UPTEC X 06 011
MAR 2006

ISSN 1401-2138

LARS PERSSON

hERG modelling using 3D-pharmacophores

Master's degree project

UPPSALA UNIVERSITET

Molecular Biotechnology Programme

Uppsala University School of Engineering

UPTEC X 06 011	Date of issue 2006-03
Author **Lars Persson**	
Title (English) **hERG modelling using 3D-pharmacophores**	
Title (Swedish)	
Abstract Eleven pharmacophores for the cardiac K^+ channel hERG were developed using the modelling software Catalyst and evaluated with multivariate analysis. The pharmacophores will be used as visual feedback in drug design and as descriptors in predictive modelling. A pharmacophore-based automatic sorting scheme for hERG-compounds was generated and new approaches for classification modelling were explored.	
Keywords hERG, pharmacophores, exclusion volumes, structure-activity relationships, PLS-DA, descriptors	
Supervisors **Mats Svensson** **AstraZeneca R&D, Södertälje**	
Scientific reviewer **Johan Åqvist** **Department of Cell and Molecular biology, Uppsala University**	
Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **40**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

hERG modelling using 3D-pharmacophores

Lars Persson

Sammanfattning

hERG är en jonkanal i hjärtat som är inblandad i hjärtats pumpfunktion. Många läkemedel från många olika läkemedelsklasser har visat sig ha som biverkning att de förutom att binda sitt farmakologiska målprotein även blockerar hERG. Detta kan störa hjärtrytmen och i värsta fall orsaka hjärtflimmer. Läkemedelsföretagen satsar därför stora resurser på utveckling av olika metoder att upptäcka hERG-problem så tidigt som möjligt i utvecklingen av nya läkemedel. Om inriktningen på ett projekt behöver ändras eller om det måste läggas ned, blir det mer ekonomiskt ju tidigare detta beslut kan tas.

En tilltalande metod är datormodellering av hERG-bindning. Om modelleringen är tillförlitlig kan stark hERG-bindning förutsägas och man kan undvika kemisk syntes av blockerare. Syftet med det här projektet var att ta fram farmakoforer utifrån ett stort dataset med föreningar med känd och varierande bindningsstyrka till hERG. En farmakofor är en sammanfattning av vilka egenskaper en molekyl måste ha för att påverka ett målprotein och består av ett antal kemiska funktioner och deras inbördes koordinater. Farmakoforer är ett visuellt hjälpmedel för läkemedelskemister och kan även användas för prediktion av bindning. Efter statistisk utvärdering av farmakoforerna utvecklades matematiska modeller för prediktion av hERG-aktivitet. Modellerna kopplar ihop olika beräknade kemiska, fysiska och strukturella egenskaper en molekyl har, bl.a. passning till farmakoforerna, till en prediktion av hur stark bindning till hERG den har.

Examensarbete 20 p i Molekylär bioteknikprogrammet

Uppsala universitet mars 2006

1. Introduction

1.1 QT prolongation and hERG

Long QT syndrome (LQTS) is an abnormality of cardiac muscle repolarisation that is characterised by the prolongation of the QT interval in the electrocardiogram [1]. LQTS is associated with increased risk for torsades de points, a ventricular tachyarrhythmia that may degenerate to ventricular fibrillation and sudden death [2]. Several congenital and acquired disorders can lead to prolongation of the QT interval. Of special interest is the fact that numerous agents, belonging to different drug classes, have been associated with QT prolongation and torsades de pointes [3]. A number of drugs have been withdrawn from the market or restricted in availability as a result of their association with LQTS [4]. This has resulted in health concerns for patients as well as in great revenue-losses for the pharmaceutical industry. Before approval of a human pharmaceutical by regulatory authorities, potential for QT prolongation must now be thoroughly evaluated [5]. LQTS is a highly unwanted side-effect for drugs.

All known LQTS related to drug exposure can be traced to one specific mechanism – blockage of the voltage-gated cardiac potassium channel hERG (human ether-a-go-go-related gene) [6, 7]. The inner cavity of the hERG K^+ channel is large and hydrophobic and can trap a variety of ligands and many that other K^+ channels cannot trap [1]. The association of hERG with LQTS has launched a massive effort on the part of the pharmaceutical companies to understand how drugs interact with hERG on the molecular level and how interaction may be eliminated. Early detection of hERG blockers is an important aim since it will save a lot of time and money. An early failure is a cheap failure. Early awareness of hERG affinity for a lead compound can also guide the lead development in a direction away from hERG activity and save the project.

One interesting approach for early detection of hERG blockers is to use in silico techniques to filter out potential blockers in the context of virtual compound libraries. Compounds predicted to have high hERG affinity could then be avoided and resources could be concentrated to synthesis of compounds that meet this safety concern. In this work, the structure-activity relationships governing hERG-drug interactions were investigated and different approaches of predictive modelling were examined.

1.2 Classes of molecules that block hERG

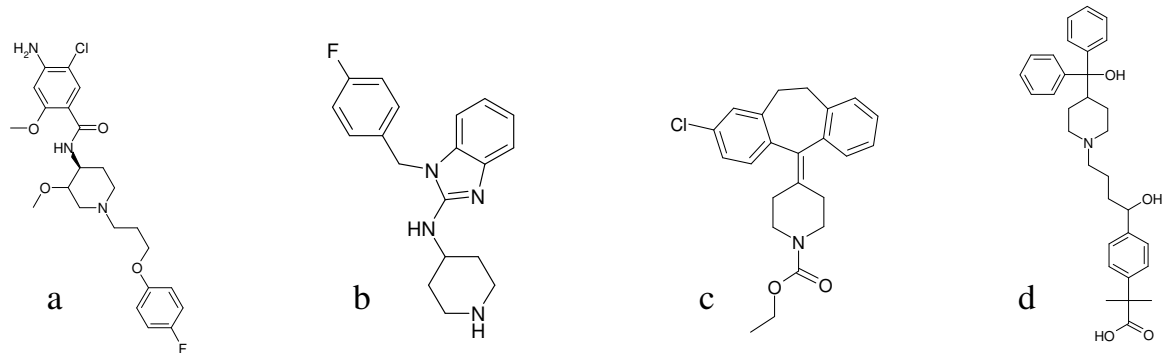

Figure 1. Drugs representing molecule classes. pIC_{50} is a measurement of binding affinity. (a) Cisapride, central amine, hERG pIC_{50}=8.19. (b) Norastemizole, terminal amine, hERG pIC_{50}=7.55. (c) Loratadine, neutral, hERG pIC_{50}=6.76. (d) Fexofenadine, acid, hERG pIC_{50}=4.67. All activities are from reference [3].

The hERG channel is promiscuous. A lot of drug-like molecules have affinity for it and the structural diversity among the binders are large. The classic hERG blocker is a compound with a central basic nitrogen between two lipophilic regions (Figure 1a). Several pharmacophores for central amine compounds have been published earlier [3, 8, 9]. A second class of hERG blockers known from the literature [8, 9] are terminal amines (Figure 1b). These compounds have generally not as high affinity for hERG as the central amines, but still results in QT-prolongation. During AZ hERG screening a third class of blockers have emerged – neutral compounds (Figure 1c). There is very little published on neutral hERG binders and the pIC50-values of the most potent compounds are often in the medium range defined below.

Since there are so many structurally diverse compounds that bind to hERG it is interesting to study the problem from the opposite direction - what properties do hERG non-blockers have? One modification that reduces hERG affinity is the introduction of an acidic group. Acids often have low or not measurable affinity. In this work, acids and zwitterions were treated as one separate class of compounds (Figure 1d).

1.3 SAR

SAR (Structure-Activity Relationship) is a common concept in medicinal chemistry. It can be defined as the association between the chemical composition of a molecule and its biological effect.

1.4 Pharmacophores

Pharmacophores are sets of molecular features and their relative coordinates. The pharmacophore for a certain macromolecular target is developed to describe the necessary features a ligand need for activity at that target. Typical features are hydrophobic centres, aromatic rings, charges, H-bond acceptors and donors. They are generated from a set of structurally diverse known active compounds and are conjunctions of their features. In other words, pharmacophores are the largest set of features with relative distances that the active training compounds have in common. Pharmacophores can also have exclusion volumes at certain positions relative to the chemical function features. The exclusion volumes represent regions which cannot contain any topology because it might impinge sterically on the macromolecular target. At AstraZeneca pharmacophores are used in virtual screening, lead identification and lead optimisation.

1.5 Task

The task was to construct new hERG-pharmacophores and to use them in hERG-modelling and classification. Besides their use as descriptors in multivariate modelling the pharmacophores can provide valuable visual feedback for synthetic chemists and help develop lead compounds away from hERG affinity. It was important that the classification protocol could be automated and run as a script from a web interface (webtool).

1.6 Aim

The primary aim for this project was to generate pharmacophores which provide good feedback and enrichment. The secondary aim was design of a model that could achieve 80% correct classification (into the three classes high, medium and low) on an external test set.

2. Material & Methods

2.1 General methodology

The general methodology was to develop pharmacophores for one type of compound at a time, use these pharmacophores to create a rule that automatically could filter compounds of this type out from a test set and then go on to work with the next type. In sequence Central amine, Terminal amine and Neutral pharmacophores were generated. PLS and PLS-DA [10] was used to evaluate the pharmacophores and for classification modelling. Both General, Central amine and Terminal amine models were developed.

2.2 Hardware and Software

All computations were carried out on a SGI server with 32 processors (MIPS R12000 400 MHz), running Irix 6.5. Clustering of compounds was performed by the in-house AstraZeneca program PC Flush 2.1.5 [11]. 1D & 2D-descriptors of the compounds were generated with SELMA [12], an in-house AstraZeneca program. hERG Smarts [13, 14] for the compounds were generated with an in-house AstraZeneca program. Conformational models, pharmacophores and database screening were performed with Catalyst version 4.11 [15]. Selection of compounds for training sets was carried out by BigPicker [11], an in-house AstraZeneca program. PLS and PLS-DA were performed with Simca-P+ version 10.0.2.0 [10].

2.3 Definition of classes

Table 1. Activity class definitions

High	$pIC_{50} \geq 6$
Medium	$4.5 \leq pIC_{50} \leq 6$
Low	$pIC_{50} \leq 4.5$

IC_{50} (Inhibition concentration 50%) represents the concentration of an inhibitor that is required for 50% inhibition of an enzyme in vitro.

There is a safety guideline at AstraZeneca saying that no compound entering late phases should have an IC_{50} for hERG lower than 30µm, corresponding to a pIC_{50} of 4.5. Therefore, 4.5 was a logical limit between low and medium for this classification model (Table 1). Leads that have medium or high affinity to hERG have to be developed towards the secure low affinity interval, with this work as one

aid. The limit between high and medium affinity was somewhat arbitrarily chosen set to 1µm. An advantage of choosing a 3-class design is that the medium class separates high and low, so even if there are classification errors, very few of them should be double faults. Especially important is that compounds classified as low should not be high affinity binders. The opposite is not good either because compounds that are predicted to be high, but is screened anyway and turns out to be low affinity binders will undermine the confidence in the model.

2.4 Datasets

The original dataset was comprised of 7071 AstraZeneca in-house compounds from various projects. The number of projects was large and between-project compound structural diversity was also large. Previous publications on hERG modelling [3, 8, 16, 17] has used datasets containing 20-400 compounds with activity data often collected from different sources within the literature. Activity data from different assays may not be comparable, and is an additional source of errors. In this work all pIC_{50}-values were measured in the same assay, a proprietary method within AstraZeneca. Compounds that did not have a measurable pIC_{50} were given the value of 4.5, so that they could be used in multivariate analysis. Apart from pIC_{50}-values, descriptors available were hERG Smarts, and Selma parameters and, after pharmacophore generation, fit-values to eleven different pharmacophores. Smarts are structure fragments combined with logical expressions. Selma parameters are physical-chemical properties, topological properties and counts of number of rings, atoms, h-bond acceptors etc for a compound using 2D-structure as input. The 7071 compounds were divided into 1473 clusters using PC Flush 2.1.5 with maximum Tanimoto distance 0.3 to aid SAR investigation and selection of compounds for pharmacophore generation. A second dataset of 3218 AZ compounds was saved as a pure test set. This set is in this text called Test set B.

2.5 Selection of template molecules for pharmacophore generation

The selection of template compounds for pharmacophore generation was performed by visual inspection in Spotfire® DecisionSite 7.3 [18]. One cluster of compounds at a time was investigated for SAR. Since molecules within the same cluster are structurally similar it is possible to find minor changes or substitutions in a series which result in large and interesting hERG activity differences. Most interesting is to compare compounds that differ in hERG pIC_{50}, but have similar clogp, which is a calculated descriptor that models hydrophobicity (Figure 2). Then activity differences are probably not dependent on hydrophobicity differences. Hydrophobicity is often a strong driving force for ligand

binding. Ideal compounds for pharmacophore generation are highly active, not too hydrophobic, structurally diverse compounds which have associated SAR.

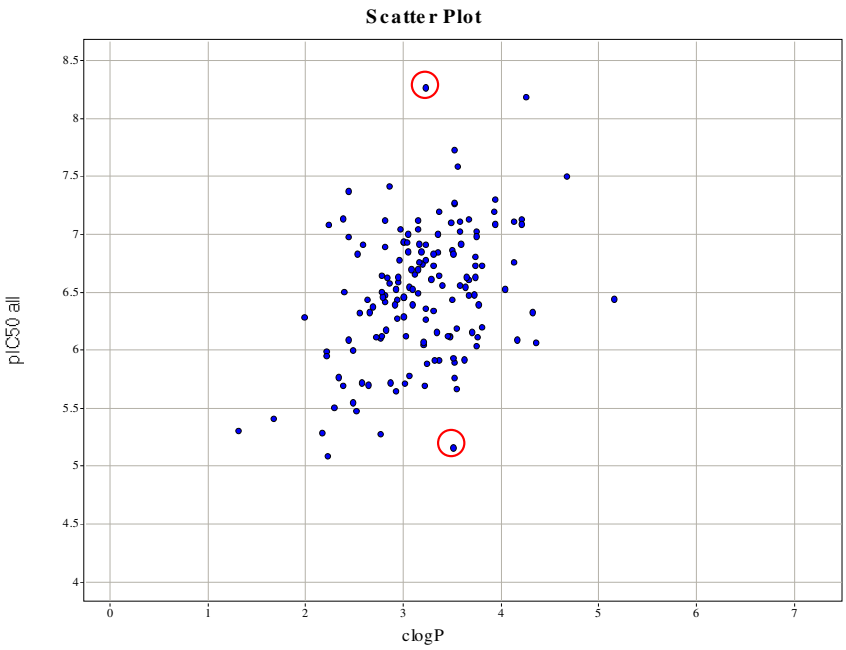

Figure 2. Plot of pIC_{50} vs. clogp for a cluster of compounds. The compounds marked with rings are interesting to compare for SAR.

Inactive compounds selected for generation of pharmacophores with exclusion volumes should have other properties. First of all they need to be low active and not too hydrophilic. If they are too hydrophilic, non-binding might depend on poor membrane permeability rather than SAR. Further they must align as well as the highly active compounds to a pharmacophore without exclusion volumes, but protrude in some region not occupied by the high activity compounds. The rationale for the exclusion volumes are then that this region is occupied by the macromolecule in ligand binding.

2.6 Conformational models

Conformers of each compound were generated in Catalyst using the default 20kcal/mol range limit and the fast search option. The maximum number of conformers was 250.

2.7 Pharmacophore generation

All pharmacophores were produced using the Catalyst program, version 4.11 (Accelrys Inc., San Diego, CA, USA). Totally over 80 pharmacophores were generated and evaluated with multivariate

analysis. In the end three non-correlating top queries for each of the molecule classes Central amines, Terminal amines and Neutrals were selected. Also sorted out for filtering purposes were the two queries **Negion** and **Posion** resulting in a set of eleven pharmacophores for use in classification and modelling.

If not stated otherwise the feature options in Hypothesis generation were H-bond acceptor (A), H-bond donor (D), hydrophobic (H), ring aromatic (A) and positive ionisable (P). When using HipHopRefine, active compounds had the number 2 in the principal column of the spreadsheet and inactives the number 0. Maximum Omitted Features was globally set to 0.

The P feature was modified because the default definition did not include amino pyridines and amino pyrimidines. The nitrogen in these rings is also protonated at physiological pH. Figure 3 depicts the added rules and also shows which nitrogen is protonated.

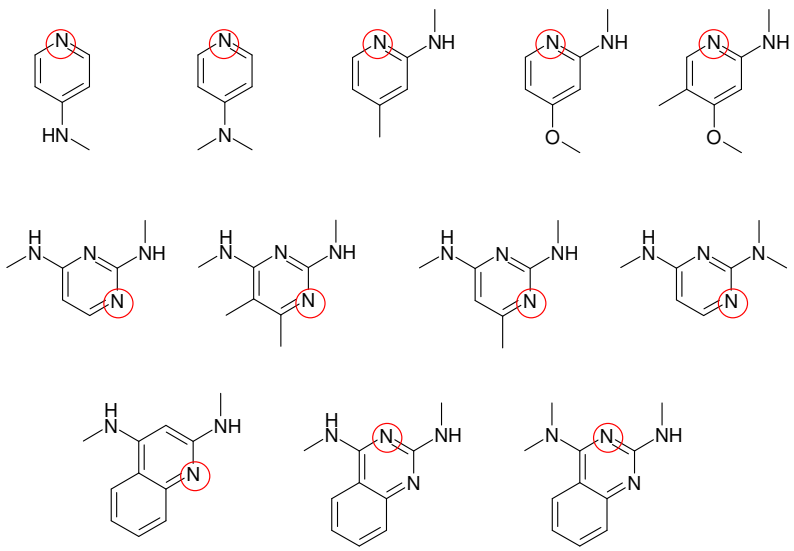

Figure 3. Added rules to the predefined chemical function Positive Ionisable (P) used in this work. The rings mark the association. All aromatic, not bridgehead, carbons have a defined hydrogen count of 1 and all terminal carbons are defined to have coordination 4.

The quality of the mapping of a compound to a pharmacophore is indicated by a fit-value. This is a kind of minimized sum of square displacements measure. For how fit-values are computed, see reference [15]. The maximum fit-value is the number of features in the hypothesis (i.e. R+H+P+A=4). If some feature is weighted, the max fit-value is the sum of the weights (i.e. R+H+P (weight 2)+A=5). If a conformer of a compound enters an exclusion volume when mapping to a pharmacophore, that alignment is blocked. If it just enters an exclusion volume slightly, the fit-value is only reduced. A shape constraint is a drug-shaped volume in a pharmacophore. To fit a pharmacophore with a shape

constraint, conformers of compounds must fit the shape better than a certain threshold value, a similarity tolerance. Only conformers that fulfil this initial condition will be considered for mapping to the chemical function features in the pharmacophore. The minimum fit-value for search is a user-defined threshold. If the fit-value of a compound to a hypothesis is higher than the minfit-value, the compound is considered as a hit and this speeds up screening and can easily be automated. Hit is set to 1 and not hit is set to 0 in the responding datasheet-column. Minfit-values were determined by visual inspection in Spotfire® DecisionSite 7.3. Since fit to pharmacophores is not an exact method to measure biological activity this conversion from continuous to binary data may not be disadvantageous.

For the Compare/Fit function in Catalyst, the energy limit was 20kcal/mol, maximum omitted features were 0 and Fast fit was used. Maximum omitted features 0 means that a compound must, at least slightly, map all features in a pharmacophore to gain a fit-value by the Compare/Fit function.

The rough optimisation of exclusion volume tolerances has been evaluated with multivariate analysis.

2.7.1 Filtering pharmacophores

Negion is identical to Catalyst's predefined chemical function Negative Ionisable. Max and min fit-value was 1.

Posion is the above defined modified version of the predefined chemical function Positive Ionisable. Max and min fit-value was 1.

2.7.2 Nomenclature for molecule class pharmacophores

The first capital letters in the pharmacophore names represents the features present in the hypothesis. The same letters as in Catalyst are used. R is ring aromatic, H is hydrophobic, P is **Posion**, the modified version of the Catalyst feature Positive Ionisable defined above, and A is H-bond acceptor. The next letter or letters stands for which molecule class the pharmacophore is developed for. kl is central amines, t is terminal amines and neu or n is neutrals. ex means that there are exclusion volumes in the pharmacophore, neg means that it is a negative pharmacophore and sh means that there is a shape constraint in the pharmacophore. Italic letter combinations are used for all properties of a pharmacophore not describing which Catalyst chemical features it contains. The names of the eleven selected pharmacophores are written in bold face.

2.7.3 Central amine pharmacophores

RHPklex1 was generated using the HipHop algorithm in Catalyst with five AZ-compounds as actives. The feature selection was set to give a RHP-pharmacophore. The top query was optimised with hypoopt v4.0 [19] and the exclusion volumes were added manually. Volumes were added to block away or lower the fit-value for one flexible inactive AZ-compound, but the highest priority was to not lower the fit-values for the five active compounds mentioned above. The inactive compound was very similar to one of the high activity compounds, but actually more hydrophobic. The Positive Ionisable feature was given a weight of 2. Max fit was 4 and min fit 1.5.

RHPklex2 was generated using the HipHopRefine algorithm in Catalyst. Five AZ-compounds were used as actives and seven other AZ-compounds were used as inactives. The top RRHP query was optimised with hypoopt v4.0. A second crude optimisation was performed by changing the tolerances of the exclusion volumes from the default 120 to 60 picometers. Finally one R feature situated next to the H feature was removed. This because RHP-pharmacophores were good, making RHP with exclusion volumes very promising, but no good RHPexclvol-query could be automatically generated by Catalyst. Max fit was 3 and min fit 1.5.

RHPAklex was generated with HipHop using six AZ-compounds and optimised with hypoopt v4.0. The Positive Ionisable feature was given a weight of 2. The exclusion volumes were added manually in the same way as for **RHPklex1**. Max fit was 5 and min fit 3.

2.7.4 Terminal amine pharmacophores

RHPtex1 was generated using the HipHopRefine algorithm. Actives were eight AZ-compounds. Inactives were five other AZ-compounds. Two queries were chosen for development, one of them ended up as **RHPtex1** and another as **RHPtex2**. To allow features to be moved during optimisation, the tolerances for the exclusion volumes for **RHPtex1** were first reduced to 60pm before optimisation with hypoopt v4.0. Then the tolerances for the exclusion volumes were roughly optimised from 120 to 80pm. Since there were gaps between exclusion volumes that did not harmonize with my SAR hypothesis for terminal amines, extra volumes were added manually to fill these gaps for blocking out inactives. For this, 16 actives and 20 inactives were used and spaces where only inactive compounds mapped were closed with exclusion volumes. Max fit was 3 and min fit 1.5.

RHPtex2 came out from the same HipHopRefine run as **RHPtex1**. The two queries had the same RHP features lined up in the same order, but different geometries and exclusion volume patterns. This

pharmacophore was optimised with hypoopt v4.0 with default tolerance on exclusion volumes and these were then reduced to 80pm. Max fit was 3 and min fit 1.5.

RRHPtneg was generated from the mapping of one inactive AZ-compound to the query RRPterm, a pharmacophore that was not selected for modelling. The extra H feature was placed on a terminal hydrophobic centre of the inactive AZ-compound situated at the other end of the molecule relative to the basic nitrogen (Figure 12). The rationale behind this was that visual inspection in Spotfire® suggested that long hydrophobic chains (about 14 bonds) with a terminal amine had less hERG affinity than terminal amines with semi long (about 11 bonds) hydrophobic chains. RRPterm was generated with the HipHop algorithm using the same active compounds as the other two terminal pharmacophores and was optimised with hypoopt v4.0. Max fit was 4 and min fit 1.5.

2.7.5 Neutral pharmacophores

RHHHneu was generated with HipHop and optimised with hypoopt v.4.0. Actives were nine neutral AZ-compounds. Max fit was 4 and min fit 2.5.

RHHHAneu was generated prior to my arrival at AstraZeneca by an in-house computational chemist. Max fit was 5 and min fit 2.5.

RHHAnexsh was generated with HipHopRefine with ten AZ-compounds as actives and six other AZ-compounds as inactives. After optimisation with hypoopt v4.0, tolerances for exclusion volumes were reduced to 80pm and some were manually deleted to raise fit-values for the ten active compounds. Finally one of the active AZ-compounds was converted to a shape when aligned to the pharmacophore and the shape and pharmacophore were merged into one combined hypothesis. For the shape min/max percent extent and box volume match were 0.7/1.6 and min/max similarity tolerance 0.4/1. Max fit was 4 and min fit 1.5.

2.8 Screening of databases

Screening of compounds against pharmacophores was performed with the Fast Flexible Search algorithm in Catalyst. Maximum search hits were 10000.

2.9 Separation of compounds into molecule classes

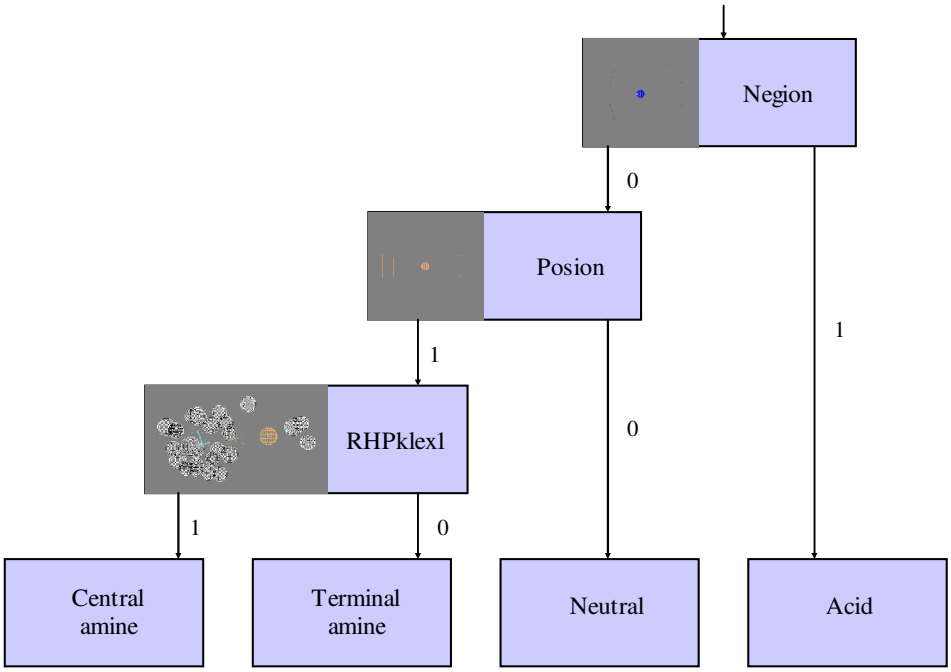

Figure 4. Flow chart for molecule classification. Depending on if a compound fit to the pharmacophores **Negion**, **Posion** and **RHPklex1**, it is automatically sorted into the molecule classes Central amines, Terminal amines, Neutrals or Acids.

For pharmacophore evaluation and the construction of Central amine and Terminal amine classifiers, it was important to generate a method to separate Central amines, Terminal amines, Neutrals and Acids. The filtering needs to be automatic to be robust and possible to integrate into a webtool. Fit to three pharmacophores, **Posion**, **Negion** and **RHPklex1**, were used as rules. See Figure 4 for the flow cart.

2.10 Training set and test sets

2.10.1 General model

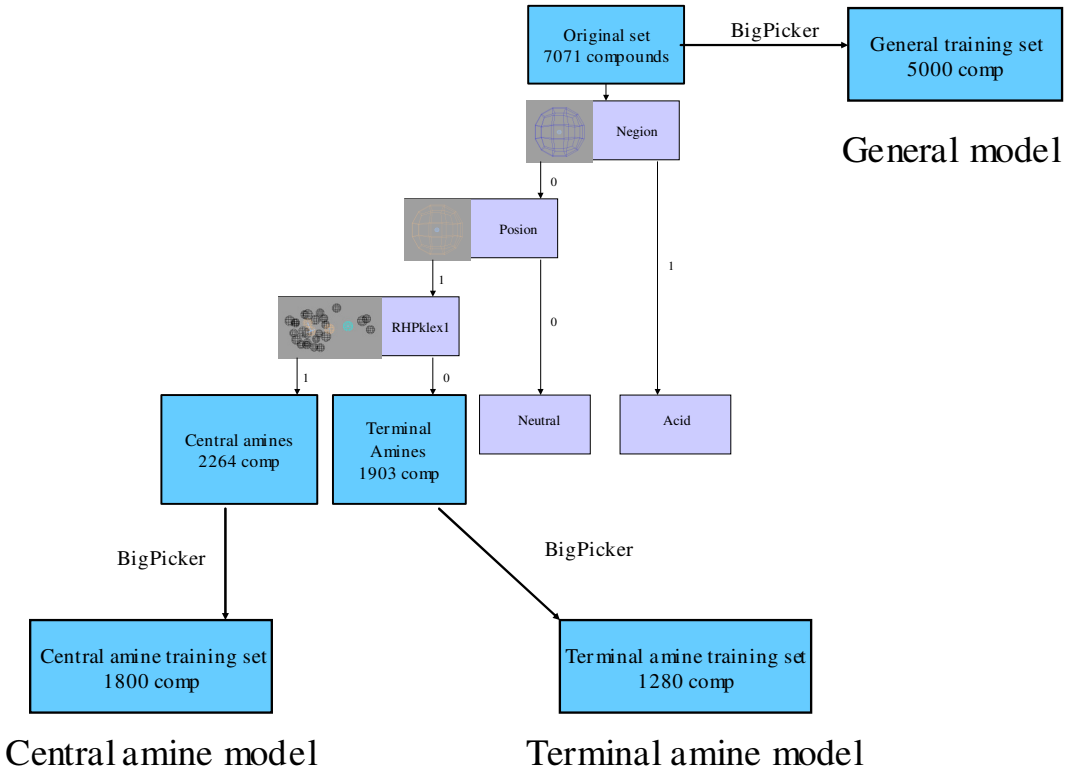

Figure 5. Flow chart over of how the General, Central amine and Terminal amine training sets were generated. The test sets A, C and T are the 2071, 464 and 623 compounds not selected by BigPicker. Note that these are not represented by a box in the figure.

Table 2. Number of compounds in each activity class for the original dataset and the General model training and test sets. X*Y means that X compounds are present in Y copies in the Training set for weighting reasons.

		Original set	Training set	Test set A	Test set B
High	$pIC_{50} \geq 6$	837	500*6	337	131
Medium	$4.5 \leq pIC_{50} \leq 6$	4025	3000	1025	1669
Low	$pIC_{50} \leq 4.5$	2209	1500*2	709	1418
Sum		7071	9000	2071	3218

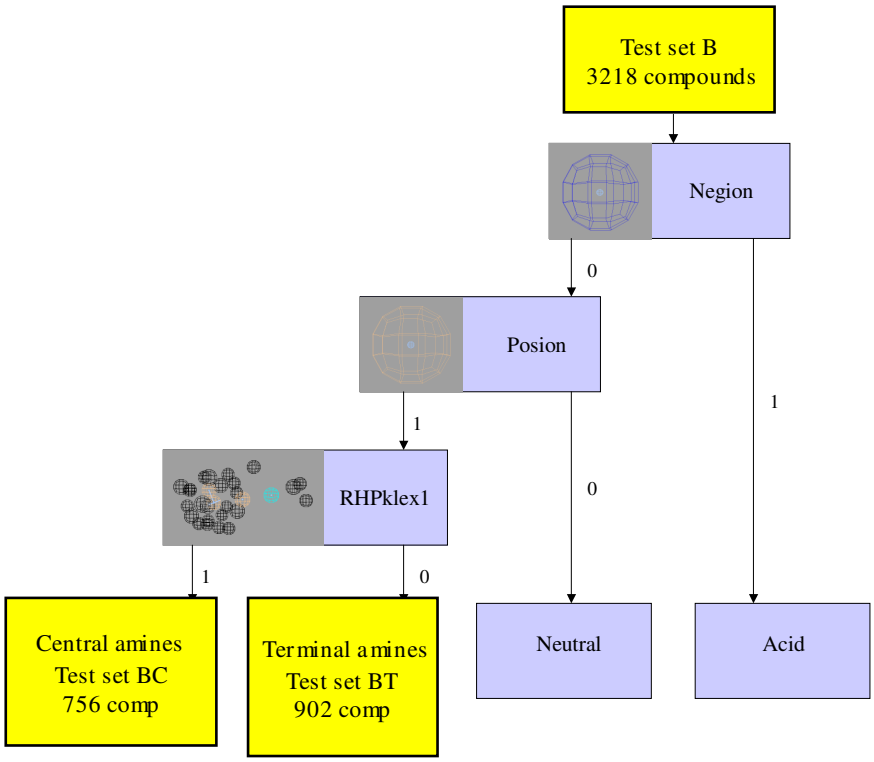

Figure 6. Flow chart over how the test sets B, BC and BT are related to each other.

The compounds were not evenly distributed across the activity range (Table 2), a majority had medium activity and only 12% were high. If the aim is to develop a model that gives equally good recall for all classes, the training set should contain an equal number of compounds from each class. To save some high and low compounds for Test set A and to still obtain a large training set, first 500 high, 3000 medium and 1500 low compounds were selected by the AZ in-house program BigPicker, which picks out structurally diverse subsets (Figure 5). The rows in the datasheet containing highs and lows were then copied 5 times respectively 1 time giving a training set of 9000 compounds, 3000 unique mediums, 6 copies each of 500 highs and 2 copies each of 1500 lows. The 2071 compounds that were not selected by BigPicker now constituted Test set A. Approximately 300 out of the 837 high activity compounds originated from the same project and were therefore structurally similar. The choice of 500 selected high compounds was made to reduce the models bias towards these series. Since BigPicker selects molecules by structural diversity, a majority of these compounds ended up in Test set A. The number of compounds from each molecule class found in each activity class in each dataset in Table 2 can be found in Appendix 6.1.

2.10.2 Central amine model

Table 3. Number of compounds in each activity class for the central amine original dataset and the Central amine model training and test sets. X*Y means that X compounds are present in Y copies in the Central Amine Training set for weighting reasons.

		Central amines Original set	Central amine Training set	Test set C	Test set BC
High	$pIC_{50} \geq 6$	706	400*3	306	109
Medium	$4.5 \leq pIC_{50} \leq 6$	1318	1200	118	418
Low	$pIC_{50} \leq 4.5$	240	200*6	40	229
Sum		2264	3600	464	756

The central amine original dataset is comprised of the 2264 central amines filtered out from the original dataset of 7071 compounds (Figure 5). The central amine training set was prepared in the same way as the original training set and the numbers of compounds from each activity class and multiplications is found in Table 3. Test set C is all central amines in the original dataset that was not selected by BigPicker and Test set BC is all central amines in Test set B (Figure 6). The performances of the General and the Central amine model on Test set BC can readily be compared.

2.10.3 Terminal amine model

Table 4. Number of compounds in each activity class for the terminal amine original dataset and the Terminal amine model training and test sets. X*Y means that X compounds are present in Y copies in the Terminal Amine Training set for weighting reasons.

		Terminal amines Original set	Terminal amine Training set	Test set T	Test set BT
High	$pIC_{50} \geq 6$	104	80*10	24	15
Medium	$4.5 \leq pIC_{50} \leq 6$	1234	800	434	550
Low	$pIC_{50} \leq 4.5$	565	400*2	165	337
Sum		1903	1280	623	902

The terminal amine original dataset is comprised of the 1903 terminal amines filtered out from the original dataset of 7071 compounds. The terminal amine training set was prepared in the same way as the original and central amine training set (Figure 5) and the numbers of compounds from each activity class and multiplications is found in Table 4. Test set T is all terminal amines in the original dataset that was not selected by BigPicker and Test set BT is all terminal amines in Test set B (Figure

6). The performances of the General and the Terminal amine model on Test set BT can readily be compared.

2.11 Classification with PLS-DA

The General, Central and Terminal amine PLS-DA models were all generated in Simca-P+ v.10.0.2.0 with the same protocol. Work set was the respective training set, all variables except pharmacophore fit, Smarts, Selma parameters and three random variables were excluded, the classes were set from the activity classes, model type in Simca was changed to PLS-DA and a first model was generated with autofit. Since several observations were present in several copies, the default validation based on Q^2-values suggested overfitted models. These models has no problem to predict a left out observation that there is another copy of in the work set and the resulting Q^2-value of such a validation is therefore too high. For this reason after autofit of a model, the last components were deleted. Usually the first five components were left after inspection of R^2-values, Q^2-values and number of iterations for the last components. After the first model was generated, all variables that did not have a VIP-value higher than all the three random variables were deleted along with the random variables. Finally a second model was generated with autofit and the last components were deleted as above.

3 Results and discussion

3.1 Pharmacophores

Pharmacophore features are coloured as follows: ring aromatic (R), two adjacent orange spheres; poison (P), orange; hydrophobic (H), blue; hydrogen bond acceptors (A), two adjacent green spheres; exclusion volumes (ex), black. The aligned molecules in the figures are drugs that are on, or have been withdrawn from, the market [3].

3.1.1 Central amine pharmacophores

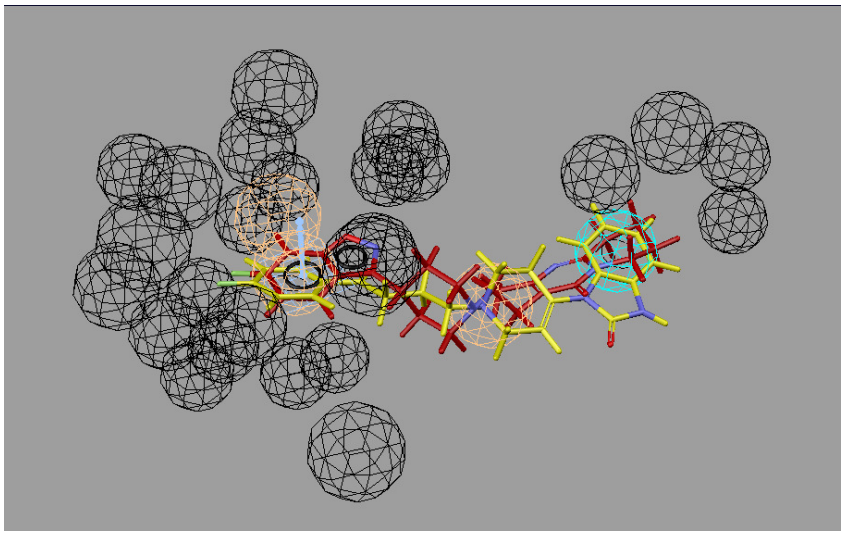

Figure 7. Droperidol (yellow) and Risperidon (red) aligned with **RHPklex1**.

RHPklex1 (Figure 7) consists of one central **Posion** (P) feature between one ring aromatic (R) and one hydrophobic (H) feature. The topology is slightly bent. Similar pharmacophores have previously been published [3, 8]. A novel feature with this hypothesis is the addition of a number of exclusion volumes that blocks out compounds branched in the ring aromatic part of the molecule. The rationale behind exclusion volumes is that they represent a subset of the volume where protein residues are situated when binding to the ligand. In **RHPklex1** these volumes are manually placed further away from the R and P features compared to the automatically generated **RHPklex2**, allowing larger and more substituted molecules to map the pharmacophore. Because of this more generously allowed volume, practically all central amines fit the query and it can be used for filtering, but does not provide excellent enrichment among central amines. The volumes are still blocking out most terminal amines that could map the RHP query without exclusive volumes in twisted and high-energy conformations. The P feature as an experiment got a weight of 2 early during development since this feature is known

[20] to be very important for hERG binding. It is not though thoroughly investigated how big the impact of this weighting is on performance.

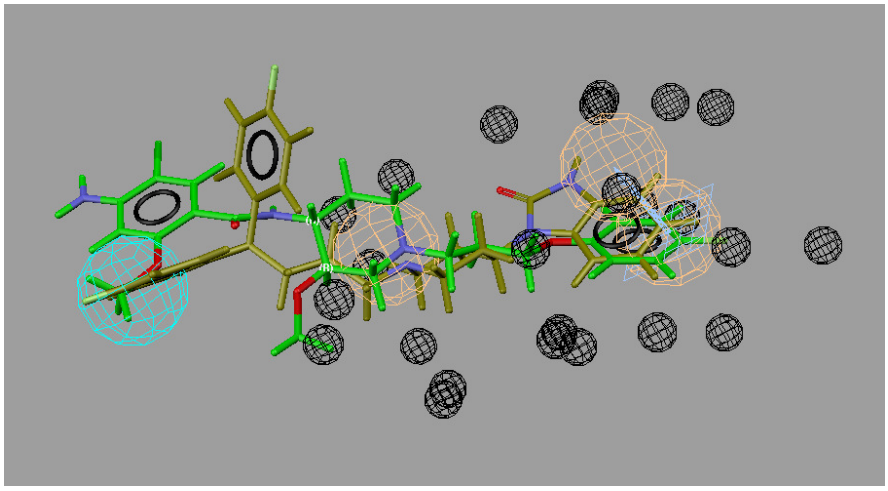

Figure 8. Pimozide (brown) and Cisapride (green) aligned with **RHPklex2**.

RHPklex2 (Figure 8) is the most enriching pharmacophore, both for the entire dataset and the central amines. It has a topology similar to **RHPklex1**, but the distance between the features is slightly longer and they are nearly linearly aligned. The HipHopRefine-generated exclusion volumes surrounds the entire R & P half of the query and are placed closer to them than in **RHPklex1**. This results in a small allowed volume around the features that blocks out R or P-branched compounds. The exclusion volumes are rather small and gaps between them allow substitutions, but these compounds often get their fit-values reduced below minfit.

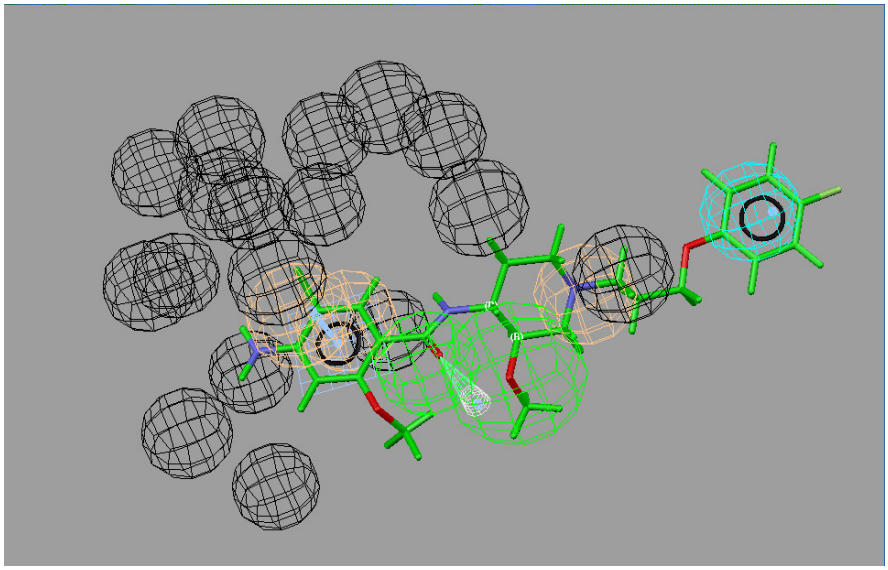

Figure 9. Cisapride aligned with **RHPAklex**.

RHPAklex (Figure 9) is very similar to **RHPklex1**, but has an additional H-bond acceptor (A) feature situated next to the aromatic ring. Hydrogen bonding to residues in the selectivity filter of the hERG channel has previously [21, 22] been reported and a RPA pharmacophore similar to **RHPAklex** without the exclusion volumes and the H feature has been published [8]. The exclusion volumes are situated in similar positions as those in **RHPklex1**. The P feature has a weight of 2 for the same reason as **RHPklex1**.

3.1.2 Terminal amine pharmacophores

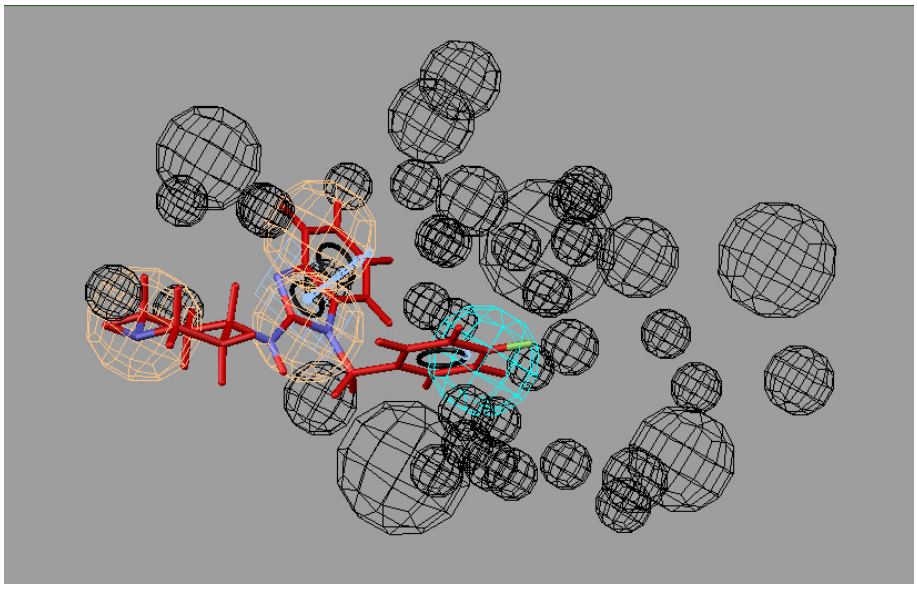

Figure 10. Norastemizole aligned with **RHPtex1**.

RHPtex1 (Figure 10) consists of one ring aromatic (R) feature between one **Posion** (P) and one hydrophobic (H) feature. The three features are arranged almost linearly in space and the R and particularly H part of the query are surrounded by exclusion volumes since there was SAR for that branching in this area was negatively correlating with hERG activity.

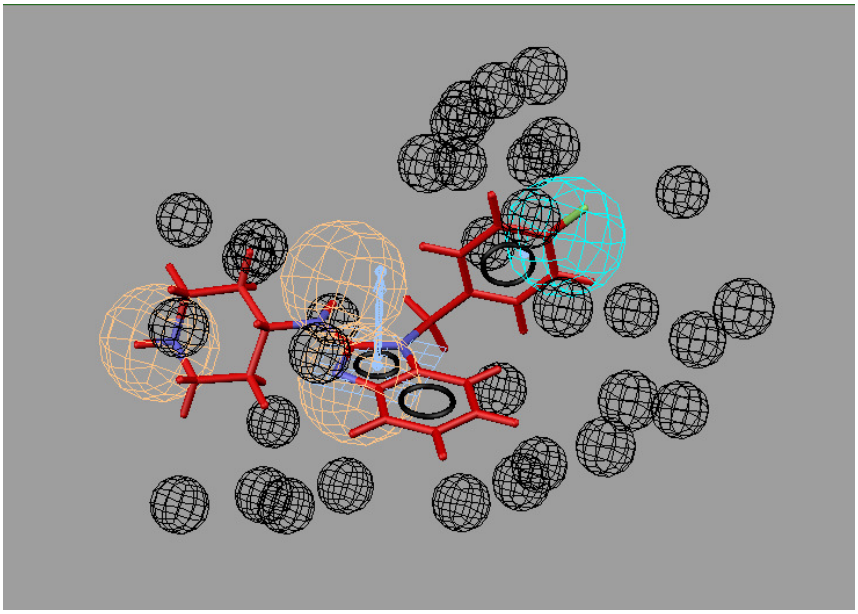

Figure 11. Norastemizole aligned with **RHPtex2**.

RHPtex2 (Figure 11) has the same features as **RHPtex1**, but they are arranged in a bent orientation instead of a linear. The exclusion volumes are fewer, but closer, to the main features resulting in a more difficult pharmacophore to fit than **RHPtex1**. The space beyond the H feature is also less closed than in **RHPtex1**. Pharmacophores resembling the two RHPtex hypothesis, but without exclusion volumes and with the H feature positioned next to the R feature at the same distance from P, has been reported [8, 9]. One interesting property of **RHPtex2** is that it functions as a negative pharmacophore for central amines. The R & H features can represent an aromatic ring branched in a direction away from the basic nitrogen. Branches like this are blocked by the exclusion volumes around the central amine pharmacophores. For this reason **RHPtex2** is less correlated with hERG activity in the General model. This contributes to the bad performance of the General model in predicting highly active terminal amines (presented in the modelling section, Table 4).

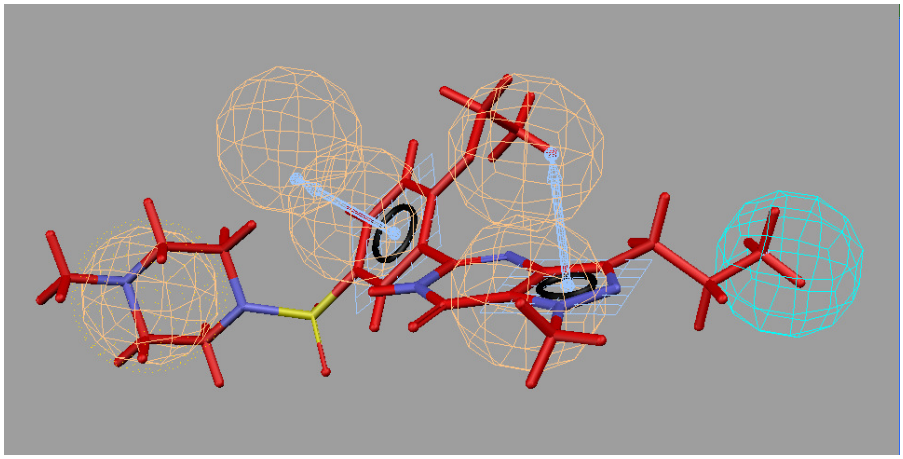

Figure 12. Sildenafil aligned with **RRHPtneg**.

Visual inspection of pIC50 vs. fit values to the pharmacophore RRPterm in Spotfire® suggested that addition of a hydrophobic feature would produce a negative pharmacophore – a pharmacophore that mostly non-actives fit. **RRHPtneg** (Figure 12) was generated and fit to this pharmacophore did indeed correlate negatively with pIC50 for terminal amines. That compounds with this topology are generally not hERG active is also supported by Aronov [20].

3.1.3 Neutral pharmacophores

No neutral pharmacophores have previously been reported. Finding SAR among the neutral compounds is difficult and the neutral pharmacophores are not as enriching as the central and terminal amine pharmacophores, meaning that they don't discriminate as well between actives and inactives.

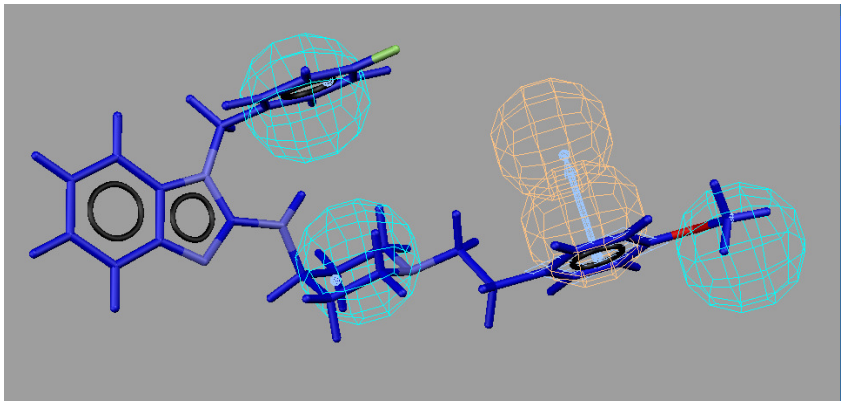

Figure 13. Astemizole aligned with **RHHHneu**. Note that Astemizole is not a neutral compound.

RHHHneu (Figure 13) consists of three hydrophobic (H) and one ring aromatic (R) features. A lot of compounds fit this pharmacophore.

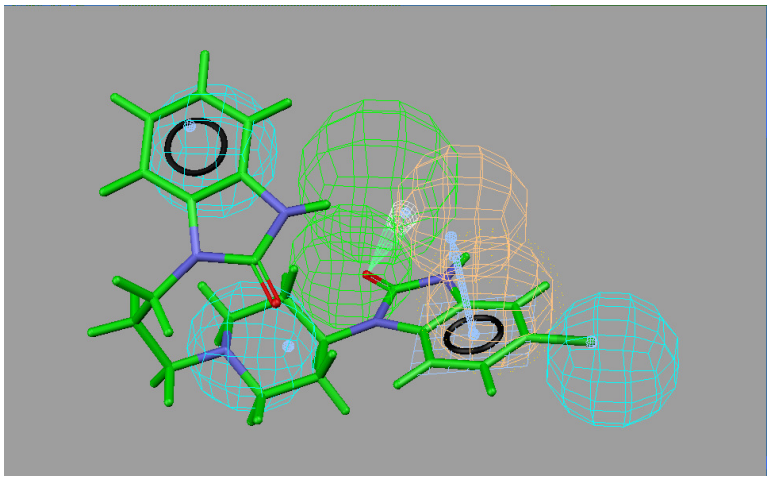

Figure 14. Domperidone aligned with **RHHHAneu**. Note that Domperidone is not a neutral compound.

RHHHAneu (Figure 14) is quite similar to **RHHHneu**, but has an additional H-bond acceptor (A) feature. This extra feature makes it more difficult to fit.

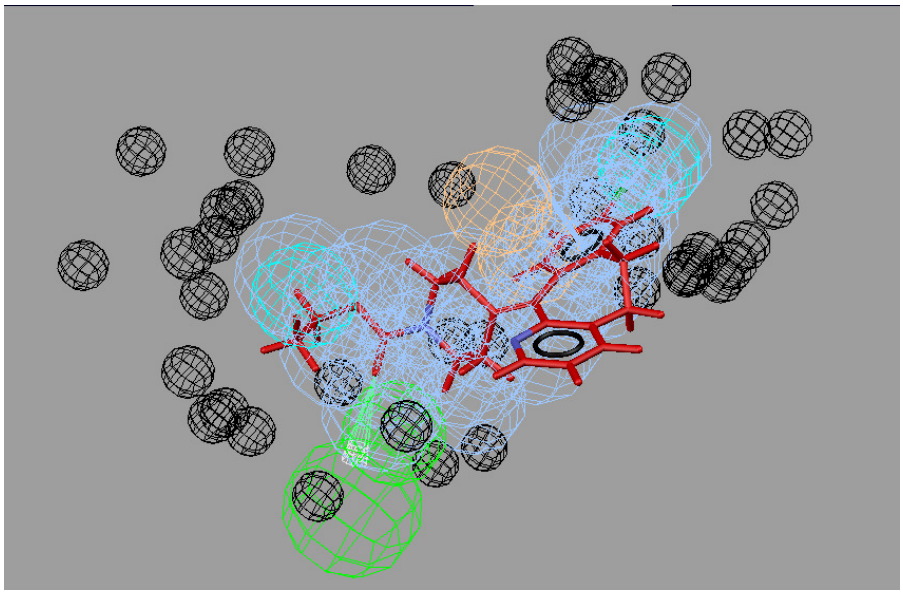

Figure 15. Loratadine aligned with **RHHAnexsh**. The light-blue volume is the shape constraint.

RHHAnexsh (Figure 15) is a complex pharmacophore comprised of two hydrophobic (H), one ring aromatic (R) and one H-bond acceptor features (A), exclusion volumes and a shape restriction. The exclusion volumes block both hydrophobic ends from branching and the shape constraint punishes excursions from the mapping of the highly active AZ-compound which was template for the shape constraint. The SAR behind the exclusions was not as solid as in the central and terminal amine case. The shape restriction slows down screening of compound-databases.

3.2 Classification with PLS-DA

3.2.1 General model

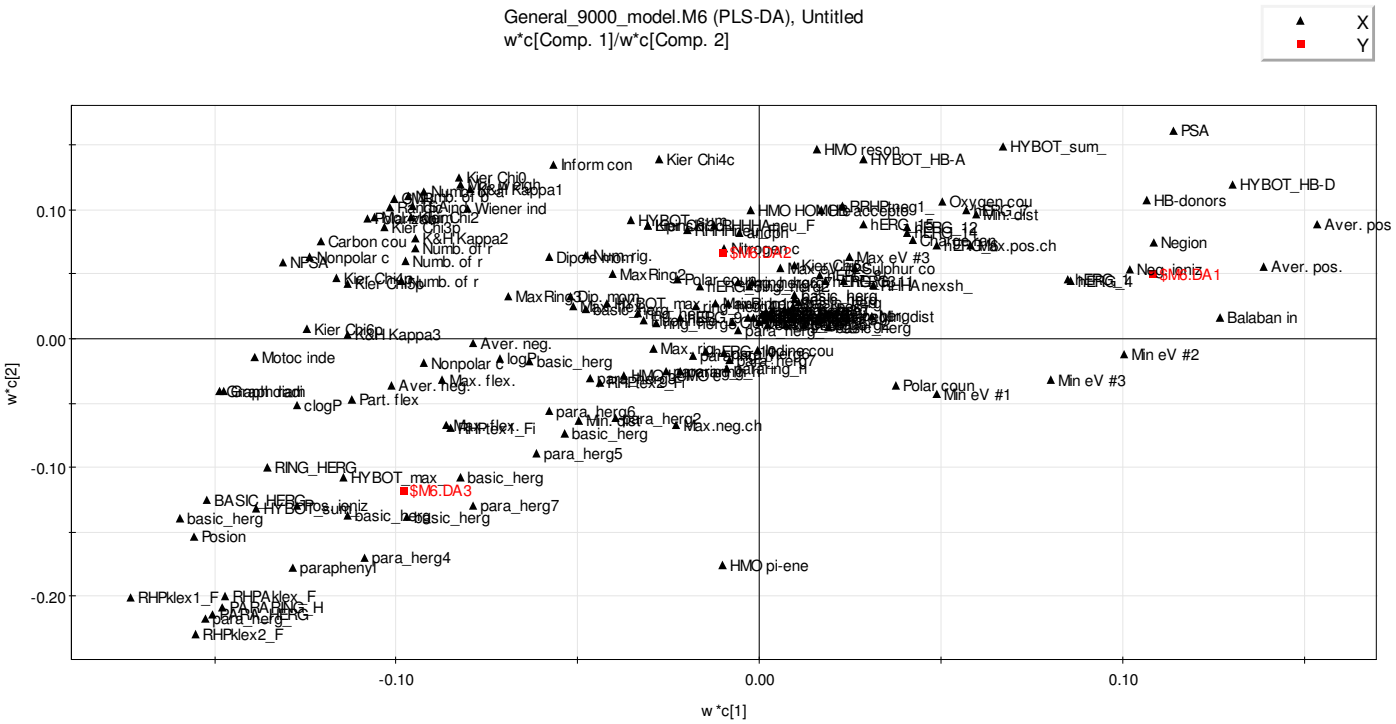

Figure 16. Loading plot for the General model. The class variables are marked in red.

The General model contained 4 components. Most important variables were **RHPklex2**, **RHPklex1**, the Smart para_herg_3 and **RHPAklex**. All correlated positively with pIC_{50} (Figure 16). These variables are in top because central amines, which are in majority among highly active compounds, fit to them. Also in the top are the descriptors **Negion**, **Posion** and the Selma parameters polar surface area (PSA) and clogp. **Negion** and PSA are negatively correlated to hERG pIC_{50} and **Posion** and clogp positively correlated. It has previously been reported that positive ionisable, hydrophobic compounds block hERG channels [20].

3.2.1.1 Test set results

Table 5. General model results on Test set A. Recall for an activity class is the percentage of compounds in that class that is predicted correctly. Precision for an activity class is the percentage of compounds correctly predicted to belong to that class.

		% Correct	66.2		**Predicted**		
Precision (%)	**Recall (%)**			low	med	high	sum
69.3	57.8		low	410	244	55	709
69.4	66.2	**Observed**	med	181	679	165	1025
56.1	83.4		high	1	55	281	337
			sum	592	978	501	2071

Table 6. General model results on Test set B.

		% Correct	59.6		**Predicted**		
Precision (%)	**Recall (%)**			low	med	high	sum
68.1	54.2		low	769	554	95	1418
64.3	62.8	**Observed**	med	358	1048	263	1669
21.8	76.3		high	2	29	100	131
			sum	1129	1631	458	3218

The results on Test set A (Table 5) are generally better than on Test set B (Table 6). This is not surprising since BigPicker was used for division of the original dataset into training set and Test set A. The compounds of Test set A should be within the structural space of the training set. Another reason for the better result is that the proportion of high affinity compounds is lower in Test set B, and the General model is good at predicting the high class compounds.

Recall for an activity class is the percentage of compounds in that class that is predicted correctly. For example recall for the high activity class is the percentage of the highly active compounds that are predicted to be highly active by the classification model. Precision for an activity class is the percentage of compounds correctly predicted to belong to that class.

Two important figures are recall for the high class and precision for the low class. These are good on both test sets. High recall for high activity compounds is important, because then you know that compounds predicted as low or medium are not highly active. These compounds can then be considered as safe or at least possible to develop away from hERG affinity. Development of a compound series from medium to low hERG activity is much more likely to succeed than development from high to low hERG activity. In the latter case such comprehensive structural changes may be needed that affinity to the pharmacological target may be hard to maintain. High precision for the low class is important because then you can trust that the compounds predicted as low are low and not medium or highly hERG active. But high precision for the low class is not as valuable if not the

recall for the low class are high. The recall for the low class is only 54% and 58% respectively for Test set A & B. Especially serious are double-faults, and particularly high class compounds that are predicted as low active. A model that plays it safe and overestimates all activities are not good either, because it will reduce freedom to operate, produce a lot of misclassifications and will not be trusted by the users.

The results of the General model on the central and terminal amine compounds of Test set B (Test set BC and BT) will be presented and discussed in the central and terminal amine model chapters.

Table 7. General model results on neutral compounds in Test set B.

		% Correct	61.7		**Predicted**		
Precision (%)	**Recall (%)**			low	med	high	sum
62.2	64.3		low	440	242	2	684
61.7	59.6	**Observed**	med	267	401	5	673
0.0	0.0		high	0	7	0	7
			sum	707	650	7	1364

Table 8. General model results on acidic compounds in Test set B.

		% Correct	75.5		**Predicted**		
Precision (%)	**Recall (%)**			low	med	high	sum
95.5	88.1		low	148	1	19	168
0.0	0.0	**Observed**	med	7	0	21	28
0.0	0.0		high	0	0	0	0
			sum	155	1	40	196

In Table 7 and 8, the General model results on neutral and acidic compounds in Test set B are presented. The results are not bad, but a problem is that the General model associates the high class with central amines since those are in majority and high class neutral compounds are not predicted correctly. Some acidic compounds are central amine zwitterions and many of those are incorrectly predicted to be highly active. The results on neutrals and acids in the training set and Test set A (not presented here) were very similar.

3.2.2 Central amine model

Figure 17. (a) Loading plot for the Central amine model. The class variables are marked in red. (b) Scatter plot of hERG pIC_{50} vs. clogp for the compounds in the central amine training set.

The Central amine model contained 5 components. Most important variables were the positively correlated clogp, **RHPklex2**, the two Smarts para_herg_3 and PARA_HERG and the negatively correlated Selma parameter HB-donors (Figure 17a). Figure 17b depicts hERG pIC_{50} vs. clogp for the compounds in the central amine training set. There is correlation, but there seems to be an optimum clogp range from 2 to 5. This was also reported previously [20]. Clogp and pharmacophores with exclusion volumes are good descriptors to combine in a prediction model since they span different dimensions of the property space (Figure 17a), but are both strongly correlating with hERG pIC_{50}. Descriptors associated with the low activity class were primarily different measures of polarity and

some smarts (Figure 17a). With these results in mind, an interesting approach for lowering hERG activity for highly active compounds is to substitute them with polar branches that protrude into volumes blocked by exclusion volumes.

3.2.2.1 Test set results

Table 9. Central amine model results on Test set C.

		% Correct	79.3		**Predicted**		
Precision (%)	**Recall (%)**			low	med	high	sum
44.7	85.0		low	34	3	3	40
67.4	54.2	**Observed**	med	34	64	20	118
92.2	88.2		high	8	28	270	306
			sum	76	95	293	464

Table 10. Central amine model results on Test set BC.

		% Correct	57.3		**Predicted**		
Precision (%)	**Recall (%)**			low	med	high	sum
50.4	73.8		low	169	43	17	229
76.1	39.7	**Observed**	med	164	166	88	418
48.3	89.9		high	2	9	98	109
			sum	335	218	203	756

Table 11. General model results on central amines in Test set B, in other words Test set BC.

		% Correct	45.8		**Predicted**		
Precision (%)	**Recall (%)**			low	med	high	sum
64.7	24.0		low	55	115	59	229
60.8	45.7	**Observed**	med	29	191	198	418
28.0	91.7		high	1	8	100	109
			sum	85	314	357	756

The Central amine model results on Test set C and BC again highlights the impact of test set composition on percent correctly predicted compounds. The result for Test set C (Table 9) was 79% correct. The recall for the high and low class was excellent, but the medium class recall was an average 54%. The proportion of high activity compounds in Test set C was large. The number of double-faults was also higher than desired. In Test set BC (Table 10), medium compounds were in majority and that effects the percent correctly predicted. The recalls were also lower on this test set.

It is interesting to compare the performances of the General (Table 11) and Central amine (Table 10) models on Test set BC. The General model associates the low activity class with neutrals and acids and had very low recall for low activity central amines. The activities were generally overestimated and only 46% were predicted correctly. The corresponding figure for the Central amine model was

57%. The big difference between the General and the Central amine model results was that the latter has a much higher recall for low class compounds, 74% compared to 24%.

3.2.3 Terminal amine model

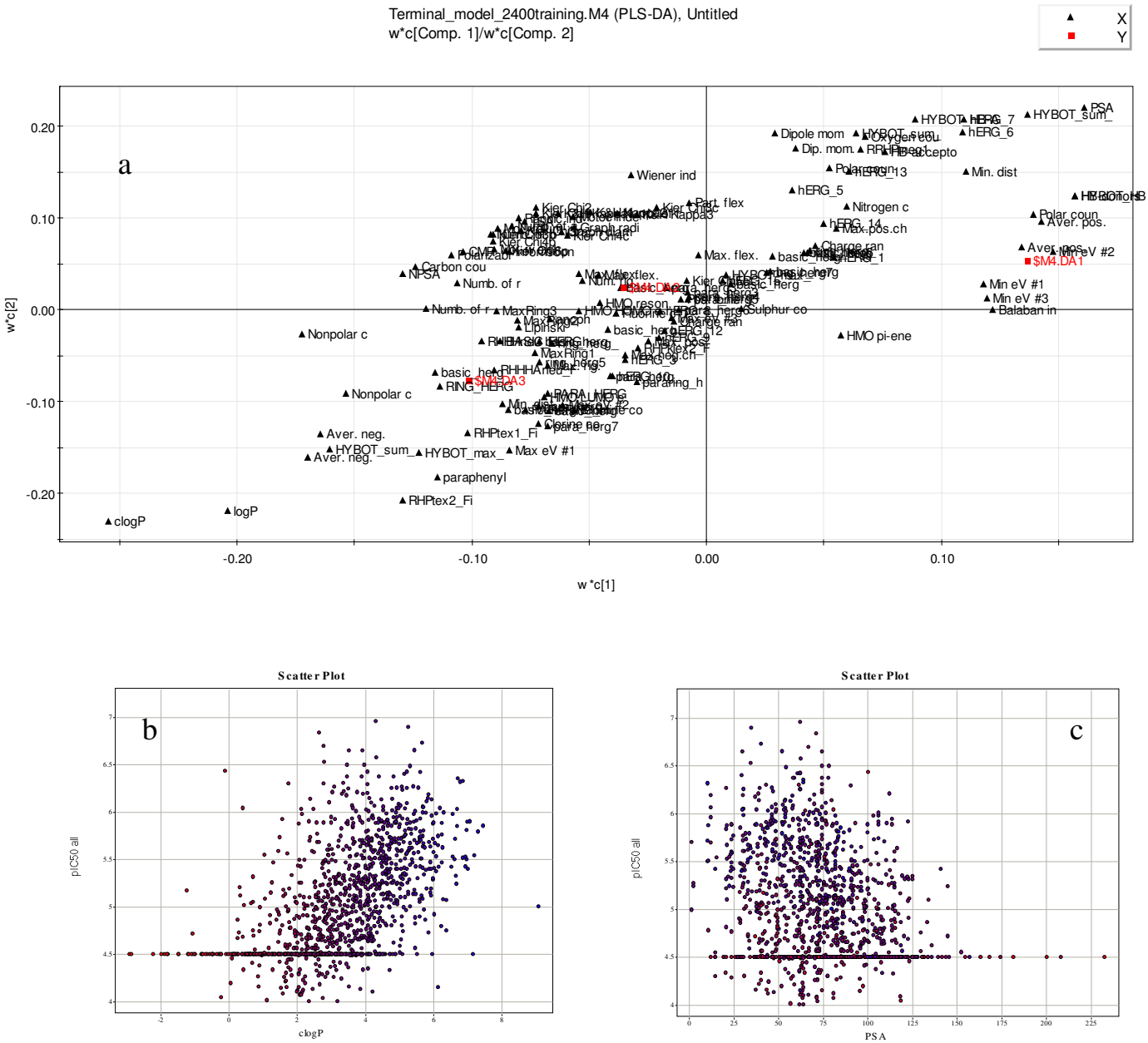

Figure 18. (a) Loading plot for the Terminal amine model. The class variables are marked in red. (b) Scatter plot of hERG pIC_{50} vs. clogp for the compounds in the terminal amine training set. (c) Scatter plot of hERG pIC_{50} vs. PSA for the same compounds. Note that hERG pIC_{50} and PSA is anti-correlated.

The Terminal model contained 5 components. Most important variables were clogp, **RHPtex2** and polar surface area (PSA). The first three were positively correlated with hERG pIC_{50} and PSA was negatively correlated (Figure 18a). Figure 18b is a scatter plot of pIC_{50} vs. clogp for the compounds in

the terminal training set. It is visible in the plot that the correlation is stronger for terminal amines than central amines (Figure 17b). Figure 18c depicts the negative correlation between hERG pIC_{50} and PSA for the same compounds.

3.2.3.1 Test set results

Table 12. Terminal amine model results on Test set T.

		% Correct	50.4		**Predicted**		
Precision (%)	**Recall (%)**			low	med	high	sum
50.7	86.1		low	142	15	8	165
90.3	34.3	**Observed**	med	138	149	147	434
12.9	95.8		high	0	1	23	24
			sum	280	165	178	623

Table 13. Terminal amine model results on Test set BT.

		% Correct	48.0		**Predicted**		
Precision (%)	**Recall (%)**			low	med	high	sum
66.5	64.1		low	216	73	48	337
72.8	37.5	**Observed**	med	109	206	235	550
3.7	73.3		high	0	4	11	15
			sum	325	283	294	902

Table 14. General model results on terminal amines in Test set B, in other words Test set BT.

		% Correct	64.5		**Predicted**		
Precision (%)	**Recall (%)**			low	med	high	sum
69.2	37.4		low	126	196	15	337
68.5	82.9	**Observed**	med	55	456	39	550
0.0	0.0		high	1	14	0	15
			sum	182	666	54	902

The results of the Terminal amine model on Test set T (Table 12) and BT (Table 13) shows a similar pattern as the Central amine model performance. Recalls were high or acceptable for the low and high class. Since the number of high activity compounds in the test sets was low and mediums were in majority, the percent correctly predicted were only about 50%. The model can discriminate between highs and lows much better than the General model (Table 14) that predicts almost all terminal amines to be medium active. Since most terminal amines are medium active, 65% of Test set BT was predicted correctly.

3.3 Additional models

Development of Neutral and Acid models was not prioritized due to lack of time and possible weighting problems arising from the low number of high activity compounds among the neutral and acidic compounds.

3.3.1 PLS

PLS-models were also developed in Simca-P+ v.10.0.2.0 using the same training sets as in the PLS-DA models. The general performance of PLS in hERG activity classification was that recall for the medium class was excellent, around 80-90%, for the high class it was around 50-70% and for the low activity class it was mediocre - around 20%. One reason for these results is that PLS is a regression and not a classification method. Another reason for the poor recall for low class compounds is that pIC50 were set to 4.5 for compounds whose hERG activity were not measurable in the assay. 4.5 is the defined limit between the poor and the medium class. If a compound with a pIC_{50} of 4.5 is predicted to 4.51 it will be misclassified, but the numerical error will be very small.

PLS predictions may still be useful as additional information in a webtool. Results not presented here show that high activity compounds that are classified as medium by PLS-DA are in over 90% of the cases predicted by PLS to have a $pIC_{50} \geq 5$. Thus, a compound that is classified as medium and have a PLS-predicted $pIC_{50} < 5$ is highly unlikely to be highly active.

3.3.2 PLS-DA with two classes

PLS-DA models that classifies compounds as more or less active than $pIC_{50}=5$ were also generated. Results are not presented in this report. The percent correct predictions was high, usually 70-80%. Since the task for these classifiers were simpler than for the 3-class PLS-DA models, this was no surprise. The problem is that there is no medium class separating the high and low activity class. The high class is then not as well separated from the low class by the model and the risk for double-faults is larger.

3.3.3 RDS

Non-linear modelling was also performed with RDS (Rule Discovery System) [23] by my supervisor at AstraZeneca. Preliminary results were very good almost meeting the 80% aim, with additional conditions such as high recall for high and low active compounds also fulfilled. More extensive hERG modelling will be performed by computational chemists at AstraZeneca, using RDS, the new pharmacophores and the molecule class separation system developed in this work.

4 Conclusions

Eleven new pharmacophores for hERG binding were generated. They were developed for three different molecule classes – central amines, terminal amines and neutral compounds. They give good enrichment and provide visual feedback to chemists trying to avoid hERG active compounds in lead development. By using these pharmacophores and the scheme presented above (Figure 4), compounds can now be automatically sorted into four molecule classes – the three mentioned above and acids. Since these classes have distinctly different distributions of hERG pIC_{50}, this is a practical way of dividing the dataset into subsets. Due to the automated sorting, these subsets can easily be treated and analysed separately or in groups. This may be very useful. For example the usually low active acids can be filtered out from the other three molecule classes and, maybe with some reservation, be predicted as low active. Then a model might be generated for the three remaining molecule classes or the highly active central amines may be filtered out from the other two classes and predicted separately. There are many different approaches and subsettings to investigate. Many of the different approaches of hERG modelling using subsets were not tested properly in this work, due to lack of time and because the software RDS was not available for use in this study.

Some conclusions may be drawn from the PLS-DA classification modelling performed in this work. When the training sets are evenly weighted as in this case, Central amine and Terminal amine models are better than a General model in separating high and low activity compounds within their respective molecule class, but may produce a larger number of misclassifications. This because they have lower recall for the medium activity class, which most compounds in the dataset belong to. Most terminal amines are medium active and if a separate model is to be used for these, it should be carefully weighted or a non-linear modelling method should be used.

The General model generalises and associates acids with the low activity class, neutrals with the low or medium activity class, terminal amines with the medium activity class and central amines with the medium or high activity class. This is generally true and that is important, but the General model will not recognise the highly active neutrals or terminal amines or the low activity central amines that exist. One important reason for these generalisations is that the molecule classes are not evenly weighted in each activity class in the General model training set. To perform this kind of weighting in Simca is very time-demanding and was therefore not executed. One argument for separate models for separate molecule classes is that pharmacophores for the terminal amines are negative for central amines. Also, if the different molecule classes bind to different binding sites in hERG, separate modelling may be most appropriate since the SAR will be different. In the literature, there is very little published on

hERG binding for neutral compounds, but for amines many research groups agree on the position of the basic nitrogen of a compound when it binds to hERG [17, 20, 24]

Looking forward this work opens up for several new approaches for hERG modelling. The secondary aim of this work – to generate a classification model that could achieve 80% accuracy in prediction – was not met, but there is hope for the future. RDS classification models will definitely give higher recalls. The aim should also be rephrased to at least 80% recall for all activity classes. Now there are eleven new pharmacophores to use as descriptors and this may also have impact on model performance. Though test model results not presented here suggests that a lot of this structural information is already covered by the Smarts. Subset modelling will be very interesting and may improve results. Already in this work, separate central amine modelling was shown to be rewarding.

5 Acknowledgements

I would like to thank my supervisor Mats Svensson for giving me excellent support throughout the project, Ulf Norinder for great additional support and advises and Markus Haeberlein for giving me the opportunity to carry out my degree project at AstraZeneca R&D Södertälje. I would also like to thank Johan Åqvist for reviewing my report and the people in the Computational Medicinal Chemistry group at AZ for advises and great coffee breaks. Finally I would like to thank all the people at the AZ Chemistry Department for being so nice and making my stay so enjoyable.

6 References

[1] Keating, M.T. and Sanguinetti, M.C. (2001) "Molecular and cellular mechanisms of cardiac arrhythmias", *Cell* **104**, 569-580.

[2] Viskin, S. (1999) "Long QT syndromes and torsade de pointes", *Lancet* **354**, 1625-1633.

[3] Cavalli, A., Poluzzi, E., De Ponti, F. and Recanatini, M. (2002) "Toward a pharmacophore for drugs inducing the long QT syndrome: insights from a CoMFA study of HERG K^+ channel blockers", *J. Med Chem.* **45**, 3844-3853.

[4] Song, M. and Clark, M. (2006) "Development and evaluation of an in silico model for HERG binding", *J. Chem Inf Model* **46**, 392-400.

[5] European Agency for the Evaluation of Medicinal Products (EMEA) (2005), "ICH S7B - The nonclinical evaluation of the potential for delayed ventricular repolarization (QT interval prolongation) by human pharmaceuticals"

[6] Pearlstein, R., Vaz, R. and Rampe, D. (2003) "Understanding the structure-activity relationship of the human ether-a-go-go-related gene cardiac K^+ channel. A model for bad behavior", *J. Med Chem.* **46**, 2017-2022.

[7] Sanguinetti, M.C., Jiang, C., Curran, M.E. and Keating, M.T. (1995) "A mechanistic link between an inherited and an acquired cardiac arrhythmia: HERG encodes the I_{Kr} potassium channel", *Cell* **81**, 299-307.

[8] Aronov, A.M. and Goldman B.B. (2004) "A model for identifying HERG K+ channel blockers", *Bioorg Med Chem.* **12**, 2307-2315.

[9] http://www.neurionpharma.com, Neurion Pharmaceuticals Inc., Pasadena, CA, USA (17 Jan. 2006).

[10] Eriksson, L., Johansson, E., Kettaneh-Wold, N. and Wold, S. Multi- and Megavariate Data Analysis, *Umetrics* (2001)

[11] AstraZeneca in house software

[12] Olsson, T. and Sherbukhin, V. (1999) Synthesis and Structure Administration (SaSA), AstraZeneca R&D Mölndal, Sweden

[13] Strandlund, G., Blomberg, N., Hasselgren Arnby, C. and Gavaghan McKee, C., AstraZeneca R&D Mölndal.

[14] http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (24 Jan. 2006).

[15] Catalyst 4.11 User Guide and On-line Help, Accelrys Inc., San Diego, CA, USA, 2005

[16] Ekins, S., Crumb, W.J., Sarazan, R.D., Wikel, J.H. and Wrighton, S.A. (2002) "Three-dimensional quantitative structure-activity relationship for inhibition of human ether-a-go-go-related gene potassium channel", *J. Pharmacol Exp Ther.* **301**, 427-434.

[17] Pearlstein, R.A., Vaz, R.J., Kang, J., Chen, X.L., Preobrazhenskaya, M., Shchekotikhin, A.E., Korolev, A.M., Lysenkova, L.N., Miroshnikova, O.V., Hendrix, J. and Rampe, D. (2003) "Characterization of HERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches", *Bioorg Med Chem Lett.* **13**, 1829-1835.

[18] Spotfire Inc., Cambridge, MA, USA

[19] Norinder, U., AstraZeneca R&D Södertälje

[20] Aronov, A.M. (2005) "Predictive in silico modeling for hERG channel blockers", *Drug Discov. Today* **10**, 149-155.

[21] Österberg, F. and Åqvist, J. (2005) "Exploring blocker binding to a homology model of the open hERG K+ channel using docking and molecular dynamics methods", *FEBS Lett.* **579**, 2939-2944.

[22] Perry, M., de Groot, M.J., Helliwell, R., Leishman, D., Tristani-Firouzi, M., Sanguinetti, M.C. and Mitcheson, J. (2004) "Structural determinants of HERG channel block by clofilium and ibutilide", *Mol Pharmacol.* **66**, 240-249.

[23] http://www.compumine.com, Compumine AB, Uppsala (25 Jan. 2006).

[24] Fernandez, D., Ghanta, A., Kauffman, G.W. and Sanguinetti, M.C. (2004) "Physicochemical features of the HERG channel drug binding site", *J. Biol Chem.* **279**, 10120-10127.

7 Appendix

7.1 Distribution of activity classes and molecule classes in General model datasets

Original set

		Central amines	Terminal amines	Neutrals	Acids	Sum
High	$pIC_{50} \geq 6$	706	104	25	2	837
Medium	$4.5 \leq pIC_{50} \leq 6$	1318	1234	1344	129	4025
Low	$pIC_{50} \leq 4.5$	240	565	964	440	2209
Sum		2264	1903	2333	571	7071

Training set

		Central amines	Terminal amines	Neutrals	Acids	Sum
High	$pIC_{50} \geq 6$	411*6	73*6	14*6	2*6	3000
Medium	$4.5 \leq pIC_{50} \leq 6$	994	919	991	96	3000
Low	$pIC_{50} \leq 4.5$	166*2	337*2	651*2	346*2	3000
Sum		3792	2031	2377	800	9000

Test set A

		Central amines	Terminal amines	Neutrals	Acids	Sum
High	$pIC_{50} \geq 6$	295	31	11	0	337
Medium	$4.5 \leq pIC_{50} \leq 6$	324	315	353	33	1025
Low	$pIC_{50} \leq 4.5$	74	228	313	94	709
Sum		693	574	677	127	2071

Test set B

		Central amines	Terminal amines	Neutrals	Acids	Sum
High	$pIC_{50} \geq 6$	109	15	7	0	131
Medium	$4.5 \leq pIC_{50} \leq 6$	418	550	673	28	1669
Low	$pIC_{50} \leq 4.5$	229	337	684	168	1418
Sum		756	902	1364	196	3218