálvez, J. (2006). Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review. *Combinatorial Chemistry & High Throughput Screening*, *9*(3), 213–228.

Ehrhardt, C. & Kim, K. (2008). *Drug absorption studies : in situ, in vitro and in silico models*. New York : Springer.

Eklund, M., Norinder, U., Boyer, S. & Carlsson, L. (2014). Choosing Feature Selection and Learning Algorithms in QSAR. *Journal of Chemical Information and Modeling*, *54*(3), 837–843.

Fenza, A. Di, Alagona, G., Ghio, C. & Leonardi, R. (2007). Caco-2 cell permeability modelling: a neural network coupled genetic algorithm approach. *Journal of Computer-aided molecular design*, *21*(4), pp.207–221.

Fourches, D., Muratov, E. & Tropsha, A. (2010). Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *Journal of Chemical Information and Modeling*, *50*(7), 1189–1204.

Fredlund, L., Winiwarter, S. & Hilgendorf, C. (2017). *In Vitro* Intrinsic Permeability: A Transporter-Independent Measure of Caco-2 Cell Permeability in Drug Design and Development. *Molecular Pharmaceutics*, *14*(5), 1601–1609.

Fujiwara, S., Yamashita, F. & Hashida, M. (2002). Prediction of Caco-2 cell permeability using a combination of MO-calculation and neural network. *International Journal of pharmaceutics*, *237*(1), 95–105.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. (2nd ed.). London:

Academic press.

Furusjö, E., Svenson, A., Rahmberg, M., & Andersson, M. (2006). The importance of outlier detection and training set selection for reliable environmental QSAR predictions. *Chemosphere*, *63*(1), 99–108.

Gadaleta, D., Mangiatordi, G. F., Catto, M., Carotti, A. & Nicolotti, O. (2016). Applicability domain for QSAR models: where theory meets reality. *International Journal of Quantitative Structure-Property Relationships (IJQSPR)*, *1*(1), pp.45–63.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey S., Michalovich, D., Al-Lazikani, B. & Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, *40*(Database issue), D1100–7.

Gavaghan, C., Arnby, C. & Blomberg, N. (2007). Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data. *Journal of Computer-aided molecular design, 21*(4), pp.189–206.

Geppert, H., Vogt, M. & Bajorath, J. (2010). Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *Journal of Chemical Information and Modelling*, *50*(2), 206–216.

Goñi, F. (2014). The basic structure and dynamics of cell membranes: an update of the Singer–Nicolson model. *Biochimica et Biophysica Acta (BBA)-Biomembranes, 1838*(6), 1467–1476.

Goodarzi, M. & Dejaegher, B. (2012). Feature selection methods in QSAR studies. *Journal of AOAC International*, *95*(3), 636-651.

Gozalbes, R. & Doucet, J. (2002). Application of topological descriptors in QSAR and drug design: history and new trends. *Current Drug Targets-Infectious Disorders*, *2*(1), 93–102.

Guangli, M. & Yiyu, C. (2006). Predicting Caco-2 permeability using support vector machine and chemistry development kit. *J. Pharm. Pharm. Sci*, *9*, 210–221.

Gupta, J. K., Adams, D. J. & Berry, N. G. (2016). Will it gel? Successful computational prediction of peptide gelators using physicochemical properties and molecular fingerprints. *Chemical Science*, *7*(7), 4713–4719.

Harding, A. & Wedge, D. (2009). pKa Prediction from “Quantum Chemical Topology” Descriptors. *Journal of Chemical Information and Modeling*, *49*(8), 1914–1924.

Hartmann, T. & Schmitt, J. (2004). Lipophilicity – beyond octanol/water: a short comparison of modern technologies. *Drug Discovery Today: Technologies*, *1*(4), 431–439.

Hemmateenejad, B., Miri, R. & Elyasi, M. (2012). A segmented principal component analysis-regression approach to QSAR study of peptides. *Journal of Theoretical Biology*, *305*, 37–44.

Heshmati, N., Wagner, B., Cheng, X., Scholz, T., Kansy, M., Eisenbrand, G. & Fricker, G.

(2013). Physicochemical characterization and *in vitro* permeation of an indirubin derivative. *European Journal of Pharmaceutical Sciences*, *50*(3–4), 467–475.

Hou, T., Li, Y., Zhang, W. & Wang, J. (2009). Recent Developments of *In Silico* Predictions of Intestinal Absorption and Oral Bioavailability. *Combinatorial Chemistry & High Throughput Screening*, *12*(5), 497–506.

Hou, T., Zhang, W., Xia, K. & Qiao, X. (2004). ADME evaluation in drug discovery. 5. Correlation of Caco-2 permeation with simple molecular properties. *Journal of chemical information and computer sciences*, *44*(5), 1585–1600.

Hughes, J. D., Blagg, J., Price, D. A., Bailey, S., DeCrescenzo, G. A., Devraj, R. V., Ellsworth, E., Fobian, Y.M., Gibbs, M.E., Gilles, R.W., Greene, N., Huang, E., Krieger-Bruke, T., Loesel, J., Wager, T., Whiteley, L. & Zhang, Y. (2008). Physiochemical drug properties associated with *in vivo* toxicological outcomes. *Bioorganic & Medicinal Chemistry Letters*, *18*(17), 4872–4875.

Hughes, L., Palmer, D., Nigsch, F. & Mitchell, J.B. (2008). Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and Log P. *Journal of chemical information and modeling*. *48*(1), 220–232.

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. (2012). ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling*, *52*(7), 1757–1768.

Jagla, B., Wiswedel, B. & Coppée, J.-Y. (2011). Extending KNIME for next-generation sequencing data analysis. *Bioinformatics (Oxford, England)*, *27*(20), 2907–2909.

Jasial, S., Hu, Y., Vogt, M. & Bajorath, J. (2016). Activity-relevant similarity values for fingerprints and implications for similarity searching. *F1000Research*. *5*.

Jaworska, J., Nikolova-Jeliazkova, N. & Aldenberg, T. (2005). QSAR applicability domain estimation by projection of the training set in descriptor space: A review. *ATLA Alternatives to Laboratory Animals*, *33*(5), 445–459.

Jung, S., Choi, S., Um, S., Kim, J. & Choo, H. (2006). Prediction of the permeability of drugs through study on quantitative structure–permeability relationship. Journal o*f pharmaceutical and biomedical analysis, 41*(2), 469–475.

Kalliokoski, T., Kramer, C., Vulpetti, A. & Gedeck, (2013). Comparability of Mixed IC50 Data – A Statistical Analysis. *PLoS ONE*, *8*(4), e61007.

Kaneko, H. & Funatsu, K. (2014). Applicability domain based on ensemble learning in classification and regression analyses. *Journal of chemical information and modeling*, *54*(9), 2469–2482.

Karelson, M., Karelson, G., Tamm, T., Tulp, I., Janes, J., Tamm, K., Lomaka, A., Savchenko, D. & Dobchev, D. (2009). QSAR study of pharmacological permeabilities. *ARKIVOC: Online Journal of Organic Chemistry,* 218-238.

Kell, D., Dobson, P. & Oliver, S. (2011). Pharmaceutical drug transport: the issues and the implications that it is essentially carrier-mediated only. *Drug Discovery Today*, *16*(15), 704–714.

Kell, D. & Oliver, S. (2014). How drugs get into cells: tested and testable predictions to help discriminate between transporter-mediated uptake and lipoidal bilayer diffusion. *Frontiers in Pharmacology*. *5*, 1-32.

Khan, M. T.H. (Ed.) (2012). *Recent Trends on QSAR in the Pharmaeutical Perceptions.* Bentham Science Publishers.

Khanna, I. (2012). Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug Discovery Today*, *17*(19–20),1088–1102.

KNIME. (2016). *KNIME | Open for Innovation.* Retrieved December, 31, 2016, from https://www.knime.org/

Kogej, T., Blomberg, N., Greasley, P. & Mundt, S. (2013). Big pharma screening collections: more of the same or unique libraries? The AstraZeneca–Bayer Pharma AG case. *Drug Discovery Today*, *18*(19), 1014–1024.

Kore, P. P., Mutha, M. M., Antre, R. V, Oswal, R. J. & Kshirsagar, S. S. (2012). Computer-Aided Drug Design: An Innovative Tool for Modeling. *Open Journal of Medicinal Chemistry. 2*, 139–148.

Kovacs, P. (2016). *Class ECFP*. Retrieved September, 27, 2016, from https://www.chemaxon.com/jchem/doc/dev/java/api/chemaxon/descriptors/ECFP.html.

Ku, M. S. (2008). Use of the Biopharmaceutical Classification System in Early Drug Development. *The AAPS Journal, 10*(1), 208–212.

Kubinyi, H., Folkers, G. & Mannhold, R. (2008). *Molecular drug properties: measurement and proprties*. Weinheim: John Wiley & Sons.

Kulkarni, A., Han, Y., & Hopfinger, A. (2002). Predicting Caco-2 cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis. *Journal of chemical information and computer sciences*, *42*(2), 331–342.

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. *Drug Discovery Today, 20*(3), 318–331.

Leach, A. & Gillet, V. (2007). *An introduction to chemoinformatics.* New York: Springer Science & Business Media.

Li, A. P. (2001). Screening for human ADME/Tox drug properties in drug discovery. *Drug Discovery Today*, *6*(7), 357–366.

Li, J., Volpe, D., Wang, Y., Zhang, W., Bode, C., Owen, A. & Hidalgo, I. J. (2011). Use of transporter knockdown Caco-2 cells to investigate the *in vitro* efflux of statin drugs. *Drug Metabolism and Disposition*. *111*, 1–27.

Li, Q., Cheng, T., Wang, Y. & Bryant, S. H. (2010). PubChem as a public resource for drug discovery. *Drug Discovery Today*, *15*(23), 1052–1057.

Li, Q., Jørgensen, F., Oprea, T., Brunak, S. & Taboureau, O. (2008). hERG classification model based on a combination of support vector machine method and GRIND descriptors. *Molecular pharmaceutics*, *5*(1), 117–127.

Liao, C., Sitzmann, M., Pugliese, A. & Nicklaus, M. C. (2011). Software and resources for computational medicinal chemistry. *Future Medicinal Chemistry*, *3*(8), 1057–85.

Liu, P. & Long, W. (2009). Current Mathematical Methods Used in QSAR/QSPR Studies. *International Journal of Molecular Sciences*, *10*(5), 1978–1998.

Livingstone, D. & Davis, A. M. (Eds). (2012). *Drug design strategies : quantitative approaches*. Cambridge: Royal Society of Chemistry.

Lodhi, H. (Ed.). (2010). *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques: Complex Computational Methods and Collaborative Techniques*. Hershey: Medical Information Science Reference.

Low, Y. W. I., Blasco, F. & Vachaspati, P. (2016). Optimised method to estimate octanol water distribution coefficient (logD) in a high throughput format. *European Journal of Pharmaceutical Sciences*, 92, 110–116.

Lu, S., Jessen, B., Strock, C. & Will, Y. (2012). The contribution of physicochemical properties to multiple *in vitro* cytotoxicity endpoints. *Toxicology in Vitro*, *26*(4), 613–620.

Mandagere, A. & Thompson, T. (2002). Graphical model for estimating oral bioavailability of drugs in humans and other species from their Caco-2 permeability and *in vitro* liver enzyme metabolic stability. *Journal of medicinal chemistry, 45*(2), 304-311.

Martin, T. M., Harten, P., Young, D. M., Muratov, E. N., Golbraikh, A., Zhu, H. & Tropsha, A. (2012). Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *Journal of Chemical Information and Modeling*, *52*(10), 2570–2578.

Mathea, M., Klingspohn, W. & Baumann, K. (2016). Chemoinformatic Classification Methods and their Applicability Domain. *Molecular Informatics*, *35*(5), 160–180.

Mazanetz, P.M., Marmon, J.R., Reisser, B.T.C. & Morao, I. (2012). Drug Discovery Applications for KNIME: An Open Source Data Mining Platform. *Current Topics in Medicinal Chemistry*, *12*(18), 1965–1979.

Melagraki, G. & Afantitis, A. (2013). Enalos KNIME nodes: Exploring corrosion inhibition of steel in acidic medium. *Chemometrics and Intelligent Laboratory Systems*, *123*, 9–14.

Melagraki, G., Afantitis, A., Sarimveis, H., & Koutentis, P. (2009). Predictive QSAR workflow for the *in silico* identification and screening of novel HDAC inhibitors. *Molecular diversity, 13*(3), 301–311.

Miller, J. M., & Miller, J. C. (2010). *Statistics and Chemometrics for Analytical Chemistry*. *Technometrics*. (5th ed.). Harlow: Pearson Education.

Mitchell, J. B. O. (2014). Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *4*(5), 468–481.

Moroy, G., Martiny, V. Y., Vayer, P., Villoutreix, B. O. & Miteva, M. A. (2012). Toward *in silico* structure-based ADMET prediction in drug discovery. *Drug Discovery Today*, *17*(1–2), 44–55.

Netzeva, T. I., Worth, A. P., Aldenberg, T., Benigni, R., Mark, T. D., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D.W., Schultz, T.W., Stanton, D.T., Van de Sandt, J.J.M., Tong, W., Veith, G. & Yang, C. (2005). Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. *ATLA*, *33*, pp.155–173.

Nordqvist, A., Nilsson, J., Lindmark, T., Eriksson, A., Garberg, P. & Kihlén, M. (2004). A General Model for Prediction of Caco-2 Cell Permeability. *Molecular Informatics, 23*(5), 303–310.

Norinder, U., Österberg, T. & Artursson, P. (1997). Theoretical calculation and prediction of Caco-2 cell permeability using MolSurf parametrization and PLS statistics. *Pharmaceutical research*, *14*(12), 1786–1791.

Oprea, T. (Ed.). (2006). *Chemoinformatics in Drug Discovery*. Weinheim: Willey VCH.

Paixão, P., Gouveia, L. & Morais, J. (2010). Prediction of the *in vitro* permeability determined in Caco-2 cells by using artificial neural networks. *European Journal of Pharmaceutical Sciences*, *41*(1), 107–117.

Pajouhesh, H. & Lenz, G. R. (2005). Medicinal chemical properties of successful central nervous system drugs. *NeuroRx : The Journal of the American Society for Experimental NeuroTherapeutics*, *2*(4), 541–53.

Palmer, D. S., O'Boyle, N. M., Glen, R. C. & Mitchell, J. B. (2007). Random forest models to predict aqueous solubility. Journal of chemical information and modeling, *47*(1), 150–158.

Papadatos, G., Gaulton, A., Hersey, A., & Overington, J. P. (2015). Activity, assay and target data curation and quality in the ChEMBL database. *Journal of Computer-aided molecular design*, *29*(9), 885–896.

Papadatos, G. & Overington, J. P. (2014). The ChEMBL database: a taster for medicinal chemists. *Future Medicinal Chemistry*, *6*(4), 361–364.

Park, J. H., Carlin, K. P., Wu, G., Ilyin, V. I., Musza, L. L., Blake, P. R. & Kyle, D. J. (2014). Studies Examining the Relationship between the Chemical Structure of Protoxin II and Its Activity on Voltage Gated Sodium Channels. *Journal of Medicinal Chemistry, 57*(15), 6623–6631.

Patrick, G. L. (2013). *An Introduction to Medicinal Chemistry* (5th ed.). Oxford: Oxford University Press.

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R. & Schacht, A. L. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, *9*(3), 203.

Peres-Neto, P. R., Jackson, D. A. & Somers, K. M. (2005). How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, *49*(4), 974–997.

Pham-The, H., Casañola-Martin, G., Garrigues, T., Bermejo, M., González-Álvarez, I.,

Nguyen-Hai, N., Cabrera-Pérez, M. Á. & Le-Thi-Thu, H. (2016). Exploring different strategies for imbalanced ADME data problem: case study on Caco-2 permeability modeling. *Molecular Diversity*, *20*(1), 93–109.

Pillai, G. (2015). *Computational Modelling of Diverse Chemical, Biochemical and Biomedical Properties.* Doctoral Dissertation, University of Tartu, Tartu.

Poulin, P. & Theil, F.-P. (2002). Prediction of Pharmacokinetics Prior to *In Vivo* Studies. II. Generic Physiologically Based Pharmacokinetic Models of Drug Disposition. *Journal of Pharmaceutical Sciencesl, 91*(5),1358–1370.

Press, B. (2011). Optimization of the Caco-2 permeability assay to screen drug compounds for intestinal absorption and efflux. *Permeability Barrier: Methods and Protocols,* 763*,* 139–154.

Puzyn, T., Leszczynski, J. & Cronin, M. T. (Eds.). (2010). *Recent advances in QSAR studies: methods and applications* (Vol. 8). London: Springer Science & Business Media.

Qi, Y. (2012). Random forest for bioinformatics. In C. Zhang, & Y. Ma (Eds.), *Ensemble Machine Learning*. (pp. 307-323). New York: Springer US.

Rajkhowa, S. & Deka, R. (2014). DFT based QSAR/QSPR models in the development of novel anti-tuberculosis drugs targeting Mycobacterium tuberculosis. *Current Pharmaceutical Design*, *20*(27), 4455-4473.

Raschka, S. (2015). *Python machine learning*. Birmingham: Packt Publishing Ltd.

Riley, R. J., Parker, A. J., Trigg, S. & Manners, C. N. (2001). Development of a Generalized, Quantitative Physicochemical Model of CYP3A4 Inhibition for Use in Early Drug Discovery. *Pharmaceutical Research, 18*(5), 652–655.

Ringnér, M. & Ringner, M. (2008). What is principal component analysis? *Nature Biotechnology, 26*(3), 303–304.

Rodgers, S. L., Davis, A. M., Tomkinson, N. P. & van de Waterbeemd, H. (2011). Predictivity of Simulated ADME AutoQSAR Models over Time. *Molecular Informatics*, 30(2–3), 256–266.

Rogers, D., & Hahn, M. (2010). Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, *50(*5), 742–754.

Roy, K., Kar, S. & Das, R. (2015). *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. Amsterdam: Academic press.

Roy, P. P., Kovarich, S., & Gramatica, P. (2011). QSAR model reproducibility and applicability: A case study of rate constants of hydroxyl radical reaction models applied to polybrominated diphenyl ethers and (benzo-) triazoles. *Journal of computational chemistry, 32*(11), 2386–2396.

RStudio (2016). *Open source and enterprise-ready professional software for R.* Retrieved July, 2, 2017, from https://www.rstudio.com/

Sahigara, F., Ballabio, D., Todeschini, R. & Consonni, V. (2013). Defining a novel k-nearest

neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. *Journal of Cheminformatics*, *5*(5), 1–9.

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V. & Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules*, *17*(5), 4791–4810.

Sampson, K. E., Brinker, A., Pratt, J., Venkatraman, N., Xiao, Y., Blasberg, J., Steiner, T., Bourner, M. & Thompson, D. C. (2014). Zinc Finger Nuclease–Mediated Gene Knockout Results in Loss of Transport Activity for P-Glycoprotein, BCRP, and MRP2 in Caco-2 Cells. *Drug Metabolism and Disposition, 43*(2), 199–207.

Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L. & Nolan, G. P. (2010). Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics*, *11*(9), 647–657.

Schroeter, T. S., Schwaighofer, A., Mika, S., Ter Laak, A., Suelzle, D., Ganzer, U., Heinrich, N. & Müller, K. R. (2007). Estimating the domain of applicability for machine learning QSAR models: A study on aqueous solubility of drug discovery molecules. *Journal of Computer-aided molecular design*, *21*(12), 651–664.

Schroeter, T., Schwaighofer, A., Mika, S., Ter Laak, A., Suelzle, D., Ganzer, U., Heinrich, N. & Müller, K. R. (2007). Machine learning models for lipophilicity and their domain of applicability. *Molecular pharmaceutics*, *4*(4), 524-538.

Seiler, K. P., George, G. A., Happ, M. P., Bodycombe, N. E., Carrinski, H. A., Norton, S., Brudz, S., Sullivan, J.P., Muhlich, J., Feerraiolo, M.S.P., Tolliday, N.J., Schreiber, S.L. & Clemons, P. A. (2008). ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Research*, *36*(Database issue), D351-359.

Sethi, N. S. (2012). A Review on Computational Methods in Developing Quantitative Structure-Activity Relationship (Qsar). *International Journal of Drug Research and Technology*, *2*(2), 189–197.

Smith, L. (2002). A tutorial on principal components analysis. *Cornell University, USA, 51*(52), 65.

Statnikov, A., Wang, L. & Aliferis, C. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics, 9*(1), 319.

Tan, J. J., Cong, X. J., Hu, L. M., Wang, C. X., Jia, L. & Liang, X.-J. (2010). Therapeutic strategies underpinning the development of novel techniques for the treatment of HIV infection. *Drug Discovery Today*, *15*(5–6), 186–97.

Tao, L., Zhang, P., Qin, C., Chen, S. Y., Zhang, C., Chen, Z., Zhu, F., Yang, S.Y., Wei, Y.Q. & Chen, Y. Z. (2015). Recent progresses in the exploration of machine learning methods as *in-silico* ADME prediction tools. *Advanced Drug Delivery Reviews*, *86*, 83–100.

Tarcsay, Á., Nyíri, K. & Keserű, G. M. (2012). Impact of Lipophilic Efficiency on Compound Quality. *Journal of Medicinal Chemistry*, *55*(3), 1252–1260.

Testa, B. & Turski, L. (Eds.). (2006). *Virtual ADMET Assessment in Target Selection and Maturation* (Vol. 6). Amsterdam: IOS Press.

Tetko, I. & Poda, G. (2004). Application of ALOGPS 2.1 to predict log D distribution coefficient for Pfizer proprietary compounds. *Journal of medicinal chemistry, 47*(23), 5601–5604.

Tetko, I. V. & Bruneau, P. (2004). Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database. *Journal of Pharmaceutical Sciences*, *93*(12), 3103–3110.

Thomas, S., Brightman, F., Gill, H., Lee, S., & Pufong, B. (2008). Simulation Modelling of Human Intestinal Absorption using Caco-2 Permeability and Kinetic Solubility Data for Early Drug Discovery. *Journal of Pharmaceutical Sciences*, *97*(10), 4557–4574.

Thompson, T. (2000). Early ADME in Support of Drug Discovery: The Role of Metabolic Stability Studies. *Current Drug Metabolism*, *1*(3), 215–241.

TIBCO Spotfire. (2016). *Data Visualization & Analytics Software - TIBCO Spotfire*. Retrieved January, 4, 2017, from http://spotfire.tibco.com/

Todeschini, R., & Consonni, V. (2009). *Molecular descriptors for chemoinformatics, volume 41* (2 volume set). Weinhem: John Wiley & Sons.

Tran, L., Vu, V. & Wang, K. (2013). Sparse random graphs: Eigenvalues and eigenvectors. *Random Structures & Algorithms, 4*2(1), 110-134.

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*, *29*(6–7), 476–488.

Tsaioun, K. (2007). The Emerging Role of ADME in Venture Capital. *Drug Discovery Technologies*, 20–21.

Tsaioun, K. & Kates, S. (2011). *ADMET for medicinal chemists: a practical guide*. Hoboken: John Wiley & Sons.

Ulenberg, S., Belka, M., Król, M. & Herold, F. (2015). Prediction of overall *in vitro* microsomal stability of drug candidates based on molecular modeling and support vector machines. Case study of novel. *PloS One*, *10*(3), e0122772.

Valko, K., Chiarparin, E., Nunhuck, S., Montanari, D., Lipinski, C. A., Lombardo, F., Benet, L. Z. (2012). *In vitro* measurement of drug efficiency index to aid early lead optimization. *Journal of Pharmaceutical Sciences*, *101*(11), 4155–4169.

van Breemen, R. B. & Li, Y. (2005). Caco-2 cell permeability assays to measure drug absorption. *Expert Opinion on Drug Metabolism & Toxicology*, *1*(2), 175–185.

Varmuza, K. & Filzmoser, P. (2016). *Introduction to multivariate statistical analysis in chemometrics.* Boca Raton: CRC Press.

Venkatapathy, R., & Wang, N. C. Y. (2013). Developmental Toxicity Prediction. *Computational Toxicology,* 3, 305–340.

Ventura, C., Latino, D. A. R. S. & Martins, F. (2013). Comparison of Multiple Linear Regressions and Neural Networks based QSAR models for the design of new antitubercular compounds. *European Journal of Medicinal Chemistry*, *70*, 831–845.

Volpe, D. A. (2008). Variability in Caco-2 and MDCK Cell-Based Intestinal Permeability Assays. *Journal of Pharmaceutical Sciences*, *97*(2), 712–725.

Wale, N., Watson, I. A. & Karypis, G. (2008). Comparison of descriptor spaces for chemical compound retrieval and classification. *Knowledge and Information Systems*, *14*(3), 347–375.

Wang, J., Cao, D., Zhu, M. & Yun, Y. (2015). *In silico* evaluation of logD7. 4 and comparison with other prediction methods. *Journal of Chemometrics, 29*(2), 389–398.

Wang, J. & Hou, T. (2015). Advances in computationally modeling human oral bioavailability. *Advanced Drug Delivery Reviews*, *86*, 11–16.

Wang, J. & Urban, L. (2004). The impact of early ADME profiling on drug discovery and development strategy. *Drug Discovery World*, *5*(4), 73–86.

Wang, N.-N., Dong, J., Deng, Y.-H., Zhu, M.-F., Wen, M., Yao, Z.-J., Lu, A.P., Wang, J.P. & Cao, D.-S. (2016). ADME Properties Evaluation in Drug Discovery: Prediction of Caco-2 Cell Permeability Using a Combination of NSGA-II and Boosting. *Journal of Chemical Information and Modeling*, *56*(4), 763–773.

Wang, Y., Bolton, E., Dracheva, S., Karapetyan, K., Shoemaker, B. A., Suzek, T. O., Wang, J., Xiao, J., Zhang, J. & Bryant, S. H. (2010). An overview of the PubChem BioAssay resource. *Nucleic Acids Research*, *38*(Database issue), D255-66.

Wang, Z., Kim, S., Quinney, S. K., Guo, Y., Hall, S. D., Rocha, L. M. & Li, L. (2009). Literature mining on pharmacokinetics numerical data: A feasibility study. *Journal of Biomedical Informatics*, 42(4), 726–735.

Warr, W. A. (2012). Scientific workflow systems: Pipeline Pilot and KNIME. *Journal of Computer-aided molecular design*, *26*(7), 801–804.

Weaver, S. & Gleeson, M. P. (2008). The importance of the domain of applicability in QSAR modeling. *Journal of Molecular Graphics and Modelling*, *26*(8), 1315–1326.

Wenlock, M. C., & Carlsson, L. A. (2014). How experimental errors influence drug metabolism and pharmacokinetic QSAR/QSPR models. *Journal of Chemical Information and Modeling, 55*(1), 125-134.

Wermuth, C. (Ed). (2008). *The practice of medicinal chemistry*. (3rd ed.) London: Academic Press.

Will, Y., McDuffie, J., Olaharski, A. & Jeffy, B. (2016). *Drug Discovery Toxicology: From Target Assessment to Translational Biomarkers*. New Jersey: John Wiley & Sons.

Xing, J. J., Luo, R. M., Guo, H. L., Li, Y. Q., Fu, H. Y., Yang, T. M. & Zhou, Y. P. (2014). Radial basis function network-based transformation for nonlinear partial least-squares as optimized by particle swarm optimization: application to QSAR studies. *Chemometrics and Intelligent Laboratory Systems, 130*, 37-44.

Yamashita, F., Wanchana, S. & Hashida, M. (2002). Quantitative structure/property relationship analysis of Caco-2 permeability using a genetic algorithm-based partial least squares method. *Journal of pharmaceutical Sciences, 91*(10), 2230–2239.

Yang, P., Hwa Yang, Y., B Zhou, B. & Y Zomaya, A. (2010). A review of ensemble methods in bioinformatics. *Current Bioinformatics, 5*(4), 296–308.

Yeagle, P. (2011). *The structure of biological membranes.* (3rd ed.) Boca Raton: CRC Press.

Yee, L. C. & Wei, Y. C. (2012). Current Modeling Methods Used in QSAR/QSPR. *Statistical Modelling of Molecular Descriptors in QSAR/QSPR, 2*, 1–31.

Young, D., Martin, T., Venkatapathy, R. & Harten, P. (2008). Are the Chemical Structures in Your QSAR Correct? *QSAR & Combinatorial Science*, *27*(11–12), 1337–1345.

Yousefinejad, S., Bagheri, M. & Moosavi-Movahedi, A. A. (2015). Quantitative sequence–activity modeling of antimicrobial hexapeptides using a segmented principal component strategy: an approach to describe and predict activities of peptide drugs containing l/d and unnatural residues. *Amino Acids*, *47*(1), 125–134.

Zhang, D., Luo, G., Ding, X. & Lu, C. (2012). Preclinical experimental models of drug metabolism and disposition in drug discovery and development. *Acta Pharmaceutica Sinica B*, *2*(6), 549–561.

Zhang, D. & Surapaneni, S. (Eds.) (2012). *ADME-Enabling Technologies in Drug Design and Development*. New Jersey: John Wiley & Sons, Inc.

6 Appendix

Table S1: Java code used to calculate electric state with the “Java Snippet (simple)” node in KNIME.

KNIME node	Java code
Java Snippet (simple)	String state="na"; if (($neg_ionazible_groups$ == 0) && ($pos_ionazible_groups$ == 0)) {state = "neutral";} else if ($neg_ionazible_groups$ == 1 && $pos_ionazible_groups$ == 0) {state = "acid";} else if ($neg_ionazible_groups$ > 1 && $pos_ionazible_groups$ == 0) {state = "acid";} else if ($neg_ionazible_groups$ == 0 && $pos_ionazible_groups$ == 1) {state = "base";} else if ($neg_ionazible_groups$ == 0 && $pos_ionazible_groups$ > 1) {state = "base";} else if ($neg_ionazible_groups$ == $pos_ionazible_groups$ && $pos_ionazible_groups$ >= 1) {state = "zwitterion";} else if ($formal charge$ >= 1) {state = "base";} else if ($formal charge$ <= -1) {state = "acid";} else if ($formal charge$ == 0) {state = "zwitterion";} return state;

Table S2: list of R code that has been used.

Task	R-Code		R-Package
Continuous Random Forest (RF)	R-Learner library(randomForest) knime.model <- randomForest(Y ~ ., ntree=500, data = knime.in)	R-Predictor library(randomForest) levels(knime.in$electric_state) <- c("acid", "base", "neutral", "zwitterion") prediction <- as.data.frame(predict(knime.model, knime.in)) colnames(prediction) <- "prediction" knime.out <- as.data.frame(cbind(knime.in, prediction)) #knime.out <- as.data.frame(cbind(knime.in, predict(knime.model, knime.in)))	randomForest
Partial Least Squares (PLS)	model_rmsep <- c() numSamples <- 100 for (i in 1:numSamples) {model <- plsr(Y ~ ., data=knime.in, validation="CV", ncomp=40)	levels(knime.in$electric_state) <- c("acid", "base", "neutral", "zwitterion") knime.out <- knime.in knime.out$prediction <- predict(knime.model, ncomp = x, newdata=knime.in) Note: where x is the number of significant components defined	pls

	current_rmsep <- RMSEP(model)$val[2, ,] model_rmsep <- cbind(model_rmsep, current_rmsep) } model_rmsep rmsep_mean <- rowMeans(model_rmsep) rmsep_sem <- apply(model_rmsep, 1, sd)/sqrt(numSamples) rmsep_mean min(rmsep_mean) highest_performing_model <- which(rmsep_mean == min(rmsep_mean)) highest_performing_model z_values <- (rmsep_mean-rmsep_mean[[highest_performing_model]])/sqrt(rmsep_sem^2+rmsep_sem[[highest_performing_model]]^2) z_values minimal_model_component_count <-	by the learner and used to train the model.	

	min(which(z_values<= qnorm(.95)))-1 minimal_model_component_count newmod <- plsr(Y ~ ., data=knime.in, validation="CV", ncomp=minimal_model_component_count) summary(newmod) knime.model <- plsr(Y ~ ., data=knime.in, validation="CV", ncomp=minimal_model_component_count)		
Support Vector Regression (SVR)	model<- svm(Y~ ., data=data, kernel='radial') tc<-tune.control(cross="5") tuneResult<tune(svm, Y~ ., data=newdata, ranges=list(epsilon=seq(0,1, by=0.05), cost=list(1,2,5,10,15,20,25,30,40,50,60,70,80,90,100,125, 150,175,200,250,300, 400,500,750,1000,1250,1500,1800), gamma=list(0.0001,0.0002,0.0004,0.0006, 0.0008,0.001,0.002,0.004,0.006,0.008,0.01,0.02,0.04,0.06 ,0.08,0.1,0.2,0.4,0.6, 0.8,1.0), tunecontrol=tc, best.model=TRUE)		e1071
Principal Component Analysis (PCA)	knime.out <- knime.in arc.pca1 <-princomp(knime.in, cor=TRUE, scores=TRUE) summary (arc.pca1) print(arc.pca1)		

	arc.pca1$scores loadings(arc.pca1) knime.out <- data.frame(arc.pca1$scores) knime.model <-princomp(knime.in, cor=TRUE, scores=TRUE)	
Mahalanobis Distance (MD)	knime.out <- knime.in colMeans (knime.in) Sx<-cov(knime.in) D2<-mahalanobis(knime.in,colMeans(knime.in),cov(knime.in)) mean(D2) D2 knime.out<-data.frame(D2)	stats

Table S3: Correlation of the Average Distance bins for the Euclidean distance. Distance was calculated with the descriptors.

Row ID	Avg_1_distance [Binned]	Avg_3_distance [Binned]	Avg_5_distance [Binned]	Avg_10_distance [Binned]	Avg_20_distance [Binned]	Avg_30_distance [Binned]
Avg_1_distance [Binned]	1	0.968	0.957	0.944	0.931	0.922
Avg_3_distance [Binned]	0.968	1	0.988	0.969	0.954	0.946
Avg_5_distance [Binned]	0.957	0.988	1	0.979	0.963	0.955
Avg_10_distance [Binned]	0.944	0.969	0.979	1	0.983	0.975
Avg_20_distance [Binned]	0.931	0.954	0.963	0.983	1	0.991
Avg_30_distance [Binned]	0.922	0.946	0.955	0.975	0.991	1

Table S4: Correlation of the Average Distance bins for the Euclidean distance. Distance was calculated with the first 27 Principal Components.

Row ID	Avg_1_distance [Binned]	Avg_3_distance [Binned]	Avg_5_distance [Binned]	Avg_10_distance [Binned]	Avg_20_distance [Binned]	Avg_30_distance [Binned]
Avg_1_distance [Binned]	1	0.958	0.947	0.933	0.92	0.914
Avg_3_distance [Binned]	0.958	1	0.986	0.97	0.953	0.946
Avg_5_distance [Binned]	0.947	0.986	1	0.982	0.964	0.956
Avg_10_distance [Binned]	0.933	0.97	0.982	1	0.981	0.972
Avg_20_distance [Binned]	0.92	0.953	0.964	0.981	1	0.989
Avg_30_distance [Binned]	0.914	0.946	0.956	0.972	0.989	1

Table S5: Correlation of the Average distance bins for the Manhattan distance. Distance was calculated with the descriptors.

Row ID	Avg_1_distance [Binned]	Avg_3_distance [Binned]	Avg_5_distance [Binned]	Avg_10_distance [Binned]	Avg_20_distance [Binned]	Avg_30_distance [Binned]
Avg_1_distance [Binned]	1	0.967	0.956	0.943	0.927	0.919
Avg_3_distance [Binned]	0.967	1	0.984	0.969	0.953	0.944
Avg_5_distance [Binned]	0.956	0.984	1	0.983	0.966	0.956
Avg_10_distance [Binned]	0.943	0.969	0.983	1	0.981	0.97
Avg_20_distance [Binned]	0.927	0.953	0.966	0.981	1	0.989
Avg_30_distance [Binned]	0.919	0.944	0.956	0.97	0.989	1

Table S6: Correlation of the Average Distance bins for the Manhattan distance. Distance was calculated with the first 27 Principal Components.

Row ID	Avg_1_distance [Binned]	Avg_3_distance [Binned]	Avg_5_distance [Binned]	Avg_10_distance [Binned]	Avg_20_distance [Binned]	Avg_30_distance [Binned]
Avg_1_distance [Binned]	1	0.96	0.946	0.932	0.921	0.914
Avg_3_distance [Binned]	0.96	1	0.984	0.967	0.953	0.945
Avg_5_distance [Binned]	0.946	0.984	1	0.98	0.965	0.957
Avg_10_distance [Binned]	0.932	0.967	0.98	1	0.982	0.973
Avg_20_distance [Binned]	0.921	0.953	0.965	0.982	1	0.99
Avg_30_distance [Binned]	0.914	0.945	0.957	0.973	0.99	1

Table S7: Correlation of the Average Distance bins for the Tanimoto coefficient in chemical space.

Row ID	Avg_1_similarity [Binned]	Avg_3_Similarity [Binned]	Avg_5_Similarity [Binned]	Avg_10_Similarity [Binned]	Avg_20_Similarity [Binned]	Avg_30_Similarity [Binned]
Avg_1_similarity [Binned]	1	0.921	0.884	0.844	0.8	0.783
Avg_3_Similarity [Binned]	0.921	1	0.965	0.929	0.889	0.871
Avg_5_Similarity [Binned]	0.884	0.965	1	0.964	0.926	0.908
Avg_10_Similarity [Binned]	0.844	0.929	0.964	1	0.966	0.949
Avg_20_Similarity [Binned]	0.8	0.889	0.926	0.966	1	0.983
Avg_30_Similarity [Binned]	0.783	0.871	0.908	0.949	0.983	1

Table S8: Correlation of the Average Distance bins for the Dice coefficient in chemical space.

Row ID	Avg_1_Dice [Binned]	Avg_3_distance [Binned]	Avg_5_distance [Binned]	Avg_10_distance [Binned]	Avg_20_distance [Binned]	Avg_30_distance [Binned]
Avg_1_Dice [Binned]	1	0.921	0.881	0.842	0.797	0.779
Avg_3_distance [Binned]	0.921	1	0.964	0.927	0.888	0.869
Avg_5_distance [Binned]	0.881	0.964	1	0.965	0.926	0.908
Avg_10_distance [Binned]	0.842	0.927	0.965	1	0.965	0.948
Avg_20_distance [Binned]	0.797	0.888	0.926	0.965	1	0.983
Avg_30_distance [Binned]	0.779	0.869	0.908	0.948	0.983	1

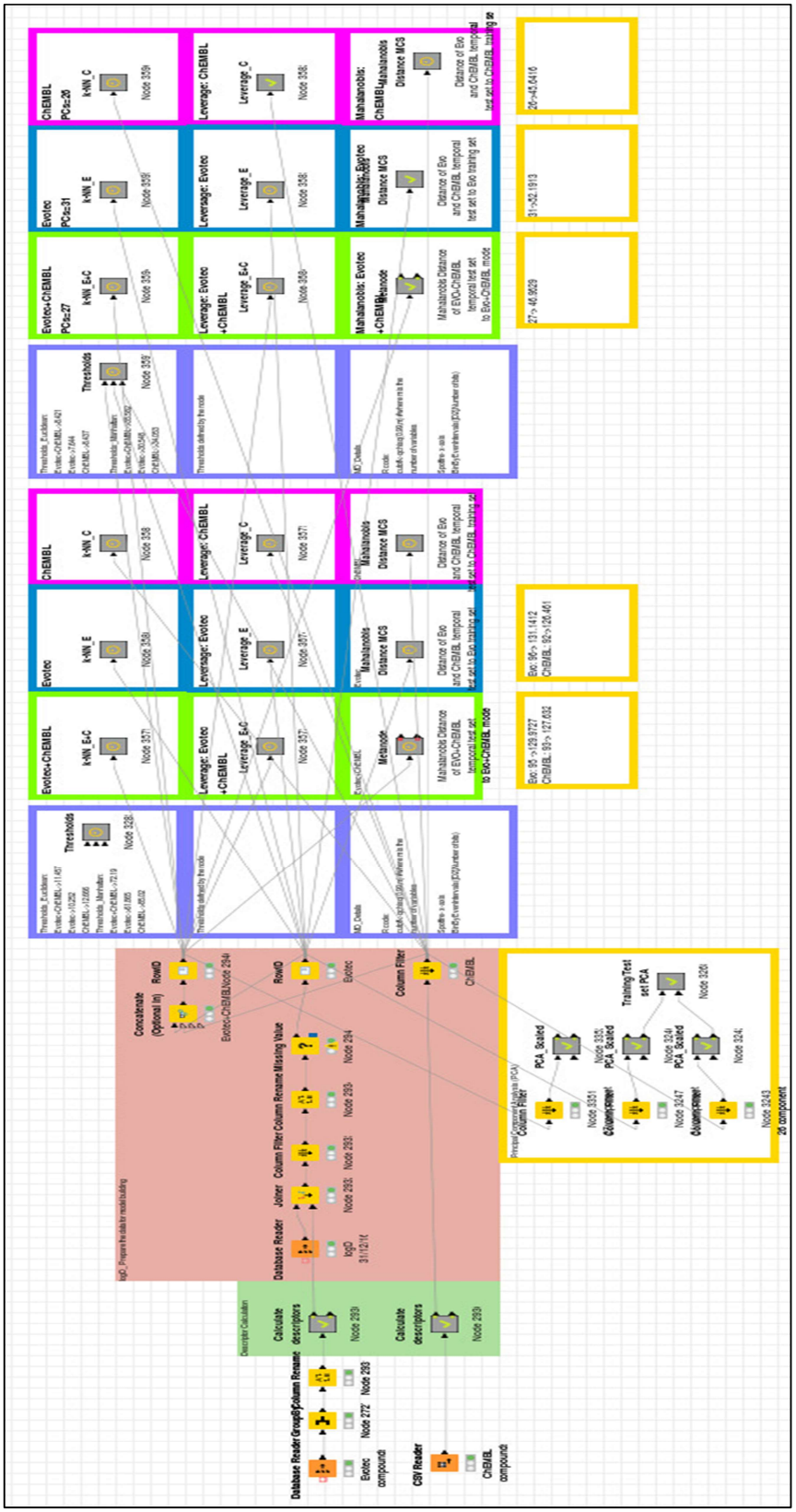

Figure S1: Screenshot of the KNIME workflow, created for the estimation of the Applicability Domain (AD) of the models and the calculation of principal components of the descriptors' matrix.

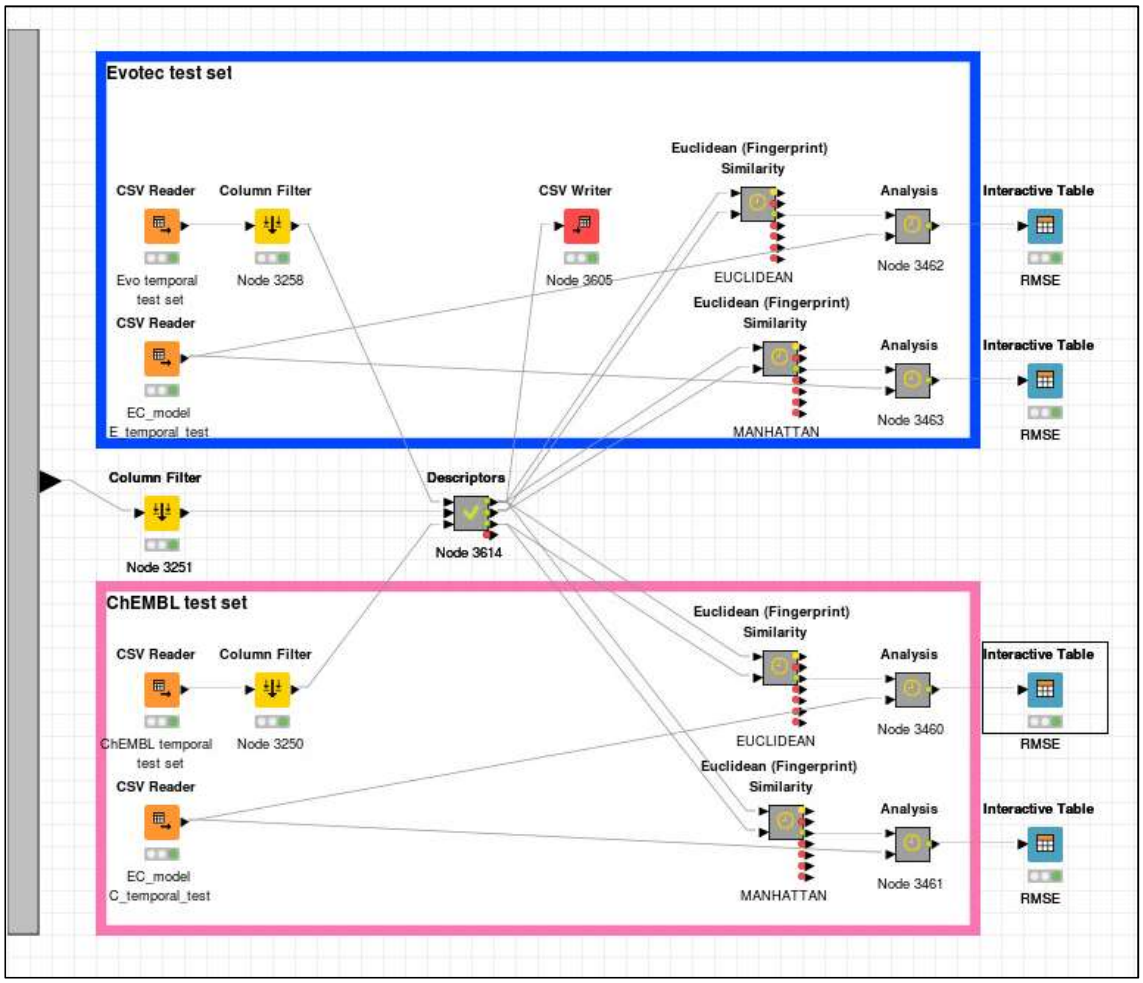

Figure S2: Screenshot of the metanode created to calculate the Applicability Domain of the models with the kNN with Euclidean and Manhattan distance function.

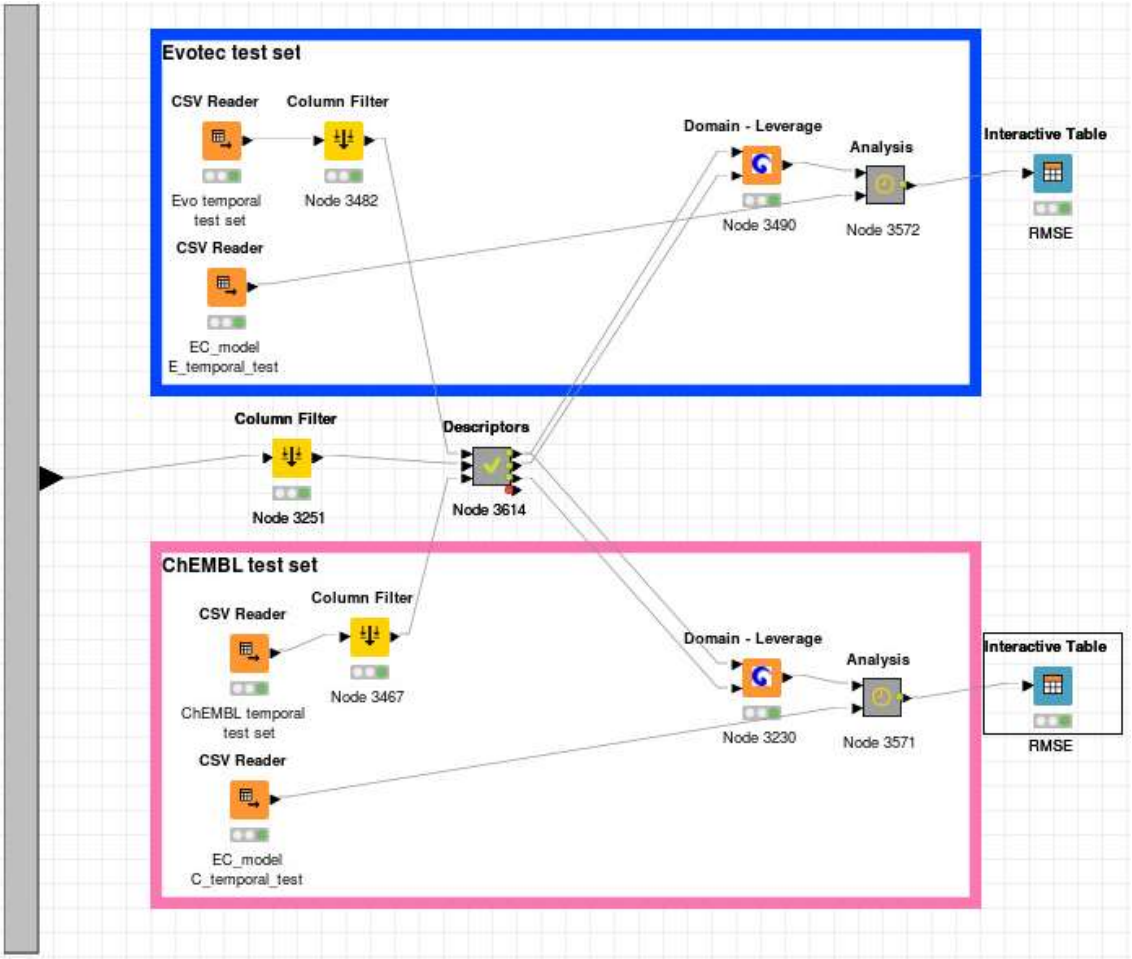

Figure S3: Screenshot of the metanode created to calculate the Applicability Domain of the models with the leverage method.

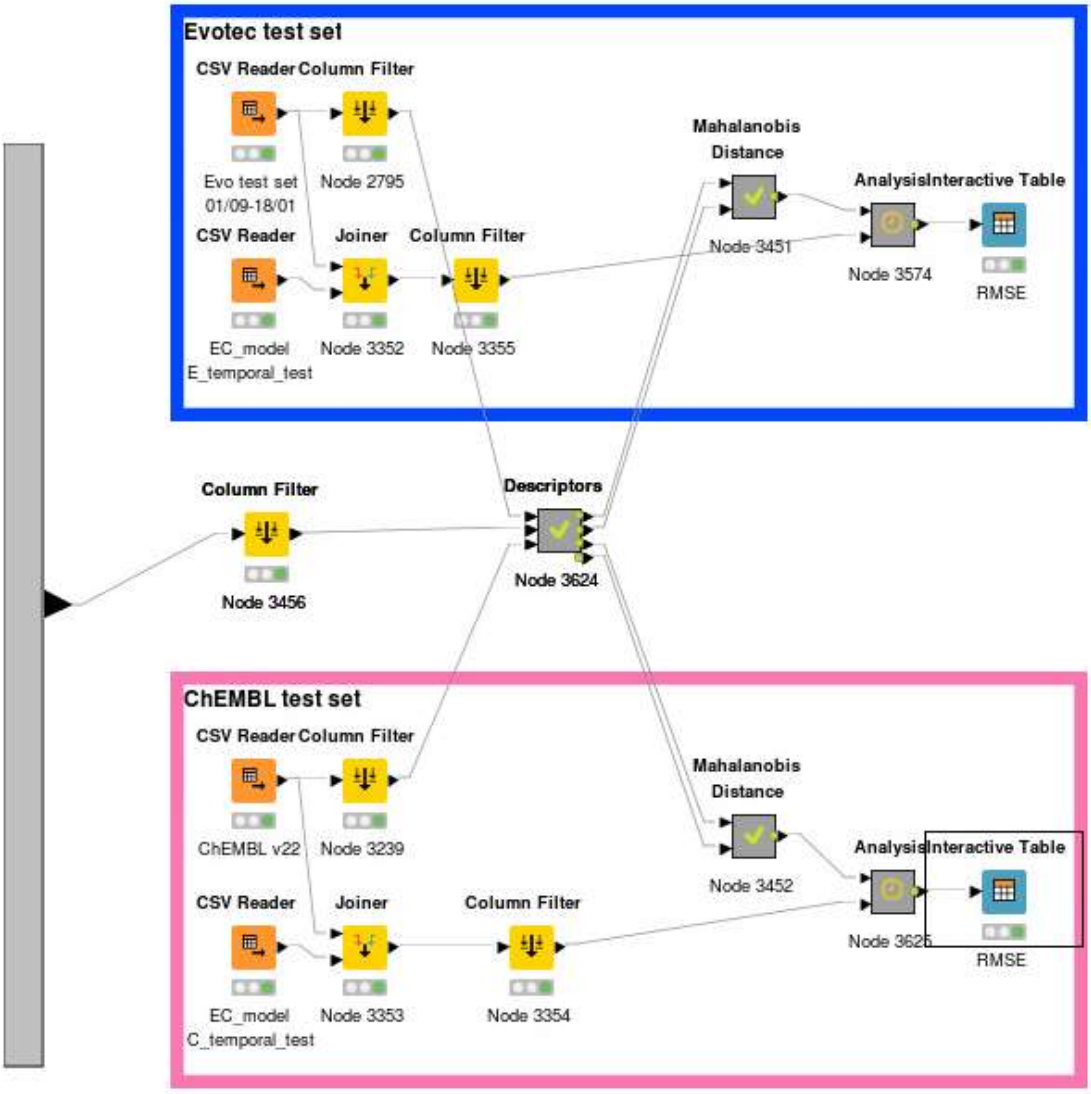

Figure S4: Screenshot of the metanode created to calculate the Applicability Domain of the models with the Mahalanobis Distance.